A New Separation Index and Classification Techniques Based on Shannon Entropy

Methodology and Computing in Applied Probability (2023) 25:78
https://doi.org/10.1007/s11009-023-10055-w

RESEARCH

A New Separation Index and Classification Techniques Based on Shannon Entropy

Jorge Navarro[1] · **Francesco Buono**[2,3] · **Jorge M. Arevalillo**[4,5]

Received: 22 September 2022 / Revised: 25 July 2023 / Accepted: 29 August 2023 /
Published online: 22 September 2023

Abstract

The purpose is to use Shannon entropy measures to develop classification techniques and an index which estimates the separation of the groups in a finite mixture model. These measures can be applied to machine learning techniques such as discriminant analysis, cluster analysis, exploratory data analysis, etc. If we know the number of groups and we have training samples from each group (supervised learning) the index is used to measure the separation of the groups. Here some entropy measures are used to classify new individuals in one of these groups. If we are not sure about the number of groups (unsupervised learning), the index can be used to determine the optimal number of groups from an entropy (information/uncertainty) criterion. It can also be used to determine the best variables in order to separate the groups. In all the cases we assume that we have absolutely continuous random variables and we use the Shannon entropy based on the probability density function. Theoretical, parametric and non-parametric techniques are proposed to get approximations of these entropy measures in practice. An application to gene selection in a colon cancer discrimination study with a lot of variables is provided as well.

Keywords Shannon entropy · Discriminant analysis · Cluster analysis · Kernel density estimation · Omic data

Francesco Buono and Jorge M. Arevalillo contributed equally to this work.

✉ Jorge Navarro
jorgenav@um.es

Francesco Buono
francesco.buono3@unina.it

Jorge M. Arevalillo
jmartin@ccia.uned.es

[1] Department Statistics and Operational Research, Universidad de Murcia, Facultad de Matemáticas, Murcia 30100, Spain

[2] Università di Napoli Federico II, Via Cintia, Napoli I-80126,, Italy

[3] RWTH Aachen University, Pontdriesch 14-16, Aachen 52062, Germany

[4] UC3M-Santander Big Data Institute, Madrid Street 135, Getafe 28903, Madrid, Spain

[5] Department of Statistics and Operational Research, University Nacional Educación a Distancia (UNED), Juan del Rosal 10, Madrid 28040, Spain

Mathematics Subject Classification 62N05 · 90B25

1 Introduction

The measure of the uncertainty associated to a random variable is a task of great and increasing interest. Since the pioneering work of Shannon (1948), in which the concept of Shannon entropy was defined as the average level of information or uncertainty related to a random event, several measures of uncertainty with different purposes have been defined and studied. The Shannon (differential) entropy associated to a random vector **X** with an absolutely continuous distribution is a good way to measure the uncertainty of the data from **X**. It is defined by

$$H(\mathbf{X}) = E(-\log f(\mathbf{X})),$$

where f is the probability density function of **X** and log is the natural log, see Shannon (1948). Several generalizations and extensions of the Shannon entropy have been proposed in the literature with the scope of better analyzing the uncertainty in different scenarios. Among them we recall the weighted entropy (Di Crescenzo and Longobardi 2006), the cumulative entropies (Balakrishnan et al. 2022; Rao et al. 2004), Tsallis entropy (Tsallis 1988) and Rényi entropy (Rényi 1961).

In many applications, the distribution of **X** is a finite mixture of s distributions with some probabilities $p_1, \ldots, p_s \geq 0$ such that $p_1 + \cdots + p_s = 1$. Maybe, the main application nowadays is the assessment of differential expression from high dimensional genomic data. There are a lot of other applications. For example, some applications to information of additive noise models in communication channels or thermodynamic of computations can be seen in Melbourne et al. (2022) and in the references therein. Results for Gaussian (normal) mixtures in several scenarios where the transmitters utilize pulse amplitude modulation constellations can be seen in Moshksar and Khandani (2016).

Several criteria are available in the literature to determine the optimal number of groups in a mixture model. An entropy criterion called NEC (normalized entropy criterion) to estimate the number of clusters arising from a mixture model was proposed in Celeux and Soromenho (1996). There it is compared with other popular indices such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). A recent modification was proposed in Biernacki et al. (1999). The main advantage of the entropy indices is that they do not depend on the number of unknown parameters in the mixture model (as AIC and BIC do).

In this paper we propose a new index based on Shannon entropy to measure the separation of the groups in a mixture. As the NEC, this index does not depend on the number of unknown parameters and it can be estimated by using non-parametric techniques. This index can be used to determine the optimal number of groups in a mixture model both in discriminant (supervised methods) or cluster analysis (unsupervised methods). This approach is also used to propose discriminant criteria to classify new individuals in one of these groups. The procedures are illustrated by using simulated examples (which show the accuracy of the empirical measures) and real data examples. A real case study dealing with the selection of discriminant variables in a colon cancer data set is provided as well.

The rest of the paper is organized as follows. In Sect. 2, we introduce the notation, the main definitions and the preliminary results. The new index and the applications to discriminant analysis are placed in Sect. 3. The examples are in Sect. 4. The application to the colon cancer data set is done in Sect. 5. Finally, Sect. 6 contains the conclusions and some tasks for future research projects.

2 Notation and Preliminary Results

First, we present the notions for the univariate case. The Shannon entropy associated to a random variable X with probability density function (PDF) f is defined by

$$H(X) = E(-\log f(X)) = -\int_{\mathbb{R}} f(x) \log f(x) dx,$$

where log represents the natural log and, by convention, $0 \log 0 = 0$. The value $H(X)$ is used to measure the uncertainty (dispersion) in the values of X, see Shannon (1948). The random variable $-\log(f(\mathbf{X}))$ is called *information content* of X in Di Crescenzo et al. (2021). For other entropy measures see Balakrishnan et al. (2022), Buono and Longobardi (2020), Di Crescenzo and Longobardi (2006), Rao et al. (2004), Rényi (1961), Tsallis (1988) and the references therein.

If $X_1, \ldots, X_n$ is a sample of independent and identically distributed (IID) random variables from X, then $H(X)$ can be estimated with

$$\widehat{H}(X) = -\frac{1}{n} \sum_{i=1}^{n} \log \widehat{f_n}(X_i), \tag{1}$$

where $\widehat{f_n}$ is an estimator for the PDF f based on $X_1, \ldots, X_n$. Here we can use both parametric and non-parametric estimators. In the first case, we assume a known functional form for f (e.g. exponential or Gaussian) with some unknown parameters (mean, variance, etc.) that are estimated from the sample. In the second case, an empirical estimator for f is used (e.g. a kernel density estimator).

If we fix a value t and we consider the values of X below and above t, that is, we consider the conditional random variables $(X|X \le t)$ and $(X|X > t)$, then the entropy of X can be rewritten (see Proposition 2.1 in Di Crescenzo and Longobardi 2002) as

$$H(X) = F(t)H(X|X \le t) + \bar{F}(t)H(X|X > t) + H(G), \tag{2}$$

where $F(t) = \Pr(X \le t)$ is the distribution function of X, $\bar{F}(t) = 1 - F(t) = \Pr(X > t)$ is the reliability (or survival) function of X,

$$H(X|X \le t) = -\int_{-\infty}^{t} \frac{f(x)}{F(t)} \log \frac{f(x)}{F(t)} dx$$

is the entropy of the past lifetime $(X|X \le t)$,

$$H(X|X > t) = -\int_{t}^{\infty} \frac{f(x)}{\bar{F}(t)} \log \frac{f(x)}{\bar{F}(t)} dx$$

is the entropy of the residual lifetime $(X - t|X > t)$, and

$$H(G) = -F(t) \log F(t) - \bar{F}(t) \log \bar{F}(t)$$

is the entropy of the discrete (Bernoulli) random variable necessary to distinguish between the two groups. A similar representation holds for any partition of the support of X with s disjoint sets (groups). A bivariate version of (2) was obtained in Ahmadi et al. (2015).

If X contains two groups $G = 1$ and $G = 0$ with respective PDF f_1 and f_0, then $f = pf_1 + (1 - p)f_0$, where $p = \Pr(G = 1)$. Hence the entropy with the two groups together (i.e. the entropy of the mixture) is

$$H(X) = -\int_{\mathbb{R}} (pf_1(x) + (1-p)f_0(x)) \log(pf_1(x) + (1-p)f_0(x))dx. \tag{3}$$

Expression (2) can be used to define the entropy in X with the two groups as follows.

Definition 1 If X has a mixture PDF $f = pf_1 + (1-p)f_0$, then the *entropy of the two groups* is

$$H^{(2)}(X) = pH(X|G=1) + (1-p)H(X|G=0), \tag{4}$$

where

$$H(X|G=1) = -\int_{\mathbb{R}} f_1(x) \log f_1(x) dx$$

is the entropy of the first group and

$$H(X|G=0) = -\int_{\mathbb{R}} f_0(x) \log f_0(x) dx$$

is the entropy of the second group. The *efficiency* in the division made by the groups is defined as

$$Eff^{(2)}(X) = H(X) - H^{(2)}(X),$$

where $H(X)$ is given in (3).

The efficiency is also called the *concavity deficit* in Melbourne et al. (2022) and can be interpreted as a generalization of the Jensen-Shannon divergence measure (see Briët and Harremoës 2009).

If $S_i = \{x : f_i(x) > 0\}$, $i = 0, 1$ are the supports of the two groups and $S_1 \cap S_0 = \emptyset$ (the two groups are completely separated), then $p = \Pr(X \in S_1)$,

$$H^{(2)}(X) \leq pH(X|G=1) + (1-p)H(X|G=0) + H_2(G) = H(X)$$

and $Eff^{(2)}(X) = H_2(G) \geq 0$, that is, the efficiency coincides with the entropy to distinguish between the two groups given by

$$H_2(G) = -p\log(p) - (1-p)\log(1-p) \geq 0.$$

Even more, if $0 < p < 1$, then $H^{(2)}(X) < H(X)$ and the division in two groups is effective since it decreases the uncertainty (the entropy).

Another extreme case is when $f_1 = f_0$ (identically distributed groups), where

$$H^{(2)}(X) = pH(X|G=1) + (1-p)H(X|G=0) = pH(X) + (1-p)H(X) = H(X)$$

for any $p \in [0, 1]$. Here, $H^{(2)}(X) = H(X)$ and $Eff^{(2)}(X) = 0$ tell us that it is not a good idea to consider two groups since the uncertainty does not change.

Thus, we can say that the division in two groups *is efficient* if $H(X) > H^{(2)}(X)$ since, in this case, it decreases the uncertainty in X. This is always the case when $H(X) > H(X|G = i)$ for $i = 0, 1$, that is, when the uncertainties in the groups are smaller than the uncertainty in the mixed population (a reasonable property).

The following proposition shows that the efficiency is related with the Kullback–Leibler (KL) divergence measure between the densities. If f and g are two PDF, the KL-divergence measure (or the relative entropy) is defined as

$$KL(f|g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx,$$

see e.g. Melbourne et al. (2022). It can be proved that $KL(f|g) \geq 0$ and that $KL(f|g) = 0$

if and only if $f = g$ (a.e.). This proposition also shows that the efficiency is non-negative and bounded by $H_2(G)$. This bound can be traced back to Grandvalet and Bengio (2005) and Cover and Thomas (2006). Some improvements of this bound were obtained in Melbourne et al. (2022) and in Moshksar and Khandani (2016) (Gaussian mixtures). The relationship with KL divergence measure can be seen in, e.g., (22) of Melbourne et al. (2022). To get a self contained paper, we provide a proof in the Appendix since it is a key result for our purposes.

Proposition 1 *Let $f = pf_1 + (1-p)f_0$ be the PDF of X, then*

$$0 \leq Eff^{(2)}(X) = pKL(f_1|f) + (1-p)KL(f_0|f) \leq H_2(G). \tag{5}$$

Moreover, if $f_1 \neq f_0$ (a.e.) and $0 < p < 1$, then $Eff^{(2)}(X) > 0$.

Equation (5) implies that $H(X) \geq H^{(2)}(X)$ under a mixture model with two groups.

A straightforward calculation leads to the generalization to the mixture model with s groups, stated in the following relationship:

$$H(X) - H^{(s)}(X) = \sum_{i=1}^{s} p_i \; KL(f_i \mid f), \tag{6}$$

where $H^{(s)}(X) = \sum_{i=1}^{s} p_i H(X|G = i)$. Therefore, $Eff^{(s)}(X) = H(X) - H^{(s)}(X)$ can be formulated as a weighted sum of KL divergences between the class conditional PDF of the groups and the PDF of the mixture. Expression (6) proves the non-negativeness of $Eff^{(s)}(X)$ whenever there exists an underlying group structure for the variable X. It can also be used to assess the overlapping of the class structure. An investigation is needed in order to elucidate the usefulness of the quantity $Eff^{(s)}(X)$ in the statistical practice. Some applications would include: its use as an auxiliary tool that may help to determine the number of groups in clustering analysis or its application in genomic studies for the selection of genomic variables having the potential to discriminate a clinical outcome, just to name a couple of applications.

The next proposition proves that the efficiency increases when we divide a group in two subgroups. The proof is given in the Appendix.

Proposition 2 *Let $f = p_1 f_1 + (1-p_1) f_0$ be the PDF of X and let us assume that $f_0 = qf_2 + (1-q) f_3$ for some $q \in [0, 1]$. Then $H^{(2)}(X) \geq H^{(3)}(X)$ and*

$$0 \leq Eff^{(2)}(X) \leq Eff^{(3)}(X) \leq H_3(G),$$

where $H_3(G) = -\sum_{i=1}^{3} p_i \log p_i$, $p_2 = (1-p_1)q$ and $p_3 = (1-p_1)(1-q)$.

Now we can state the results for the k-dimensional case. Let $\mathbf{X} = (X_1, \ldots, X_k)$ be a random vector with an absolutely continuous joint distribution and joint PDF f. Then the (multivariate) Shannon entropy is defined by

$$H(\mathbf{X}) = E(-\log f(\mathbf{X})) = -\int_{\mathbb{R}^k} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x}. \tag{7}$$

The estimator for $H(\mathbf{X})$ is defined as in the univariate case. An expression similar to (2) can be obtained for $H(\mathbf{X})$ when the support of $\mathbf{X}$ is divided in s disjoint sets (groups).

The results for this general case are stated below. They are completely analogous to the results for the univariate case, so we omit the proofs.

Definition 2 If $\mathbf{X}$ has a joint PDF $f = \sum_{i=1}^{s} p_i f_i$, the entropy of the s groups is

$$H^{(s)}(\mathbf{X}) = \sum_{i=1}^{s} p_i H(\mathbf{X}|G = i), \tag{8}$$

where

$$H(\mathbf{X}|G = i) = -\int_{\mathbb{R}^k} f_i(\mathbf{x}) \log f_i(\mathbf{x}) d\mathbf{x}$$

is the entropy of the ith group for $i = 1, \ldots, s$. The efficiency in the division made by the s groups is defined as

$$Eff^{(s)}(\mathbf{X}) = H(\mathbf{X}) - H^{(s)}(\mathbf{X}),$$

where $H(\mathbf{X})$ is given in (7).

Proposition 3 *Let $f = \sum_{i=1}^{s} p_i f_i$ be the PDF of $\mathbf{X}$, then*

$$0 \leq Eff^{(s)}(\mathbf{X}) = \sum_{i=1}^{s} p_i KL(f_i|f) \leq H_s(G), \tag{9}$$

where $H_s(G) = -\sum_{i=1}^{s} p_i \log p_i \geq 0$. Moreover, if, for some i, $f_i \neq f$ (a.e) and $0 < p_i < 1$, then $Eff^{(s)}(X) > 0$.

3 New Results

Following the idea of the normalized entropy criterion (NEC) defined in Celeux and Soromenho (1996) and Biernacki et al. (1999), we can define the following index to measure the relative efficiency of the division made by the s groups. This index can be used to decide about the optimal number of groups (clusters) in a finite mixture model (including the case of no groups).

Proposition 2 proves that the efficiency is not a good value if we want to determine the optimal number of groups since it is always increasing when a group is divided in two. So we use the upper bound in (9) to propose a relative efficient measure.

Definition 3 If $f = \sum_{i=1}^{s} p_i f_i$ is the PDF of $\mathbf{X}$, then we define the relative efficiency of the division in s groups, shortly denoted as $RED(s)$, by

$$RED(s) = \frac{Eff^{(s)}(\mathbf{X})}{H_s(G)} = \frac{H(\mathbf{X}) - H^{(s)}(\mathbf{X})}{H_s(G)}.$$

Note that $0 \leq RED(s) \leq 1$ and that we should choose the value of s which leads to a maximum of $RED(s)$. If this value is close to zero, then we should consider just one group (i.e. no groups). In practice, these values will be replaced with their estimations (see next section). Note that the indices $RED(2), RED(3), \ldots$ are not necessarily ordered.

Theorem 1 in Melbourne et al. (2022) provides an upper bound for $RED(s)$ based on $f_1, \ldots, f_s$ written as

$$0 \leq RED(s) \leq T_s$$

where $T_s = \max_{j=1,\ldots,s} \left\| f_j - \widehat{f_j} \right\|_{TV}$, $\|g\|_{TV} = \frac{1}{2}\int_{\mathbb{R}} |g(x)|dx$ is the *Total Variation (TV) distance* and

$$\widehat{f_j}(x) = \sum_{i \neq j} \frac{p_i}{1 - p_j} f_i(x)$$

is the mixture complement of f_j. Note that $f = p_j f_j + (1 - p_j)\widehat{f_j}$.

Let us see now how to apply this approach to classify new individuals in one of these groups. As mentioned above let us assume here that our population is divided into s groups and that we want to use the numerical random variables $X_1, \ldots, X_k$ to classify new individuals into one these groups. To simplify the notation let us assume that $k = 1$ and $s = 2$ but the same techniques can be applied for $k > 1$ and $s > 2$ (see Example 3).

Let us assume first that the PDF of the two groups f_1 and f_0 are known. Then we need to determine two disjoint regions R_1 and R_0 such that $R_1 \cup R_0 = \mathbb{R}$ in order to classify an individual with a value X in the first (second) group when $X \in R_1$ ($X \in R_0$). Two typical (classical) solutions are the maximum likelihood criterion which defines R_1 as

$$R_1^{ML} = \{x : f_1(x) \geq f_0(x)\}$$

and the maximum posterior probability criterion with

$$R_1^{MPP} = \{x : pf_1(x) \geq (1-p)f_0(x)\},$$

where $p = \Pr(G = 1)$. We want to provide an alternative option based on entropy.

The ideal case is when the respective supports of the groups $S_1 = \{x : f_1(x) > 0\}$ and $S_0 = \{x : f_0(x) > 0\}$ are disjoint sets. In that case, the entropy can be written from (2) as

$$H(X) = pH(X|G=1) + (1-p)H(X|G=0) + H_2(G) = H^{(2)}(X) + H_2(G),$$

where $p = \Pr(G = 1) = \Pr(X \in S_1)$, $1 - p = \Pr(G = 0) = \Pr(X \in S_0)$,

$$H(X|G=1) = -\int_{S_1} f_1(x) \log f_1(x) dx,$$

$$H(X|G=0) = -\int_{S_0} f_0(x) \log f_0(x) dx$$

and $H_2(G) = -p \log p - (1-p) \log(1-p)$. In this case $RED(2) = 1$.

This case is unrealistic since usually the populations have values in common regions. So we might try to determine the region R_1 that minimizes

$$H(R_1) := -p_1 \int_{R_1} f_1(x) \log f_1(x) dx - p_0 \int_{R_0} f_0(x) \log f_0(x) dx, \tag{10}$$

where $p_1 = \Pr(X \in R_1)$ and $p_0 = 1 - p_1 = \Pr(X \in R_0)$.

In the ideal case with $S_1 \cap S_0 = \emptyset$, the optimal region is $R_1^{opt} = S_1$ and we have

$$H(X) - H(R_1^{opt}) = H_2(G) \geq 0.$$

Clearly, for $R_1 = \mathbb{R}$ we get $p_1 = 1$ and $H(R_1)$ coincides with the entropy of group 1. For $R_1 = \emptyset$, $p_1 = 0$ and we get the entropy of group 0. Hence, $H(R_1^{opt}) \leq H(X|G = i)$ for $i = 0, 1$. So, from (4) and for the optimal region R_1^{opt} we get

$$\begin{aligned} H^{(2)}(X) &= pH(X|G=1) + (1-p)H(X|G=0) \\ &\geq pH(R_1^{opt}) + (1-p)H(R_1^{opt}) \\ &= H(R_1^{opt}). \end{aligned}$$

Hence $H(X) \geq H^{(2)}(X) \geq H(R_1^{opt})$. Thus, if we define the effectiveness of R_1 as

$$Eff(R_1) = H(X) - H(R_1)$$

we get $Eff(R_1^{opt}) \geq Eff(R_1)$ and $Eff(R_1^{opt}) \geq Eff^{(2)}(X) \geq 0$.

We must say that it is not easy to solve the theoretical problem that leads to the optimal region R_1^{opt}. In the univariate case, if the mean of the first group is bigger than the one of the second, we might assume $R_1 = [t, \infty)$ and then $H(R_1)$ is just a function of t that could be plotted (numerically) in order to find its minimum value.

In order to simplify the calculations in practice, if we have two IID samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ from f_1 and f_0, respectively, the entropies in the groups can be approximated by

$$H(X|G=1) = E(-\log f_1(X)|G=1) \approx \widehat{H}(X|G=1) := -\frac{1}{n}\sum_{i=1}^{n} \log f_1(X_i)$$

and

$$H(X|G=0) = E(-\log f_1(X)|G=0) \approx \widehat{H}(X|G=0) := -\frac{1}{m}\sum_{i=1}^{m} \log f_0(Y_i)$$

when the PDF f_1 and f_0 are known. If they are unknown, they will be replaced by parametric or non-parametric estimations. Therefore, $H^{(2)}(X)$ can be estimated with

$$\widehat{H}^{(2)}(X) := -\frac{p}{n}\sum_{i=1}^{n} \log f_1(X_i) - \frac{1-p}{m}\sum_{i=1}^{m} \log f_0(Y_i), \tag{11}$$

when f_1, f_0 and p are known.

The entropy determined by the region $R_1 = [t, \infty)$ can be written as,

$$H(t) = H(R_1) = -p_1 \int_t^{\infty} f_1(x) \log f_1(x) dx - p_0 \int_{-\infty}^{t} f_0(x) \log f_0(x) dx,$$

where $p_1 = \Pr(X \in R_1) = \Pr(X > t)$. Hence, it can be approximated with

$$\begin{aligned}\widehat{H}(t) = &- \widehat{p}_1 \frac{1}{n+m}\left(\sum_{i=1}^{n} 1(X_i > t) \log f_1(X_i) + \sum_{i=1}^{m} 1(Y_i > t) \log f_1(Y_i)\right) \\ &- \widehat{p}_0 \frac{1}{n+m}\left(\sum_{i=1}^{n} 1(X_i < t) \log f_0(X_i) + \sum_{i=1}^{m} 1(Y_i < t) \log f_0(Y_i)\right),\end{aligned} \tag{12}$$

where $\widehat{p}_1 = \left(\sum_{i=1}^{n} 1(X_i > t) + \sum_{i=1}^{m} 1(Y_i > t)\right)/(n+m)$ and $\widehat{p}_0 = 1 - \widehat{p}_1$.

These sample entropy measures can be used to define a new classification criterion. Thus, if we have a new individual with value $Z = t$, we can compute this entropy by considering that Z belongs to the group 1

$$\widehat{H}_1(t) = \widehat{H}^{(2)}(X|Z \in G_1) = -\frac{p}{n+1}\left(\log f_1(t) + \sum_{i=1}^{n} \log f_1(X_i)\right) - \frac{1-p}{m}\sum_{i=1}^{m} \log f_0(Y_i)$$

or to group 0,

$$\widehat{H}_0(t) = \widehat{H}^{(2)}(X|Z \in G_0) = -\frac{p}{n}\left(\sum_{i=1}^{n} \log f_1(X_i)\right) - \frac{1-p}{m+1}\left(\log f_0(t) + \sum_{i=1}^{m} \log f_0(Y_i)\right).$$

It should be classified into the group with the minimum entropy. If $p = 0.5$, $n = m$ and we replace $n + 1 = m + 1$ with n, then this criterion is equivalent to the maximum likelihood criterion. For an arbitrary probability p, if we are still replacing $n+1$ and $m+1$ with n and m, respectively, $\widehat{H}_1(t) \leq \widehat{H}_0(t)$ holds if and only if

$$-\frac{p}{n}\log f_1(t) \leq -\frac{1-p}{m}\log f_0(t).$$

This is also a reasonable criterion to determine R_1 similar to that based on the posterior probabilities. As mentioned above, in practice, the unknown PDF f_1 and f_0 should be replaced with parametric or non-parametric estimations. If p is unknown and it is estimated with $\widehat{p} = n/(n+m)$, then this criterion is again equivalent to the maximum likelihood criterion. The same can be done in the k-dimensional case or when we have more than two groups. Let us see some examples.

4 Examples

In the first example we consider a population having a mixture of two (univariate) exponential distributions. The purpose is to show the accuracy of the empirical measures.

Example 1 Let us assume that the two groups have exponential distribution functions $F_i(t) = 1 - \exp(-t/\mu_i)$ for $t \geq 0$ and $i = 0, 1$ with means $\mu_1 = 1$ and $\mu_0 = 0.5$. The entropy of the exponential model is

$$H(\mu) = -\int_0^\infty \frac{1}{\mu}e^{-t/\mu}\log\left(\frac{1}{\mu}e^{-t/\mu}\right) = 1 + \log\mu.$$

As expected, it is increasing with μ since its variance is μ^2. Hence, the entropy of the groups are $H(X|G=1) = H(1) = 1$ and $H(X|G=0) = H(0.5) = 1 - \log 2 = 0.3068528$. The values of the first group are more dispersed (i.e. X has a bigger uncertainty in that group).

Let us consider a fifty-fifty mixture of these two groups, that is,

$$f(t) = 0.5f_1(t) + 0.5f_0(t) = 0.5e^{-t} + e^{-2t}$$

for $t \geq 0$. A straightforward calculation shows that its entropy is $H(X) = 0.7072083$. This value is between the values of the entropies in the two groups. This is due to the facts that the first group has a big uncertainty (comparing with the other) and that the two groups share similar values (the supports are not disjoint sets). So, when we mix them, the uncertainty decreases. The entropy with two groups defined by (4) is then

$$H^{(2)}(X) = 0.5H(X|G=1) + 0.5H(X|G=0) = 0.5 + 0.5 \cdot 0.3068528 = 0.6534264.$$

In this case, the division in two groups is effective $H(X) > H^{(2)}(X)$ and the RED index is

$$RED(2) = \frac{H(X) - H^{(2)}(X)}{H_2(G)} = \frac{0.7072083 - 0.6534264}{\log 2} = 0.07759088.$$

Its closeness to 0 confirms that the two groups are really mixed (as mentioned above).

Now we simulate two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ from these distributions with $n = m = 50$ IID data in each group. The approximations of the entropies obtained with these samples and the exact PDF are

$$H(X|G=1) = 1 \approx \widehat{H}(X|G=1) = -\frac{1}{50}\sum_{i=1}^{50}\log f_1(X_i) = 1.130371$$

and

$$H(X|G=0)=0.3068528\approx\widehat{H}(X|G=0)=-\frac{1}{50}\sum_{i=1}^{50}\log f_0(Y_i)=0.2679194.$$

The entropy in the mixed population can be estimated in a similar way with

$$H(X)=0.7072083\approx\widehat{H}(X)=-\frac{1}{100}\left(\sum_{i=1}^{50}\log f(X_i)+\sum_{i=1}^{50}\log f(Y_i)\right)=0.7816074.$$

The entropy with two groups $H^{(2)}(X)$ can be approximated (by assuming $p=0.5$) with

$$H^{(2)}(X)=0.6534264\approx\widehat{H}^{(2)}(X)=0.5\cdot 1.130371+0.5\cdot 0.2679194=0.6991452.$$

The RED index is then approximated as

$$RED(2)\approx\frac{0.7816074-0.6991452}{\log 2}=0.1189678.$$

These approximations can be improved by increasing n and m. Note that p can also be estimated from the sample sizes (if we have a sample from the mixed population).

If we assume that the means of the exponential distributions are unknown and we estimate them with the sample means $\bar{X}=1.130371$ and $\bar{Y}=0.4805333$, we get the approximations $H(X|G=1)=1\approx 1.122546$ and $H(X|G=0)=0.3068528\approx 0.2671412$. The values $H(X)$ and $H^{(2)}(X)$ can be approximated in a similar way obtaining 0.7822224 and 0.6948435, respectively. Then the approximation of the RED index is 0.1260611. If we do not know that they come from exponential models, we can use empirical kernel estimators for f_1 and f_0 based on the respective samples.

Let us determine now the optimal regions to separate these two groups. In this example we can assume $R_1=[t,\infty)$ since $\mu_1=1>\mu_0=0.5$. The value of t for the optimal region under the maximum likelihood (or the maximum posterior probability) criterion is obtained by solving $f_1(t)=f_0(t)$ for $t>0$. This equation leads to the value $t_{ML}=\log 2=0.6931472$. The exact misclassification probabilities are

$$\Pr(X<t_{ML}|G=1)=F_1(\log 2)=1-e^{-\log 2}=1-0.5=0.5,$$

$$\Pr(X>t_{ML}|G=0)=1-F_0(\log 2)=e^{-2\log 2}=0.25$$

and the total misclassification probability with (known) prior probabilities $p=0.5$ and $1-p=0.5$ is

$$p\Pr(X<t_{ML}|G=1)+(1-p)\Pr(X>t_{ML}|G=0)=\frac{3}{8}=0.375.$$

If we want to use the criterion based on the entropy given in (10) for $R_1=[t,\infty)$, we get

$$H(t):=H(R_1)=-p_1\int_t^{\infty}f_1(x)\log f_1(x)dx-p_0\int_0^t f_0(x)\log f_0(x)dx,$$

where

$$p_1=\Pr(X\in R_1)=\Pr(X>t)=\frac{1}{2}\bar{F}_1(t)+\frac{1}{2}\bar{F}_0(t)=\frac{1}{2}e^{-t}+\frac{1}{2}e^{-2t}$$

and $p_0=1-p_1=\Pr(X\in R_0)=\Pr(X<t)$. A direct calculation leads to

$$H(t)=p_1\,(t+1)\,e^{-t}+p_0\left(1-\log 2+(-1+\log 2-2t)e^{-2t}\right).$$

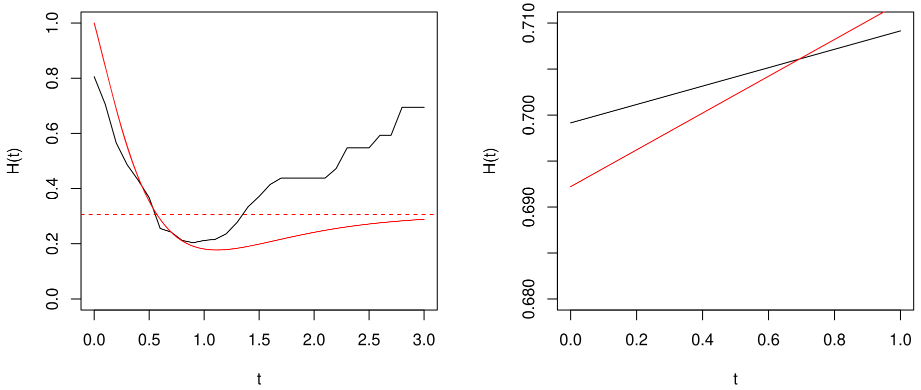

Fig. 1 Entropy function $H(t)$ (left, red) for $R_1 = [t, \infty)$. The red dashed line represents the entropy of the group $G = 0$. Empirical entropy function $\widehat{H}(t)$ (left, black) for the exponential distributions in Example 1. Empirical entropy functions $\widehat{H}_1(t)$ (right, black) and $\widehat{H}_0(t)$ (right, red) in this example

The plot can be seen in Fig. 1 (left, red). As stated above $H(\infty) = 0.3068528 = 1-\log(2) = H(X|G = 0)$ and $H(0) = 1 = H(X|G = 1)$. The optimal value with the minimum entropy criterion is $t_{ME} = 1.115213$ getting $H(t_{ME}) = 0.1776504$. With this value, the exact misclassification probabilities are

$$\Pr(X < t_{ME}|G = 1) = F_1(1.115213) = 0.6721546,$$

$$\Pr(X > t_{ME}|G = 0) = 1 - F_0(1.115213) = 0.1074826$$

and 0.3898186. The total misclassification probability is greater than that obtained with the ML criterion. By using the approximated version of this criterion given in (12), we obtain $\hat{t}_{ME} = 0.86534$. The plot can be seen in Fig. 1 (left, black). With this value, the exact misclassification probabilities are

$$\Pr(X < \hat{t}_{ME}|G = 1) = F_1(0.86534) = 0.57909,$$

$$\Pr(X > \hat{t}_{ME}|G = 0) = 1 - F_0(0.86534) = 0.17716$$

and 0.37813 which is again a little bit greater than the error obtained with the ML criterion.

If we use the criterion based on the empirical entropy (with known means) for the samples obtained above by replacing $n + 1$ with n, we get the functions $\widehat{H}_1$ and $\widehat{H}_0$ plotted in Fig. 1, right. In this case the values that lead to a classification in group 1 belong to $R_1 = [0.6931472, \infty)$. It coincides with the region determined by the maximum likelihood criterion since $p = 0.5 = n/(n + m)$.

In the second example we consider a mixture of two univariate normal (Gaussian) distributions. In this case, we replace the exact calculations with approximations.

Example 2 In the first case we consider a population obtained by mixing two normal distributions with means $\mu_1 = 2$ and $\mu_0 = -2$ and a common variance $\sigma^2 = 1$. To approximate the entropy functions we simulate two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from these distributions with $n = m = 50$ IID data in each group. The approximations of the entropies obtained with these samples and the exact PDF are

$$\widehat{H}(X|G = 1) = -\frac{1}{50}\sum_{i=1}^{50} \log f_1(X_i) = 1.339576$$

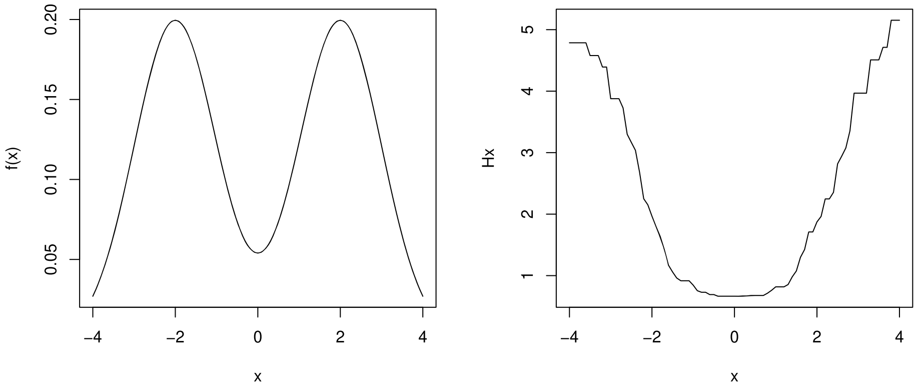

Fig. 2 Probability density function (left) of the mixture of two normal distributions considered in Example 2. Entropy function $\widehat{H}(x)$ (right) for $R_1 = [x, \infty)$

and

$$\widehat{H}(X|G=0) = -\frac{1}{50}\sum_{i=1}^{50} \log f_0(Y_i) = 1.331375.$$

The (common) exact value is $0.5 + 0.5\log(2\pi) = 1.418939$. The entropy in the mixed population can be estimated in a similar way with

$$\widehat{H}(X) = -\frac{1}{100}\left(\sum_{i=1}^{50} \log f(X_i) + \sum_{i=1}^{50} \log f(Y_i)\right) = 1.992107,$$

where $f = 0.5 f_1 + 0.5 f_0$. In this case, the mixed population has more uncertainty than the subpopulations, that is, $H(X) > H(X|G=i)$ for $i = 0, 1$. So we can consider the entropy with two groups approximated from (11) as

$$\widehat{H}^{(2)}(X) = -\frac{1}{100}\left(\sum_{i=1}^{50} \log f_1(X_i) + \sum_{i=1}^{50} \log f_0(Y_i)\right) = 1.335475.$$

Therefore, the division is effective $H(X) > H^{(2)}(X)$ and the approximated RED index is

$$RED(2) \approx \frac{1.992107 - 1.335475}{\log 2} = 0.9473197.$$

This value close to 1 indicates that the groups are well separated (as expected).

As $\mu_1 > \mu_0$ we can consider again the region $R_1 = [t, \infty)$ for the classification in the first group. Clearly, by using the maximum likelihood criterion, we get the optimal region $R_1 = [0, \infty)$ (see Fig. 2, left). To apply the minimum entropy criterion, we consider the function $H(t)$ approximated with (12), obtaining the plot given in Fig. 2, right. The minimum of this function is $\hat{t} = -0.01138$, a value close to the expected one ($t = 0$).

However, if we use the criterion based on the empirical entropies $\widehat{H}_1$ and $\widehat{H}_0$ with known means and variances, we get $t = -0.00205$, which is very close to the value obtained with the maximum likelihood criterion. Their plots can be seen in Fig. 3, left. If we replace $n+1$ with n we get $t = 0$. If the exact means and variances are replaced by their estimations from the samples we get $t = 0.06873$. We omit the plot since it is very similar to the one in Fig. 3, left.

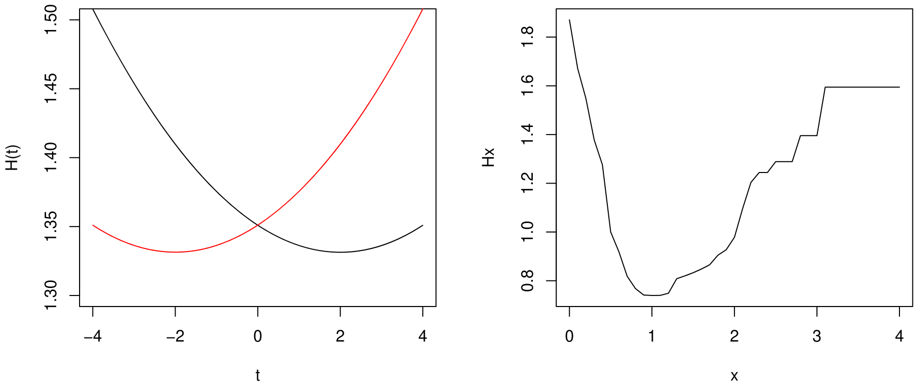

Fig. 3 Empirical entropy functions $\widehat{H}_1(t)$ (left, black) and $\widehat{H}_0(t)$ (left, red) for the normal distributions in Example 2 with known parameters. Empirical entropy function $\widehat{H}(t)$ (right) for $R_1 = [-t, t]$

In the second case, we consider $\mu_1 = \mu_0 = 0$ and $\sigma_1^2 = 1 < \sigma_0^2 = 4$. Then we can use the region $R_1 = [-t, t]$ for the classification in the first group. The estimations of the entropies in the groups and in the mixed populations are $H(X|G = 1) \approx 1.339576$, $H(X|G = 0) \approx 2.024522$ and $H(X) \approx 1.762427$. As in the first example, the entropy (uncertainty) in the second group is bigger than that in the mixed population (since the first population reduces uncertainty). The approximation for the entropy with two groups is $H^{(2)}(X) \approx 1.682049$. It reduces a little bit the global entropy $H(X)$ in the mixed population and the RED index is 0.1159609. By using $\widehat{H}(t)$ we get the region $R_1 = [-1.02890, 1.02890]$, see Fig. 3, right.

If we estimate $H_1(t)$ and $H_0(t)$, we obtain the plots given in Fig. 4 by using n (left) or $n+1$ (right). Note that in this case the results are very different. In the first case the optimal region is $R_1 = [-1.35956, 1.35956]$ (that coincides with the region of the maximum likelihood criterion) while in the second $R_1 = [-0.14788, 0.14788]$. The total misclassification probabilities are 0.3386627 and 0.4706897, respectively. The first value is actually the minimum error.

In the next example we show how to work with a real data set with four numerical variables and three groups.

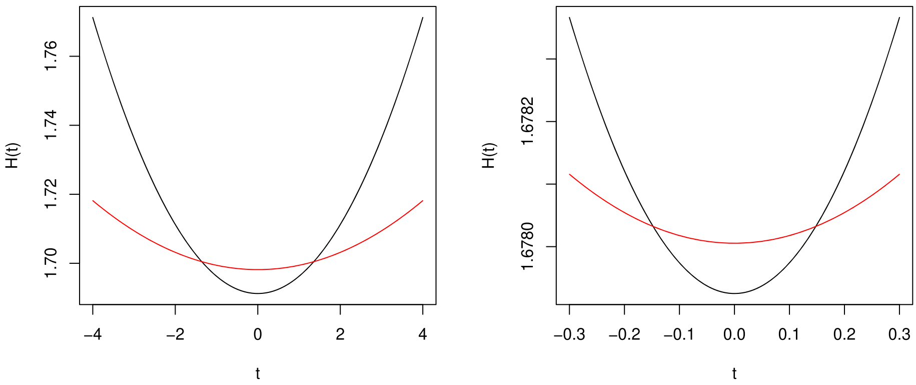

Fig. 4 Empirical entropy functions $\widehat{H}_1(t)$ (black) and $\widehat{H}_0(t)$ (red) for the normal distributions in Example 2 by using n (left) or $n+1$ (right) with known parameters

Example 3 Let us consider the `iris` data set available in the statistical program R. It contains the values in four variables (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) measured in 150 iris flowers from three different species: setosa ($G = 1$), versicolor ($G = 2$), and virginica ($G = 3$). There are 50 data from each specie with $\mathbf{X}_1, \dots, \mathbf{X}_{50} \in G_1$, $\mathbf{X}_{51}, \dots, \mathbf{X}_{100} \in G_2$, and $\mathbf{X}_{101}, \dots, \mathbf{X}_{150} \in G_3$. For another analysis of this data set using Deng extropy see Buono and Longobardi (2020).

Let us assume a Normal (Gaussian) distribution for these data in each group. As we know that there are three groups, we proceed as follows:

- We estimate the means and variance-covariance matrices in each group (by using only the data in each group).
- We use them to estimate the PDF f_i in each group by using normal PDF with these parameter values. The estimations of the respective PDF are represented by $\widehat{f_i}$ for $i = 1, 2, 3$.
- We approximate the entropies in the groups from (7) and in analogy with (1) for the multivariate case.

By using this procedure we obtain the following entropy values:

$$H(\mathbf{X}|G = 1) \approx \widehat{H}(\mathbf{X}|G = 1) := -\frac{1}{50}\sum_{i=1}^{50} \log \widehat{f}_1(\mathbf{X}_i) = -0.897926,$$

$$H(\mathbf{X}|G = 2) \approx \widehat{H}(\mathbf{X}|G = 2) := -\frac{1}{50}\sum_{i=51}^{100} \log \widehat{f}_2(\mathbf{X}_i) = 0.1985916,$$

and

$$H(\mathbf{X}|G = 3) \approx \widehat{H}(\mathbf{X}|G = 3) := -\frac{1}{50}\sum_{i=101}^{150} \log \widehat{f}_3(\mathbf{X}_i) = 1.172225.$$

These entropies show that the values of the flowers from the third group are more dispersed.

Analogously, we can estimate the PDF of the mixed population with $\widehat{f} = (\widehat{f}_1 + \widehat{f}_2 + \widehat{f}_3)/3$. We use this function to estimate the entropy of all the data (mixed population) with

$$H(\mathbf{X}) \approx \widehat{H}(\mathbf{X}) := -\frac{1}{150}\sum_{i=1}^{150} \log \widehat{f}(\mathbf{X}_i) = 1.219902.$$

Note that $\widehat{H}(\mathbf{X}) > \widehat{H}(\mathbf{X}|G = i)$ for $i = 1, 2, 3$ (although it is closed to $\widehat{H}(\mathbf{X}|G = 3)$). This may confirm the existence of the three groups.

We can also compare this entropy with the entropy without groups estimated as

$$H_{wg}(\mathbf{X}) \approx -\frac{1}{150}\sum_{i=1}^{150} \log \widehat{f}_{wg}(\mathbf{X}_i) = 2.532809,$$

where f_{wg} is the normal PDF with the mean and the variance-covariance matrix estimated with all the data together. As $\widehat{H}_{wg}(\mathbf{X}) >> \widehat{H}(\mathbf{X})$, this fact confirms the existence of the three groups.

Next we compare it with the entropy with three groups defined as in (4) with

$$H^{(3)}(\mathbf{X}) = p_1 H(\mathbf{X}|G = 1) + p_2 H(\mathbf{X}|G = 2) + p_3 H(\mathbf{X}|G = 3), \tag{13}$$

where $p_i = \Pr(G = i)$ for $i = 1, 2, 3$ are the prior probabilities. By assuming $p_i = 1/3$ for $i = 1, 2, 3$, we estimate it as

$$\widehat{H}^{(3)}(\mathbf{X}) = \frac{1}{3}\widehat{H}(\mathbf{X}|G = 1) + \frac{1}{3}\widehat{H}(\mathbf{X}|G = 2) + \frac{1}{3}\widehat{H}(\mathbf{X}|G = 3) = 0.1576302.$$

As $\widehat{H}^{(3)}(\mathbf{X}) < \widehat{H}(\mathbf{X})$, this fact might also confirm that the uncertainty is reduced by considering three groups. Hence

$$0 < Eff^{(3)}(\mathbf{X}) \approx \widehat{H}(\mathbf{X}) - \widehat{H}^{(3)}(\mathbf{X}) = 1.062272 \leq H_3(G) = \log(3) = 1.098612$$

and $RED(3) = 0.9669213$. This value confirms that the three groups can be separated.

We might wonder what happen if we just consider two groups. Note that in this case the estimation for $\widehat{H}(\mathbf{X})$ also changes (since we estimate f in different ways). The most efficient option is to join the groups two and three. The entropy of the new group is then 1.638049, obtaining $\widehat{H}^{(2)}(X) = 0.7927237$, $\widehat{H}(X) = 1.429234$ and

$$0 < Eff^{(2)}(X) \approx 0.6365099 < Eff^{(3)}(X) \approx 1.062272 \leq 1.098612.$$

Hence $RED(2) = 0.9182897 < RED(3) = 0.9669213$. With the other groups we get $Eff^{(2)}(X) \approx 0.5660699$ (join groups one and two) or 0.5355315 (join groups one and three). Therefore it is not a good idea to join these groups and it is better to consider the three initial groups.

We could also study what happens by considering just the two first groups (which are the least dispersed) by including the data of group 3 in groups 1 or 2. By applying the maximum likelihood criteria to do so, all the data from group 3 go to group 2 and so the result is the same as that stated above with $Eff^{(2)}(X) \approx 0.6365099$.

If we just consider the first 100 data, that belong to groups 1 and 2, then we get $\widehat{H}(X) = 0.34348$, $\widehat{H}^{(2)}(X) = -0.3496672$ and $RED(2) = 1$. Therefore, these two groups are completely separated. This is not the case if we just consider the data from groups 2 and 3. In this case we get $\widehat{H}(X) = 0.8302966$, $\widehat{H}^{(2)}(X) = 0.6854083$ and $RED(2) = 0.2090297$. Therefore, these two groups are mixed.

The PCA plot with the two first principal components for these three groups can be seen in Fig. 5. Note that our conclusions based on the RED index are consistent with the different groups in that figure.

If we want to use these entropy functions to classify a new flower with measures $\mathbf{z} = (z_1, z_2, z_3, z_4)$, we just compute the entropy $\widehat{H}^{(3)}$ by assuming that $\mathbf{z}$ belongs to each group. Then it is classified in the group with the minimum entropy. However, it is not easy to determine the classification regions in $\mathbb{R}^4$ obtained with this criterion for each group.

For example, for the first flower in this data set with $\mathbf{z} = (5.1, 3.5, 1.4, 0.2)$ and replacing $n + 1$ with n, we get the approximations

$$\begin{aligned} H^{(3)}(X|\mathbf{z} \in G_1) \approx\ & \frac{\widehat{H}(X|G = 1) + \widehat{H}(X|G = 2) + \widehat{H}(X|G = 3)}{3} \\ & - \frac{\log(\widehat{f_1}(\mathbf{z}))}{150} = 0.1400744, \\ H^{(3)}(X|\mathbf{z} \in G_2) \approx\ & \frac{\widehat{H}(X|G = 1) + \widehat{H}(X|G = 2) + \widehat{H}(X|G = 3)}{3} \\ & - \frac{\log(\widehat{f_1}(\mathbf{z}))}{150} = 0.5285691, \end{aligned}$$

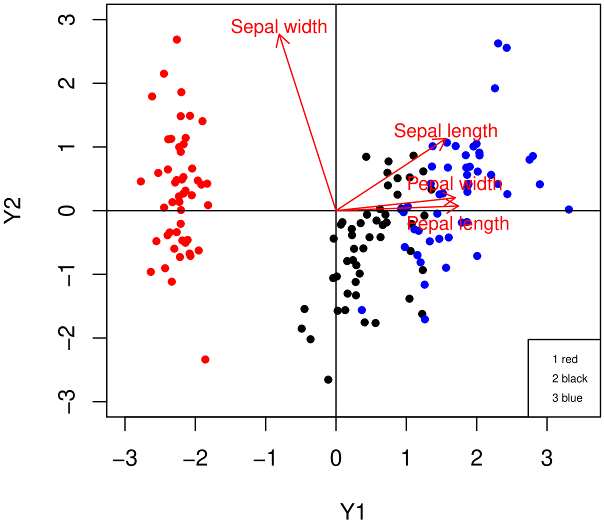

Fig. 5 PCA plot of the iris data set studied in Example 3

and

$$H^{(3)}(X|\mathbf{z} \in G_3) \approx \frac{\widehat{H}(X|G=1)+\widehat{H}(X|G=2)+\widehat{H}(X|G=3)}{3} - \frac{\log(\widehat{f_1}(\mathbf{z}))}{150} = 0.7621647.$$

Hence with the minimum entropy criterion it is classified (correctly) in the first group. As the sample sizes of the groups in the training sample coincide and the prior probabilities are equal, this classification criterion is equivalent to the maximum likelihood criterion (under normality) and to the classical Quadratic Discriminant Analysis (QDA) since we have used the normal PDF. This is not the case if the prior probabilities are unequal. It is also different if the PDF of the groups are estimated with nonparametric techniques.

If we do not replace n with $n+1$, we get the estimations

$$H^{(3)}(X|\mathbf{z} \in G_1) \approx \frac{\widehat{H}(X|G=2)+\widehat{H}(X|G=3)}{3} + \frac{50\widehat{H}(X|G=1)-\log(\widehat{f_1}(\mathbf{z}))}{153} = 0.1462874,$$

$$H^{(3)}(X|\mathbf{z} \in G_2) \approx \frac{\widehat{H}(X|G=1)+\widehat{H}(X|G=3)}{3} + \frac{50\widehat{H}(X|G=2)-\log(\widehat{f_2}(\mathbf{z}))}{153} = 0.5199978,$$

and

$$H^{(3)}(X|\mathbf{z} \in G_3) \approx \frac{\widehat{H}(X|G=1)+\widehat{H}(X|G=2)}{3} + \frac{50\widehat{H}(X|G=3)-\log(\widehat{f_3}(\mathbf{z}))}{153} = 0.7426495,$$

where $\mathbf{z}$ is not used to compute $\widehat{f_i}$ (i.e. to compute the mean and the covariance matrix of group i). Again, it is classified correctly in the first group. These entropy values play a role similar to the role played by the posterior probabilities in the classical QDA showing the "reliability" (margins) of these classifications. Note that by adding just one data in a wrong group might increase the entropy considerably. We do the same with all the 150 flowers of the data set. If we replace $n+1$ with n, the classification is correct in 147 cases. In particular, it fails for two flowers in the second group (classified in the third group) and one in the third one (classified in the second group). If we do not replace $n+1$ with n, 147 flowers are classified correctly and the three failures occur for flowers in the second group that are classified in the third one. In this case, the group with the biggest entropy may attract more data (since their values are more dispersed).

In the last example, as in Biernacki et al. (1999), we consider bivariate Gaussian distributions to study the evolution of the RED index when we change the means.

Example 4 Let us consider a mixture model with equal proportions of two Gaussian distributions. We simulate a sample of size 100 from a bivariate normal distribution with mean $\mu_1 = (0, 0)$ and variance-covariance matrix $\Sigma_1 = I_2$, that is $\mathbf{X}_1, \ldots, \mathbf{X}_{100} \in G_1$, and samples of size 100 from bivariate normal distributions with means $\mu_2 = (d, 0)$ and variance-covariance matrix $\Sigma_2 = I_2$, by varying d from 0 to 5 in steps of 0.1, i.e., $\mathbf{X}_{101}, \ldots, \mathbf{X}_{200} \in G_2$. We use the data in each group to estimate the means and the variance-covariance matrices and then to obtain the estimated PDF $\widehat{f_1}$ and $\widehat{f_2}$ by using normal distributions with these parameters. Then, we estimate the PDF $\widehat{f}$ of the mixed population by the arithmetic mean of the estimated PDF (since we are assuming a mixture model with equal proportions). Thus, we can estimate the entropies in the groups by

$$H(\mathbf{X}|G=1) \approx \widehat{H}(\mathbf{X}|G=1) := -\frac{1}{100}\sum_{i=1}^{100} \log \widehat{f_1}(\mathbf{X}_i),$$

$$H(\mathbf{X}|G=2) \approx \widehat{H}(\mathbf{X}|G=2) := -\frac{1}{100}\sum_{i=101}^{200} \log \widehat{f_2}(\mathbf{X}_i)$$

and the entropy of the mixed population with

$$H(\mathbf{X}) \approx \widehat{H}(\mathbf{X}) := -\frac{1}{200}\sum_{i=1}^{200} \log \widehat{f}(\mathbf{X}_i).$$

Then, we estimate $H^{(2)}(\mathbf{X})$ by

$$\widehat{H}^{(2)}(\mathbf{X}) = \frac{\widehat{H}(\mathbf{X}|G=1) + \widehat{H}(\mathbf{X}|G=2)}{2}$$

and the relative efficiency of the division in two groups as

$$RED(2) \approx \frac{\widehat{H}(\mathbf{X}) - \widehat{H}^{(2)}(\mathbf{X})}{\log 2}.$$

The results are shown in Fig. 6, left, as a function of d (black points). Moreover, we can estimate the mean and the variance-covariance matrix without assuming the existence of groups and then obtain an estimate of the entropy without groups as

$$H_{wg}(\mathbf{X}) \approx \widehat{H}_{wg}(\mathbf{X}) := -\frac{1}{200}\sum_{i=1}^{200} \log \widehat{f_{wg}}(\mathbf{X}_i).$$

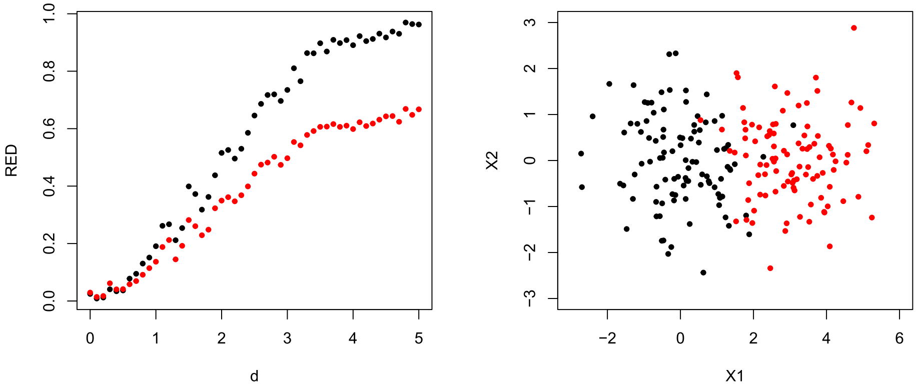

Fig. 6 Relative efficiency of division (left) in two (black) and three (red) groups for simulated fifty-fifty mixtures from bivariate normal distributions with means $\mu_1 = (0, 0)$, $\mu_2 = (d, 0)$ and variance-covariance matrices $\Sigma_1 = \Sigma_2 = I_2$. In the right plot we can see the sample values for $d = 3$ with $RED(2) = 0.7347749$

Hence, we compare the values of $\widehat{H}_{wg}(\mathbf{X})$ and $\widehat{H}(\mathbf{X})$ and obtain that the former is lower than the latter only for d equal to $0.1,\ 0.2,\ 0.4,\ 0.6$ and 1.9 confirming the existence of two groups with the increase of d. Further, we may suppose the existence of a third group and divide the data of the second group in two groups of 50 data, that is $\mathbf{X}_{101}, \dots, \mathbf{X}_{150} \in G_2$ and $\mathbf{X}_{151}, \dots, \mathbf{X}_{200} \in G_3$. In analogy with what we have done above, we estimate the mean and variance-covariance matrices of the new groups and then the entropies of the groups. In the mixture, the second and the third group have a weight of 0.25, so the estimate of $H^{(3)}(\mathbf{X})$ is given by

$$\widehat{H}^{(3)}(\mathbf{X}) = 0.5 \cdot \widehat{H}(\mathbf{X}|G = 1) + 0.25 \cdot \widehat{H}(\mathbf{X}|G = 2) + 0.25 \cdot \widehat{H}(\mathbf{X}|G = 3),$$

and the relative efficiency of the division in three groups is

$$RED(3) \approx \frac{\widehat{H}(\mathbf{X}) - \widehat{H}^{(3)}(\mathbf{X})}{0.5 \log 2 + 0.25 \log 4 + 0.25 \log 4}.$$

In Fig. 6, left, we also plot the values of $RED(3)$ (red points) as a function of d and we can compare them with the values of $RED(2)$ (black points). We note that, as expected, the values of $RED(3)$ are dominated by the values of $RED(2)$ and the former is slightly higher than the latter only for small choices of d $(0,\ 0.1,\ 0.2,\ 0.3,\ 0.4$ and $0.5)$. In Fig. 6, right, we plot the samples for $d = 3$. Note that the value $RED(2) = 0.7347749$ for $d = 3$ allows us to detect the existence of the two groups even when they are really close.

We repeat the same experiment by choosing $\mu_2 = (d, d)$, varying d from 0 to 5 in steps of 0.1. In this case, the value of the entropy without groups is lower than the value of the entropy with two groups for $d \in \{0, 0.1, 0.2, 0.4, 0.5, 0.6, 0.7\}$. Moreover, we again consider the possibility of dividing the second group in two groups and we obtain a value of $RED(3)$ higher than $RED(2)$ only with $d = 0.1$. The results are shown in Fig. 7, left, where we also plot the samples (right) for $d = 2$. Again the value $RED(2) = 0.7930995$ for $d = 2$ shows the existence of the two groups even when they are really close.

By comparing the values of $RED(2)$ in Fig. 7, left, and in Fig. 6, left, it is possible to observe a faster tendency to one in the case in which $\mu_2 = (d, d)$ due to the higher distance between the means of the mixed populations.

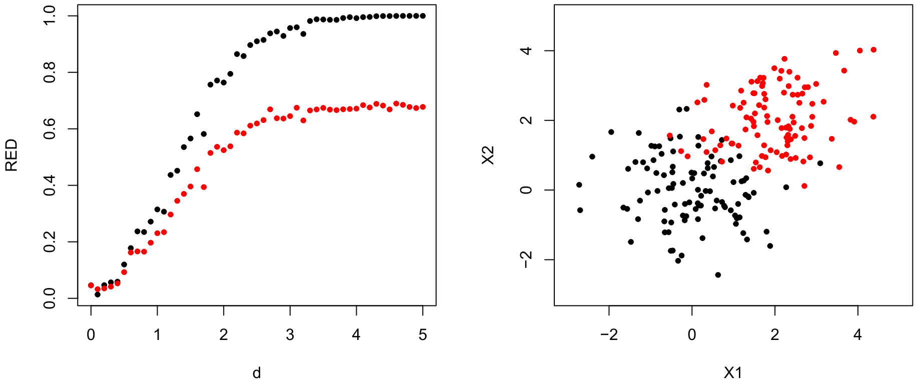

Fig. 7 Relative efficiency of division RED (left) in two (black) and three (red) groups for simulated fifty-fifty mixtures from bivariate normal distributions with means $\mu_1 = (0, 0)$, $\mu_2 = (d, d)$ and variance-covariance matrices $\Sigma_1 = \Sigma_2 = I_2$. In the right plot we can see the sample values for $d = 2$ with $RED(2) = 0.7930995$

5 Application to Variable Selection in Omic Data

In this section we study the performance of the proposed RED index when applied to variable selection in biological omic data. One of the main characteristics of omic data sources is concerned with the high dimensionality of the data sets due to the development of high-throughput technologies that allow the simultaneous monitoring of hundreds or thousands of biological variables from different layers of biological information such as genes, proteins, RNA and metabolites. Actually, the data sources generated by these technologies have given rise to the so-called omic data sources as well as the need of ad hoc exploratory data analysis tools for analyzing such high-dimensional data. One of the challenges settled by biologists and geneticists is concerned with the identification of the most informative omic variables for explaining a specific clinical outcome such as disease or the evolution of a disease in the response of patients to a specific drug. Hence, the challenge is to carry out variable selection for identifying those variables that discriminate the outcome and, as a result, eliminate the noisy inputs. In this section we show how the RED index can be used as a tool for variable selection when applied to a well-known microarray gene expression colon cancer data set.

The genomic study consists of gene expression levels for 40 tumor and 22 normal tissue samples collected by the Affymetrix oligonucleotide Hum-6000 array complementary to more than 6500 human genes from which only 2000 *genes* with the highest minimal intensity across samples are retained, see Alon et al. (1999). Hence, we end up with data set containing the expression levels for 2000 genes arranged in a matrix with 2000 columns and 62 rows, along with a clinical outcome related to the status of each tissue sample: tumor versus healthy. This gene expression data set is a classic in the literature and can be downloaded from the R package colonCA, see Sylvia (2019).

Some data preprocessing about robust normalization of gene expression measures following previous work by Arevalillo and Navarro (2013) is carried out. Then the RED index is estimated for the two group case (tumor and healthy outcomes) in order to generate a ranking that helps to sort the genes in accordance to their relevance for discriminating the clinical outcome. The results are provided by the gene ranking appearing in Fig. 8 which displays the whole ranking (left) and the top 13 genes with $RED > 0.5$ (right).

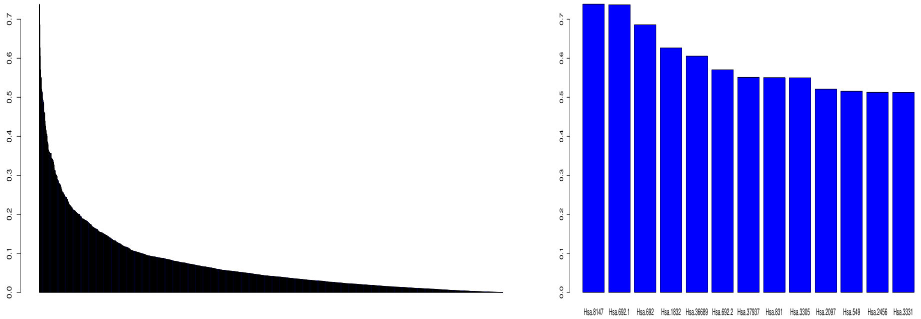

Fig. 8 Ranking of the genes obtained according to the RED index (left) and top ranked genes (right)

The genes at the top of the ranking have been previously described as relevant biomarkers of colon cancer. Table 1 shows the Hsa identifiers and the gene descriptions of the top genes having RED greater than 0.5.

The genes with identifiers Hsa.8147, Hsa.692, Hsa. 692.1 and Hsa.692.2 exhibit a high degree of co-expression as measured by correlation coefficients around 0.9. We now assess the RED score that results by considering pairs of genes in order to elucidate whether their joint behavior has a stronger impact than their marginal behavior at discriminating the clinical outcome. As highly correlated genes convey redundant expression measures, we only consider gene pairings having correlations below the 0.90 threshold for estimating their RED scores; this is achieved by selecting pairwise gene associations corresponding to the top RED parings having correlations lower than the 0.90 threshold.

The scatter plots depicted by Fig. 9 show the selected gene pairings; in all the cases the RED score is higher than the individual RED values previously obtained for the genes Hsa.36689, Hsa.692.1, Hsa.8147 and Hsa.2456 given by 0.605, 0.737, 0.739 and 0.513

Table 1 Hsa identifiers and gene description of the top genes selected from the RED ranking when $RED(2) > 0.5$

Hsa Id	Gene description	*RED*(2)
Hsa.8147	Human desmin gene, complete cds	0.739
Hsa.692.1	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0.737
Hsa.692	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0.686
Hsa.1832	Myosin regulatory light chain 2, smooth muscle isoform (human)	0.626
Hsa.36689	H.sapiens mRNA for GCAP-II/uroguanylin precursor	0.605
Hsa.692.2	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0.570
Hsa.37937	Myosin heavy chain, nonmuscle	0.551
Hsa.831	Mitochondrial matrix protein P1 precursor (human)	0.550
Hsa.3305	Tropomyosin alpha chain, smooth muscle (human)	0.550
Hsa.2097	Single-strand binding protein (Escherichia coli)	0.521
Hsa.549	Transcription factor IIIA	0.515
Hsa.2456	Human MaxiK potassium channel beta subunit mRNA complete cds	0.513
Hsa.3331	Nucleoside diphosphate kinase (human)	0.512

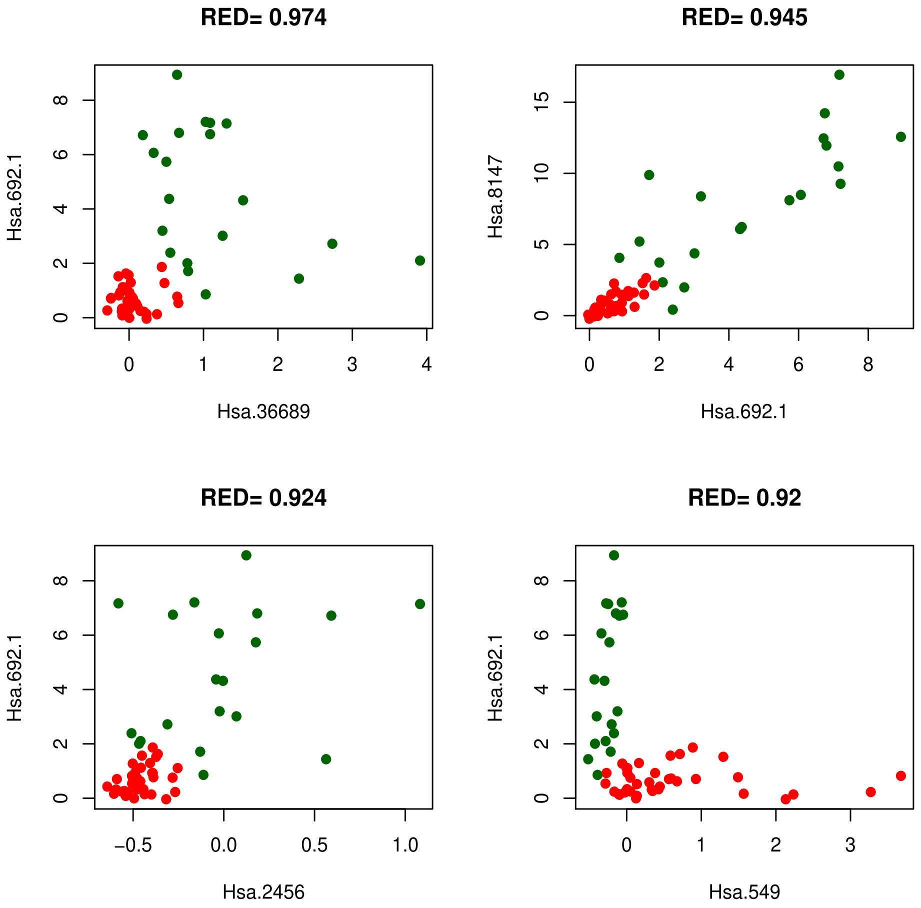

Fig. 9 Gene pairings with the highest RED values

respectively. Note that with just two genes, Hsa.36689 and Hsa.692.1, we get a RED index equal to 0.974 and a very good separation of these two groups. The other pairs also show high RED indices which informed about the strong bivariate differential expression patterns depicted by the scatter plots of Fig. 9.

6 Conclusions

We have provided new tools based on Shannon entropy to study data from a population with groups. This paper is just a first step and the potential applications are countless. The main one is the RED index. The illustrative examples show that this is a good tool to measure the separation of the groups. The main advantage is that it does not depend on the number of unknown parameters in the model. The new classification techniques also lead to promising results (similar to the ones obtained with classical discrimination measures).

There are several tasks for future research. Maybe, the main one could be to apply these tools to cluster analysis (unsupervised techniques) in order to determine the optimal number of clusters in a population. Applications to specific data sets in different research areas are obvious chores to be done.

Appendix. Proofs

Proof of Proposition 1 If S_1 and S_0 are the respective supports of f_1 and f_0, the entropy of the two groups can be written as

$$\begin{aligned} H^{(2)}(X) &= pH(X|G=1) + (1-p)H(X|G=0) \\ &= -p\int_{S_1} f_1(x)\log f_1(x)dx - (1-p)\int_{S_0} f_0(x)\log f_0(x)dx \\ &= -p\int_{S_1} f_1(x)\log\frac{f_1(x)}{f(x)}dx - p\int_{S_1} f_1(x)\log f(x)dx \\ &\quad -(1-p)\int_{S_0} f_0(x)\log\frac{f_0(x)}{f(x)}dx - (1-p)\int_{S_0} f_0(x)\log f(x)dx \\ &= -p\int_{S_1\cup S_0} f_1(x)\log\frac{f_1(x)}{f(x)}dx - (1-p)\int_{S_0\cup S_1} f_0(x)\log\frac{f_0(x)}{f(x)}dx \\ &\quad -p\int_{S_1\cup S_0} f_1(x)\log f(x)dx - (1-p)\int_{S_0\cup S_1} f_0(x)\log f(x)dx \\ &= H(X) - p\,KL(f_1|f) - (1-p)\,KL(f_0|f). \end{aligned}$$

Therefore

$$Eff^{(2)}(X) = H(X) - H^{(2)}(X) = p\,KL(f_1|f) + (1-p)\,KL(f_0|f).$$

Hence $Eff^{(2)}(X) \geq 0$ since the KL-measure is non-negative.

To get the upper bound we note that

$$\begin{aligned} KL(f_1|f) &= \int_{S_1\cup S_0} f_1(x)\log\frac{pf_1(x)}{pf(x)}dx \\ &= \int_{S_1\cup S_0} f_1(x)\log\frac{pf_1(x)}{f(x)}dx - \int_{S_1\cup S_0} f_1(x)\log(p)dx \\ &\leq -\int_{S_1\cup S_0} f_1(x)\log(p)dx = -\log(p), \end{aligned}$$

where the inequality holds since $0 \leq pf_1(x) \leq f(x)$. Analogously, it can be proved that $KL(f_0|f) \leq -\log(1-p)$. Hence, from (5), we get

$$\begin{aligned} Eff^{(2)}(X) &= pKL(f_1|f) + (1-p)KL(f_0|f) \\ &\leq -p\log(p) - (1-p)\log(1-p) = H_2(G). \end{aligned}$$

Moreover, if the $f_1 \neq f_0$ (a.e), then $f_i \neq f$ (a.e) and $KL(f_i|f) > 0$ for $i = 1, 2$. Then $Eff^{(2)}(X) > 0$ for all $p \in (0, 1)$.

Proof of Proposition 2 From the definition we have

$$H^{(2)}(X) = p_1H(X|G=1) + (1-p_1)H(X|G=0).$$

On the other hand, from Proposition 1, we get

$$H(X|G=0) \geq qH(X|G=2) + (1-q)H(X|G=3).$$

Replacing $H(X|G=0)$ with this expression we get $H^{(2)}(X) \geq H^{(3)}(X)$. Hence, the result for the efficiency also holds. The bounds are obtained as in Proposition 1.

Acknowledgements The authors thank an anonymous reviewer for several helpful suggestions.

Author Contributions The authors contributed equally to this work.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. JN is partially supported by "Ministerio de Ciencia e Innovación" of Spain under grant PID2019-103971GB-I00/AEI/10.13039/501100011033. FB is member of the research group GNAMPA of INdAM (Istituto Nazionale di Alta Matematica) and is partially supported by MIUR-PRIN 2017, project "Stochastic Models for Complex Systems", no. 2017 JFFHSH. JMA acknowledges NextGenerationEU support.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no conflict of interest.

References

Ahmadi J, Di Crescenzo A, Longobardi M (2015) On dynamic mutual information for bivariate lifetimes. Adv Appl Probab 47:1157–1174

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 96:6745–6750. https://doi.org/10.1073/pnas.96.12.6745

Arevalillo JM, Navarro H (2013) Exploring correlations in gene expression microarray data for maximum predictive-minimum redundancy biomarker selection and classification. Comput Biol Med 43:1437–1443

Balakrishnan N, Buono F, Longobardi M (2022) On cumulative entropies in terms of moments of order statistics. Methodol Comput Appl Probab 24:345–359

Biernacki C, Celeux G, Govaert G (1999) An improvement of the NEC criterion for assessing the number of clusters in a mixture model. Pattern Recogn Lett 20:267–272

Briët J, Harremoës P (2009) Properties of classical and quantum Jensen? Shannon divergence. Phys Rev A 79:283–304

Buono F, Longobardi M (2020) A dual measure of uncertainty: the Deng extropy. Entropy 22:582. https://doi.org/10.3390/e22050582

Celeux G, Soromenho G (1996) An entropy criterion for assessing the number of clusters in a mixture model. J Classif 13:195–212

Cover TM, Thomas JA (2006) Elements of Information Theory, 2nd edn. Wiley, Hoboken, NJ, USA

Di Crescenzo A, Longobardi M (2002) Entropy-based measure of uncertainty in past lifetime distributions. J Appl Probab 39:434–440

Di Crescenzo A, Longobardi M (2006) On weighted residual and past entropies. Sci Math Jpn 64(2):255–266

Di Crescenzo A, Paolillo L, Suárez-Llorens A (2021) Stochastic comparisons, differential entropy and varentropy for distributions induced by probability density functions. https://doi.org/10.48550/arXiv.2103.1108

Grandvalet Y, Bengio Y (2005) Semi-supervised learning by entropy minimization. Proc Adv Neural Inf Process Syst 529–536

Melbourne J, Talukdar S, Bhaban S, Madiman M, Salapaka MV (2022) The differential entropy of mixtures: New bounds and applications. IEEE Trans Inf Theory 68:2123–2146

Moshksar K, Khandani AK (2016) Arbitrarily tight bounds on differential entropy of Gaussian mixtures. IEEE Trans Inf Theory 62:3340–3354

Rao M, Chen Y, Vemuri B, Wang F (2004) Cumulative residual entropy: a new measure of information. IEEE Trans Inf Theory 50:1220–1228

Rényi A (1961) On measures of information and entropy. In: Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability pp 547–561

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:279–423

Sylvia M (2019) colonCA: exprSet for Alon et al. (1999) colon cancer data. R package version 1.28.0

Tsallis C (1988) Possible generalization of Boltzmann-Gibbs statistic. J Stat Phys 52:479–487

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.