machine learning & knowledge extraction

MDPI

Review

Entropic Statistics: Concept, Estimation, and Application in Machine Learning and Knowledge Extraction

Jialin Zhang

Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS 39762, USA; jzhang@math.msstate.edu

Abstract: The demands for machine learning and knowledge extraction methods have been booming due to the unprecedented surge in data volume and data quality. Nevertheless, challenges arise amid the emerging data complexity as significant chunks of information and knowledge lie within the non-ordinal realm of data. To address the challenges, researchers developed considerable machine learning and knowledge extraction methods regarding various domain-specific challenges. To characterize and extract information from non-ordinal data, all the developed methods pointed to the subject of Information Theory, established following Shannon's landmark paper in 1948. This article reviews recent developments in entropic statistics, including estimation of Shannon's entropy and its functionals (such as mutual information and Kullback–Leibler divergence), concepts of entropic basis, generalized Shannon's entropy (and its functionals), and their estimations and potential applications in machine learning and knowledge extraction. With the knowledge of recent development in entropic statistics, researchers can customize existing machine learning and knowledge extraction methods for better performance or develop new approaches to address emerging domain-specific challenges.

Keywords: discrete data; non-ordinal data; non-parametric estimation; entropic statistics; information-theoretic quantity

Citation: Zhang, J. Entropic Statistics: Concept, Estimation, and Application in Machine Learning and Knowledge Extraction. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 865–887. https://doi.org/10.3390/make4040044

Academic Editor: Andreas Holzinger

Received: 29 August 2022
Accepted: 26 September 2022
Published: 30 September 2022

1. Introduction of Entropic Statistics

Entropic statistics is a collection of statistical procedures that characterize information from non-ordinal spaces with Shannon's entropy and its generalized functionals. Such procedures includes but not limited to statistical methods involving Shannon's entropy (entropy) and Mutual Information (MI) [1], Kullback–Leibler divergence (KL) [2], entropic basis and diversity index [3,4], and Generalized Shannon's Entropy (GSE) and Generalized Mutual Information (GMI) [5]. The field of entropic statistics is at the intersection of information theory and statistics. Entropic statistics quantities are also referred as information-theoretic quantities [6,7].

There are two general data types—ordinal and non-ordinal (nominal). Ordinal data are data with an inherent numerical scale. For example, {52 F, 50 F, 49 F, 53 F}—a set of daily high temperatures at Nuuk, Greenland—is ordinal. Ordinal data are generated from random variables (which map outcomes from sample space to the real numbers). For ordinal data, classical concepts, such as moments (mean, variance, covariance, etc.) and characteristics functions, are powerful tools to induce various statistical methods, including but not limited to regression analysis [8] and analysis of variance (ANOVA) [9].

Non-ordinal data are data without an inherent numerical scale. For example, {androgen receptor, clock circadian regulator, epidermal growth factor, Werner syndrome RecQ helicase-like}—a subset of human genes names—is a set of data without inherent numerical scale. Non-ordinal data are generated from random elements (which map outcomes from sample space to alphabet). Due to the absence of inherent numerical scale, the concept of random variable is undefined according to its definition. Therefore, statistical concepts involving ordinal scale (e.g., mean, variance, covariance, and characteristic functions) no

longer exist. For example, consider the mentioned data of human genes names; what is the mean or variance of the data? Such questions cannot be answered because the concepts of mean and variance do not exist, while in practice, researchers need to measure the level of dependence in non-ordinal joint space between gene types and genetic phenotype to study the gene's functionalities. One would use covariance and its generated methods in ordinal data. However, the concept of covariance no longer exists in such non-ordinal space. Furthermore, all well-established statistical methods that require ordinal scale (e.g., regression and ANOVA) cannot be directly applied anymore.

Non-ordinal data have several variant names, such as categorical data, qualitative data, and nominal data. A common situation is a dataset is mixed with ordinal and non-ordinal data. On such a dataset, a common practice is to introduce coded (dummy) variables [10]. However, introducing dummy variables is equivalent to separating the mixed dataset according to the classes in non-ordinal variables to induce multiple purely ordinal subsets and then utilizing ordinal methods (such as regression analysis) case-by-case on the induced subsets. Unfortunately, this approach sometimes could be impractical because of the curse of dimensionality, particularly when there are too many categorical variables or when some categorical variable has too many categories (classes).

With the challenges from non-ordinal data, entropic statistics methods focus on underlying probability distribution instead of associated labels. As a result, all the entropic statistical quantities are location (permutation) invariant. The main strengths of entropic statistics lie within non-ordinal alphabets, or a mixture data space that significant bulk of information lies within the non-ordinal sub-space. For ordinal spaces, although ordinal variables can be binned as categorical variables, the strength of entropic statistics are generally incapable of overcoming the loss of ordinal information during discretization. Therefore, ordinal statistical methods are preferred when they are capable of the needs. In summary, potential scenarios for entropic statistics are:

1. The data lie within non-ordinal space.
2. The data are a mixture of ordinal and non-ordinal spaces, and the non-ordinal space is expected to carry unneglectable bulk of information.
3. The data lie within ordinal space, yet the performance of ordinal statistics methods fails to meet the expectation.

The following notations are used throughout the article. They are listed here for convenience.

1. Let $\mathcal{X} = \{x_i; i = 1, 2, \cdots\}$ and $\mathcal{Y} = \{y_j; j = 1, 2, \cdots\}$ be two countable alphabets with cardinalities $K_1 \leq \infty$ and $K_2 \leq \infty$, respectively.
2. Let the Cartesian product $\mathcal{X} \times \mathcal{Y}$ be with a joint probability distribution $\mathbf{p}_{XY} = \{p_{i,j}\}$.
3. Let the two marginal distributions be respectively denoted by $\mathbf{p}_X = \{p_{i,\cdot}\}$ and $\mathbf{p}_Y = \{p_{\cdot,j}\}$ where $p_{i,\cdot} = \sum_j p_{i,j}$ and $p_{\cdot,j} = \sum_i p_{i,j}$; hence X is a variable on $\mathcal{X}$ with distribution $\mathbf{p}_X$ and Y is a variable on $\mathcal{Y}$ with distribution $\mathbf{p}_Y$.
4. For uni-variate situations, K stands for K_1, and $\mathbf{p}$ stands for $\mathbf{p}_X$.
5. Let $\{X_1, X_2, \cdots, X_n\}$ be an independent and identically distributed (*i.i.d.*) random sample of size n from $\mathcal{X}$. Let $C_r = \sum_{i=1}^{n} 1[X_i = l_r]$; hence C_r is the count of occurrence of letter l_r in a sample. Let $\hat{\mathbf{p}} = \{\hat{p}_1, \hat{p}_2, \cdots\} = \{C_1/n, C_2/n, \cdots\}$. $\hat{\mathbf{p}}$ is called the plug-in estimator of $\mathbf{p}$. Similarly, one can construct the plug-in estimators for $\mathbf{p}_Y$ and $\mathbf{p}_{XY}$ and name them as $\hat{\mathbf{p}}_Y$ and $\hat{\mathbf{p}}_{XY}$, respectively.
6. For any two functions f and g taking values in $(0, \infty)$ with $\lim_{n\to\infty} f(n) = \lim_{n\to\infty} g(n) = 0$, the notation $f(n) = \mathcal{O}(g(n))$ means

$$0 < \liminf_{n\to\infty} \frac{g(n)}{f(n)} \leq \limsup_{n\to\infty} \frac{g(n)}{f(n)} < \infty.$$

7. For any two functions f and g taking values in $(0, \infty)$, the notation $f(n) = \mathcal{O}(g(n))$ means
$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

Many concepts discussed in the following sections have continuous counterparts under the same concept name. The results reviewed in this article focus on non-ordinal data space. Therefore, some notable results on ordinal space are not reviewed (for example, [11–13]). In Section 2, estimation on some classic entropic statistics quantities are discussed. Section 3 reviews estimation results and properties for some recently developed information-theoretic quantities. Entropic statistics' application potentials in machine learning (ML) and knowledge extraction are discussed in Section 4. Finally, some remarks are given in Section 5.

2. Classic Entropic Statistics Quantities and Estimation

This section reviews three classic entropic concepts and their estimations, including Shannon's entropy (Section 2.1.1) and mutual information (Section 2.1.2), and Kullback–Leibler divergence (Section 2.2). These three concepts are among the earliest entropic concepts and have been intensively studied over the past decades. Enormous amounts of statistical methods and computational algorithms are designed based on these three concepts [14–16]. Nevertheless, most of those methods and algorithms use naive plug-in estimation, which could be improved for a smaller estimation bias and better performance. For this reason, this section reviews several notable estimation methods as a reference. Some asymptotic properties are also presented as a reference. The asymptotic properties provide a theoretical guarantee for the corresponding estimators with statistical procedures such as hypothesis testing and confidence intervals.

2.1. Shannon's Entropy and Mutual Information

2.1.1. Shannon's Entropy

Established by Shannon in his landmark paper [1], the concept of entropy is the first and still the most important building brick in characterizing information from non-ordinal spaces. Many of the established information-theoretic quantities are linear functions of entropy. Shannon's entropy, H, is defined as
$$H = -\sum_i p_i \ln p_i.$$

Some remarkable properties of entropy are:

Property 1 (Entropy)**.**
1. *H is a measurement of dispersion. It is always non-negative by definition.*
2. *$H = 0$ if and only if the probability of a letter l in $\mathcal{X}$ is 1; hence no dispersion.*
3. *For a finite alphabet with cardinality K, H is bounded from the above by $\ln K$, and the maximum is achieved when its distribution is uniform ($p_i = 1/K, i = 1, 2, \cdots, K$); hence maximum dispersion.*
4. *For a countably infinite alphabet, H may not exist (See Example 4 in Section 3).*

Entropy Estimation-The Plug-in Estimator

Estimation of entropy has been a core research topic for decades. Due to the curse of "High Dimensionality" and "Discrete and Non-ordinal Nature", entropy estimation is a technically difficult problem. Advances in this area have been slow to come. The plug-in estimator of entropy (also known as empirical entropy estimator), $\hat{H}$, defined as
$$\hat{H} = -\sum_i \hat{p}_i \ln \hat{p}_i,$$

is inarguably the most naive entropy estimator. $\hat{H}$ has been studied thoroughly in recent decades. Ref. [17] provided the asymptotic properties for $\hat{H}$ when K is finite, namely,

Theorem 1 (Asymptotic property of $\hat{H}$ when K is finite)**.**

$$\sqrt{n}\left(\hat{H}-H\right)/\hat{\sigma} \xrightarrow{D} N(0,1),$$

where $\hat{\sigma}=\sqrt{\sum_i \hat{p}_i \ln^2 \hat{p}_i - \hat{H}^2}$.

Ref. [18] derived the bias of $\hat{H}$ for finite K

$$\mathrm{E}(\hat{H})-H=-\frac{K-1}{2n}+\frac{1}{12n^2}\left(1-\sum_{i=1}^{K}\frac{1}{p_i}\right)+\mathcal{O}\left(n^{-3}\right). \tag{1}$$

Ref. [19] derived the asymptotic properties for $\hat{H}$ when K is countable infinite. Namely,

Theorem 2 (Asymptotic property of $\hat{H}$ when K is countable infinite)**.** *For any nonuniform distribution* $\{p_i; i \geq 1\}$ *satisfying* $\sum_i\left(p_i \ln^2 p_i\right) < \infty$*, if there exists an integer-valued function* $K(n)$ *such that, as* $n \to \infty$*,*

1. $K(n) \to \infty$,
2. $K(n) = o(\sqrt{n})$*, and*
3. $\sqrt{n}\sum_{i \geq K(n)} p_i \ln p_i \to 0$;

then

$$\sqrt{n}\left(\hat{H}-H\right)/\hat{\sigma} \xrightarrow{D} N(0,1),$$

where $\hat{\sigma}=\sqrt{\sum_k \hat{p}_k \ln^2 \hat{p}_k - \hat{H}^2}$.

As discussed in [19], the conditions with $K(n)$ hold if $p_i \sim 1/\left(i^2 \ln^2 i\right)$; the conditions do not hold if $p_i \sim 1/\left(i^2 \ln i\right)$.

Entropy Estimation-The Miller–Madow and Jackknife Estimators

$\hat{H}_{MM}$ [20] and $\hat{H}_{JK}$ [21] are two notable entropy estimators with bias adjustments. Namely,

$$\hat{H}_{MM}=\hat{H}+\frac{\hat{K}-1}{2n}, \tag{2}$$

where $\hat{K}$ is the observed sample cardinality. For finite K, the bias of $\hat{H}_{MM}$ is

$$\mathrm{E}(\hat{H}_{MM})-H=\mathcal{O}\left(n^{-2}\right).$$

$\hat{H}_{JK}$ is calculated in three steps:

1. for each $i \in \{1,2,\ldots,n\}$, construct $\hat{H}^{(i)}$, which is a plug-in estimator based on a sub-sample of size $n-1$ obtained by leaving the ith observation out;
2. obtain $\hat{H}_{(i)}=n\hat{H}-(n-1)\hat{H}^{(i)}$ for $i=1,\cdots,n$; and then
3. compute the jackknife estimator

$$\hat{H}_{JK}=\frac{\sum_{i=1}^{n}\hat{H}_{(i)}}{n}. \tag{3}$$

Equivalently, (3) can be written as

$$\hat{H}_{\mathbb{JK}} = n\hat{H} - (n-1)\frac{\sum_{i=1}^{n}\hat{H}^{(i)}}{n}.$$

When $K < \infty$, it can be shown that the bias of $\hat{H}_{\mathbb{JK}}$ is

$$\mathrm{E}(\hat{H}_{\mathbb{JK}}) - H = \mathcal{O}\left(n^{-2}\right).$$

Asymptotic properties for $\hat{H}_{MM}$ and $\hat{H}_{\mathbb{JK}}$ were derived in [22]. $\hat{H}_{MM}$ and $\hat{H}_{\mathbb{JK}}$ reduce the rate of bias to a higher order power-decaying. Ref. [23] proved the convergence of $\hat{H}$ could be arbitrarily slow. Ref. [24] proved that for finite K, an unbiased estimator for entropy does not exist. As a result, it is only possible to reduce the bias to a smaller extent.

Entropy Estimation-The Z-Estimator

Recent studies on entropy estimation have reduced the bias to exponentially decaying. For example,

$$\hat{H}_z = \sum_{v=1}^{n-1}\left\{\frac{1}{v}\frac{n^{1+v}[n-(1+v)]!}{n!}\sum_i\left[\hat{p}_i\prod_{j=0}^{v-1}\left(1-\hat{p}_i-\frac{j}{n}\right)\right]\right\}$$

is the entropy estimator provided in [25] with an exponentially decaying bias (Interested readers may refer to [26] for discussion on an entropy estimator that is algebraically equivalent to $\hat{H}_z$). Ref. [27] derived the asymptotic properties for $\hat{H}_z$. Namely,

Theorem 3 (Asymptotic property of $\hat{H}_z$ when K is finite)**.**

$$\sqrt{n}(\hat{H}_z - H)/\hat{\sigma} \xrightarrow{D} N(0,1),$$

where $\hat{\sigma} = \sqrt{\sum_i \hat{p}_i \ln^2 \hat{p}_i - \hat{H}^2}$.

The following asymptotic properties for $\hat{H}_z$ when K is countable infinite were provided in [28].

Theorem 4 (Asymptotic property of $\hat{H}_z$ when K is countable infinite)**.** *For a nonuniform distribution* $\{p_i; i \geq 1\} \in \mathscr{P}$ *satisfying* $\sum_i\left(p_i \ln^2 p_i\right) < \infty$*, if there exists an integer-valued function* $K(n)$ *such that, as* $n \to \infty$*,*

1. $K(n) \to \infty$,
2. $K(n) = o(\sqrt{n}/\ln n)$*, and*
3. $\sqrt{n}\sum_{i \geq K(n)} p_i \ln p_i \to 0$*;*

then

$$\sqrt{n}(\hat{H}_z - H)/\hat{\sigma} \xrightarrow{D} N(0,1),$$

where $\hat{\sigma} = \sqrt{\hat{H}_{2z} - \hat{H}_z^2}$ *and* $\hat{H}_{2z} = \sum_i\left\{\hat{p}_i\sum_{v=1}^{n-n\hat{p}_i}\left[\left(\sum_{s=1}^{v-1}\frac{1}{s(v-s)}\right)\prod_{j=1}^{v}\left(1-\frac{n\hat{p}_i-1}{n-j}\right)\right]\right\}$.

The sufficient condition given in Theorem 4 for the normality of $\hat{H}_z$ is slightly more restrictive than that of the plug-in estimator $\hat{H}$ as stated in Theorem 2, and consequently supports a smaller class of distributions. The sufficient conditions of Theorem 4 still holds for $p_i = C_\lambda i^{-\lambda}$ where $\lambda > 2$, but not for $p_i = C/\left(i^2 \ln^2 i\right)$, which satisfies the sufficient conditions of Theorem 2. However, it is discussed in [28] that simulation results indicate that the asymptotic normality of $\hat{H}_z$ in Theorem 4 may still hold for $p_i = C/\left(i^2 \ln^2 i\right)$ for $i \geq 1$ though not covered by the sufficient condition.

Remarks

Another perspective of entropy estimation is to combine $\hat{H}_z$ and $\hat{H}_{JK}$. Namely, one could use $\hat{H}_z$ in place of each $\hat{H}_{(i)}$ in (3). Interested readers may refer to [25] where a single layer combination of $\hat{H}_z$ and $\hat{H}_{JK}$ was discussed. In addition, ref. [29] presented a non-parametric entropy estimator ($\hat{H}_{chao}$) when there are unseen species in the sample. $\hat{H}_{chao}$ has a smaller sample root mean squared error than $\hat{H}_{MM}$ and a smaller bias than $\hat{H}_{JK}$, according to the simulation study. Unfortunately, the bias decaying rate for $\hat{H}_{chao}$ was not theoretically offered. Based on their simulation study of $\hat{H}_{chao}$, it seems that the bias decaying rate is $\mathcal{O}(1/n^2)$, which is slower than $\hat{H}_z$. Asymptotic properties of $\hat{H}_{chao}$ are not developed in the literature.

There are several parametric entropy estimators for specific interests. For example, Dirichlet prior Bayesian estimator of entropy [30,31] and shrinkage estimator of entropy [32]. This review article focuses on results from non-parametric estimation methods. To conclude this section, a small scale comparison between $\hat{H}$ and $\hat{H}_z$ from [33] is provided in Table 1.

Table 1. Estimation comparison between $\hat{H}$ and $\hat{H}_z$ via simulation. In the simulation, the real underlying distribution is $p_i = i/2001000$, where $i = 1, 2, \ldots, 2000$ (i.e., a triangle distribution). Under this setting, the true entropy $H = 7.408005$. To compare the two estimators, 10,000 samples were independently generated following the triangle distribution for each of the six sample size settings in the table (i.e., we generate 60,000 random samples in total). The average values of $\hat{H}$ and $\hat{H}_z$ under different sample sizes are summarized and reported in the table. The simulation shows that $\hat{H}$ would consistently underestimate H more than $\hat{H}_z$. The underestimation is more severe when the sample size is smaller.

n	**100**	**300**	**500**	**1000**	**1500**	**2000**
avg. of $\hat{H}$	4.56	5.57	6.00	6.51	6.75	6.89
avg. of $\hat{H}_z$	5.11	6.09	6.49	6.92	7.11	7.21

2.1.2. Mutual Information

In the same paper defining Shannon's entropy, the concept of Mutual Information (MI) was also described [1]. Shannon's entropies for $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$ are defined as

$$H(X) = -\sum_i p_{i,\cdot} \ln p_{i,\cdot},$$
$$H(Y) = -\sum_j p_{\cdot,j} \ln p_{\cdot,j},$$
$$H(X,Y) = -\sum_i \sum_j p_{i,j} \ln p_{i,j},$$

and MI between $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$MI(X,Y) = H(X) + H(Y) - H(X,Y).$$

Some notable properties of MI are:

Property 2 (Mutual Information)**.**

1. *MI is a measurement of dependence. It is always non-negative by definition.*
2. *$MI = 0$ if and only if the two marginals are independent.*
3. *$MI > 0$ if and only if the two marginals are dependent.*
4. *A non-zero MI does not always indicate the degree (level) of dependence.*
5. *MI may not exist when the cardinality of joint space is countably infinite.*

MI Estimation-The Plug-in Estimator and Z-Estimator

Since MI is a function of entropy, estimation of MI is essentially entropy estimation. Let $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ be an *i.i.d.* random sample of size n from the joint alphabet $(\mathcal{X}, \mathcal{Y})$. Based on the sample, plug-in estimators of the component entropy of MI can be obtained. Namely,

$$\begin{aligned}\hat{H}(X) &= -\sum_i \hat{p}_{i,\cdot} \ln \hat{p}_{i,\cdot},\\ \hat{H}(Y) &= -\sum_j \hat{p}_{\cdot,j} \ln \hat{p}_{\cdot,j},\\ \hat{H}(X,Y) &= -\sum_i \sum_j \hat{p}_{i,j} \ln \hat{p}_{i,j},\end{aligned}$$

where $\hat{p}_{i,\cdot}$ is the plug-in estimator for $p_{i,\cdot}$, $\hat{p}_{\cdot,j}$ is the plug-in estimator for $p_{\cdot,j}$, and $\hat{p}_{i,j}$ is the plug-in estimator for $p_{i,j}$. Then the plug-in estimator of mutual information between $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$\widehat{MI}(X,Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X,Y).$$

With various entropy estimation methods, one could estimate MI by replacing $\hat{H}$ with a different entropy estimator. For example, using the entropy estimator with the fastest bias decaying rate, $\hat{H}_z$, the resulting estimator ($\widehat{MI}_z$) also has a bias with an exponentially decaying rate [34], namely,

$$\widehat{MI}_z = \hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y).$$

The asymptotic properties for $\widehat{MI}$-s ($\widehat{MI}$ and $\widehat{MI}_z$) shall be discussed under two situations: (1) $MI = 0$, and (2) $MI > 0$.

The first situation of $MI = 0$ is used for testing independence. For example, in feature selection, irrelevant (to the outcome) non-ordinal features shall be dropped, and a feature is irrelevant if it is independent of the outcome. Let A be the potential irrelevant feature and B be the outcome; hence one must test $H_0 : MI(A,B) = 0$ against $H_a : MI(A,B) > 0$. To test such a hypothesis, one needs the asymptotic properties of $\widehat{MI}$-s under the null hypothesis: $MI = 0$, derived in [35]. Namely,

Theorem 5 (Asymptotic properties of $\widehat{MI}$ and $\widehat{MI}_z$ when $MI = 0$)**.** *Provided that* $MI = 0$,

$$2n\widehat{MI} \xrightarrow{D} \chi^2_{(K_1-1)(K_2-1)}$$

and

$$2n\widehat{MI}_z + (K_1-1)(K_2-1) \xrightarrow{D} \chi^2_{(K_1-1)(K_2-1)},$$

where n is the sample size and $\chi^2_{(K_1-1)(K_2-1)}$ *stands for chi-squared distribution with degrees of freedom* $(K_1-1)(K_2-1)$.

For the second situation of $MI > 0$ (recall that $MI > 0$ if and only if the two marginals are dependent), the following asymptotic properties were due to [34].

Let

$$\begin{aligned}\hat{v} &= \left(\hat{p}_1, \cdots, \hat{p}_{K_1K_2-1}\right)^\tau\\ &= \left(\hat{p}_{1,1}, \hat{p}_{1,2}, \cdots, \hat{p}_{1,K_2}, \hat{p}_{2,1}, \hat{p}_{2,2}, \cdots, \hat{p}_{2,K_2}, \cdots, \hat{p}_{K_1,1}, \hat{p}_{K_1,2}, \cdots, \hat{p}_{K_1,K_2-1}\right)^\tau\end{aligned}$$

be the enumeration of joint probabilities plug-in estimators. Let

$$\Sigma(\hat{v}) = \begin{pmatrix} \hat{p}_1(1-\hat{p}_1) & -\hat{p}_1\hat{p}_2 & \cdots & -\hat{p}_1\hat{p}_{K_1K_2-1} \\ -\hat{p}_2\hat{p}_1 & \hat{p}_2(1-\hat{p}_2) & \cdots & -\hat{p}_2\hat{p}_{K_1K_2-1} \\ \cdots & \cdots & \cdots & \cdots \\ -\hat{p}_{K_1K_2-1}\hat{p}_1 & -\hat{p}_{K_1K_2-1}\hat{p}_2 & \cdots & \hat{p}_{K_1K_2-1}(1-\hat{p}_{K_1K_2-1}) \end{pmatrix},$$

and

$$g(\hat{v}) = \begin{cases} \ln(\hat{p}_{K_1,\cdot}\hat{p}_{\cdot,K_2}\hat{p}_k) - \ln(\hat{p}_{i,\cdot}\hat{p}_{\cdot,j}\hat{p}_{K_1,K_2}), & \text{if } k\text{'s corresponding } \{i,j\} \text{ satisfying} \\ & i \neq K_1 \text{ and } j \neq K_2 \\ \ln(\hat{p}_{\cdot,K_2}\hat{p}_k) - \ln(\hat{p}_{\cdot,j}\hat{p}_{K_1,K_2}), & \text{if } k\text{'s corresponding } \{i,j\} \text{ satisfying} \\ & i = K_1 \text{ and } j \neq K_2 \\ \ln(\hat{p}_{K_1,\cdot}\hat{p}_k) - \ln(\hat{p}_{i,\cdot}\hat{p}_{K_1,K_2}), & \text{if } k\text{'s corresponding } \{i,j\} \text{ satisfying} \\ & i \neq K_1 \text{ and } j = K_2 \end{cases},$$

where $k \in \{1, 2, \cdots, K_1K_2 - 1\}$. Then,

Theorem 6 (Asymptotic properties of $\widehat{MI}$ and $\widehat{MI}_z$ when $MI > 0$)**.** *Provided that* $MI > 0$,

$$\sqrt{n}(\widehat{MI} - MI)[g^\tau(\hat{v})\Sigma(\hat{v})g(\hat{v})]^{-\frac{1}{2}} \xrightarrow{L} N(0,1)$$

and

$$\sqrt{n}\Big(\widehat{MI}_z - MI\Big)[g^\tau(\hat{v})\Sigma(\hat{v})g(\hat{v})]^{-\frac{1}{2}} \to N(0,1).$$

The following examples describe a proper use of MI and properties in Theorems 5 and 6.

Example 1 (Genes TMEM30A and MTCH2—data and descriptions are in Example 1 of [34])**.** *In the example, data were from two different genes in 191 patients. It has been calculated in [34] that* $\widehat{MI}_z = 0.0552$*. The hypothesis test in Example 1 of [35] gave a p-value of 0.0567, which suggests* $MI = 0$ *at* $\alpha = 0.05$*. However, one shall use the property in Theorem 6 to obtain a confidence interval of MI. One must not use the property in Theorem 5 for the purpose of the confidence interval in this situation (because the asymptotic distribution in Theorem 5 assumes a specific location for MI under the null hypothesis).*

Example 2 (Genes ENAH and ENAH—data and descriptions are in Example 2 of [34])**.** *In the example, data were from different probes of the same genes on 191 patients. It has been calculated in [34] that* $\widehat{MI}_z = 0.1157$*. The hypothesis test in Example 2 of [35] gave a p-value of 0.0012, which suggests* $MI > 0$ *at* $\alpha = 0.05$*. Furthermore, one shall use the property in Theorem 6 to obtain a confidence interval of MI. One must not use the property in Theorem 5 for the purpose of the confidence interval in this situation.*

Example 3 (Compare the MI between Examples 1 and 2)**.** *From Examples 1 and 2,* $\widehat{MI}_z(TMEM30A, MTCH2) = 0.0552$ *and* $\widehat{MI}_z(ENAH_1, ENAH_2) = 0.1157$*. Although the second estimation value is higher, one cannot conclude that the level of dependence between* $ENAH_1$ *and* $ENAH_2$ *is higher than that between TMEM30A and MTCH2 due to the limitation described in the 4-th property in Property 2. To compare the level of dependence, one shall refer to the standardized mutual information in Section 3.1.*

Recall that MI is always non-negative. For the same reason, $\widehat{MI}$ is always non-negative (note that $\widehat{MI}$ can be viewed as the MI for the distribution $\hat{\mathbf{p}}$). Nevertheless, $\widehat{MI}_z$ can be negative under some scenarios. A negative $\widehat{MI}_z$ suggests the level of dependence between the two random elements is extremely weak. If one uses the results from Theorem 5 to test

if $H_0 : MI = 0$, a negative $\widehat{MI}_z$ would lead to a fail-to-reject for most settings of α (level of significance).

Remarks

There is another line of research on multivariate information-theoretic methods, the Partial Information Decomposition (PID) framework [36–38]. The PID may be viewed as a direct extension of MI to a measures of information provided by two or more variables about a third. Interesting applications of the PID are, for example, in explaining representation learning in neural networks [39] or in feature selection from dependent features [40]. PID aims to characterize redundancy with information decomposition. Another approach to characterize redundancy is to utilize MI on a joint feature space [33]. Additional research to compare the two approaches is needed.

2.2. Kullback–Leibler Divergence

Kullback–Leibler divergence (KL) [2], also known as relative entropy, is the distance between two probability distributions, introduced by [2], and is an important measure of information in information theory. The notations to define KL and describe its properties differ slightly from other sections. Let $P = \{p_k : k = 1, \cdots, K\}$ and $Q = \{q_k : k = 1, \cdots, K\}$ be two discrete probability distributions on the same finite alphabet, $\mathcal{X} = \{\ell_k : k = 1, \cdots, K\}$, where $K \geq 2$ is a finite integer. KL is defined to be

$$KL = KL(P\|Q) = \sum_{k=1}^{K} p_k \ln(p_k/q_k) = \sum_{k=1}^{K} p_k \ln(p_k) - \sum_{k=1}^{K} p_k \ln(q_k).$$

Note that many also use D as the notation of KL, namely, $D(P\|Q)$. KL is not a metric since it does not satisfy the triangle inequality and is not symmetric. Some notable properties of KL are:

Property 3 (Kullback–Leibler divergence)**.**

1. *KL is a measurement of non-metric distance between two distributions on the same alphabet (with the same discrete support). It is always non-negative because of Gibbs' inequality.*
2. *$KL = 0$ if and only if the two underlying distributions are the same. Namely, $P = Q$ for each $k = 1, \cdots, K$.*
3. *$KL > 0$ if and only if the two underlying distributions are different. Namely, $p_k \neq q_k$ for some k.*

The use of KL has several variants, including but not limited to, (1) P and Q are unknown; (2) Q is known; (3) P and Q are continuous distributions. The second variant is an alternative method of the Pearson goodness-of-fit test. Interested readers may refer to [41] for more discussion on the second variant. Although utilizing entropic statistics on continuous spaces is generally not recommended, interested readers may refer to [42,43] for discussions on the third variant.

2.2.1. KL Point Estimation-The Plug-in Estimator, Augmented Estimator, and Z-Estimator

Although KL is not exactly a function of entropy, it still carries many similarities with entropy. For that reason, KL estimation is very similar to entropy estimation. For example, KL can be estimated from a plug-in perspective. Let $\hat{p}_k$ be the plug-in estimator of p_k and $\hat{q}_k$ be the plug-in estimator of q_k, then the KL plug-in estimator is

$$\widehat{KL} = \widehat{KL}(P\|Q) = \sum_{k=1}^{K} \hat{p}_k \ln(\hat{p}_k) - \sum_{k=1}^{K} \hat{p}_k \ln(\hat{q}_k).$$

Because $\widehat{KL}$ could have an infinite bias [44], an augmented plug-in estimator of KL was presented in [44]:

$$\widehat{KL}^* = \sum_{i=1}^{K} \hat{p}_k \ln(\hat{p}_k) - \sum_{k=1}^{K} \hat{p}_k \ln(\hat{q}_k^*),$$

where

$$\hat{q}_k^* = \hat{q}_k + \frac{1[\hat{q}_k = 0]}{m},$$

and m is the sample size of the sample from Q. The bias of $\widehat{KL}$ is no faster than $\mathcal{O}(1/n)$, where n is the sample size of the sample from P [44].

Since the $\widehat{KL}$ could have an infinite bias, its estimation in perspectives of $\hat{H}_{MM}$ or $\hat{H}_{JK}$ will not help in reducing the bias to a finite extent. In the perspective of $\hat{H}_z$, a KL estimator with exponentially decaying bias was offered in [44]:

$$\widehat{KL}_z = \widehat{KL}_z(P\|Q) = \sum_{k=1}^{K} \hat{p}_k \left[\sum_{v=1}^{m-m\hat{q}_k} \frac{1}{v} \prod_{j=1}^{v} \left(1 - \frac{m\hat{q}_k}{m-j+1}\right) - \sum_{v=1}^{n-n\hat{p}_k} \frac{1}{v} \prod_{j=1}^{v} \left(1 - \frac{n\hat{p}_k - 1}{n-j}\right) \right].$$

2.2.2. Symmetrized KL and Its Point Estimation

As mentioned in the first property of Property 3, KL is generally an asymmetric measurement. For certain interests that require a symmetric measurement, a symmetrized KL is defined to be

$$\begin{aligned} S = S(P,Q) &= \frac{1}{2}[KL(P\|Q) + KL(Q\|P)] \\ &= \frac{1}{2}\left(\sum_{k=1}^{K} p_k \ln(p_k) - \sum_{k=1}^{K} p_k \ln(q_k)\right) + \frac{1}{2}\left(\sum_{k=1}^{K} q_k \ln(q_k) - \sum_{k=1}^{K} q_k \ln(p_k)\right). \end{aligned}$$

The symmetrized KL S, as a function of KL, can be similarly estimated in the perspective of $\widehat{KL}$, $\widehat{KL}^*$, and $\widehat{KL}_z$. The respective estimators are

$$\hat{S} = \frac{1}{2}\left(\sum_{k=1}^{K} \hat{p}_k \ln(\hat{p}_k) - \sum_{k=1}^{K} \hat{p}_k \ln(\hat{q}_k)\right) + \frac{1}{2}\left(\sum_{k=1}^{K} \hat{q}_k \ln(\hat{q}_k) - \sum_{k=1}^{K} \hat{q}_k \ln(\hat{p}_k)\right),$$

$$\hat{S}^* = \frac{1}{2}\left(\sum_{k=1}^{K} \hat{p}_k \ln(\hat{p}_k) - \sum_{k=1}^{K} \hat{p}_k \ln(\hat{q}_k^*)\right) + \frac{1}{2}\left(\sum_{k=1}^{K} \hat{q}_k \ln(\hat{q}_k) - \sum_{k=1}^{K} \hat{q}_k \ln(\hat{p}_k^*)\right),$$

where $\hat{p}_k^* = \hat{p}_k + 1[\hat{p}_k = 0]/n$ (n is the sample size of the sample from P), and

$$\begin{aligned} \hat{S}_z =& \frac{1}{2}\sum_{k=1}^{K} \hat{p}_k \left[\sum_{v=1}^{m-m\hat{q}_k} \frac{1}{v} \prod_{j=1}^{v} \left(1 - \frac{m\hat{q}_k}{m-j+1}\right) - \sum_{v=1}^{n-n\hat{p}_k} \frac{1}{v} \prod_{j=1}^{v} \left(1 - \frac{n\hat{p}_k - 1}{n-j}\right) \right] \\ &+ \frac{1}{2}\sum_{k=1}^{K} \hat{q}_k \left[\sum_{v=1}^{n-n\hat{p}_k} \frac{1}{v} \prod_{j=1}^{v} \left(1 - \frac{n\hat{p}_k}{n-j+1}\right) - \sum_{v=1}^{m-m\hat{q}_k} \frac{1}{v} \prod_{j=1}^{v} \left(1 - \frac{m\hat{q}_k - 1}{m-j}\right) \right]. \end{aligned}$$

2.2.3. Asymptotic Properties for KL and Symmetrized KL Estimators

The asymptotic properties for $\widehat{KL}$, $\widehat{KL}^*$, $\widehat{KL}_z$, $\hat{S}$, $\hat{S}^*$, and $\hat{S}_z$ are all presented in [44]. All the asymptotic properties therein require $P \neq Q$ (namely, $KL > 0$). When $P = Q$, the asymptotic property of KL (or S) estimators are currently missing from the literature. The derivation of such asymptotic properties is not complicated yet unnecessary. The only purpose of such asymptotic property under $P = Q$ is to test if $H_0 : P = Q$ against $H_a : P \neq Q$. For such a purpose, the two-sample goodness-of-fit chi-squared test can be used (see p. 616 in [45]).

3. Recently Developed Entropic Statistics Quantities and Estimation

In this section, various recently developed entropic statistics quantities are introduced and discussed. Some quantities are quite new with limited estimation properties developed other than the plug-in estimation. Therefore, some of the following discussions focus on conceptual spirits and application potentials.

3.1. Standardized Mutual Information

Mutual information between two random elements (on non-ordinal alphabets) is similar to the covariance between two random variables (on ordinal spaces) regarding properties and drawbacks. For example, the covariance does not provide general information on the degree of correlation, and the concept correlation of coefficient was defined to fill the gap. Similarly, recall the fourth property of MI that MI generally does not provide information about the degree of dependence, standardized mutual information (SMI), κ, has been studied and defined in various ways. To name a few, provided $H(X,Y) < \infty$,

$$\begin{aligned}
\kappa_1 &= \frac{MI(X,Y)}{H(X,Y)}, \\
\kappa_2 &= \frac{MI(X,Y)}{\min\{H(X),H(Y)\}}, \\
\kappa_3 &= \frac{MI(X,Y)}{\sqrt{H(X)H(Y)}}, \\
\kappa_4 &= \frac{MI(X,Y)}{(H(X)+H(Y))/2}, \\
\kappa_5 &= \frac{MI(X,Y)}{\max\{H(X),H(Y)\}}, \\
\kappa_6 &= \frac{MI(X,Y)}{H(X)}.
\end{aligned} \tag{4}$$

The quantity κ_6 is also called information gain ratio [46]. The benefits of SMI are supported by Theorem 7.

Theorem 7 (Theorem 5.4 in [28]). *Suppose $H(X,Y) < \infty$. Then*

$$0 \le \kappa \le 1$$

Moreover, (1) $\kappa = 0$ if and only if X and Y are independent, and (2) $\kappa = 1$ if and only if X and Y have a one-to-one correspondence.

Interested readers may refer to [47–50] for discussions on SMI. A detailed discussion of the estimation of various SMI may be found in [51].

3.2. Entropic Basis: A Generalization from Shannon's Entropy

Shannon's entropy and MI are powerful tools to quantify dispersion and dependence on non-ordinal space. More concepts and statistical tools are needed to characterize non-ordinal space information from different perspectives.

Generalized Simpson's diversity indices were established in [52] and coined in [3].

Definition 1 (Generalized Simpson's Diversity Indices). *For a given $\mathbf{p} = \{p_k; k \ge 1\}$ and an integer pair $(u \ge 1, v \ge 0)$, let $\zeta_{u,v} = \sum_{k\ge 1} p_k^u (1- p_k)^v$. Let*

$$\boldsymbol{\zeta} = \{\zeta_{u,v}; u \ge 1, v \ge 0\}$$

be defined as the family of generalized Simpson's diversity indices.

Generalized Simpson's Diversity Indices are the foundation of entropic basis and entropic moment. Interested readers may refer to [53] for discussions on entropic moments and a goodness-of-fit test under permutation based on entropic moments. In estimating $\zeta_{u,v}$,

$$z_{u,v} = \frac{[n-(u+v)]!n^{u+v}}{n!} \times \sum_{k\geq 1}\left\{1\left[\hat{p}_k \geq \frac{u}{n}\right]\left[\prod_{i=0}^{u-1}\left(\hat{p}_k - \frac{i}{n}\right)\right]\left[\prod_{j=0}^{v-1}\left(1-\hat{p}_k - \frac{j}{n}\right)\right]\right\}$$

was derived in [52], where n is the sample size; u and v are given constants; $\hat{p}_k$ is the sample proportion of the k-th letter(category); $1[\cdot]$ stands for indicator function. $z_{u,v}$ is an uniformly minimum-variance unbiased estimator (UMVUE) of $\zeta_{u,v}$ for any combination of (u,v) non-negative integers pair as long as $u+v \leq n$, where n is the corresponding sample size.

Based on $\zeta_{u,v}$, ref. [3] defined the entropic basis.

Definition 2 (Entropic Basis). *Given Definition 1, the entropic basis is the sub-family*

$$\boldsymbol{\zeta}_1 = \{\zeta_{1,v}; v \geq 0\}$$

of $\boldsymbol{\zeta}$.

All diversity indices can be represented as a function of $\boldsymbol{\zeta}_1$ [3] (most representations are due to Taylor's expansion). For example,

1. Simpson's index [54]:
$$\lambda = \sum_{k\geq 1} p_k^2 = \zeta_{1,0} - \zeta_{1,1}$$
2. Gini–Simpson index [54,55]:
$$1-\lambda = \sum_{k\geq 1} p_k(1-p_k) = \zeta_{1,1}$$
3. Shannon's entropy:
$$H = -\sum_{k\geq 1} p_k \ln(p_k) = \sum_{v=1}^{\infty}\frac{1}{v}\zeta_{1,v}$$
4. Rényi equiv. entropy [56]:
$$h_r = \sum_{k\geq 1} p_k^r = \zeta_{1,0} + \sum_{v=1}^{\infty}\prod_{i=1}^{v}\left(\frac{i-r}{i}\right)\zeta_{1,v}$$
5. Emlen's index [57]:
$$D = \sum_{k\geq 1} p_k e^{-p_k} = \sum_{v=0}^{\infty}\frac{e^{-1}}{v!}\zeta_{1,v}$$
6. Richness index (population size):
$$K = \sum_{k\geq 1} 1[p_k > 0] = \sum_{v=0}^{\infty}\zeta_{1,v}$$
7. Generalized Simpson's index:
$$\zeta_{u,m} = \sum_{k\geq 1} p_k^u(1-p_k)^m = \sum_{v=0}^{u-1}(-1)^v\begin{pmatrix} u-1 \\ v \end{pmatrix}\zeta_{1,m+v}.$$

In practice, plug-in estimation is used in estimating diversity indices. The representations of the diversity index on an entropic basis allow a new estimation method with a smaller bias. Namely, $z_{1,v}$, the UMVUE for $\zeta_{1,v}$, exist for all v up to $n-1$. If one replaces $\{\zeta_{1,0}, \zeta_{1,1}, \cdots, \zeta_{1,n-1}\}$ with $\{z_{1,0}, z_{1,1}, \cdots, z_{1,n-1}\}$, and let all the other ζs (namely, $\{\zeta_{1,n}, \zeta_{1,n+1}, \cdots\}$) to be zero, then the resulting estimator is exactly the same as plug-in estimator. However, the estimation can be further improved if one estimate $\{\zeta_{1,n}, \zeta_{1,n+1}, \cdots\}$ based on $\{z_{1,0}, z_{1,1}, \cdots, z_{1,n-1}\}$.

For example, let $\hat{K}$ (the observed number of categories) be the plug-in estimator of K. Meanwhile, $\sum_{v=0}^{n-1} z_{1,v}$ (the estimator in perspective of entropic basis representation) is algebraically equivalent to $\hat{K}$ [58]. Namely,

$$K = \sum_{k\geq 1} 1[p_k > 0] = \sum_{v=0}^{\infty} \zeta_{1,v} = \sum_{v=0}^{n-1} \zeta_{1,v} + \sum_{v=n}^{\infty} \zeta_{1,v}$$

is a decomposition of K, then

$$\hat{K} = \sum_{k\geq 1} 1[\hat{p}_k > 0] = \sum_{v=0}^{n-1} \zeta_{1,v};$$

whereas

$$\hat{K}_{entropic} = \sum_{v=0}^{n-1} \zeta_{1,v} + \sum_{v=n}^{\infty} \zeta_{1,v} = \hat{K} + \text{estimator of } \sum_{v=n}^{\infty} \zeta_{1,v}.$$

$\hat{K}_{entropic}$ has a smaller bias than $\hat{K}$. Interested readers may refer to [58] for details on the estimation of $\sum_{v=n}^{\infty} \zeta_{1,v}$. Similar estimation could benefit the estimation of Rényi equiv. entropy, Emlen's index, and any other diversity indices or theoretical quantities which contain the terms $\sum_{v=n}^{\infty} \zeta_{1,v}$ after Taylor's expansion.

3.3. Generalized Shannon's Entropy and Generalized Mutual Information

Because of the advantages in characterizing information in non-ordinal space, Shannon's entropy and MI have become the building blocks of information theory and essential aspects of ML methods. Yet, they are only finitely defined for distributions with fast decaying tails on a countable alphabet.

Example 4 (Unbounded entropy)**.** *Let $X = \{2, 3, 4, \cdots\}$ be a random variable following the distribution $P(X = k) = p_k = c/(k \ln^2 k)$, where $k = 2, 3, 4, \cdots$, and c is the constant to make the distribution valid (total probability add up to 1). Such a constant uniquely exists because the summation $\sum_{i=2}^{\infty} 1/(i \ln^2 i)$ converges. Then*

$$\begin{aligned} H(X) &= \sum_{k=2}^{\infty} p_k \ln \frac{1}{p_k} = \sum_{k=2}^{\infty} \left(\frac{c}{k \ln^2 k} \ln \frac{k \ln^2 k}{c} \right) \\ &= \sum_{k=2}^{\infty} \left(\frac{c}{k \ln^2 k} (\ln k + 2 \ln \ln k - \ln c) \right) \\ &= \sum_{k=2}^{\infty} \frac{c}{k \ln k} + \sum_{k=2}^{\infty} \left(\frac{c(2 \ln \ln k - \ln c)}{k \ln^2 k} \right) = \infty, \end{aligned}$$

because

$$\sum_{k=2}^{\infty} \frac{c}{k \ln k} = \infty$$

and

$$\sum_{k=2}^{\infty} \left(\frac{c(2 \ln \ln k - \ln c)}{k \ln^2 k} \right) < \infty.$$

Therefore the entropy $H(X)$ is unbounded.

The unboundedness of Shannon's entropy and MI over the general class of all distributions on an alphabet prevents their potential utility from being fully realized. Ref. [5] proposed GSE and GMI, which are finitely defined everywhere. To state the definition of GSE and GMI, Definition 3 is stated first.

Definition 3 (Conditional Distribution of Total Collision (CDOTC)). *Given $\mathcal{X} = \{x_i; i \geq 1\}$ and $\mathbf{p} = \{p_i\}$, consider the experiment of drawing an identically and independently distributed (iid) sample of size m ($m \geq 2$). Let C_m denote the event that all observations of the sample take on the same letter in $\mathcal{X}$, and let C_m be referred to as the event of a total collision. The conditional probability, given C_m, that the total collision occurs at the letter x_i is*

$$p_{m,i} = \frac{p_i^m}{\sum_{s\geq 1} p_s^m},$$

where $m \geq 2$. $\mathbf{p}_m = \{p_{m,i}\}$ is defined as the m-th order CDOTC.

The idea of CDOTC is to adopt a special member of the family of the escort distributions introduced in [59]. The utility of CDOTC is endorsed by Lemmas 1 and 2, which are proved in [5].

Lemma 1. *For any order m, $\mathbf{p}$ and $\mathbf{p}_m$ uniquely determine each other.*

Lemma 2. *For any order m, $\mathbf{p}_{XY} = \{p_{i,j}\} = \{p_{i,.} \times p_{.,j}\}$ if and only if $\mathbf{p}_{XY,m} = \{p_{m,i,j}\} = \{p_{m,i,.} \times p_{m,.,j}\}$.*

It is clear that $\mathbf{p}_m$ is a probability distribution induced from $\mathbf{p} = \{p_k\}$. An example is provided to help understand Definition 3.

Example 5 (The second-order CDOTC). *Let $\mathbf{p} = \{p_i\} = \{6i^{-2}/\pi^2; i = 1, 2, 3, \dots\}$, the second-order CDOTC is then defined as*

$$\mathbf{p}_2 = \{p_{2,i}\},$$

where

$$p_{2,i} = \frac{p_i^2}{\sum_{s\geq 1} p_s^2} = \frac{36i^{-4}/\pi^4}{\sum_{s\geq 1}[36s^{-4}/\pi^4]} = \frac{i^{-4}}{\sum_{s\geq 1} s^{-4}}$$

for $i = 1, 2, 3, \dots$.

Based on Definition 3, GSE and GMI are defined as follows.

Definition 4 (Generalized Shannon's Entropy (GSE)). *Given $\mathcal{X} = \{x_i; i \geq 1\}$, $\mathbf{p}_X = \{p_i\}$, and $\mathbf{p}_{X,m} = \{p_{m,i}\}$, generalized Shannon's entropy (GSE) is defined as*

$$H_m(X) = -\sum_{i\geq 1} p_{m,i} \ln p_{m,i},$$

where $p_{m,i}$ is defined in Definition 3, and $m = 2, 3, \dots$ is the order of GSE. GSE with order m is called the m-th order GSE.

Definition 5 (Generalized Mutual Information (GMI)). *Let $\mathcal{X} = \{x_i; i \geq 1\}$ and $\mathcal{Y} = \{y_j; j \geq 1\}$. Let $\mathbf{p}_{XY} = \{p_{i,j}\}$ and $\mathbf{p}_{XY,m} = \{p_{m,i,j}\}$. Let X^* and Y^* be the marginal distributions of $\mathbf{p}_{XY,m}$. The m-th order generalized mutual information (GMI) between X and Y is defined as*

$$MI_m(X, Y) = H(X^*) + H(Y^*) - H_m(X, Y);$$

or equivalently

$$MI_m(X,Y) = H(X^*) + H(Y^*) - H(X^*,Y^*);$$

To help understand Definitions 4 and 5, Examples 6 and 7 are provided as follows.

Example 6 (The second-order GSE). *Let* $\mathcal{X} = \{x_i; i \geq 1\}$ *and* $\mathbf{p} = \{p_i\} = \{6i^{-2}/\pi^2; i = 1,2,3,\dots\}$. *The second-order GSE,* $H_2(X)$, *is then defined as*

$$H_2(X) = -\sum_{i\geq 1} p_{2,i} \ln p_{2,i},$$

where

$$p_{2,i} = \left\{ \frac{i^{-4}}{\sum_{s\geq 1} s^{-4}}; i = 1,2,\dots \right\}$$

is given in Example 5.

Example 7 (The second-order GMI). *Let* $\mathcal{X} = \{x_{i,.}; i \geq 1\}$, $\mathcal{Y} = \{x_{.,j}; j \geq 1\}$, *and* $\mathcal{X} \times \mathcal{Y} = \{x_{i,j}; i \geq 1, j \geq 1\}$. *Let*

$$\mathbf{p}_{XY} = \{p_{i,j}\} = \left\{ \frac{24\left[(i+j-1)^2 + i - j + 1\right]^{-2}}{\pi^2}; i \geq 1, j \geq 1 \right\},$$

and

$$\mathbf{p}_{XY,2} = \{p_{2,i,j}\} = \left\{ \frac{p_{i,j}^2}{\sum_{s\geq 1, t\geq 1} p_{s,t}^2}; i \geq 1, j \geq 1 \right\}.$$

Further, let

$$\mathbf{p}_{X^*} = \left\{ \sum_j p_{2,i,j}; i \geq 1 \right\}$$

and

$$\mathbf{p}_{Y^*} = \left\{ \sum_i p_{2,i,j}; j \geq 1 \right\}.$$

The second-order GMI, $MI_2(X,Y)$, *is then defined as*

$$MI_2(X,Y) = H(X^*) + H(Y^*) - H(X^*,Y^*),$$

where $H(X^*)$, $H(Y^*)$, *and* $H(X^*,Y^*)$ *are Shannon's entropy based on* $\mathbf{p}_{X^*}$, $\mathbf{p}_{Y^*}$, *and* $\mathbf{p}_{XY,2}$, *respectively.*

GSE's and GMI's plug-in estimators are stated in Definitions 6 and 7.

Definition 6 (GSE's plug-in estimators). *Let* $X_1, X_2, \dots, X_n$ *be i.i.d. random variables taking values in* $\mathcal{X} = \{x_i; i \geq 1\}$ *with distribution* $\mathbf{p}_X$. *The plug-in estimator for* $\mathbf{p}_X$ *is* $\hat{\mathbf{p}}_X = \{\hat{p}_i\}$. *The plug-in estimator for the m-th order GSE,* $\hat{H}_m(X)$, *is*

$$\begin{aligned} \hat{H}_m(X) &= -\sum_{i\geq 1} [\hat{p}_{m,i} \ln \hat{p}_{m,i}] \\ &= -\sum_{i\geq 1} \left[\frac{\hat{p}_i^m}{\sum_{s\geq 1} \hat{p}_s^m} \ln \frac{\hat{p}_i^m}{\sum_{s\geq 1} \hat{p}_s^m} \right]. \end{aligned}$$

Definition 7 (GMI's plug-in estimators)**.** *Let* $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ *be i.i.d. random variables taking values in* $\mathcal{X} \times \mathcal{Y} = \{(x_i, y_j); i \geq 1, j \geq 1\}$ *with distribution* $\mathbf{p}_{XY} = \{p_{i,j}\}$*. Let* $\hat{\mathbf{p}}_{XY} = \{\hat{p}_{i,j}\}$ *be the plug-in estimator of* $\mathbf{p}_{XY}$*. The plug-in estimator for the m-th order GMI,* $\widehat{MI}_m(X, Y)$*, is*

$$\widehat{MI}_m(X, Y) = \hat{H}(X^*) + \hat{H}(Y^*) - \hat{H}(X^*, Y^*),$$

where

$$\hat{H}(X^*) = -\sum_{i\geq 1}\left[\left(\sum_{j\geq 1}\frac{\hat{p}_{i,j}^m}{\sum_{s\geq 1,t\geq 1}\hat{p}_{s,t}^m}\right)\ln\left(\sum_{j\geq 1}\frac{\hat{p}_{i,j}^m}{\sum_{s\geq 1,t\geq 1}\hat{p}_{s,t}^m}\right)\right],$$

$$\hat{H}(Y^*) = -\sum_{j\geq 1}\left[\left(\sum_{i\geq 1}\frac{\hat{p}_{i,j}^m}{\sum_{s\geq 1,t\geq 1}\hat{p}_{s,t}^m}\right)\ln\left(\sum_{i\geq 1}\frac{\hat{p}_{i,j}^m}{\sum_{s\geq 1,t\geq 1}\hat{p}_{s,t}^m}\right)\right],$$

$$\hat{H}(X^*, Y^*) = -\sum_{i\geq 1, j\geq 1}\left[\frac{\hat{p}_{i,j}^m}{\sum_{s\geq 1,t\geq 1}\hat{p}_{s,t}^m}\ln\frac{\hat{p}_{i,j}^m}{\sum_{s\geq 1,t\geq 1}\hat{p}_{s,t}^m}\right].$$

The following asymptotic properties for GSE's plug-in estimators are given in [60].

Theorem 8. *Let* $\mathbf{p}_X = \{p_k; k \geq 1\}$ *be a probability distribution on a countably infinite alphabet* $\mathcal{X}$*, without any further conditions,*

$$\sqrt{n}\left(\frac{\hat{H}_m(X) - H_m(X)}{\hat{\sigma}_m}\right) \xrightarrow{d} N(0, 1),$$

where

$$\hat{\sigma}_m^2 = \sum_{k=1}^{\infty}\left[\frac{m^2}{\hat{p}_k}\left(\hat{p}_{m,k}\ln\hat{p}_{m,k} + \hat{p}_{m,k}\hat{H}_m(Z)\right)\right]^2. \tag{5}$$

Theorem 9. *Let* $\mathbf{p}_X = \{p_k; k = 1, 2, \ldots, K\}$ *be a non-uniform probability distribution on a countably finite alphabet* $\mathcal{X}$*, without any further conditions,*

$$\sqrt{n}\left(\frac{\hat{H}_m(X) - H_m(X)}{\hat{\sigma}_m}\right) \xrightarrow{d} N(0, 1),$$

where

$$\hat{\sigma}_m^2 = \sum_{k=1}^{K}\left[\frac{m^2}{\hat{p}_k}\left(\hat{p}_{m,k}\ln\hat{p}_{m,k} + \hat{p}_{m,k}\hat{H}_m(Z)\right)\right]^2.$$

The properties in Theorems 8 and 9 allow interval estimation and hypothesis testing with $\hat{H}_m$. The advantage of shifting the original distribution to an escort distribution is reflected in Theorem 8-the asymptotic normality requires no assumption on a countable infinite alphabet. Theorem 9 can be viewed as a special case of Theorem 8 under a finite situation, where the uniform distribution shall be omitted because a uniform distribution has no variation between different category probabilities and hence results in a zero GSE and degenerate asymptotic distribution.

Nevertheless, suppose one is certain that the cardinality of distribution is finite. In that case, one shall use Shannon's entropy instead of GSE because Shannon's entropy always exists under a finite distribution. There are various well-studied estimation methods with Shannon's entropy (whereas only plug-in estimation on GSE has been studied by far).

Asymptotic properties for GMI plug-in estimator have not been studied yet. Nonetheless, a test of independence with modified GMI [61] has been studied. The test does not require the knowledge of the number of columns or rows of a contingency table; hence it yielded an alternative other than Pearson's chi-squared test of independence, particularly when a contingency table is large or sparse.

4. Application of Entropic Statistics in Machine Learning and Knowledge Extraction

Applications of entropic statistics in ML and knowledge extraction can be clustered in two directions. The first direction is to solve an existing question from a new perspective by creating a new information-theoretic quantity [61] or revisiting an existing information-theoretic quantity for additional insights [62]. The second direction is to use different estimation methods in existing methods to improve the performance by reducing bias and/or variation [32]. Application potentials in the second direction are very promising because theoretical results from recent-developed estimation methods suggest the performance of many existing ML methods could be improved, yet not much research has been conducted in the direction. In this section, many established ML and knowledge extraction methods are discussed with their potential to improve in the second direction.

4.1. An Entropy-Based Random Forest Model

Ref. [63] proposed an entropy-importance-based random forest model for power quality feature selection and disturbance classification. The method used a greedy search based on entropy and information gain for node segmentation. Nevertheless, only the plug-in estimation of entropy and information gain was considered. The method could be improved by replacing the plug-in estimation with smaller bias estimation methods, such as $\hat{H}_z$ in [25]. Further, one can also combine $\hat{H}_z$ with the jackknife procedure in (3) to obtain

$$\hat{H}_{z\mathbb{JK}} = \frac{\sum_{i=1}^{n} \hat{H}_{z(i)}}{n},$$

where $\hat{H}_{z(i)} = n\hat{H}_z - (n-1)\hat{H}_z^{(i)}$, and use $\hat{H}_{z\mathbb{JK}}$ in place of the adopted plug-in estimation. The benefit of using $\hat{H}_{zJK}$ is the potential smaller bias, and variance [25]. However, asymptotic properties for $\hat{H}_{zJK}$ are yet developed. When asymptotic properties are desired (e.g., for confidence interval or hypothesis testing purposes), one shall consider estimators with established asymptotic properties (also called theoretical guarantee), such as $\hat{H}$, $\hat{H}_{MM}$, $\hat{H}_{JK}$, and $\hat{H}_z$.

4.2. Feature Selection Methods

In [16,64], various information-theoretic feature selection methods were reviewed and discussed. The two review articles did not mention that all the discussed methods adopted plug-in estimators for the corresponding information-theoretic quantities. Improving the performance with different estimation methods is possible, and investigation is needed. For example, some of the discussed methods are summarized in Table 2 with suggestions to utilize smaller bias and/or variance estimation methods.

Table 2. Selected information-theoretic feature selection methods reviewed in [16,64] with potential perspectives to improve the performance using smaller bias and/or variance estimation methods. The proposed criterion uses the same notations as their original forms for readers' ease in tracing them back to the original articles, except some terms are equivalently re-denoted as $\widehat{MI}$ or $\hat{\kappa}$ to help readers follow the same notations in the previous sections. To further clarify the notations, $\widehat{MI}(X_1, X_2)$ and $\widehat{MI}(X_1; X_2)$ are the same. $\widehat{MI}(X_1, X_2|Y)$ is the plug-in estimator of conditional mutual information. By its definition, $\widehat{MI}(X_1, X_2|Y) = \hat{H}(X_1, Y) + \hat{H}(X_2, Y) - \hat{H}(X_1, X_2, Y) - \hat{H}(Y)$. Similarly, $\widehat{MI}_z(X_1, X_2|Y) = \hat{H}_z(X_1, Y) + \hat{H}_z(X_2, Y) - \hat{H}_z(X_1, X_2, Y) - \hat{H}_z(Y)$, where each $\hat{H}_z$ could be further replaced using $\hat{H}_z$ with jackknife procedure.

MIM [65]	Proposed Criterion (Score)	$\widehat{MI}$
MIFS [66]		
JMI [67]	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure
IF [68]	Proposed Criterion (Score)	$\widehat{MI}(C, F)$
	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure
FCBF [69]	Proposed Criterion (Score)	$\hat{\kappa}_4$
	Different Estimation Method	Use $\hat{\kappa}_{4z}$ [51] with jackknife procedure
AMIFS [70]	Proposed Criterion (Score)	$\widehat{MI}(C; f_i) - \beta \sum_{s \in S} \hat{\kappa}_6(C; f_s) \widehat{MI}(f_s; f_i)$
	Different Estimation Method	Use $\widehat{MI}_z$ and $\hat{\kappa}_{6z}$ [51] with jackknife procedure
CMIM [71]	Proposed Criterion (Score)	$\widehat{MI}(Y; X_n)$ and $\widehat{MI}(Y; X_n \| X_m)$
	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure
MRMR [72]	Proposed Criterion (Score)	$\max\left[\widehat{MI}(x_i; c) - \frac{1}{\|s\|} \sum_{x_i, x_j \in S} \widehat{MI}\left(x_i, x_j\right)\right]$
	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure
ICAP [73]	Proposed Criterion (Score)	$\arg\max_{X \in \overline{\mathcal{A}}}\left(\widehat{MI}(X; Y) + \sum_{A \in \mathcal{A}} \min\{0, \widehat{MI}(X; Y; A)\}\right)$
	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure
CIFE [74]	Proposed Criterion (Score)	$\underset{\theta_t}{\operatorname{argmax}}\left\{\widehat{MI}\left(y^{(t)}; c\right) - \sum_{u=1}^{t-1}\left[\widehat{MI}\left(y^{(u)}; y^{(t)}\right) - \widehat{MI}\left(y^{(u)}; y^{(t)} \mid c\right)\right]\right\}$
	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure
DISR [75]	Proposed Criterion (Score)	$\arg\max_{X_i \in X_{-s}}\left\{\sum_{X_j \in X_S} \hat{\kappa}_1\left(X_{i,j}; Y\right)\right\}$
	Different Estimation Method	Use $\hat{\kappa}_{1z}$ [51] with jackknife procedure
IGFS [76]	Proposed Criterion (Score)	$\arg\max_{X \in X_{-s}}\left(\widehat{MI}(X_i; Y) + \frac{1}{d} \sum_{X_j \in X_S} \widehat{MI}\left(X_i; X_j; Y\right)\right)$
	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure
SOA [77]	Proposed Criterion (Score)	$\sum_i \widehat{MI}(X_i, Y) - \sum_i \sum_{j>i} \widehat{MI}\left(X_i, X_j\right) + \sum_i \sum_{j>i} \widehat{MI}\left(X_i, X_j \mid Y\right)$
	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure
CMIFS [78]	Proposed Criterion (Score)	$\widehat{MI}(C; f_i \mid f_1) + \left[\widehat{MI}(f_n; f_i \mid C) - \widehat{MI}(f_n; f_i \mid f_1)\right]$
	Different Estimation Method	Use $\widehat{MI}_z$ with jackknife procedure

4.3. A Keyword Extraction Method

Ref. [79] proposed a keyword extraction method with Rényi's entropy

$$S_R(w, q) = \frac{1}{1-q} \log_2 \sum_{i=1}^{F_w} p_i^q,$$

and used the plug-in estimator therein. Namely, $\hat{S}_R(w,q) = \frac{1}{1-q}\log_2 \sum_{i=1}^{F_w} \hat{p}_i^q$. Nevertheless, $S_R(w,q)$ can be represented as

$$\begin{aligned} S_R(w,q) =& \frac{1}{1-q}\log_2\left[\zeta_{1,0} + \sum_{v=1}^{\infty}\left(\prod_{j=1}^{v}\left(\frac{j-q}{j}\right)\zeta_{1,v}\right)\right] \\ =& \frac{1}{1-q}\log_2\left[\zeta_{1,0} + \sum_{v=1}^{n-1}\left(\prod_{j=1}^{v}\left(\frac{j-q}{j}\right)\zeta_{1,v}\right) + \sum_{v=n}^{\infty}\left(\prod_{j=1}^{v}\left(\frac{j-q}{j}\right)\zeta_{1,v}\right)\right], \end{aligned}$$

where $\zeta_{1,v}$ has the UMVUE $z_{1,v}$ for $v = 0, 1, 2, \cdots, n-1$. For $v \geq n$, $\zeta_{1,v}$ could be estimated based on regression analysis [58]. Hence, $S_R(w,q)$ can be estimated as

$$\hat{S}_{Rz}(w,q) = \frac{1}{1-q}\log_2\left[z_{1,0} + \sum_{v=1}^{n-1}\left(\prod_{j=1}^{v}\left(\frac{j-q}{j}\right)z_{1,v}\right) + \sum_{v=n}^{\infty}\left(\prod_{j=1}^{v}\left(\frac{j-q}{j}\right)z_{1,v}^*\right)\right],$$

where the construction of $z_{1,v}^*$ needs investigation using regression analysis [58]. The resulting $\hat{S}_{Rz}(w,q)$ would have a smaller bias than that of $\hat{S}_R(w,q)$ to help improve the established keyword extraction method. Note that if one wishes to use $z_{1,v}$ up to $n-1$ only, the resulting estimator would become

$$\hat{S}_{Rz}^*(w,q) = \frac{1}{1-q}\log_2\left[z_{1,0} + \sum_{v=1}^{n-1}\left(\prod_{j=1}^{v}\left(\frac{j-q}{j}\right)z_{1,v}\right)\right]. \tag{6}$$

(6) is the same as $\hat{H}_r^\sharp(n)$ in [3]. Asymptotic properties for $\hat{H}_r^\sharp(n)$ were provided therein (Corollary 3 in [3]) for interested readers.

5. Conclusions

Entropic statistics is effective in characterizing information from non-ordinal space. Meanwhile, it is essential to realize that non-ordinal information is inherently difficult to identify due to its non-ordinal and permutation invariant nature. This survey article aims to provide a comprehensive review of recent advances in entropic statistics, including classic entropic concepts estimation, recent-developed entropic statistics quantities, and their applications potentials in ML and knowledge extraction. This article first introduces the concept of entropic statistics and emphasizes challenges from non-ordinal data. Then this article reviews the estimation for classic entropic quantities. These classic entropic concepts, including Shannon's entropy, MI, and KL, are widely used in established machine learning and knowledge extraction methods. Most, if not all, of the established methods use plug-in estimation, which is computation efficient yet with a large bias. The surveyed different estimation methods would help researchers to potentially improve existing methods' performance by adopting a different estimation method or adding theoretical guarantee to the existing methods. Recent-developed entropic statistics concepts are also reviewed with their estimation and applications. These new concepts not only allow researchers to estimate existing quantities in a new perspective, but also support additional aspects in characterizing non-ordinal information. In particular, the generalized Simpson's diversity indices (with the induced entropic basis and entropic moments) have significant application and theoretical potential to either customize existing ML and knowledge extraction methods or to establish new methods considering domain-specific challenges. Further, this article provides some examples of how to apply the surveyed results to some of the existing methods, including a random forest model, fourteen feature selection methods, and a keyword extraction model. It should be mentioned that the aim of the survey is not to claim the superiority of some estimation methods over others but to provide a comprehensive list of recent advances in entropic statistics research. Specifically, although an estimator with a faster-decaying bias seems theoretically preferred, it has a longer calculation time even with the convenient R functions, particularly when multiple layers of jackknife (boot-

strap) are involved. The preference of estimation varies case by case—some may prefer an estimator with a smaller bias, some may prefer one with a smaller variance, while some may need a trade-off between them. Furthermore, the article focuses on non-parametric estimation, while parametric estimation would perform better if the specified model fits the domain-specific reality. In summary, one should always investigate if a new estimation method fits the needs.

Enormous additional works are still needed in entropic statistics. For example, (1) the asymptotic properties for many established estimators (such as $\hat{H}_{chao}$ and $\hat{H}_{zJK}$) are not clear when cardinality is infinite. (2) With the transition from original distribution to escort distribution, GSE and GMI fill the void left by Shannon's entropy and MI. However, only plug-in estimations of GSE and GMI have been studied. The biases of these plug-in estimators have not been studied, and additional estimation methods are undoubtedly needed. (3) Calculations for many entropic statistics are not yet supported in R, such as entropic basis, GSE, and GMI. Furthermore, more work is needed to implement the new entropic statistics concepts in programming software other than R (some of the reviewed estimators are implemented in R and are listed in Appendix A as a reference), particularly in Python. With additional theoretical development and application support, entropic statistics methods would be a more efficient tool to characterize more non-ordinal information and better serve the demands arose from the emerging domain-specific challenges.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

ANOVA	Analysis of Variance
GSE	Generalized Shannon's Entropy
GMI	Generalized Mutual Information
i.i.d.	independent and identically distributed
KL	Kullback–Leibler Divergence
MI	Mutual Information
ML	Machine Learning
PID	Partial Information Decomposition
SMI	Standardized Mutual Information
UMVUE	Uniformly Minimum-Variance Unbiased Estimator

Appendix A. R Functions

Statistic	R Package Name	Function Name
$\hat{H}$	entropy [80]	entropy.plugin
$\hat{H}_{MM}$	entropy	entropy.MillerMadow
$\hat{H}_{JK}$	bootstrap [81]	jackknife
$\hat{H}_{chao}$	entropy	entropy.ChaoShen
$\hat{H}_z$	EntropyEstimation [82]	Entropy.z
$\hat{\sigma}$ in Theorems 1 and 3	EntropyEstimation	Entropy.sd
$\widehat{MI}$	entropy	mi.plugin
$\widehat{MI}_z$	EntropyEstimation	MI.z
$\left[g^\tau(\hat{v})\Sigma(\hat{v})g(\hat{v})\right]^{\frac{1}{2}}$ in Theorem 6	EntropyEstimation	MI.sd
$\widehat{KL}$	entropy	KL.plugin
$\widehat{KL}_z$	EntropyEstimation	KL.z
$\hat{S}$	EntropyEstimation	SymKL.plugin
$\hat{S}_z$	EntropyEstimation	SymKL.z
$\hat{H}_r^s(n)$	EntropyEstimation	Renyi.z

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
2. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
3. Zhang, Z.; Grabchak, M. Entropic representation and estimation of diversity indices. *J. Nonparametr. Stat.* **2016**, *28*, 563–575. [CrossRef]
4. Grabchak, M.; Zhang, Z. Asymptotic normality for plug-in estimators of diversity indices on countable alphabets. *J. Nonparametr. Stat.* **2018**, *30*, 774–795. [CrossRef]
5. Zhang, Z. Generalized Mutual Information. *Stats* **2020**, *3*, 158–165. [CrossRef]
6. Burnham, K.P.; Anderson, D.R. Practical use of the information-theoretic approach. In *Model Selection and Inference*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 75–117.
7. Dembo, A.; Cover, T.M.; Thomas, J.A. Information theoretic inequalities. *IEEE Trans. Inf. Theory* **1991**, *37*, 1501–1518. [CrossRef]
8. Chatterjee, S.; Hadi, A.S. *Regression Analysis by Example*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
9. Speed, T. What is an analysis of variance? *Ann. Stat.* **1987**, *15*, 885–910. [CrossRef]
10. Hardy, M.A. *Regression with Dummy Variables*; Sage: Newcastle upon Tyne, UK, 1993; Volume 93.
11. Kent, J.T. Information gain and a general measure of correlation. *Biometrika* **1983**, *70*, 163–173. [CrossRef]
12. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
13. Van Erven, T.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
14. Sethi, I.K.; Sarvarayudu, G. Hierarchical classifier design using mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **1982**, *4*, 441–445. [CrossRef] [PubMed]
15. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660–674. [CrossRef]
16. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–45. [CrossRef]
17. Basharin, G.P. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Probab. Appl.* **1959**, *4*, 333–336. [CrossRef]
18. Harris, B. *The Statistical Estimation of Entropy in the Non-Parametric Case*; Technical Report; Wisconsin Univ-Madison Mathematics Research Center: Madison, WI, USA, 1975.
19. Zhang, Z.; Zhang, X. A normal law for the plug-in estimator of entropy. *IEEE Trans. Inf. Theory* **2012**, *58*, 2745–2747. [CrossRef]
20. Miller, G.A.; Madow, W.G. *On the Maximum Likelihood Estimate of the Shannon-Weiner Measure of Information*; Operational Applications Laboratory, Air Force Cambridge Research Center, Air Research and Development Command, Bolling Air Force Base: Washington, DC, USA, 1954.
21. Zahl, S. Jackknifing an index of diversity. *Ecology* **1977**, *58*, 907–913. [CrossRef]
22. Chen, C.; Grabchak, M.; Stewart, A.; Zhang, J.; Zhang, Z. Normal Laws for Two Entropy Estimators on Infinite Alphabets. *Entropy* **2018**, *20*, 371. [CrossRef]
23. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms* **2001**, *19*, 163–193. [CrossRef]
24. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
25. Zhang, Z. Entropy estimation in Turing's perspective. *Neural Comput.* **2012**, *24*, 1368–1389. [CrossRef]
26. Schürmann, T. A note on entropy estimation. *Neural Comput.* **2015**, *27*, 2097–2106. [CrossRef] [PubMed]
27. Zhang, Z. Asymptotic normality of an entropy estimator with exponentially decaying bias. *IEEE Trans. Inf. Theory* **2013**, *59*, 504–508. [CrossRef]
28. Zhang, Z. *Statistical Implications of Turing's Formula*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
29. Chao, A.; Shen, T.J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **2003**, *10*, 429–443. [CrossRef]
30. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. *arXiv* **2001**, arXiv:physics/0108025.
31. Agresti, A.; Hitchcock, D.B. Bayesian inference for categorical data analysis. *Stat. Methods Appl.* **2005**, *14*, 297–330. [CrossRef]
32. Hausser, J.; Strimmer, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.* **2009**, *10*, 1469–1484.
33. Shi, J.; Zhang, J.; Ge, Y. CASMI—An Entropic Feature Selection Method in Turing's Perspective. *Entropy* **2019**, *21*, 1179. [CrossRef]
34. Zhang, Z.; Zheng, L. A mutual information estimator with exponentially decaying bias. *Stat. Appl. Genet. Mol. Biol.* **2015**, *14*, 243–252. [CrossRef]
35. Zhang, J.; Chen, C. On "A mutual information estimator with exponentially decaying bias" by Zhang and Zheng. *Stat. Appl. Genet. Mol. Biol.* **2018**, *17*, 20180005. [CrossRef]
36. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
37. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [CrossRef]
38. Griffith, V.; Koch, C. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 159–190.

39. Tax, T.M.; Mediano, P.A.; Shanahan, M. The partial information decomposition of generative neural network models. *Entropy* **2017**, *19*, 474. [CrossRef]
40. Wollstadt, P.; Schmitt, S.; Wibral, M. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *arXiv* **2021**, arXiv:2105.04187.
41. Mori, T.; Nishikimi, K.; Smith, T.E. A divergence statistic for industrial localization. *Rev. Econ. Stat.* **2005**, *87*, 635–651. [CrossRef]
42. Wang, Q.; Kulkarni, S.R.; Verdú, S. Divergence estimation for multidimensional densities via *k*-Nearest-Neighbor distances. *IEEE Trans. Inf. Theory* **2009**, *55*, 2392–2405. [CrossRef]
43. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861. [CrossRef]
44. Zhang, Z.; Grabchak, M. Nonparametric estimation of Küllback-Leibler divergence. *Neural Comput.* **2014**, *26*, 2570–2593. [CrossRef]
45. Press, W.H.; Teukolsky Saul, A. *Numerical Recipes in Fortran: The Art of Scientific Computing*; Cambridge University Press: Cambridge, UK, 1993.
46. De Mántaras, R.L. A distance-based attribute selection measure for decision tree induction. *Mach. Learn.* **1991**, *6*, 81–92. [CrossRef]
47. Kvalseth, T.O. Entropy and correlation: Some comments. *IEEE Trans. Syst. Man Cybern.* **1987**, *17*, 517–519. [CrossRef]
48. Strehl, A.; Ghosh, J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.
49. Yao, Y. Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 115–136.
50. Vinh, N.X.; Epps, J.; Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **2010**, *11*, 2837–2854.
51. Zhang, Z.; Stewart, A.M. *Estimation of Standardized Mutual Information*; Technical Report; UNC Charlotte Technical Report: Charlotte, NC, USA, 2016.
52. Zhang, Z.; Zhou, J. Re-parameterization of multinomial distributions and diversity indices. *J. Stat. Plan. Inference* **2010**, *140*, 1731–1738. [CrossRef]
53. Chen, C. Goodness-of-Fit Tests under Permutations. Ph.D. Thesis, The University of North Carolina at Charlotte, Charlotte, NC, USA, 2019.
54. Simpson, E.H. Measurement of diversity. *Nature* **1949**, *163*, 688. [CrossRef]
55. Gini, C. Measurement of inequality of incomes. *Econ. J.* **1921**, *31*, 124–126. [CrossRef]
56. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 1 January 1961; Volume 1.
57. Emlen, J.M. *Ecology: An Evolutionary Approach*; Addison-Wesley: Boston, MA, USA, 1977.
58. Zhang, Z.; Chen, C.; Zhang, J. Estimation of population size in entropic perspective. *Commun.-Stat.-Theory Methods* **2020**, *49*, 307–324. [CrossRef]
59. Beck, C.; Schögl, F. *Thermodynamics of Chaotic Systems*; Cambridge University Press: Cambridge, UK, 1995.
60. Zhang, J.; Shi, J. Asymptotic Normality for Plug-In Estimators of Generalized Shannon's Entropy. *Entropy* **2022**, *24*, 683. [CrossRef]
61. Zhang, J.; Zhang, Z. A Normal Test for Independence via Generalized Mutual Information. *arXiv* **2022**, arXiv:2207.09541.
62. Kontoyiannis, I.; Skoularidou, M. Estimating the directed information and testing for causality. *IEEE Trans. Inf. Theory* **2016**, *62*, 6053–6067. [CrossRef]
63. Huang, N.; Lu, G.; Cai, G.; Xu, D.; Xu, J.; Li, F.; Zhang, L. Feature selection of power quality disturbance signals with an entropy-importance-based random forest. *Entropy* **2016**, *18*, 44. [CrossRef]
64. Brown, G.; Pocock, A.; Zhao, M.J.; Luján, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
65. Lewis, D.D. Feature selection and feature extraction for text categorization. In Proceedings of the Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, NY, USA, 23–26 February 1992.
66. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [CrossRef] [PubMed]
67. Yang, H.; Moody, J. Feature selection based on joint mutual information. In Proceedings of the International ICSC Symposium on Advances in Intelligent Data Analysis, Rochester, NY, USA, 22–25 June 1999; Volume 1999, pp. 22–25.
68. Ullman, S.; Vidal-Naquet, M.; Sali, E. Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* **2002**, *5*, 682–687. [CrossRef] [PubMed]
69. Yu, L.; Liu, H. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.* **2004**, *5*, 1205–1224.
70. Tesmer, M.; Estévez, P.A. AMIFS: Adaptive feature selection by using mutual information. In Proceedings of the 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541), Budapest, Hungary, 25–29 July 2004; Volume 1, pp. 303–308.
71. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **2004**, *5* , 1531–1555.
72. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
73. Jakulin, A. Machine Learning Based on Attribute Interactions. Ph.D. Thesis, Univerza v Ljubljani, Ljubljana, Slovenia, 2005.

74. Lin, D.; Tang, X. Conditional infomax learning: An integrated framework for feature extraction and fusion. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 68–82.
75. Meyer, P.E.; Bontempi, G. On the use of variable complementarity for feature selection in cancer classification. In Proceedings of the Workshops on Applications of Evolutionary Computation, Budapest, Hungary, 10–12 April 2006; pp. 91–102.
76. El Akadi, A.; El Ouardighi, A.; Aboutajdine, D. A powerful feature selection approach based on mutual information. *Int. J. Comput. Sci. Netw. Secur.* **2008**, *8*, 116.
77. Guo, B.; Nixon, M.S. Gait feature subset selection by mutual information. *IEEE Trans. Syst. Man-Cybern.-Part Syst. Hum.* **2008**, *39*, 36–46.
78. Cheng, G.; Qin, Z.; Feng, C.; Wang, Y.; Li, F. Conditional Mutual Information-Based Feature Selection Analyzing for Synergy and Redundancy. *Etri J.* **2011**, *33*, 210–218. [CrossRef]
79. Singhal, A.; Sharma, D. Keyword extraction using Renyi entropy: A statistical and domain independent method. In Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 19–20 March 2021; Volume 1, pp. 1970–1975.
80. R Package Entropy. Available online: https://cran.r-project.org/web/packages/entropy/index.html (accessed on 27 September 2022).
81. R Package Bootstrap. Available online: https://cran.r-project.org/web/packages/bootstrap/index.html (accessed on 27 September 2022).
82. R Package EntropyEstimation. Available online: https://cran.r-project.org/web/packages/EntropyEstimation/index.html (accessed on 27 September 2022).