Statistical inference for entropy-type density functionals

David Källberg
Umeå university

Third Northern Triangular Seminar
St. Petersburg, April 13, 2011

Coauthors:

Nikolaj Leonenko
School of Mathematics
Cardiff University

Oleg Seleznjev
Department of Mathematics and Mathematical Statistics
Umeå university

Outline

- Introduction
- Basic notation
- Main results
- Numerical experiments
- Applications
- References

Introduction

Let X and Y be d-dimensional random vectors with distributions $\mathcal{P}_X$ and $\mathcal{P}_Y$.

For the discrete case $\mathcal{P}_X = \{p_X(k), k \in N^d\}$ and $\mathcal{P}_Y = \{p_Y(k), k \in N^d\}$.

In the continuous case let $\mathcal{P}_X$ and $\mathcal{P}_Y$ be with densities $p_X(x), p_Y(x), x \in R^d$, respectively.

Entropy

In information theory and statistics, there are various generalizations of Shannon entropy, characterizing uncertainty, e.g.,

- the Rényi entropy,

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

- the (differentiable) variability for approximate record matching in random databases

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

Bregman distance

An example of statistical distance between distributions is given by the (nonsymmetric) Bregman distance

$$B_s(p_X, p_Y) = \int_{R^d} \left[p_X(x)^s + \frac{1}{s-1} p_Y(x)^s - \frac{s}{s-1} p_X(x) p_Y(x)^{s-1} \right] dx,$$

for $s \neq 1$. When $s = 2$, we get the second order distance

$$B_2(p_X, p_Y) = \int_{R^d} [p_X(x) - p_Y(x)]^2 dx.$$

Entropy-type functionals

For non-negative integers $r_1, r_2 \geq 0$ and $\mathbf{r} := (r_1, r_2)$, we consider *Rényi entropy functionals*

$$q_{\mathbf{r}} = q_{r_1,r_2} := \int_{R^d} p_X(x)^{r_1} p_Y(x)^{r_2} dx,$$

for continuous distributions, and

$$q_{\mathbf{r}} = q_{r_1,r_2} := \sum_k p_X(k)^{r_1} p_Y(k)^{r_2},$$

for discrete distributions.

Note that

- the Rényi entropy $h_s = h_{s,0} = \log(q_{s,0})/(1-s)$.
- the variability $v = h_{1,1} = -\log(q_{1,1})$.
- the second order Bregman distance $K_2 = q_{2,0} + q_{0,2} - 2q_{1,1}$.

Our aim

Estimation of entropy-type functionals $q_{\mathbf{r}}$ and related characteristics for $\mathcal{P}_X$ and $\mathcal{P}_Y$ from mutually independent and identically distributed samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$.

Basic notation

Basic notation

Let the $d(x,y)$ denote the Euclidean distance in R^d and $B_\epsilon(x) := \{y : d(x,y) \leq \epsilon\}$ be an ϵ-ball in R^d with center at x, radius ϵ, and volume $b_\epsilon(d) = \epsilon^d b_1(d)$. Define the ϵ-ball probability

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

Write $I(C)$ for the indicator of an event C.

Denote $\mathbf{n} := (n_1, n_2)$, $n := n_1 + n_2$, and say that $\mathbf{n} \to \infty$ if $n_1, n_2 \to \infty$ and let $p_\mathbf{n} := n_1/n \to p, 0 < p < 1$, as $\mathbf{n} \to \infty$.

Coincidence probability

Our method relies on estimating the ϵ-coincidence probability

$$\begin{aligned} q_{\mathbf{r},\epsilon} &:= P(d(X_1, X_i) \leq \epsilon, d(X_1, Y_j) \leq \epsilon, i = 2, \ldots, r_1, j = 2, \ldots, r_2) \\ &= \mathrm{E} p_{X,\epsilon}(X)^{r_1 - 1} p_{Y,\epsilon}(X)^{r_1}. \end{aligned}$$

In the discrete case, we put $\epsilon = 0$. Hence

$$q_{\mathbf{r},0} = q_{\mathbf{r}} = P(X_1 = X_i = Y_j, i = 2, \ldots, r_1, j = 1, \ldots, r_2)$$

is the coincidence probability.

Estimator

Let $\mathcal{S}_{m,k}$ be the set of all k-subsets of $\{1,\ldots,m\}$. For $S \in \mathcal{S}_{n_1,r_1}$, $T \in \mathcal{S}_{n_2,r_2}$, and $i \in S$, define

$$\psi_{\mathbf{n}}^{(i)}(S;T) := I(d(X_i, X_j) \leq \epsilon, d(X_i, Y_k) \leq \epsilon, \forall j \in S, \forall k \in T),$$

i.e., the indicator of the event that all elements in $\{X_j, j \in S\}$ and $\{Y_k, k \in T\}$ are ϵ-close to X_i.

Estimator

Note that

$$\mathrm{E}\psi_{\mathbf{n}}^{(i)}(S;T) = \mathrm{E}p_{X,\epsilon}(X)^{r_1-1}p_{Y,\epsilon}(X)^{r_2} = q_{\mathbf{r},\epsilon}.$$

A generalized U-statistic for $q_{\mathbf{r},\epsilon}$ is given by

$$Q_{\mathbf{n}} = Q_{\mathbf{n},\mathbf{r},\epsilon} := \binom{n_1}{r_1}^{-1}\binom{n_2}{r_2}^{-1}\sum_{(n_1,r_1)}\sum_{(n_2,r_2)}\psi_{\mathbf{n}}(S;T)$$

with the symmetrized kernel

$$\psi_{\mathbf{n}}(S;T) := \frac{1}{r_1}\sum_{i\in S}\psi_{\mathbf{n}}^{(i)}(S;T).$$

Asymptotic variance

For discrete and continuous distributions, we define

$$\begin{aligned}
\zeta_{1,0} &:= \operatorname{Var}(p_X(X)^{r_1-1}p_Y(X)^{r_2}) = q_{2r_1-1,2r_2} - q_{r_1,r_2}^2, \\
\zeta_{0,1} &:= \operatorname{Var}(p_X(Y)^{r_1}p_Y(Y)^{r_2-1}) = q_{2r_1,2r_2-1} - q_{r_1,r_2}^2, \\
\kappa &:= p^{-1}r_1^2\zeta_{1,0} + (1-p)^{-1}r_2^2\zeta_{0,1}.
\end{aligned}$$

Idea: The asymptotic variance κ takes the form of an entropy-type functional, and hence it can be estimated by the same method.

Main results

Exact coincidences ($\epsilon = 0$) are considered. Then

$$\psi_{\mathbf{n}}(S;T) = \psi_{\mathbf{n}}^{(i)}(S;T),$$

and $Q_{\mathbf{n}}$ is an unbiased estimator of $q_{\mathbf{r}}$. Let $Q_{\mathbf{n},\mathbf{r}} := Q_{\mathbf{n},\mathbf{r},0}$,

$$\begin{aligned} K_{\mathbf{n}} := p_{\mathbf{n}}^{-1} r_1^2 (Q_{\mathbf{n},2r_1-1,2r_2} - Q_{\mathbf{n},\mathbf{r}}^2) \\ + (1-p_{\mathbf{n}})^{-1} r_2^2 (Q_{\mathbf{n},2r_1,2r_2-1} - Q_{\mathbf{n},\mathbf{r}}^2), \end{aligned}$$

and $k_{\mathbf{n}} := \max(K_{\mathbf{n}}, 1/n)$ be an estimator of κ.

Denote by $H_{\mathbf{n}} := \log(\max(Q_{\mathbf{n}}, 1/n))/(1-r)$ the estimator of $h_{\mathbf{r}} := \log(q_{\mathbf{r}})/(1-r)$.

Theorem

If $\zeta_{1,0}, \zeta_{0,1} > 0$, *then*

$$\sqrt{n}(Q_{\mathbf{n}} - q_{\mathbf{r}}) \xrightarrow{\text{D}} N(0, \kappa) \;\; \textit{and} \;\; \sqrt{n}(Q_{\mathbf{n}} - q_{\mathbf{r}})/k_{\mathbf{n}}^{1/2} \xrightarrow{\text{D}} N(0, 1);$$

$$\sqrt{n}(1 - r)\frac{Q_{\mathbf{n}}}{k_{\mathbf{n}}^{1/2}}(H_{\mathbf{n}} - h_{\mathbf{r}}) \xrightarrow{\text{D}} N(0, 1) \;\; \textit{as} \;\; \mathbf{n} \to \infty.$$

Continuous distributions

- Denote by
$$\tilde{Q}_{\mathbf{n}} := Q_{\mathbf{n}}/b_{\epsilon}(d)^{r-1}$$
the estimator of $q_{\mathbf{r}}$.
- Let $\tilde{q}_{\mathbf{r},\epsilon} := \mathrm{E}\tilde{Q}_{\mathbf{n}} = q_{\mathbf{r},\epsilon}/b_{\epsilon}(d)^{r-1}$ and $v_{\mathbf{n}}^2 := \mathrm{Var}(\tilde{Q}_{\mathbf{n}})$.
- Assume that $\epsilon = \epsilon(\mathbf{n}) \to 0$ as $\mathbf{n} \to \infty$.

Continuous distributions

Theorem

Let $p_X(x)$ and $p_Y(x)$ be bounded and continuous or with a finite number of discontinuity points.

(i) $v_{\mathbf{n}}^2 = \mathrm{O}(n^{-1}\epsilon^{d(1/r-1)})$ *and* $\mathrm{E}\tilde{Q}_{\mathbf{n}} \to q_{\mathbf{r}}$ *as* $\mathbf{n} \to \infty$, *and hence if* $n\epsilon^{d(1-1/r)} \to \infty$ *as* $\mathbf{n} \to \infty$, *then* $\tilde{Q}_{\mathbf{n}}$ *is a consistent estimator of* $q_{\mathbf{r}}$.

(ii) *If* $n\epsilon^d \to \infty$ *as* $\mathbf{n} \to \infty$ *and* $\zeta_{1,0}, \zeta_{0,1} > 0$, *then*

$$\sqrt{n}(\tilde{Q}_{\mathbf{n}} - \tilde{q}_{\mathbf{r},\epsilon}) \xrightarrow{\mathrm{D}} N(0,\kappa) \text{ as } \mathbf{n} \to \infty.$$

Smoothness conditions

The estimator is biased, so we introduce smoothness conditions to evaluate $q_{\mathbf{r}}$.

Denote by $H^{(\alpha)}(C)$ $, 0 < \alpha \leq 2, C > 0$, a linear space of bounded and continuous functions in R^d satisfying α-Hölder condition if $0 < \alpha \leq 1$ or if $1 < \alpha \leq 2$ with continuous partial derivatives satisfying $(\alpha - 1)$-Hölder condition with constant C.

Continuous distributions

- Let

$$K_{\mathbf{n}} := p_{\mathbf{n}}^{-1} r_1^2 (\tilde{Q}_{\mathbf{n},2r_1-1,2r_2,\epsilon} - \tilde{Q}^2_{\mathbf{n},\mathbf{r},\epsilon}) \\ + (1-p_{\mathbf{n}})^{-1} r_2^2 (\tilde{Q}_{\mathbf{n},2r_1,2r_2-1,\epsilon} - \tilde{Q}^2_{\mathbf{n},\mathbf{r},\epsilon}),$$

  and define $k_{\mathbf{n}} := \max(K_{\mathbf{n}}, 1/n)$.

- Denote by $H_{\mathbf{n}} := \log(\max(\tilde{Q}_{\mathbf{n}}, 1/n))/(1-r)$ the estimator of $h_{\mathbf{r}} := \log(q_{\mathbf{r}})/(1-r)$.

- Let $L(n)$ be a slowly varying function.

Theorem

Let $p_X(x), p_Y(x) \in H^{(\alpha)}(C)$.
(i) Then the bias $|\tilde{q}_{\mathbf{r},\epsilon} - q_{\mathbf{r}}| \leq C_1\epsilon^{\alpha}, C_1 > 0$.
(ii) If $0 < \alpha \leq d/2$ *and* $\epsilon \sim cn^{-\alpha/(2\alpha+d(1-1/r))}, 0 < c < \infty$, *then*

$$\begin{aligned} \tilde{Q}_{\mathbf{n}} - q_{\mathbf{r}} &= \mathrm{O}_{\mathrm{P}}(n^{-\alpha/(2\alpha+d(1-1/r))}); \\ H_{\mathbf{n}} - h_{\mathbf{r}} &= \mathrm{O}_{\mathrm{P}}(n^{-\alpha/(2\alpha+d(1-1/r))}) \text{ as } \mathbf{n} \to \infty. \end{aligned}$$

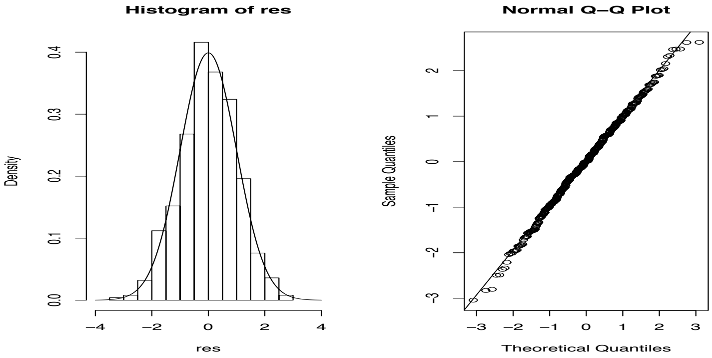
(iii) If $\alpha > d/2$ *and* $\epsilon \sim L(n)n^{-1/d}$ *and* $n\epsilon^d \to \infty$, *then*

$$\sqrt{n}(\tilde{Q}_{\mathbf{n}} - q_{\mathbf{r}}) \xrightarrow{\mathrm{D}} N(0,\kappa) \text{ and } \sqrt{n}(\tilde{Q}_{\mathbf{n}} - q_{\mathbf{r}})/k_{\mathbf{n}}^{1/2} \xrightarrow{\mathrm{D}} N(0,1);$$

$$\sqrt{n}(1-r)\frac{\tilde{Q}_{\mathbf{n}}}{k_{\mathbf{n}}^{1/2}}(H_{\mathbf{n}} - h_{\mathbf{r}}) \xrightarrow{\mathrm{D}} N(0,1) \text{ as } \mathbf{n} \to \infty.$$

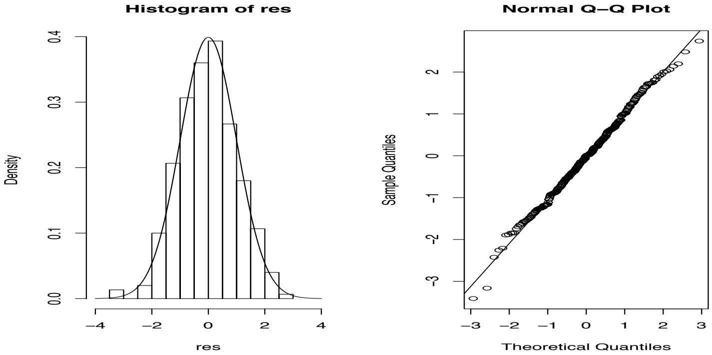
Numerical experiments

Numerical experiments



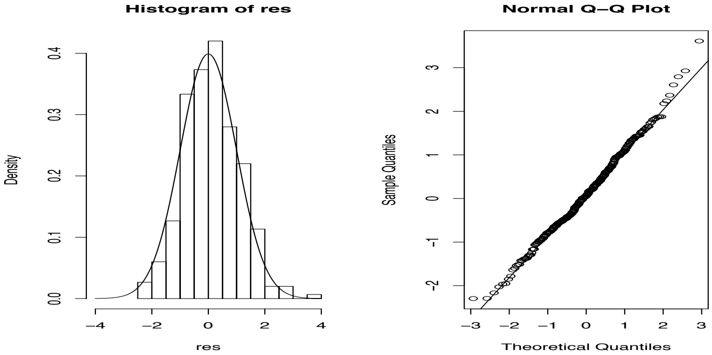
Cubic Rényi entropy $h_{3,0}$ for the Bernoulli d-dimensional distribution; $d = 3$, $Be(p)$-i.i.d. components, $p = 0.8$, sample size $n = 200, N_{sim} = 500$.

Numerical experiments



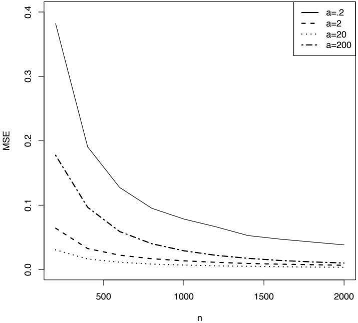
Variability $v = h_{1,1}$ for two Gaussian distributions; $N(0, 3/2)$, $N(2, 1/2)$, $n_1 = 100, n_2 = 200, \epsilon = 1/10, N_{sim} = 300$.

Numerical experiments



Bivariate normal distribution with $N(0,1)$-i.i.d. components; sample size $n = 300, \epsilon = 1/2, N_{sim} = 300$.

Numerical experiments



Bregman distance $B_2(p, q)$ for two exponential distributions $p(x) = \beta_1 e^{-\beta_1 x}, x > 0$, and $q(x) = \beta_2 e^{-\beta_2 x}, x > 0$, with rate parameters $\beta_1 = 1, \beta_2 = 3$, and equal sample size n, with $n\epsilon = a$ for different values of n and a.

Applications

Approximate matching in stochastic databases

Let tables (in a *relational database*) T_1 and T_2 be matrices with m_1 and m_2 i.i.d. random tuples (or records), respectively. The basic database operation *join* combines two tables into a third one by matching values for given columns (attributes).

For the approximate join, we match ϵ-close tuples, say, $d(t_1(j), t_2(i)) \leq \epsilon, t_k(j) \in T_k, k = 1, 2,$ with a specified distance.

The cost of join operations is usually proportional to the size of the intermediate results and so the joining order is a primary target for join-optimizers for multiple (large) tables.

The distribution of the ϵ-join size N_ϵ is thus of importance. With some conditions, it can be shown that the average size

$$\mathrm{E}N_\epsilon = m_1 m_2 q_{1,1,\epsilon} = m_1 m_2 \epsilon^d b_1(d)(q_{1,1} + \mathrm{o}(1)) \text{ as } \epsilon \to 0,$$

that is the asymptotically optimal (in average) pairs of tables are amongst the tables with minimal value of the functional $q_{1,1}$. The estimators of $q_{1,1}$ can be used for samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$.

Entropy maximizing distributions

For a positive definite and symmetric matrix Σ, $s \neq 1$, define the constants

$$m = d + 2/(s-1), \qquad \mathbf{C}_s = (m+2)\Sigma,$$

and

$$A_s = \frac{1}{|\pi \mathbf{C}_s|^{1/2}} \frac{\Gamma(m/2+1)}{\Gamma((m-d)/2+1)}.$$

Among all densities with mean μ and covariance matrix Σ, the Rényi entropy h_s, $s = 2, \ldots,$ is uniquely maximized by the density.

$$p_s^*(x) = \begin{cases} A_s(1-(x-\mu)^T \mathbf{C}_s^{-1}(x-\mu))^{1/(s-1)}, & x \in \Omega_s \\ 0, & x \notin \Omega_s, \end{cases} \qquad (1)$$

with support

$$\Omega_s = \{x \in R^d : (x-\mu)^T \mathbf{C}_s^{-1}(x-\mu) \leq 1\}.$$

Entropy maximizing distributions

The distribution given by $p_s^*(x)$ belongs to the class of *Student-r* distributions.

Let $\mathcal{K}$ be a class of d-dimensional density functions with positive definite covariance matrix.

The proposed estimator of h_s can be used to test the null hypothesis $H_0 : X_1, \ldots, X_n$ *is a sample from a Student-r distribution of type* (1) against the alternative $H_1 : X_1, \ldots, X_n$ *is a sample from any other member of* $\mathcal{K}$.

References

- Källberg, D., Leonenko, N., Seleznjev, O.: Statistical Inference for Rényi Entropy Functionals (submitted) `arXiv:1103.4977v1`
- Leonenko, N., Pronzato, L., Savani, V.: A class of Rényi information estimators for multidimensional densities. Ann. Stat. **36** (2008) 2153–2182. Corrections, Ann. Stat., 38, N6 (2010) 3837-3838
- Leonenko, N., Seleznjev, O.: Statistical inference for the ϵ-entropy and the quadratic Rényi entropy. Jour. Multivariate Analysis **101** (2010) 1981–1994
- Seleznjev, O., Thalheim, B.: Random databases with approximate record matching. Methodol. Comput. Appl. Prob. **12** (2008) 63–89