Boundary effect correction in k-nearest-neighbor estimation

A. R. Alizad Rahvar* and M. Ardakani†

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G2V4

(Received 4 January 2011; revised manuscript received 28 March 2011; published 19 May 2011)

The problem of the boundary effect for the k-nearest-neighbor (kNN) estimation is addressed, and a correction method is suggested. The correction is proposed for bounded distributions, but it can be used for any set of bounded samples. We apply the proposed correction to entropy estimation of multidimensional distributions and time series, and this correction reduces considerably the bias and statistical errors in the estimation. For a small sample size or high-dimensional data, the corrected estimator outperforms the uncorrected estimator significantly. This advantage makes the kNN method applicable to more real-life situations, e.g., the analysis of biological and molecular data.

DOI: 10.1103/PhysRevE.83.051121 PACS number(s): 02.50.−r, 89.70.Cf, 05.10.−a, 05.45.Tp

I. INTRODUCTION

The k-nearest-neighbor (kNN) method is a nonparametric technique with successful performance in estimation [1–3], classification and pattern recognition [4,5], and data mining [6]. The basic advantage of the kNN method is its simplicity.

Nonparametric entropy estimation is an important application of the kNN method, which has received much attention in diverse study areas such as neuroscience [7], molecular sciences [8,9], and genetics [3]. The work of Kozachenko and Leonenko [2] presents the kNN entropy estimator according to the first-neighbor distance ($k = 1$) and proves its mean-square consistency for multidimensional data. Asymptotically unbiased and consistent estimators for $k \geqslant 1$ are proposed in [3,10,11]. Moreover, the kNN estimators of the mutual information and divergence are presented in [3] and [12], respectively.

Despite the large body of literature on kNN estimation, most studies neglect the boundary effect of bounded distributions as it is negligible asymptotically. However, in some practical setups, we may have a small sample size. Thus, the results will be influenced significantly by the boundary effect. For example, considering the boundary effect, the moments of the kNN distance distribution are obtained analytically in [13]. As this study shows, for a three-dimensional uniform distribution, the moments are dramatically more accurate than when the boundary effect is ignored.

The main focus of this paper is to consider the boundary effect in kNN estimation. Indeed, we propose a correction method for kNN estimators in which the boundary effect is eliminated. In particular, this correction is investigated for the kNN entropy estimator. Our approach, however, can be used for other estimation tasks, e.g., estimation of density and mutual information. As the simulations show, the proposed correction method completely eradicates the adverse effect of the boundary for the uniform distribution. For nonuniform distributions, it seems that the boundary effect is eliminated and the remaining errors have other causes, such as nonuniformity. Furthermore, we see a remarkable increase in the accuracy of the entropy estimation for cases with a small sample size and high dimensional data. Hence, by this correction, the kNN estimator can be applied successfully to more practical problems.

The rest of the paper is organized as follows. Section II is devoted to the required background and includes an introduction to the kNN entropy estimator. The proposed correction method is presented in Sec. III. In Sec. IV, we apply the proposed correction method to kNN entropy estimation, and we provide simulation results. Finally, conclusions are drawn in Sec. V.

II. BACKGROUND AND MOTIVATION

A. kNN entropy estimation method

Let $X_1, \ldots, X_N$ be a sequence of N independent and identically distributed sample vectors $X_i = (x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^d$ from an unknown d-dimensional probability density function $f(X)$. Here, $x_{i,m} (1 \leqslant m \leqslant d)$ represents the mth element of the vector X_i. Various norms, e.g., the Euclidean norm or maximum norm, can be used to measure the distance between X_i and X_j. In this paper, the maximum norm is used, which is defined as follows:

$$\|X_i - X_j\| = \max_{1 \leqslant m \leqslant d} \{|x_{i,m} - x_{j,m}|\}. \tag{1}$$

For $1 \leqslant k \leqslant N-1$, let $X_{i(k)}$ represent the kth nearest neighbor of X_i. For notational consistency, let us denote the sample point X_i by $X_{i(0)}$. We also define $\rho_{i,k}$ as twice the distance between X_i and $X_{i(k)}$. The set of vectors whose distance from X_i is a given constant, c, forms the surface of a hypercube centered at X_i with edge length $2c$. In particular, let $B_{i,k}$ represent a hypercube of edge length $\rho_{i,k}$ centered at X_i with $X_{i(k)}$ on its face.

The probability measure of $B_{i,k}$ is given by

$$P_{i,k} = \int_{B_{i,k}} f(\xi)\, d\xi. \tag{2}$$

Although $f(X)$ is unknown, the distribution of $P_{i,k}$, denoted by $f_{i,k}(p)$, can be obtained [14]. In general, $f_{i,k}(p)$ for any sample point X_i is a beta$(k, N-k)$ distribution. Notice that this property holds regardless of the underlying distribution $f(X)$ [14].

*alizad@ece.ualberta.ca
†ardakani@ece.ualberta.ca

In kNN estimators, it is commonly assumed that $f(X)$ is uniform over the entire $B_{i,k}$. Consequently, Eq. (2) can be approximated by

$$P_{i,k} \approx \rho_{i,k}^{d} f(X_i), \tag{3}$$

where $\rho_{i,k}^{d}$ is the volume of $B_{i,k}$. Hence, $f(X_i)$ can be estimated by

$$\widehat{f}(X_i) = P_{i,k}/\rho_{i,k}^{d}. \tag{4}$$

The approximation by Eq. (3) is the basis of a vast number of studies that use kNN for various estimations. However, we will see in Sec. II B that in some cases, Eq. (3) is not accurate; hence, it can adversely affect the accuracy of any estimation made based on it. For example, Eq. (3) is used in [3] to estimate Shannon's differential entropy H_X of a continuous distribution $f(X)$. H_X is defined as [15]

$$H_X = -\int f(\xi)\log f(\xi)d\xi \tag{5}$$

$$= -E[\log f(X)], \tag{6}$$

where $E[\cdot]$ denotes the expected value, log is the natural logarithm, and H_X is measured in nats.

An estimator of H_X can be realized by

$$\widehat{H}_X = -\frac{1}{N}\sum_{i=1}^{N}\log \widehat{f}(X_i). \tag{7}$$

According to the distribution of $P_{i,k}$, i.e., beta$(k, N-k)$, we have

$$E[\log P_{i,k}] = \psi(k) - \psi(N). \tag{8}$$

Here, $\psi(x) = [d\Gamma(x)/dx]/\Gamma(x)$ is the digamma function, where $\Gamma(x)$ is the gamma function. Replacing Eqs. (4) and (8) in (7), we obtain

$$\widehat{H}_X^{(k)} = \psi(N) - \psi(k) + \frac{d}{N}\sum_{i=1}^{N}\log \rho_{i,k}. \tag{9}$$

This method is very well known and is widely used for multidimensional distributions [16]. However, in some situations, the distribution of $P_{i,k}$ calculated by approximation (3) is not beta$(k, N-k)$; hence, Eq. (8) is invalid, making Eq. (9) inaccurate.

B. Non-beta distribution at boundaries

As mentioned in Sec. II A, $f_{i,k}(p)$ is beta$(k, N-k)$. However, we will show that if X_i is located near the boundaries of a bounded distribution, the approximation (3) is inaccurate. Consequently, the distribution of the approximated $P_{i,k}$ is not beta$(k, N-k)$.

Example: Consider a one-dimensional uniform distribution with the range from $X_{\min} = 0$ to $X_{\max} = 10$, $N = 1000$, and $k = 300$. The normalized histograms of the approximated $P_{i,k}$, denoted by $\tilde{f}(P_{i,k})$, are depicted in Fig. 1(a) at different sample points $X_{i(0)} \in \{0, 1, \ldots, 10\}$. As Fig. 1(a) shows, $\tilde{f}(P_{i,k})$ is not matched with beta$(k, N-k)$ for $X_{i(0)} \leqslant 1$ and $X_{i(0)} \geqslant 9$ (points close to the boundaries). In fact, for these values of $X_{i(0)}$, the approximated values of $P_{i,k}$ by using Eq. (3) differ from the actual values obtained by Eq. (2).

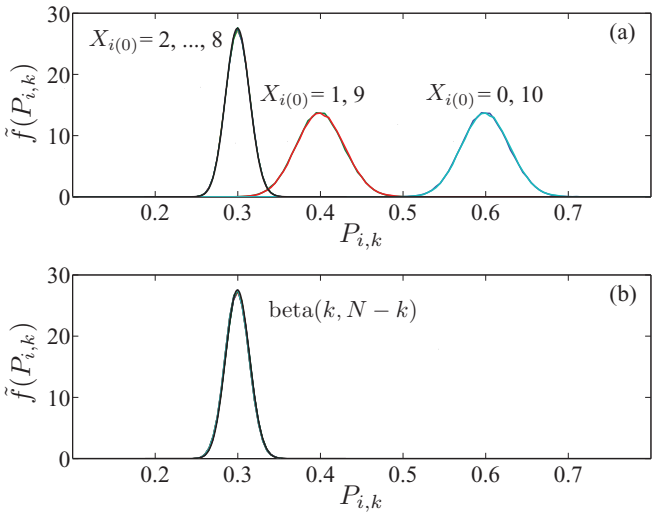

FIG. 1. (Color online) $\tilde{f}(P_{i,k})$ for a one-dimensional uniform distribution in $[0, 10]$, $N = 1000$, $k = 300$, and $X_{i(0)} \in \{0, 1, \ldots, 10\}$. (a) Without correction. (b) With correction.

In the next section, we propose a correction method resulting in an accurate approximation of $P_{i,k}$ matched to beta$(k, N-k)$. First, we will investigate the reason for the non-beta behavior of $\tilde{f}(P_{i,k})$.

III. PROPOSED CORRECTION METHOD

A. Reason for non-beta distribution

As was mentioned in Sec. II A, Eq. (3) assumes that $f(X)$ takes the constant value $f(X_i)$ in $B_{i,k}$. Hence, in a one-dimensional case, where $B_{i,k} = [X_i - \rho_{i,k}/2, X_i + \rho_{i,k}/2]$, we have $P_{i,k} \approx \rho_{i,k} f(X_i)$. Now see Fig. 2, which shows a point X_i close to the boundary of the uniform distribution of the previous example. Here, $X_i - \rho_{i,k}/2$, which is the lower border of $B_{i,k}$, is smaller than $X_{\min}$, and $f(X)$ is zero for $X < X_{\min}$. Indeed, the distribution of $f(X)$ cannot be assumed uniform in $B_{i,k}$. Hence, Eq. (2) yields

$$P_{i,k} = \int_{X_{\min}}^{X_i+\rho_{i,k}/2} f(\xi)d\xi$$

$$\approx \underbrace{(X_i + \rho_{i,k}/2 - X_{\min})}_{\rho'_{i,k}} f(X_i), \tag{10}$$

where $\rho'_{i,k}$ is the effective length of $B_{i,k}$, and obviously is smaller than $\rho_{i,k}$. In fact, in two cases, (a) $X_i - \rho_{i,k}/2 < X_{\min}$ and (b) $X_i + \rho_{i,k}/2 > X_{\max}$, $f(x)$ cannot be assumed constant

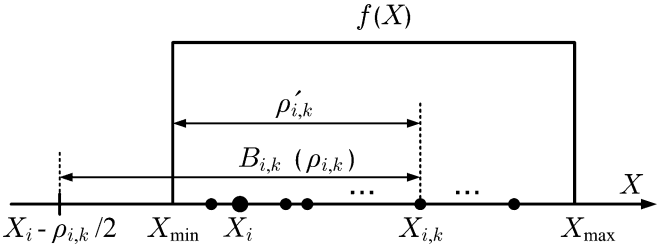

FIG. 2. $B_{i,k}$ goes beyond the boundary of the one-dimensional bounded distribution $f(X)$.

throughout the entire $B_{i,k}$, and consequently, $\rho_{i,k}$ is larger than $\rho'_{i,k}$. Hence, the approximated $P_{i,k}$ by using Eq. (3) is greater than the actual value obtained by Eq. (2), and therefore its distribution does not match with beta$(k, N-k)$.

B. Correction for one-dimensional distributions

Here, for one-dimensional distributions, we propose a correction method to approximate $P_{i,k}$ around the boundaries of $f(X)$. The generalization to multidimensional cases will be provided later.

The proposed correction method is summarized as follows:
(i) Find $X_{i(k)}$ and calculate $\rho_{i,k}$ for X_i.
(ii) Calculate the effective length of $B_{i,k}$, i.e., $\rho'_{i,k}$:
(a) $X_i - \rho_{i,k}/2 < X_{\min} \rightarrow \rho'_{i,k} = (X_i + \rho_{i,k}/2) - X_{\min}$;
(b) $X_i + \rho_{i,k}/2 > X_{\max} \rightarrow \rho'_{i,k} = X_{\max} - (X_i - \rho_{i,k}/2)$;
(c) otherwise, $\rho'_{i,k} = \rho_{i,k}$.
(iii) The corrected approximation of $P_{i,k}$ is given by

$$P_{i,k} \approx \rho'_{i,k} f(X_i). \tag{11}$$

Noticeably, cases (a) and (b) cannot occur simultaneously because $X_{i(k)}$, which is one of the borders of $B_{i,k}$, is always between the boundaries of $f(X)$. Thus, only one border can go beyond the boundaries. Moreover, in case (c), the whole of $B_{i,k}$ is inside the boundaries of $f(X)$. Hence, $\rho'_{i,k}$ equals $\rho_{i,k}$, and Eq. (11) simplifies to Eq. (3).

We should mention that as $f(X)$ is unknown, we do not know $X_{\min}$ and $X_{\max}$ for use in the correction method. Nonetheless, the minimum and maximum values of X_i $(i = 1, \ldots, N)$ can be used as the estimation of the boundaries of $f(X)$. Therefore, we can treat any set of samples that are limited to a minimum and maximum value as the samples of a bounded or truncated distribution. Hence, we can use the correction method for any set of samples.

Let us apply this correction method to the previous example. As depicted in Fig. 1(b), $\tilde{f}(P_{i,k})$ is now matched with beta$(k, N-k)$ for all values of $X_{i(0)}$. This result indicates that the boundary effect does not harm the new approximation method.

C. Correction for d-dimensional distributions

In this section, we generalize the proposed correction method for d-dimensional bounded distributions.

As mentioned in Sec. III B, for one-dimensional distributions, cases (a) and (b) cannot occur at the same time. In other words, $B_{i,k}$ cannot exceed both boundaries of $f(X)$ simultaneously. For higher dimensions, however, the boundaries of one dimension can be much smaller than those of another dimension, resulting in situations in which $B_{i,k}$ goes beyond both boundaries of the smaller dimension. Therefore, both cases (a) and (b) can occur for a given dimension at the same time. Consequently, the generalized correction method for d-dimensional distributions should deal with both boundaries of each dimension simultaneously.

In Sec. III A, we discussed the effect of going beyond the boundaries in the case of a one-dimensional distribution. Now, we will consider the effect when d dimensions are involved. If some faces of the hypercube $B_{i,k}$ are outside the boundaries of $f(X)$, the distribution of the $P_{i,k}$ calculated by Eq. (3) will not

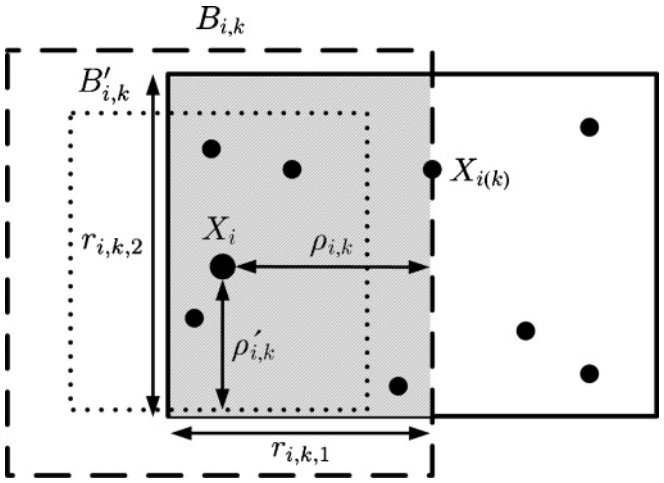

FIG. 3. A two-dimensional bounded distribution. Boundaries of $f(X)$, solid-line rectangle; $B_{i,k}$, dashed square; $B'_{i,k}$, dotted square; and $B^{\cap}_{i,k}$, shaded rectangle.

be beta$(k, N-k)$ because $f(X)$ is zero outside the boundaries. Hence, to approximate $P_{i,k}$, the volume of the intersection of $B_{i,k}$ and the space surrounded by the boundaries of $f(X)$ should be considered. This intersection, in which $f(X)$ is not zero, is denoted by $B^{\cap}_{i,k}$. For example, in Fig. 3, the boundaries of a two-dimensional $f(X)$ are represented by the solid-line rectangle, and $B_{i,k}$ is depicted by the dashed square. Hence, $B^{\cap}_{i,k}$ is the shaded rectangle.

Generally, $B^{\cap}_{i,k}$ is a hyper-rectangle whose volume is given by

$$V^{\cap}_{i,k} = \prod_{m=1}^{d} r_{i,k,m}, \tag{12}$$

where $r_{i,k,m}$ indicates the edge length of $B^{\cap}_{i,k}$ in the mth dimension and is calculated by

$$\begin{aligned} r_{i,k,m} = {} & \min\{x_{i,m} + \rho_{i,k}/2\,,\, x_{m(\max)}\} \\ & - \max\{x_{i,m} - \rho_{i,k}/2\,,\, x_{m(\min)}\}. \end{aligned} \tag{13}$$

Here, $x_{m(\max)}$ and $x_{m(\min)}$ are the maximum and minimum boundaries of $f(X)$ in the mth dimension, respectively. Similar to the one-dimensional correction method, for an unknown $f(X)$, the d-dimensional method can use the maximum and minimum values of $x_{i,m}$ $(i = 1, \ldots, N)$ as the estimation of $x_{m(\max)}$ and $x_{m(\min)}$, respectively. Also, note that Eq. (13) takes care of both boundaries of each dimension of $f(X)$ simultaneously, which is needed for correction of d-dimensional distributions.

For simplification, the hyper-rectangle $B^{\cap}_{i,k}$ can be represented by a hypercube, $B'_{i,k}$, centered at X_i whose volume is $V^{\cap}_{i,k}$. For a two-dimensional distribution, $B'_{i,k}$ is depicted by the dotted square in Fig. 3. The edge length of $B'_{i,k}$ is obtained by

$$\rho'_{i,k} = (V^{\cap}_{i,k})^{1/d}. \tag{14}$$

By this representation, $B^{\cap}_{i,k}$ and its volume can be simply determined from $\rho'_{i,k}$. Moreover, this representation is necessary to correct the kNN entropy estimator explained in Sec. IV C.

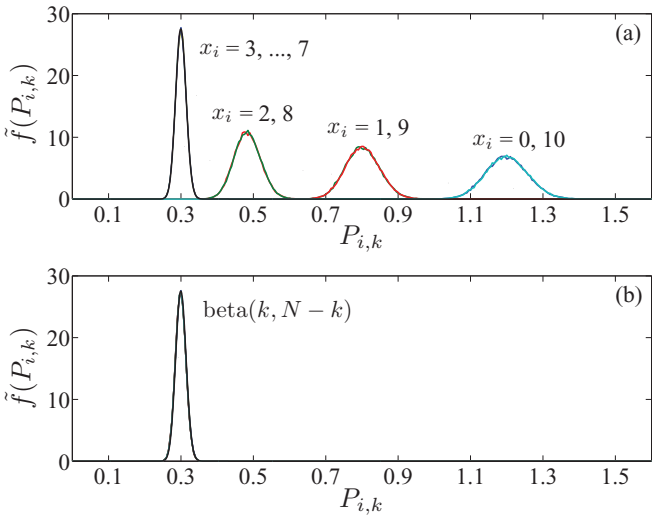

FIG. 4. (Color online) $\tilde{f}(P_{i,k})$ for a two-dimensional uniform distribution in $[0,10]^2$, $N = 1000$, $k = 300$, and $X_{i(0)} = (x_i, x_i)$, where $x_i \in \{0,1,\ldots,10\}$. (a) Without correction. (b) With correction.

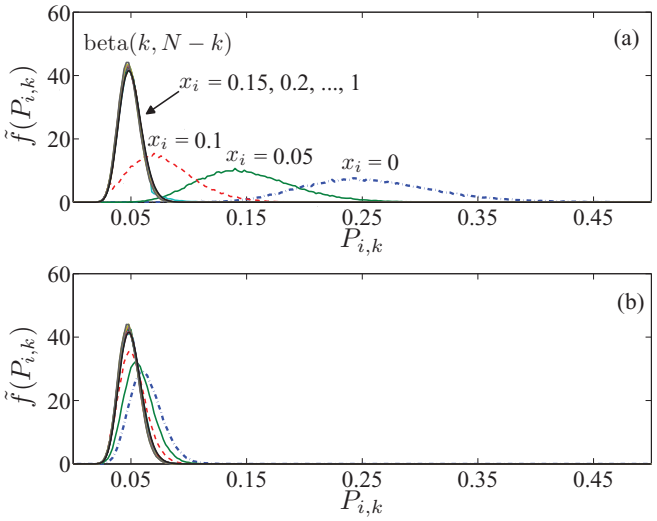

FIG. 5. (Color online) The effect of the nonuniformity of $f(x)$ on the correction method is shown by a two-dimensional exponential distribution for $N = 500$, $k = 25$, and $X_{i(0)} = (x_i, x_i)$, where $x_i \in \{0, 0.05, \ldots, 1\}$. (a) Without correction. (b) With correction.

Briefly, the corrected approximation of $P_{i,k}$ for d-dimensional $f(X)$ is

$$P_{i,k} \approx \int_{B'_{i,k}} f(X_i)\, d\xi \approx (\rho'_{i,k})^d f(X_i). \quad (15)$$

Obviously, if $B_{i,k}$ is inside the boundaries of $f(X)$, $B'_{i,k}$ and $\rho'_{i,k}$ are equivalent to $B_{i,k}$ and $\rho_{i,k}$, respectively, and Eq. (15) gives rise to Eq. (3). Now we can summarize the correction method to approximate $P_{i,k}$ for d-dimensional distributions as follows:

(i) Find $X_{i(k)}$ and calculate $\rho_{i,k}$ for X_i.

(ii) Calculate the edge lengths of the hyper-rectangle $B^{\cap}_{i,k}$ and $V^{\cap}_{i,k}$ by using Eqs. (13) and (12), respectively.

(iii) Calculate $\rho'_{i,k}$ by Eq. (14).

(iv) $P_{i,k} \approx (\rho'_{i,k})^d f(X_i)$.

IV. RESULTS AND APPLICATIONS

A. Effect of correction on $\tilde{f}(P_{i,k})$

In this section, we provide numerical results to demonstrate the performance of the proposed correction method. We have already seen one example for a one-dimensional uniform distribution. To evaluate the correction method in higher dimensions, consider a two-dimensional uniform distribution in $[0,10]^2$, $N = 1000$, $k = 300$, and $X_{i(0)} = (x_i, x_i)$, where $x_i \in \{0,1,\ldots,10\}$. We can observe again in Fig. 4(a) that for $x_i \leqslant 2$ and $x_i \geqslant 8$, $\tilde{f}(P_{i,k})$ obtained by Eq. (3) is not the same as beta$(k, N-k)$. However, in Fig. 4(b), $\tilde{f}(P_{i,k})$ obtained by the corrected formula (15) matches with beta$(k, N-k)$ for all values of $X_{i(0)}$.

In the example given above, $f(X)$ is constant over the entire $B_{i,k}$; hence, the correction method works extremely well. Now, let us examine the correction method for a nonuniform distribution. Consider a two-dimensional exponential distribution whose marginal distributions are independent with parameter $\lambda = 1$. In Fig. 5(a), $\tilde{f}(P_{i,k})$ given by Eq. (3) is plotted for $N = 1000$, $k = 25$, and $X_{i(0)} = (x_i, x_i)$, where $x_i \in \{0, 0.05, \ldots, 1\}$. Figure 5(b) provides $\tilde{f}(P_{i,k})$ by using our proposed correction method. As Fig. 5(b) reveals, $\tilde{f}(P_{i,k})$ is not perfectly matched with beta$(k, N-k)$ for $x_i \in \{0, 0.05, 0.1\}$ because for the exponential distribution, the assumption of a strictly fixed $f(X)$ over $B_{i,k}$ is not valid. This mismatch is most significant for small values of x_i where the rate of change of $f(X)$ is large. Consequently, for small values of x_i, Eq. (15) is not accurate. However, the distribution derived by the correction method is much more similar to beta$(k, N-k)$ than the corresponding distribution derived by using Eq. (3). In other words, it seems that the boundary effect is successfully eliminated and that only the effect of nonuniformity remains.

B. Effect of sample size on boundary estimation

It is mentioned in Sec. III that the boundaries of each dimension of $f(X)$ can be estimated by using the minimum and maximum values of the sample set in each dimension. Certainly, a larger sample size leads to more accurate estimates of the boundaries. Obviously, the minimum (maximum) of the data is equal to or greater (smaller) than the minimum (maximum) boundary of $f(X)$. Consequently, $\rho'_{i,k}$ calculated from the estimated boundaries would be smaller than the value obtained from the real boundaries. As a result, the $P_{i,k}$ calculated by Eq. (11) or (15), for the points close to the boundaries, is smaller than its actual value. This effect becomes noticeable for a very small sample size, because it is more likely than a larger sample size to misestimate the boundaries.

Figure 6 reveals the effect of the small sample size. Here, we have again the two-dimensional uniform distribution described in Sec. IV A for $N = 25$ and $k = 5$. Figure 6(a) depicts $\tilde{f}(P_{i,k})$ resulting from the correction method using the estimated boundaries. The non-beta histograms correspond to the points near the boundaries of $f(X)$. Figure 6(a) shows

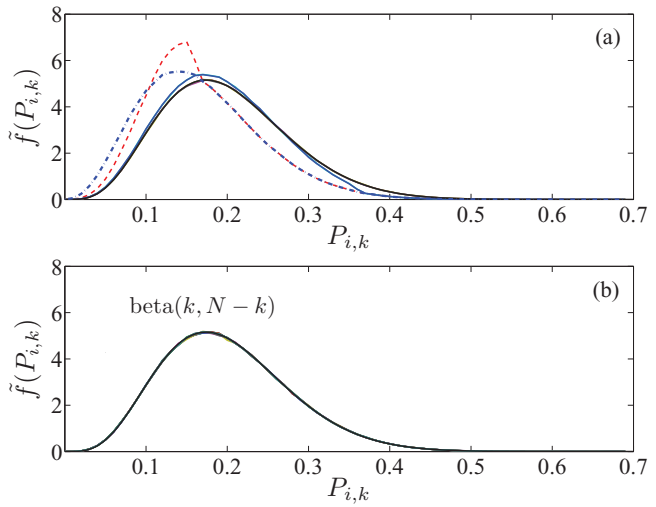

FIG. 6. (Color online) The effect of the boundary estimation on the correction method for small sample sizes. Here, $f(X)$ is uniform in $[0,10]^2$, $N = 25$, $k = 5$, and $X_{i(0)} = (x_i, x_i)$, where $x_i \in \{0,1,\ldots,10\}$. (a) Unknown boundaries. (b) Known boundaries (all the curves are over each other).

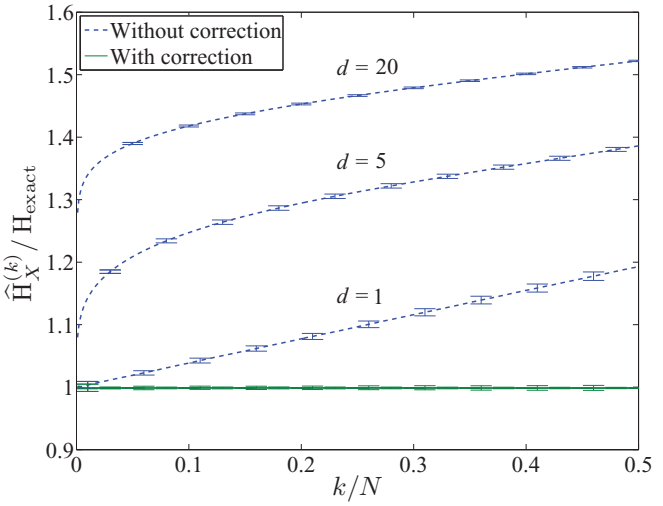

FIG. 7. (Color online) Comparison of the corrected and uncorrected estimates of entropy for uniform distribution against k/N with 1000 samples. For both methods, each curve corresponds to a fixed dimension d, with $d = 1$, 5, and 20. The curves of the corrected estimation are over each other.

that the non-beta histograms are shifted to the left compared to beta$(k, N-k)$. This result means that the $P_{i,k}$ calculated by Eq. (15) is smaller than the expected value because of the smaller amount of $\rho'_{i,k}$. Now assume that we know the real boundaries and that we use them in the correction method. The result is shown in Fig. 6(b), where all the histograms are perfectly matched with beta$(k, N-k)$. This result means that if the boundaries are known in some applications, we can benefit greatly from them, particularly for very small sample size data.

C. Application: kNN entropy estimation

As explained in Sec. II A, the beta distribution of $P_{i,k}$ is used for kNN entropy estimation. However, the boundary effect leads to the non-beta distribution of the approximated $P_{i,k}$ when using Eq. (3). Hence, the kNN entropy estimator (9) is also adversely affected. We can use the proposed correction method to resolve this problem.

The entropy estimator (9) is based on the edge length of the hypercube $B_{i,k}$ centered at X_i. To resolve the problem, we must use the edge length of $B'_{i,k}$ (i.e., $\rho'_{i,k}$), which contains the effective volume of $B_{i,k}$, instead of the edge length of $B_{i,k}$ (i.e., $\rho_{i,k}$) in Eq. (9), as follows:

$$\widehat{H}_X^{(k)} = \psi(N) - \psi(k) + \frac{d}{N}\sum_{i=1}^{N} \log \rho'_{i,k}. \qquad (16)$$

In the following subsections, we examine the corrected estimator (16) with uniform and nonuniform distributions and time series. Finally, we investigate the effect of correlated dimensions on the performance of the proposed correction.

1. Uniform distribution

Let us first investigate the uniform distribution described in Sec. IV A and generalize it to $[0,10]^d$. In Fig. 7, the estimated entropy is depicted for different values of k/N for $N = 1000$ and $d = 1, 5$, and 20. Here, the estimated entropy is normalized by the real value of the entropy, H_{exact}. In all figures, the error bars indicate the standard deviation of the estimation error normalized by H_{exact}. As Fig. 7 reveals, increasing k or d biases the uncorrected estimation drastically. In contrast, the estimated entropy by using Eq. (16) is not biased for any value of k and d. Notice that the entropy estimator (9) is proved to be asymptotically unbiased, but only if the distribution of the approximated $P_{i,k}$ is beta$(k, N-k)$ [3]. Consequently, because our correction method resolves the mismatch with the beta distribution, it provides an unbiased entropy estimation. Furthermore, the standard deviation of the corrected estimation is roughly two to five times smaller than that of the uncorrected estimate.

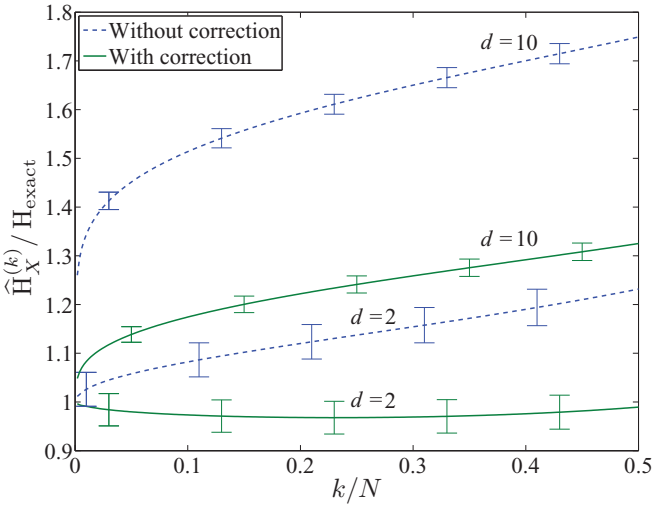

FIG. 8. (Color online) Entropy estimation for exponential distribution using the corrected and uncorrected estimators. The normalized entropies are plotted against k/N for $N = 500$. For both methods, each curve corresponds to a fixed dimension $d = 2$ and 10.

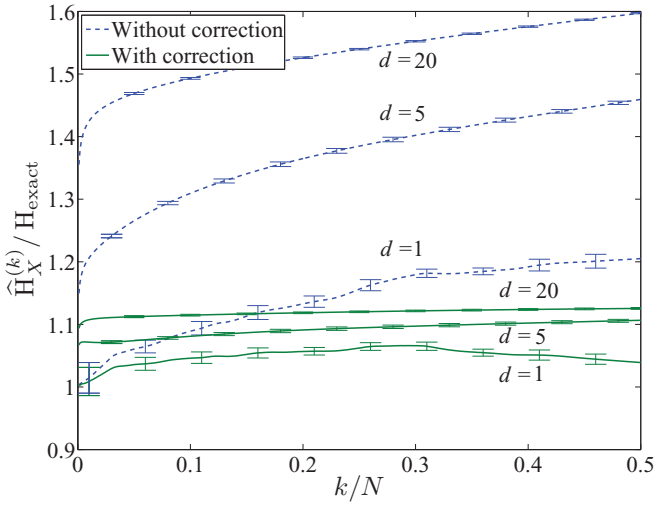

FIG. 9. (Color online) The normalized estimations of the entropy corresponding to the Henon map, with and without correction, are depicted against k/N with $N = 1000$. Curves are plotted for different dimensions, i.e., $d = 1$,5, and 20.

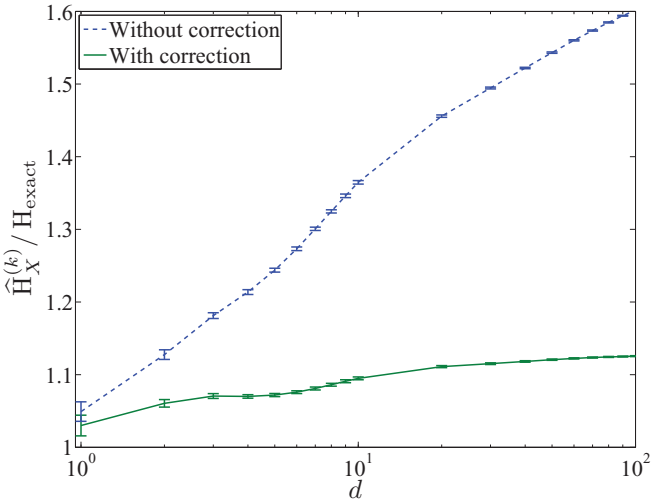

FIG. 10. (Color online) Comparison of the corrected and uncorrected kNN entropy estimators against different numbers of dimensions for the Henon map with $N = 1000$ and $k = 32$.

2. *Nonuniform distributions*

Consider again the two-dimensional exponential distribution described in Sec. IV A and generalize it to 10 dimensions. In Fig. 8, the estimated entropies using the corrected and uncorrected estimators are shown. As explained in Sec. IV A, the distribution of the corrected approximation of $P_{i,k}$ does not match perfectly with beta$(k, N-k)$ due to the nonuniformity of $f(X)$. Hence, the corrected estimation is biased, but considerably less than the uncorrected one.

3. *Entropy estimation of time series*

Entropy estimation is commonly used in time-series analysis. Most of the practical time series are bounded; hence, we can deal with them as the samples of a bounded or truncated distribution. Therefore, the proposed correction method can be performed to improve the accuracy of the estimation in these cases.

Now, let us consider the Henon map [17], which is a well-known dynamical system with equations

$$x_{t+1} = 1 - a\,x_t^2 + y_t, \;\; y_{t+1} = b\,x_t, \tag{17}$$

where $a = 1.4$, $b = 0.3$, and the subscript t denotes the time index. Here, the time series of x_t obtained by Eq. (17), represented by $\{x_t\}$, is transformed to have zero mean and unit variance. $X_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,d})$ is the d-dimensional sample vector whose elements belong to d separate time series with randomly selected different initial values. Note that for $k = 1$, the kNN estimator (9) is the estimator proposed by Kozachenko and Leonenko [2], which is asymptotically unbiased. Hence, we average the estimated entropy of $\{x_t\}$ with 10^5 samples and $d = 1$ over 50 trials, and the obtained value of $\widehat{H}_X^{(1)} = 1.126 \pm 0.002$ is used as the real entropy of $\{x_t\}$ in simulations. For d-dimensional samples, $d\,H_X^{(1)}$ is considered as H_{exact}.

Figure 9 shows the trend of the normalized estimation of the entropy against k/N for $N = 1000$ and $d = 1$,5, and 20. As Fig. 9 shows, the estimation with the correction has significantly less bias than the uncorrected one. In addition, unlike the bias of the uncorrected estimation, that of the corrected estimator does not increase noticeably with k, except for small values of k/N. In other words, in the Henon map, the corrected estimator is insensitive to k for higher dimensions. This property reduces the noise effect in data because for noisy data, we need a large value of k to have a small enough variance [18], but the bias of the uncorrected estimator (9) increases with k. Using the corrected estimator, one can increase k without increasing the bias.

In Fig. 10, 1000 samples are considered with the heuristic formula given in [11], $k = \lfloor \sqrt{N} + 0.5 \rfloor$, and the estimators are compared together for various dimension size d. As Fig. 10 indicates, for the uncorrected estimator, the normalized estimate increases about 55% when d changes from 1 to 100, whereas for the corrected estimator, this increase is about 9%.

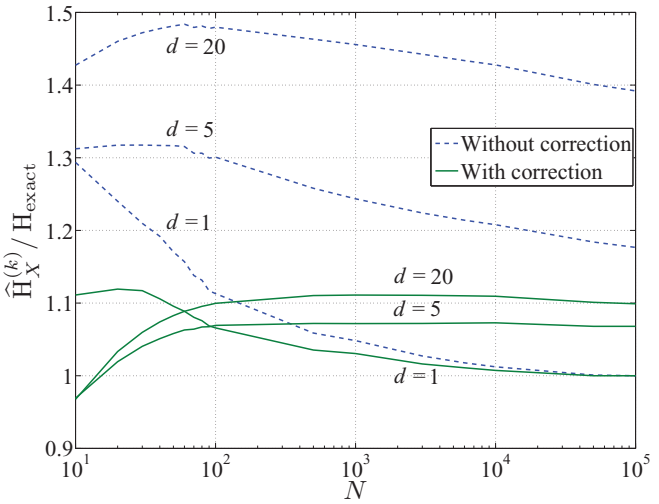

FIG. 11. (Color online) Comparison of the corrected and uncorrected kNN entropy estimators against the sample size for the Henon map with $k = \lfloor \sqrt{N} + 0.5 \rfloor$, and $d = 1$, 5, and 20 for each method.

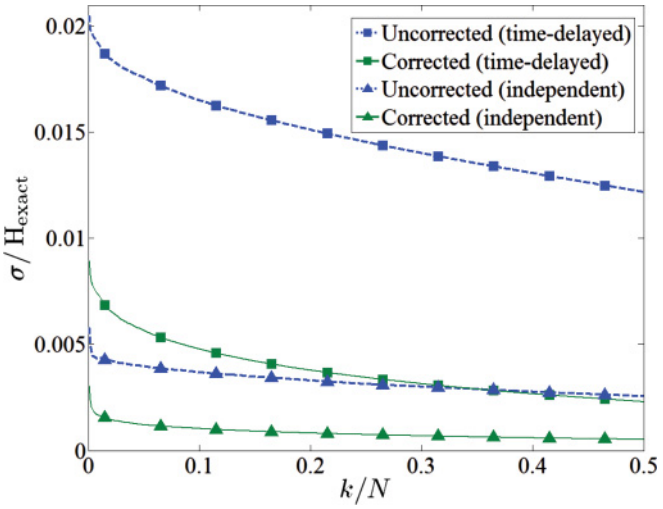

FIG. 12. (Color online) Comparison of the normalized standard deviation of the estimation error for corrected and uncorrected methods with independent and correlated (time-delayed) dimensions. The estimates are for the time-delayed vectors of a uniform distribution in [0,10] with $N = 1000$ and $d = 20$.

The bias of the corrected estimator for $d > 10$ is four to five times smaller than that of the uncorrected method, while both methods have a similar standard deviation of the estimation error. Consequently, the corrected estimator can be successfully used for high-dimensional problems.

Now, let us investigate the effect of the number of samples on both methods, as depicted in Fig. 11. For more clarity, the error bars are not drawn, but obviously the standard deviation of error decreases as N increases. For each value of N, $k = \lfloor \sqrt{N} + 0.5 \rfloor$. In the uncorrected estimation, the bias reduces when N increases for $N > 100$. However, even with 10^5 samples, we still have about 20% and 40% error for $d = 5$ and 20, respectively. On the other hand, with 100 samples or more, the corrected estimator has about 7% and 11% bias for $d = 5$ and 20, respectively. Moreover, for $d > 1$, the bias of the corrected estimator with 100 samples is approximately equal to the bias with 10^5 samples. These results mean that the corrected estimator is capable of accurately estimating the entropy with small sample sizes.

As explained in Sec. IV B, for a very small sample size, there exists an underestimation in the boundaries of $f(x)$, resulting in a smaller $\rho'_{i,k}$. Hence, for a very small sample size, we expect an underestimation in entropy. This result is seen for $N < 100$ and $d > 1$ in Fig. 11.

4. Effect of correlated dimensions

In all the examples given above, the different dimensions of X_i were independent. Now, we illustrate the performance of the proposed method for correlated dimensions. To investigate the effect of correlated dimensions on our method, we consider uniform distributions to avoid errors due to nonuniformity.

Consider a sample set of a one-dimensional uniform distribution as a time series, i.e., $\{x_t\}$. The d-dimensional vector X_i is constructed by using the time-delayed samples of $\{x_t\}$; i.e., $X_i = (x_i, x_{i-1}, \ldots, x_{i-d+1})$. This time-delayed vector has many applications in time-series data analysis, such as in transfer entropy calculation [19].

By applying the corrected and uncorrected entropy estimators to the time-delayed vectors, we find that each method has the same bias as the case of independent dimensions (Fig. 7). In other words, the same improvement reported in the case of independent dimensions is achieved again. Thus, the proposed correction also eradicates the boundary effect for correlated dimensions. It should be noted that, for correlated dimensions, the standard deviation of the estimation error is greater for both methods. For example, the normalized standard deviation of the estimation against k/N is depicted in Fig. 12 for a uniform distribution in [0,10] with $N = 1000$ and $d = 20$. As Fig. 12 shows, for both methods in the case of correlated dimensions, the standard deviation is roughly five times greater than that of independent dimensions. However, the corrected estimator has three to five times smaller standard deviation than the uncorrected one.

V. CONCLUSION AND DISCUSSION

In this paper, a correction method for the boundary effect in kNN estimators was proposed. Although this approach was developed for bounded distributions, it can be used for any set of samples from a distribution or a time series.

We applied this correction approach to multidimensional entropy estimation. It was observed that the corrected estimator was unbiased for uniform distributions and had less bias than the uncorrected estimator for nonuniform distributions. Furthermore, for different distributions, the statistical error of the corrected estimation was less than that of the estimation without correction.

It is shown in [18] that the kNN entropy estimator is promising for very small sample sizes. In addition, [3] shows that the kNN entropy estimator is capable of estimating the entropy of high-dimensional data. Here, we show that the corrected estimator performs more successfully than the original approach for small data sets and high-dimensional problems. Therefore, the proposed correction allows for further applying the kNN method to more challenging cases, e.g., the analysis of biological data where a small sample size and high dimensionality exist at the same time.

[1] D. Loftsgaarden and C. Quesenberry, Ann. Math. Statist. **36**, 1049 (1965).
[2] L. F. Kozachenko and N. N. Leonenko, Probl. Inf. Trans. **23**, 95 (1987).
[3] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).
[4] *Nearest Neighbor Pattern Classification Techniques*, edited by B. V. Dasarathy (IEEE Computer Society Press, Los Alamitos, CA, 1991).
[5] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition* (Springer, New York, 1996).

[6] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, Knowledge Inf. Syst. **14**, 1 (2008).
[7] J. D. Victor, Phys. Rev. E **66**, 051903 (2002).
[8] E. J. Harner, H. Singh, S. Li, and J. Tan, in *Computing Science and Statistics* (2003), p. 35.
[9] N. Misra, H. Singh, and V. Hnizdo, Entropy **12**, 1125 (2010).
[10] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, Am. J. Math. Manage. Sci. **23**, 301 (2003).
[11] M. Goria, N. Leonenko, V. Mergel, and P. Inverardi, J. Nonparametr. Statist. **17**, 277 (2005).
[12] Q. Wang, S. R. Kulkarni, and S. Verdú, IEEE Trans. Inf. Theory **55**, 2392 (2009).
[13] E. Liitiäinen, A. Lendasse, and F. Corona, Random Struct. Algorithms **37**, 223 (2010).
[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition (Computer Science and Scientific Computing Series)*, 2nd ed. (Academic Press, San Diego, 1990).
[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley Interscience, Hoboken, NJ, 2006).
[16] Q. Wang, S. R. Kulkarni, and S. Verdú, Found. Trends Commun. Inf. Theory **5**, 265 (2009).
[17] M. Hénon, Commun. Math. Phys. **50**, 69 (1976).
[18] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson, V. Protopopescu, and G. Ostrouchov, Phys. Rev. E **76**, 026209 (2007).
[19] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).