Information Estimators for Weighted Observations

Hideitsu Hino*, Noboru Murata

Department of Computer Science, University of Tsukuba, 1-1-1 Tennodai, Tukuba, Ibaraki, 305-8573, Japan

School of Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, Japan

Abstract

The Shannon information content is a valuable numerical characteristic of probability distributions. The problem of estimating information content from an observed dataset is very important in the fields of statistics, information theory, and machine learning. The contribution of the present paper is in proposing information estimators, and showing some of their applications. When the given data are associated with weights, each datum contributes differently to the empirical average of statistics. The proposed estimators can deal with this kind of weighted data. Similar to other conventional methods, the proposed information estimator contains a parameter to be tuned, and is computationally expensive. To overcome these problems, the proposed estimator is further modified so that it is more computationally efficient and has no tuning parameter. The proposed methods are also extended so as to estimate the cross entropy, entropy and KL divergence. Simple numerical experiments show that the information estimators work properly. Then, the estimators are applied to two specific problems, distribution preserving data compression, and weight optimization for ensemble regression.

Keywords: information estimation, entropy estimation, weighted data

1. Introduction

In information theory (Shannon, 1948; Cover & Thomas, 1991), one of the most important quantities is the Shannon information content (an information metric)

$$I_f(x) = -\log f(x), \tag{1}$$

*Corresponding author. Tel.:+81 298 53 5538

Email address: hinohide@cs.tsukuba.ac.jp (Hideitsu Hino)

where $f(x)$ is a probability density function (pdf) of a random variable X, and $x \in \mathbb{R}^d$ is its realization, which is called a datum in this paper. The Shannon differential entropy is defined by averaging the information content $I_f(x)$ with its pdf $f(x)$ as

$$H(f) = \mathrm{E}_f[I_f(X)] = -\int f(x) \log f(x) \mathrm{d}x, \tag{2}$$

where $\mathrm{E}_f[\ \cdot\ ]$ is a mean operator. When the information content $I_g(x) = -\log g(x)$ of a datum x generated from a pdf $g(x)$ is averaged with respect to another pdf $f(x)$, it is called the cross entropy:

$$H(f, g) = \mathrm{E}_f[I_g(X)] = -\int f(x) \log g(x) \mathrm{d}x. \tag{3}$$

The difference between the cross entropy and the entropy is called the Kullback-Leibler (KL) divergence or the relative entropy (Kullback & Leibler, 1951). These quantities are also calculated from the information content (1). The information content, entropy, and KL divergence play important roles in various literatures. For example, in independent component analysis (see, e.g., Hyvärinen et al., 2001), entropy-based criteria are often used for recovering independent signals from mixed signals (Comon, 1994; Learned-Miller & Fisher, 2004). Also, for modeling neural networks, Linsker (1987, 2005) proposed *infomax* as a neural information processing principle, and Bell & Sejnowski (1995) utilized the principle for blind source separation under a neural network model. A computationally efficient and stable entropy estimator is indispensable for neural network studies. In discriminant problems, the cross entropy and entropy are used as optimization objectives (Mannor et al., 2005; Hino & Murata, 2010). Thus, many important quantities in statistics and information theory are derived from the information content (1).

Many attempts have been made to estimate the Shannon information content for a newly observed datum z, which we call an *inspection point*, with an observed dataset $\mathcal{D} = \{x_i\}_{i=1}^n$. Consider a *weighted dataset* defined by

$$\begin{aligned} \mathfrak{D} = \{\mathcal{D}, \mathcal{W}\}, \quad \mathcal{D} = \{x_i\}_{i=1}^n, \mathcal{W} = \{w_i\}_{i=1}^n, \\ x_i \in \mathbb{R}^d, \ \sum_{i=1}^n w_i = 1,\ 0 < w_i < 1, \end{aligned} \tag{4}$$

where $\mathcal{D}$ is a collection of data and $\mathcal{W}$ is a collection of positive valued weights, and each element $w_i \in \mathcal{W}$ is assigned to each datum $x_i \in \mathcal{D}$. One can consider a variety of generative mechanisms of this kind of weighted datasets. As a simple example, we can assume that x_i and w_i are sampled from a certain joint distribution $\tilde{p}(x, w)$. A special case of $\mathcal{W}$ is that all weights have the same value, which we will denote by $\mathcal{U} = \{1/n, \ldots, 1/n\}$. With an equally weighted dataset $\mathfrak{D} = \{\mathcal{D},\mathcal{U}\}$, an empirical average of a function $G(x)$ is calculated as $E_{\mathfrak{D}}[G(X)] = \frac{1}{n}\sum_{i=1}^{n} G(x_i)$, which is the ordinal sample mean of $G(x)$ with $\mathcal{D}$. With a general weighted dataset $\mathfrak{D} = \{\mathcal{D}, \mathcal{W}\}$, the empirical average of a function $G(x)$ is defined by

$$E_{\mathfrak{D}}[G(X)] = \sum_{i=1}^{n} w_i G(x_i). \tag{5}$$

Hereafter, a set of weights in Eq. (4) defines the average operation Eq. (5), and we identify *the distribution of* $\mathfrak{D} = \{\mathcal{D}, \mathcal{W}\}$ as Eq. (5). More formally, following the definition of weights in Cook & Nachtsheim (1994), let Θ_n denote the set of probability measures on the observed dataset $\mathcal{D}$, and any $\theta_n \in \Theta_n$ defines a distribution function on $\mathcal{D}$. The empirical distribution function corresponding to this distribution function places weights $w_i = \theta_n(x_i)$ on each datum x_i. The averaging operation with this empirical distribution is defined by Eq. (5).

To illustrate the fact that the weights play important roles in engineering, we consider the following two problems.

Problem 1 (Sample sets matching problem).
Consider two weighted datasets $\mathfrak{D} = \{\mathcal{D}, \mathcal{W}\} = \{(x_i, w_i)\}_{i=1}^{n}$ *and* $\mathfrak{D}' = \{\mathcal{D}', \mathcal{W}'\} = \{(x'_j, w'_j)\}_{j=1}^{m}$, *where* $\mathcal{D}, \mathcal{W}$, *and* $\mathcal{D}'$ *are given. Then, the problem is to find weights* $\mathcal{W}'$ *for* $\mathcal{D}'$ *such that* $\mathfrak{D}$ *and* $\mathfrak{D}'$ *are as close as possible in a certain statistical sense.*

This problem is inspired by the *two sample test* (Gretton et al., 2007), which is a test that evaluates whether two distributions are different based on samples drawn from each distribution. We consider the situation where the data distribution is specified by some data producing mechanism for X and a weight for each realization of X, which defines the averaging operation by Eq. (5). Then, we will tailor distributions by modifying the weights assigned to the data, and consider an optimization problem of the distribution with respect to the weights.

When $\mathcal{W} = \mathcal{U}$ and $|\mathcal{D}'| \ll |\mathcal{D}|$, the problem is considered as a specific data compression problem in which the original dataset $\mathcal{D}$ is reduced to $\mathcal{D}'$ while preserving the probability distribution. There are a lot of established methods for data compression such as k-means clustering (Jain et al., 1999) and Learning Vector Quantization (LVQ; Kohonen et al., 1996). Most data compression methods minimize approximation errors (or, energy, in the clustering literature), but do not preserve the distribution of the original dataset in the sense of statistical dispersion measure such as the KL divergence estimated by weighted data sets $\mathfrak{D}$ and $\mathfrak{D}'$[1]. Thus, problem 1 is regarded as a problem where one tries to achieve distribution-preserving data compression by optimizing $\mathcal{W}'$ of $\mathcal{D}'$.

Problem 2 (Weighted ensemble regression).
Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ be a given set of explanatory and response variables. To relate x and y, a set of regressors $\mathcal{C} = \{c_j | c_j : \mathbb{R}^d \to \mathbb{R}\}_{j=1}^m$ is also given. We assign a weight $w_j \in \mathcal{W}$ to c_j and construct a predictive distribution of the response variable with weighted particle approximation of regressors $\mathcal{C}(x) = \{c_j(x)\}_{j=1}^m$. Then, the problem becomes one of optimizing $\mathcal{W}$ such that the distribution over y represented by $\mathfrak{C}(x_i) = \{\mathcal{C}(x_i), \mathcal{W}\}$ conditioned by x_i is as close to the distribution of y_i conditioned by the same x_i as possible.

In problem 2, we wish to approximate the distribution of y_i by a set of weighted predictions $\{c_j(x_i), w_j\}_{j=1}^m$ (see figure 1). For example, the mean and the variance of y_i are approximated by $\sum_{j=1}^m w_j c_j(x_i)$ and $\sum_{j=1}^m w_j (c_j(x_i) - \sum_{l=1}^m w_l c_l(x_i))^2$, respectively. If we had a sufficient number of response variables $y_{i,k}, k = 1, \ldots$ corresponding to each x_i, it would be possible to optimize $w_j, j = 1, \ldots, m$ using, for example, the moment matching method (Ramberg & Schmeiser, 1974; Song et al., 2008). However, because there is only one response y_i for each explanatory variable x_i, we cannot use standard distribution adjustment techniques.

We will propose information estimators $I_{\mathfrak{D}}(x)$ based on a weighted dataset $\mathfrak{D}$, which enables us to solve these problems within a unified framework. With an information estimator $I_{\mathfrak{D}}(x)$, we can estimate the KL divergence between two distributions from observed data as

$$D_{KL}(\mathfrak{D}, \mathfrak{D}') = E_{\mathfrak{D}}[I_{\mathfrak{D}'}(X)] - E_{\mathfrak{D}}[I_{\mathfrak{D}}(X)].$$

Then, we can solve problem 1 by estimating the KL divergence between empirical distributions of $\mathfrak{D}$ and $\mathfrak{D}'$, and minimizing the KL divergence with respect to weights $\mathcal{W}'$. As for problem 2, the information estimator is used

[1] We note that clustering by Gaussian mixture modeling can be seen to preserve the distribution of the original data, and the mixture ratio parameter and the *weight* in our study have similarity. However, Gaussian mixture modeling assigns a Gaussian distribution for each cluster, while we do not assume any form of distribution functions for neither individual clusters nor the whole observed data.

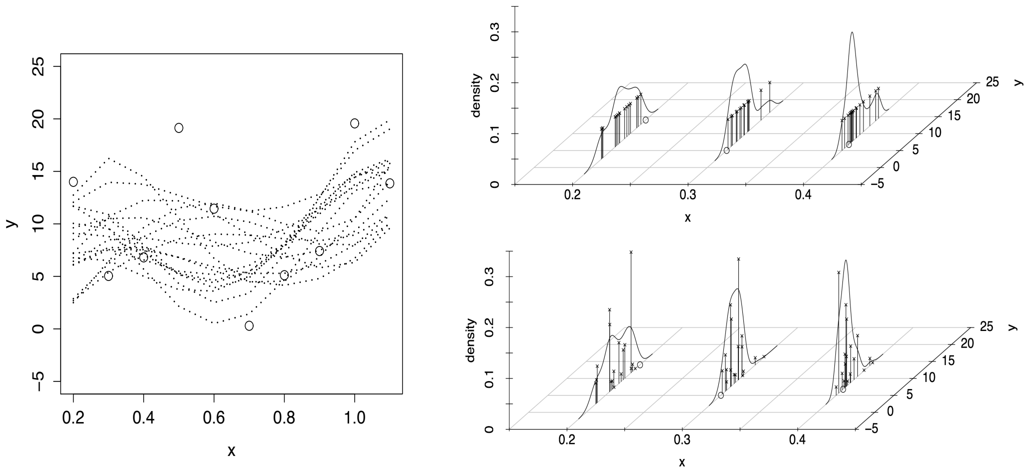


Figure 1: An illustrative example of the weighted regressors' distribution. Left: response variables are plotted with ∘. The outputs of 17 regressors are shown as dotted lines. Right top: The output value of each of the regressors is plotted at three input locations with × and a vertical segment. The height of the symbol × represents the weight for the regressor. Density functions estimated using the regressors' outputs are also shown (density functions are estimated by a Gaussian kernel density estimator). Right bottom: Different weights are assigned to different regressors. The weights affect the densities of the regressors' outputs.

to estimate the information content of the actual observed value y_i under the distribution of the weighted predictions for x_i, that is, $\mathfrak{C}(x_i) = \{\mathcal{C}(x_i), \mathcal{W}\}$. Then, the sum of information estimates is minimized with respect to $\mathcal{W}$, as a small information content is equivalent to large likelihood, and it implies that it is highly likely that y_i is an outcome of $\mathfrak{C}(x_i)$.

The remainder of this paper is organized as follows. In section 2, we will show a brief literature survey on estimators of information related quantities. Section 3 describes the notion of the quantile, and derives an estimator for the information content with weighted data. By considering stability and computational cost, this basic estimator is extended to create more efficient estimators. In section 4, the cross entropy and entropy estimators are derived from the proposed information estimators. In section 5, the proposed information estimator is validated by a simple experiment with artificial data. Thereafter, applications of the information estimators for problems 1 and 2 are shown. The last section offers concluding remarks. A short and preliminary version of this paper appeared in ICANN 2011, with another application example (Hino & Murata, 2011).

2. Related Works on Estimators of Information Related Quantities

In this section, we will show a concise survey of related works on estimators of information related quantities. In a parametric approach, a pdf is often approximated by a mixture of Gaussian distributions (Leiva-Murillo & Artes-Rodriguez, 2004; Peltonen et al., 2007). In non-parametric approaches, there are two representative non-parametric density estimation methods, the kernel density estimator (KDE) and the k-nearest neighbor (k-NN) based estimator, (see Wand & Jones (1994); Beirlant et al. (1997); Györfi & van der Meulen (1987); Paninski (2003) for examples). Several attempts have been made to estimate information theoretic quantities based on k-NN. An entropy estimator using 1-NN is described in Kozachenko & Leonenko (1987), and its mean-square consistency is proved for data of any dimension. This result is extended to develop a k-NN based estimator (Goria et al., 2005). Divergence estimators based on k-NN are investigated in, for example, Pérez-Cruz (2008) and Wang et al. (2009). The former study proved almost sure convergence of the divergence estimator using a waiting time analysis technique. The latter proved asymptotic unbiasedness and mean-square consistency of the divergence estimator. For estimating mutual information, estimators based on k-NN have also been proposed (Kraskov et al., 2004). Some researchers treated k as a non-decreasing function of n, and proposed density and entropy estimators using a variable k (Loftsgaarden & Quesenberry, 1965; Mnatsakanov et al., 2008). Specifically, in Wang et al. (2009), the authors generalized the k-NN method and allowed k to vary for each datum. Furthermore, Wang et al. (2009) and Kraskov et al. (2004) proposed that, instead of fixing a constant value for k, the number of samples included in a ball of fixed radius centered at an inspection point is used to estimate KL divergence and mutual information. We note that density, entropy, and divergence estimators with variable k depending on the data points are proposed in Loftsgaarden & Quesenberry (1965); Mnatsakanov et al. (2008), and Wang et al. (2009).

The quantile-based formulation used in the present paper is essentially the same as conventional k-NN based methods. However, in the case in which the given data are weighted, it is not clear (at least, it is not a trivial problem) how to find the k-th nearest point taking into account the weights, whereas we can naturally define the α-quantile point for weighted data. We note that kernel density estimators can be easily extended to deal with weighted data. However, it is known that applying KDE to high dimensional data is difficult because it involves the problem of bandwidth selection in high dimensionality.

Finally we also note that the notion of quantile is effectively utilized in the literature of regression (Koenker, 2005). Motivated by robust regression, quantile regression was advocated in Koenker & Bassett (1978), and various extensions of quantile regression were then proposed. For example, quantile regression for censored data was proposed in Portnoy (2003) and Wang & Wang (2009). A nonparametric extension of quantile regression was proposed in Takeuchi et al. (2006).

3. Information Estimator

In this section, three different estimators for the information content $-\log f(z)$ where $f(z)$ is a probability density function (pdf) are developed. The first one is based on the notion of quantile, and the second one is obtained by fixing the radius of the ball in data space centered at the inspection point, which is eventually shown to be equivalent to the kernel density estimator with a hard window kernel. The last estimator extends the first one by averaging out the quantile to obtain a stable and computationally feasible estimator. Information estimators developed in this section will be used to derive entropy and cross-entropy estimators in section 4.

3.1. Information Estimator Based on Quantiles

Let $\|z-x\|$ be the Euclidean distance between z and x in $\mathbb{R}^d$, and $b(z,\varepsilon) = \{x \in \mathbb{R}^d; \|z-x\| < \varepsilon\}$ be an ε-ball centered at z. The volume of this ε-ball is $|b(z,\varepsilon)| = c_d \varepsilon^d$, where $c_d = \pi^{d/2}/\Gamma(1+d/2)$ and $\Gamma(x)$ is the gamma function. Denote the probability mass contained within the ε-ball centered at the inspection point z by $P_z(\varepsilon)$, i.e.,

$$P_z(\varepsilon) = \int_{x \in b(z,\varepsilon)} f(x)\mathrm{d}x. \tag{6}$$

We consider a dataset $\mathfrak{D}$ as given in Eq. (4). Sorting points $x_i \in \mathcal{D}$ in ascending order of $\|z - x_i\|$, we denote the index of the i-th nearest point to z by (i). For the sake of simplicity, we assume that the distances of all points in the observed data $\mathcal{D}$ from the inspection point are different. The *quantile* of $x_{(i)}$ with respect to z is defined by $\sum_{j=1}^{i} w_{(j)}$. If the weight $w_i \in \mathcal{W}$ associated with $x_i \in \mathcal{D}$ is identical for all i, $i = 1, \ldots, n$, then the quantile of $x_{(i)}$ is $\sum_{j=1}^{i} w_{(j)} = i/n$, which is simply a proportion of the whole dataset size. Conversely, when an inspection point z and a quantile α are specified, the point $x_{\iota(\alpha)_z} \in \mathcal{D}$ where $\iota(\alpha)_z = \arg\max_k \sum_{j=1}^{k} w_{(j)} \le \alpha$, is called the *$\alpha$-quantile point*. See figure 2 for an illustrative example of the notion of α-quantile points. From this example, we see that the α-quantile points differ according to weights.

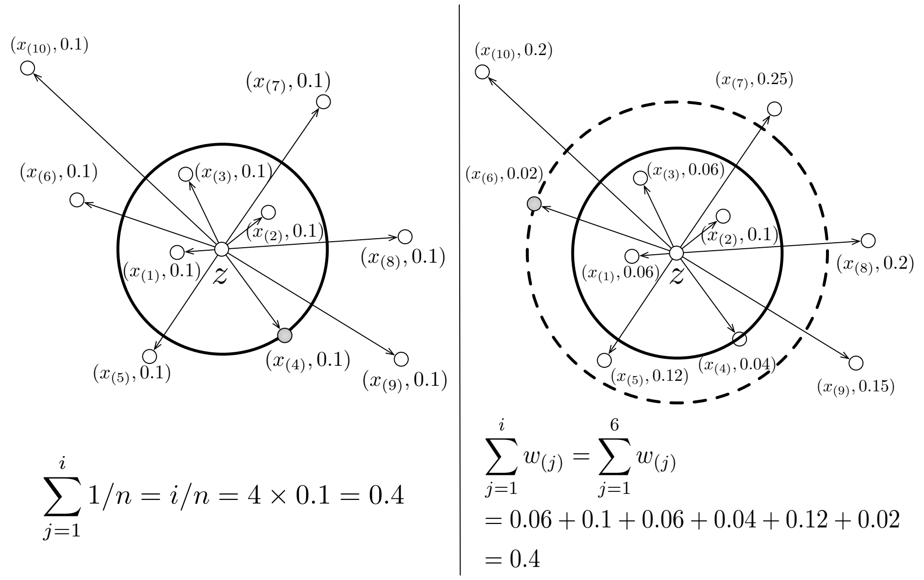


Figure 2: An illustrative example of the notion of α-quantile points. Circles drawn by solid and dashed lines denote d-dimensional hyperspheres centered at the inspection point z. We let $\alpha = 0.4$, and consider the 0.4-quantile point in a given dataset of size $n = 10$. Left: All the weights are equal to $1/n = 0.1$, and the 0.4-quantile point is $x_{(4)}$. Right: Each datum has its own weight. As the sum of the weights up to the 6-th nearest point amounts to 0.4, the 0.4-quantile point is $x_{(6)}$.

Let $\varepsilon_{\alpha,z} = \|z - x_{\iota(\alpha)_z}\|$ be the distance between z and its α-quantile point. When the inspection point z is clear from the context, we simply denote $\varepsilon_{\alpha,z}$ and $\iota(\alpha)_z$ as ε_α and $\iota(\alpha)$, respectively.

Figure 3 (left) is an illustrative example of the relationship between the distance ε and the probability mass function $P_z(\varepsilon)$ for $n = 17$. When the number of data points is finite, the index $\iota(\alpha)$ of the α-quantile point in $\mathfrak{D}$ does not satisfy the equation $\sum_{j=1}^{\iota(\alpha)} w_{(j)} = \alpha$ in general. In this case, there is a gap between ε_α and $\varepsilon_{\hat{\alpha}}$, where $\hat{\alpha} = \sum_{j=1}^{\iota(\alpha)} w_{(j)}$. In the case of large values of n, the staircase pattern in figure 3 (left) becomes a smooth curve, as shown in figure 3 (right). As $n \to \infty$, the gap between ε_α and $\varepsilon_{\hat{\alpha}}$ goes to zero consistently. That is, the probability density function (pdf) of $\varepsilon_{\alpha,z} = \|z - x_{\iota(\alpha)}\|$ is written in terms of the probability mass $P_z(\varepsilon)$, and the variance of ε goes to zero in the large sample limit:

Theorem 1. *Let $p_{\varepsilon_{\alpha,z}}(\varepsilon)$ be the pdf of the distance ε between z and its α-quantile point in a given dataset $\mathfrak{D}$ of size n where α is fixed. Assume that there exists $L \in \mathbb{R}^+$ which depends on α such that $\frac{\iota(\alpha)}{n} \overset{n\to\infty}{\longrightarrow} L < \infty$. Then,*

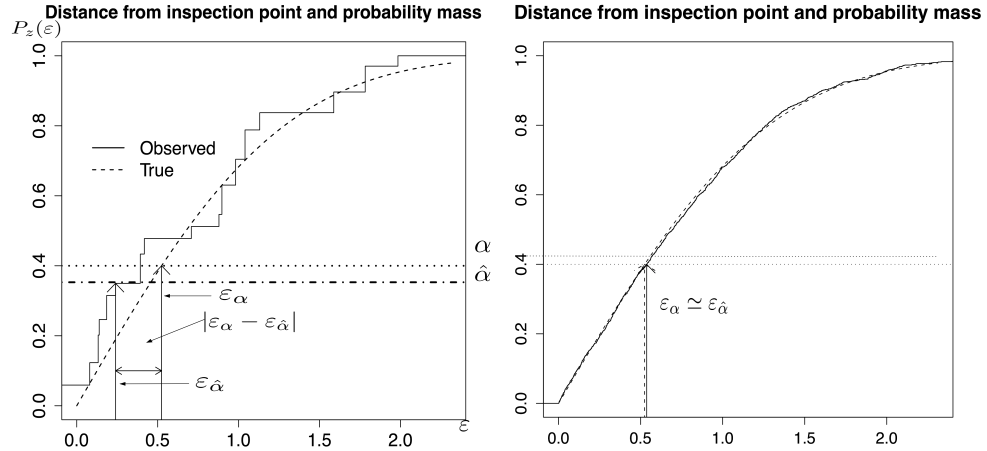


Figure 3: An illustrative example of the probability mass function and the distance from the inspection point x to its α-quantile point when $\alpha = 0.4$. Left: A staircase function for $n = 17$ (solid line segments) and a theoretical curve of $P_x(\varepsilon)$ (dashed line) are shown. In this case, $\hat{\alpha} = \sum_{j=1}^{i} w_{(j)} = 0.353 \neq \alpha$ and the gap $|\varepsilon_{\alpha} - \varepsilon_{\hat{\alpha}}|$ is relatively large. Right: The size of data is large ($n = 1500$) and the gap $|\varepsilon_{\alpha} - \varepsilon_{\hat{\alpha}}|$ is negligible.

for large n,

$$p_{\varepsilon_{\alpha,z}}(\varepsilon) = \phi(P_z(\varepsilon);\beta,\sigma^2)\frac{\mathrm{d}P_z(\varepsilon)}{\mathrm{d}\varepsilon} \tag{7}$$

holds, where $\phi(P_z(\varepsilon);\beta,\sigma^2)$ is the pdf of the Gaussian distribution with mean $\beta = \alpha + \sigma^2\left(\frac{\iota(\alpha)}{\alpha} - \frac{n-\iota(\alpha)}{1-\alpha}\right)$ and variance $\sigma^2 = \left(\frac{\iota(\alpha)}{\alpha^2} + \frac{n-\iota(\alpha)}{(1-\alpha)^2}\right)^{-1}$. Under this distribution, the expectation value of ε is $\varepsilon_\beta = P_z^{-1}(\beta) \overset{n\to\infty}{\longrightarrow} P_z^{-1}(\alpha)$ and the variance goes to zero as $n \to \infty$.

The proof is given in Appendix A. Regarding the weight, we have the following corollary:

Corollary 1. *If all the weights are equal, that is, $\iota(\alpha) = \lfloor \alpha n \rfloor$, where $\lfloor y \rfloor$ for $y \in \mathbb{R}$ is the largest integer not greater than y, then $\beta = \alpha$ and $\sigma^2 = \frac{\alpha(1-\alpha)}{n}$.*

Assuming that

$$\varepsilon_\alpha \ll 1, \tag{8}$$

we obtain the following approximation formula by Taylor's expansion

$$\alpha \;=\; P_z(\varepsilon_\alpha) = \int_{b(z,\varepsilon_\alpha)} f(x)\mathrm{d}x = \int_{b(z,\varepsilon_\alpha)} f(z + (x - z))\mathrm{d}x$$

$$= \quad |b(z,\varepsilon_\alpha)|(f(z)+O(\varepsilon_\alpha^2)) \sim c_d \varepsilon_\alpha^d f(z),$$

where $O(\varepsilon^2)$ denotes the terms higher than or equal to ε^2. Taking the logarithm of both sides of $\alpha \sim c_d \varepsilon_\alpha^d f(z)$, we obtain an estimator for the information content $-\log f(z)$ at an inspection point $z \in \mathbb{R}^d$ as

$$I_\alpha(z;\mathfrak{D}) = \log c_d - \log\alpha + d\log\varepsilon_{\alpha,z}. \tag{9}$$

We call this estimator a *Quantile Information Estimator (QIE)* for z with a weighted dataset $\mathfrak{D} = \{\mathcal{D},\mathcal{W}\}$.

When α is fixed, the bias of the QIE depends on assumption (8), and this assumption is not always valid. When α goes to zero as $n \to \infty$ and αn is fixed as $\alpha n = M < \infty, M \in \mathbb{R}^+$, then $\varepsilon_\alpha \overset{n\to\infty}{\to} 0$. In this case, the estimation bias caused by violation of assumption (8) vanishes as $n \to \infty$, although the bias caused by unequal weights does not vanish. Regarding this bias, the following theorem holds. To show the dependency of $\mathfrak{D}$ on the number of data n, we write the data in Eq. (4) as $\mathfrak{D}_n$.

Theorem 2. *When $\alpha n = M < \infty, M \in \mathbb{R}^+$ is fixed for any value of n, the bias of the QIE is given by*

$$E[I_\alpha(z;\mathfrak{D}_n)] - I_f(z) \overset{n\to\infty}{\to} -\psi(M) + \psi(\iota(\alpha)), \tag{10}$$

where $\psi(\cdot)$ is the digamma function. That is,

$$I_\alpha^{unbiased}(z;\mathfrak{D}) = \log c_d - \log\alpha + \psi(M) - \psi(\iota(\alpha)) + d\log\varepsilon_\alpha \tag{11}$$

is an asymptotically unbiased information estimator.

The proof is given in Appendix B.

The assumption $\alpha n = M < \infty$ means that we always focus on the limited portion of data near the inspection point. For example, since $\alpha n < \infty$ implies $\alpha = O(1/n)$, the assumption is satisfied when $\iota(\alpha) = constant$ and $\max\{w_i\} \leq O(1/n)$. If all of the weights are equal to $1/n$, $\alpha n = M = \iota(\alpha)$ holds because $\alpha = \frac{\iota(\alpha)}{n}$. This is a simple example of the ideal case where the bias vanishes as $n \to \infty$.

3.2. Fixed Radius Quantile Information Estimator

To use the QIE appropriately, we must determine the value of α, depending on the given dataset $\mathfrak{D} = \{\mathcal{D},\mathcal{W}\}$. Furthermore, to find the α-quantile point in $\mathfrak{D}$, we must sort all the data with a computational cost on the order of $O(n\log n)$.

A simple solution of avoiding the sort operation is to consider the ratio of data points within balls of fixed radius to estimate the information content. Let $\alpha_R(z;\mathfrak{D})$ be the sum of weights associated with data within an R-ball $b(z,R) = \{x_i \in \mathcal{D} | \|z - x_i\| < R\}$. Then, we obtain a *fixed Radius Quantile Information Estimator (RQIE)* with a weighted dataset $\mathfrak{D}$ as

$$I_{\mathrm{R}}(z;\mathfrak{D}) = \log c_d - \log \alpha_R(z;\mathfrak{D}) + d \log R, \tag{12}$$

where

$$\alpha_R(z;\mathfrak{D}) \quad = \sum_{i:\|z-x_i\|<R} w_i.$$

The value of R for the RQIE should be determined based on the distribution of the given dataset $\mathfrak{D}$. When an appropriate R value is chosen, we can avoid violation of assumption (8) and reduce computational cost from $O(n \log n)$ to $O(n)$, because $\alpha_R(z;\mathfrak{D})$ can be calculated without sorting the data.

Finally, we mention the relationship between the proposed RQIE and kernel density estimator with hard window (uniform) kernel (Wand & Jones, 1994). The information content $I_R(z;\mathfrak{D})$ is an estimate of the quantity $-\log f(z)$, and we denote $f^R_{\mathfrak{D}}(z) = e^{-I_R(z;\mathfrak{D})}$. From Eq. (12), it is written as

$$f^R_{\mathfrak{D}}(z) = \frac{\alpha_R(z;\mathfrak{D})}{c_d R^d}. \tag{13}$$

When all the weights w_i are equal to $1/n$, the quantile $\alpha_R(z;\mathfrak{D})$ is reduced to the number of points in a ball of radius R centered at z, and the above equation becomes

$$f^R_{\mathcal{D}}(z) = \frac{\#\{x_i \in \mathcal{D} \mid \|(z - x_i)/R\| < 1\}}{c_d}, \tag{14}$$

which is nothing but the kernel density estimator with a hard window kernel, and the radius R is considered to be the kernel bandwidth. This particular case is the crossroad of the information estimator based on the notion of quantile and the kernel density estimator. In other words, the RQIE gives a natural extension of the KDE with a hard window to a weighted dataset.

3.3. Mean Quantile Information Estimator

Another approach for reducing computational cost is marginalization of α in the QIE $I_\alpha(z;\mathfrak{D})$ as

$$\int_0^1 I_\alpha(z;\mathfrak{D})\mathrm{d}\alpha = \log c_d - \int_0^1 \log \alpha \mathrm{d}\alpha + d \int_0^1 \log \varepsilon_\alpha \mathrm{d}\alpha$$

$$= \log c_d + 1 + d \int_0^1 \log \varepsilon_\alpha \mathrm{d}\alpha. \tag{15}$$

We note that $\int_0^1 \log \alpha \mathrm{d}\alpha$ is defined in the sense of improper integration as $\int_u^1 \log \alpha \mathrm{d}\alpha = [\alpha \log \alpha - \alpha]_u^1 = -1 - u \log u + u \overset{u \to 0+}{\longrightarrow} -1$. In actual calculation of (15), the integration $\int \log \varepsilon_\alpha \mathrm{d}\alpha$ is approximated by a summation of values at the observed data points. When the weight for each datum is $1/n$, this approximated integration is calculated by summation of rectangles of equal width, $1/n$, as shown in figure 4 (left). On the other hand, when each datum x_i has its own weight w_i, each datum contributes to the sum at the rate of its weight w_i, and the approximation is calculated by summation of rectangles of different widths w_i, as shown in figure 4 (right). As α increases from 0 to 1, each point in $\mathfrak{D}$ becomes the α-quantile point once for each α. Then, we obtain an information estimator by integrating $I_\alpha(z; \mathfrak{D})$ with respect to α as

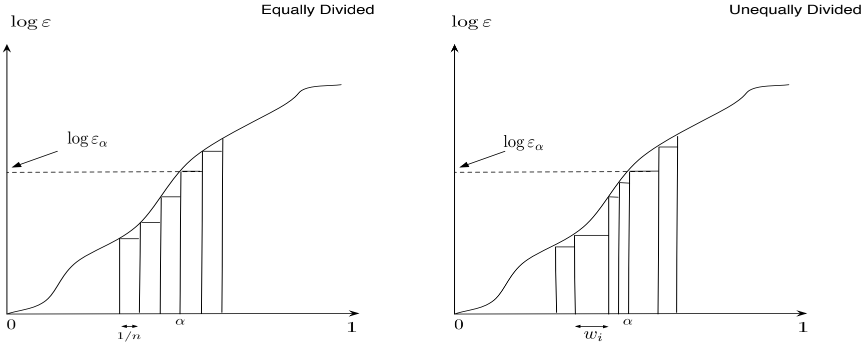


Figure 4: Difference of quadrature. Left: Approximation of integral by sum of equal width rectangles. Right: Approximation of integral by sum of unequal width rectangles.

$$\begin{aligned} I_{\mathrm{MQ}}(z; \mathfrak{D}) &= \log c_d + 1 + d \sum_{i=1}^n w_i \log \|z - x_i\| \\ &= \log c_d + 1 + d E_{\mathfrak{D}} \left[\log \|z - X\| \right]. \end{aligned} \tag{16}$$

We call this estimator a *Mean Quantile Information Estimator (MQIE)* with a weighted dataset $\mathfrak{D}$. We note that if all of the weights are equal to $1/n$, then the MQIE is a quantile analogue of the MeanNN method (Faivishevsky & Goldberger, 2009), which modifies the k-nearest neighbor entropy estimator to a parameter-free estimator by taking the sum of the estimators for all values of k.

To satisfy the assumption (8), α must be sufficiently small. However, the MQIE is the average of the QIE for $\alpha \in [0,1]$, and the assumption can be violated for a large value of α. Consequently, the estimate is biased. On the other hand, the MQIE requires neither calculating the α-quantile point nor sorting the data according to $\|z - x_i\|$, and, as a result, its computational cost is on the order of $O(n)$. We can introduce a weight function $w(\alpha)$ so that $I_\alpha(z; \mathfrak{D})$ with large α does not contribute to the integral $\int_0^1 I_\alpha(z;\mathfrak{D}) w(\alpha) \mathrm{d}\alpha$. With careful choice of the weight function for α, we observed that the estimation bias can be reduced; however, we will be faced with another problem of optimizing the weight function. In the present paper, we only consider uniform weight for α in the integral. We note that using a smaller α or weighting the α will effectively reduce the amount of data used for the estimation, and there would be a trade off between the bias and the variance due to using few data. This conjecture is numerically supported by the observation that the variance of the MQIE is consistently small compared to that of the QIE, but the MQIE may contain consistent bias as shown in section 5.

4. Cross Entropy and Entropy Estimator

The cross entropy and entropy are defined as averages of the information content. Their estimators are derived from estimators proposed in the previous section.

4.1. Entropy Estimation Based on the Mean Quantile Information Estimator

Given two sets of weighted data

$$\begin{aligned} \mathfrak{D} &= \{\mathcal{D}_x, \mathcal{W}\} = \{(x_i, w_i)\}_{i=1}^n, &(17)\\ \mathfrak{D}' &= \{\mathcal{D}_y, \mathcal{V}\} = \{(y_j, v_j)\}_{j=1}^m, &(18)\end{aligned}$$

we will derive cross entropy estimators. Let $y_{\iota(\alpha)_{x_i}} \in \mathcal{D}_y$ be the α-quantile point in $\mathfrak{D}'$ from a point $x_i \in \mathcal{D}_x$. We note that the definition of the α-quantile is the same as in section 3.1, and the only difference is that the inspection point x_i is taken from another dataset $\mathcal{D}_x$. With a weighted dataset $\mathfrak{D}'$, following (9), the QIE at the inspection point x_i is given by

$$I_\alpha(x_i; \mathfrak{D}') = \log c_d - \log \alpha + d \log \|y_{\iota(\alpha)_{x_i}} - x_i\| \simeq -\log f_{\mathfrak{D}'}(x_i), \qquad (19)$$

where we formally denote the pdf estimated using $\mathfrak{D}$ by $f_{\mathfrak{D}}$. Averaging this estimate with respect to $x_i \in \mathcal{D}_x$, we obtain the *Quantile Cross Entropy Estimator*:

$H_\alpha(\mathfrak{D}, \mathfrak{D}')$

$$= \log c_d - \log \alpha + d \sum_{i=1}^{n} w_i \log \|y_{\iota(\alpha)_{x_i}} - x_i\| \simeq E_{f_{\mathfrak{D}}}[-\log f_{\mathfrak{D}'}(X)]. \tag{20}$$

Furthermore, integrating α out taking account of the weight for $y_{\iota(\alpha)_{x_i}}$ as in (16), we obtain the *Mean Quantile Cross Entropy Estimator* as

$$H_{\mathrm{MQ}}(\mathfrak{D}, \mathfrak{D}') = \log c_d + 1 + d \sum_{i=1}^{n} \sum_{j=1}^{m} w_i v_j \log \|y_j - x_i\|. \tag{21}$$

We note that $H_{\mathrm{MQ}}(\mathfrak{D}, \mathfrak{D}')$ is regarded as an empirical average of $I_{\mathrm{MQ}}(x_i; \mathfrak{D}') \simeq -\log f_{\mathfrak{D}'}(x_i)$ with a weighted observed dataset $\mathfrak{D}$.

To estimate the entropy with the dataset $\mathfrak{D} = \{\mathcal{D}, \mathcal{W}\} = \{(x_i, w_i)\}_{i=1}^{n}$, we can use the leave-one-out estimation procedure (Hastie et al., 2001; Vapnik, 1998). Let $\bar{\mathfrak{D}}_i = \{\mathcal{D} \backslash x_i, \frac{1}{1-w_i}(\mathcal{W} \backslash w_i)\}$ be a renormalized weighted dataset excluding $\{(x_i, w_i)\}$. With this weighted dataset $\bar{\mathfrak{D}}_i$, the α-quantile information estimator is written as

$$I_\alpha(x_i; \bar{\mathfrak{D}}_i) = \log c_d - \log \alpha + d \log \|x_i - x_{\iota(\alpha)_{x_i}}\|, \tag{22}$$

where $x_{\iota(\alpha)_{x_i}} \in \mathcal{D} \backslash x_i$ is the α-quantile point of the inspection point $x_i \in \mathcal{D}$. By averaging $I_\alpha(x_i; \bar{\mathfrak{D}}_i)$ with respect to x_i by the leave-one-out estimation procedure, the α-quantile entropy estimator is given by

$$H_\alpha(\mathfrak{D}) = \log c_d - \log \alpha + d \sum_{i=1}^{n} w_i \log \|x_i - x_{\iota(\alpha)_{x_i}}\|. \tag{23}$$

By integrating $H_\alpha(\mathfrak{D})$ with respect to α, we obtain the *Mean Quantile Entropy Estimator*

$$H_{\mathrm{MQ}}(\mathfrak{D}) = \log c_d + 1 + d \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \frac{w_i w_j}{1 - w_i} \log \|x_i - x_j\|. \tag{24}$$

This estimator inherits properties of the MQIE, namely, there is no need to specify an appropriate α at the cost of introducing a bias by violation of the assumption (8).

4.2. Entropy Estimation Based on the Fixed Radius Quantile Information Estimator

We can also extend the RQIE defined in (12) to estimate the cross entropy and entropy. Suppose weighted datasets $\mathfrak{D}$ and $\mathfrak{D}'$ defined by (17) and (18) are given. Let the information estimate at the inspection point x_i with the weighted dataset $\mathfrak{D}'$ be

$$I_R(x_i;\mathfrak{D}') = \log c_d - \log \alpha_R(x_i;\mathfrak{D}') + d\log R. \tag{25}$$

Averaging the above information estimator with respect to $x_i \in \mathcal{D}_x$, we obtain a cross entropy estimator

$$H_{\mathrm{R}}(\mathfrak{D},\mathfrak{D}') \;=\; \log c_d - \sum_{i=1}^{m} w_i \log \alpha_R(x_i;\mathfrak{D}') + d\log R, \tag{26}$$

where

$$\alpha_R(x_i;\mathfrak{D}') = \sum_{j:\|y_j - x_i\|<R} v_j. \tag{27}$$

We next derive an entropy estimator with a fixed radius R. Taking care of the renormalized weight $w'_j = \frac{w_j}{1-w_i}$, the RQIE at the inspection point $x_i \in \mathcal{D}$ with the weighted dataset $\bar{\mathfrak{D}}_i$ becomes

$$I_R(x_i;\bar{\mathfrak{D}}_i) = \log c_d - \log \alpha_R(x_i;\bar{\mathfrak{D}}_i) + d\log R, \tag{28}$$

where

$$\alpha_R(x_i;\bar{\mathfrak{D}}_i) = \sum_{\substack{j:\|x_j - x_i\|<R \\ j\neq i}} \frac{w_j}{1-w_i}. \tag{29}$$

Then, averaging this information estimate with respect to x_i yields an entropy estimator

$$H_R(\mathfrak{D}) = \log c_d - \sum_{i=1}^{n} w_i \log \alpha_R(x_i;\bar{\mathfrak{D}}_i) + d\log R. \tag{30}$$

5. Numerical Experiments

In this section, we first verify that the proposed information estimators work properly using artificial data. Then, in sections 5.2 and 5.3, instances of problem 1 and problem 2 are presented with real-world data sets, and are solved by weight optimization.

5.1. Information Estimation Experiments on Artificial Data

In this subsection, we show simple experimental results for artificial data. We compare the proposed information estimators $I_\alpha(z;\mathfrak{D})$ in Eq. (9), $I_R(z;\mathfrak{D})$ in Eq. (12), $I_{MQ}(z;\mathfrak{D})$ in Eq. (16), and a classical non-parametric method $-\log \hat{f}_G(z;\mathfrak{D})$ with the pdf $\hat{f}_G(z;\mathfrak{D})$ estimated by the KDE with a Gaussian kernel. We calculated the average and standard deviation from ground truth values and computational time as indices of comparison.

We sampled 300 points from 5 different one dimensional Gaussian distributions with mean 0 and variances $\sigma^2 \in \{1, \dots, 5\}$ to be used as the datasets for the experiments. We created 200 independent datasets of size 300 sampled from a Gaussian distribution $\phi(x; 0, \sigma^2)$, and we estimated the information content at inspection points $z = 0$ and $z = \sigma$ using four information estimators. In figure 5, the ground truth information contents at $z = 0$ and $z = \sigma$ are depicted with solid and dashed lines, respectively. For the KDE, we performed tests with both a Gaussian kernel and an Epanechinikov kernel; we adopted the Gaussian kernel as it showed more accurate results. We note that the RQIE (12) is equivalent to the KDE with a hard window kernel. We consider both unweighted datasets and weighted datasets. For the weighted data experiment, the dataset $\mathcal{D}$ is first sampled from a uniform distribution on the interval $[-2\sigma, 2\sigma]$. Thereafter, for each $x_i \in \mathcal{D}$, a weight is given by $w_i = \phi(x_i; 0, \sigma^2)/\sum_{j=1}^{300} \phi(x_j; 0, \sigma^2)$, where $\phi(x; 0, \sigma^2)$ is the pdf of the ground truth Gaussian distribution. With these weights, each datum x_i contributes to the information estimate as if it were distributed as a (truncated) Gaussian distribution with mean 0 and variance σ^2. Intuitively speaking, the weighted data generated by this procedure is the same as the (truncated) Gaussian distribution in terms of the averaging operation (5). A kernel density estimator with the Gaussian kernel function is written as $\hat{f}_G(z; \mathcal{D}) = (2\pi)^{-d/2}(nh)^{-1} \sum_{x_i \in \mathcal{D}} \exp(-\|(z - x_i)/h\|^2/2)$, where h is the bandwidth parameter, and replacing $1/n$ by w_i, its weighted version becomes

$$\hat{f}_G(z; \mathfrak{D}) = (2\pi)^{-d/2}(h)^{-1} \sum_{(x_i, w_i) \in \mathfrak{D}} w_i \exp(-\|(z - x_i)/h\|^2/2). \tag{31}$$

In the above estimators, the KDE, QIE, and RQIE have one tuning parameter (the bandwidth of the Gaussian kernel, quantile, and radius, respectively). In this experiment, we searched the optimal parameters by 10-fold cross validation. That is, the parameter value which maximizes the likelihood (i.e., minus of the Shannon information content) of retained samples is adopted for estimating the information of inspection points z. The experimental results are shown in table 1 and figure 5. We also performed information estimation experiments using datasets from Poisson distributions. We created 200 independent one-dimensional integer-valued data sets of size 300 sampled from the Poisson distribution $\pi(x; p) = p^x e^{-p}/x!$, $x \in \mathbb{N}$, and we estimated the information content at inspection points $z = p$ and $z = \lfloor 1.2 \times p \rfloor + 1$. We performed experiments using 5 different values of the location parameter $p \in \{1, 5, 15, 20, 30\}$. The information contents are estimated by using four different information estimators. For weighted data experiments, we first sampled 300 data points from a discrete uniform distribution on $\{0, 1, 2, \ldots, 2p\}$. Then, a weight w_i for each x_i was set to $w_i = \pi(x_i; p)/\sum_{j=1}^{300} \pi(x_j; p)$. The averaged biases and standard deviations of estimation are averaged for all parameters p and reported in table 2. For discrete distributions such as Poisson, it would be better to use specially designed kernels for discrete distributions (Hall & Titterington, 1989; Rajagopalan & Lall, 1995). We simply used the Gaussian kernel for both continuous and discrete observations.

To investigate the influence of the dimension of the data to the accuracy of estimation, we also conducted the above-mentioned experiment using datasets generated from d-dimensional Gaussian distributions with mean zero and spherical covariance matrices $\sigma^2 I_d,\ d = 1, 2, \ldots, 20$. There are some methods to extend the KDE to deal with multidimensional data. We adopt a single bandwidth kernel estimator $\hat{f}_G(z; \mathfrak{D}) = (2\pi)^{-d/2}(h)^{-d} \sum_{(x_i,w_i)\in\mathfrak{D}} w_i \exp(-\|(z-x_i)/h\|^2/2)$, because the QIE and RQIE also have only one parameter to be selected. 300 samples are generated from a d-dimensional Gaussian distribution, and tuning parameters for the KDE, RQIE and QIE are optimized by 10-fold cross validation. Then, the Shannon information contents at $z = 0 \in \mathbb{R}^d$ and $z = \sigma 1_d \in \mathbb{R}^d$ are estimated, where 1_d is a d dimensional vector with all ones. This procedure is repeated 200 times using independently sampled datasets to calculate mean absolute errors of the estimates. Figure 6 (a) shows the relationship between mean absolute errors of estimates and the dimensions of the data when $z = 0 \in \mathbb{R}^d$. We also show averaged biases and standard deviations for 20 dimensional data in table 1.

Finally, in figure 6 (b), we show the averages and one standard deviations of computational time for each method to estimate the information contents using 300 samples from a one dimensional standard Gaussian distribution. The averages and standard deviations are calculated using 200 independent trials.

From these results, we can deduce the following properties of the information estimators:

1. As shown in figures 5 (a) and (c), and in table 1 for $d = 1$, biases and standard deviations of the KDE with a Gaussian kernel and the RQIE are similar when unweighted datasets are used. This similarity is plausible because the RQIE is equivalent to the KDE with a hard window kernel as mentioned in section 3.2.
2. As shown in figures 5 (b) and (f) versus figures 5 (c) and (g), and in table 1, the variance of the QIE is relatively large compared to that of the RQIE. This is accountable by observing that the optimized α's are

small, and only a small part of data are used to estimate the information contents. Consequently, the variances became large due to using few data.

3. The biases of the MQIE are relatively large; however, its standard deviations are always the smallest among all estimators. This is because the MQIE is derived by integrating the QIE with respect to α, which is essentially the same as averaging many QIEs for different values of α. This averaging results in variance reduction. However, as a side effect, condition (8) is violated for large α and this may be the reason for relatively large and consistent biases of the MQIE where the ground truth values are not within a standard deviation of the estimates.
4. As shown in figures 5 (e) to (h) and in table 1 and table 2, all estimators are able to evaluate the weighted data. The influences of weights are different for each estimator. For example, for one dimensional data, biases and standard deviations of the KDE and RQIE increase when weights are considered, while biases of the QIE and MQIE decrease and standard deviations stay about the same when weights are considered. For 20 dimensional data, the effects of the weights seem to be different from the case of one dimensional data. Further investigation to explain the effects of weights to each estimator is remained as an important future work.
5. Figure 6 reveals interesting tendencies of various estimators. When we deal with one dimensional data, the KDE with a Gaussian kernel and the RQIE (which is a KDE with a hard window kernel) offers small bias. However, the accuracies of both the KDE and the RQIE deteriorate quickly as the dimension increases, while the accuracy of the QIE does not deteriorate so much. As stated in section 2, this can be accounted by the difficulty of bandwidth and radius tuning for the KDE and RQIE in high-dimensional situations. From this result, we see that the QIE is suitable for high-dimensional data. The MQIE is in between the KDE and the QIE.
6. As shown in figure 6 (b), the KDE method is the fastest method tested in this experiment[2]. The QIE is relatively slow, as we must sort all the data in $\mathcal{D}$ to calculate the α-quantile point. The MQIE can reduce the computational cost and standard deviation of the estimate at the cost of an additional bias. We note that the computational time reported

[2] All numerical experiments in this paper are implemented with R-language version 2.9.1 (R Development Core Team, 2010) and processed on an Intel machine with 2.93 GHz dual processors, 8 GB memory, and the operating system is Mac OS X version 10.6.5

here exclude the parameter optimization phase by cross validation. The MQIE does not require parameter optimization, while other methods need parameter optimization to obtain acceptable estimation accuracy, which may strongly increase the total computational time.

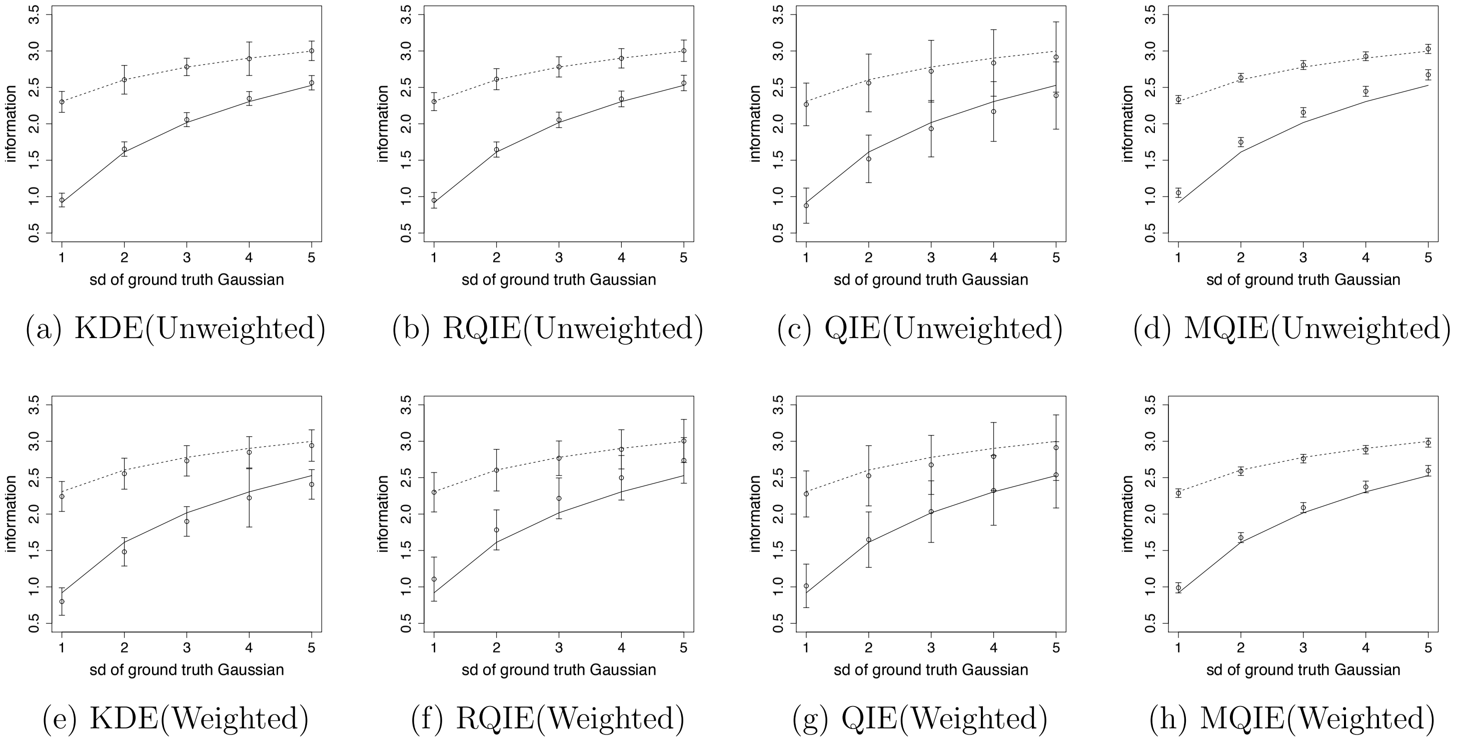
Table 1: Estimating the information with data from Gaussian distributions. Averaged biases and standard deviations for $\sigma^2 = 1, \ldots, 5$ of various information estimators at the inspection points $z = 0$ and $z = \sigma$. Each value is associated with its standard deviation in parentheses. Results for one and 20 dimensions are shown.

dimension=1				
Unweighted	KDE	RQIE	QIE	MQIE
Averaged Bias(z=0)	0.038(0.003)	**0.034**(0.003)	0.099(0.040)	0.139(0.004)
Averaged SD (z=0)	0.097(0.002)	0.106(0.002)	0.365(0.085)	**0.066**(0.003)
Averaged Bias(z=σ)	0.011(0.008)	**0.010**(0.004)	0.136(0.038)	0.064(0.006)
Averaged SD (z=σ)	0.165(0.047)	0.137(0.009)	0.411(0.073)	**0.060**(0.003)
Weighted	KDE	RQIE	QIE	MQIE
Averaged Bias(z=0)	0.115(0.018)	0.192(0.014)	**0.035**(0.034)	0.067(0.002)
Averaged SD (z=0)	0.238(0.091)	0.295(0.017)	0.407(0.071)	**0.072**(0.005)
Averaged Bias(z=σ)	0.129(0.016)	**0.020**(0.011)	0.192(0.072)	0.042(0.004)
Averaged SD (z=σ)	0.211(0.004)	0.272(0.022)	0.364(0.074)	**0.059**(0.001)
dimension=20				
Unweighted	KDE	RQIE	QIE	MQIE
Averaged Bias(z=0)	12.974(1.563)	9.268(1.492)	**2.803**(1.110)	8.397(0.204)
Averaged SD (z=0)	0.416(0.072)	1.277(0.038)	1.206(0.044)	**0.204**(0.008)
Averaged Bias(z=σ)	8.020(2.612)	3.306(0.682)	**1.302**(0.762)	5.449(0.163)
Averaged SD (z=σ)	0.744(0.299)	0.686(0.029)	1.014(0.022)	**0.163**(0.006)
Weighted	KDE	RQIE	QIE	MQIE
Averaged Bias(z=0)	15.208(3.103)	11.284(1.277)	3.137(0.830)	**2.713**(0.528)
Averaged SD (z=0)	0.640(0.067)	1.492(0.042)	0.843(0.028)	**0.534**(0.047)
Averaged Bias(z=σ)	2.058(1.350)	2.030(0.651)	**1.010**(0.671)	3.036(0.333)
Averaged SD (z=σ)	1.115(0.522)	0.680(0.016)	1.112(0.071)	**0.333**(0.010)

Table 2: Estimating the information with data from one dimensional Poisson distributions. Averaged biases and standard deviations for $p = 1, 5, 15, 20, 30$ of various information estimators at the inspection points $z = p$ and $z = \lfloor 1.2 \times p \rfloor + 1$. Each value is associated with its standard deviation in parentheses.

Unweighted	KDE	RQIE	QIE	MQIE
Averaged Bias(z=0)	0.320(0.054)	0.388(0.181)	**0.246**(0.160)	0.358(0.071)
Averaged SD (z=0)	0.039(0.032)	0.448(0.082)	0.281(0.163)	**0.031**(0.006)
Averaged Bias(z=$\lfloor 1.2 \times p \rfloor + 1$)	**0.129**(0.063)	0.282(0.167)	0.141(0.076)	0.242(0.092)
Averaged Bias(z=$\lfloor 1.2 \times p \rfloor + 1$)	0.040(0.016)	0.308(0.076)	0.114(0.020)	**0.032**(0.005)

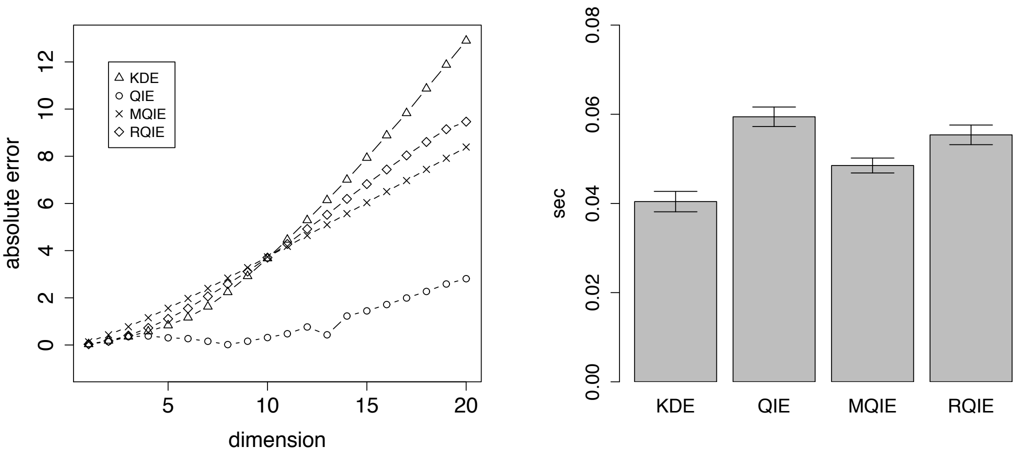
Weighted	KDE	RQIE	QIE	MQIE
Averaged Bias(z=0)	0.769(0.202))	1.23(0.439)	1.062(0.219)	**0.351**(0.070)
Averaged SD (z=0)	0.050(0.021)	0.310(0.169)	0.111(0.028)	**0.040**(0.016)
Averaged Bias(z=$\lfloor 1.2 \times p \rfloor + 1$)	0.483(0.100)	0.966(0.285)	0.670(0.144)	**0.237**(0.090)
Averaged SD (z=$\lfloor 1.2 \times p \rfloor + 1$)	0.090(0.003)	0.230(0.174)	0.119(0.050)	**0.042**(0.014)



(a) KDE(Unweighted) (b) RQIE(Unweighted) (c) QIE(Unweighted) (d) MQIE(Unweighted)

(e) KDE(Weighted) (f) RQIE(Weighted) (g) QIE(Weighted) (h) MQIE(Weighted)

Figure 5: The result of estimation of the information content with artificial data. Averaged estimates in 200 trials are plotted with one standard deviation error bars. Ground truth information contents are shown in solid lines (at inspection point $z = 0$) and dashed lines (for $z = \sigma$).



(a) Error and dimension

(b) Computational time

Figure 6: (a): Mean absolute errors of the estimates of the information contents as functions of dimensionality of the data. (b): average and one standard deviation of computational time for each method.

The experimental results demonstrate that the proposed information estimators are shown to work correctly. In general, from table 1 and table 2, it is seen that the QIE and MQIE perform better than the KDE and RQIE, which are based on kernel density estimation. Particularly, the MQIE is experimentally shown to work well in both weighted and unweighted settings. Compared to the KDE, the QIE is computationally inefficient, but by averaging QIEs with respect to α, the computational cost is reduced, and moreover the variance of the estimate is reduced by a large margin. In practical problems, it is often more important to estimate the information content or entropy in a stable manner. In particular, when we want to optimize the information content or entropy by variable transformation or some other methods, the derivative of the estimator is often used. In such situations, as is argued in Faivishevsky & Goldberger (2009), the stability of estimates is more important than obtaining exact values of the information contents.

5.2. Application to Distribution Preserving Data Compression

We consider problem 1 presented in the introduction, that is, the problem of optimizing weights $\mathcal{W}'$ for the compressed dataset $\mathcal{D}'$, which we call a set of codebook vectors, to let the distributions of $\mathfrak{D}'$ and $\mathfrak{D}$ be as close as possible. As a measure of similarity of distributions, we consider the KL divergence between $\mathfrak{D}' = \{\mathcal{D}', \mathcal{W}'\} = \{(x'_j, w'_j)\}_{j=1}^m$ and $\mathfrak{D} = \{\mathcal{D}, \mathcal{U}\} = \{(x_i, 1/n)\}_{i=1}^n$, which is estimated by the difference between the mean quantile cross entropy estimator (21) and the mean quantile entropy estimator (24):

$$\begin{aligned} D_{KL}(\mathfrak{D}',\mathfrak{D}) &= H(\mathfrak{D}',\mathfrak{D}) - H(\mathfrak{D}') \simeq H_{\mathrm{MQ}}(\mathfrak{D}',\mathfrak{D}) - H_{\mathrm{MQ}}(\mathfrak{D}') \\ &= \frac{d}{n}\sum_{j=1}^{m}\sum_{i=1}^{n} w'_j \log\|x'_j - x_i\| \\ &\quad -d\sum_{j=1}^{m}\sum_{k\neq j}\frac{w'_j w'_k}{1-w'_j}\log\|x'_j - x'_k\|. \end{aligned} \tag{32}$$

This estimate is the objective function of minimization with respect to $\mathcal{W}' = \{w'_j\}_{j=1}^m$ with constraints $\sum_{j=1}^m w'_j = 1, w'_j > 0$. In the experiments shown below, we minimized $D_{KL}(\mathfrak{D}',\mathfrak{D})$ by a quasi-Newton method (the BFGS method; Nocedal & Wright, 2006) equipped with R's `constrOptim` function.

We consider optimizing weights for codebook vectors. We compress the original data $\mathcal{D}$ into the small sized codebook vectors $\mathcal{D}'$ by LVQ, and only store the codebook vectors so as to conserve the storage resources. We suppose the original dataset consists of two classes. Compression by LVQ is performed in each class. Then, in each class, weights for codebook vectors are optimized so that the KL divergence between the codebook vectors and original data is minimized. The obtained weights for class 1 are denoted by $W' = \{w'_1, w'_2, \dots\}$ and the weights for class 2 are denoted by $V' = \{v'_1, v'_2, \dots\}$. Thereafter, using the stored data, we classify the incoming data using the k-NN classifier (Duda et al., 2000). Let x_e be the test point, and $\mathcal{C}'^e = \{\mathcal{C}'^e_1, \mathcal{C}'^e_2\}$ be an index set of codebook vectors which consists of the k-th nearest codebook vectors from the test point x_e. $\mathcal{C}'^e_1$ and $\mathcal{C}'^e_2$ are subsets of $\mathcal{C}'^e$ which correspond to codebook vectors with class labels C_1 and C_2, respectively. The test point is classified as class C_1 if $\sum_{i\in\mathcal{C}'^e_1} w'_i \geq \sum_{j\in\mathcal{C}'^e_2} v'_j$ and as class C_2 if $\sum_{i\in\mathcal{C}'^e_1} w'_i < \sum_{j\in\mathcal{C}'^e_2} v'_j$. To see the effectiveness of the proposed weight optimization for small sized codebook vectors, we employ the IDA datasets, which are the standard binary classification datasets originally used in Rätsch et al. (2001). The original data are compressed to one-tenth by LVQ, and the weights for the codebook vectors are optimized. In this experiment, we set $k = 7$. We compared two weighting schemes; one using the weights obtained by counting the proportion of training samples that are assigned to each codebook vector, and the other using the optimized weights by the proposed method. In table 3, for many datasets, we see that the k-NN classification accuracy has improved with the proposed method. In the k-NN classification, there are some test instances which are misclassified when all codebook vectors have the same weight. We infer that some such instances are correctly classified when we assign appropriate weights for the codebook

Table 3: In the left part, the dimensions of feature vectors, the numbers of training data and test data, and the numbers of realizations (pairs of training and test datasets) of the IDA are shown. In the right part, averages and standard deviations of misclassification rates (in percent) by k-NN classification with full data, LVQ compressed data, LVQ compressed data with weights obtained by counting the proportion of training samples assigned to each codebook vector, and LVQ compressed data with optimized weights are shown. The numbers in parentheses denote standard deviations. Comparing the LVQ, LVQ+count and LVQ+opt, the best results are shown in boldface.

	specifications				Results			
Data name	dim	# train	# test	# sets	Original	LVQ	LVQ+count	LVQ+opt
banana	2	400	4900	100	11.44(0.54)	24.43(5.08)	24.56(2.78)	**23.93**(4.53)
breast-cancer	9	200	77	100	27.53(4.21)	**26.66**(4.50)	30.03(6.92)	26.94(4.83)
diabetes	8	468	300	100	27.02(.200)	27.04(2.10)	30.14(2.90)	**26.65**(2.17)
flare-solar	9	666	400	100	35.46(2.01)	36.80(3.18)	39.97(3.69)	**36.48**(3.30)
german	20	700	300	100	25.62(2.45)	26.42(2.30)	27.57(2.68)	26.42(2.52)
heart	13	170	100	100	17.55(3.47)	18.47(3.51)	21.29(3.59)	**18.27**(3.42)
image	18	1300	1010	20	5.17(0.70)	13.93(2.18)	18.33(2.06)	**13.65**(2.12)
ringnorm	20	400	7000	100	45.03(1.18)	47.67(1.90)	**41.36**(6.06)	47.60(2.06)
splice	60	1000	2175	20	26.43(2.02)	21.36(2.20)	33.40(4.41)	**21.00**(2.06)
thyroid	5	140	75	100	8.71(2.72)	27.48(3.67)	**22.72**(4.49)	24.36(4.42)
titanic	3	150	2051	100	22.91(0.84)	22.67(0.62)	22.68(0.62)	**22.65**(0.50)
twonorm	20	400	7000	100	3.42(0.23)	3.30(0.35)	4.18(1.06)	**3.28**(0.37)
waveform	21	1000	1000	100	12.09(0.56)	11.13(0.96)	14.36(1.73)	**11.11**(0.98)

vectors. We also note that LVQ with weight assigned by counting performs worse than LVQ without weight. This can be because the LVQ codebook is optimized for classification rather than for density estimation. This result suggests the importance of weight optimization.

5.3. Application to Weighted Ensemble Regression

We now consider problem 2 described in the introduction, that is, the problem of optimizing weights $\mathcal{W}$ for an ensemble of regressors. In problem 2, the regressors are supposed to be given. In our experiments, base regressors are generated by the Bagging method (Breiman, 1996). Let $\mathfrak{C} = \{\mathcal{C}, \mathcal{W}\}$ be a pair of a regressor set and a weight set. The regressors $c_j : \mathbb{R}^d \to \mathbb{R}$ are trained using a training dataset $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$. Then, the problem is to optimize weights $\mathcal{W} = \{w_j\}_{j=1}^m$ for regressors such that the average information content of the response variable is minimized under the empirical distribution of $\mathfrak{C}$. In this study, we adopt rpart (recursive partitioning and regression trees; E. J. Atkinson & Therneau 1997) as a base regressor, and adopt the MQIE as an information estimator. We note that in optimizing the KL divergence, the objective function (32) for $\{w_j\}_{j=1}^m$

contains the term $w_j/(1-w_i)$, which prevents w_i from approaching 1. In the case of minimization of the average information content, there is no natural regularization; therefore, we add a log regularization term to the objective function, and solve the following problem:

$$\min_{\{w_j\}_{j=1}^m} \sum_{i=1}^{n} I_{\mathrm{MQ}}(y_i;\mathfrak{C}) - \lambda \sum_{j=1}^{m} \log w_j,\ \lambda \geq 0. \tag{33}$$

The regularization term $\lambda \sum_{j=1}^{m} \log w_j$ is regarded as the minus of log likelihood of the Dirichlet distribution. That is, it is equivalent to place a Dirichlet prior for the weights defined by Eq. (B.5) with $\gamma = \lambda + 1$. This minimization problem is solved by a constrained quasi-Newton method. In this experiment, λ is determined by 10-fold cross validation using the training data such that the mean squared error of the prediction is minimized.

We first show an experimental result based on artificial data. We consider the base model $y(x) = 50\sin(x) + x^2$. From this model, we generated 101 points $\{x_i\}_{i=1}^{101}$ within $[-5, 5]$ at intervals of 0.1 and corresponding response values $y(x_i)$. Then, we added Gaussian noise with mean 0 and standard deviation 10 to $\{y(x_i)\}_{i=1}^{101}$ to construct the observed response values $\{y_i\}_{i=1}^{101}$ with noise, and thereby obtained a training dataset $\{(x_i, y_i)\}_{i=1}^{101}$. Using this data, we trained 20 rpart regression trees and predicted the response variable from their unweighted and weighted averages using optimized weights by minimizing the regularized average information content (33). The maximum depth of rpart regression trees is arbitrary, and we set it to 5 in this paper. The mean squared error of the predicted values and corresponding true values were 6.907 in the unweighted case and 6.764 in the weighted average case. The averaged standard deviations of the outputs of 20 regression trees were 7.919 in the unweighted case and 6.578 in the weighted average case. From this simple experimental result, weight optimization is shown to contribute to reducing the variance of the prediction. In figure 7, white circles denote observed data. The dotted line represents the base model (without noise), the blue dashed line represents the unweighted average of regressors, and the red solid line represents the weighted average of regressors. The rectangular region $[0, 3] \times [-15, 55]$ in figure 7 (a) is magnified and presented in figure 7 (b) with error bars at one standard deviation. We note that the standard deviation at each point x_i is estimated by optimized weights as

$$sd(x_i) = \sqrt{\sum_{j=1}^{20} w_j \left(c_j(x_i) - \sum_{l=1}^{20} w_l c_l(x_i) \right)^2}. \tag{34}$$

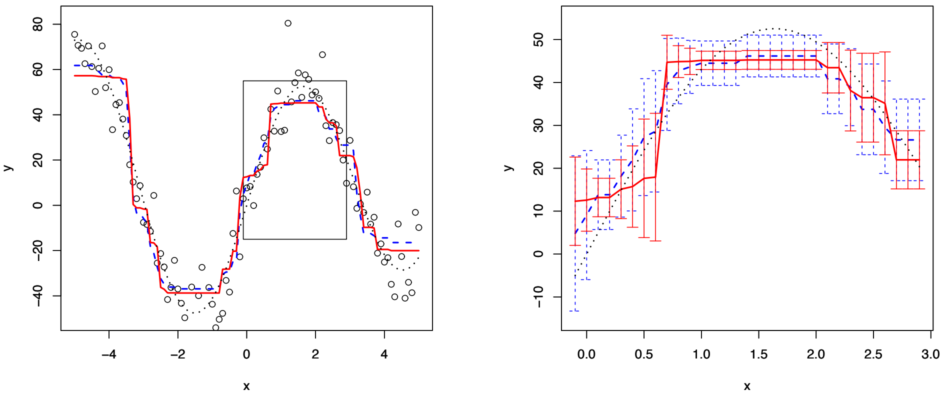


Figure 7: Unweighted and weighted average of tree regressors. Left: The observed data are plotted by white circles. The dotted line represents the base model (without noise), the dashed blue line represents the unweighted average of regressors, and the solid red line represents the weighted average of regressors. Right: The rectangle region in the left figure is magnified to show the effect of variance reduction.

Then, we show experimental results using other datasets. The Boston, imports85, airquality, Ozone, and abalone datasets are real world data from the UCI machine learning repository (Murphy & Aha, 1994); Friedman1 to Friedman3 are artificial data presented in Friedman (1991). Table 4 presents data specifications and the results of regression experiments. To estimate test errors, each experiment is repeated 100 times with random resampling except for the abalone dataset. For the abalone dataset, the data size is large; therefore, we repeated the regression experiment only 10 times. For each dataset, response variable and all explanatory variables are scaled to have zero mean and unit variance. We report mean squared errors (MSEs) of regressions, and standard deviations of regressors. We calculate standard deviations of output values of 20 trees, and averaged over all test points as:

$$sd = \frac{1}{\#\{\text{test samples}\}} \sum_{i \in \text{test samples}} sd(x_i). \tag{35}$$

Intuitively speaking, this quantity sd corresponds to the averaged width of confidence intervals, and a smaller sd implies a reliable prediction. Since we have 100 repetitions of random data sampling, we have 100 different values of sds. The average and standard deviation of sds over 100 repetitions are shown in table 4 for each dataset. From this table, we see that for most datasets, there are no significant differences in MSE between Bagging alone and Bagging with optimized weights, whereas the standard deviations of regressors, that is, the variances of the empirical distributions of the response variables based on regressors, are reduced by weight optimization. This leads us to a more reliable prediction via the ensemble of regressors.

Table 4: Data specifications and results of regression using Bagging and Bagging with optimized weights. MSEs of prediction are reported with one standard deviations, and averaged standard deviations of regressors are reported with one standard deviation. The better results with significance level of 5% in t-test are shown in boldface.

	Specifications			Results			
	dim	# train	# test	Bagging		Bagging+weight	
Data name				MSE	SD of $\{c_j(x)\}$	MSE	SD of $\{c_j(x)\}$
Boston	13	506	51	0.422(0.105)	0.246(0.171)	0.421(0.105)	**0.239**(0.166)
imports85	25	193	20	0.327(0.073)	0.178(0.128)	0.327(0.073)	0.172(0.122)
airquality	5	111	11	0.585(0.289)	0.311(0.198)	0.585(0.289)	0.299(0.191)
Ozone	12	330	33	0.542(0.089)	0.333(0.175)	0.543(0.089)	0.325(0.170)
abalone	8	4177	1045	0.705(0.022)	0.229(0.142)	0.705(0.022)	0.222(0.139)
Friedman1	10	200	2000	0.546(0.022)	0.459(0.146)	0.547(0.022)	**0.445**(0.142)
Friedman2	4	200	2000	0.386(0.015)	0.278(0.133)	0.386(0.014)	**0.269**(0.128)
Friedman3	4	200	2000	0.567(0.032)	0.335(0.253)	0.567(0.033)	0.325(0.246)

6. Conclusion

In this paper, we proposed estimators of the Shannon information content with weighted observations, and extended those estimators to allow estimation of the cross entropy and entropy. In the kernel density estimation framework, weight for each datum is easily taken into account. However, to the authors' knowledge, this is the first attempt to introduce the weight to the quantile or nearest neighbor based estimators of the information content or entropy. The proposed quantile-based estimators are similar to well known k-NN methods. However, the proposed methods are based on the notion of the quantile, and can naturally take weights into account.

There are many practical circumstances in which data are associated with weights. For example, in the context of selection bias correction (see, e.g., Heckman, 1979), empirical likelihood method (Owen, 2001) is applied to correct the selection bias caused by unobserved response variables (Qin et al., 2002). Let y and x be response and explanatory variables, respectively. Consider we have complete observation $\{(y_i, x_i)\}_{i=1}^{n_c}$ of size n_c and incomplete observation $\{x_j\}_{j=n_c+1}^{n}$ of size $n - n_c$. In this situation, the estimate $(1/n_c)\sum_{i=1}^{n_c} y_i$ of the average of response variable $E_{p_Y}[Y]$ is known to be inconsistent when the event whether y is observed or not depends on both the value of y and x. We can introduce weights w_i for y_i, defined using both $\{(y_i, x_i)\}_{i=1}^{n_c}$ and $\{x_j\}_{j=n_c+1}^{n}$, to obtain a consistent estimator $\sum_{i=1}^{n_c} w_i y_i$. The analysis of incomplete data is an important problem not only in engineering but in economics, politics, pedagogics, sociology, and medical science, for example, and the proposed estimators in this paper would be useful to estimate information theoretic quantities of weighted data in these fields of researches. Weighted data may also arise in collaborative filtering systems (Adomavicius & Tuzhilin, 2005). Collaborative filtering systems attempt to present items that are likely of interest to a target user automatically. For this purpose, the rating of items by the target user is estimated by ratings already done by other users to the item. These ratings by other users could be weighted by the influence or reliability of each user. For example, a rating given by a user could be weighted proportional to the number of ratings by the user. As another example, in time series analysis, it is natural to assign weights to past sequences so as to reduce the influence of past events; smaller weights are assigned to sequences that are farther away from the current time of interest. Our ongoing study on change point detection in time series is based on the proposed information estimators.

By optimizing weights in estimation of the KL divergence, we can compress data, preserving the original distribution. Furthermore, based on the proposed information estimators, we proposed a method of constructing weighted ensemble regressors and showed that the proposed method can reduce the variance of the regression. The accuracy of a classification by weighted LVQ is improved as a whole. The rate of improvement is limited and further efforts to improve the accuracy by weight optimization will be made. We will also consider and analyse the effect of weights for data compression preserving the original distribution. To reduce the computational load of query processing in database systems, it is important to compress the original data without degradation of neighborhood search accuracy. Variance of the ensemble regressor is reduced moderately. Further investigation into the application of weighted ensemble regressors is an important objective of future work. For example, we can apply the proposed method to the prediction of stock prices. According to the predictive distributions composed of weighted ensemble of regressors, we are developing strategies of buying and selling stocks.

Although we discussed some of the theoretical properties of the proposed quantile information estimator $I_\alpha(z; \mathfrak{D})$, further exploration into the theoretical aspects of other estimators such as $I_{\mathrm{MQ}}(z; \mathfrak{D})$ and $I_R(z; \mathfrak{D})$ is needed.

Acknowledgements

The authors would like to express our special thanks to the editor and reviewers whose valuable comments led to improvements of the manuscripts. Part of this work was supported by MEXT Kakenhi No.22800067 and No.25870811, and Grant-in-Aid for Challenging Exploratory Research No.23656072.

Appendix A. Proof of theorem 1

Let $x_i \in \mathcal{D}$ be the α-quantile point of an inspection point z. Consider a pdf $p_{\varepsilon_{\alpha,z}}(\varepsilon)$ of the distance $\varepsilon_{\alpha,z} = \|z - x_{\iota(\alpha)}\|$ between the inspection point z and its α-quantile in the data $\mathfrak{D}$ of size n. That is, $p_{\varepsilon_{\alpha,z}}(\varepsilon)\mathrm{d}\varepsilon$ represents the probability that, regarding to the distance ε from z, one and only one point $x_i \in \mathcal{D}$ falls in an interval $[\varepsilon, \varepsilon + \mathrm{d}\varepsilon]$, other $\iota(\alpha)_z - 1$ points fall in $[0, \varepsilon)$, and the remaining $n - \iota(\alpha)_z$ points fall in $(\varepsilon + \mathrm{d}\varepsilon, \infty)$. Then, following the construction of the entropy estimator in Kozachenko & Leonenko (1987), the probability $p_{\varepsilon_{\alpha,z}}(\varepsilon)\mathrm{d}\varepsilon$ is expressed as a trinomial distribution:

$$\begin{aligned} & p_{\varepsilon_{\alpha,z}}(\varepsilon)\mathrm{d}\varepsilon \\ & = \frac{n!}{1!(\iota(\alpha)-1)!(n-\iota(\alpha))!} \left(P_z(\varepsilon)\right)^{\iota(\alpha)-1}(1-P_z(\varepsilon))^{n-\iota(\alpha)}\mathrm{d}P_z(\varepsilon), \end{aligned} \tag{A.1}$$

where we again wrote $\iota(\alpha)_z$ as $\iota(\alpha)$ for the sake of notational simplicity. We note that $\iota(\alpha)$ is defined by $\arg\max_k \sum_{j=1}^{k} w_{(j)} \leq \alpha$, and Eq. (A.1) holds whenever $\iota(\alpha) \geq 1$ for a give weighted dataset $\mathfrak{D}$. We also note that under the assumption $\frac{\iota(\alpha)}{n} \overset{n\to\infty}{\longrightarrow} L$, $\iota(\alpha)$ must increase in the same rate as n. In Eq. (A.1), we consider the expression $P_z^{\iota(\alpha)-1}(1-P_z)^{n-\iota(\alpha)}$ where $P_z(\varepsilon)$ is replaced by P_z for the sake of notational simplicity. Using second-order Taylor series expansion of $\log P_z$ and $\log(1-P_z)$ around α and $1-\alpha$, we obtain

$$\begin{aligned} \log P_z &= \log(P_z - \alpha + \alpha) = \log \alpha \left(\frac{P_z - \alpha}{\alpha} + 1 \right) \\ &\sim \log \alpha + \frac{1}{\alpha}(P_z - \alpha) - \frac{1}{2\alpha^2}(P_z - \alpha)^2, \\ \log(1 - P_z) &= \log(1 - P_z + \alpha - \alpha) = \log(1-\alpha)\left(1 - \frac{P_z - \alpha}{1-\alpha}\right) \\ &\sim \log(1-\alpha) - \frac{1}{1-\alpha}(P_z - \alpha) - \frac{1}{2(1-\alpha)^2}(P_z - \alpha)^2. \end{aligned}$$

Then, logarithm of $P_z^{\iota(\alpha)-1}(1-P_z)^{n-\iota(\alpha)}$ is approximated as

$$(\iota(\alpha) - 1)\left(\frac{1}{\alpha}(P_z - \alpha) - \frac{1}{2\alpha^2}(P_z - \alpha)^2\right)$$

$$
\begin{aligned}
&+ (n - \iota(\alpha))\left(-\frac{P_z - \alpha}{1 - \alpha} - \frac{(P_z - \alpha)^2}{2(1 - \alpha)^2}\right) \\
&= -\frac{1}{2\sigma^2}(P_z - \alpha - \sigma^2\tau)^2 + \frac{\sigma^2\tau^2}{2} = -\frac{1}{2\sigma^2}(P_z - \beta)^2 + \frac{\alpha^2\tau^2}{2},
\end{aligned}
$$

where we ignored terms irrelevant to P_z and defined

$$\beta = \alpha + \sigma^2\tau, \quad \sigma^2 = \left(\frac{\iota(\alpha) - 1}{\alpha^2} + \frac{n - \iota(\alpha)}{(1 - \alpha)^2}\right)^{-1}, \quad \tau = \frac{\iota(\alpha) - 1}{\alpha} - \frac{n - \iota(\alpha)}{1 - \alpha}.$$

It is easy to see that $\lim_{n\to\infty} \sigma^2 = 0$ and

$$\lim_{n\to\infty} \sigma^2\tau \quad = \quad \lim_{n\to\infty} \frac{(\iota(\alpha) - 1)/\alpha - (n - \iota(\alpha))/(1 - \alpha)}{(\iota(\alpha) - 1)/\alpha^2 + (n - \iota(\alpha))/(1 - \alpha)^2} = \frac{\alpha(1 - \alpha)(L - \alpha)}{L - 2\alpha L + \alpha^2}.$$

If $L = \alpha$, for example, then $\beta = \alpha$ because $\sigma^2\tau = 0$ as $n \to \infty$. Then, the distribution (A.1) is approximated by

$$p_{\varepsilon_{\alpha,z}}(\varepsilon) = C(\beta)\phi(P_z(\varepsilon); \beta, \sigma^2)\frac{\mathrm{d}P_z(\varepsilon)}{\mathrm{d}\varepsilon} \tag{A.2}$$

where $C(\beta)$ is a normalizing factor. That is,

$$
\begin{aligned}
\int_0^\infty p_{\varepsilon_{\alpha,z}}(\varepsilon)\mathrm{d}\varepsilon &= C(\beta)\int_0^\infty \phi(P_z; \beta, \sigma^2)\frac{\mathrm{d}P_z}{\mathrm{d}\varepsilon}\mathrm{d}\varepsilon = C(\beta)\int_0^1 \phi(P_z; \beta, \sigma^2)\mathrm{d}P_z \\
&= C(\beta)\int_{-\beta}^{1-\beta} \phi(P_z; 0, \sigma^2)\mathrm{d}P_z = C(\beta)\int_{-\beta/\sigma}^{(1-\beta)/\sigma} \phi(P_z; 0, 1)\mathrm{d}P_z \\
&= C(\beta) \times \Big(\int_{-\infty}^\infty \phi(P_z; 0, 1)\mathrm{d}P_z \\
&\qquad - \int_{-\infty}^{-\beta/\sigma} \phi(P_z; 0, 1)\mathrm{d}P_z - \int_{-\infty}^{(\beta-1)/\sigma} \phi(P_z; 0, 1)\mathrm{d}P_z\Big) \\
&= C(\beta)\left(1 - \Phi(-\beta/\sigma) - \Phi((\beta - 1)/\sigma)\right)
\end{aligned}
$$

where $\Phi(x)$ is the cumulative distribution function of the standard Gaussian distribution. In the above equations, we used the fact that

$$
\begin{aligned}
&\int_{-\infty}^\infty \phi(x; 0, 1)\mathrm{d}x \\
&= \int_{-\infty}^{-\beta/\sigma} \phi(x; 0, 1)\mathrm{d}x + \int_{-\beta/\sigma}^{(1-\beta)/\sigma} \phi(x; 0, 1)\mathrm{d}x + \int_{(1-\beta)/\sigma}^{\infty} \phi(x; 0, 1)\mathrm{d}x \\
&= \int_{-\infty}^{-\beta/\sigma} \phi(x; 0, 1)\mathrm{d}x + \int_{-\beta/\sigma}^{(1-\beta)/\sigma} \phi(x; 0, 1)\mathrm{d}x + \int_{-\infty}^{(\beta-1)/\sigma} \phi(x; 0, 1)\mathrm{d}x.
\end{aligned}
$$

Then, we obtain

$$C(\beta) = \frac{1}{1 - \Phi(-\beta/\sigma) - \Phi((\beta - 1)/\sigma)}.$$

For example, when $\beta = 0.5$, $C(\beta) \simeq 1.002$ for $n = 10$, and $C(\beta) \simeq 1 + 1 \times 10^{-23}$ for $n = 100$. Hence, in this paper, we always treat $C(\beta)$ as unity.

Noting that $\phi(P_z; \beta, \sigma^2) \overset{n\to\infty}{\longrightarrow} \delta(P_z - \alpha)$ where $\delta(\cdot)$ is the Dirac's delta function, the expectation of ε under the pdf $p_{\varepsilon_{\alpha,z}}(\varepsilon)$ is calculated as

$$\begin{aligned} &\int \varepsilon \phi(P_z(\varepsilon); \beta, \sigma^2) \frac{\mathrm{d}P_z(\varepsilon)}{\mathrm{d}\varepsilon} \mathrm{d}\varepsilon = \int \varepsilon \phi(P_z(\varepsilon); \beta, \sigma^2) \mathrm{d}P_z(\varepsilon) \\ &\overset{n\to\infty}{\longrightarrow} \int \varepsilon \delta(P_z(\varepsilon) - \alpha) \mathrm{d}P_z(\varepsilon) = P_z^{-1}(\alpha) = \varepsilon_\alpha, \end{aligned}$$

where $P_z^{-1}(\cdot)$ is the inverse function of the probability mass (6). From Eq. (A.2), where the right hand side of Eq. (A.2) includes the pdf of the Gaussian distribution with $\sigma^2 \overset{n\to\infty}{\to} 0$, we see that the variance of $p_{\varepsilon_{\alpha,z}}(\varepsilon)$ goes to zero as $n \to \infty$ and this proves the theorem 1 ■.

Intuitively speaking, the result implies that the empirical cumulative density function tends to the true one when the sample size tends to infinity from Glivenko-Cantelli theorem: (Vapnik, 1998). Theorem 1 states this fact in terms of pdf with explicit dependency on weights for the given data.

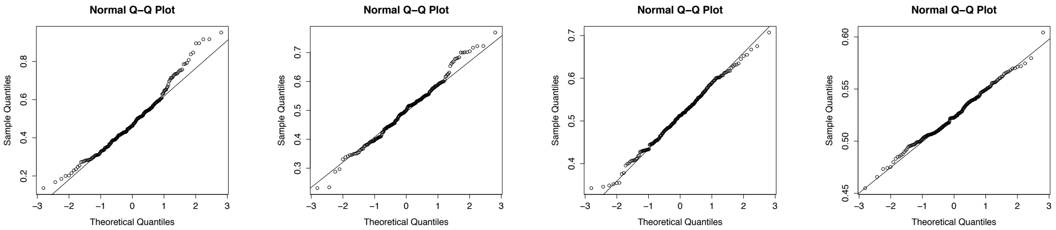
To grasp the characteristics of Eq. (A.2), we show two simple experimental results on variance convergence property of $p_{\varepsilon_{\alpha,z}}(\varepsilon)$.

Firstly, we check the normality of $p_{\varepsilon_{\alpha,z}}(\varepsilon)$. We generate 200 datasets $\mathcal{D} = \{x_i\}_{i=1}^n$ from the ground truth distribution $\phi(x; 0, 1)$. The size of datasets n are set to $n = 20, 50, 100$, and 1000. The weight w_i for each datum x_i is determined as follows. From a set $\{1, 2, \ldots, \lfloor n/3 \rfloor\}$, we sampled $\{\hat{w}_i\}_{i=1}^n$ at random with replacement, and defined the weight for x_i by $w_i = \hat{w}_i / \sum_{i=1}^n \hat{w}_i$. This weighting scheme may result in relatively large w_i, but in our experiment, the order of the largest weight is $1/n$. Then, we calculate $\varepsilon_{\hat{\alpha}}$, where $\alpha = 0.4$ and the inspection point is set to zero. In table A.5, we show p-values of the Kolmogorov-Smirnov test and the Shapiro-Wilk test on sample distributions of $\varepsilon_{\hat{\alpha}}$ for each n. We also show Q-Q plots on sample distributions for each n in figure A.8. From the table and figures, we see that the sample distribution of $\varepsilon_{\hat{\alpha}}$ get closer to a Gaussian distribution as the data size n increases.

Secondly, from the ground truth distribution $\phi(x; 0, 1)$, we generate datasets 1000 times for $n = 20, 40, \ldots, 1000$ (increasing by 20). Then, we calculate $\varepsilon_{\hat{\alpha}}$, with the quantile $\alpha = 0.4$ and the inspection point $x = 0$. Because

Table A.5: The p-values of the Kolmogorov-Smirnov test and Shapiro-Wilk test on the sample distributions of distance $\varepsilon_{\hat{\alpha}}$ between $\hat{\alpha}$-quantile point to the inspection point $x = 0$, when $n = 20, 50, 100, 1000$. For each setting of n, 200 datasets are generated.

	n=20	n=50	n=100	n=1000
Kolmogorov-Smirnov	1.349e-08	0.007355	0.08062	0.1541
Shapiro-Wilk	0.01254	0.3122	0.2478	0.9997



(a) $n = 20$ (b) $n = 50$ (c)$n = 100$ (d) $n = 1000$

Figure A.8: Q-Q plots of sample distributions of $\varepsilon_{\hat{\alpha}}$ with 200 trials. (a): $n = 20$. (b): $n = 50$. (c): $n = 100$. (d): $n = 1000$.

$f(x) = \phi(x; 0, 1)$ is known, we can write the probability mass function explicitly as

$$\begin{aligned} P_{z=0}(\varepsilon) &= \int_{b(0,\varepsilon)} \phi(x; 0, 1)\mathrm{d}x = 2\int_0^{\varepsilon} \phi(x; 0, 1)\mathrm{d}x \\ &= 2\left(\Phi(\varepsilon) - \Phi(0)\right) = 2\Phi(\varepsilon) - 1, \end{aligned}$$

and the expectation value of ε is calculated using the inverse function of $2\Phi(\varepsilon) - 1$. In figure A.9 (left), we show the mean absolute values of $|\varepsilon_\alpha - \varepsilon_{\hat{\alpha}}|$ and their one standard deviations with 1000 trials for various n. From this figure, we see that the average of $\varepsilon_{\hat{\alpha}}$ converges to the theoretical value ε_α. In figure A.9 (right), we show the number of samples and empirical standard deviations of $\varepsilon_{\hat{\alpha}}$ in a log-log plot. From this figure, we can see that the standard deviation decays in $\sqrt{n}$ order.

Appendix B. proof of theorem 2

The proof of theorem 2 is based on the technique used in Kozachenko & Leonenko (1987) and Goria et al. (2005) to derive the bias of the k-NN based entropy estimator. We introduce a lemma regarding the limit of probability mass in shrinking open balls.

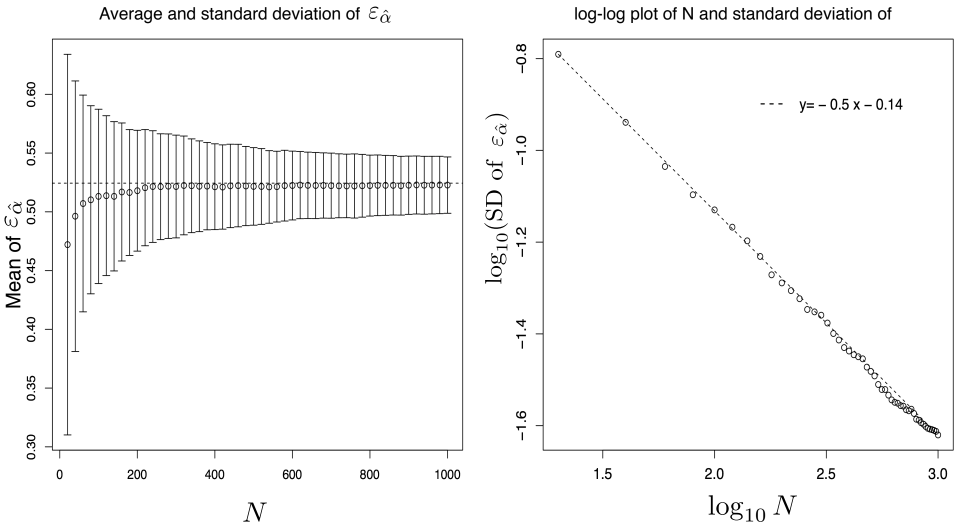


Figure A.9: Left: Average and one standard deviation of $\varepsilon_{\hat{\alpha}}$ with various data size ($n = 20, 40, 60, \ldots, 1000$). The theoretical value of ε_{α} is shown by a dashed line. Right: Standard deviation of $\varepsilon_{\hat{\alpha}}$ and the number of samples in a log-log plot. A fitted line with tangent -0.5 is depicted in dashed line.

Lemma 1 (Lebesgue's Differentiation Theorem). *If $g(x)$ is an absolutely integrable function on $\mathbb{R}$, then for any sequences of open balls $b(x, r_k)$ of radius $r_k \overset{k\to\infty}{\longrightarrow} 0$ and for almost all $x \in \mathbb{R}^d$,*

$$\lim_{k\to\infty} \frac{1}{|b(x, r_k)|} \int_{b(x,r_k)} g(y)\mathrm{d}y = g(x). \tag{B.1}$$

Proof 1. *See (Stein & Shakarchi, 2005), for example.*

Let αn be fixed arbitrarily, say, $M \in \mathbb{R}^+$. Let $r_n(u) = (u\alpha e^{\psi(M)}/(Mc_d))^{1/d}$, $u \in \mathbb{R}$ be an element of decreasing series of open ball's radii. Then, the volume of the ball is $|b(x, r_n(u))| = c_d(u\alpha e^{\psi(M)}/(Mc_d)) = u(e^{\psi(M)}/M)\alpha \overset{n\to\infty}{\longrightarrow} 0$ because α goes to zero when n goes to infinity. Now, we consider a random variable $\xi_{n,\alpha,z} = e^{I_\alpha(z;\mathfrak{D}_n)}$. The distribution function of $\xi_{n,\alpha,z}$ is

$$F_{n,\alpha,z}(u) = \Pr(e^{I_\alpha(z;\mathfrak{D}_n)} < u) = 1 - \sum_{m=0}^{\iota(\alpha)-1} q_m(n, u), \tag{B.2}$$

where

$$q_m(n, u) = \binom{n}{m} (P_z(r_n(u)))^m (1 - P_z(r_n(u)))^{n-m} \tag{B.3}$$

is a probability that m out of n points lay within a ball of radius $r_n(u)$ centered at z. We note that if $w_i = 1/n$, then $M \simeq \iota(\alpha)$. However, the equality does not hold in general because of the weights. For notational simplicity, we write $b(z, r_n(u))$ as b. By lemma 1 and noting that $b(z, r_n(u)) = ue^{\psi(M)}/n$, we get

$$
\begin{aligned}
\lim_{n\to\infty} q_m(n,u) = & \lim_{n\to\infty} \frac{n!}{m!(n-m)!} |b(z,r_n(u))|^m \left(\frac{1}{|b(z,r_n(u))|} \int_{b(z,r_n(u))} f(x)\mathrm{d}x \right)^m \\
& \times \left(1 - |b(z,r_n(u))| \frac{1}{|b(z,r_n(u))|} \int_{b(z,r_n(u))} f(x)\mathrm{d} \right)^{n-m} \\
= & \lim_{n\to\infty} \frac{1}{m!} n \cdot (n-1) \cdots (n-m+1) |b|^m \\
& \times \left(\frac{1}{|b|} \int_b f(x)\mathrm{d}x \right)^m \left(1 - u\frac{e^{\psi(M)}}{n} \left(\frac{1}{|b|} \int_b f(x)\mathrm{d}x \right) \right)^{n-m} \\
= & \lim_{n\to\infty} \frac{1}{m!} (un\frac{e^{\psi(M)}}{n})(un\frac{e^{\psi(M)}}{n} - u\frac{e^{\psi(M)}}{n}) \\
& \cdots (un\frac{e^{\psi(M)}}{n} - um\frac{e^{\psi(M)}}{n} + u\frac{e^{\psi(M)}}{n}) \\
& \times \left(\frac{1}{|b|} \int_b f(x)\mathrm{d}x \right)^m \left(1 - u\frac{e^{\psi(M)}}{n} \left(\frac{1}{|b|} \int_b f(x)\mathrm{d}x \right) \right)^{n-m} \\
= & \frac{(f(z)e^{\psi(M)}u)^m}{m!} e^{-f(z)e^{\psi(M)}u}, \qquad \text{(B.4)}
\end{aligned}
$$

and

$$
\lim_{n\to\infty} F_{n,\alpha,z}(u) = F_z(u) = 1 - \sum_{m=1}^{\iota(\alpha)-1} \frac{(f(z)e^{\psi(M)}u)^m}{m!} e^{-f(z)e^{\psi(M)}u}.
$$

To derive Eq. (B.4), we used

$$
(ue^{\psi(M)})(ue^{\psi(M)} - u\frac{1}{n}e^{\psi(M)}) \cdots (ue^{\psi(M)} - um\frac{1}{n}e^{\psi(M)} + u\frac{1}{n}e^{\psi(M)}) \overset{n\to\infty}{\longrightarrow} (ue^{\psi(M)})^m,
$$

and

$$
(1-\zeta/n)^{-m}(1-\zeta/n)^n \overset{n\to\infty}{\longrightarrow} 1 \cdot e^{-\zeta}, \quad \zeta = f(z)e^{\psi(M)}u
$$

combined with the lemma 1, which gives $\frac{1}{|b|}\int_b f(x)\mathrm{d}x \overset{n\to\infty}{\longrightarrow} f(z)$.

We next consider a random variable ξ_z with distribution function $F_z(u)$. The pdf of ξ_z is given by the derivative of $F_z(u)$ as

$$
f_z(u) \;=\; \frac{\mathrm{d}}{\mathrm{d}u} \left(1 - \sum_{m=1}^{\iota(\alpha)-1} \frac{(f(z)e^{\psi(M)}u)^m}{m!} e^{-f(z)e^{\psi(M)}u} \right)
$$

$$
\begin{aligned}
&= f(z)e^{\psi(M)}e^{-f(z)e^{\psi(M)}u}\sum_{m=0}^{\iota(\alpha)-1}\frac{(f(z)e^{\psi(M)})^m}{m!}u^m \\
&\quad -f(z)e^{\psi(M)}e^{-f(z)e^{\psi(M)}u}\sum_{m=1}^{\iota(\alpha)-1}\frac{(f(z)e^{\psi(M)})^{m-1}}{(m-1)!}u^{m-1} \\
&= f(z)e^{\psi(M)}e^{-f(z)e^{\psi(M)}u}\frac{(f(z)e^{\psi(M)})^{\iota(\alpha)-1}}{(\iota(\alpha)-1)!}u^{\iota(\alpha)-1} \\
&= \frac{f(z)e^{\psi(M)}}{(\iota(\alpha)-1)!}e^{-f(z)e^{\psi(M)}u}(f(z)e^{\psi(M)}u)^{\iota(\alpha)-1}.
\end{aligned}
$$

Finally, we get

$$
\begin{aligned}
&\lim_{n\to\infty} E[I_\alpha(z;\mathfrak{D}_n)] \\
&= E_{f_z}[\log\xi_z] = \int_0^\infty f_z(u)\log u \mathrm{d}u \\
&= \frac{f(z)e^{\psi(M)}}{(\iota(\alpha)-1)!}\int_0^\infty (\log u)e^{-f(z)e^{\psi(M)}u}(f(z)e^{\psi(M)}u)^{\iota(\alpha)-1}\mathrm{d}u \\
&= \frac{1}{(\iota(\alpha)-1)!}\int_0^\infty \log\left(\frac{t}{f(z)e^{\psi(M)}}\right)e^{-t}t^{\iota(\alpha)-1}\mathrm{d}t \\
&= \frac{1}{(\iota(\alpha)-1)!}\left\{\int_0^\infty t^{\iota(\alpha)-1}e^{-t}\log t\mathrm{d}t - (\log f(z)e^{\psi(M)})\int_0^\infty t^{\iota(\alpha)-1}e^{-t}\mathrm{d}t\right\} \\
&= \frac{1}{(\iota(\alpha)-1)!}\left\{\Gamma'(\iota(\alpha)) - (\log f(z)e^{\psi(M)})\Gamma(\iota(\alpha))\right\} \\
&= -\log f(z) - \log e^{\psi(M)} + \frac{\Gamma'(\iota(\alpha))}{(\iota(\alpha)-1)!} = -\log f(z) + \psi(\iota(\alpha)) - \psi(M),
\end{aligned}
$$

which proves theorem 2 ■.

As mentioned in section 1, one of the simplest generative mechanisms of weighted dataset (4) is that $\mathfrak{D}$ is a set of realization from a certain joint distribution $\tilde{p}(x,w)$. In this case, Eq. (10) evaluates the gap between $I_f(z)$ and expectation of $I_\alpha(z;\mathfrak{D}_n)$ with respect to $\tilde{p}(x,w)$. If we assume $\tilde{p}(x,w) = f(x)h(w)$, where f is an unknown data distribution and h is a weight distribution, further analysis regarding the dispersion of the term $\psi(\iota(\alpha))$ is possible. We show a simple example bellow:

Example 1 (Dirichlet Sampling). *When the set of weights is a realization of a Dirichlet distribution*

$$
h(w;\gamma) = \frac{1}{Z(\gamma)}\prod_{j=1}^{n} w_j^{\gamma-1}, \tag{B.5}
$$

where $w_j \geq 0$, $\sum_{j=1}^{n} w_j = 1$, $\gamma > 0$ and the normalization factor $Z(\gamma)$ is defined by $Z(\gamma) = \frac{\prod_{j=1}^{n}\Gamma(\gamma)}{\Gamma(n\gamma)}$. Then, mean and variance of $\iota(\alpha)$ are αn and $\alpha(1-\alpha)/\gamma$, respectively.

In statistics, it is important to investigate how the bias behaves. Although it is difficult in general, theoretical estimation of the bias of the QIE is possible under a rather unrealistic assumption $\tilde{p}(x, w) = f(x)h(w)$. In this case, we can estimate the fluctuation of the bias of the QIE under the assumption that weights are realizations from a Dirichlet distribution.

We show a rough sketch of the proof of the assertion in Example 1. Assuming that n is large, we identify $\hat{\alpha}$ and α, and identify $\lfloor n\gamma \rfloor$ and $n\gamma$. For simplicity, we denote $\iota(\alpha)$ by ι, and define $w_{(0)} = 0$. It is known that a partial sum of components of a Dirichlet distribution is a beta random variable (Haas & Formery, 2002). That is, let $\xi = \iota\gamma \in \{0, \gamma, \ldots, n\gamma\} = \Lambda$ be fixed, the distribution of $\alpha = \sum_{i=0}^{\iota} w_{(i)}$ is

$$f(\alpha|\xi) = \mathrm{Beta}(\xi + 1, n\gamma - \xi + 1)^{-1}\alpha^{\xi}(1-\alpha)^{n\gamma-\xi}. \tag{B.6}$$

For notational simplicity, we consider ξ instead of ι henceforth. We assume uniform prior for $\alpha \in [0, 1]$. Accordingly, it is natural to define the prior distribution of ξ as a uniform distribution in Λ. Then, from the Bayes' theorem, we obtain

$$\begin{aligned} f(\xi|\alpha) &= f(\alpha|\xi)f(\xi)/f(\alpha) = f(\alpha|\xi)\frac{1}{n+1} \\ &= \mathrm{Beta}(\xi + 1, n\gamma - \xi + 1)^{-1}\alpha^{\xi}(1-\alpha)^{n\gamma-\xi}\frac{1}{n+1}. \end{aligned}$$

Considering the generalized binomial distribution

$$1 = \{\alpha + (1-\alpha)\}^{n\gamma} \simeq \sum_{k\in\Lambda} \frac{\gamma\Gamma(n\gamma+1)}{\Gamma(k+1)\Gamma(n\gamma-k+1)}\alpha^{k}(1-\alpha)^{n\gamma-k},$$

we see that the summation of

$$f(\xi|\alpha) = \frac{1}{n\gamma+1}\frac{\alpha^{\xi}(1-\alpha)^{n\gamma-\xi}}{\mathrm{Beta}(\xi+1, n\gamma-\xi+1)} \simeq \frac{\gamma\Gamma(n\gamma+1)\alpha^{\xi}(1-\alpha)^{n\gamma-\xi}}{\Gamma(\xi+1)\Gamma(n\gamma-\xi+1)}$$

for all ξ becomes 1. As is the case with the binomial distribution, the mean and variance of ξ are $\alpha n\gamma$ and $\alpha(1-\alpha)n\gamma$, respectively. Since $\xi = \iota\gamma$, the mean and variance of ι are shown to be αn, $\alpha(1-\alpha)n/\gamma$.

The weights $\{w_i\}_{i=1}^{n}$ are supposed to be realizations of a Dirichlet distribution (B.5), where the variance of w_i is given by $(n-1)/(n^2(1+n\gamma))$. This shows that when γ is large, the variance of each weight w_i is small. So, it is natural that the variance of ι is proportional to the reciprocal of γ.

References

G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol.17(6), pp. 734–749, 2005.

E. J. Atkinson, T. M. Therneau, "An introduction to recursive partitioning using the rpart routines," Technical report, Mayo Clinic, 61, 1997.

J. Beirlant, E. J. Dudewicz, L. Györfi, E. C. Meulen, "Nonparametric entropy estimation: An overview," *International Journal of the Mathematical Statistics Sciences*, vol. 6, pp. 17–39, 1997.

L. Breiman, "Bagging predictors," *Machine Learning*, vol.24 (2), pp. 123–140, 1996.

P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36 (3), pp. 287–314, 1994.

R. D. Cook, C. J. Nachtsheim, "Reweighting to achieve elliptically contoured covariates in regression," *Journal of American Statistical Association*, vol. 89, pp. 592–599, 1994.

T. M. Cover, J. A. Thomas, *Elements of information theory.* John Wiley and Sons, Inc, 1991.

R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification.* Wiley-Interscience Publication, 2000.

L. Faivishevsky, J. Goldberger, "ICA based on a smooth estimation of the differential entropy," In: *Proceedings of Advances in Neural Information Processing Systems 21*, pp. 433–440, 2009.

J. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics*, vol. 19 (1), pp. 1–67, 1991.

M. N. Goria, N. N. Leonenko, V. V. Mergel, P. L. N. Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Journal of Nonparametric Statistics*, vol. 17 (3), pp. 277–297, 2005.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, "A kernel method for the two sample problem," In: *Proceedings of Advances in Neural Information Processing Systems 19*, pp. 513–520, 2007.

L. Györfi, E. C. van der Meulen, "Density-free convergence properties of various estimators of entropy," *Comput. Stat. Data Anal.*, vol. 5 (4), pp. 425–436, 1987.

A. Haas, P. Formery, "Uncertainties in facies proportion estimation I. theoretical framework: The Dirichlet distribution," *Mathematical Geology*, vol. 34, pp. 679–702, 2001.

P. Hall, D. M. Titterington, "On smoothing sparse multinomial data," *Australian Journal of Statistics*, vol. 29, pp. 19–37, 1989.

T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning — Data Mining, Inference, and Prediction, 1st Edition.* Springer, 2001.

J. Heckman, "Sample Selection Bias as a Specification Error", *Econometrica*, vol. 47(1), pp.153–161, 1979.

H. Hino, N. Murata, "A conditional entropy minimization criterion for dimensionality reduction and multiple kernel learning," *Neural Comput.*, vol. 22(11), pp. 2887–2923, 2010.

H. Hino, N. Murata, "A computationally efficient information estimator for weighted data," In: *International conference on neural networks*, pp. 301–308, 2011.

A. Hyvärinen, J. Karhunen, E. Oja, *Independent Component Analysis.* J. Wiley, New York, 2001.

A. Jain, M. Murty, P. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31 (3), pp. 264–323, 1999.

R. Koenker, *Quantile Regression.* Cambridge University Press, 2005.

R. Koenker, G. Bassett, "Regression Quantiles," *Econometrica,* vol. 46, pp. 33–50, 1978.

T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, K. Torkkola, "LVQ pak:the learning vector quantization program package," Technical report, Helsinki University of Technology, Laboratory of Computer and Information Science, A30, 1996.

L. F. Kozachenko, N. N. Leonenko, "Sample estimate of entropy of a random vector," *Problems of Information Transmission,* vol. 23, pp. 95–101, 1987.

A. Kraskov, H. Stögbauer, P. Grassberger, "Estimating mutual information," *Physical review. E, Statistical, nonlinear, and soft matter physics,* vol. 69 (6 Pt 2), 2004.

S. Kullback, R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics,* vol. 22 (1), pp. 79–86 1951.

E. G. Learned-Miller, J. W. Fisher III, "ICA using spacings estimates of entropy," *J. Mach. Learn. Res.,* vol. 4 (7-8), pp. 1271–1295, 2004.

J. M. Leiva-Murillo, A. Artes-Rodriguez, "A Gaussian mixture based maximization of mutual information for supervised feature extraction," In: *Proceedings of International Conference on Independent Component Analysis,* pp. 271–278, 2004.

R. Linsker, "Towards an Organizing Principle for a Layered Perceptual Network," In: *Proceedings of Neural Information Processing Systems 1,* pp. 485–494, 1987.

R. Linsker, "Improved local learning rule for information maximization and related applications," *Neural Networks*, vol. 18(3), pp. 261–265 2005.

D. O. Loftsgaarden, C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *The Annals of Mathematical Statistics,* vol. 36 (3), pp. 1049–1051, 1965.

S. Mannor, D. Peleg, R. Y. Rubinstein, "The cross entropy method for classification," In: *Proceedings of International Conference on Machine Learning,* pp. 561–568, 2005.

R. M. Mnatsakanov, N. Misra, S. Li, E. J. Harner, "K-nearest neighbor estimators of entropy," *Mathematical Methods of Statistics*, vol. 17 (3), pp. 261–277, 2008.

P. Murphy, D. Aha, UCI Repository of machine learning databases. Tech. rep., University of California, Department of Information and Computer Science, Irvine, CA, US, 1994.

J. Nocedal, S. Wright, *Numerical Optimization (2nd ed.).* Springer, 2006.

A. B. Owen, *Empirical Likelihood*, Chapman & Hall, 2001.

L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.,* vol. 15, pp. 1191–1253, 2003.

J. Peltonen, J. Goldberger, S. Kaski, "Fast semi-supervised discriminative component analysis," In: *Proceedings of the International conference on Machine Learning for Signal Processing,* pp. 312–317, 2007.

F. Pérez-Cruz, "Estimation of information theoretic measures for continuous random variables," In: *Proceedings of Neural Information Processing Systems 20,* pp. 1257–1264, 2008.

S. Portnoy, "Censored Quantile Regression," *Journal of the American Statistical Association,* vol. 98(464), pp. 1001–1012, 2003.

J. Qin, D. Leung, J. Shao, "Estimation With Survey Data Under Nonignorable Nonresponse or Informative Sampling," *Journal of the American Statistical Association*, vol. 97, pp. 193–200, 2002.

R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2010.

B. Rajagopalan, U. Lall, "A kernel estimator for discrete distributions," *Journal of Nonparametric Statistics*, vol. 4, pp. 409–426, 1995.

J. S. Ramberg, B. W. Schmeiser, "An approximate method for generating asymmetric random variables," *Commun. ACM,* vol. 17, pp. 78–82, 1974.

G. Rätsch, T. Onoda, K.-R. Müller, "Soft margins for adaboost," *Machine Learning,* vol. 42 (3), pp. 287–320, 2001.

A. J. Bell, and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation,* vol. 7 (6), pp. 1129–1159, 1995.

C. E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal,* vol. 27, pp. 379–423,623–656, 1948.

L. Song, X. Zhang, A. Smola, A. Gretton, B. Schölkopf, "Tailoring density estimation via reproducing kernel moment matching," In: *Proceedings of the 25th international conference on Machine learning*, pp. 992–999, 2008.

E. M. Stein, R. Shakarchi, *Real Analysis.* Princeton Lectures in Analysis, III. Princeton University Press. Princeton University Press, 2005.

I. Takeuchi, Q. V. Le, T. D. Sears, A. J. Smola, "Nonparametric Quantile Estimation," *Journal of Machine Learning Research,* vol. 7, pp.1231-1264, 2006.

V. N. Vapnik, *Statistical Learning Theory.* Wiley-Interscience, 1998.

M. P. Wand, M. C. Jones, *Kernel Smoothing.* Chapman & Hall/CRC, 1994.

Q. Wang, S. R. Kulkarni, S. Verdú, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Trans. Inf. Theor.,* vol. 55, pp. 2392–2405, 2009.

H. J. Wang, L. Wang, "Locally Weighted Censored Quantile Regression," *Journal of the American Statistical Association,* vol. 104(487), pp. 1117–1128, 2009.