

Estimation of entropy-type integral functionals

D. Källberg^{a,*}, O. Seleznev^b

^a *Department of Mathematics and Mathematical Statistics, Umeå University, SE-901 87 Umeå, Sweden*

^b *Department of Mathematics and Mathematical Statistics, Umeå University, SE-901 87 Umeå, Sweden*

Abstract

Integrated powers of densities of one- or two-multidimensional random variables appear in a variety of problems in mathematical statistics, information theory, and computer science. We study U -statistic estimators for a class of such integral functionals based on the ϵ -close vector observations in the corresponding independent and identically distributed samples. We show some asymptotic properties of these estimators (e.g., consistency and asymptotic normality). The results can be used in a variety of problems in mathematical statistics and computer science (e.g., distribution identification problems, approximate matching for random databases, two-sample problems).

Keywords: U -statistics, estimation of divergence, density power divergence, asymptotic normality, entropy estimation, Rényi entropy

1. Introduction

Let the distributions \mathcal{P}_X and \mathcal{P}_Y of the d -dimensional random variables X and Y have densities $p_X(x)$ and $p_Y(x)$, $x \in R^d$, respectively. Various characteristics in mathematical statistics, information theory, and computer science, say *entropy-type integral functionals*, are expressed in terms of integrated powers of $p_X(x)$ and $p_Y(x)$. For example, a widely accepted measure of closeness between \mathcal{P}_X and \mathcal{P}_Y is the (quadratic) density power divergence (Basu et al., 1998)

$$D_2 = D_2(\mathcal{P}_X, \mathcal{P}_Y) := \int_{R^d} (p_X(x) - p_Y(x))^2 dx.$$

Other examples include the Rényi entropy for quantifying uncertainty in \mathcal{P}_X (Rényi, 1970)

$$h_s = h_s(\mathcal{P}_X) := \frac{1}{1-s} \log \left(\int_{R^d} p_X(x)^s dx \right), \quad s \neq 1,$$

*Corresponding author

Email addresses: david.kallberg@math.umu.se (D. Källberg), oleg.seleznev@math.umu.se (O. Seleznev)

and the differential variability for some database problems (Seleznev and Thalheim, 2010)

$$v = v(\mathcal{P}_X, \mathcal{P}_Y) := -\log\left(\int_{R^d} p_X(x)p_Y(x)dx\right).$$

Henceforth we use $\log x$ to denote the natural logarithm of x . For non-negative integers $k_1, k_2 \geq 0$, $\mathbf{k} := (k_1, k_2)$, we consider the Rényi entropy functionals (Källberg et al., 2012)

$$q_{\mathbf{k}} = q_{k_1, k_2} := \int_{R^d} p_X(x)^{k_1} p_Y(x)^{k_2} dx, \quad k_1 + k_2 \geq 2.$$

Moreover, given a set of constants $\mathbf{a} := \{a_0, a_1, a_2\}$, we introduce the related quadratic functionals

$$q_2 = q_2(\mathbf{a}) := a_0 q_{2,0} + a_1 q_{1,1} + a_2 q_{0,2}.$$

Note that the quadratic divergence $D_2 = q_{2,0} - 2q_{1,1} + q_{0,2}$, the Rényi entropy $h_k = \log(q_{k,0})/(1-k)$, $k = 2, 3, \dots$, and the variability $v = -\log(q_{1,1})$. Some applications of Rényi entropy and divergence measures can be found, e.g., in information theoretic learning (Principe, 2010). More applications of entropy and divergence in statistics (e.g., classification, distribution identification problems, and statistical inference), computer science (e.g., average case analysis for random databases, pattern recognition, and image matching), and econometrics are discussed, e.g., in Kapur (1989), Kapur and Kesavan (1992), Pardo (2006), Leonenko et al. (2008), Escolano et al. (2009), Seleznev and Thalheim (2003, 2010), Thalheim (2000), Leonenko and Seleznev (2010), Neemuchwala et al. (2005), and Ullah (1996). The divergence D_2 belongs to a subclass of the Bregman divergences that find various applications in statistics (see, e.g., Basseville, 2010, and references therein).

In this paper, to demonstrate the general approach, we study non-parametric estimation of some entropy-type integral functionals, e.g., $q_{\mathbf{k}}$ and $q_2(\mathbf{a})$, using independent samples from \mathcal{P}_X and \mathcal{P}_Y . Some new asymptotic results are presented for a class of U -statistic estimators for these functionals. These estimators are based on the ϵ -close observations in the corresponding samples. We generalize some results and techniques proposed in Leonenko and Seleznev (2010) and Källberg et al. (2012). In particular, we obtain consistency of the corresponding estimators for a more wide class of distributions and prove asymptotic normality of the estimators for the quadratic functionals $q_2(\mathbf{a})$.

Leonenko et al. (2008) study asymptotic properties of nearest-neighbor estimators for $q_{\mathbf{k}}$, and obtain consistency when the densities are bounded. Giné and Nickl (2008) show asymptotical normality for a kernel estimator of $q_{2,0}$ in the one-dimensional case. Ahmad and Cerrito (1993) and Li (1996, 1999) use kernel estimates of the quadratic divergence D_2 as test statistics for the two-sample problem, and obtain asymptotically normal null distribution. For a certain kernel estimator, we prove asymptotic normality under different conditions. The

number of small interpoint distances in a random sample is among the most studied examples of U -statistics with kernels varying with the sample size (see, e.g., Weber, 1983, Jammalamadaka and Janson, 1986, Penrose, 1995). A significant feature of such characteristics is that one can obtain normal limit laws even in some degenerate cases. We generalize this approach for two-sample statistics. This extension enables some statistical applications where the degeneracy condition might be crucial, e.g., estimation of divergence.

First we introduce some notation. Throughout the paper, we assume that the random vectors X and Y are independent. Let $d(x, y) := |x - y|$ denote the Euclidean distance in R^d and define $B_\epsilon(x) := \{y : d(x, y) < \epsilon\}$ to be an open ϵ -ball in R^d with center at x and radius ϵ . Denote by $b_\epsilon(d) := \epsilon^d b_1(d)$, $b_1(d) = 2\pi^{d/2}/(d\Gamma(d/2))$, the volume of the ϵ -ball. Define the ϵ -ball probability as

$$p_{X,\epsilon}(x) := P\{X \in B_\epsilon(x)\}.$$

We say that the vectors x and y are ϵ -close, if $d(x, y) < \epsilon$, for some $\epsilon > 0$. Let X_1, \dots, X_{n_1} and Y_1, \dots, Y_{n_2} be mutually independent samples of independent and identically distributed (i.i.d.) observations from \mathcal{P}_X and \mathcal{P}_Y , respectively. Define $\mathbf{n} := (n_1, n_2)$, $n := n_1 + n_2$, and say that $\mathbf{n} \rightarrow \infty$ if $n_1, n_2 \rightarrow \infty$. Let $\epsilon = \epsilon(\mathbf{n}) \rightarrow 0$ as $\mathbf{n} \rightarrow \infty$.

In what follows, we consider estimation problems for both one and two samples. However, in statements of results and the proofs it is assumed, for sake of space and clarity, that two samples are available, i.e., $n_1, n_2 > 0$. This can be done without loss of generality, because in the one-sample case, e.g., estimation of $q_{k_1,0}$, $k_1 \geq 2$, from X_1, \dots, X_{n_1} , an auxiliary sample Y_1, \dots, Y_{n_2} can be considered.

Denote by \xrightarrow{D} and \xrightarrow{P} convergence in distribution and probability, respectively. For a sequence of random variables $U_n, n \geq 1$, we write $U_n = O_P(1)$ as $n \rightarrow \infty$ if for any $\delta > 0$ and large enough $n \geq 1$, there exists $C > 0$ such that $P(|U_n| > C) \leq \delta$. Moreover, for a numerical sequence $w_n, n \geq 1$, let $U_n = O_P(w_n)$ as $n \rightarrow \infty$ if $U_n/w_n = O_P(1)$ as $n \rightarrow \infty$.

The remaining part of the paper is organized as follows. In Section 2, we consider estimation of the Rényi entropy functional $q_{\mathbf{k}}$. The asymptotic results for estimation of the quadratic functional $q_2(\mathbf{a})$ are given in Section 3. In Section 4, we discuss applications of the obtained results to estimation of density power divergence, the two-sample problem, and statistical inference for some entropy-type characteristics. Several numerical examples illustrate the rate of convergence of the asymptotic results. Section 5 contains the proofs of the statements from the previous sections.

2. Estimation of the Rényi entropy functional $q_{\mathbf{k}}$

We introduce the U -statistic estimators of $q_{\mathbf{k}}$ proposed by Källberg et al. (2012). If r is a non-negative integer, define $\mathcal{S}_{m,r}$ to be the set of all r -subsets of $\{1, \dots, m\}$. Let $S \in \mathcal{S}_{n_1, k_1}$ and $T \in \mathcal{S}_{n_2, k_2}$. When $k_1 \geq 1$, we define

$$\psi_{\mathbf{k}, \mathbf{n}, \epsilon}^{(i)}(S; T) = I(d(X_i, X_j) < \epsilon, d(X_i, Y_l) < \epsilon, \forall j \in S, \forall l \in T), \quad i \in S,$$

i.e., the indicator of the event that the observations $\{X_j, j \in S\}$ and $\{Y_l, l \in T\}$ are ϵ -close to X_i . By conditioning, we have

$$q_{\mathbf{k},\epsilon} := \mathbb{E}(\psi_{\mathbf{k},\mathbf{n}}^{(i)}(S; T)) = \mathbb{E}(p_{X,\epsilon}(X)^{k_1-1} p_{Y,\epsilon}(X)^{k_2}).$$

In a similar way, when $k_1 = 0$ and $k_2 \geq 1$, we define

$$\psi_{\mathbf{k},\mathbf{n},\epsilon}^{(i)}(T) = I(d(Y_i, Y_j) < \epsilon, \forall j \in T), \quad i \in T,$$

and

$$q_{\mathbf{k},\epsilon} := \mathbb{E}(\psi_{\mathbf{k},\mathbf{n}}^{(i)}(T)) = \mathbb{E}(p_{Y,\epsilon}(Y)^{k_2-1}).$$

Now, a U -statistic for $q_{\mathbf{k},\epsilon}$ (see, e.g., Ch. 2, Lee, 1990) is given by

$$Q_{\mathbf{k},\mathbf{n}} = Q_{\mathbf{k},\mathbf{n},\epsilon} := \binom{n_1}{k_1}^{-1} \binom{n_2}{k_2}^{-1} \sum_{S \in \mathcal{S}_{n_1, k_1}} \sum_{T \in \mathcal{S}_{n_2, k_2}} \psi_{\mathbf{k},\mathbf{n}}(S; T),$$

with the *kernel* $\psi_{\mathbf{k},\mathbf{n}}(S; T)$ defined by the symmetrization

$$\psi_{\mathbf{k},\mathbf{n}}(S; T) = \psi_{\mathbf{k},\mathbf{n},\epsilon}(S; T) := \begin{cases} \frac{1}{k_1} \sum_{i \in S} \psi_{\mathbf{k},\mathbf{n},\epsilon}^{(i)}(S; T), & \text{if } k_1 \geq 1, \\ \frac{1}{k_2} \sum_{i \in T} \psi_{\mathbf{k},\mathbf{n},\epsilon}^{(i)}(T), & \text{if } k_1 = 0, k_2 \geq 1. \end{cases}$$

By definition, $Q_{\mathbf{k},\mathbf{n}}$ is an unbiased estimator of $q_{\mathbf{k},\epsilon}$. Let $k := k_1 + k_2, k \geq 2$, and define the estimator of $q_{\mathbf{k}}$ as

$$\tilde{Q}_{\mathbf{k},\mathbf{n}} := Q_{\mathbf{k},\mathbf{n}}/b_\epsilon(d)^{k-1}.$$

Källberg et al. (2012) obtain consistency of $\tilde{Q}_{\mathbf{k},\mathbf{n}}$ when the densities are bounded and continuous. The following theorem yields weaker density conditions for consistency.

Theorem 1. *If $p_X, p_Y \in L_{2k-1}(R^d)$, $n\epsilon^{d(1-1/k)} \rightarrow \infty$, and $n_1/n \rightarrow \rho, 0 < \rho < 1$, then*

$$\tilde{Q}_{\mathbf{k},\mathbf{n}} \xrightarrow{P} q_{\mathbf{k}} \text{ as } \mathbf{n} \rightarrow \infty.$$

3. Estimation of the quadratic functional $q_2(\mathbf{a})$

The following linear combination is a sensible estimator for the quadratic functional q_2

$$\tilde{Q}_{2,\mathbf{n}} = \tilde{Q}_{2,\mathbf{n}}(\mathbf{a}) := a_0 \tilde{Q}_{2,0,\mathbf{n}} + a_1 \tilde{Q}_{1,1,\mathbf{n}} + a_2 \tilde{Q}_{0,2,\mathbf{n}}.$$

For $0 < \rho < 1$, introduce the characteristics

$$\begin{aligned} \kappa = \kappa_\rho(\mathbf{a}) &:= \frac{4}{\rho} \text{Var} \left(a_0 p_X(X) + \frac{a_1}{2} p_Y(X) \right) + \frac{4}{1-\rho} \text{Var} \left(a_2 p_Y(Y) + \frac{a_1}{2} p_X(Y) \right), \\ \eta = \eta_\rho(\mathbf{a}) &:= \frac{2}{b_1(d)} \left(\frac{a_0^2}{\rho^2} q_{2,0} + \frac{a_2^2}{(1-\rho)^2} q_{0,2} + \frac{a_1^2}{2\rho(1-\rho)} q_{1,1} \right). \end{aligned}$$

Let

$$\tilde{q}_{2,\epsilon} := \mathbb{E}(\tilde{Q}_{2,\mathbf{n}}) = a_0\tilde{q}_{2,0,\epsilon} + a_1\tilde{q}_{1,1,\epsilon} + a_2\tilde{q}_{0,2,\epsilon},$$

where $\tilde{q}_{\mathbf{k},\epsilon} := \mathbb{E}(\tilde{Q}_{\mathbf{k},\mathbf{n}}) = b_\epsilon(d)^{-1}q_{\mathbf{k},\epsilon}$. We get the following theorem for the asymptotic normality of $\tilde{Q}_{2,\mathbf{n}}$.

Theorem 2. *Let $p_X, p_Y \in L_3(\mathbb{R}^d)$, and $n_1/n = \rho, 0 < \rho < 1$.*

(i) *If $n\epsilon^d \rightarrow \beta, 0 < \beta < \infty$, then*

$$\sqrt{n}(\tilde{Q}_{2,\mathbf{n}} - \tilde{q}_{2,\epsilon}) \xrightarrow{D} N(0, \kappa + \eta/\beta) \text{ as } \mathbf{n} \rightarrow \infty.$$

(ii) *If $n\epsilon^d \rightarrow 0$ and $n^2\epsilon^d \rightarrow \infty$, then*

$$n\epsilon^{d/2}(\tilde{Q}_{2,\mathbf{n}} - \tilde{q}_{2,\epsilon}) \xrightarrow{D} N(0, \eta) \text{ as } \mathbf{n} \rightarrow \infty.$$

From a practical point of view, the unknown asymptotic variances in Theorem 2 have to be estimated, i.e., we need consistent estimators for κ and η . By expanding the terms in κ , we see that it is a function of ρ and the functionals $\{q_{i,j} : 2 \leq i + j \leq 3\}$, i.e., $\kappa = \kappa(\rho, \{q_{i,j} : 2 \leq i + j \leq 3\})$. Since Theorem 1 yields that $\{\tilde{Q}_{i,j,\mathbf{n}} : 2 \leq i + j \leq 3\}$ are consistent estimators of these functionals, we set up the plug-in estimator

$$\kappa_{\mathbf{n}} := \kappa(\rho_{\mathbf{n}}, \{\tilde{Q}_{i,j,\mathbf{n}} : 2 \leq i + j \leq 3\})$$

for κ , where $\rho_{\mathbf{n}} := n_1/n$. Similarly, denote by $\eta_{\mathbf{n}}$ the corresponding estimator for η and define $\nu_{\mathbf{n}} := \kappa_{\mathbf{n}} + \eta_{\mathbf{n}}/\beta_{\mathbf{n}}, \beta_{\mathbf{n}} := n\epsilon^d$, to be an estimator of $\kappa + \eta/\beta$. It follows from the Slutsky theorem that $\eta_{\mathbf{n}}$ and $\nu_{\mathbf{n}}$ are consistent estimators of η and $\kappa + \eta/\beta$, respectively.

To ensure a sufficient rate of decay for the bias term $\tilde{q}_{2,\epsilon} - q_2$, we propose smoothness conditions for the densities. Denote by $H_2^{(\alpha)}(K), 0 < \alpha \leq 1, K > 0$, a linear space of functions in $L_5(\mathbb{R}^d)$ that satisfy a α -Hölder condition in L_2 -norm with constant K , i.e., if $p \in H_2^{(\alpha)}(K)$ and $h \in B_1(d)$, then

$$\|p(\cdot + h) - p(\cdot)\|_2 \leq K|h|^\alpha. \quad (1)$$

Note that (1) holds if, e.g., for some function $g \in L_2(\mathbb{R}^d)$,

$$|p(x + h) - p(x)| \leq g(x)|h|^\alpha.$$

The density smoothness can be introduced in different ways, e.g., by the point-wise Hölder conditions (Källberg et al., 2012) or the Fourier characterization (Giné and Nickl, 2008).

A bound for the bias and the rate of convergence in probability are presented in the following theorem. Additionally, we obtain asymptotically pivotal quantities which can be used, e.g., to construct asymptotic confidence intervals for the functional q_2 . Let $L(n), n \geq 1$, be a slowly varying function as $n \rightarrow \infty$.

Theorem 3. Let $p_X, p_Y \in H_2^{(\alpha)}(K)$ and $n_1/n = \rho, 0 < \rho < 1$.

(i) Then for the bias, we have

$$|\tilde{q}_{2,\epsilon} - q_2| \leq C\epsilon^{2\alpha}, C > 0.$$

(ii) If $0 < \alpha \leq d/4$ and $\epsilon \sim cn^{-1/(2\alpha+d/2)}, c > 0$, then

$$\tilde{Q}_{2,\mathbf{n}} - q_2 = O_P(n^{-2\alpha/(2\alpha+d/2)}) \text{ as } \mathbf{n} \rightarrow \infty.$$

(iii) If $\alpha > d/4$ and $n\epsilon^d \rightarrow \beta, 0 < \beta < \infty$, then

$$\sqrt{n}(\tilde{Q}_{2,\mathbf{n}} - q_2) \xrightarrow{D} N(0, \kappa + \eta/\beta) \quad \text{and} \quad \frac{\sqrt{n}}{\sqrt{v_{\mathbf{n}}}}(\tilde{Q}_{2,\mathbf{n}} - q_2) \xrightarrow{D} N(0, 1)$$

as $\mathbf{n} \rightarrow \infty$.

(iv) If $\epsilon \sim L(n)n^{-2/d}$ and $L(n) \rightarrow \infty$, i.e., $n^2\epsilon^d \rightarrow \infty$, then

$$n\epsilon^{d/2}(\tilde{Q}_{2,\mathbf{n}} - q_2) \xrightarrow{D} N(0, \eta) \quad \text{and} \quad \frac{n\epsilon^{d/2}}{\sqrt{\eta_{\mathbf{n}}}}(\tilde{Q}_{2,\mathbf{n}} - q_2) \xrightarrow{D} N(0, 1)$$

as $\mathbf{n} \rightarrow \infty$.

Remark 1. It is worth noting that in this paper, we do not require $\kappa > 0$ (cf. Källberg et al, 2012). The condition $\kappa > 0$ corresponds to the non-degeneracy condition commonly used for proving asymptotic normality of U -statistics by the conventional techniques, e.g., using the H -decomposition (see, e.g., Lee, 1990, Koroljuk and Borovskich, 1994). For example, when considering, e.g., the divergence D_2 , the condition $\kappa > 0$ implies that $p_X(x) \neq p_Y(x)$ on a set of positive measure. This assumption may be too restrictive in some statistical applications whenever the distributions of X and Y are too close.

Remark 2. In Theorem 3(iv), we get asymptotic normality for an arbitrary dimension. This is an improvement of some results in Leonenko and Seleznev (2010) and Källberg et al. (2012). Note, however, that the rate of convergence $n\epsilon^{d/2} \sim L(n)^{d/2}$ can be slower than \sqrt{n} in this case.

Remark 3. The condition $n_1/n = \rho, 0 < \rho < 1$, in Theorems 2 and 3 is technical and we claim that it can be replaced with the slightly weaker condition $n_1/n \rightarrow \rho, 0 < \rho < 1$.

Remark 4. In the one-sample case, the results in Theorems 2 and 3 are essentially independent of ρ . In fact, consider, e.g., the estimator $\tilde{Q}_{2,0,\mathbf{n}}$ of $q_{2,0}$, i.e., $\mathbf{a} = \{1, 0, 0\}$, $\kappa = 4\text{Var}(p_X(X))/\rho$, and $\eta = 2b_1(d)^{-1}q_{2,0}/\rho^2$. We have $n = n_1/\rho$, so if $n_1\epsilon^d \rightarrow \lambda, 0 < \lambda < \infty$, then $n\epsilon^d \rightarrow \lambda/\rho =: \beta$. Hence, it follows from Theorem 3(iii) that

$$\sqrt{n_1}(\tilde{Q}_{2,0,\mathbf{n}} - q_{2,0}) \xrightarrow{D} N\left(0, 4\text{Var}(p_X(X)) + \frac{2}{b_1(d)}q_{2,0}/\lambda\right) \text{ as } n_1 \rightarrow \infty.$$

Therefore, we obtain a result with $\sqrt{n_1}$ -scaling that does not depend on ρ as desired. A similar modification can be done for the $n\epsilon^{d/2}$ -scaling.

4. Applications

4.1. Estimation of density power divergence

The introduced quadratic divergence D_2 belongs to the wide class of density power divergences (Basu et al., 1998), defined by

$$\begin{aligned} D_s &= D_s(\mathcal{P}_X, \mathcal{P}_Y) := \int_{R^d} \left(\frac{1}{s-1} p_X(x)^s - \frac{s}{s-1} p_X(x) p_Y(x)^{s-1} + p_Y(x)^s \right) dx \\ &= \frac{1}{s-1} q_{s,0} - \frac{s}{s-1} q_{1,s-1} + q_{0,s}, \quad s > 1. \end{aligned}$$

When $s = r$ is a non-negative integer and $p_X, p_Y \in L_{2r-1}(R^d)$, then Theorem 1 implies that

$$\hat{D}_{r,\mathbf{n}} := \frac{1}{r-1} \tilde{Q}_{r,0,\mathbf{n}} - \frac{r}{r-1} \tilde{Q}_{1,r-1,\mathbf{n}} + \tilde{Q}_{0,r,\mathbf{n}}$$

is a consistent estimator of D_r . Moreover, Theorem 3 gives conditions for asymptotic normality of the quadratic estimator $\hat{D}_{2,\mathbf{n}}$.

The quadratic divergence D_2 can be used as a dissimilarity measure to investigate pairwise differences among M populations or objects. Let the features of population l be represented by the random feature vector V_l with density $p_{V_l}(x), x \in R^d, l = 1, \dots, M$. Using independent samples from populations $V_l, l = 1, \dots, M$, we apply, e.g., the Bonferroni method and calculate the $\binom{M}{2}$ approximate simultaneous confidence intervals $\{I_{lm}\}$ for the quadratic divergences $\{D_{2,l,m}\}, D_{2,l,m} := D_2(\mathcal{P}_{V_l}, \mathcal{P}_{V_m})$, for a given confidence level. Now, the intervals $\{I_{lm}\}$ can be used to determine which populations are different with respect to their feature densities.

Example 1. We consider estimation of the quadratic density power divergence $D_2(\mathcal{P}_X, \mathcal{P}_Y)$ between two three-dimensional distributions. The distribution of the components of X and Y are $t(3)$ -i.i.d. and $N(1,1)$ -i.i.d., respectively. In this case it holds that $D_2 \approx 0.018$. We simulate $N_{sim} = 500$ independent and normalized residuals $R_{\mathbf{n}}^{(i)} := \sqrt{n}(\hat{D}_{2,\mathbf{n}} - D_2)/\sqrt{\nu_{\mathbf{n}}}, i = 1, \dots, N_{sim}$, with $n_1 = n_2 = 500$, and $\epsilon = 1/4$. Figure 1 illustrates the performance of the normal approximation of $R_{\mathbf{n}}^{(i)}$ indicated by Theorem 3. The histogram, normal quantile plot, and p-value (0.41) for the Kolmogorov-Smirnov test also support the assumption of standard normality for the residuals.

4.2. The two-sample problem

A general null hypothesis of closeness between \mathcal{P}_X and \mathcal{P}_Y is given by

$$H_0 : p_X(x) = p_Y(x) \text{ a.e.}$$

We consider the problem of testing H_0 against the alternative H_1 that $p_X(x)$ and $p_Y(x)$ differ on a set of positive measure (often referred to as the two-sample problem). Note that the alternative can be written as $H_1 : D_2 > 0$. Hence, we

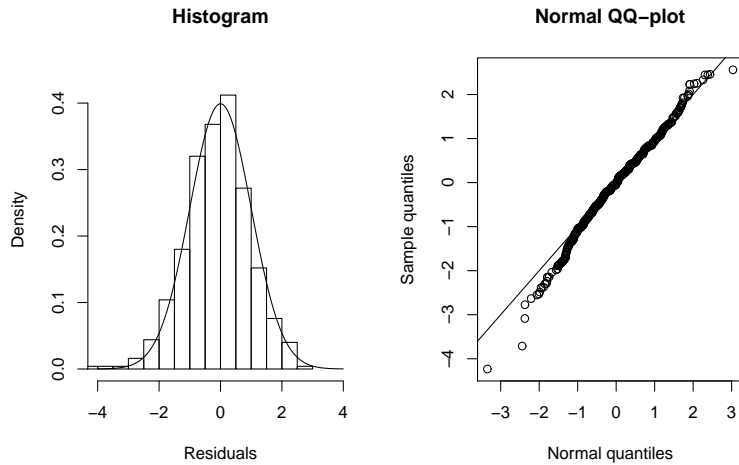


Figure 1: Three-dimensional distributions with $t(3)$ -i.i.d. and $N(1, 1)$ -i.i.d. components, respectively; sample sizes $n_1 = n_2 = 500$ and $\epsilon = 1/4$. Standard normal approximation for the normalized residuals; $N_{sim} = 500$.

define a test statistic based on the estimator $\hat{D}_{2,\mathbf{n}}$ for D_2 (see, e.g., Li, 1996), according to

$$T_{\mathbf{n}} := \frac{n\epsilon^{d/2}}{\sqrt{\eta_{\mathbf{n}}}} \hat{D}_{2,\mathbf{n}}.$$

The asymptotics for the distribution of $T_{\mathbf{n}}$ are presented in the following proposition. Let $\{u_{\mathbf{n}}\}$ be a numerical sequence such that $u_{\mathbf{n}} = o(n\epsilon^{d/2})$ as $\mathbf{n} \rightarrow \infty$.

Proposition 4. *Assume that $p_X, p_Y \in L_3(\mathbb{R}^d)$, $n^2\epsilon^d \rightarrow \infty$, and $n_1/n = \rho$.*

(i) *Under H_0 , we have*

$$n\epsilon^{d/2} \hat{D}_{2,\mathbf{n}} \xrightarrow{D} N(0, \eta) \quad \text{and} \quad T_{\mathbf{n}} \xrightarrow{D} N(0, 1) \quad \text{as } \mathbf{n} \rightarrow \infty.$$

(ii) *Under H_1 , we have*

$$P(T_{\mathbf{n}} > u_{\mathbf{n}}) \rightarrow 1 \quad \text{as } \mathbf{n} \rightarrow \infty.$$

Thus, we reject H_0 if $T_{\mathbf{n}} > \lambda_{\alpha}$, where λ_{α} is the α -quantile of the standard normal distribution. It follows that this test has asymptotic significance level α and is consistent against all alternatives that satisfy $p_X, p_Y \in L_3(\mathbb{R}^d)$.

Remark 5. Since $q_{2,0} = q_{1,1} = q_{0,2}$ under H_0 , the asymptotic variance in Proposition 4(i) is reduced to

$$\eta \stackrel{H_0}{=} \eta_0 := \frac{2q_{2,0}}{b_1(d)\rho^2(1-\rho)^2}.$$

Therefore, the test might be more accurate if $T_{\mathbf{n}}$ is redefined by means of replacing $\eta_{\mathbf{n}}$ with an estimate of η_0 based on the pooled sample $\{Z_1, \dots, Z_n\} := \{X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}\}$ (cf. Li, 1999).

4.3. Estimation of Rényi entropy and differential variability

Consider the class of functionals

$$h_{\mathbf{k}} = h_{\mathbf{k}}(\mathcal{P}_X, \mathcal{P}_Y) := \frac{1}{1-k} \log(q_{\mathbf{k}}), \quad k \geq 2.$$

When $\mathcal{P}_X = \mathcal{P}_Y$, we get the Rényi entropy $h_{k,0}$, a family of functions for measuring uncertainty or randomness of a system (Rényi, 1970). Another important example is the differential variability $v = h_{1,1}$, a characteristic for modeling some random databases (Seleznev and Thalheim, 2010). When the densities are bounded and continuous, the results in Källberg et al. (2012) imply consistency of the truncated plug-in estimator

$$H_{\mathbf{k},\mathbf{n}} := \frac{1}{1-k} \log(\max(\tilde{Q}_{\mathbf{k},\mathbf{n}}, 1/n))$$

for $h_{\mathbf{k}}$. It follows from Theorem 1 and the Slutsky theorem that $H_{\mathbf{k},\mathbf{n}}$ is consistent under the weaker condition $p_X, p_Y \in L_{2k-1}(R^d)$.

In the quadratic cases $k = 2$, i.e., $\mathbf{k} = (2, 0), (1, 1)$, the asymptotic normality properties of $H_{\mathbf{k},\mathbf{n}}$ are studied by Leonenko and Seleznev (2010) and Källberg et al. (2012). The following proposition generalizes some of these results (see also Remarks 1-4).

Proposition 5. *Assume that $k = 2$. Let $p_X, p_Y \in H_2^{(\alpha)}(K)$ and $n_1/n = \rho$, $0 < \rho < 1$.*

(i) *If $0 < \alpha \leq d/4$ and $\epsilon \sim cn^{-1/(2\alpha+d/2)}$, $c > 0$, then*

$$H_{\mathbf{k},\mathbf{n}} - h_{\mathbf{k}} = \text{O}_{\text{P}}(n^{-2\alpha/(2\alpha+d/2)}) \text{ as } \mathbf{n} \rightarrow \infty.$$

(ii) *If $\alpha > d/4$ and $n\epsilon^d \rightarrow \beta$, $0 < \beta < \infty$, then*

$$\sqrt{n}(H_{\mathbf{k},\mathbf{n}} - h_{\mathbf{k}}) \xrightarrow{\text{D}} N(0, \kappa + \eta/\beta) \quad \text{and} \quad \sqrt{n} \frac{\tilde{Q}_{\mathbf{k},\mathbf{n}}}{\sqrt{\nu_{\mathbf{n}}}}(H_{\mathbf{k},\mathbf{n}} - h_{\mathbf{k}}) \xrightarrow{\text{D}} N(0, 1)$$

as $\mathbf{n} \rightarrow \infty$.

(iii) *If $\epsilon \sim L(n)n^{-2/d}$ and $L(n) \rightarrow \infty$, i.e., $n^2\epsilon^d \rightarrow \infty$, then*

$$n\epsilon^{d/2}(H_{\mathbf{k},\mathbf{n}} - h_{\mathbf{k}}) \xrightarrow{\text{D}} N(0, \eta) \quad \text{and} \quad n\epsilon^{d/2} \frac{\tilde{Q}_{\mathbf{k},\mathbf{n}}}{\sqrt{\eta_{\mathbf{n}}}}(H_{\mathbf{k},\mathbf{n}} - h_{\mathbf{k}}) \xrightarrow{\text{D}} N(0, 1)$$

as $\mathbf{n} \rightarrow \infty$.

The estimator $H_{\mathbf{k},\mathbf{n}}$ of $h_{\mathbf{k}}$ can be used, e.g., for distribution identification problems and approximate matching in stochastic databases (for a description, see Källberg et al., 2012).

Example 2. Let X and Y be one-dimensional uniform random variables, i.e., $X \sim U(0,1)$ and $Y \sim U(0,\sqrt{2})$, and consider estimation of the differential variability $v = h_{1,1} = \log(2)/2$. We simulate independent and normalized residuals $R_{\mathbf{n}}^{(i)} := \sqrt{n}\tilde{Q}_{1,1,\mathbf{n}}(H_{1,1,\mathbf{n}} - h_{1,1})/\sqrt{v_{\mathbf{n}}}$, $i = 1, \dots, N_{sim}$, with $n_1 = n_2 = 300$, $\epsilon = 1/100$, and $N_{sim} = 600$. Figure 2 illustrates the normal approximation for these residuals indicated by Proposition 5(ii). The histogram, quantile plot, and p-value (0.36) for the Kolmogorov-Smirnov test allow to accept the hypothesis of standard normality.

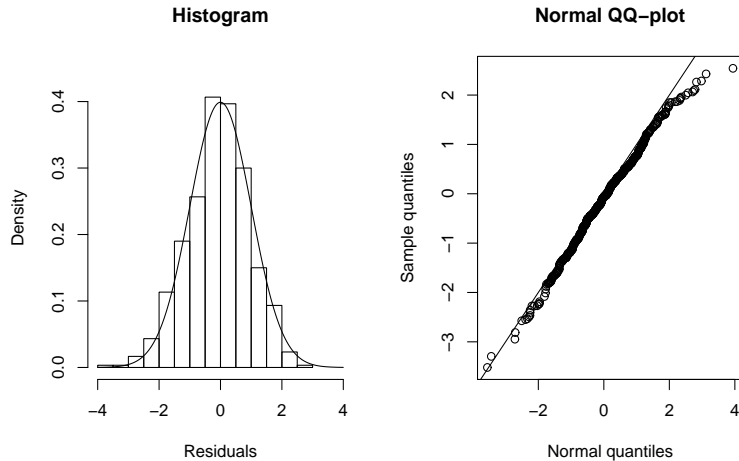


Figure 2: Uniform distributions, $U(0,1)$ and $U(0,\sqrt{2})$; sample sizes $n_1 = n_2 = 300$ and $\epsilon = 1/100$. Standard normal approximation for the normalized residuals; $N_{sim} = 600$.

5. Proofs

The following lemma is used in the subsequent proofs.

Lemma 1. For $a, b \geq 0$, assume that $p_X, p_Y \in L_{a+b+1}(R^d)$. Then

$$b_{\epsilon}(d)^{-(a+b)}\mathbb{E}(p_{X,\epsilon}(X)^a p_{Y,\epsilon}(X)^b) \rightarrow q_{a+1,b} \text{ as } \epsilon \rightarrow 0.$$

Proof. Let $\tilde{p}_{X,\epsilon}(x) := b_{\epsilon}(d)^{-1}p_{X,\epsilon}(x)$, $\tilde{p}_{Y,\epsilon}(x) := b_{\epsilon}(d)^{-1}p_{Y,\epsilon}(x)$. Consider the

decomposition

$$\begin{aligned}
b_\epsilon(d)^{-(a+b)}\mathbb{E}(p_{X,\epsilon}(X)^a p_{Y,\epsilon}(X)^b) &= \int_{R^d} \tilde{p}_{X,\epsilon}(x)^a \tilde{p}_{Y,\epsilon}(x)^b p_X(x) dx \\
&= \int_{R^d} p_X(x)^{a+1} p_Y(x)^b dx + \int_{R^d} (\tilde{p}_{Y,\epsilon}(x)^b - p_Y(x)^b) p_X(x)^{a+1} dx \quad (2) \\
&\quad + \int_{R^d} (\tilde{p}_{X,\epsilon}(x)^a - p_X(x)^a) \tilde{p}_{Y,\epsilon}(x)^b p_X(x) dx
\end{aligned}$$

Hence, the assertion follows if the last two terms in (2) tend to 0 as $\epsilon \rightarrow 0$. By the extension of Hölder's inequality (see, e.g., Ch. 2, Bogachev, 2007),

$$\begin{aligned}
&\left| \int_{R^d} (\tilde{p}_{X,\epsilon}(x)^a - p_X(x)^a) \tilde{p}_{Y,\epsilon}(x)^b p_X(x) dx \right| \quad (3) \\
&\leq \|\tilde{p}_{X,\epsilon}(\cdot)^a - p_X(\cdot)^a\|_{(a+b+1)/a} \|\tilde{p}_{Y,\epsilon}(\cdot)^b\|_{(a+b+1)/b} \|p_X(\cdot)\|_{a+b+1}
\end{aligned}$$

The Lebesgue differentiation theorem implies

$$\tilde{p}_{X,\epsilon}(x)^{a+b+1} \rightarrow p_X(x)^{a+b+1} \text{ as } \epsilon \rightarrow 0 \text{ a.e.} \quad (4)$$

If $V = (V_1, \dots, V_d)'$ is an auxiliary random vector uniformly distributed in the unit ball $B_1(d)$, then $\tilde{p}_{X,\epsilon}(x) = \mathbb{E}(p_X(x - \epsilon V))$, and by Jensen's inequality,

$$\begin{aligned}
(\tilde{p}_{X,\epsilon}(x)^a)^{(a+b+1)/a} &\leq g_\epsilon(x) := \mathbb{E}(p_X(x - \epsilon V)^{a+b+1}) \quad (5) \\
&= \frac{1}{b_\epsilon(d)} \int_{B_\epsilon(x)} p_X(y)^{a+b+1} dy.
\end{aligned}$$

Since $p_X \in L_{a+b+1}(R^d)$, it follows from the Lebesgue differentiation theorem that

$$g_\epsilon(x) \rightarrow g(x) := p_X(x)^{a+b+1} \text{ as } \epsilon \rightarrow 0 \text{ a.e.} \quad (6)$$

Furthermore, Fubini's theorem yields

$$\int_{R^d} g_\epsilon(x) dx = \int_{R^d} g(x) dx. \quad (7)$$

We get from (4)-(7) and a generalization of the dominated convergence theorem (see, e.g., Ch. 2, Bogachev, 2007) that

$$\|\tilde{p}_{X,\epsilon}(\cdot)^a\|_{(a+b+1)/a} \rightarrow \|p_X(\cdot)^a\|_{(a+b+1)/a} \text{ as } \epsilon \rightarrow 0. \quad (8)$$

Similarly,

$$\|\tilde{p}_{Y,\epsilon}(\cdot)^b\|_{(a+b+1)/b} \rightarrow \|p_Y(\cdot)^b\|_{(a+b+1)/b} \text{ as } \epsilon \rightarrow 0. \quad (9)$$

Now we use the following result (see, e.g., Ch. 1, Kallenberg, 1997): For a sequence of functions $f_n \in L_p(R^d)$, $p \geq 1$, $n = 1, \dots$, with $f_n(x) \rightarrow f(x)$ a.e., $f \in L_p(R^d)$, it holds that

$$\|f_n\|_p \rightarrow \|f\|_p \text{ iff } \|f_n - f\|_p \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (10)$$

Note that (4), (8), and (10) imply

$$\|\tilde{p}_{X,\epsilon}(\cdot)^a - p_X(\cdot)^a\|_{(a+b+1)/a} \rightarrow 0 \text{ as } \epsilon \rightarrow 0. \quad (11)$$

Finally, it follows from (3), (9), and (11) that

$$\int_{\mathbb{R}^d} (\tilde{p}_{X,\epsilon}(x)^a - p_X(x)^a) \tilde{p}_{Y,\epsilon}(x)^b p_X(x) dx \rightarrow 0 \text{ as } \epsilon \rightarrow 0.$$

In a similar way, we obtain

$$\int_{\mathbb{R}^d} (\tilde{p}_{Y,\epsilon}(x)^b - p_Y(x)^b) p_X(x)^{a+1} dx \rightarrow 0 \text{ as } \epsilon \rightarrow 0.$$

This completes the proof. \square

Proof of Theorem 1. For $l = 0, \dots, k_1$, and $m = 0, \dots, k_2$, let

$$\begin{aligned} \psi_{\mathbf{k},l,m,\mathbf{n}}(x_1, \dots, x_l; y_1, \dots, y_m) \\ := \mathbb{E}(\psi_{\mathbf{k},\mathbf{n}}(x_1, \dots, x_l, X_{l+1}, \dots, X_{k_1}; y_1, \dots, y_m, Y_{m+1}, \dots, Y_{k_2})) \end{aligned}$$

and

$$\sigma_{\mathbf{k},l,m,\mathbf{n}}^2 := \text{Var}(\psi_{\mathbf{k},l,m,\mathbf{n}}(X_1, \dots, X_l; Y_1, \dots, Y_m)),$$

where we define $\sigma_{\mathbf{k},0,0,\mathbf{n}}^2 = 0$. From the conventional theory of U -statistics (see, e.g., Ch. 2, Lee, 1990), we have

$$\text{Var}(\tilde{Q}_{\mathbf{k},\mathbf{n}}) = b_\epsilon(d)^{-2(k-1)} \sum_{l=0}^{k_1} \sum_{m=0}^{k_2} \frac{\binom{k_1}{l} \binom{k_2}{m} \binom{n_1-k_1}{k_1-l} \binom{n_2-k_2}{k_2-m}}{\binom{n_1}{k_1} \binom{n_2}{k_2}} \sigma_{\mathbf{k},l,m,\mathbf{n}}^2. \quad (12)$$

First, assume that $k_1 \geq 1$. Following the argument in Källberg et al. (2012), it is straightforward to show that

$$\sigma_{\mathbf{k},l,m,\mathbf{n}}^2 \leq \mathbb{E}(p_{X,3\epsilon}(X)^{2k_1-l-1} p_{Y,3\epsilon}(X)^{2k_2-m}),$$

so Lemma 1 implies

$$\sigma_{\mathbf{k},l,m,\mathbf{n}}^2 = \mathcal{O}(b_\epsilon(d)^{2k-l-m-1}) \text{ as } \mathbf{n} \rightarrow \infty. \quad (13)$$

For $l = 0, \dots, k_1$ and $m = 0, \dots, k_2$, we obtain

$$\begin{aligned} b_\epsilon(d)^{-2(k-1)} \frac{\binom{k_1}{l} \binom{k_2}{m} \binom{n_1-k_1}{k_1-l} \binom{n_2-k_2}{k_2-m}}{\binom{n_1}{k_1} \binom{n_2}{k_2}} \sigma_{\mathbf{k},l,m,\mathbf{n}}^2 \\ \sim C_{l,m} \frac{b_\epsilon(d)^{-(2k-l-m-1)} \sigma_{\mathbf{k},l,m,\mathbf{n}}^2}{n^{l+m} \epsilon^{d(l+m-1)}}, \end{aligned} \quad (14)$$

for some constant $C_{l,m} > 0$. Since $l + m \leq k$, we have

$$n^{l+m} \epsilon^{d(l+m-1)} = (n \epsilon^{d(1-1/(l+m))})^{l+m} \geq (n \epsilon^{d(1-1/k)})^{l+m}. \quad (15)$$

Now, if $n\epsilon^{d(1-1/k)} \rightarrow c, 0 < c \leq \infty$, it follows from (12)-(15) that

$$\text{Var}(\tilde{Q}_{\mathbf{k},\mathbf{n}}) = O\left(\frac{1}{n\epsilon^{d(1-1/k)}}\right) \text{ as } \mathbf{n} \rightarrow \infty. \quad (16)$$

In the same way, it can be shown that (16) holds when $k_1 = 0$. In particular, if $n\epsilon^{d(1-1/k)} \rightarrow \infty$, we get

$$\text{Var}(\tilde{Q}_{\mathbf{k},\mathbf{n}}) \rightarrow 0 \text{ as } \mathbf{n} \rightarrow \infty. \quad (17)$$

Moreover, it follows from Lemma 1 that $E(\tilde{Q}_{\mathbf{k},\mathbf{n}}) = b_\epsilon(d)^{-(k-1)}q_{\mathbf{k},\epsilon} \rightarrow q_{\mathbf{k}}$, so we obtain from (17) that $\tilde{Q}_{\mathbf{k},\mathbf{n}} \xrightarrow{P} q_{\mathbf{k}}$ as $\mathbf{n} \rightarrow \infty$. This completes the proof. \square

Proof of Theorem 2. Note that (i) and (ii) can be expressed together as follows: If $n\epsilon^d \rightarrow \beta, 0 \leq \beta < \infty$, and $n^2\epsilon^d \rightarrow \infty$, then

$$n\epsilon^{d/2}(\tilde{Q}_{2,\mathbf{n}} - \tilde{q}_{2,\epsilon}) \xrightarrow{D} N(0, \eta + \beta\kappa).$$

So we prove the theorem using the scaling $n\epsilon^{d/2}$ for both (i) and (ii). If $n_3 = n_3(\mathbf{n})$ is defined to be the greatest common divisor of n_1 and n_2 , then $n_1 = n_3l$ and $n_2 = n_3m$, where l and m are positive integers that satisfy $l/(l+m) = \rho$. Consider the following *pooled* random vectors in $R^{d(l+m)}$

$$Z_i := (X_{l(i-1)+1}, \dots, X_{li}, Y_{m(i-1)+1}, \dots, Y_{mi}), \quad i = 1, \dots, n_3.$$

The method of proof relies on the decomposition

$$n\epsilon^{d/2}(\tilde{Q}_{2,\mathbf{n}} - \tilde{q}_{2,\epsilon}) = n\epsilon^{d/2} \binom{n_3}{2}^{-1} b_\epsilon(d)^{-1}(U_{\mathbf{n}} - E(U_{\mathbf{n}})) + R_{\mathbf{n}}, \quad (18)$$

where $U_{\mathbf{n}}$, to be defined later, essentially is a *one-sample U-statistic* with respect to the i.i.d. sample $\{Z_1, \dots, Z_{n_3}\}$. The idea is to prove that the remainder $R_{\mathbf{n}}$ tends to 0 in probability and use the corresponding result from Jammalamadaka and Janson (1986) to show asymptotic normality for the first term in (18).

For $z_i := (x_{l(i-1)+1}, \dots, x_{li}, y_{m(i-1)+1}, \dots, y_{mi}), i = 1, \dots, n_3$, introduce the kernels

$$\begin{aligned} \phi_{\mathbf{n}}^{(1)}(z_i, z_j) &:= \sum_{s=1}^l \sum_{t=1}^l I(d(x_{l(i-1)+s}, x_{l(j-1)+t}) < \epsilon), \\ \phi_{\mathbf{n}}^{(2)}(z_i, z_j) &:= \sum_{s=1}^m \sum_{t=1}^m I(d(y_{m(i-1)+s}, y_{m(j-1)+t}) < \epsilon), \\ \phi_{\mathbf{n}}^{(3)}(z_i, z_j) &:= \sum_{s=1}^l \sum_{t=1}^m I(d(x_{l(i-1)+s}, y_{m(j-1)+t}) < \epsilon) \\ &\quad + \sum_{s=1}^l \sum_{t=1}^m I(d(x_{l(j-1)+s}, y_{m(i-1)+t}) < \epsilon). \end{aligned}$$

Furthermore, define

$$\begin{aligned}
f_{\mathbf{n}}(z_i, z_j) &:= a_0 l^{-2} \phi_{\mathbf{n}}^{(1)}(z_i, z_j) + a_2 m^{-2} \phi_{\mathbf{n}}^{(2)}(z_i, z_j) + a_1 (2lm)^{-1} \phi_{\mathbf{n}}^{(3)}(z_i, z_j), \\
\mu_{\mathbf{n}} &:= \mathbb{E}(f_{\mathbf{n}}(Z_1, Z_2)) = b_{\epsilon}(d) \tilde{q}_{2, \epsilon}, \\
g_{\mathbf{n}}(z_i) &:= \mathbb{E}(f_{\mathbf{n}}(z_i, Z_j)) - \mu_{\mathbf{n}} \\
&= \frac{a_0}{l} \sum_{s=1}^l p_{X, \epsilon}(x_{l(i-1)+s}) + \frac{a_2}{m} \sum_{s=1}^m p_{Y, \epsilon}(y_{m(i-1)+s}) \\
&\quad + \frac{a_1}{2} \left(\frac{1}{l} \sum_{s=1}^l p_{Y, \epsilon}(x_{l(i-1)+s}) + \frac{1}{m} \sum_{s=1}^m p_{X, \epsilon}(y_{m(i-1)+s}) \right) - \mu_{\mathbf{n}}.
\end{aligned} \tag{19}$$

Let

$$\begin{aligned}
M_{\mathbf{n}} &:= \sum_{i < j} I(d(X_i, X_j) < \epsilon), \quad V_{\mathbf{n}} := \sum_{i < j} I(d(Y_i, Y_j) < \epsilon), \\
W_{\mathbf{n}} &:= \sum_{i, j} I(d(X_i, Y_j) < \epsilon),
\end{aligned}$$

and note that

$$b_{\epsilon}(d) \tilde{Q}_{2, \mathbf{n}} = a_0 \binom{n_1}{2}^{-1} M_{\mathbf{n}} + a_2 \binom{n_2}{2}^{-1} V_{\mathbf{n}} + a_1 (n_1 n_2)^{-1} W_{\mathbf{n}}. \tag{20}$$

Now consider the decompositions

$$M_{\mathbf{n}} = M_{\mathbf{n}}^{(1)} + M_{\mathbf{n}}^{(2)}, \quad V_{\mathbf{n}} = V_{\mathbf{n}}^{(1)} + V_{\mathbf{n}}^{(2)}, \quad W_{\mathbf{n}} = W_{\mathbf{n}}^{(1)} + W_{\mathbf{n}}^{(2)},$$

where

$$M_{\mathbf{n}}^{(1)} := \sum_{i < j} \phi_{\mathbf{n}}^{(1)}(Z_i, Z_j), \quad V_{\mathbf{n}}^{(1)} := \sum_{i < j} \phi_{\mathbf{n}}^{(2)}(Z_i, Z_j), \quad W_{\mathbf{n}}^{(1)} := \sum_{i < j} \phi_{\mathbf{n}}^{(3)}(Z_i, Z_j),$$

and define

$$U_{\mathbf{n}} := \frac{a_0}{l^2} M_{\mathbf{n}}^{(1)} + \frac{a_2}{m^2} V_{\mathbf{n}}^{(1)} + \frac{a_1}{2lm} W_{\mathbf{n}}^{(1)} = \sum_{i < j} f_{\mathbf{n}}(Z_i, Z_j).$$

With this definition of $U_{\mathbf{n}}$, it follows from (20) that the decomposition (18) holds with remainder

$$R_{\mathbf{n}} = R_{\mathbf{n}}^{(1)} + R_{\mathbf{n}}^{(2)} + R_{\mathbf{n}}^{(3)} + R_{\mathbf{n}}^{(4)} + R_{\mathbf{n}}^{(5)} + R_{\mathbf{n}}^{(6)},$$

where

$$\begin{aligned}
R_{\mathbf{n}}^{(1)} &:= a_0 \left(\binom{n_1}{2}^{-1} - l^{-2} \binom{n_3}{2}^{-1} \right) n \epsilon^{d/2} b_\epsilon(d)^{-1} M_{\mathbf{n}}^{(1)}, \\
R_{\mathbf{n}}^{(2)} &:= a_2 \left(\binom{n_2}{2}^{-1} - m^{-2} \binom{n_3}{2}^{-1} \right) n \epsilon^{d/2} b_\epsilon(d)^{-1} V_{\mathbf{n}}^{(1)}, \\
R_{\mathbf{n}}^{(3)} &:= a_1 \left((n_1 n_2)^{-1} - (2lm)^{-1} \binom{n_3}{2}^{-1} \right) n \epsilon^{d/2} b_\epsilon(d)^{-1} W_{\mathbf{n}}^{(1)}, \\
R_{\mathbf{n}}^{(4)} &:= a_0 \binom{n_1}{2}^{-1} n \epsilon^{d/2} b_\epsilon(d)^{-1} M_{\mathbf{n}}^{(2)}, \\
R_{\mathbf{n}}^{(5)} &:= a_2 \binom{n_2}{2}^{-1} n \epsilon^{d/2} b_\epsilon(d)^{-1} V_{\mathbf{n}}^{(2)}, \\
R_{\mathbf{n}}^{(6)} &:= a_1 (n_1 n_2)^{-1} n \epsilon^{d/2} b_\epsilon(d)^{-1} W_{\mathbf{n}}^{(2)}.
\end{aligned}$$

By the conventional U -statistic theory for one-sample U -statistics (see, e.g., Ch. 1, Lee, 1990), we have

$$\text{Var}(M_{\mathbf{n}}^{(1)}) = \binom{n_3}{2} (2(n_3 - 2)\xi_{1,\mathbf{n}} + \xi_{2,\mathbf{n}}), \quad (21)$$

where

$$\xi_{1,\mathbf{n}} := \text{Cov}(\phi_{\mathbf{n}}^{(1)}(Z_1, Z_2), \phi_{\mathbf{n}}^{(1)}(Z_1, Z_3)), \quad \xi_{2,\mathbf{n}} := \text{Var}(\phi_{\mathbf{n}}^{(1)}(Z_1, Z_2)).$$

It follows that

$$\xi_{2,\mathbf{n}} \leq \mathbb{E}(\phi_{\mathbf{n}}^{(1)}(Z_1, Z_2)^2) \leq l^2 \mathbb{E}(\phi_{\mathbf{n}}^{(1)}(Z_1, Z_2)) = l^4 P(d(X_1, X_2) < \epsilon),$$

and hence $\xi_{1,\mathbf{n}}, \xi_{2,\mathbf{n}} = O(b_\epsilon(d))$. Therefore, we get from (21) that

$$\text{Var}(M_{\mathbf{n}}^{(1)}) = O(n_3^3 b_\epsilon(d)) \text{ as } \mathbf{n} \rightarrow \infty. \quad (22)$$

Since

$$\binom{n_1}{2}^{-1} - l^{-2} \binom{n_3}{2}^{-1} = \binom{n_1}{2}^{-1} - l^{-2} \binom{n_1/l}{2}^{-1} \sim \frac{C}{n_1^3} \text{ as } \mathbf{n} \rightarrow \infty,$$

it follows from (22) that

$$\text{Var}(R_{\mathbf{n}}^{(1)}) = O(n^{-1}) \rightarrow 0 \text{ as } \mathbf{n} \rightarrow \infty. \quad (23)$$

Similarly, for $i = 2, 3$,

$$\text{Var}(R_{\mathbf{n}}^{(i)}) \rightarrow 0 \text{ as } \mathbf{n} \rightarrow \infty. \quad (24)$$

Moreover, if a kernel is defined as

$$\theta_{\mathbf{n}}(z_i) := \sum_{1 \leq j < k \leq l} I(d(x_{l(i-1)+j}, x_{l(i-1)+k}) < \epsilon), \quad i = 1, \dots, n_3,$$

then

$$M_{\mathbf{n}}^{(2)} = \sum_{i=1}^{n_3} \theta_{\mathbf{n}}(Z_i)$$

and by Lemma 1,

$$\text{Var}(M_{\mathbf{n}}^{(2)}) = n_3 \text{Var}(\theta_{\mathbf{n}}(Z_1)) \sim n_3 \binom{l}{2} b_{\epsilon}(d) q_{2,0} = O(n_3 \epsilon^d) \text{ as } \mathbf{n} \rightarrow \infty.$$

This yields

$$\text{Var}(R_{\mathbf{n}}^{(4)}) = O(n^{-1}) \rightarrow 0 \text{ as } \mathbf{n} \rightarrow \infty. \quad (25)$$

In a similar way, for $i=5, 6$,

$$\text{Var}(R_{\mathbf{n}}^{(i)}) \rightarrow 0 \text{ as } \mathbf{n} \rightarrow \infty. \quad (26)$$

Since $E(R_{\mathbf{n}}) = 0$, it follows from (23)-(26) that

$$R_{\mathbf{n}} \xrightarrow{P} 0 \text{ as } \mathbf{n} \rightarrow \infty. \quad (27)$$

Next we prove asymptotic normality for $U_{\mathbf{n}}$. Let

$$\sigma_{\mathbf{n}}^2 := \frac{n_3^2}{2} \text{Var}(f_{\mathbf{n}}(Z_1, Z_2)) + n_3^3 \text{Var}(g_{\mathbf{n}}(Z_1)). \quad (28)$$

Using Lemma 1, it is straightforward to show that, as $\mathbf{n} \rightarrow \infty$,

$$\begin{aligned} \sigma_{\mathbf{n}}^2 &\sim \frac{n_3^2 b_{\epsilon}(d)}{2} \left(\frac{a_0^2}{l^2} q_{2,0} + \frac{a_2^2}{m^2} q_{0,2} + \frac{a_1^2}{2lm} q_{1,1} \right) \\ &+ n_3^3 b_{\epsilon}(d)^2 \left(\frac{1}{l} \text{Var} \left(a_0 p_X(X) + \frac{a_1}{2} p_Y(X) \right) + \frac{1}{m} \text{Var} \left(a_2 p_Y(Y) + \frac{a_1}{2} p_X(Y) \right) \right). \end{aligned} \quad (29)$$

Since $n^2 \epsilon^d \rightarrow \infty$ implies $n_3^2 b_{\epsilon}(d) \rightarrow \infty$, we get from (29) that

$$\sigma_{\mathbf{n}} \rightarrow \infty \text{ as } \mathbf{n} \rightarrow \infty,$$

and hence

$$\sup_{z_1, z_2} |f_{\mathbf{n}}(z_1, z_2)| \leq |a_0| + |a_2| + |a_1| = o(\sigma_{\mathbf{n}}) \text{ as } \mathbf{n} \rightarrow \infty. \quad (30)$$

Moreover, note that

$$\begin{aligned} E(|f_{\mathbf{n}}(z_1, Z_2)|) &\leq \frac{|a_0|}{l} \sum_{i=1}^l p_{X,\epsilon}(x_i) + \frac{|a_2|}{m} \sum_{i=1}^m p_{Y,\epsilon}(y_i) \\ &+ \frac{|a_1|}{2l} \sum_{i=1}^l p_{Y,\epsilon}(x_i) + \frac{|a_1|}{2m} \sum_{i=1}^m p_{X,\epsilon}(y_i). \end{aligned} \quad (31)$$

By Hölder's inequality,

$$\begin{aligned} p_{X,\epsilon}(x) &= \int_{|y-x|<\epsilon} p_X(y)dy \leq \left(\int_{|y-x|<\epsilon} dy \right)^{1/2} \left(\int_{|y-x|<\epsilon} p_X(y)^2 dy \right)^{1/2} \\ &= b_\epsilon(d)^{1/2} \left(\int_{|y-x|<\epsilon} p_X(y)^2 dy \right)^{1/2}, \end{aligned}$$

where the last integral tends to 0 uniformly in x as $\epsilon \rightarrow 0$. The corresponding results can be shown for the other terms in (31). Hence, we obtain

$$\sup_{z_1} \mathbb{E}(|f_{\mathbf{n}}(z_1, Z_2)|) = o(b_\epsilon(d)^{1/2}) = o(\sigma_{\mathbf{n}}/n_3). \quad (32)$$

Now, it follows from (30) and (32) that the conditions of Theorem 2.1. in Jamalamadaka and Janson (1986) hold. Consequently,

$$\frac{U_{\mathbf{n}} - \mathbb{E}(U_{\mathbf{n}})}{\sigma_{\mathbf{n}}} = \frac{1}{\sigma_{\mathbf{n}}} \left(\sum_{i<j} f_{\mathbf{n}}(Z_i, Z_j) - \binom{n_3}{2} \mu_{\mathbf{n}} \right) \xrightarrow{D} N(0, 1) \text{ as } \mathbf{n} \rightarrow \infty. \quad (33)$$

Moreover, since $n\epsilon^d \rightarrow \beta, 0 \leq \beta < \infty$, it follows from (29) and Lemma 1 that

$$n^2 \epsilon^d \binom{n_3}{2}^{-2} b_\epsilon(d)^{-2} \sigma_{\mathbf{n}}^2 \rightarrow \eta + \beta\kappa \text{ as } \mathbf{n} \rightarrow \infty. \quad (34)$$

Finally, (18), (27), (33), (34), and the Slutsky theorem yield

$$n\epsilon^{d/2}(\tilde{Q}_{2,\mathbf{n}} - \tilde{q}_{2,\epsilon}) \xrightarrow{D} N(0, \eta + \beta\kappa) \text{ as } \mathbf{n} \rightarrow \infty.$$

This completes the proof. \square

Proof of Theorem 3. (i) Let $V := (V_1, \dots, V_d)'$ be an auxiliary random vector uniformly distributed in the unit ball $B_1(0)$. By definition, we have $\tilde{q}_{1,1,\epsilon} = b_\epsilon(d)^{-1} \mathbb{E}(p_{X,\epsilon}(Y)) = \mathbb{E}(p_X(Y - \epsilon V))$, and hence

$$\begin{aligned} \tilde{q}_{1,1,\epsilon} - q_{1,1} &= \int_{\mathbb{R}^d} \mathbb{E}(p_X(y - \epsilon V)) p_Y(y) dy - \int_{\mathbb{R}^d} p_X(y) p_Y(y) dy \\ &= \mathbb{E} \left(\int_{\mathbb{R}^d} (p_X(y - \epsilon V) - p_X(y)) p_Y(y) dy \right) \\ &= \mathbb{E} \left(\int_{\mathbb{R}^d} (p_X(y - \epsilon V) - p_X(y)) (p_Y(y) - p_Y(y - \epsilon V)) dy \right) \\ &\quad + \mathbb{E} \left(\int_{\mathbb{R}^d} (p_X(y - \epsilon V) - p_X(y)) p_Y(y - \epsilon V) dy \right). \end{aligned}$$

For the last term, by the change of variables $z = y - \epsilon V$ and symmetry $V \stackrel{D}{=} -V$, we obtain

$$\begin{aligned} & \mathbb{E} \left(\int_{R^d} (p_X(y - \epsilon V) - p_X(y)) p_Y(y - \epsilon V) dy \right) \\ &= \mathbb{E} \left(\int_{R^d} (p_X(z) - p_X(z + \epsilon V)) p_Y(z) dz \right) \\ &= \mathbb{E} \left(\int_{R^d} (p_X(z) - p_X(z - \epsilon V)) p_Y(z) dz \right) = -(\tilde{q}_{1,1,\epsilon} - q_{1,1}). \end{aligned}$$

We get

$$2(\tilde{q}_{1,1,\epsilon} - q_{1,1}) = \mathbb{E} \left(\int_{R^d} (p_X(y - \epsilon V) - p_X(y)) (p_Y(y) - p_Y(y - \epsilon V)) dy \right),$$

and, by Hölder's inequality and the density smoothness condition,

$$\begin{aligned} |\tilde{q}_{1,1,\epsilon} - q_{1,1}| &\leq \frac{1}{2} \mathbb{E} \left(\left(\int_{R^d} (p_X(y - \epsilon V) - p_X(y))^2 dy \right)^{1/2} \right. \\ &\quad \times \left. \left(\int_{R^d} (p_Y(y) - p_Y(y - \epsilon V))^2 dy \right)^{1/2} \right) \\ &\leq \frac{1}{2} K^2 \mathbb{E}(|V|^{2\alpha}) \epsilon^{2\alpha} \leq \frac{1}{2} K^2 \epsilon^{2\alpha}. \end{aligned}$$

Similar inequalities can be obtained for $\tilde{q}_{2,0,\epsilon}$ and $\tilde{q}_{0,2,\epsilon}$. Now it follows directly that, for some $C > 0$,

$$|\tilde{q}_{2,\epsilon} - q_2| \leq C \epsilon^{2\alpha}.$$

This proves the assertion.

(ii) Note that the conditions $\epsilon \sim cn^{-1/(2\alpha+d/2)}$ and $0 < \alpha \leq d/4$ imply

$$n\epsilon^d = c^d n^{-\frac{d/2-2\alpha}{2\alpha+d/2}} \rightarrow \beta, 0 \leq \beta < 0, \text{ as } \mathbf{n} \rightarrow \infty. \quad (35)$$

From Jammalamadaka and Janson (1986), we get

$$\text{Var}(\tilde{Q}_{2,0,\mathbf{n}}) \sim 4n_1^{-1}(q_{3,0} - q_{2,0}^2) + 2n_1^{-2}b_\epsilon(d)^{-1}q_{2,0} \text{ as } \mathbf{n} \rightarrow \infty, \quad (36)$$

and the corresponding result for $\text{Var}(\tilde{Q}_{0,2,\mathbf{n}})$. Furthermore, (12) yields

$$\text{Var}(\tilde{Q}_{1,1,\mathbf{n}}) = \frac{b_\epsilon(d)^{-2}}{n_1 n_2} ((n_1 - 1)\sigma_{1,1,0,1,\mathbf{n}}^2 + (n_2 - 1)\sigma_{1,1,1,0,\mathbf{n}}^2 + \sigma_{1,1,1,1,\mathbf{n}}^2), \quad (37)$$

where, by Lemma 1,

$$\begin{aligned} b_\epsilon(d)^{-2}\sigma_{1,1,0,1,\mathbf{n}}^2 &= b_\epsilon(d)^{-2}\text{Var}(p_{X,\epsilon}(Y_1)) \rightarrow q_{2,1} - q_{1,1}^2, \\ b_\epsilon(d)^{-2}\sigma_{1,1,1,0,\mathbf{n}}^2 &= b_\epsilon(d)^{-2}\text{Var}(p_{Y,\epsilon}(X_1)) \rightarrow q_{1,2} - q_{1,1}^2, \\ b_\epsilon(d)^{-1}\sigma_{1,1,1,1,\mathbf{n}}^2 &= b_\epsilon(d)^{-1}\text{Var}(I(d(X_1, Y_1) < \epsilon)) \rightarrow q_{1,1} \text{ as } \mathbf{n} \rightarrow \infty. \end{aligned} \quad (38)$$

It follows from (35)-(38) that

$$\begin{aligned}\text{Var}(\tilde{Q}_{2,\mathbf{n}}) &\leq 3 \left(a_0^2 \text{Var}(\tilde{Q}_{2,0,\mathbf{n}}) + a_2^2 \text{Var}(\tilde{Q}_{0,2,\mathbf{n}}) + a_1^2 \text{Var}(\tilde{Q}_{1,1,\mathbf{n}}) \right) \\ &= O \left(\frac{1}{n^2 \epsilon^d} \right) = O \left(n^{-\frac{4\alpha}{2\alpha+d/2}} \right) \text{ as } \mathbf{n} \rightarrow \infty.\end{aligned}\quad (39)$$

We get from (i) that $(\tilde{q}_{2,\epsilon} - q_2)^2 \leq C n^{-\frac{4\alpha}{2\alpha+d/2}}$, $C > 0$, so (39) implies

$$\text{Var}(\tilde{Q}_{2,\mathbf{n}}) + (\tilde{q}_{2,\epsilon} - q_2)^2 = O \left(n^{-\frac{4\alpha}{2\alpha+d/2}} \right) \text{ as } \mathbf{n} \rightarrow \infty.$$

Hence, for some $C_2 > 0$, any $A > 0$, and large enough n_1, n_2 , we obtain

$$P \left(|\tilde{Q}_{2,\mathbf{n}} - q_2| > A n^{-\frac{2\alpha}{2\alpha+d/2}} \right) \leq n^{\frac{4\alpha}{2\alpha+d/2}} \frac{\text{Var}(\tilde{Q}_{2,\mathbf{n}}) + (\tilde{q}_{2,\epsilon} - q_2)^2}{A^2} \leq \frac{C_2}{A^2},$$

and the assertion follows.

(iii) We have

$$\sqrt{n}(\tilde{Q}_{2,\mathbf{n}} - q_2) = \sqrt{n}(\tilde{Q}_{2,\mathbf{n}} - \tilde{q}_{2,\epsilon}) + \sqrt{n}(\tilde{q}_{2,\epsilon} - q_2). \quad (40)$$

Now, when $n\epsilon^d \rightarrow \beta$, $0 \leq \beta < \infty$ and $\alpha > d/4$, then (i) imply

$$|\sqrt{n}(\tilde{q}_{2,\epsilon} - q_2)| \leq C n^{1/2} \epsilon^{2\alpha} = C (n\epsilon^d)^{1/2} \epsilon^{2\alpha-d/2} \rightarrow 0 \text{ as } \mathbf{n} \rightarrow \infty,$$

so the assertion follows from Theorem 2(i) and the Slutsky theorem.

(iv) From (i) and the assumption $\epsilon \sim L(n)n^{-2/d}$, we get

$$|n\epsilon^{d/2}(\tilde{q}_{2,\epsilon} - q_2)| \leq CL(n)^{d/2+2\alpha} n^{-4\alpha/d} \rightarrow 0 \text{ as } \mathbf{n} \rightarrow \infty.$$

Therefore, similarly as above, the assertion follows by using the decomposition corresponding to (40), Theorem 2(ii), and the Slutsky theorem. This completes the proof. \square

Proof of Proposition 4. (i) When $n^2\epsilon^d \rightarrow \infty$ and $n\epsilon^d \rightarrow \beta$, $0 \leq \beta < \infty$, Theorem 2 can be applied with $\tilde{Q}_{2,\mathbf{n}} = \hat{D}_{2,\mathbf{n}}$. Under H_0 , we have $\tilde{q}_{2,\epsilon} = E(\hat{D}_{2,\mathbf{n}}) = 0$ and $\kappa = 0$, so Theorem 2 yields

$$n\epsilon^{d/2} \hat{D}_{2,\mathbf{n}} \xrightarrow{D} N(0, \eta) \text{ as } \mathbf{n} \rightarrow \infty \quad (41)$$

in this case. Hence, we need to show that, for $\hat{D}_{2,\mathbf{n}}$ under H_0 , the proof of Theorem 2 can be modified so that the assumption $n\epsilon^d \rightarrow \beta$, $0 \leq \beta < \infty$ is unnecessary. In fact, this assumption is only needed for the convergence property (34) of $\sigma_{\mathbf{n}}^2$. Under H_0 , we get from the definition (19) that $g_{\mathbf{n}}(z) = 0$ and hence $\text{Var}(g_{\mathbf{n}}(Z_1)) = 0$. Therefore, the definition (28) of $\sigma_{\mathbf{n}}^2$ implies

$$\sigma_{\mathbf{n}}^2 \sim \frac{n_3^2 b_{\epsilon}(d)}{2} \left(\frac{a_0^2}{l^2} q_{2,0} + \frac{a_2^2}{m^2} q_{0,2} + \frac{a_1^2}{2lm} q_{1,1} \right) \text{ as } \mathbf{n} \rightarrow \infty,$$

and hence (34) can be written

$$n^2 \epsilon^d \binom{n_3}{2}^{-2} b_\epsilon(d)^{-2} \sigma_{\mathbf{n}}^2 \rightarrow \eta \text{ as } \mathbf{n} \rightarrow \infty,$$

which does not require convergence of $n\epsilon^d$. Thus, the assertion follows from (41) and the Slutsky theorem.

(ii) Under H_1 , we get from Theorem 1 and the Slutsky theorem that $\hat{D}_{2,\mathbf{n}}/\sqrt{\eta_{\mathbf{n}}} \xrightarrow{P} D_2/\sqrt{\eta}$. Hence,

$$P(T_{\mathbf{n}} > u_{\mathbf{n}}) = 1 - P(D_2/\sqrt{\eta} - u_{\mathbf{n}}/n\epsilon^{d/2} \leq D_2/\sqrt{\eta} - \hat{D}_{2,\mathbf{n}}/\sqrt{\eta_{\mathbf{n}}}) \rightarrow 1 \text{ as } \mathbf{n} \rightarrow \infty.$$

This completes the proof. \square

Proof of Proposition 5. The assertion follows straightforwardly from Theorem 3 and in a similar way as in Leonenko and Seleznev (2010). \square

Acknowledgment

The second author is partly supported by the Swedish Research Council grant 2009-4489. We would like to thank professor N. Leonenko for valuable comments and discussions.

References

- Ahmad, I.A., Cerrito, P.B., 1993, Goodness of fit tests based on the L_2 -norm of multivariate probability density functions, *J. Nonparametr. Stat.* 2, 169-181.
- Basseville, M., 2010, Divergence measures for statistical data processing, Technical Report 1961, IRISA.
- Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C., 1998, Robust and efficient estimation by minimising a density power divergence, *Biometrika* 85, 549-559.
- Bogachev, V.I., 2007, *Measure Theory*, vol. I, Springer-Verlag, Berlin.
- Escolano, F., Suau, P., Bonev, B., 2009, *Information Theory in Computer Vision and Pattern Recognition*, Springer, New York.
- Giné, E., Nickl, R., 2008, A simple adaptive estimator for the integrated square of a density, *Bernoulli* 14, 47-61.
- Jammalamadaka, S.R., Janson, S., 1986, Limit theorems for a triangular scheme of U -statistics with applications to inter-point distances, *Ann. Probab.* 14, 1347-1358.
- Kallenberg, O., 1997, *Foundations of Modern Probability*, Springer-Verlag, New York.
- Kapur, J.N., 1989, *Maximum-entropy Models in Science and Engineering*, Wiley, New York.

- Kapur, J.N., Kesavan, H.K., 1992, *Entropy Optimization Principles with Applications*, Academic Press, New York.
- Koroljuk, V.S., Borovskich, Y.V., 1994, *Theory of U -statistics*, Kluwer, Dordrecht.
- Källberg, D., Leonenko, N., Seleznev, O., 2012, Statistical inference for Rényi entropy functionals, *Lecture Notes in Comput. Sci.* 7260, 36-51.
- Lee, A.J., 1990, *U -Statistics: Theory and Practice*, Marcel Dekker, New York.
- Leonenko, N., Pronzato, L., Savani, V., 2008, A class of Rényi information estimators for multidimensional densities. *Ann. Statist.* 36, 2153-2182. Corrections, 2010, *Ann. Statist.* 38, 3837-3838.
- Leonenko, N., Seleznev, O., 2010, Statistical inference for the ϵ -entropy and the quadratic Rényi entropy. *J. Multivariate Anal.* 101, 1981-1994.
- Li, Q., 1996, Nonparametric testing of closeness between two unknown distribution functions, *Econometric Rev.* 15, 261-274.
- Li, Q., 1999, Nonparametric testing the similarity of two unknown density functions: local power and bootstrap analysis, *J. Nonparametr. Stat.* 11, 189-213.
- Neemuchwala, H., Hero, A., Carson, P., 2005, Image matching using alpha-entropy measures and entropic graphs, *Signal Processing* 85, 277-296.
- Pardo, L., 2006, *Statistical Inference Based on Divergence Measures*, Chapman & Hall, Boca Raton.
- Penrose, M., 1995, Generalized two-sample U -statistics and a two-species reaction-diffusion model, *Stochastic Process. Appl.* 55, 57-64.
- Principe, J.C., 2010, *Information Theoretic Learning*, Springer, New York.
- Rényi, A., 1970, *Probability Theory*, North-Holland, Amsterdam.
- Seleznev, O., Thalheim, B., 2003, Average case analysis in database problems, *Methodol. Comput. Appl. Prob.* 5, 395-418.
- Seleznev, O., Thalheim, B., 2010, Random databases with approximate record matching, *Methodol. Comput. Appl. Prob.* 12, 63-89.
- Thalheim, B., 2000, *Entity-Relationship Modeling. Foundations of Database Technology*, Springer-Verlag, Berlin.
- Ullah, A., 1996, Entropy, divergence and distance measures with econometric applications. *J. Statist. Plann. Inference* 49, 137-162.
- Weber, N.C., 1983, Central limit theorems for a class of symmetric statistics, *Math. Proc. Cambridge Philos. Soc.* 94, 307-313.