

# A Nearest-Neighbor Approach to Estimating Divergence between Continuous Random Vectors

Qing Wang, Sanjeev R. Kulkarni, Sergio Verdú  
 Department of Electrical Engineering  
 Princeton University  
 Princeton, NJ 08544 USA  
 Email: {qingwang, kulkarni, verdu}@princeton.edu

**Abstract**—A method for divergence estimation between multi-dimensional distributions based on nearest neighbor distances is proposed. Given i.i.d. samples, both the bias and the variance of this estimator are proven to vanish as sample sizes go to infinity. In experiments on high-dimensional data, the nearest neighbor approach generally exhibits faster convergence compared to previous algorithms based on partitioning.

## I. INTRODUCTION

Suppose  $P$  and  $Q$  are two probability distributions on  $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$ . The divergence between  $P$  and  $Q$  is defined as [1]

$$D(P\|Q) \equiv \int_{\mathbb{R}^d} dP \log \frac{dP}{dQ}, \quad (1)$$

when  $P$  is absolutely continuous with respect to  $Q$ , and  $+\infty$  otherwise. If the densities of  $P$  and  $Q$  with respect to Lebesgue measure exist, denoted by  $p(x)$  and  $q(x)$  respectively, with  $p(x) = 0$  for  $P$ -almost every  $x$  such that  $q(x) = 0$  and  $0 \log \frac{0}{0} \equiv 0$ , then

$$D(p\|q) \equiv \int_{\mathbb{R}^d} p(x) \log \frac{p(x)}{q(x)} dx. \quad (2)$$

Divergence gauges how differently two random variables are distributed and it provides a useful measure of discrepancy between distributions. The key role of divergence in information theory and large deviations is well known. There has been a growing interest in applying divergence to various fields of science and engineering for the purpose of estimation, classification, etc [3]-[11].

Despite its wide range of applications, relatively limited work has been done on the universal estimation of divergence, see [12] and references therein. The traditional approach is to use histograms with equally sized bins to estimate the densities  $p(x)$  and  $q(x)$  and substitute the density estimates  $\hat{p}(x)$  and  $\hat{q}(x)$  into (2). Reference [12] proposes an estimator based on data-dependent partitioning. Instead of estimating the two densities separately, this method estimates the Radon-Nikodym derivative  $dP/dQ$  using frequency counts on a statistically equivalent partition of  $\mathbb{R}^d$ . As commented in [13], the estimation bias of this method originates from two sources: finite resolution and finite sample sizes. The basic estimator

in [12] can be improved by choosing the number of partitions adaptively or by correcting the error due to finite sample sizes. Algorithm G from [13] is the most advanced version which combines these two schemes. Although this algorithm is applicable to estimating divergence for multi-dimensional data, the computational complexity is exponential in  $d$  and the estimation accuracy deteriorates quickly as the dimension increases. The intuition is that to attain a fixed accuracy, the required number of samples grows exponentially with the dimension. However, in many applications, e.g. neural coding [14], only moderately large sizes of high-dimensional data are available. This motivates us to search for alternative approaches to efficiently estimate divergence for data in  $\mathbb{R}^d$ .

In this paper, we present a new estimator based on nearest neighbor (NN) distances which bypasses the difficulties associated with partitioning in a high-dimensional space. The nearest neighbor method, since its inception in 1951 [15], has been shown to be a powerful nonparametric technique for classification [16] [17], density estimation [18] and regression estimation [19] [20]. In [21], Kozachenko and Leonenko used NN distances to estimate differential entropy and proved the mean-square consistency of the resulting estimator for data of any dimension. Tsybakov and van der Meulen [22] considered a truncated version of the differential entropy estimator and showed its  $1/\sqrt{n}$ -rate of convergence for a class of one-dimensional densities with unbounded support and exponentially decreasing tails. In [14] and [23], the NN estimator of differential entropy is applied to estimate mutual information  $I(X; Y)$ , via

$$I(X; Y) = h(X) + h(Y) - h(X, Y), \quad (3)$$

$$\text{where } h(X) \equiv - \int_{\mathbb{R}^d} p_X(x) \log p_X(x) dx, \quad (4)$$

is the differential entropy of a random variable  $X$  distributed according to the density function  $p_X$ . Our work focuses on the estimation of divergence using NN distances, which is inspired by the  $k$ -NN density estimation. The  $k$ -NN density estimate at a point  $x$  can be expressed as

$$\hat{p}_k(x) = \frac{k/n}{V_{x,k}}, \quad (5)$$

<sup>1</sup>This work was supported in part by ARL MURI under Grant number DAAD19-00-1-0466, Draper Laboratory under IR&D 6002 grant DL-H-546263, and the National Science Foundation under grant CCR-0312413.

where  $n$  is the total number of samples and  $V_{x,k}$  is the volume of the ball centered at estimation point  $x$  with radius equal to distance of  $x$  to its  $k$ -nearest neighbor in the given samples.

Let  $\{X_i\}$  and  $\{Y_i\}$  denote i.i.d. samples generated independently according to densities  $p$  and  $q$  respectively. Let  $\hat{p}_k$  and  $\hat{q}_k$  be the corresponding  $k$ -NN density estimates. Suppose we evaluate  $\hat{p}_k$  and  $\hat{q}_k$  at  $\{X_i\}_{i=1,\dots,n}$ . Then, by the law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n \log \frac{\hat{p}_k(X_i)}{\hat{q}_k(X_i)} \quad (6)$$

will give us a consistent estimate of  $D(p||q)$  provided that the density estimates  $\hat{p}_k$  and  $\hat{q}_k$  satisfy some consistency conditions. For the  $k$ -NN density estimates to be consistent,  $k$  should go to  $\infty$  with the sample size [18]. However, in the construction of our NN divergence estimator,  $k$  can be any constant. In this paper,  $k$  is fixed to be 1 and the resulting estimator is still consistent in the sense that both the bias and the variance vanish as sample sizes increase. Detailed development and convergence analysis of this estimator are given in Sections II and III respectively. Section IV provides experimental results to compare the performance of the NN method and Algorithm G.

## II. ESTIMATES OF DIVERGENCE BASED ON NEAREST NEIGHBOR DISTANCES

Let  $\{X_1, \dots, X_n\}$  and  $\{Y_1, \dots, Y_m\}$  be i.i.d.  $d$ -dimensional samples drawn independently from the densities  $p$  and  $q$  respectively. The distance of  $X_i$  to its nearest neighbor in  $\{X_j\}_{j \neq i}$  is defined as

$$\rho_n(i) = \min_{j=1,\dots,n, j \neq i} \|X_i - X_j\|, \quad (7)$$

where  $\|\cdot\|$  is the  $L^2$  norm in  $\mathbb{R}^d$ .

Our goal is to estimate the divergence between  $p$  and  $q$  given i.i.d. samples  $\{X_i\}$  and  $\{Y_i\}$ . The relationship between these two sets of samples is employed in the estimation of divergence. Namely, in addition to  $\rho_n(i)$  as defined above, we also use the distance of  $X_i$  to its nearest neighbor in  $\{Y_j\}$

$$\nu_m(i) = \min_{j=1,\dots,m} \|X_i - Y_j\|. \quad (8)$$

Now consider the  $d$ -dimensional open ball centered at  $X_i$  with radius  $\rho_n(i)$ , denoted as  $B(X_i, \rho_n(i))$ . Among the samples  $\{X_j\}_{j \neq i}$ , only one  $X_j$  falls into the closure of  $B(X_i, \rho_n(i))$ . Empirically, the density estimate of  $p$  at  $X_i$  is

$$\hat{p}(X_i) = \frac{1/(n-1)}{c_1(d)\rho_n^d(i)}, \quad (9)$$

where  $c_1(d)\rho_n^d(i)$  is the volume of  $B(X_i, \rho_n(i))$ ,  $c_1(d) = \pi^{d/2}/\Gamma(d/2+1)$ . Similarly, given  $\{Y_j\}_{j=1,\dots,m}$ , only one  $Y_j$  is contained in the closure of  $B(X_i, \nu_m(i))$ , the density estimate of  $q$  evaluated at  $X_i$  is

$$\hat{q}(X_i) = \frac{1/m}{c_1(d)\nu_m^d(i)}. \quad (10)$$

Inspired by (6), we propose the following NN divergence estimator:

$$\begin{aligned} \hat{D}_{n,m}(p||q) &= \frac{1}{n} \sum_{i=1}^n \log \frac{\frac{1/(n-1)}{c_1(d)\rho_n^d(i)}}{\frac{1/m}{c_1(d)\nu_m^d(i)}} \\ &= \frac{d}{n} \sum_{i=1}^n \log \frac{\nu_m(i)}{\rho_n(i)} + \log \frac{m}{n-1}. \end{aligned} \quad (11)$$

The convergence properties of the divergence estimator in (11) are established in the following section.

## III. ANALYSIS

In this section, we prove that the bias (Theorem 1) and the variance (Theorem 2) of the NN divergence estimator (11) vanish as sample sizes increase, provided that some regularity conditions are satisfied. In contrast to our previous results on partitioning estimators in [12], in this analysis, we assume that  $\{X_1, \dots, X_n\}$  is independent of  $\{Y_1, \dots, Y_m\}$  in order to establish mean-square consistency (Theorem 2) whereas this assumption is not required for showing the asymptotic unbiasedness (Theorem 1).

*Theorem 1:* Suppose that the probability density functions  $p, q$  satisfy the following conditions:

$$\int_{\mathbb{R}^d} |\log p(x)|^{1+\epsilon} p(x) dx < \infty, \quad (12)$$

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\log \|x - y\||^{1+\epsilon} p(x)p(y) dx dy < \infty, \quad (13)$$

$$\int_{\mathbb{R}^d} |\log q(x)|^{1+\epsilon} p(x) dx < \infty, \quad (14)$$

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\log \|x - y\||^{1+\epsilon} p(x)q(y) dx dy < \infty, \quad (15)$$

for some  $\epsilon > 0$ . Then the divergence estimator (11) is asymptotically unbiased, i.e.

$$\lim_{n,m \rightarrow \infty} \mathbb{E} \hat{D}_{n,m}(p||q) = D(p||q). \quad (16)$$

*Proof:* Rewrite  $\hat{D}_{n,m}(p||q)$  as

$$\hat{D}_{n,m}(p||q) = \frac{1}{n} \sum_{i=1}^n [\log(m\nu_m^d(i)) - \log((n-1)\rho_n^d(i))].$$

$$\text{Let } \psi_m(i) \triangleq \log(m\nu_m^d(i)), \quad \zeta_n(i) \triangleq \log((n-1)\rho_n^d(i)).$$

$$\text{Then } \hat{D}_{n,m}(p||q) = \frac{1}{n} \sum_{i=1}^n [\psi_m(i) - \zeta_n(i)]. \quad (17)$$

Since  $\psi_m(i) - \zeta_n(i)$ ,  $i = 1, \dots, n$  are identically distributed,

$$\mathbb{E} \hat{D}_{n,m}(p||q) = \mathbb{E} [\psi_m(i) - \zeta_n(i)]. \quad (18)$$

Now it suffices to show that the right hand side of (18) converges to  $D(p||q)$  as  $m, n \rightarrow \infty$ . The proof techniques are from [21]. Let us first consider the expectation of  $\psi_m(i)$ , which can be obtained by finding the conditional distribution of  $\exp(\psi_m(i))$  given  $X_i = x$ . In fact, for almost all  $x \in \mathbb{R}^d$ ,

$$\begin{aligned} G_{m,x}(u) &\triangleq P\{\exp(\psi_m(i)) < u | X_i = x\} \\ &= P\{\nu_m(i) < (u/m)^{1/d} | X_i = x\} \\ &= 1 - P\{\cap_{j=1}^m \{Y_j \notin B(x, (u/m)^{1/d})\}\} \\ &= 1 - \left(1 - \int_{B(x, (u/m)^{1/d})} q(y) dy\right)^m, \end{aligned} \quad (19)$$

where  $u \in R_+$ . Note that for almost all  $x \in \mathbb{R}^d$  and any sequences of open balls  $B(x, r_k)$  of radius  $r_k \rightarrow 0$ ,

$$\lim_{k \rightarrow \infty} \frac{1}{\lambda(B(x, r_k))} \int_{B(x, r_k)} q(y) dy = q(x), \quad (20)$$

if  $q(x) \in L^1(R^d)$  ( $\lambda$  represents the Lebesgue measure). Therefore, as  $m \rightarrow \infty$ ,

$$G_{m,x}(u) \rightarrow 1 - \exp(-c_1(d)q(x)u), \quad (21)$$

where  $c_1(d) = \pi^{d/2}/\Gamma(d/2 + 1)$ .

Let  $\xi_{m,x}$  be a random variable with the distribution function  $G_{m,x}(u)$  and  $\xi_x$  a random variable with the distribution function  $G_x(u) \triangleq 1 - \exp(-c_1(d)q(x)u)$ . Then for  $q(x) > 0$ ,

$$\begin{aligned} & E \log \xi_x \\ &= \int_0^\infty \log u \exp(-c_1(d)q(x)u) c_1(d)q(x) du \\ & \stackrel{t=q(x)c_1(d)u}{=} \int_0^\infty \log [t/(c_1(d)q(x))] e^{-t} dt \\ &= -\log q(x) - \log c_1(d) - \gamma, \end{aligned} \quad (22)$$

where  $\gamma = -\int_0^\infty \log t e^{-t} dt \approx 0.5772$  is the Euler-Mascheroni constant.

Also note that  $E \log \xi_{m,x} = E\{\psi_m(i)|X_i = x\}$ . Therefore

$$\lim_{m \rightarrow \infty} E\{\psi_m(i)|X_i = x\} = -\log q(x) - \log c_1(d) - \gamma, \quad (23)$$

for any  $x$  such that

$$\lim_{m \rightarrow \infty} E \log \xi_{m,x} = E \log \xi_x. \quad (24)$$

Furthermore, since we already know that  $\xi_{m,x} \xrightarrow{\mathcal{D}} \xi_x$ , (24) holds, if we have  $E|\log \xi_{m,x}|^{1+\epsilon} < C$  for some  $\epsilon > 0$  and some  $C > 0$ , according to [[25], vol. 2, pp. 251], which can be verified using (15).

Now we only need to show that for  $m \rightarrow \infty$ ,

$$\begin{aligned} E\psi_m(i) &= \int_{\mathbb{R}^d} E(\psi_m(i)|X_i = x)p(x)dx \\ &\rightarrow \int_{\mathbb{R}^d} (-\log q(x) - \log c_1(d) - \gamma)p(x)dx, \end{aligned} \quad (25)$$

which can be proven by condition (14) and Reference [24].

Using the same approach and the conditions (12) and (13), we can obtain that

$$\lim_{n \rightarrow \infty} E\zeta_n(i) = \int_{\mathbb{R}^d} (-\log p(x) - \log c_1(d) - \gamma)p(x)dx. \quad (26)$$

Combining (25) and (26), we have

$$\lim_{n, m \rightarrow \infty} E[\psi_m(i) - \zeta_n(i)] = \int_{\mathbb{R}^d} p(x) \log \frac{p(x)}{q(x)} dx. \quad \square$$

Theorem 2 shows that the NN estimator (11) is consistent in  $L^2$ .

*Theorem 2:* Suppose that the probability density functions  $p, q$  satisfy the following conditions:

$$\int_{\mathbb{R}^d} |\log p(x)|^{2+\epsilon} p(x) dx < \infty, \quad (27)$$

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\log \|x - y\||^{2+\epsilon} p(x)p(y) dx dy < \infty, \quad (28)$$

$$\int_{\mathbb{R}^d} |\log q(x)|^{2+\epsilon} p(x) dx < \infty, \quad (29)$$

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\log \|x - y\||^{2+\epsilon} p(x)q(y) dy < \infty, \quad (30)$$

for some  $\epsilon > 0$ . Then

$$\lim_{n, m \rightarrow \infty} E \left( \hat{D}_{n,m}(p||q) - D(p||q) \right)^2 = 0. \quad (31)$$

*Proof:* By the triangle inequality, we have

$$\begin{aligned} & \sqrt{E \left( \hat{D}_{n,m}(p||q) - D(p||q) \right)^2} \\ & \leq \sqrt{E \left( \hat{D}_{n,m}(p||q) - E\hat{D}_{n,m}(p||q) \right)^2} \\ & \quad + \sqrt{E \left( E\hat{D}_{n,m}(p||q) - D(p||q) \right)^2}. \end{aligned} \quad (32)$$

Theorem 1 implies that the second term on the right hand side of (32) will vanish as  $n, m$  increase. Thus it suffices to show that  $\text{Var} \hat{D}_{n,m}(p||q) \rightarrow 0$  with  $n, m \rightarrow \infty$ . By the alternative expression (17) for  $\hat{D}_{n,m}(p||q)$ ,  $\text{Var} \hat{D}_{n,m}$  can be written in terms of  $\zeta_n(i)$  and  $\psi_m(i)$ , i.e.

$$\begin{aligned} & \text{Var} \hat{D}_{n,m} \\ &= \left[ \sum_{i=1}^n \text{Var} \zeta_n(i) + \sum_{i \neq j} \text{Cov}(\zeta_n(i), \zeta_n(j)) + \sum_{i=1}^n \text{Var} \psi_m(i) + \sum_{i \neq j} \text{Cov}(\psi_m(i), \psi_m(j)) - \sum_{i,j} \text{Cov}(\zeta_n(i), \psi_m(j)) \right] / n^2 \\ &= \text{Var} \zeta_n(i)/n + \sum_{i \neq j} \text{Cov}(\zeta_n(i), \zeta_n(j))/n^2 \\ & \quad + \text{Var} \psi_m(i)/n + \sum_{i \neq j} \text{Cov}(\psi_m(i), \psi_m(j))/n^2 \\ & \quad - \text{Cov}(\zeta_n(i), \psi_m(i))/n - \sum_{i \neq j} \text{Cov}(\zeta_n(i), \psi_m(j))/n^2. \end{aligned} \quad (33)$$

As shown in [21], the first two terms go to zero as  $n \rightarrow \infty$ , given (27) and (28). Following the same line of proof, the next two terms can also be shown to diminish as  $n, m$  increase when conditions (29) and (30) are satisfied. Now let us consider the last two terms.

Since the samples from  $p$  and  $q$  are assumed to be independent,  $\zeta_n(i)$  and  $\psi_m(j)$  are independent given  $X_i$  and  $X_j$  ( $i$  can be equal to  $j$ ). We have

$$\begin{aligned} & E\{\zeta_n(i)\psi_m(j)|X_i = x_i, X_j = x_j\} \\ &= E\{\zeta_n(i)|X_i = x_i, X_j = x_j\} E\{\psi_m(j)|X_j = x_j\}, \end{aligned} \quad (34)$$

where

$$\lim_{n \rightarrow \infty} E\{\zeta_n(i)|X_i = x_i, X_j = x_j\} = -\log(p(x_i)c_1(d)e^\gamma) \quad (35)$$

$$\lim_{m \rightarrow \infty} E\{\psi_m(j)|X_j = x_j\} = -\log(q(x_j)c_1(d)e^\gamma). \quad (36)$$

(35) is intuitive since the influence of  $X_j = x_j$  will be wiped out as  $n \rightarrow \infty$  and (36) is a result from the proof of Theorem 1.

If  $i \neq j$ , we have  $\lim_{n,m \rightarrow \infty} E[\zeta_n(i)\psi_m(j)] =$

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \log(p(x_i)c_1(d)e^\gamma) \log(q(x_j)c_1(d)e^\gamma) p(x_i)p(x_j) dx_i dx_j,$$

which is equal to  $\lim_{n,m \rightarrow \infty} E\zeta_n(i)E\psi_m(j)$ . Namely,  $\lim_{n,m \rightarrow \infty} \text{Cov}(\zeta_n(i), \psi_m(j)) = 0$  if  $i \neq j$ .

If  $i = j$ , by conditions (27) and (29), we obtain  $\int_{\mathbb{R}^d} |\log p(x)|^{1+\epsilon/2} |\log q(x)|^{1+\epsilon/2} p(x) dx < \infty$  for some  $\epsilon > 0$ .

Thus,  $\lim_{n,m \rightarrow \infty} E[\zeta_n(i)\psi_m(i)] < \infty$ , which implies  $\lim_{n,m \rightarrow \infty} \text{Cov}(\zeta_n(i), \psi_m(i)) < \infty$ , since we already have (25) and (26). Therefore, the last two terms of (33) are guaranteed to vanish as  $n, m \rightarrow \infty$ .  $\square$

#### IV. EXPERIMENTS

The advantage of the NN divergence estimator is that it is more easily generalized and implemented for higher-dimensional data as compared to our previous algorithms via data-dependent partitions [12]. However, the NN method also suffers from the curse of high dimensions, in a way different from methods based on partitioning. In [26], Hinneburg *et al.* noted that nearest neighbor search would become unreliable in a high dimensional space due to the sparsity of the data objects. Their work put forward a new notion of nearest neighbor search, which does not treat all dimensions equally but uses a quality criterion to select relevant dimensions with respect to a given query. Another disadvantage of the NN method is that finding the nearest neighbor is a very time-consuming process, particularly for large sample sizes. The problem of designing an efficient algorithm for nearest neighbor searching has been investigated in the literature. Fukunaga and Narendra [27] presented a branch and bound algorithm for computing  $k$ -nearest neighbors. Another procedure based on  $k$ -d tree was proposed by Bentley [28] at about the same time. Since then, there have been a number of improvements and variants of these algorithms. By using these procedures, we can significantly reduce the running time for the computation of the nearest neighbor.

The following experiments are performed on simulated data to compare the NN method with Algorithm G from [13]. Recall that Algorithm G combines locally adaptive partitioning and finite-sample-size error correction (Algorithm C and Algorithm E respectively in [12]). The curves in all the figures show the estimation average of 25 independent runs. Also  $n$  and  $m$  are equal in the experiments.

Figure 1 shows the case with two exponential distributions. The NN method exhibits better convergence than Algorithm G as sample sizes increase. In general, for scalar distributions, Algorithm G suffers from relatively higher bias even when a large number of samples are available. The NN method has higher variance for small sample sizes, but as sample sizes increase, the variance decreases quickly.

In Figure 2, we have two 4-dimensional Gaussian distributions with different means and different covariance matrices. The NN estimator converges very quickly to the actual divergence whereas Algorithm G is biased downwards and takes a lot more samples to converge to the actual divergence.

In Figure 3, both distributions are 10-dimensional Gaussian with equal means but different covariance matrices. The estimates by the NN method are closer to the true value, whereas Algorithm G is seriously under-estimating.

In Figure 4, we have two identical distributions in  $\mathbb{R}^{20}$ . The NN method outperforms Algorithm G, which has a very large upward bias. Note that in the experiments on high-dimensional data, the NN estimator suffers from a larger estimation variance compared to Algorithm G, though the estimation variance shrinks with sample sizes.

In summary, the divergence estimator using the NN distances can be more efficient than partitioning-based methods, especially in high-dimensional cases when the number of samples is limited.

#### REFERENCES

- [1] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
- [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, 1991.
- [3] P. K. Bhattacharya, "Efficient estimation of a shift parameter from grouped data," *The Annals of Mathematical Statistics*, vol. 38, no. 6, pp. 1770–1787, Dec., 1967.
- [4] D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri, "Information-theoretic analysis of neural coding," *Journal of Computational Neuroscience*, vol. 10, pp. 47–69, 2001.
- [5] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi, "An information-theoretic approach to detecting changes in multi-dimensional data streams," to appear in *Proceedings of the 38th Symposium on the Interface of Statistics, Computing Science, and Applications (Interface '06)*, Pasadena, CA, May, 2006.
- [6] B. Krishnamurthy, H. V. Madhyastha, and S. Venkatasubramanian, "On stationarity in Internet measurements through an information-theoretic lens," *Proc. 1st IEEE Workshop on Networking and Database*, 2005.
- [7] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. J. Rubio, "A new Kullback-Leibler VAD for speech recognition in noise," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 266–269, February 2004.
- [8] J. R. Mathiassen, A. Skavhaug, and K. Bø, "Texture similarity measure using Kullback-Leibler divergence between Gamma distributions," *Proceedings of the 7th European Conference on Computer Vision-Part III*, pp. 133–147, 2002.
- [9] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," *Proceedings of Ninth IEEE International Conference on Computer Vision*, vol. 1, pp. 487–493, 13-16 October 2003.
- [10] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1265–1287, March, 2003.
- [11] C. Liu, and H-Y Shum, "Kullback-Leibler boosting," *Proc. of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 587–594, June, 2003.
- [12] Q. Wang, S. R. Kulkarni and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3064–3074, September 2005.
- [13] Q. Wang, S. R. Kulkarni and S. Verdú, "On bias reduction for divergence estimation of continuous distributions", Report, September 2005, available at <http://www.princeton.edu/~qingwang/bias-reduction-WKVsep2005.pdf>.
- [14] J. D. Victor, "Binless strategies for estimation of information from neural data," *Physical Review E*, **66**, 051903, 2002.

[15] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination: Consistency properties," Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX, USA, 1951.

[16] T. M. Cover, P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, January 1967.

[17] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[18] D. O. Loftsgaarden, C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *The Annals of Mathematical Statistics*, vol. 36, no. 3, pp. 1049-1051, June 1965.

[19] L. Devroye, L. Györfi, A. Krzyzak, and G. Lugosi, "On the strong universal consistency of nearest neighbor regression function estimates," *The Annals of Statistics*, vol. 22, no. 3, pp. 1371-1385, September 1985.

[20] S. R. Kulkarni, S. E. Posner, S. Sandilya, "Data-dependent  $k_n$ -NN and kernel estimators consistent for arbitrary processes," *IEEE Transactions on Information Theory*, vol. 48, no. 10, pp. 2785-2788, October 2002.

[21] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problems Inform. Transmission*, 23, pp. 95-101, 1987.

[22] A. B. Tsybakov and E. C. van der Meulen, "Root- $n$  consistent estimators of entropy for densities with unbounded support," *Scand. J. Statist.*, 23, 75-83, 1996.

[23] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, 69, 066138, 2004.

[24] M. Loève, *Probability Theory 4th Edition*, Springer-Verlag, 1977.

[25] W. Feller, *An Introduction to Probability Theory and Its Applications, 3rd Edition*, John Wiley & Sons, Inc., 1970.

[26] A. Hinneburg, C. C. Aggarwal and D. A. Keim, "What is the nearest neighbor in high dimensional spaces?" *Proceedings of the 26th VLDB Conference*, Cairo, Egypt, 2000.

[27] K. Fukunaga and P. M. Nerada, "A branch and bound algorithm for computing  $k$ -nearest neighbors," *IEEE Transactions on Computers*, vol. 24, pp. 750-753, July 1975.

[28] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509-517, 1975.

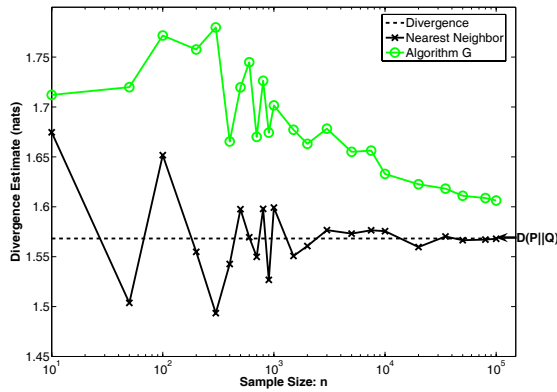


Fig. 1.  $X \sim P = \text{Exp}(1), Y \sim Q = \text{Exp}(12), D(P||Q) = 1.5682$  nats.

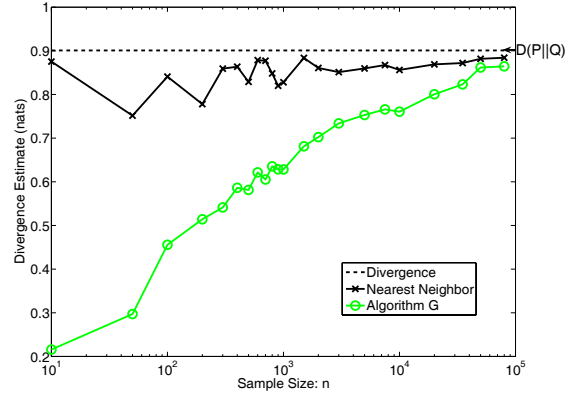


Fig. 2.  $X \sim P = \text{Gaussian}(\mu^P, \mathbf{C}^P), Y \sim Q = \text{Gaussian}(\mu^Q, \mathbf{C}^Q); \dim = 4; \mu^P = [.1 \ .3 \ .6 \ .9]^T, \mu^Q = [0 \ 0 \ 0 \ 0]^T, \mathbf{C}_{\ell,\ell}^P = 1, \mathbf{C}_{\ell,s}^P = 0.5, \mathbf{C}_{\ell,\ell}^Q = 1, \mathbf{C}_{\ell,s}^Q = 0.1, \text{ for } \ell = 1, \dots, 4, s = 1, \dots, 4, \ell \neq s; D(P||Q) = 0.9009$  nats.

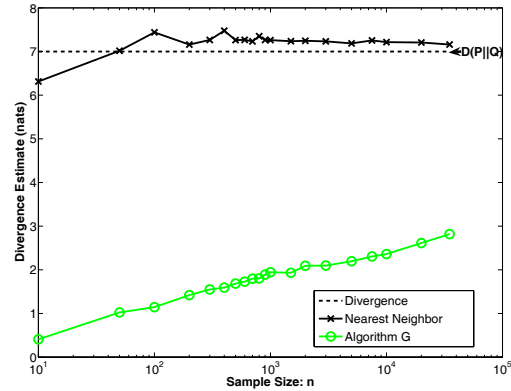


Fig. 3.  $X \sim P = \text{Gaussian}(\mu^P, \mathbf{C}^P), Y \sim Q = \text{Gaussian}(\mu^Q, \mathbf{C}^Q); \dim = 10; \mu_\ell^P = \mu_\ell^Q = 0, \mathbf{C}_{\ell,\ell}^P = 1, \mathbf{C}_{\ell,s}^P = 0.9, \mathbf{C}_{\ell,\ell}^Q = 1, \mathbf{C}_{\ell,s}^Q = 0.1, \text{ for } \ell = 1, \dots, 10, s = 1, \dots, 10, \ell \neq s; D(P||Q) = 6.9990$  nats.

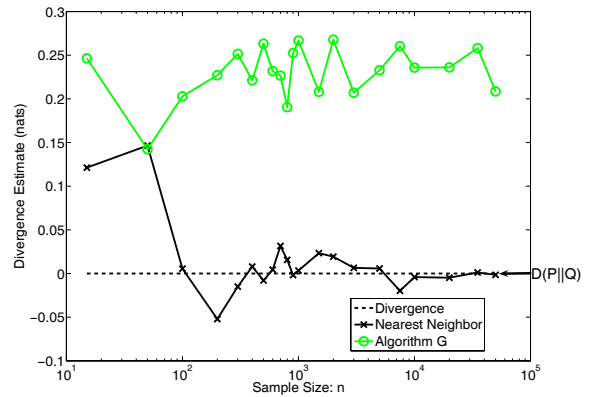


Fig. 4.  $X \sim P = \text{Gaussian}(\mu^P, \mathbf{C}^P), Y \sim Q = \text{Gaussian}(\mu^Q, \mathbf{C}^Q); \dim = 20; \mu_\ell^P = \mu_\ell^Q = 0, \mathbf{C}_{\ell,\ell}^P = \mathbf{C}_{\ell,\ell}^Q = 1, \mathbf{C}_{\ell,s}^P = \mathbf{C}_{\ell,s}^Q = 0.2, \text{ for } \ell = 1, \dots, 20, s = 1, \dots, 20, \ell \neq s; D(P||Q) = 0$  nats.