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).