Jackknife approach to the estimation of mutual information

Xianli Zeng[a], Yingcun Xia[a,b,1], and Howell Tong[b,c]

[a]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546; [b]School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China; and [c]Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, United Kingdom

Edited by Larry Wasserman, Carnegie Mellon University, Pittsburgh, PA, and approved August 20, 2018 (received for review November 16, 2017)

Quantifying the dependence between two random variables is a fundamental issue in data analysis, and thus many measures have been proposed. Recent studies have focused on the renowned mutual information (MI) [Reshef DN, et al. (2011) *Science* 334:1518–1524]. However, "Unfortunately, reliably estimating mutual information from finite continuous data remains a significant and unresolved problem" [Kinney JB, Atwal GS (2014) *Proc Natl Acad Sci USA* 111:3354–3359]. In this paper, we examine the kernel estimation of MI and show that the bandwidths involved should be equalized. We consider a jackknife version of the kernel estimate with equalized bandwidth and allow the bandwidth to vary over an interval. We estimate the MI by the largest value among these kernel estimates and establish the associated theoretical underpinnings.

jackknifed estimation | mutual information | statistical dependence | kernel density estimation

A key issue in data science is how to measure the dependence between two random variables. Pearson's correlation coefficient (1) provides a powerful measure for linear dependence, but it is incapable of detecting nonlinear association (2, 3). Thus, many other measures have been introduced to quantify complex dependence. For example, Gretton et al. (4) proposed a kernel-based independence criterion that uses the squared Hilbert–Schmidt norm of the cross-covariance operator. Székely, Rizzo, and Bakirov (5) introduced the distance correlation (dCor) which does not involve any nonparametric estimation and is free of tuning parameters. Heller, Heller, and Gorfine (6) proposed a rank-based distance measure which demonstrates good numerical performance. Although many different measures have been proposed, the mutual information introduced by Claude Shannon in 1948 (7) is not replaceable and is still of great research interest. As a fundamental measure of dependence, mutual information (MI) possesses several desirable properties and can be interpreted intuitively (8). These advantages secure MI as a very powerful measure of nonlinear dependence with very wide applications in data analysis. As such, studies have focused on its mathematical properties and its estimation efficiency (2, 9, 10).

For continuous data, there are three typical groups of estimation for MI. The first group is the "bins" method that discretizes the continuous data into different bins and estimates MI from the discretized data (11, 12). The asymptotic performance for this bins method is systematically analyzed in ref. 13. The second group is based on estimates of probability density functions, for example, the histogram estimator of ref. 14, the kernel density estimator (KDE) of ref. 15, the B-spline estimator of ref. 16, and the wavelet estimator of ref. 17. To reduce the bias at the boundary region, ref. 18 introduced the mirrored KDE and derived its exponential concentration bound. Recently, ref. 19 further applied the ensemble method in kernel estimation and derived the ensemble estimator. The third group is based on the relationship between the MI and entropies. One of the most popular estimations in this group is the k-nearest neighbors (kNN) estimator introduced in ref. 20, which was extended in ref. 10, leading to the introduction of the Kraskov–Stögbauer–Grassberger (KSG) estimator. This estimator is further discussed in refs. 21–23.

Although many different approaches have been considered, the estimation of MI, especially for continuous data, relies heavily on the choice of the tuning parameters involved such as the number of bins, the bandwidth in kernel density estimation, and the number of neighbors in the kNN estimator. As a consequence, the corresponding estimators may be very unstable or seriously biased. However, little research has been done on the selection of these parameters. In this paper, we focus on the KDE approach (24, 25), which involves at least four bandwidth matrices. Experience with tests for independence suggests that the bandwidths should be set equal. Equalization of bandwidths also helps us reduce the bias at the boundary region and thus improve the efficiency of estimation. To free the estimation from bandwidth selection, we consider a jackknife version of MI (called JMI) and show that JMI has asymptotically a unique maximum with respect to the equalized bandwidth. We adopt the maximum value as our estimator of MI and provide the necessary statistical underpinnings. Interestingly, for the very special case of independent random variables, JMI enjoys a consistency rate higher than that of root-n. Numerically, we compare the estimation efficiency of JMI vs. that of other estimation methods that include the mixed KSG of ref. 23, the copula-based KSG of ref. 21, and other KDE methods. We also construct a test for independence (2, 3, 26) based on the JMI and compare it with several popular methods, such as the dCor of ref. 5, the maximal information coefficient (MIC) of ref. 9, and the Heller–Heller–Gorfine (HHG) test of ref. 6. These comparisons demonstrate the superior performance of JMI.

Significance

As a fundamental concept in information theory, mutual information has been commonly applied to quantify the dependence between variables. However, existing estimations have unstable statistical performance since they involve a set of tuning parameters. We develop a jackknife approach that does not incur predetermined tuning parameters. The proposed approach enjoys several appealing theoretical properties and has stable numerical performance.

Author contributions: X.Z. and Y.X. designed research; X.Z. and Y.X. performed research; and X.Z., Y.X., and H.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All calculation codes used in this study have been deposited in GitHub, https://github.com/XianliZeng/JMI.

[1]To whom correspondence should be addressed. Email: yingcun.xia@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1715593115/-/DCSupplemental.

Published online September 17, 2018.

MI and Its Kernel Estimation

Consider two random variables $\mathbf{X}=(X_1, X_2, \ldots, X_P)'$ and $\mathbf{Y}=(Y_1, Y_2, \ldots, Y_Q)'$. Let us focus on the case that their joint probability density function exists. MI is defined as

$$MI(\mathbf{X},\mathbf{Y}) = E\left\{\log \frac{f_{\mathbf{XY}}(\mathbf{X},\mathbf{Y})}{f_{\mathbf{X}}(\mathbf{X})f_{\mathbf{Y}}(\mathbf{Y})}\right\},$$

where $f_{\mathbf{X}}$, $f_{\mathbf{Y}}$, and $f_{\mathbf{XY}}$ are the density functions of $\mathbf{X}$, $\mathbf{Y}$, and $(\mathbf{X},\mathbf{Y})$, respectively. This definition can be easily extended to other types of random variables that may not have density functions (13). As a measure of complex dependence, MI possesses the following desirable properties. It is always nonnegative, i.e., $MI(\mathbf{X},\mathbf{Y})\geq 0$, and equality holds if and only if the two variables are independent. Moreover, the stronger is the dependence between two variables, the larger is the MI. MI is also invariant under strictly monotonic variable transformations. More recently, Kinney and Atwal (2) proved that MI satisfies the so-called self-equitability, indicating that MI places the same importance on linear and nonlinear dependence.

Let $S=\{(\mathbf{x}_i,\mathbf{y}_i), i=1,\ldots,n\}$ with $\mathbf{x}_i=(x_{i1}, x_{i2}, \ldots, x_{iP})'$ and $\mathbf{y}_i=(y_{i1}, y_{i2}, \ldots, y_{iQ})'$ be independent samples from $(\mathbf{X},\mathbf{Y})$. Let $|\cdot|$ represent the determinant of a matrix. Consider the following multivariate KDEs:

$$\hat{f}_{\mathbf{X},\mathbf{H_X}}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{K}^P_{\mathbf{H_X}}(\mathbf{x}_j-\mathbf{x});$$

$$\hat{f}_{\mathbf{Y},\mathbf{H_Y}}(\mathbf{y}) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{K}^Q_{\mathbf{H_Y}}(\mathbf{y}_j-\mathbf{y});$$

$$\hat{f}_{\mathbf{XY},\mathbf{B_X},\mathbf{B_Y}}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{K}^P_{\mathbf{B_X}}(\mathbf{x}_j-\mathbf{x})\mathbf{K}^Q_{\mathbf{B_Y}}(\mathbf{y}_j-\mathbf{y}).$$

Here $\mathbf{H_X}$, $\mathbf{H_Y}$, $\mathbf{B_X}$, and $\mathbf{B_Y}$ are diagonal bandwidth matrices with $\mathbf{H_X}=diag(h^2_{\mathrm{X}_1}, h^2_{\mathrm{X}_2}, \ldots, h^2_{\mathrm{X}_P})$, $\mathbf{H_Y}=diag(h^2_{\mathrm{Y}_1}, h^2_{\mathrm{Y}_2}, \ldots, h^2_{\mathrm{Y}_Q})$, $\mathbf{B_X}=diag(b^2_{\mathrm{X}_1}, b^2_{\mathrm{X}_2}, \ldots, b^2_{\mathrm{X}_P})$, and $\mathbf{B_Y}=diag(b^2_{\mathrm{Y}_1}, b^2_{\mathrm{Y}_2}, \ldots, b^2_{\mathrm{Y}_Q})$; typically, for diagonal matrix $\mathbf{A}$, $\mathbf{K}^P_{\mathbf{A}}(\mathbf{x})=|\mathbf{A}|^{-1/2}\mathbf{K}^P(\mathbf{A}^{-1/2}\mathbf{x})$, where $\mathbf{K}^P$ is a P-dimensional symmetric density function with $\mathbf{K}^P(\mathbf{x})=\prod_{p=1}^{P}K(x_p)$. Based on these estimators, the KDE of MI is

$$\hat{I}_1(\mathbf{B_X},\mathbf{B_Y},\mathbf{H_X},\mathbf{H_Y}) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\hat{f}_{\mathbf{XY},\mathbf{B_X},\mathbf{B_Y}}(\mathbf{x}_i,\mathbf{y}_i)}{\hat{f}_{\mathbf{X},\mathbf{H_X}}(\mathbf{x}_i)\hat{f}_{\mathbf{Y},\mathbf{H_Y}}(\mathbf{y}_i)}. \quad [1]$$

This estimator is consistent under some mild conditions. However, as pointed out by ref. 15, its numerical performance is heavily influenced by the choice of bandwidths. Another problem is the notorious boundary effect, which becomes more serious as $\hat{f}_{\mathbf{X},\mathbf{H_X}}(\mathbf{x}_i)$ and $\hat{f}_{\mathbf{Y},\mathbf{H_Y}}(\mathbf{y}_i)$ appear in the denominator.

In checking the independence between $\mathbf{X}$ and $\mathbf{Y}$, we usually compare the product of frequencies in hypercubes $\mathcal{D}_{\mathbf{X}}=\{\mathbf{x}_i : |x_{ip}-x_p|<h_{\mathrm{X}_p}, p=1,2,\ldots,P\}$ and $\mathcal{D}_{\mathbf{Y}}=\{\mathbf{y}_i : |y_{iq}-y_q|<h_{\mathrm{Y}_q}, q=1,2,\ldots,Q\}$ with the frequency in their intersection $\mathcal{D}_{\mathbf{X}}\cap\mathcal{D}_{\mathbf{Y}}=\{(\mathbf{x}_i,\mathbf{y}_i) : |x_{ip}-x_p|<h_{\mathrm{X}_p}, |y_{iq}-y_q|<h_{\mathrm{Y}_q}, p=1,2,\ldots,P, q=1,2,\ldots,Q\}$. Let $\#$ denote the number of elements in a set. By taking $\mathbf{K}^P(\mathbf{x})=\prod_{p=1}^{P}[\mathbf{1}_{(|x_p|\leq 1)}/2]$, where $\mathbf{1}_A$ is the indicator function of set A, those frequencies are, respectively,

$$\frac{\#\mathcal{D}_{\mathbf{X}}}{n} = \frac{2^P}{n}\sum_{i=1}^{n}\prod_{p=1}^{P}K\left(\frac{x_{ip}-x_p}{h_{\mathrm{X}_p}}\right) = 2^P|\mathbf{H_X}|^{1/2}\hat{f}_{\mathbf{X},\mathbf{H_X}}(\mathbf{x}),$$

$$\frac{\#\mathcal{D}_{\mathbf{Y}}}{n} = \frac{2^Q}{n}\sum_{i=1}^{n}\prod_{q=1}^{Q}K\left(\frac{y_{iq}-y_q}{h_{\mathrm{Y}_q}}\right) = 2^Q|\mathbf{H_Y}|^{1/2}\hat{f}_{\mathbf{Y},\mathbf{H_Y}}(\mathbf{y}),$$

and

$$\frac{\#(\mathcal{D}_{\mathbf{X}}\cap\mathcal{D}_{\mathbf{Y}})}{n} = \frac{2^{P+Q}}{n}\sum_{i=1}^{n}\prod_{p=1}^{P}K\left(\frac{x_{ip}-x_p}{h_{\mathrm{X}_p}}\right)\prod_{q=1}^{Q}K\left(\frac{y_{iq}-y_q}{h_{\mathrm{Y}_q}}\right)$$
$$= 2^{P+Q}(|\mathbf{H_X}||\mathbf{H_Y}|)^{1/2}\hat{f}_{\mathbf{XY},\mathbf{H_X},\mathbf{H_Y}}(\mathbf{x},\mathbf{y}).$$

Thus, the ratio of comparison is

$$\frac{\#(\mathcal{D}_{\mathbf{X}}\cap\mathcal{D}_{\mathbf{Y}})/n}{\#\mathcal{D}_{\mathbf{X}}/n\,\#\mathcal{D}_{\mathbf{Y}}/n} = \frac{\hat{f}_{\mathbf{XY},\mathbf{H_X},\mathbf{H_Y}}(\mathbf{x},\mathbf{y})}{\hat{f}_{\mathbf{X},\mathbf{H_X}}(\mathbf{x})\hat{f}_{\mathbf{Y},\mathbf{H_Y}}(\mathbf{y})}.$$

Note that the bandwidth matrices corresponding to $\mathbf{X}$ in both estimators of $f_{\mathbf{X}}$ and $f_{\mathbf{X},\mathbf{Y}}$ are the same, and the same is true for the case of $\mathbf{Y}$. We therefore argue that $\mathbf{B_X}=\mathbf{H_X}$ and $\mathbf{B_Y}=\mathbf{H_Y}$ should be imposed on the KDE. With these equalizations of bandwidths, the joint and marginal densities are well defined; i.e., $\int\hat{f}_{\mathbf{XY},\mathbf{H_X},\mathbf{H_Y}}(\mathbf{x},\mathbf{y})\,d\mathbf{x}=\hat{f}_{\mathbf{Y},\mathbf{H_Y}}(\mathbf{y})$ and $\int\hat{f}_{\mathbf{XY},\mathbf{H_X},\mathbf{H_Y}}(\mathbf{x},\mathbf{y})\,d\mathbf{y}=\hat{f}_{\mathbf{X},\mathbf{H_X}}(\mathbf{x})$, which is an important feature in the definition of MI. Another important motivation for equalizing the bandwidths is that it can automatically reduce the estimation bias at the boundary region; see theoretical justification in the next section.

Jackknife Estimation of MI

Marginal transformation is an efficient way to improve the estimation of ref. 1 and to reduce the technical complexity (27, 28). We consider the uniform transformation $\mathbf{U}=(U_1, U_2, \ldots, U_P)'=(F_{X_1}(X_1), F_{X_2}(X_2), \ldots, F_{X_P}(X_P))'$ and $\mathbf{V}=(V_1, V_2, \ldots, V_Q)'=(F_{Y_1}(Y_1), F_{Y_2}(Y_2), \ldots, F_{Y_Q}(Y_Q))'$, where F_{X_p}, $p=1,2,\ldots,P$ and $F_{Y_q}, q=1,2,\ldots,Q$ are the cumulative distribution functions of X_p and Y_q, respectively. It is easy to see that $MI(\mathbf{U},\mathbf{V})=MI(\mathbf{X},\mathbf{Y})$. Use $c_{\mathbf{U}}(\mathbf{u})$, $c_{\mathbf{V}}(\mathbf{v})$ and $c_{\mathbf{UV}}(\mathbf{u},\mathbf{v})$ to denote the copula density functions of $\mathbf{U}$, $\mathbf{V}$, and $(\mathbf{U},\mathbf{V})$, respectively. For observed data, the corresponding transformation is $(\mathbf{u}^\star_i,\mathbf{v}^\star_i)'=(u^\star_{i1},\ldots,u^\star_{iP},v^\star_{i1},\ldots,v^\star_{iQ})'=(F_{X_1,n}(x_1),\ldots,F_{X_P,n}(x_P),F_{Y_1,n}(y_1),\ldots,F_{Y_Q,n}(y_Q))'$ with $F_{X_p,n}, p=1,2,\ldots,P$ and $F_{Y_q,n}, q=1,2,\ldots,Q$ denoting the empirical cumulative distribution functions of $\{x_{ip}, i=1,\ldots,n\}$ and $\{y_{iq}, i=1,\ldots,n\}$, respectively.

A main problem with the KDE is the selection of bandwidths. In the literature, many selection methods have been proposed (29–31). Although good properties have been proved for these selectors in the estimation of density function, applying them to the estimation of MI does not work well (15, 32).

Jackknife Estimation. Note that MI is zero when the two random variables are independent, and generally bigger MIs imply stronger dependence. Thus, the bandwidths should be selected to maximize MI to detect possible dependence. This idea was also applied in statistical tests (33, 34). However, maximizing MI tends to overfit the MI, making it possibly infinite. Instead, we use the jackknife estimation (27, 35). Let

$$\hat{I}_2(\mathbf{B_U},\mathbf{B_V},\mathbf{H_U},\mathbf{H_V}) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\hat{c}^{\setminus i}_{\mathbf{UV},\mathbf{B_U},\mathbf{B_V}}(\mathbf{u}^\star_i,\mathbf{v}^\star_i)}{\hat{c}^{\setminus i}_{\mathbf{U},\mathbf{H_U}}(\mathbf{u}^\star_i)\hat{c}^{\setminus i}_{\mathbf{V},\mathbf{H_V}}(\mathbf{v}^\star_i)} \quad [2]$$

with bandwidth matrices $\mathbf{H_U}=diag(h^2_{\mathrm{U}_1}, h^2_{\mathrm{U}_2}, \ldots, h^2_{\mathrm{U}_P})$, $\mathbf{H_V}=diag(h^2_{\mathrm{V}_1}, h^2_{\mathrm{V}_2}, \ldots, h^2_{\mathrm{V}_Q})$, $\mathbf{B_U}=diag(b^2_{\mathrm{U}_1}, b^2_{\mathrm{U}_2}, \ldots, b^2_{\mathrm{U}_P})$, $\mathbf{B_V}=diag(b^2_{\mathrm{V}_1}, b^2_{\mathrm{V}_2}, \ldots, b^2_{\mathrm{V}_Q})$, and

STATISTICS

$$\hat{c}^{\setminus i}_{\mathbf{U},\mathbf{H_U}}(\mathbf{u}) = \frac{1}{n-1}\sum_{j\neq i}\mathbf{K}^P_{\mathbf{H_U}}(\mathbf{u}^\star_j - \mathbf{u});$$

$$\hat{c}^{\setminus i}_{\mathbf{V},\mathbf{H_V}}(\mathbf{v}) = \frac{1}{n-1}\sum_{j\neq i}\mathbf{K}^Q_{\mathbf{H_V}}(\mathbf{v}^\star_j - \mathbf{v});$$

$$\hat{c}^{\setminus i}_{\mathbf{UV},\mathbf{B_U},\mathbf{B_V}}(\mathbf{u},\mathbf{v}) = \frac{1}{n-1}\sum_{j\neq i}\mathbf{K}^P_{\mathbf{B_U}}(\mathbf{u}^\star_j - \mathbf{u})\mathbf{K}^Q_{\mathbf{B_V}}(\mathbf{v}^\star_j - \mathbf{v}).$$

We estimate MI by the maximum of $\hat{I}_2(\mathbf{B_U},\mathbf{B_V},\mathbf{H_U},\mathbf{H_V})$. With four bandwidth matrices as arguments, this maximization is not easy. However, as suggested above, the bandwidths in the calculation should be set equal. The bias expansion in *Theorem 1* also indicates that this equalization of bandwidths in Eq. **2** is helpful for counteracting boundary bias.

For theoretical analysis, we define an "oracle" estimator

$$\hat{I}_0 = \frac{1}{n}\sum_{i=1}^{n}\log\frac{c_{\mathbf{UV}}(\mathbf{u}_i,\mathbf{v}_i)}{c_{\mathbf{U}}(\mathbf{u}_i)c_{\mathbf{V}}(\mathbf{v}_i)}.$$

Note that $\hat{I}_0$ is exactly 0 when the two variables are independent and, by the central limit theorem, it is root-n consistent when $0 < MI < \infty$ and $E\left[\log\frac{c_{\mathbf{UV}}(\mathbf{U},\mathbf{V})}{c_{\mathbf{U}}(\mathbf{U})c_{\mathbf{V}}(\mathbf{V})}\right]^2 < \infty$. As $\hat{I}_0$ is actually not obtainable, hereafter we use it only as a benchmark for asymptotic analysis.

Theorem 1. *Under general regularity conditions (SI Appendix, section C, Assumptions A.1, A.2, A.5, and A.6) on functions K, $c_{\mathbf{U}}$, $c_{\mathbf{V}}$, and $c_{\mathbf{UV}}$ and bandwidth matrices, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\log\hat{c}^{\setminus i}_{\mathbf{UV},\mathbf{B_U},\mathbf{B_V}}(\mathbf{u}^\star_i,\mathbf{v}^\star_i) - \frac{1}{n}\sum_{i=1}^{n}\log c_{\mathbf{UV}}(\mathbf{u}_i,\mathbf{v}_i)$$
$$= -C_1\left[\sum_{p=1}^{P}b_{\mathrm{U}_p} + \sum_{q=1}^{Q}b_{\mathrm{V}_q}\right] + o_P\left(\sum_{p=1}^{P}|b_{\mathrm{U}_p}| + \sum_{q=1}^{Q}|b_{\mathrm{V}_q}|\right),$$

where C_1 is given in in SI Appendix, Eq. **S1**. *Furthermore,*

$$\hat{I}_2(\mathbf{H_U},\mathbf{H_V},\mathbf{B_U},\mathbf{B_V}) - \hat{I}_0$$
$$= C_1\left[\sum_{p=1}^{P}(h_{\mathrm{U}_p} - b_{\mathrm{U}_p}) + \sum_{q=1}^{Q}(h_{\mathrm{V}_q} - b_{\mathrm{V}_q})\right]$$
$$+ o_P\left(\sum_{p=1}^{P}(|h_{\mathrm{U}_p}| + |b_{\mathrm{U}_p}|) + \sum_{q=1}^{Q}(|h_{\mathrm{V}_q}| + |b_{\mathrm{V}_q}|)\right).$$

The first part of *Theorem 1* indicates that a kernel-based estimate of a copula density has a bias of the same order as the bandwidth. It is caused by the boundary points since the kernel density estimation has a much faster consistency rate for the interior points (36). The second part concerns $\hat{I}_2$, which involves a divisor in the form of a product of two marginal copula densities; it shows that its bias depends on the difference of bandwidths, i.e., $h_{\mathrm{U}_p} - b_{\mathrm{U}_p}$ and $h_{\mathrm{V}_q} - b_{\mathrm{V}_q}$. If the commonly used selectors (24, 25, 31, 37) are adopted to select the bandwidths separately, then b_{U_p} and b_{V_q} are of order $n^{-1/(P+Q+4)}$ while h_{U_p} and h_{V_q} are of order $n^{-1/(P+4)}$ and $n^{-1/(Q+4)}$, respectively. Consequently, the kernel estimator of MI with those bandwidths would suffer a bias of order $n^{-1/(P+Q+4)}$. However, if we equalize the bandwidths, $\mathbf{B_U} = \mathbf{H_U}$ and $\mathbf{B_V} = \mathbf{H_V}$, the leading term in the bias will be eliminated automatically. To further simplify the bandwidth selection, we also restrict $\mathbf{H_U} = h^2\mathbf{I}_P$ and $\mathbf{H_V} = h^2\mathbf{I}_Q$ with $\mathbf{I}_n$ the $n\times n$ identity matrix. This restriction is reasonable since each component of $\mathbf{U}$ and $\mathbf{V}$ is uniformly distributed on $[0,1]$ after the transformation. Thus, we consider a jackknife estimator of MI with one common bandwidth,

$$\hat{I}_3(h) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\hat{c}^{\setminus i}_{\mathbf{UV},h^2\mathbf{I}_P,h^2\mathbf{I}_Q}(\mathbf{u}^\star_i,\mathbf{v}^\star_i)}{\hat{c}^{\setminus i}_{\mathbf{U},h^2\mathbf{I}_P}(\mathbf{u}^\star_i)\hat{c}^{\setminus i}_{\mathbf{V},h^2\mathbf{I}_Q}(\mathbf{v}^\star_i)}.$$

Our final estimator is

$$\widehat{JMI}(\mathbf{X},\mathbf{Y}) = \max_{h>0}[\hat{I}_3(h)].$$

Since this estimation procedure involves a maximization problem, the existence of a global maximum is crucial for computation. With a different relationship between $\mathbf{X}$ and $\mathbf{Y}$, the objective function $\hat{I}_3(h)$ possesses three typical shapes: (*i*) When $\mathbf{X}$ and $\mathbf{Y}$ are independent of each other, $\hat{I}_3(h)$ is an increasing function bounded by 0 as in Fig. 1 *Ci*; the maximum is achieved with a large h and tends to 0. (*ii*) When $\mathbf{X}$ and $\mathbf{Y}$ are functionally dependent (i.e., one is a function of the other), the objective function is monotone decreasing as in Fig. 1 *Cii*, suggesting a zero-valued bandwidth and that the estimated information tends to infinity. (*iii*) When $\mathbf{X}$ and $\mathbf{Y}$ are partially correlated, $\hat{I}_3(h)$ is a unimodal function as shown in Fig. 1 *Ciii*. These shapes are further justified by *Theorem 2*. In particular, when $\mathbf{X}$ and $\mathbf{Y}$ are partially correlated, *Theorem 2* indicates that the unique maximum is achieved at $h \propto n^{-1/(P+Q+3)}$.

Theorem 2. *Under general regularity conditions (SI Appendix, section C, Assumptions A.1, A.5, and A.7) on functions K, $c_{\mathbf{U}}$, $c_{\mathbf{V}}$, and bandwidth,*

a) if $MI < \infty$ and $c_{\mathbf{UV}}$ satisfies the regularity condition A.2 in SI Appendix, section C, then

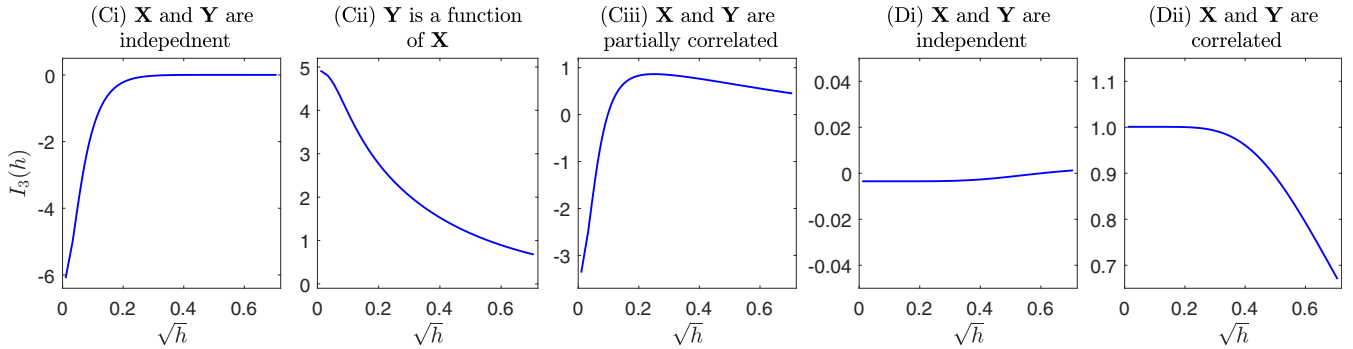

Fig. 1. Typical shapes of function $\hat{I}_3(h)$. In *Ci*, *Cii*, and *Ciii*, at least one of $\mathbf{X}$ and $\mathbf{Y}$ is a continuous variable; in *Di* and *Dii*, both $\mathbf{X}$ and $\mathbf{Y}$ are discrete.

$$\hat{I}_3(h) - \hat{I}_0 = -\frac{C_2}{2}h^3 - \frac{C_3}{2nh^{P+Q}} + o_P(h^3 + \frac{1}{nh^{P+Q}}),$$

where C_2 and C_3 are two nonnegative constants given in SI Appendix, section C. In particular, $C_2 = 0$ when $\mathbf{X}$ and $\mathbf{Y}$ are independent.

b) If $MI = \infty$, and $\mathbf{X}$ and $\mathbf{Y}$ are functionally dependent and satisfy regularity condition A.3 in SI Appendix, section C, then

$$\hat{I}_3(h) = \min(P, Q)\log\frac{1}{h} + o_P\left(\log\frac{1}{h}\right).$$

Although JMI is defined for continuous random variables, it also applies to the discrete random variables and mixed random variables (with neither purely continuous distributions nor purely discrete distributions). If both $\mathbf{X}$ and $\mathbf{Y}$ are discrete, $\hat{I}_3(h)$ depends on h as shown in Fig. 1 *Di* or *Dii*: If they are independent, $\hat{I}_3(h)$ changes only marginally with h; otherwise $\hat{I}_3(h)$ remains unchanged when h is smaller than a certain value and is decreasing thereafter, suggesting any bandwidth that is small enough will give the same estimator. If $\mathbf{X}$ and $\mathbf{Y}$ have a mixture of continuous components and discrete components, $\hat{I}_3(h)$ depends on h in the same way as in Fig. 1 *Ci*, *Cii*, and *Ciii*. More details about extension of JMI to discrete case can be found in *SI Appendix, section G*.

Compared with existing methods such as mirrored KDE (18), ensemble KDE (19), copula-based KSG (21), and mixed KSG (23), the $\widehat{JMI}(\mathbf{X}, \mathbf{Y})$ has several advantages. First, it is more computationally efficient since only one common bandwidth is introduced. Second, the procedure is completely data driven, and we provide a stable selection procedure for bandwidth h so that no tuning parameter needs to be predetermined. Third, our method does not necessitate boundary correction and yet it retains the same estimation efficiency because the boundary biases are eliminated automatically. Finally and most importantly, taking the unique maximum value makes $\widehat{JMI}(\mathbf{X}, \mathbf{Y})$ numerically stable.

Consistency of the Estimation. We have the following results for the consistency of our estimator.

Theorem 3. *Under general regularity conditions (SI Appendix, section C, Assumptions A.1, A.5, and A.8) on functions K, c_U, c_V, and bandwidth,*

a) if $MI < \infty$ and c_{UV} satisfies regularity condition A.2 in SI Appendix, section C, then

$$\widehat{JMI}(\mathbf{X}, \mathbf{Y}) - \hat{I}_0 = O_P(n^{-3/(P+Q+3)});$$

b) if $MI = \infty$ and c_{UV} satisfies regularity condition A.4 in SI Appendix, section C, then

$$\widehat{JMI}(\mathbf{X}, \mathbf{Y}) \to \infty, \quad a.s..$$

Note that $\hat{I}_0$ is the oracle estimator of MI with root-n consistency. When X and Y are both univariate random variables, *Theorem 3a* indicates that JMI has, in general, the same root-n consistency as $\hat{I}_0$, which is the minimax rate (38) for estimating MI. Interestingly, for the special case when X and Y are independent, $\hat{I}_0 = 0$ and the estimator converges to 0 at rate $n^{-3/5}$, which is even faster than root-n. We exploit this faster consistency rate for the independent case to yield a test for independence with high local power as shown in *SI Appendix, section F*.

Test for Independence

The proposed JMI estimator is a natural test statistic for independence. Tests based on asymptotic distributions require data with large sample size. In contrast, the permutation technique can give a precise distribution for even small samples (39–41). For a random sample of n observations, $S = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$, from $(\mathbf{X}, \mathbf{Y})$, let $\delta(1), \delta(2), \ldots, \delta(n)$ be a random permutation of $1, 2, \ldots, n$. Based on data $S_1 = \{(\mathbf{x}_i, \mathbf{y}_{\delta(i)}), i = 1, \ldots, n\}$, calculate the jackknife estimator, denoted by $\widehat{JMI}_1$. On repeating the above procedure N times, $T = \{\widehat{JMI}_k, k = 1, \ldots, N\}$ are obtained. The distribution of $\widehat{JMI}$, under the null hypothesis H_0, can be approximated by the empirical distribution of T. At significance level α, we reject the null hypothesis when $\widehat{JMI}$ based on the original data is greater than the $(1-\alpha)$th quantile of T.

Simulation Study

Estimation of MI. In this section, we compare the efficiency of the JMI estimator with that of existing estimators of MI. As shown in ref. 23, the mixed KSG has the best performance among the three kNN-based estimators they considered and outperforms the partitioning method in their simulation studies. We thus include the mixed KSG as a representative of kNN-based methods in the comparison. As the copula-based KSG (21) makes the same marginal transformation as ours, it is also included in the comparison. To illustrate the stability of the bandwidth selection of JMI, we compare its performance with that of mirrored KDE (18) and other KDE methods that select bandwidths, respectively, by the rule-of-thumb method (THUMB) and the plug-in method (PLUG-IN). We use the same models as in ref. 23 and their variations for multivariate cases, all together nine models, to evaluate the estimation methods. Details of these models are listed in *SI Appendix, section D*. Similar to ref. 23, the mean-squared errors (MSEs) of different methods for different models and sample sizes are calculated based on 250 replications. The results for the first five models are plotted in Fig. 2, while those for the other four models are shown in *SI Appendix, section E*. It can be seen that the other KDE methods have much bigger MSE than JMI, possibly due to the fact that unequalized bandwidths tend to cause boundary estimation bias. Copula-based KSG also performs badly as it is not applicable directly to discrete random variables. Mixed KSG shows satisfactory performance in all of the models but is worse than JMI, especially for models I and III. In comparison, our JMI has the smallest MSE for almost all models across different sample sizes, indicating its superior performance.

Test for Independence. Many statistical tests have been proposed for independence. As demonstrated in ref. 3, HHG of ref. 6 usually has the best performance among all of the existing methods. To ease visualization, we include only dCor of ref. 5, HHG of ref. 6, and MIC of ref. 9 in this simulation study. We examine 16 models that were used in refs. 2, 3, and 6 which include both univariate models and multivariate models. For each model, we further consider the additive noise and the case where the data are contaminated with pure noise. Details of these models are listed in *SI Appendix, section D*. Similar to ref. 2, we simulate data with sample size $n = 320$ and 25 different magnitudes of noise level. For models with additive noise, we increase the noise level by changing the noise ratio amplitude (NR), while for models with contaminations, we raise the noise level by introducing more contaminating observations. We plot the power curves for some of the results in Fig. 3, and the others are in *SI Appendix, section E*.

We summarize the results as follows. dCor performs well only for linear models with symmetric additive noise, but poorly for the other models. Similar to results in refs. 2 and 3, MIC

STATISTICS

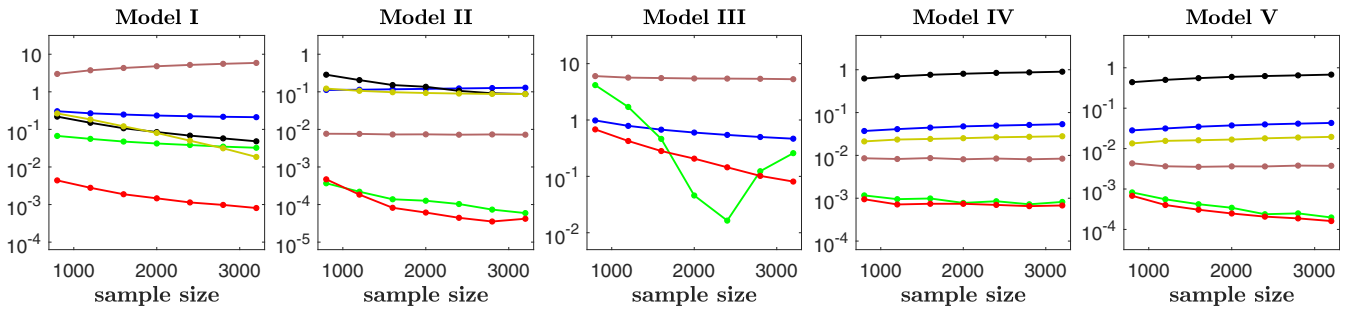

Fig. 2. In each panel, the lines represent MSEs of different estimation methods based on 250 replications. Correspondence between colors and the methods are as follows: JMI in red, mixed KSG in green, rule of thumb KDE in blue, plug-in KDE in black, mirrored KDE in dark yellow, and copula-based KSG in brown. For model III, the plug-in method and the mirrored KDE method are not calculated due to their excessive computational complexity.

performs very well in testing independence for the high-frequency sine model but it fails in other models. In most cases, HHG has relatively better performance than MIC and dCor, which is consistent with the findings in ref. 3. Generally, JMI appears to be the most stable test for independence. It has similar performance to HHG for models with additive noise but has clear superiority for data with contaminated observations.

Discussion

In this paper, we have introduced a jackknife kernel estimation for the MI (JMI). Inspired by statistical tests for independence and guided by asymptotic analysis, we propose that the bandwidths involved in the estimation should be set equal. For measuring the dependence, we estimate the MI using its maximum value with respect to the equalized bandwidth. JMI is completely data driven and does not incur predetermined tuning parameters. It enjoys very good statistical properties such as automatic bias correction and high local power in testing for independence. The superiority of JMI is also demonstrated through simulation studies. We attribute the good performance of JMI to two factors: (*i*) the definition of MI itself and (*ii*) our estimation method that enjoys several advantages mentioned above.

Materials and Methods

"Mixed KSG" was calculated by the codes in ref. 23. "Copula-based KSG" was based on the algorithm described in ref. 21. "Mirrored KDE" was

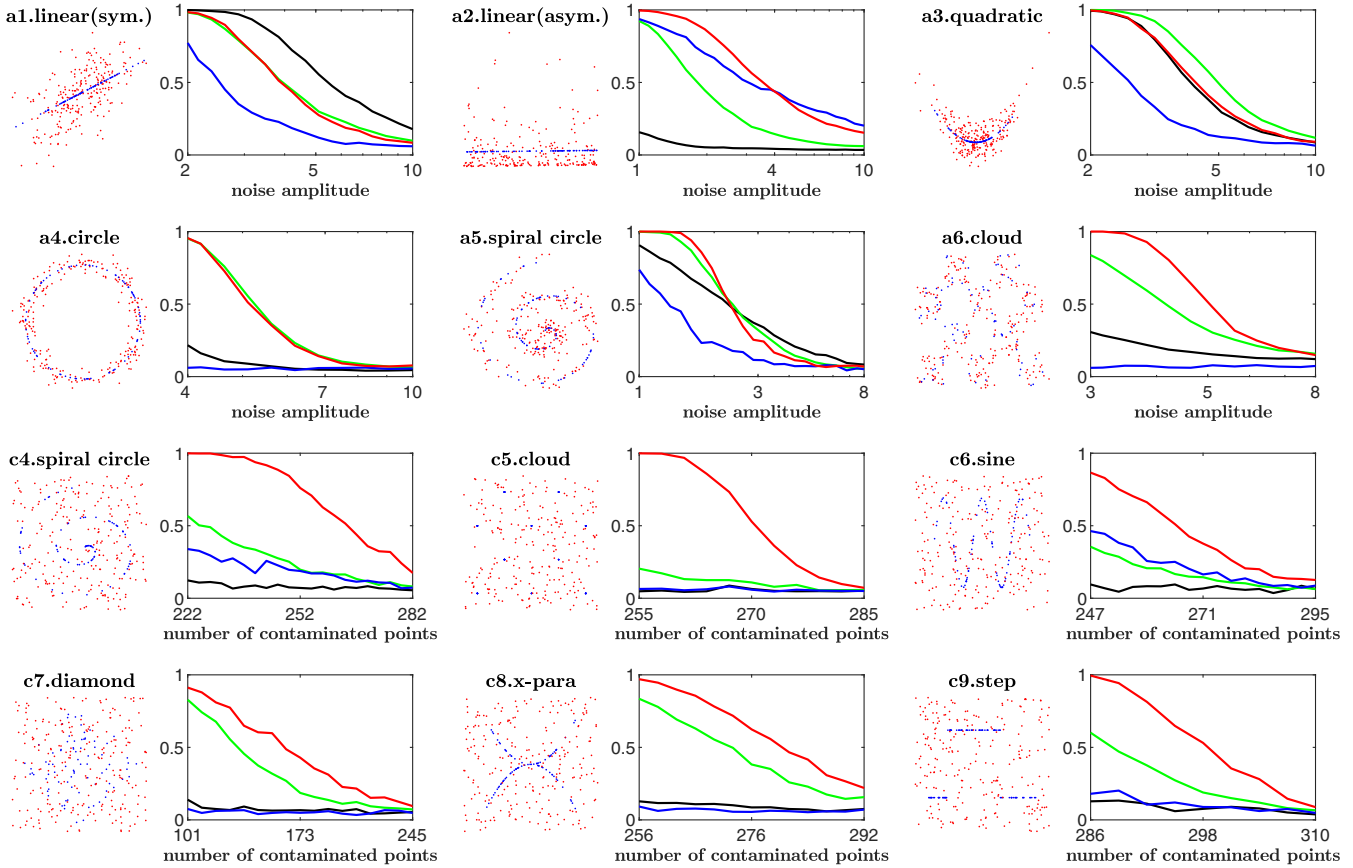

Fig. 3. In each pair of panels, *Left* shows the model and the data, and *Right* shows the power of tests at significance level $\alpha = 0.05$: Models a1–a6 with additive noise are in *Upper* panels, and models c4–c9 with contaminated noise are in *Lower* panels. For power curves, correspondence between colors and different methods is as follows: red for JMI, green for HHG, blue for MIC, and black for dCor.

computed by the algorithm introduced in ref. 42. For the other methods, the simulations were conducted in R. "dCor" and "MIC" were estimated using packages "energy" and "Minerva," respectively. The HHG test was carried out using the package "HHG." For kernel methods, the rule of thumb used the optimal Gaussian bandwidth given in ref. 24; bandwidths for "PLUG-IN" were calculated using the package "ks." "JMI" was calculated by the procedure discussed in *SI Appendix, section B*. For estimation efficiency, MSE for each model and sample size was calculated based on 250 replications. For the test of independence, we adopted the permutation test discussed above. For each sample, we randomly permuted observations of **Y** to generate samples under the null hypothesis. The power curve for each model and at each noise level was calculated based on 1,000 simulations. Our calculation codes are available at https://github.com/XianliZeng/JMI.

ACKNOWLEDGMENTS. We are most grateful to the editor and two referees for their meticulous review, valuable comments, and constructive suggestions, which have led to a substantial improvement of this paper. Y.X. is partially supported by MOE AcRF Grant of Singapore R-155-000-193-114.

1. Pearson K (1895) Note on regression and inheritance in the case of two parents. *Proc R Soc Lond* 58:240–242.
2. Kinney JB, Atwal GS (2014) Equitability, mutual information, and the maximal information coefficient. *Proc Natl Acad Sci USA* 111:3354–3359.
3. Ding AA, Li Y (2013) Copula correlation: An equitable dependence measure and extension of Pearson's correlation. arXiv:1312.7214.
4. Gretton A, Bousquet O, Smola A, Schölkopf B (2005) Measuring statistical dependence with Hilbert-Schmidt norms. *International Conference on Algorithmic Learning Theory*, eds Jain S, Simon HU, Tomita E (Springer, Berlin), pp 63–77.
5. Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. *Ann Stat* 35:2769–2794.
6. Heller R, Heller Y, Gorfine M (2012) A consistent multivariate test of association based on ranks of distances. *Biometrika* 100:503–510.
7. Shannon CE (1948) A mathematical theory of communication. *Bell Labs Tech J* 27:379–423.
8. Cover TM, Thomas JA (2006) *Elements of Information Theory* (Wiley-Interscience, New York), 2nd Ed.
9. Reshef DN, et al. (2011) Detecting novel associations in large data sets. *Science* 334:1518–1524.
10. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. *Phys Rev E* 69:066138.
11. Bialek W, Rieke F, Van Steveninck RDR, Warland D (1991) Reading a neural code. *Science* 252:1854–1857.
12. Strong SP, Koberle R, van Steveninck RRDR, Bialek W (1998) Entropy and information in neural spike trains. *Phys Rev Lett* 80:197–200.
13. Paninski L (2003) Estimation of entropy and mutual information. *Neural Comput* 15:1191–1253.
14. Fraser AM, Swinney HL (1986) Independent coordinates for strange attractors from mutual information. *Phys Rev A* 33:1134–1140.
15. Moon YI, Rajagopalan B, Lall U (1995) Estimation of mutual information using kernel density estimators. *Phys Rev E* 52:2318–2321.
16. Daub CO, Steuer R, Selbig J, Kloska S (2004) Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinformatics* 5:118.
17. Peter AM, Rangarajan A (2008) Maximum likelihood wavelet density estimation with applications to image and shape matching. *IEEE Trans Image Process* 17:458–468.
18. Singh S, Póczos B (2014) Exponential concentration of a density functional estimator. *Advances in Neural Information Processing Systems* 27, eds Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (Curran Associates, Inc., Red Hook, NY), pp 3032–3040.
19. Moon KR, Sricharan K, Hero AO (2017) Ensemble estimation of mutual information. *2017 IEEE International Symposium on Information Theory (ISIT)*, eds Durisi G, Studer C (IEEE, Aachen, Germany), pp 3030–3034.
20. Singh H, Misra N, Hnizdo V, Fedorowicz A, Demchuk E (2003) Nearest neighbor estimates of entropy. *Am J Math Manag Sci* 23:301–321.
21. Pál D, Póczos B, Szepesvàri C (2010) Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. *Advances in Neural Information Processing Systems* 23, eds Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A (Curran Associates, Inc., Red Hook, NY), pp 1849–1857.
22. Gao S, Ver Steeg G, Galstyan A (2015) Efficient estimation of mutual information for strongly dependent variables. *Artificial Intelligence and Statistics*, eds Lebanon G, Vishwanathan SVN (Proceedings of Machine Learning Research, San Diego), pp 277–286.
23. Gao W, Kannan S, Oh S, Viswanath P (2017) Estimating mutual information for discrete-continuous mixtures. *Advances in Neural Information Processing Systems*, eds Guyon I, et al. (Curran Associates, Inc., Red Hook, NY), Vol 30, pp 5986–5997.
24. Silverman BW (1986) *Density Estimation for Statistics and Data Analysis* (CRC Press, Boca Raton, FL), Vol 26.
25. Scott DW (2015) *Multivariate Density Estimation: Theory, Practice, and Visualization* (Wiley, New York).
26. Khan S, et al. (2007) Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys Rev E* 76: 026209.
27. Hong Y, White H (2005) Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica* 73:837–901.
28. Geenens G, Charpentier A, Paindaveine D (2017) Probit transformation for nonparametric kernel estimation of the copula density. *Bernoulli* 23:1848–1873.
29. Bowman AW (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 71:353–360.
30. Scott DW, Terrell GR (1987) Biased and unbiased cross-validation in density estimation. *J Am Stat Assoc* 82:1131–1146.
31. Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc Ser B Methodol* 53:683–690.
32. Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* 18:S231–S240.
33. Horowitz JL, Spokoiny VG (2001) An adaptive rate optimal test of a parametric mean regression model against a nonparametric alternative. *Econometrica* 69:599–631.
34. Heller R, Heller Y, Kaufman S, Brill B, Gorfine M (2016) Consistent distribution-free *k*-sample and independence tests for univariate random variables. *J Machine Learn Res* 17:1–54.
35. Efron B, Stein C (1981) The jackknife estimate of variance. *Ann Stat* 9:586–596.
36. Fan J, Gijbels I (1996) *Local Polynomial Modelling and its Applications: Monographs on Statistics and Applied Probability* (Chapman and Hall, London), Vol 66.
37. Wand MP, Jones MC (1994) *Kernel Smoothing* (Chapman and Hall/CRC, Boca Raton, FL).
38. Kandasamy K, Krishnamurthy A, Poczos B, Wasserman L (2015) Nonparametric von Mises estimators for entropies, divergences and mutual informations. *Advances in Neural Information Processing Systems*, eds Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (Curran Associates, Inc., Red Hook, NY), pp 397–405.
39. Fisher RA (1935) *The Design of Experiments* (Oliver and Boyd, Edinburgh, London).
40. Manly BF (2006) *Randomization, Bootstrap and Monte Carlo Methods in Biology* (Chapman and Hall/CRC, Boca Raton, FL) Vol 70.
41. Edgington E, Onghena P (2007) *Randomization Tests* (Chapman and Hall/CRC, Boca Raton, FL).
42. Singh S, Póczos B (2014) Generalized exponential concentration inequality for Rényi divergence estimation. *International Conference on Machine Learning,* eds Xing EP, Jebara T (Proceedings of Machine Learning Research, Bejing, China), pp 333–341.