Computational Statistics and Data Analysis 89 (2015) 72–84

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Non-parametric entropy estimators based on simple linear regression

Hideitsu Hino [a,*], Kensuke Koshijima [b], Noboru Murata [b]

[a] Department of Computer Science, University of Tsukuba, 1-1-1 Tennodai, Ibaraki, Tsukuba 305-8573, Japan
[b] School of Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, Japan

ARTICLE INFO

Article history:
Received 11 July 2014
Received in revised form 10 March 2015
Accepted 13 March 2015
Available online 20 March 2015

Keywords:
Entropy estimation
Non-parametric
Simple linear regression

ABSTRACT

Estimators for differential entropy are proposed. The estimators are based on the second order expansion of the probability mass around the inspection point with respect to the distance from the point. Simple linear regression is utilized to estimate the values of density function and its second derivative at a point. After estimating the values of the probability density function at each of the given sample points, by taking the empirical average of the negative logarithm of the density estimates, two entropy estimators are derived. Other entropy estimators which directly estimate entropy by linear regression, are also proposed. The proposed four estimators are shown to perform well through numerical experiments for various probability distributions.

1. Introduction

Let X be a p-dimensional random variable with probability density function (pdf) $f(x)$, then its differential entropy (Cover and Thomas, 1991; Shannon, 1948) is defined by

$$H(f) = -\int f(x)\ln f(x)\mathrm{d}x. \tag{1}$$

We assume that $H(f)$ is well-defined and finite. The differential entropy plays a central role not only in information and communication theory, but also in statistics (Tarasenko, 1968; Vasicek, 1976; Hino et al., 2013), signal processing (Comon, 1994; Learned-Miller and Fisher, 2004), machine learning and pattern recognition (Mannor et al., 2005; Rubinstein and Kroese, 2004; Hino and Murata, 2010, 2013). For a concrete example, the differential entropy is used as a criterion for independence in the literature of independent component analysis (ICA; Comon, 1994; Hyvärinen et al., 2001). In ICA, mixed signals are decomposed into statistically independent signals. A sum of the marginal entropies $\sum_{k=1}^{p} H(X_k)$ is an upper bound of the joint entropy $H(X_1, \ldots, X_p)$, where p is the number of observed source signals. Since the gap between the sum of marginal entropies and joint entropy is zero if and only if signals are independent, signal decomposition is sometimes done by transforming the p observed signals into p signals $X_k,\ k = 1, \ldots, p$ so that the quantity $\sum_{k=1}^{p} H(X_k) - H(X_1, \ldots, X_p)$ is minimized. The entropy can be estimated by plugging in the estimate of a pdf, however, density estimation for high dimensional data is difficult and computationally demanding. Direct entropy estimators often offer better results.

Consider the problem of estimating the entropy $H(f)$ using a set of observed samples $\mathcal{D} = \{x_i\}_{i=1}^{n}$, where $x_i,\ i = 1, \ldots, n$ are realizations of a random variable X with a pdf $f(x)$. Since entropy estimation is often required in exploratory data

* Corresponding author. Tel.: +81 293 53 5538; fax: +81 293 53 5538.
E-mail address: hinohide@cs.tsukuba.ac.jp (H. Hino).

http://dx.doi.org/10.1016/j.csda.2015.03.011
0167-9473/

analysis, it is preferable not to assume any specific form of probability distribution behind the data, and therefore non-parametric approach is often of the choice. There are several non-parametric methods for estimating the differential entropy of a continuous random variable. One of the simplest methods is the plug-in estimate, which is based on a density estimate $\hat{f}(x)$ of $f(x)$. Once we obtain an estimate $\hat{f}(x)$ using samples $\mathcal{D}$, the differential entropy can be estimated by numerically integrating $\hat{f}(x)\ln\hat{f}(x)$. Since numerical integration is unstable and computationally demanding when p, the dimensionality of X, is large, it is suggested by Joe (1989) to use re-substitution

$$\hat{H}(\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n}\ln\hat{f}(x_i) \tag{2}$$

instead of numerical integration. With a kernel density estimate $\hat{f}_\kappa(x)$, which will be defined later, some asymptotics are investigated for the plug-in estimator $\hat{H}(\mathcal{D})$ with univariate (Ahmad and Lin, 1976) and multivariate (Joe, 1989) cases, respectively.

Another popular approach for entropy estimation is based on the k-nearest neighbor (k-NN) method. An entropy estimator using 1-NN is proposed by Kozachenko and Leonenko (1987), and its mean-square consistency is proved for any dimension. This result is extended to develop a k-NN-based estimator (Goria et al., 2005), which includes the spacing entropy estimator (Vasicek, 1976; Dudewicz and van der Meulen, 1981; Hall, 1986) as a special case of $p = 1$. For more extension and theoretical developments, see Beirlant et al. (1997); Györfi and van der Meulen (1987); Paninski (2003); Pérez-Cruz (2008) for examples. Due to the resemblance to the proposed method, the k-NN entropy estimator is explained later in more detail.

In this paper, we propose novel non-parametric entropy estimators based on the second order expansion of probability mass function and simple linear regression. The proposed methods are conceptually simple with almost no tuning parameter.

The rest of this paper is organized as follows. Section 2 formulates the problem of density and entropy estimation. In Section 3, novel entropy estimators based on second order expansion of probability mass function and simple linear regression are proposed. Experimental results are given in Section 4. The last section is devoted to concluding remarks.

2. Preliminary and notation

As a building block of an entropy estimator, consider the problem of estimating pdf $f(z)$ at an inspection point $z \in \mathbb{R}^p$ from a set of observations $\mathcal{D} = \{x_i\}_{i=1}^n$.

Let $\|x_i - z\|$ be the Euclidean distance between the inspection point z and the ith sample x_i, and let $b(z;\varepsilon) = \{x \in \mathbb{R}^p \mid \|x - z\| < \varepsilon\}$ be an ε-ball centered at z with volume $|b(z;\varepsilon)| = c_p\varepsilon^p$, where $c_p = \pi^{p/2}/\Gamma(p/2+1)$, and $\Gamma(\cdot)$ is the gamma function. Denote the probability mass contained within the ε-ball centered at z by

$$q_z(\varepsilon) = \int_{x\in b(z;\varepsilon)} f(x)\mathrm{d}x. \tag{3}$$

Expanding the integrand, we obtain

$$\begin{aligned} q_z(\varepsilon) &= \int_{x\in b(z;\varepsilon)} \left\{f(z) + (x-z)^\top \nabla f(z) + O(\varepsilon^2)\right\}\mathrm{d}x \\ &= |b(z;\varepsilon)|\left(f(z) + O(\varepsilon^2)\right) = c_p\varepsilon^p f(z) + O(\varepsilon^{p+2}). \end{aligned}$$

In the above expansion, $(x-z)$ is of order ε because the integration is within the ε-ball. The term with first derivative of the density function vanishes due to symmetry. Ignoring the second term in the expansion and approximating the probability mass $q_z(\varepsilon)$ with the ratio of the number of points within the ε-ball, we obtain a density estimator

$$\hat{f}_\varepsilon(z) = \frac{k_\varepsilon}{nc_p\varepsilon^p}, \tag{4}$$

where k_ε is the number of samples that fall in the ε-ball. This estimator is nothing but the kernel density estimator (Wand and Jones, 1994b)

$$\hat{f}_\kappa(x) = \frac{1}{n\varepsilon^p}\sum_{i=1}^{n}\kappa(\|x - x_i\|/\varepsilon) \tag{5}$$

with the hard window kernel function

$$\kappa(x) = \frac{1}{c_p}\mathbb{1}\{\|x\| \le 1\}, \tag{6}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Here ε is the bandwidth parameter in the context of kernel density estimator. On the other hand, when k, the number of neighbors from the inspection point z, is fixed instead of ε, the estimator $\hat{f}_\varepsilon(z)$ in Eq. (4) is rewritten as $\hat{f}_k(z) = k/(nc_p\varepsilon_k^p)$, where ε_k is determined by the distance from the inspection point to its k-th nearest point. The estimator $\hat{f}_k(z)$ is called the k-NN density estimator (Loftsgaarden and Quesenberry, 1965; Mack and Rosenblatt, 1979; Moore and Yackel, 1977).

By averaging $-\ln \hat{f}_{k,i}(x_i)$, where $\hat{f}_{k,i}(x_i)$ is estimated using $\mathcal{D} \setminus \{x_i\}$, we obtain the k-NN entropy estimator

$$\hat{H}_k(\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_{k,i}(x_i). \tag{7}$$

3. Proposed entropy estimator

In this section, we will propose entropy estimators based on the second order expansion of the probability mass $q_z(\varepsilon)$, and simple linear regression.

Proposition 1. *The probability mass $q_z(\varepsilon)$ of the ε-ball centered at z is expressed in the form*

$$q_z(\varepsilon) = c_p f(z)\varepsilon^p + \frac{1}{4(p/2+1)} c_p \varepsilon^{p+2} \mathrm{Tr}\nabla^2 f(z) + O(\varepsilon^{p+4}). \tag{8}$$

See Appendix A for the proof of Proposition 1.

By empirically approximating $q_z(\varepsilon)$ in Eq. (8) by the ratio k_ε/n, and dividing the equation by $c_p\varepsilon^p$, we obtain

$$\frac{k_\varepsilon}{n c_p \varepsilon^p} = f(z) + C(z)\varepsilon^2 + O(\varepsilon^4), \quad C(z) = \frac{\mathrm{Tr}\nabla^2 f(z)}{4(p/2+1)}. \tag{9}$$

Furthermore, denoting $Y_\varepsilon = k_\varepsilon/(nc_p\varepsilon^p)$ and $X_\varepsilon = \varepsilon^2$, and ignoring the higher order term with respect to ε, we obtain the following linear relationship

$$Y_\varepsilon \simeq f(z) + C(z)X_\varepsilon \tag{10}$$

with respect to the explanatory variable X_ε and the response variable Y_ε. Based on this relationship, we derive entropy estimators in the following sections.

3.1. Leave one out estimator with least squares

The relationship in Eq. (10) is regarded as a linear equation between an explanatory variable X_ε and a response variable Y_ε. We propose to select a small number of radii of the balls, $\mathcal{E} = \{\varepsilon_1, \ldots, \varepsilon_m\}$, $m < n$, and fit a linear model to the set of points $\{(X_\varepsilon, Y_\varepsilon)\}_{\varepsilon\in\mathcal{E}}$ by minimizing the sum of squared residuals

$$R = \frac{1}{m}\sum_{\varepsilon\in\mathcal{E}} (Y_\varepsilon - f(z) - C(z)X_\varepsilon)^2 \tag{11}$$

with respect to $f(z)$ and $C(z)$. The line fitting is possible with at least two points, but for stable fitting, the number of samples m should be larger than two. Each value $\varepsilon_j \in \mathcal{E}$ must be within $[0, \max\{\|x_i - z\|\}_{i=1}^n]$. In this work, we randomly select $m = 30$ radii for the sake of simplicity and computational efficiency. More sophisticated method for selecting the set of radii $\mathcal{E}$ might improve the estimation accuracy, with possible increase of computational cost. In the fitted model, the intercept is an estimate of $f(z)$ and the coefficient of X_ε is an estimate of $C(z)$, from which we can estimate $\mathrm{Tr}\nabla^2 f(z)$ as a byproduct. We denote the fitted intercept value as $\hat{f}_s(z)$.

Now that we have density estimate $\hat{f}_s(z)$ at point z, we can derive an entropy estimator based on the density estimate by leave-one-out estimation

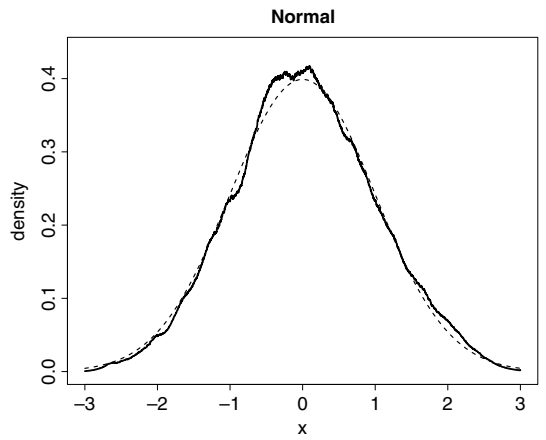
$$\hat{H}_s(\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_{s,i}(x_i), \tag{12}$$

where $\hat{f}_{s,i}(x_i)$ is the density estimated by minimizing Eq. (11) without using sample x_i. We call this entropy estimator as Simple Regression Entropy Estimator (SRE) henceforth.

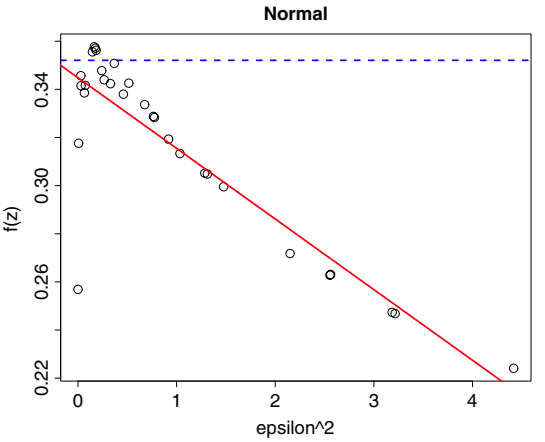
To see how the proposed method works, in Fig. 1, for some simple probability distributions, we show the estimated densities and the plot of $X_\varepsilon = \varepsilon^2$ and $Y_\varepsilon = k_\varepsilon/(nc_p\varepsilon^p)$ together with fitted lines (shown in solid lines) and the ground truth density (shown in dashed lines) at inspection points.

3.2. Leave one out estimator with weighted least squares

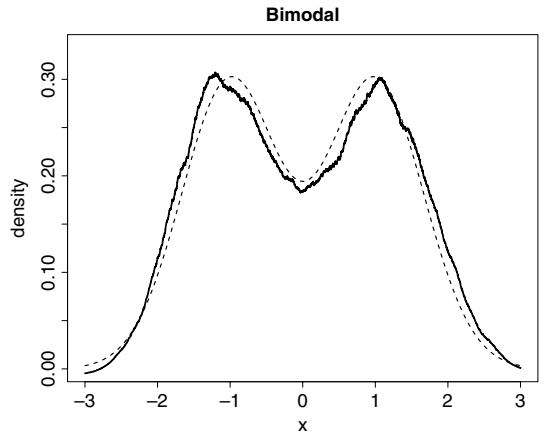
The rationale behind the k-NN density estimator Eq. (4) is the assumption that the number of points that fall within the ε-ball centered at the inspection point z follows the Bernoulli distribution with parameter θ_ε, which is empirically estimated by $\hat{\theta}_\varepsilon = k_\varepsilon/n$. The variance of the Bernoulli distribution is estimated by $\hat{\sigma}^2_\varepsilon = \hat{\theta}_\varepsilon(1-\hat{\theta}_\varepsilon) = k_\varepsilon/n(1-k_\varepsilon/n)$, and we can



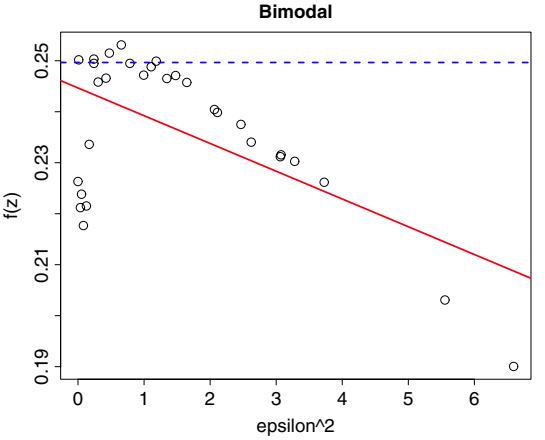
(a) Fitted density function.



(b) Fitted intercept $\hat{f}_s(z = 0.5)$.



(c) Fitted density function.



(d) Fitted intercept $\hat{f}_s(z = 0.5)$.

Fig. 1. (a) and (c): probability densities of normal and bimodal distributions. Solid curves and dashed curves are the estimated density and the ground truth densities with 300 samples, respectively. (b) and (d): for estimating $f(z)$ at $z = 0.5$, points $(X_\varepsilon, Y_\varepsilon)$ are marked as ∘, fitted regression lines are shown in solid lines, and ground truth density values $f(z = 0.5)$ are shown in horizontal dashed lines, respectively.

utilize $\hat{\sigma}_\varepsilon^2$ for weighting residuals to obtain the weighted least square formulation

$$R_w = \frac{1}{m}\sum_{\varepsilon\in\mathcal{E}} w_\varepsilon (Y_\varepsilon - f(z) - C(z)X_\varepsilon)^2, \quad w_\varepsilon = 1/\hat{\sigma}_\varepsilon^2. \tag{13}$$

The fitted intercept obtained by minimizing R_w is denoted by $\hat{f}_w(z)$. We can estimate entropy with the pdf estimated by weighted least square fitting as

$$\hat{H}_w(\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_{w,i}(x_i), \tag{14}$$

where $\hat{f}_{w,i}(x_i)$ is the estimated density obtained by minimizing Eq. (13) without using a sample x_i. We call this entropy estimator Weighted Regression Entropy Estimator (WRE).

3.3. Direct entropy estimation

We next derive another entropy estimator based on direct estimation of entropy after plugging the relationship in Eq. (10) into the re-substitution formula for the differential entropy in Eq. (2).

Consider the relationship in Eq. (10) at $z = x_i \in \mathcal{D}$ for a fixed ε. Since

$$Y_\varepsilon = \frac{k_\varepsilon}{n c_p \varepsilon^p}, \quad \text{and} \quad C(x_i) = \frac{\mathrm{Tr}\nabla^2 f(x_i)}{4(p/2+1)}$$

depend on the inspection point x_i, we denote Y_ε and $C(z)$ at $z = x_i$ as $Y_\varepsilon(x_i)$ and $C(x_i)$, respectively. To obtain an entropy estimator based on Eq. (2), we take the negative of logarithm of $Y_\varepsilon(x_i) = f(x_i) + C(x_i)X_\varepsilon$, and take the empirical average over all $x_i \in \mathcal{D}$ as

$$\begin{aligned}
-\frac{1}{n}\sum_{i=1}^{n}\ln Y_\varepsilon(x_i) &= -\frac{1}{n}\sum_{i=1}^{n}\ln\{f(x_i) + C(x_i)X_\varepsilon\} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left\{\ln f(x_i)\left(1 + \frac{C(x_i)}{f(x_i)}X_\varepsilon\right)\right\} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\ln f(x_i) - \frac{1}{n}\sum_{i=1}^{n}\ln\left(1 + \frac{C(x_i)}{f(x_i)}X_\varepsilon\right) \\
&\simeq -\frac{1}{n}\sum_{i=1}^{n}\ln f(x_i) - \frac{1}{n}\left(\sum_{i=1}^{n}\frac{C(x_i)}{f(x_i)}\right)X_\varepsilon. \qquad (15)
\end{aligned}$$

The approximation in the last line is due to the first order Taylor expansion of the function $\ln(1+\xi)$. The validity of the use of the Taylor expansion is briefly discussed in Appendix B. The first term of Eq. (15) is nothing but the re-substitution estimate of the entropy. Hence, rewriting $\bar{Y}_\varepsilon = -\frac{1}{n}\sum_{i=1}^{n}\ln Y_\varepsilon(x_i)$, $H(\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} f(x_i)$, and $\bar{C} = -\frac{1}{n}\sum_{i=1}^{n} C(x_i)/f(x_i)$, we obtain another linear relationship

$$\bar{Y}_\varepsilon = H(\mathcal{D}) + \bar{C}X_\varepsilon \qquad (16)$$

for any $\varepsilon > 0$. Now, in the same manner as the previous section, we select a small number of radii $\mathcal{E} = \{\varepsilon_1, \ldots, \varepsilon_m\}$ and fit a linear model to the set of points $\{(X_\varepsilon, \bar{Y}_\varepsilon)\}_{\varepsilon\in\mathcal{E}}$ by minimizing sum of squared residuals. The intercept of the fitted linear model is our estimate of the entropy, which is called Direct Regression Entropy Estimator (DRE) henceforth.

3.4. Weighted direct entropy estimation

It is also possible to take weights into account for least square fitting in DRE, which is referred to as Weighted DRE (WDRE). There are some possible ways for defining weights for WDRE, and we show here one simple method.

We consider the random variable η_ε^i which follows the Bernoulli distribution with parameter θ_ε^i estimated as $\hat{\theta}_\varepsilon^i = k_\varepsilon(x_i)/n$, where $k_\varepsilon(x_i)$ is the number of samples that fall within the ε-ball centered at x_i. For a fixed ε, $\bar{Y}_\varepsilon$ in DRE is summation of $Y_\varepsilon(x_i)$, $i = 1, \ldots, n$, hence, the first possibility for the weighting scheme is that we assume that the error is characterized by $\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(k_\varepsilon(x_i)/n)(1 - k_\varepsilon(x_i)/n)$. We can set weight $w_\varepsilon \propto 1/\hat{\sigma}_\varepsilon^2$ for residuals to obtain the weighted least square formulation. In our experiments, as is seen from the experimental results in the next section, the improvement of the estimation accuracy obtained by weighting is marginal because DRE includes only one linear fitting.

Before investigating the performances of our proposed methods, we note that it is possible to consider higher order expansions in Eq. (8). In our preliminary experiments, we considered 4-th and 6-th order Taylor expansions of $f(z)$, and hence fitted 2nd and 3rd order polynomials of X_ε, respectively. Basically the estimation accuracies are almost identical with those of simple regression, and these results are omitted in this paper.

4. Numerical experiments

We conducted experiments to compare the performance of the proposed entropy estimators to conventional estimators.

4.1. Univariate case

We used 15 different one-dimensional continuous distributions which are called (1) Normal, (2) Skewed, (3) Strongly Skewed, (4) Kurtotic, (5) Bimodal, (6) Skewed Bimodal, (7) Trimodal, (8) Claw, (9) 4th Power Exponential, (10) Logistic, (11) Laplace, (12) t with df = 5, (13) Mixed t, (14) Exponential, and (15) Cauchy shown in Fig. 2. Explicit pdfs for these distributions are shown in Appendix C.

We compare the proposed four entropy estimators "SRE", "WRE", "DRE", and "WDRE" with the former weighting scheme to three other entropy estimators. The first two entropy estimators are obtained by plugging in the pdf estimated by kernel density estimators and re-substitution shown in Eq. (2). Density functions are estimated by kernel density estimators with Gaussian kernel functions. For selecting the bandwidth for the kernel, we adopted the plug-in bandwidth selector proposed by Sheather and Jones (1991), and unbiased least square cross-validation proposed by Rudemo (1982) and Bowman (1984), both are popular bandwidth selectors. The entropy estimator using pdf estimated by using kernel density estimators with bandwidth selected by the plug-in method and cross-validation are indicated by "KDE.p" and "KDE.cv", respectively. The last estimator is the k-NN entropy estimator indicated by "kNN" with $k = 3$, where the value $k = 3$ is chosen according to the suggestion in Khan et al. (2007).

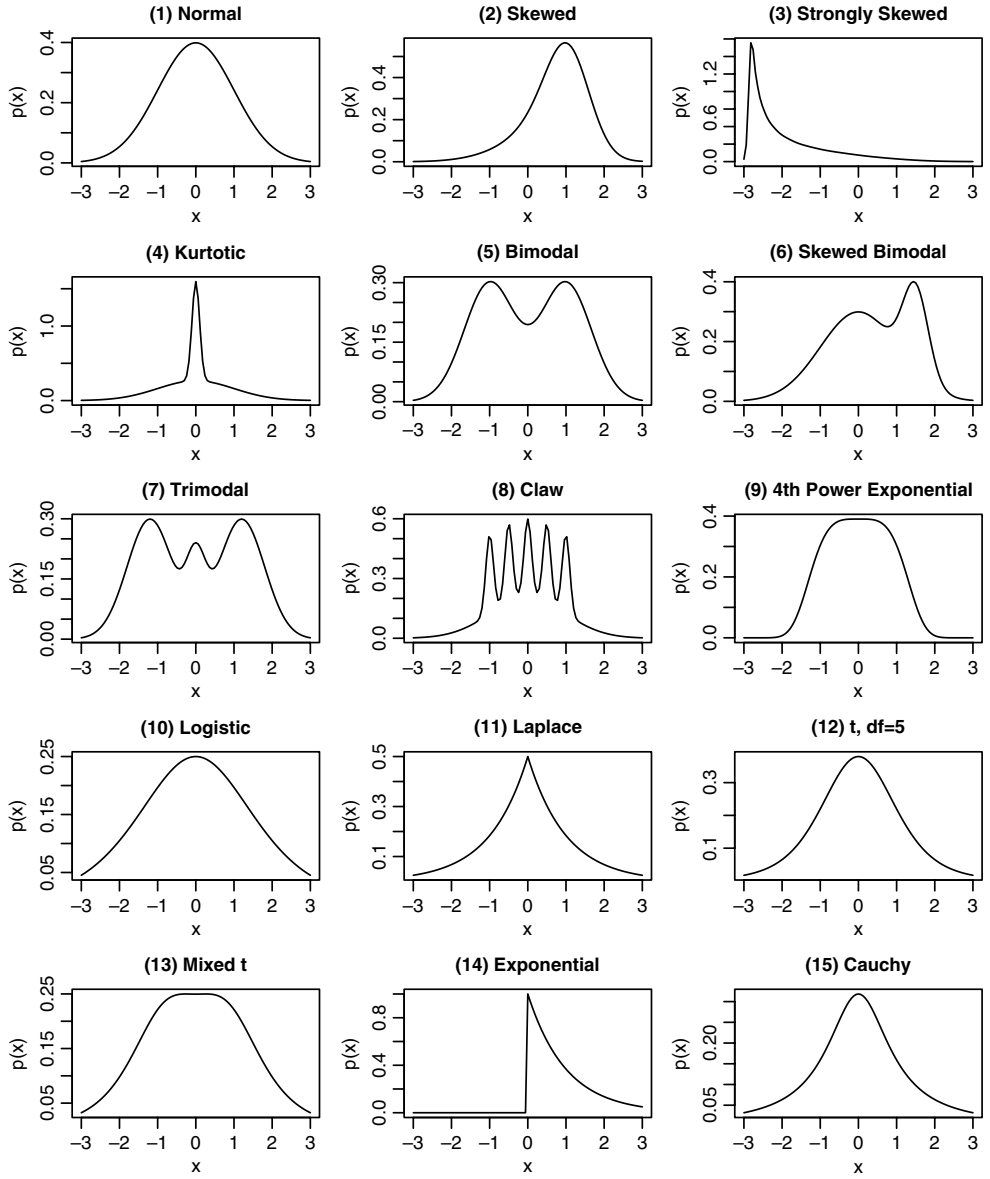


Fig. 2. Plots of 15 probability density functions for generating samples.

For evaluating the estimation performance, we calculate absolute error defined by

$$AE = |H(f) - \hat{H}(\mathcal{D})| \tag{17}$$

between the ground truth entropy $H(f)$ of the random variable with pdf $f(x)$ and the entropy $\hat{H}(\mathcal{D})$ estimated by using the observation $\mathcal{D} = \{x_i\}_{i=1}^n$ generated from $f(x)$. For some pdfs without explicit formula for the differential entropy, we computed the ground truth entropy values by numerical integration. We repeated random sampling from the underlying distribution 100 times and calculated absolute errors for each sample set. The number of observation n is set to $n = 300$. For the sake of legibility, we plot the experimental results for 15 distributions separately in Figs. 3 and 4.

From Figs. 3 and 4, and Table 1, we see that there is no single best method that consistently outperforms the others. However, we can see that plug-in entropy estimators do not perform well for many distributions.

Since the proposed methods take into account the second order derivative of the pdf, it is expected to show better performance than the conventional k-NN method when the curvature of the density function is not negligible. To see the effect of the curvature on the improvement of the estimation accuracy from that of the k-NN method, we perform a simple experiment using the Cauchy distribution,

$$f(x;\gamma) = \frac{1}{\pi\gamma(1 + (x/\gamma)^2)},$$

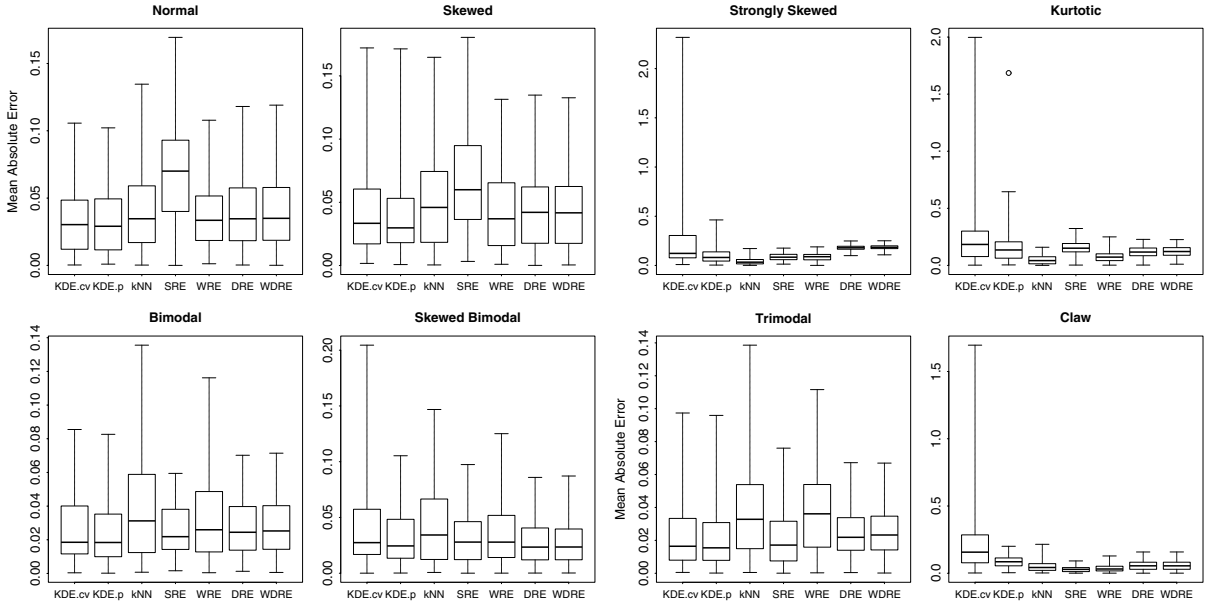


Fig. 3. Boxplots of absolute errors obtained by six different entropy estimators. The ground truth probability distributions are (1) Normal, (2) Skewed, (3) Strongly Skewed, (4) Kurtotic, (5) Bimodal, (6) Skewed Bimodal, (7) Trimodal, and (8) Claw. Sample size n is set to 300.

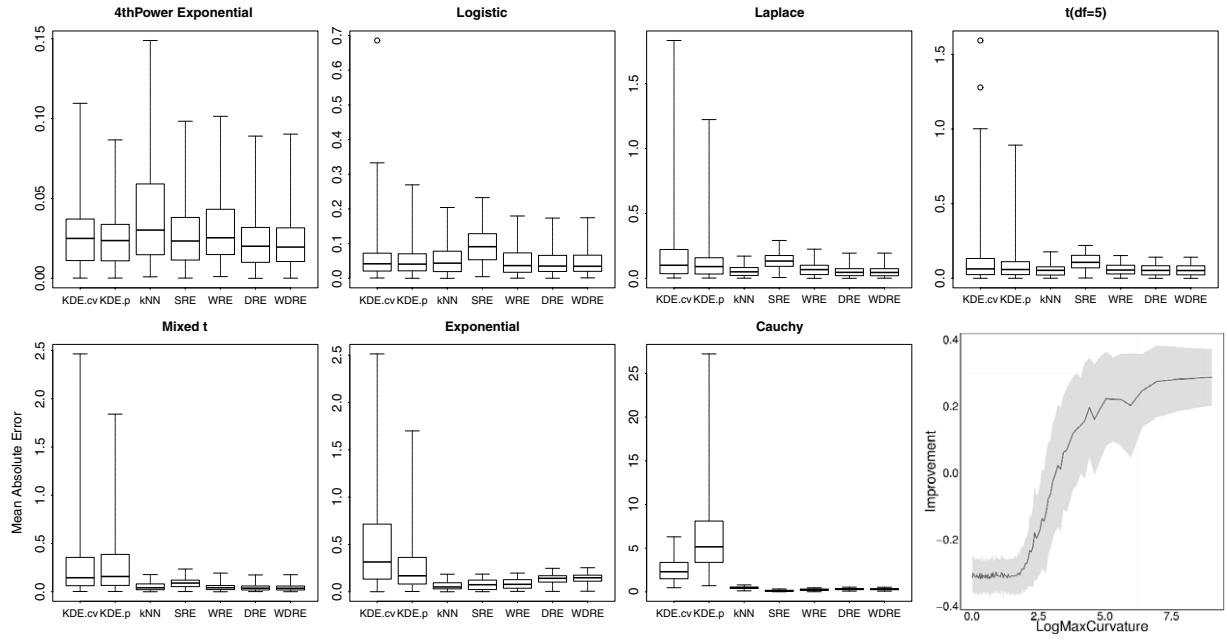


Fig. 4. Boxplots of absolute errors obtained by six different entropy estimators. The ground truth probability distributions are (9) Power Exponential, (10) Logistic, (11) Laplace, (12) t with df = 5, (13) Mixed t, (14) Exponential, and (15) Cauchy. Sample size n is set to 300. The bottom right panel shows improvement of estimation accuracies obtained by SRE over those obtained by the k-NN method, when the maximum curvature of Cauchy distribution is increased.

where $\gamma > 0$ is the scale parameter. The second derivative of this density function is

$$\nabla^2 f(x;\gamma) = \frac{2}{\pi\gamma^3}\frac{3(x/\gamma)^2 - 1}{(1+(x/\gamma)^2)^3}.$$

We varied γ from 0.01 to 0.9 and performed entropy estimation 100 times with independent datasets composed of $n = 300$ samples using both k-NN method and SRE. In the bottom right most panel of Fig. 4, we plot the averages and one standard deviations of $|\hat{H}_k(\mathcal{D}) - H(f)| - |\hat{H}_s(\mathcal{D}) - H(f)|$ evaluated by using 100 independent datasets, namely, absolute error obtained by the k-NN method minus that obtained by SRE. This quantity indicates how the estimation error is reduced by using

Table 1
Averages of absolute errors of entropy estimations for different seven methods. Sample size n is set to 300. The best results are shown in boldface.

	KDE.cv	KDE.p	kNN	SRE	WRE	DRE	WDRE
Type 1	**0.033**(0.0252)	**0.033**(0.0253)	0.042(0.0322)	0.068(0.0376)	0.038(0.0251)	0.040(0.0279)	0.040(0.0280)
Type 2	0.044(0.0347)	**0.039**(0.0307)	0.051(0.0392)	0.068(0.0416)	0.042(0.0307)	0.044(0.0310)	0.044(0.0307)
Type 3	0.262(0.3304)	0.107(0.0936)	**0.043**(0.0368)	0.086(0.0374)	0.085(0.0406)	0.180(0.0259)	0.185(0.0247)
Type 4	0.252(0.3027)	0.169(0.1960)	0.049(0.0393)	0.152(0.0538)	**0.075**(0.0448)	0.119(0.0468)	0.122(0.0458)
Type 5	0.026(0.0206)	**0.024**(0.0192)	0.039(0.0335)	0.025(0.0152)	0.033(0.0268)	0.026(0.0158)	0.026(0.0159)
Type 6	0.039(0.0340)	0.032(0.0246)	0.042(0.0342)	0.031(0.0235)	0.035(0.0275)	**0.028**(0.0207)	**0.028**(0.0205)
Type 7	0.024(0.0216)	0.023(0.0207)	0.039(0.0310)	**0.021**(0.0165)	0.038(0.0274)	0.024(0.0149)	0.025(0.0152)
Type 8	0.247(0.2936)	0.088(0.0430)	0.051(0.0382)	**0.031**(0.0220)	0.039(0.0280)	0.059(0.0354)	0.059(0.0356)
Type 9	0.027(0.0209)	0.025(0.0180)	0.039(0.0312)	0.027(0.0198)	0.031(0.0219)	0.024(0.0177)	**0.023**(0.0175)
Type 10	0.064(0.0861)	0.053(0.0467)	0.052(0.0413)	0.092(0.0524)	0.048(0.0386)	**0.047**(0.0375)	0.048(0.0380)
Type 11	0.202(0.3054)	0.143(0.1937)	0.060(0.0430)	0.133(0.0583)	0.069(0.0470)	**0.052**(0.0405)	**0.052**(0.0406)
Type 12	0.143(0.2549)	0.096(0.1247)	0.056(0.0396)	0.107(0.0533)	0.061(0.0379)	**0.054**(0.0369)	**0.054**(0.0370)
Type 13	0.290(0.3998)	0.289(0.3396)	0.054(0.0419)	0.088(0.0492)	0.047(0.0349)	**0.043**(0.0326)	**0.043**(0.0325)
Type 14	0.574(0.6199)	0.300(0.3641)	**0.064**(0.0430)	0.077(0.0524)	0.086(0.0559)	0.137(0.0537)	0.144(0.0526)
Type 15	2.595(1.4603)	6.643(4.8738)	0.475(0.1158)	**0.123**(0.0798)	0.246(0.0993)	0.315(0.0883)	0.307(0.0870)

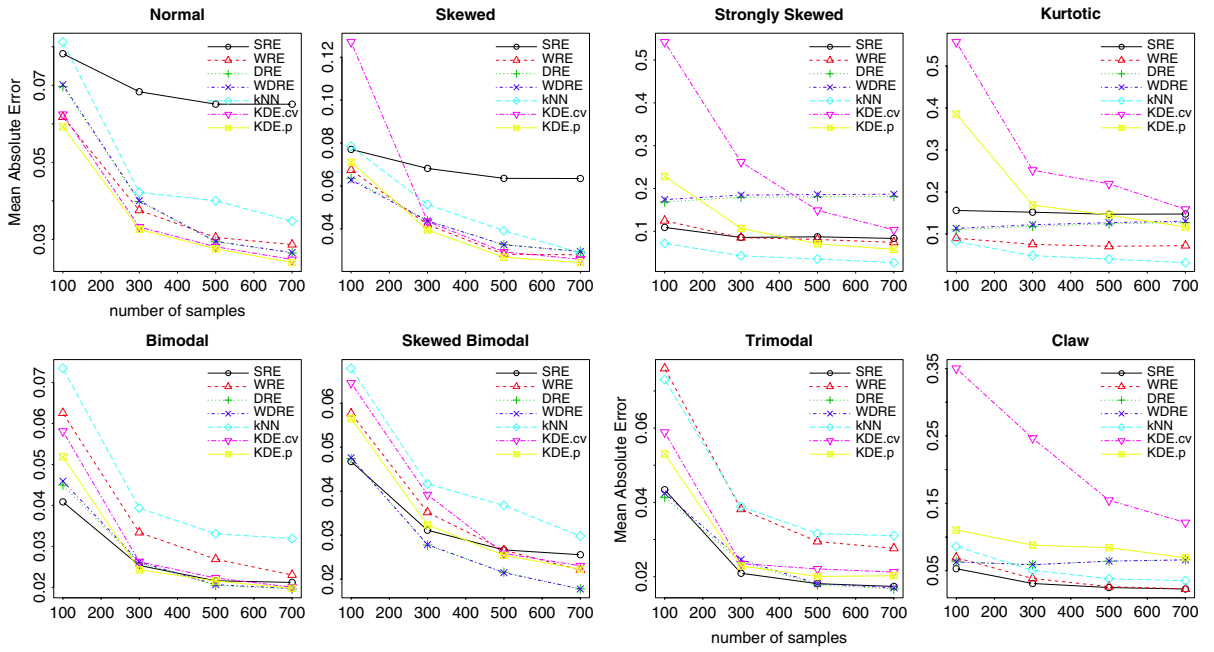


Fig. 5. Averages of absolute errors of entropy estimation when the sample sizes are varied from 100 to 700. The ground truth probability distributions are (1) Normal, (2) Skewed, (3) Strongly Skewed, (4) Kurtotic, (5) Bimodal, (6) Skewed Bimodal, (7) Trimodal, and (8) Claw.

SRE instead of the k-NN method. The horizontal axis of the figure is the logarithm of the maximum absolute curvature $\max_{x\in\mathbb{R}} \log|\nabla^2 f(x;\gamma)|$. From this result, it is clearly seen that the magnitude of improvement obtained by SRE increases as curvature increases.

From Table 1, it is seen that entropy estimators based on kernel density estimators outperform others only for some simple cases such as Normal distribution. From these experimental results, we can conclude that the proposed estimators based on simple linear regression are strong candidates for estimating the entropy from observed dataset.

We next varied the number of observations n from 100 to 700 by 200, and plotted the average of absolute errors obtained by 7 different entropy estimators in Figs. 5 and 6. These figures again support the fact that the proposed methods are comparable or superior to conventional entropy estimators.

Finally, we evaluated the computational times for conventional methods and those for our proposed methods with the increase of sample sizes. We varied the number of samples n from 100 to 900 by 200, and plotted the averages of computational time in second at the bottom right most panel of Fig. 6. The proposed methods SRE and DRE require distance computation at least $m \times n$ times, where m is the number of different radii, and n is the number of samples. Then, n linear fittings for SRE and one linear fitting for DRE are performed. Although DRE performs linear fitting only once, computation for $\bar{Y}_\varepsilon$ requires distance calculation against $n-1$ points for each x_i, hence the total computational cost is of order $O(n^2)$. In our setting, since we fixed m, the computational cost of these methods is of order $O(n^2)$, which is seen from the figure. Since

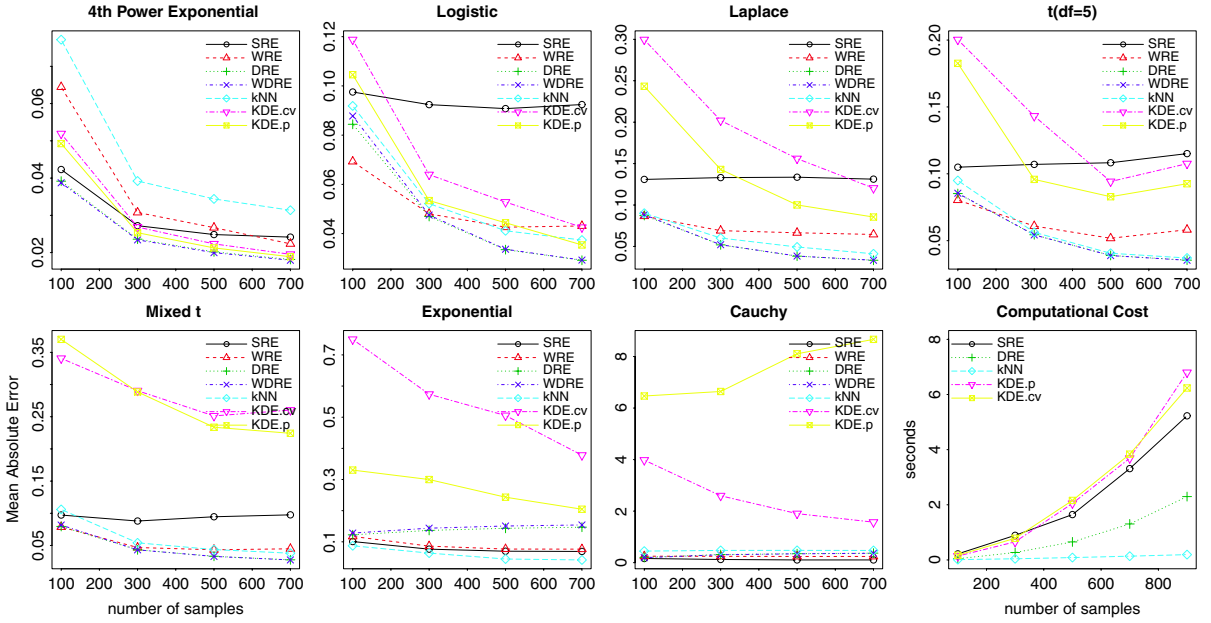


Fig. 6. Averages of absolute errors of entropy estimation when the sample sizes are varied from 100 to 700. The ground truth probability distributions are (9) 4th Power Exponential, (10) Logistic, (11) Laplace, (12) t with df = 5, (13) Mixed t, (14) Exponential, and (15) Cauchy. The bottom right panel shows computational costs for different estimators.

DRE includes linear fitting only once, it is computationally efficient compared to SRE. The gap between the computational cost of k-NN and SRE is attributed to the cost for linear fittings. The KDE based methods are relatively slow, possibly because of their computational costs for optimizing their bandwidth parameters.

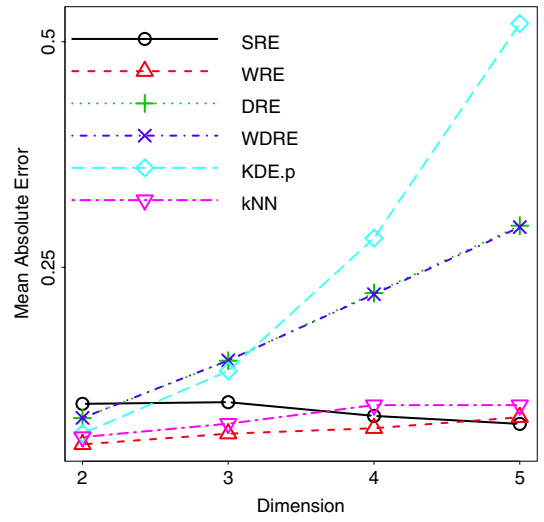
4.2. Multivariate case

To see the effect of dimensionality to the accuracies of entropy estimators, we perform entropy estimation experiments with different dimensionality. Since it is difficult to calculate the entropy of multivariate distributions in general, we restrict ourselves to examine the performance for p-dimensional Gaussian distributions with zero mean vector with the following three different covariance structures.

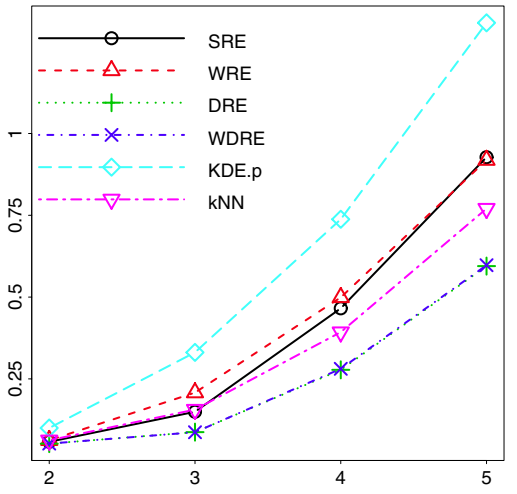
Isometric: Covariance matrix for Gaussian distribution is $\Sigma_p = I_p$, where I_p is the p-dimensional unit matrix.
Band: Covariance matrix for Gaussian distribution is a tridiagonal band matrix with main diagonal elements 1, and the first diagonal below and above the main diagonal 0.3.
Full correlation: Covariance matrix for Gaussian distribution is defined as

$$[\Sigma_p]_{ij} = 0.9^{|i-j|+1}, \quad 0 \le i, j \le p. \tag{18}$$

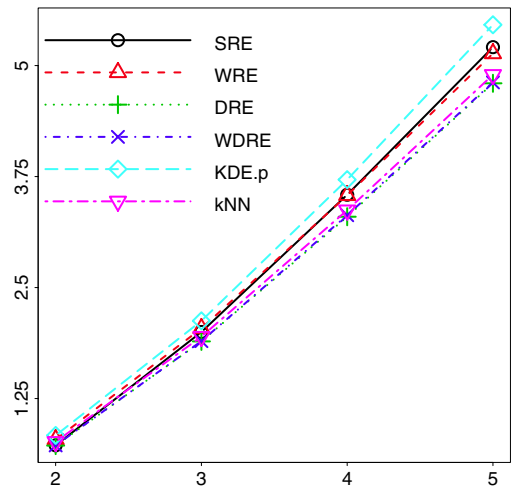
We performed experiments using different sample sizes $n = 100, 300, 500$ and 700. Since the tendencies of the relationship between accuracy and dimensionality are similar in all different sample sizes, we only show the result with $n = 300$. For kernel density estimation, we optimized full bandwidth matrices with off-diagonal elements by using both cross validation (Rudemo, 1982; Bowman, 1984) and plug-in approaches (Wand and Jones, 1994a). In all cases, the performances of cross validation-based method "KDE.cv" were far inferior to plug-in-based method "KDE.p". For the sake of legibility, we do not include the result with "KDE.cv" in the plot. The experimental result is shown in Fig. 7. From Fig. 7(a)–(c), we see that KDE based entropy estimator does not perform well in high dimensional cases. In our experiments with kernel density estimator, we optimized bandwidth matrices with off-diagonal entries. This explains why the results of "KDE.p" for **Band** and **Full correlation** datasets were relatively better. However, KDE-based estimator is inferior to other methods in all cases, and our proposed "DRE" and "WDRE" outperform others in two out of three cases. The source of the error in SRE is mainly in the error due to linear fitting in Eq. (10), which accumulates n times in summation of Eq. (12). On the other hand, the sources of the error in DRE are both fitting error and approximation error in Eq. (15). The approximation error in Eq. (15) accumulates n times in summation with respect to x_i, however, the fitting error in DRE occurs only once because the entropy is directly estimated by simple regression. The density estimation for **Band** and **Full Correlation** cases are difficult compared to the **Isometric** case, and the accumulated errors for pdf estimation severely affect to SRE. On the other hand, the effect of fitting error occurs only once in DRE, and it would contribute to better performance of DRE in the latter two



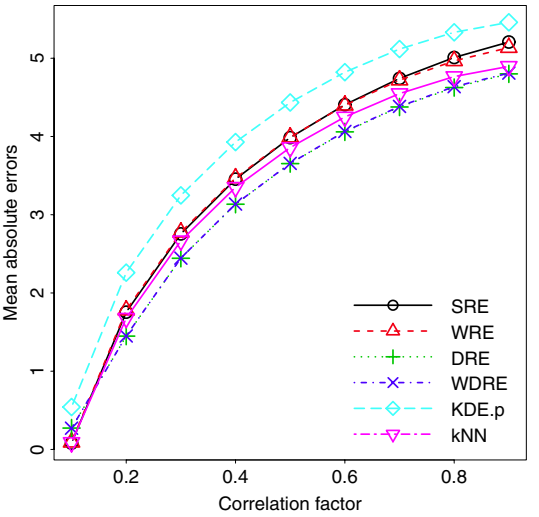
(a) Isometric.



(b) Band.



(c) Full correlation.



(d) CF and MAE ($p = 5$).

Fig. 7. (a)–(c): Averages of absolute errors of entropy estimation when p is varied from 2 to 5. The number of samples is fixed to $n = 300$, and Gaussian distributions with three different covariance matrices are investigated. (d): Mean absolute errors and correlation factors (CFs) for **Full Correlation** data.

cases. Finally, we investigated the relationship between entropy estimation errors and correlations in the **Full Correlation** case. We varied the Correlation Factor (CF) α in the covariance matrix

$$[\Sigma_p]_{ij} = \alpha^{|i-j|+1}, \quad 0 \leq i, j \leq p$$

from 0.1 to 0.9 by 0.1, and performed the same entropy estimation experiment described above. Fig. 7(d) depicts the experimental results when $p = 5$. As expected, the estimation accuracy degenerates with the increase of CF. From Fig. 7(b)–(d), it is seen that "DRE" outperforms other methods, hence we can conclude that for correlated multivariate data, the estimator "DRE" is a better alternative to conventional methods.

5. Concluding remarks

In this study, we proposed simple non-parametric estimators of the differential entropy. The estimators are based on the second order expansion of the probability mass function, and simple linear regression. We evaluated their performances with various probability distributions and compared them to those of conventional entropy estimators. The proposed methods are shown to be comparable or superior to other conventional entropy estimators.

The aim of this paper is in proposing a novel non-parametric approach for entropy estimation, and showing that the proposed entropy estimators work well for various distributions. Asymptotic theory for our estimators should be established in our future work, but we briefly mention the asymptotics of entropy estimators. The density estimate based on Eq. (4) is consistent when $k_\varepsilon \to \infty$ and $n/k_\varepsilon \to \infty$, which suggests $\varepsilon \to 0$ as $n \to \infty$ at a certain rate, and under weak regularity conditions on the pdf, the entropy estimator Eq. (7) is also asymptotically unbiased and consistent (Goria et al., 2005). In our method, in the limiting case of $n \to \infty$ hence $\varepsilon \to 0$, the second and third terms in Eq. (9) vanish and the same argument can be applied on asymptotic unbiasedness and consistency for density estimation. However, our estimator is based on simple regression and the term $O(\varepsilon^4)$, which is neglected to obtain Eq. (10), is dependent on the size of neighborhood ε, and hence dependent on the dataset $\{(X_\varepsilon, Y_\varepsilon)\}_{\varepsilon\in\mathcal{E}}$ for regression. Because of this dependency of the bias term and the fact that we cannot expect zero mean for this bias term, it is difficult to establish theoretical guarantees of the estimated pdfs and entropies.

The differential entropy considered in this paper is the Shannon entropy. There are other classes such as Rényi (Renyi, 1960) and Tsallis entropies (Tsallis, 1988). The k-NN entropy estimator is also applicable to estimate the Rényi entropy (Leonenko et al., 2008). Our proposed SRE and WRE can be applied to estimation of the Rényi and Tsallis entropies because these entropies are based on the power of pdfs, and SRE and WRE is based on the estimates of pdfs. On the other hand, since the Rényi and Tsallis entropies are defined with integral of $f^q(x)$, where $f(x)$ is the pdf and $q > 0$ is some parameter ($q \neq 1$ for Rényi entropy), it is not straightforward to derive a DRE-type estimator for general q. It is possible to derive a DRE-type estimator for a specific value of q. The Rényi and Tsallis entropies are important classes of information theoretic quantities, and estimators for these quantities based on our proposed approach remain to be investigated.

Acknowledgments

The authors would like to express their special thanks to the editor, the associate editor and three anonymous reviewers whose comments led to valuable improvements of the paper. H.H. was supported by JSPS KAKENHI Grant Number 25870811 and 26120504, and N.M. was supported by JSPS KAKENHI Grant Number 25120009.

Appendix A. Detailed calculation of second order expansion of probability mass

We prove Proposition 1 in Section 3. The probability mass contained within the ε-ball centered at inspection point z is defined as

$$q_z(\varepsilon) = \int_{x\in b(z;\varepsilon)} f(x)\mathrm{d}x. \tag{19}$$

Using second-order Taylor series expansion of $q_z(\varepsilon)$, we obtain

$$\begin{aligned} q_z(\varepsilon) &\simeq \int_{x\in b(z;\varepsilon)} \left\{ f(z) + (x-z)^\top \nabla f(z) + \frac{1}{2}(x-z)^\top \nabla^2 f(z)(x-z) \right\} \mathrm{d}x \\ &= \int_{x\in b(z;\varepsilon)} f(z)\mathrm{d}x + \nabla f(z)^\top \int_{x\in b(z;\varepsilon)} (x-z)\mathrm{d}x + \frac{1}{2}\int_{x\in b(z;\varepsilon)} (x-z)^\top \nabla^2 f(z)(x-z)\mathrm{d}x \\ &= c_p\varepsilon^p f(z) + \frac{1}{2}\mathrm{Tr}\left\{ \nabla^2 f(z) \int_{x\in b(z;\varepsilon)} (x-z)(x-z)^\top \mathrm{d}x \right\}, \end{aligned} \tag{20}$$

where $c_p = \pi^{p/2}/\Gamma(p/2+1)$. The integrand of the quadratic term is written as $\int_{x\in b(z;\varepsilon)} (x-z)(x-z)\mathrm{d}x = \int_{x\in b(0;\varepsilon)} xx^\top \mathrm{d}x$, and by symmetry, the off-diagonal elements become zero when integrated in the ball, and only diagonal elements $\int_{x\in b(0;\varepsilon)} x_i^2 \mathrm{d}x$ remain. It is easy to see that the remaining integral is

$$\int_{x\in b(0;\varepsilon)} x_i^2 \mathrm{d}x = \frac{\pi^{p/2}}{2\Gamma(p/2+2)}\varepsilon^{p+2}.$$

Substituting the above results back to Eq. (20), we obtain

$$\begin{aligned} q_z(\varepsilon) &\simeq c_p\varepsilon^p f(z) + \frac{1}{2}\mathrm{Tr}\left\{ \nabla^2 f(z) \frac{\pi^{p/2}}{2\Gamma(p/2+2)} I_p \right\} \varepsilon^{p+2} \\ &= c_p\varepsilon^p f(z) + \frac{\pi^{p/2}}{4\Gamma(p/2+2)}\varepsilon^{p+2}\mathrm{Tr}\nabla^2 f(z) \\ &= c_p f(z)\varepsilon^p + \frac{1}{4(p/2+1)} c_p\varepsilon^{p+2}\mathrm{Tr}\nabla^2 f(z), \end{aligned}$$

which proves the proposition.

Appendix B. Validity of the Taylor expansion for DRE

We consider the appropriate range of ε for DRE and WDRE based on the validity of the Taylor expansion of $\ln(1+\xi)$ with $\xi = \frac{C(x_i)}{f(x_i)} X_\varepsilon$ in Eq. (15). Since $X_\varepsilon = \varepsilon^2$ and $C(x_i) = \frac{\mathrm{Tr}\nabla^2 f(x_i)}{4(p/2+1)f(x_i)}$, ε is written in terms of ξ as

$$\varepsilon = \sqrt{\frac{4(p/2+1)f(x_i)}{\mathrm{Tr}\nabla^2 f(x_i)}\xi}. \tag{21}$$

This implies that values of $\varepsilon_m \in \mathcal{E}$ should be small enough, and they should decrease at the rate of $O(\xi^{1/2})$ with respect to the required precision of the approximation for the Taylor expansion.

The appropriate value for ε_m is a function of values of the ground truth density $f(x_i)$, $\mathrm{Tr}\nabla^2 f(x_i)$, the dimensionality p of the samples, and the value of ξ which is determined based on the required accuracy of the Taylor expansion. Since we do not know the ground truth probability density function $f(x)$, Eq. (21) cannot be used for determining the range of ε_m. However, it is possible to estimate $f(x_i)$ and $\mathrm{Tr}\nabla^2 f(x_i)$ by using SRE with a set of randomly sampled ε_m, and the estimated $f(x_i)$ and $\mathrm{Tr}\nabla^2 f(x_i)$ can be plugged into Eq. (21) to obtain an estimate of appropriate value of ε_m with additional computational cost.

Finally, we note that Eq. (21) is derived solely based on the precision of the approximation of the Taylor expansion, and it does not depend on n. In general, the radius of the ball for probability mass function should decrease with the increase of the number of samples n.

Appendix C. Probability density functions for numerical experiment

We enumerate the probability density functions of 15 distributions used in Section 4.1. For the sake of notational simplicity, we use a conventional notation $\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$.

(1) Normal: $f(x) = \mathcal{N}(x;0,1)$.
(2) Skew: $f(x) = \frac{1}{5}\mathcal{N}(x;0,1) + \frac{1}{5}\mathcal{N}(x;0.5,2/3) + \frac{3}{5}\mathcal{N}(x;13/12,5/9)$.
(3) Strongly Skewed: $f(x) = \frac{1}{8}\sum_{i=1}^{8}\mathcal{N}(x;3(2/3)^{l-1}-3,(2/3)^{l-1})$.
(4) Kurtotic: $f(x) = \frac{2}{3}\mathcal{N}(x;0,1) + \frac{1}{3}\mathcal{N}(x;0,1/10)$.
(5) Bimodal: $f(x) = \frac{1}{2}\mathcal{N}(x;-1,2/3) + \frac{1}{2}\mathcal{N}(x;1,2/3)$.
(6) Skewed Bimodal: $f(x) = \frac{3}{4}\mathcal{N}(x;0,1) + \frac{1}{4}\mathcal{N}(x;3/2,1/3)$.
(7) Trimodal: $f(x) = \frac{9}{20}\mathcal{N}(x;-6/5,3/5) + \frac{9}{20}\mathcal{N}(x;6/5,3/5) + \frac{1}{10}\mathcal{N}(x;0,1/4)$.
(8) Claw: $f(x) = \frac{1}{10}\sum_{l=1}^{5}\mathcal{N}(x;l/2-3/2,1/10) + \frac{1}{2}\mathcal{N}(x;0,1)$.
(9) 4th Power Exponential: $f(x) = \frac{1}{2p^{1/p}\Gamma(1+1/p)}\exp(-|x|^p/p),\ p=4$.
(10) Logistic: $f(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$.
(11) Laplace: $f(x) = \exp(-|x|)/2$.
(12) t with df = 5: $f_t(x;d) = \frac{\Gamma((d+1)/2)}{\sqrt{d\pi}\Gamma(d/2)}(1+x^2/d)^{-(d+1)/2},\ d=5$.
(13) Mixed t: $f(x) = \frac{1}{2}f_t(x+4/5;3) + \frac{1}{2}f_t(x-4/5;3)$.
(14) Exponential: $f(x) = \exp(-x),\ x \geq 0$.
(15) Cauchy: $f(x) = 1/(\pi(1+x^2))$.

References

Ahmad, I., Lin, P.-E., 1976. A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). IEEE Trans. Inform. Theory 22 (3), 372–375.
Beirlant, J., Dudewicz, E.J., Györfi, L., Meulen, E.C., 1997. Nonparametric entropy estimation: an overview. Int. J. Math. Stat. Sci. 6, 17–39.
Bowman, A.W., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71 (2), 353–360.
Comon, P., 1994. Independent component analysis, a new concept? Signal Process. 36 (3), 287–314.
Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. John Wiley and Sons, Inc..
Dudewicz, E.J., van der Meulen, E.C., 1981. Entropy-based tests of uniformity. J. Amer. Statist. Assoc. 76 (376), 967–974.
Goria, M.N., Leonenko, N.N., Mergel, V.V., Inverardi, P.L.N., 2005. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. J. Nonparametr. Stat. 17 (3), 277–297.
Györfi, L., van der Meulen, E.C., 1987. Density-free convergence properties of various estimators of entropy. Comput. Statist. Data Anal. 5 (4), 425–436.
Hall, P., 1986. On powerful distributional tests based on sample spacings. J. Multivariate Anal. 19 (2), 201–224.
Hino, H., Murata, N., 2010. A conditional entropy minimization criterion for dimensionality reduction and multiple kernel learning. Neural Comput. 22 (11), 2887–2923.
Hino, H., Murata, N., 2013. Information estimators for weighted observations. Neural Netw. 46 (0), 260–275.
Hino, H., Wakayama, K., Murata, N., 2013. Entropy-based sliced inverse regression. Comput. Statist. Data Anal. 67 (0), 105–114.
Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. J. Wiley, New York.
Joe, H., 1989. Estimation of entropy and other functionals of a multivariate density. Ann. Inst. Statist. Math. 41 (4), 683–697.
Khan, S., Bandyopadhyay, S., Ganguly, A.R., Saigal, S., Erickson, D.J., Protopopescu, V., Ostrouchov, G., 2007. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. Phys. Rev. E 76 (2 Pt 2), 026209.

Kozachenko, L.F., Leonenko, N.N., 1987. Sample estimate of entropy of a random vector. Probl. Inf. Transm. 23, 95–101.
Learned-Miller, E.G., Fisher III, J.W., 2004. ICA using spacings estimates of entropy. J. Mach. Learn. Res. 4 (7–8), 1271–1295.
Leonenko, N., Pronzato, L., Savani, V., 2008. A class of Rényi information estimators for multidimensional densities. Ann. Statist. 36 (5), 2153–2182.
Loftsgaarden, D.O., Quesenberry, C.P., 1965. A nonparametric estimate of a multivariate density function. Ann. Math. Statist. 36 (3), 1049–1051.
Mack, Y., Rosenblatt, M., 1979. Multivariate k-nearest neighbor density estimates. J. Multivariate Anal. 9 (1), 1–15.
Mannor, S., Peleg, D., Rubinstein, R.Y., 2005. The cross entropy method for classification. In: ICML. pp. 561–568.
Moore, D.S., Yackel, J.W., 1977. Consistency properties of nearest neighbor density function estimators. Ann. Statist. 5 (1), 143–154. 01.
Paninski, L., 2003. Estimation of entropy and mutual information. Neural Comput. 15, 1191–1253.
Pérez-Cruz, F., 2008. Estimation of information theoretic measures for continuous random variables. In: NIPS, pp. 1257–1264.
Renyi, A., 1960. On measures of information and entropy. In: 4th Berkeley Symposium on Mathematics, Statistics and Probability, pp. 547–561.
Rubinstein, R.Y., Kroese, D.P., 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer.
Rudemo, M., 1982. Empirical choice of histograms and kernel density estimators. Scand. J. Statist. 9 (2), 65–78.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. 623–656.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. Ser. B 53 (3), 683–690.
Tarasenko, F., 1968. On the evaluation of an unknown probability density function, the direct estimation of the entropy from independent observations of a continuous random variable, and the distribution-free entropy test of goodness-of-fit. Proc. IEEE 56 (11), 2052–2053.
Tsallis, C., 1988. Possible generalization of Boltzmann–Gibbs statistics. J. Stat. Phys. 52.
Vasicek, O., 1976. A test for normality based on sample entropy. J. R. Stat. Soc. Ser. B 38, 54–59.
Wand, M., Jones, M., 1994a. Multivariate plug-in bandwidth selection. Comput. Statist. 9 (2), 97–116.
Wand, M.P., Jones, M.C., 1994b. Kernel Smoothing. Chapman & Hall/CRC.