Efficient multivariate functional estimation and independence testing

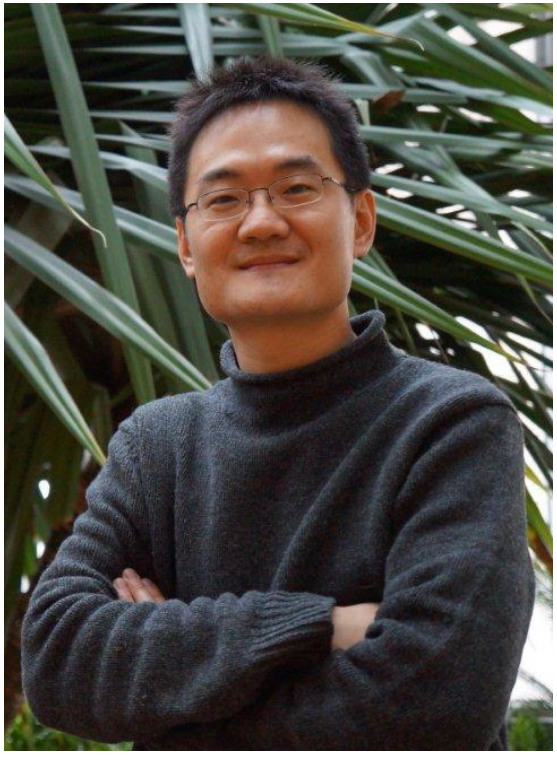
Tom Berrett
University of Cambridge

CREST Statistics Seminar

February 4th, 2019

Collaborators

Material in talk based on joint work with



Richard Samworth and Ming Yuan

Overview

1. Efficient functional estimation

2. Estimation of mutual information and tests of independence

Efficient functional estimation

Density Estimation

Given an independent and identically distributed sample $X_1, \ldots, X_n$ taking values in $\mathbb{R}^d$, a classical question in statistics is to estimate the density f.

The rates of convergence are typically slower than the parametric rate $n^{-1/2}$. For example, over Hölder balls $\Sigma(\beta, L)$ of densities supported on $[0,1]^d$ we have that

$$\inf_{\hat{f}} \sup_{f \in \Sigma(\beta, L)} \mathbb{E} \|\hat{f} - f\|_2 \asymp n^{-\frac{\beta}{2\beta + d}}.$$

Integral functional estimation

In many situations it is not the whole density f we are interested in, but a summary of the form

$$H(f) = \int_{\mathbb{R}^d} f(x)\psi\big(f(x), x\big)\, dx = \mathbb{E}\psi\big(f(X), X\big).$$

The rates of convergence here can be quicker, and in many problems we can find *efficient* estimators which satisfy

$$n^{1/2}(\hat{H}_n - H) \xrightarrow{d} N(0, \sigma_f^2),$$

where σ_f^2 is the best possible variance.

Entropies

For a random variable X with density f we can define different notions of the entropy, such as

- The Shannon entropy

$$H(X) = H(f) = -\int f \log f = -\mathbb{E} \log f(X).$$

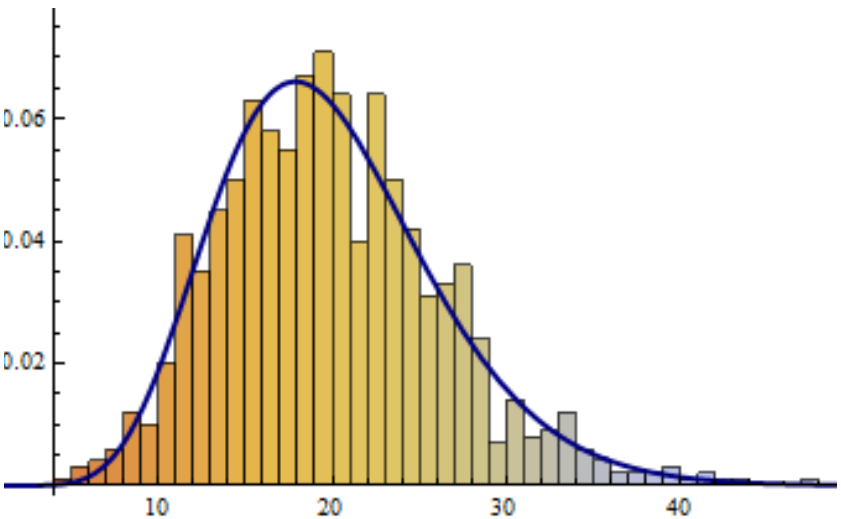
- For $\alpha \neq 1$, the Rényi entropy

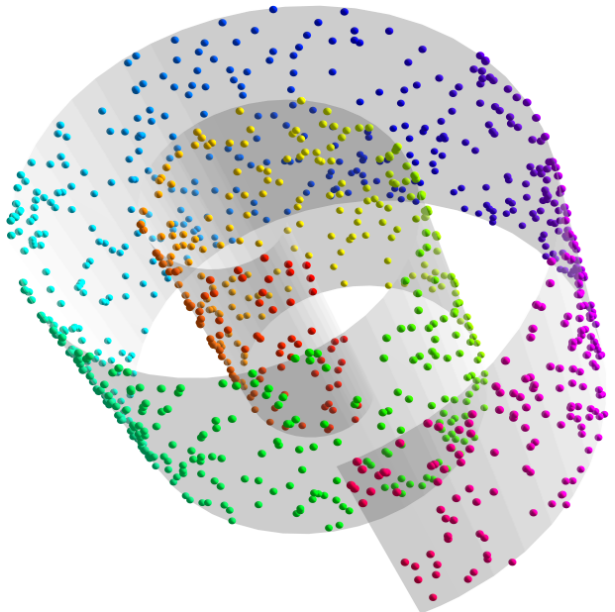
$$H_\alpha(X) = H_\alpha(f) = \frac{1}{1-\alpha} \log\left(\int f^\alpha\right).$$

These are often thought of as measures of the structure or unpredictability of the distribution of X.

Entropy estimation

Applications include tests of normality (Vasicek, 1976), dimension reduction (Huber, 1985), image alignment (Viola and Wells, 1997), independent component analysis (Comon, 1994), estimation of intrinsic dimension (Carter et al., 2010) and estimation of information flows in deep neural networks (Goldfield, Greenewald and Polyanskiy, 2018).





Two-sample functionals

We can also consider the estimation of two-sample functionals of the form

$$T(f,g) = \int_{\mathbb{R}^d} f(x)\phi\big(f(x), g(x), x\big)\, dx,$$

given i.i.d. samples $X_1, \ldots, X_m \sim f$ and $Y_1, \ldots, Y_n \sim g$.

Here we can also consider efficient estimation, and try to find the rate of convergence in m and n.

f-divergences

A class of interesting functionals of this form is the class of $f-$divergences, with

$$T(f,g) = \int_{\mathbb{R}^d} f(x)\varphi\left(\frac{g(x)}{f(x)}\right) dx$$

for some convex φ with $\varphi(1) = 0$.

- KL divergence

$$T(f,g) = \int f \log \frac{f}{g}$$

- Rényi divergence

$$T(f,g) = \frac{1}{\alpha - 1} \log\left(\int f \left(\frac{f}{g}\right)^{\alpha-1}\right).$$

Estimation of such quantities is useful is problems such as two-sample testing (Kanamori et al., 2012) and learning on spaces of distributions (Póczos et al., 2012).

Non-smooth functionals

Many of the functionals we are interested in are *non-smooth* because of their behaviour in low-density regions.

A Taylor expansion of, e.g., Shannon entropy around a density estimator $\hat{f}$ yields

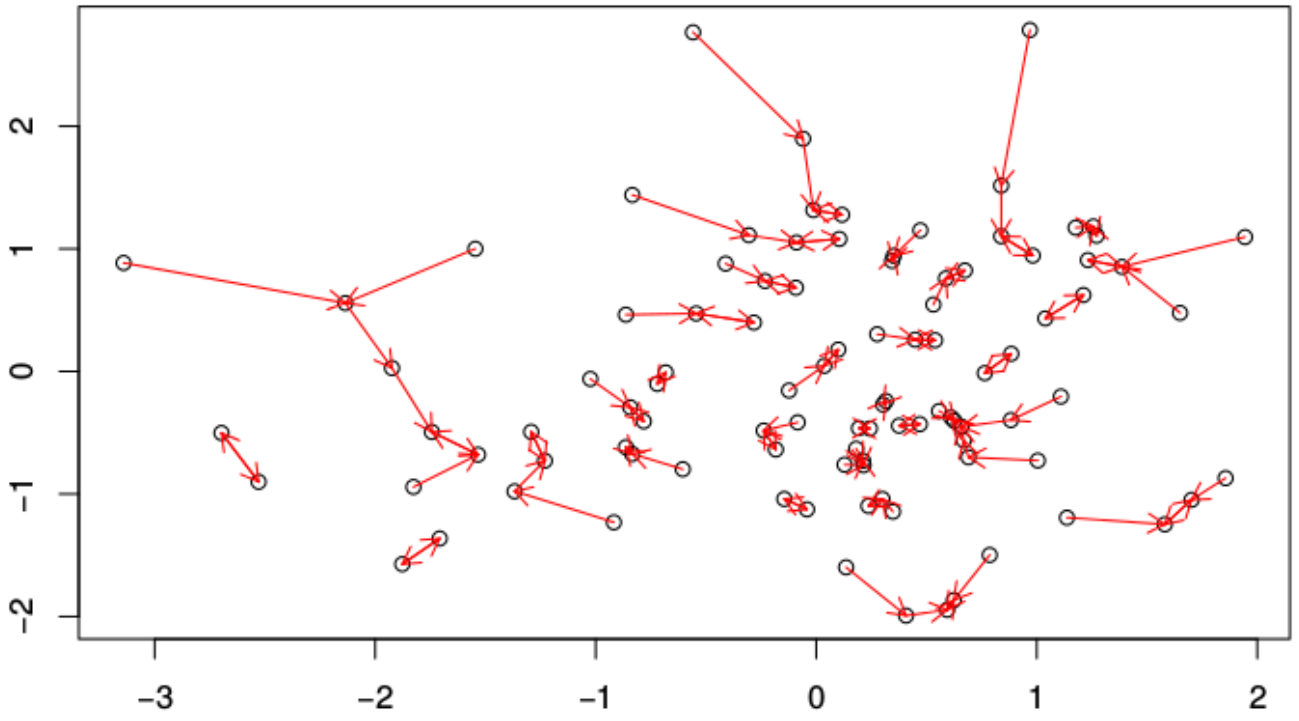
$$\begin{aligned} H(f) &= -\int_{\mathbb{R}^d} f(x) \log f(x)\, dx \\ &\approx -\int_{\mathbb{R}^d} f(x) \log \hat{f}(x)\, dx - \frac{1}{2}\left(\int_{\mathbb{R}^d} \frac{f^2(x)}{\hat{f}(x)}\, dx - 1\right). \end{aligned}$$

When f is bounded away from zero on its support, one can estimate the (smaller order) second term to obtain efficient estimators (Laurent, 1996).

We do not make such assumptions.

Nearest neighbour estimators

Nearest neighbour estimators have proved very popular in the nonparametric statistics literature for the estimation of density functionals.



Writing $\rho_{(k),i} = \|X_{(k),i} - X_i\|$ and $h_x(r) = \mathbb{P}(\|X_1 - x\| \leq r) \approx V_d f(x) r^d$, we have $h_{X_i}(\rho_{(k),i}) \stackrel{d}{=} U_{(k)} \sim \mathrm{Beta}(k, n-k)$, so that $V_d f(X_i) \rho_{(k),i}^d \approx k/n$.

Nearest neighbour estimators

In the one-sample setting we can consider

$$\hat{H}_n = \frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{k}{nV_d\rho_{(k),i}^d}, X_i\right)$$

with $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$ and $\rho_{(k),i} = \|X_{(k)}(X_i) - X_i\|$.

In the two-sample setting we can similarly consider

$$\hat{T}_{m,n} = \frac{1}{m}\sum_{i=1}^{m} \phi\left(\frac{k_X}{mV_d\rho_{(k_X),i,X}^d}, \frac{k_Y}{nV_d\rho_{(k_Y),i,Y}^d}, X_i\right)$$

with $\rho_{(k_X),i,X} = \|X_{(k_X)}(X_i) - X_i\|$ and $\rho_{(k_Y),i,Y} = \|Y_{(k_Y)}(X_i) - X_i\|$.

The Kozachenko–Leonenko estimator

In particular, the *Kozachenko–Leonenko estimator* of Shannon entropy has been extensively studied Kozachenko and Leonenko (1987); Tsybakov and Van der Meulen (1996); Biau and Devroye (2015); Singh and Póczos (2016); Delattre and Fournier (2017); Jiao, Gao and Han (2017); Gao, Oh and Viswanath (2018).

The Kozachenko–Leonenko estimator

In particular, the *Kozachenko–Leonenko estimator* of Shannon entropy has been extensively studied Kozachenko and Leonenko (1987); Tsybakov and Van der Meulen (1996); Biau and Devroye (2015); Singh and Póczos (2016); Delattre and Fournier (2017); Jiao, Gao and Han (2017); Gao, Oh and Viswanath (2018).

$$\hat{H}_{n,(k)} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n V_d \rho_{(k),i}^d}{e^{\Psi(k)}}\right) \approx -\frac{1}{n}\sum_{i=1}^{n} \log f(X_i) =: H_n^*,$$

where $\Psi(k) \sim \log k$ denotes the digamma function.

Weighted Kozachenko–Leonenko estimator

It turns out that, under regularity conditions and when $d \geq 3$, the bias of the standard Kozachenko–Leonkenko estimator satisfies

$$\mathbb{E}\hat{H}_{n,(k)} - H = -\frac{\Gamma(k + 2/d)}{2(d+2)V_d^{2/d}\Gamma(k)n^{2/d}} \int_{\mathbb{R}^d} \frac{\Delta f(z)}{f(z)^{2/d}}\, dx + o\Big(\frac{k^{2/d}}{n^{2/d}}\Big).$$

When $d \geq 4$ this mean that we cannot achieve asymptotic efficiency with this estimator.

Weighted Kozachenko–Leonenko estimator

It turns out that, under regularity conditions and when $d \geq 3$, the bias of the standard Kozachenko–Leonkenko estimator satisfies

$$\mathbb{E}\hat{H}_{n,(k)} - H = -\frac{\Gamma(k+2/d)}{2(d+2)V_d^{2/d}\Gamma(k)n^{2/d}} \int_{\mathbb{R}^d} \frac{\Delta f(z)}{f(z)^{2/d}}\, dx + o\Big(\frac{k^{2/d}}{n^{2/d}}\Big).$$

When $d \geq 4$ this mean that we cannot achieve asymptotic efficiency with this estimator.

We can consider a weighted sum $\hat{H}_n^w = \sum_{j=1}^k w_j H_{n,(j)}$. The bias can be reduced to $o(n^{-1/2})$ by considering $w \in \mathbb{R}^k$ such that

$$\sum_{j=1}^k w_j = 1 \quad \text{and} \quad \sum_{j=1}^k w_j \frac{\Gamma(j+2\ell/d)}{\Gamma(j)} = 0 \ \forall \ell = 1, \ldots, \lfloor d/4 \rfloor.$$

Similar bias reduction can be carried out with more general one- and two-sample functionals.

Controlling smoothness

For a smoothness parameter $\beta > 0$ we can define $m := \lceil \beta \rceil - 1$ and

$$M_{f,\beta}(x) := \inf\Bigg\{ M \geq 1 : \max_{t \in [m]} \left(\frac{\|f^{(t)}(x)\|}{f(x)} \right)^{1/t} \bigvee \sup_{\substack{y,z \in B_x(\{2d^{1/2}M\}^{-1}), \\ y \neq z}} \left(\frac{\|f^{(m)}(z) - f^{(m)}(y)\|}{f(y)\|z - y\|^{\beta - m}} \right)^{1/\beta} \leq M \Bigg\}.$$

This provides a measure of the smoothness of f at x, such that $|f(y)/f(x) - 1| \leq 1/2$ whenever

$$\|y - x\| M_{f,\beta}(x) \leq (8d^{1/2})^{-1/(\beta \wedge 1)}.$$

Classes of densities

For $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \lambda, \nu) \in (0, \infty)^4$ let $\mathcal{F}_d$ denote the set of densities on $\mathbb{R}^d$ and

$$\mathcal{F}_{d,\theta} = \left\{ f \in \mathcal{F}_d : \mu_\alpha(f) \leq \nu, \|f\|_\infty \leq \nu, \int_{\mathbb{R}^d} f(x) \left\{ \frac{M_{f,\beta}(x)^d}{f(x)} \right\}^\lambda \leq \nu \right\},$$

where $\mu_\alpha(f) = \int \|x\|^\alpha f(x)\, dx$ and $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} f(x)$.

Examples

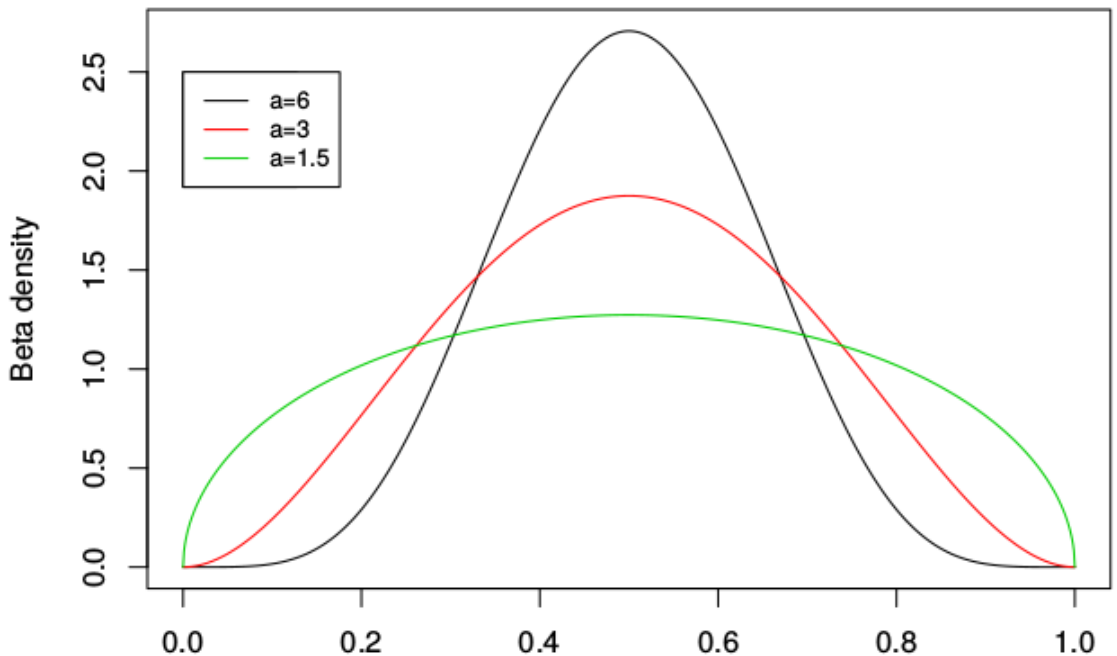
- The $N_d(0, I_d)$ density belongs to $\mathcal{F}_{d,\theta}$ for any $\alpha, \beta > 0, \lambda \in (0, 1)$ and sufficiently large ν.

Examples

- The $N_d(0, I_d)$ density belongs to $\mathcal{F}_{d,\theta}$ for any $\alpha, \beta > 0, \lambda \in (0, 1)$ and sufficiently large ν.
- The multivariate-t density with τ degrees of freedom belongs to $\mathcal{F}_{d,\theta}$ for any $\alpha \in (0, \tau)$, $\beta > 0, \lambda \in (0, \tau/(\tau + d))$ and sufficiently large ν.

Examples

- The $N_d(0, I_d)$ density belongs to $\mathcal{F}_{d,\theta}$ for any $\alpha, \beta > 0, \lambda \in (0, 1)$ and sufficiently large ν.
- The multivariate-t density with τ degrees of freedom belongs to $\mathcal{F}_{d,\theta}$ for any $\alpha \in (0, \tau)$, $\beta > 0, \lambda \in (0, \tau/(\tau + d))$ and sufficiently large ν.



- The Beta-type density $f(x) \propto \|x\|^{a-1}(1 - \|x\|)^{b-1}$ belongs to $\mathcal{F}_{d,\theta}$ for any $\alpha, \beta > 0, \lambda \in (0, b/(b + d - 1))$ and sufficiently large ν.

Assumptions on the functional

Recall in the one-sample setting we estimate $H(f) = \int f(x)\psi(f(x), x)\, dx$. We must make assumptions about the function $\psi(u, x)$.

For $\beta^* > 0$ let $m^* = \lceil \beta^* \rceil - 1$ and suppose that the m^*th partial derivative $\psi_{m^*} = \partial^{m^*}\psi / \partial u^{m^*}$ exists, and that

$$\left| \psi(u + \epsilon u, x) - \sum_{\ell=0}^{m^*} \frac{(u\epsilon)^\ell}{\ell!} \psi_\ell(u, x) \right| \leq L(u^{-\kappa} \vee u^K)|\epsilon|^{\beta^*}$$

for $\epsilon \in (-\epsilon_0, \epsilon_0)$. Also suppose that $|u^\ell \psi_\ell(u, x)| \leq L(u^{-\kappa} \vee u^K)$ for $\ell = 0, 1, \ldots, m^*$.

Suppose that

$$\sup_{\epsilon \in (-r, r)} \sup_{u > 0, x \in \mathbb{R}^d} \max\left\{ \left| \frac{\psi(u, x + \epsilon x)}{\psi(u, x)} - 1 \right|, \left| \frac{\psi_1(u, x + \epsilon x)}{\psi_1(u, x)} - 1 \right| \right\} \to 0$$

as $r \searrow 0$.

Assumption on the functional

This includes weighted Shannon entropy $\psi(u, x) = -w(x) \log u$ for any $\beta^*, \kappa, K > 0$ and uniformly continuous w.

This also includes weighted Rényi entropy $\psi(u, x) = w(x) u^{\alpha-1}$ for any $\beta^* > 0$, $\kappa = (\alpha - 1)_-$, $K = (\alpha - 1)_+$ and uniformly continuous w.

Risk of the weighted estimator

Define

$$\zeta = \max\left(\frac{K}{\lambda}, \frac{\kappa}{\lambda} + \frac{d\kappa}{\alpha}\right)$$

$$\tau = 1 - \max\left(\frac{d}{2\beta}, \frac{d}{2(\beta \wedge 2) + d}, \frac{d}{4\beta^*}, \frac{1}{2\lambda(1-\zeta)}\right)$$

Theorem

Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \lambda, \nu) \in (0, \infty)^4$ with $\zeta < 1/2$ and $\tau > 1/\beta^*$. If $kn^{-1/\beta^*} \to \infty$, $kn^{-\tau} \to 0$ and our weights are chosen suitably then

$$\sup_{f \in \mathcal{F}_{d,\theta}} \left| n\mathbb{E}_f\left\{(\hat{H}_n - H)^2\right\} - V \right| \to 0,$$

where $V(f) := \mathrm{Var}\big(\psi(f(X), X) + f(X)\psi_1(f(X), X)\big)$.

Risk of the weighted estimator

Define

$$\zeta = \max\left(\frac{K}{\lambda}, \frac{\kappa}{\lambda} + \frac{d\kappa}{\alpha}\right)$$

$$\tau = 1 - \max\left(\frac{d}{2\beta}, \frac{d}{2(\beta \wedge 2) + d}, \frac{d}{4\beta^*}, \frac{1}{2\lambda(1-\zeta)}\right)$$

Theorem

Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \lambda, \nu) \in (0, \infty)^4$ with $\zeta < 1/2$ and $\tau > 1/\beta^*$. If $kn^{-1/\beta^*} \to \infty$, $kn^{-\tau} \to 0$ and our weights are chosen suitably then

$$\sup_{f \in \mathcal{F}_{d,\theta}} \left| n\mathbb{E}_f\left\{(\hat{H}_n - H)^2\right\} - V \right| \to 0,$$

where $V(f) := \mathrm{Var}\big(\psi(f(X), X) + f(X)\psi_1(f(X), X)\big)$.

For functionals with $\beta^* = \infty$ and densities with $\lambda = 1$ and $\alpha, \beta = \infty$ the result applies provided $\max(\kappa, K) < 1/2$.

Asymptotic Normality

If, in addition to the previous conditions, we have $V_1 \geq \nu^{-1}$ and $\int f(x)^{2-4\kappa}\,dx \leq \nu$ then

$$\sup_{f \in \bar{\mathcal{F}}_{d,\theta}} d_{\mathrm{K}}\left(\mathcal{L}\left(\frac{n^{1/2}\{\hat{H}_n^w - H(f)\}}{V(f)^{1/2}}\right), N(0,1)\right) \to 0$$

as $n \to \infty$. Here $d_{\mathrm{K}}(F,G) = \sup_{t \in \mathbb{R}} |F(t) - G(t)|$.

This allows us to construct uniformly asymptotically valid confidence intervals for $H(f)$.

Super-oracle phenomenon

For $\alpha \in (1/2, 1)$, consider

$$H_\alpha(f) := \int_{\mathcal{X}} f(x)^\alpha \, dx = \mathbb{E} f(X)^{\alpha-1},$$

for which $\psi(f) = f^{\alpha-1}, f\psi_1(f) = (\alpha-1)f^{\alpha-1}$ and

$$n\mathbb{E}\{(\hat{H}_n - H_\alpha(f))^2\} \to \alpha^2 \mathrm{Var}(f(X)^{\alpha-1}).$$

Remarkably, this outperforms the natural oracle estimator

$$H_n^* = n^{-1} \sum_{i=1}^n f(X_i)^{\alpha-1}.$$

Two-sample functionals

Recall the two-sample functional

$$T(f,g) = \int_{\mathbb{R}^d} f(x)\phi(f(x), g(x), x)\, dx.$$

For a similar weighted nearest neighbour estimator $\hat{T}_{m,n}$, we also have that

$$\frac{\hat{T}_{m,n} - T(f,g)}{m^{-1}V_1(f,g) + n^{-1}V_2(f,g)} \xrightarrow{d} N(0,1),$$

uniformly over suitable classes of densities (f,g) and functions ϕ, where

$$V_1(f,g) := \mathrm{Var}\Big(\phi\big(f(X,g(X),X\big) + f(X)\phi_{10}\big(f(X),g(X),X\big)\Big)$$
$$V_2(f,g) := \mathrm{Var}\Big(f(Y)\phi_{01}\big(f(Y),g(Y),Y\big)\Big).$$

Sketch of asymptotic normality proof

Consider the unweighted estimator

$$\hat{T}_{m,n} = \frac{1}{m}\sum_{i=1}^{m} \phi\left(\frac{k_X}{mV_d\rho^d_{(k_X),i,X}}, \frac{k_Y}{nV_d\rho^d_{(k_Y),i,Y}}, X_i\right).$$

Writing

$$\tilde{\phi}(u,v,x) = \phi(u,v,x) - \phi(u,g(x),x) - \phi(f(x),v,x) + \phi(f(x),g(x),x)$$

we have that

$$\tilde{\phi}(f(x),g(x),x) \equiv \tilde{\phi}_{10}(f(x),g(x),x) \equiv \tilde{\phi}_{01}(f(y),g(y),y) \equiv 0,$$

and so $\tilde{V}_1 = \tilde{V}_2 = 0$.

Sketch of asymptotic normality proof

We can therefore linearise $\hat{T}_{m,n}$, and write

$$\hat{T}_{m,n} - \mathbb{E}\hat{T}_{m,n} = \hat{T}_m^{(1)} - \mathbb{E}\hat{T}_m^{(1)} + \hat{T}_{m,n}^{(2)} - \mathbb{E}\hat{T}_{m,n}^{(2)} + o_p(m^{-1/2} + n^{-1/2}),$$

where

$$\hat{T}_m^{(1)} := \frac{1}{m}\sum_{i=1}^{m} \phi\left(\frac{k_X}{m V_d \rho^d_{(k_X),i,X}}, g(X_i), X_i\right)$$

$$\hat{T}_{m,n}^{(2)} := \frac{1}{m}\sum_{i=1}^{m} \left\{\phi\left(f(X_i), \frac{k_Y}{n V_d \rho^d_{(k_Y),i,Y}}, X_i\right) - \phi\Big(f(X_i), g(X_i), X_i\Big)\right\}.$$

Sketch of asymptotic normality proof

We can therefore linearise $\hat{T}_{m,n}$, and write

$$\hat{T}_{m,n} - \mathbb{E}\hat{T}_{m,n} = \hat{T}_m^{(1)} - \mathbb{E}\hat{T}_m^{(1)} + \hat{T}_{m,n}^{(2)} - \mathbb{E}\hat{T}_{m,n}^{(2)} + o_p(m^{-1/2} + n^{-1/2}),$$

where

$$\hat{T}_m^{(1)} := \frac{1}{m}\sum_{i=1}^{m} \phi\left(\frac{k_X}{mV_d\rho^d_{(k_X),i,X}}, g(X_i), X_i\right)$$

$$\hat{T}_{m,n}^{(2)} := \frac{1}{m}\sum_{i=1}^{m}\left\{\phi\left(f(X_i), \frac{k_Y}{nV_d\rho^d_{(k_Y),i,Y}}, X_i\right) - \phi\left(f(X_i), g(X_i), X_i\right)\right\}.$$

Moreover, $\hat{T}_{m,n}^{(2)} - \mathbb{E}\hat{T}_{m,n}^{(2)} = \hat{T}_n^{(2)} - \mathbb{E}T_n^{(2)} + o_p(m^{-1/2})$, where

$$\hat{T}_n^{(2)} := \mathbb{E}\left[\phi\left(f(X_1), \frac{k_Y}{nV_d\rho^d_{(k_Y),1,Y}}, X_1\right) \,\middle|\, Y_1, \ldots, Y_n\right] - T(f,g).$$

Sketch of asymptotic normality proof

So

$$\hat{T}_{m,n} - \mathbb{E}\,\hat{T}_{m,n} = \hat{T}_m^{(1)} - \mathbb{E}\,\hat{T}_m^{(1)} + \hat{T}_n^{(2)} - \mathbb{E}\,\hat{T}_n^{(2)} + o_p(m^{-1/2} + n^{-1/2}),$$

and we have approximated $\hat{T}_{m,n}$ as the sum of a random variable that only depends on $X_1, \ldots, X_m$ and a random variable that only depends on $Y_1, \ldots, Y_n$.

It remains to show that

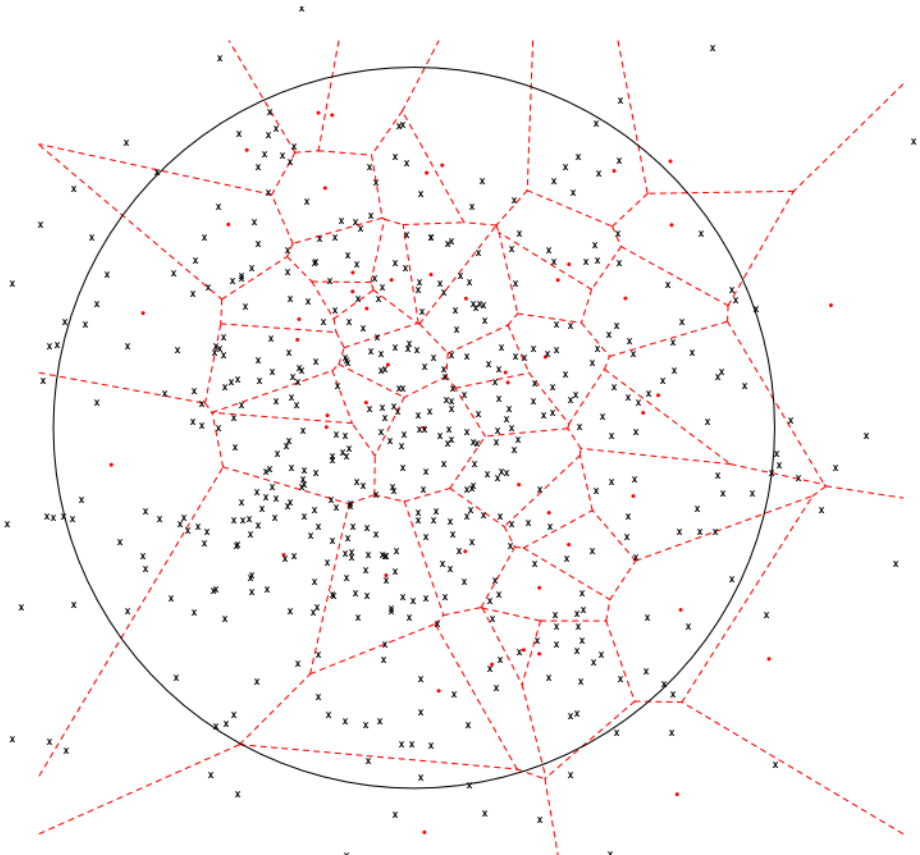
$$m^{1/2}(\hat{T}_m^{(1)} - \mathbb{E}\,\hat{T}_m^{(1)}) \xrightarrow{d} N(0, V_1) \quad \text{and} \quad n^{1/2}(\hat{T}_n^{(2)} - \mathbb{E}\,\hat{T}_n^{(2)}) \xrightarrow{d} N(0, V_2).$$

Sketch of asymptotic normality proof

Consider

$$\hat{T}_m^{(1)} = \frac{1}{m}\sum_{i=1}^{m} \phi\left(\frac{k_X}{m V_d \rho_{(k_X),i,X}^d}, g(X_i), X_i\right).$$

We can partition $\mathbb{R}^d$ using the Voronoi cells associated to a Poisson process with intensity $\frac{m}{k_X} f(\cdot)$. Distant cells are roughly independent.



Local asymptotic minimax lower bound

Fix f satisfying our conditions. For $t \in \mathbb{R}$ and a smooth function p, let

$$f_t(x) := c(t)p\big(th(x)\big)f(x)$$

where $c(t)$ is a normalising constant and

$$h(x) = \psi(f(x), x) + f(x)\psi_1(f(x), x) - \mathbb{E}\big\{\psi(f(X), X) + f(X)\psi_1(f(X), X)\big\}$$

so that $\mathbb{E}h(X)^2 = V(f)$.

Local asymptotic minimax lower bound

Fix f satisfying our conditions. For $t \in \mathbb{R}$ and a smooth function p, let

$$f_t(x) := c(t)p\big(th(x)\big)f(x)$$

where $c(t)$ is a normalising constant and

$$h(x) = \psi(f(x), x) + f(x)\psi_1(f(x), x) - \mathbb{E}\big\{\psi(f(X), X) + f(X)\psi_1(f(X), X)\big\}$$

so that $\mathbb{E}h(X)^2 = V(f)$.

If $\mathcal{I}$ denotes the set of finite subsets of $\mathbb{R}$, then for any estimator sequence $(\tilde{H}_n)$,

$$\sup_{I \in \mathcal{I}} \liminf_{n\to\infty} \max_{t \in I} n\mathbb{E}_{f_{n^{-1/2}t}}\Big[\big\{\tilde{H}_n - H(f_{n^{-1/2}t})\big\}^2\Big] \geq V(f).$$

An analogous result holds for two-sample functionals.

Local asymptotic minimax lower bound

To conclude that our estimators are efficient we must show that they achieve this lower bound. It suffices to show that $f_t(\cdot)$ satisfies our conditions for t sufficiently small.

This is true for Shannon and Rényi entropies, KL divergence and Rényi divergence and other, sufficiently regular, functionals.

Summary

- Standard nearest neighbour estimators can be efficient for $d \leq 3$, but are typically not when $d \geq 4$.
- By incorporating weights to cancel bias terms, we obtain efficient estimators in arbitrary dimensions, subject to appropriate regularity conditions.

Independence testing

Independence testing

Measuring dependence and testing independence are fundamental problems in statistics, and are essential for model building, certain goodness-of-fit tests, feature selection, independent component analysis and more.

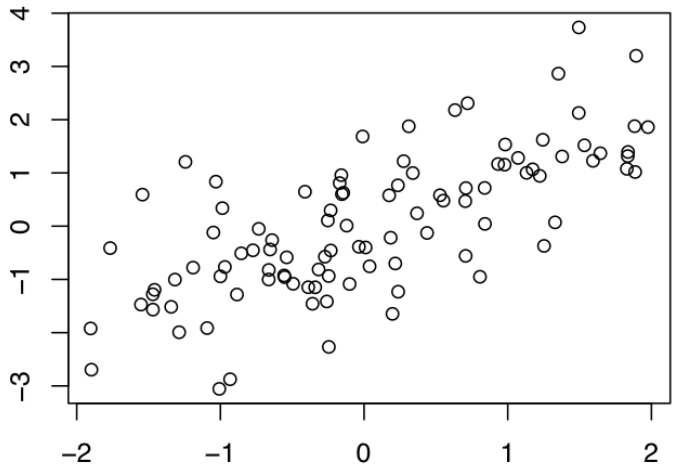
Independence testing

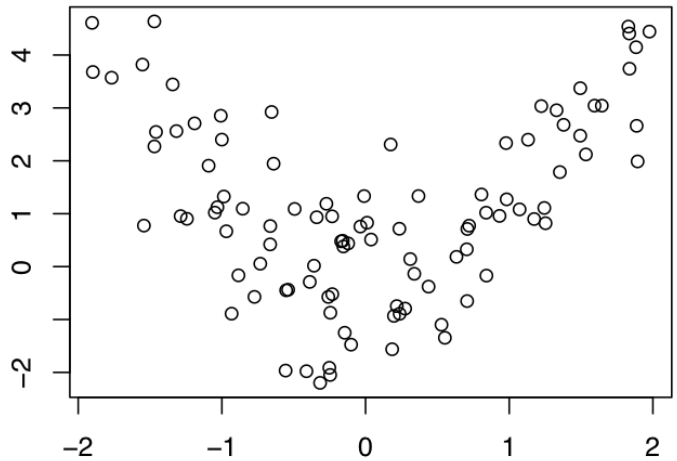
Measuring dependence and testing independence are fundamental problems in statistics, and are essential for model building, certain goodness-of-fit tests, feature selection, independent component analysis and more.

Classical measures include:

- Pearson's correlation (e.g. Pearson, 1920);
- Kendall's tau (Kendall, 1938);
- Hoeffding's D (Hoeffding, 1948).

These are limited to linear or monotonic dependence, or bivariate settings.





Independence testing

Modern datasets often exhibit complex dependence which is not well captured by these classical measures;

Independence testing

Modern datasets often exhibit complex dependence which is not well captured by these classical measures; see examples in bioinformatics (Steuer et al., 2002), climate science (Donges et al., 2009), neuroscience (Vicente et al., 2011), computer security (Amiri et al., 2011) and linguistics (Nguyen and Eisenstein, 2017).

Independence testing

Modern datasets often exhibit complex dependence which is not well captured by these classical measures; see examples in bioinformatics (Steuer et al., 2002), climate science (Donges et al., 2009), neuroscience (Vicente et al., 2011), computer security (Amiri et al., 2011) and linguistics (Nguyen and Eisenstein, 2017).

As a result, many new measures and tests have been proposed and studied recently:

- Distance covariance (Székely, Rizzo and Bakirov, 2007; Székely and Rizzo, 2013);
- RKHS norms (Bach and Jordan, 2002; Gretton et al., 2005; Sejdinovic et al., 2013);
- Multivariate rank-based tests (Weihs, Drton and Meinshausen, 2018);
- Empirical copula processes (Kojadinovic and Holmes, 2009);
- Sample space partitioning (Gretton and Györfi, 2010; Heller et al., 2016).

Independence testing

Modern datasets often exhibit complex dependence which is not well captured by these classical measures; see examples in bioinformatics (Steuer et al., 2002), climate science (Donges et al., 2009), neuroscience (Vicente et al., 2011), computer security (Amiri et al., 2011) and linguistics (Nguyen and Eisenstein, 2017).

As a result, many new measures and tests have been proposed and studied recently:

- Distance covariance (Székely, Rizzo and Bakirov, 2007; Székely and Rizzo, 2013);
- RKHS norms (Bach and Jordan, 2002; Gretton et al., 2005; Sejdinovic et al., 2013);
- Multivariate rank-based tests (Weihs, Drton and Meinshausen, 2018);
- Empirical copula processes (Kojadinovic and Holmes, 2009);
- Sample space partitioning (Gretton and Györfi, 2010; Heller et al., 2016).

Each of these has its own advantages and disadvantages, and no universally accepted measure exists.

Problem statement

Let $Z = (X, Y)$ have a density f with repect to Lebesgue measure on $\mathbb{R}^d$, and let f_X and f_Y be the marginal densities of X and Y with respect to Lebesgue measure on $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$ respectively.

Given independent and identically distributed observations $Z_1, \ldots, Z_n$ of Z, we wish to test the hypotheses

$$H_0 : X \perp\!\!\!\perp Y \quad \text{vs.} \quad H_1 : X \not\perp\!\!\!\perp Y.$$

Mutual information

We measure dependence by the mutual information (Shannon, 1948)

$$I(X;Y) = \int \int f(x,y) \log \frac{f(x,y)}{f_X(x) f_Y(y)} \, dx \, dy.$$

Mutual information

We measure dependence by the mutual information (Shannon, 1948)

$$I(X;Y) = \int \int f(x,y) \log \frac{f(x,y)}{f_X(x) f_Y(y)} \, dx \, dy.$$

This is the KL divergence between f and $f_X f_Y$, so is non-negative and zero if and only if $X \perp\!\!\!\perp Y$.

Mutual information

We measure dependence by the mutual information (Shannon, 1948)

$$I(X;Y) = \int \int f(x,y) \log \frac{f(x,y)}{f_X(x) f_Y(y)} \, dx \, dy.$$

This is the KL divergence between f and $f_X f_Y$, so is non-negative and zero if and only if $X \perp\!\!\!\perp Y$.

A consequence of the data processing inequality is that

$$I(\phi(X);Y) = I(X;Y)$$

whenever X and Y are conditionally independent given $\phi(X)$ (e.g. Kinney and Atwal, 2014). Mutual information is *self-equitable*.

Mutual information and entropy

Provided $H(X), H(Y)$ and $H(X, Y)$ are finite, we can write

$$I(X; Y) = H(X) + H(Y) - H(X, Y).$$

So, if we can estimate entropies then we can estimate mutual information.

Estimation of mutual information

We may estimate $I(X;Y)$ using

$$\hat{I}_n = \hat{H}_n^X + \hat{H}_n^Y - \hat{H}_n^Z,$$

where, e.g., $\hat{H}_n^Z = \hat{H}_{n,k}^{w_Z}(Z_1, \ldots, Z_n)$ is a weighted Kozachenko–Leonenko estimator of $H(Z)$.

Estimation of mutual information

We may estimate $I(X;Y)$ using

$$\hat{I}_n = \hat{H}_n^X + \hat{H}_n^Y - \hat{H}_n^Z,$$

where, e.g., $\hat{H}_n^Z = \hat{H}_{n,k}^{w_Z}(Z_1, \ldots, Z_n)$ is a weighted Kozachenko–Leonenko estimator of $H(Z)$.

By previous theory we have

$$n^{1/2}\{\hat{I}_n - I(X;Y)\} \xrightarrow{d} N\big(0, V(X;Y)\big),$$

where $V(X;Y) = \mathrm{Var} \log \frac{f(X,Y)}{f_X(X) f_Y(Y)}$, for suitable choices of k and weights.

One known marginal

Suppose f_Y known. Generate $\{Y_i^{(b)} : i = 1, \ldots, n, b = 1, \ldots, B\}$ and calculate

$$\hat{I}_n^{(b)} := \hat{I}_n\big((X_1, Y_1^{(b)}), \ldots, (X_n, Y_n^{(b)})\big).$$

One known marginal

Suppose f_Y known. Generate $\{Y_i^{(b)} : i = 1, \ldots, n, b = 1, \ldots, B\}$ and calculate

$$\hat{I}_n^{(b)} := \hat{I}_n\big((X_1, Y_1^{(b)}), \ldots, (X_n, Y_n^{(b)})\big).$$

We can now estimate a critical value for our test by

$$\hat{C}_q^{(n),B} = \inf\left\{ r \in \mathbb{R} : 1 + \sum_{b=1}^{B} \mathbb{1}_{\{\hat{I}_n^{(b)} \geq r\}} \leq (B+1)q \right\},$$

the $(1-q)$th quantile of $\{\hat{I}_n, \hat{I}_n^{(1)}, \ldots, \hat{I}_n^{(B)}\}$. We refer to the test that rejects H_0 if and only if $\hat{I}_n > \hat{C}_q^{(n),B}$ by `MINTknown`(q).

Power of MINTknown

We may use earlier results on entropy estimation to perform a local power analysis on MINTknown. For $d_X, d_Y \in \mathbb{N}$ and $\vartheta = (\theta, \theta_Y)$ define

$$\mathcal{F}_{d_X,d_Y,\vartheta} := \left\{(f, g_Y) \in \mathcal{F}_{d_X+d_Y,\theta} \times \mathcal{F}_{d_Y,\theta_Y} : f_Y \in \mathcal{F}_{d_Y,\theta_Y}, f_X g_Y \in \mathcal{F}_{d_X+d_Y,\theta}\right\}$$

and, for $b \geq 0$, let

$$\mathcal{F}_{d_X,d_Y,\vartheta}(b) = \left\{(f, g_Y) \in \mathcal{F}_{d_X,d_Y,\vartheta} : I(f) > b\right\}.$$

Power of MINTknown

We may use earlier results on entropy estimation to perform a local power analysis on MINTknown. For $d_X, d_Y \in \mathbb{N}$ and $\vartheta = (\theta, \theta_Y)$ define

$$\mathcal{F}_{d_X,d_Y,\vartheta} := \left\{(f, g_Y) \in \mathcal{F}_{d_X+d_Y,\theta} \times \mathcal{F}_{d_Y,\theta_Y} : f_Y \in \mathcal{F}_{d_Y,\theta_Y}, f_X g_Y \in \mathcal{F}_{d_X+d_Y,\theta}\right\}$$

and, for $b \geq 0$, let

$$\mathcal{F}_{d_X,d_Y,\vartheta}(b) = \left\{(f, g_Y) \in \mathcal{F}_{d_X,d_Y,\vartheta} : I(f) > b\right\}.$$

Theorem

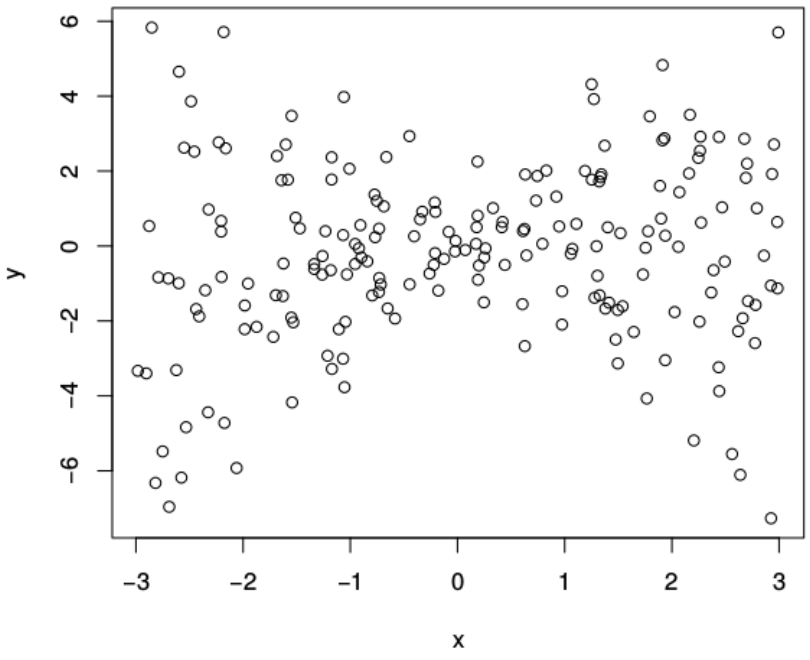
For suitable ϑ and choices of tuning parameters there exists a sequence (b_n) such that $b_n = o(n^{-1/2})$ and for each $q \in (0, 1)$

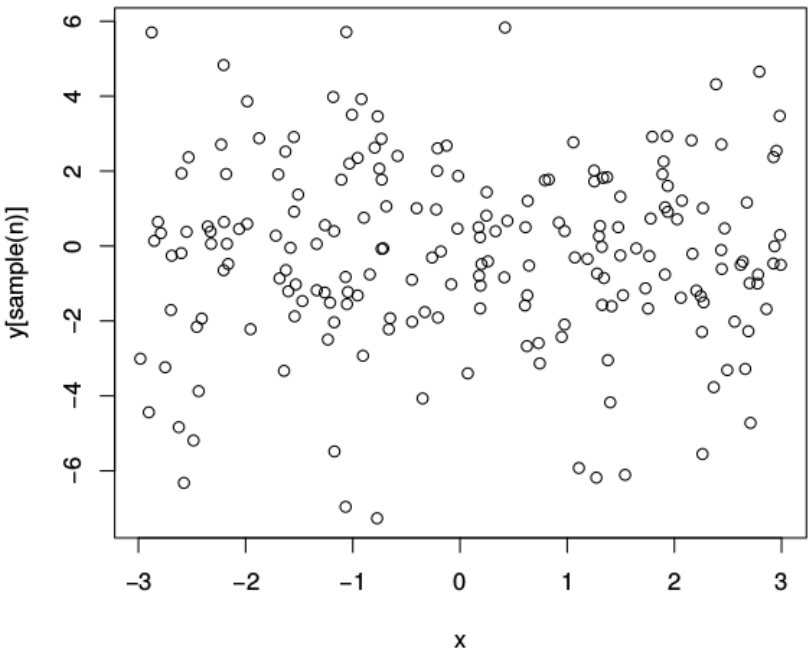
$$\inf_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}(b_n)} \mathbb{P}_f(\hat{I}_n > \hat{C}_q^{(n),B}) \to 1.$$

Permutation test

If we do not have an approximation to either marginal distribution then we may instead use a permutation test. We generate $\pi_1, \ldots, \pi_B$ uniformly from the permutation group S_n and calculate

$$\hat{I}_n^{(b)} = \hat{I}_n\big((X_1, Y_{\pi_b(1)}), \ldots, (X_n, Y_{\pi_b(n)})\big)$$

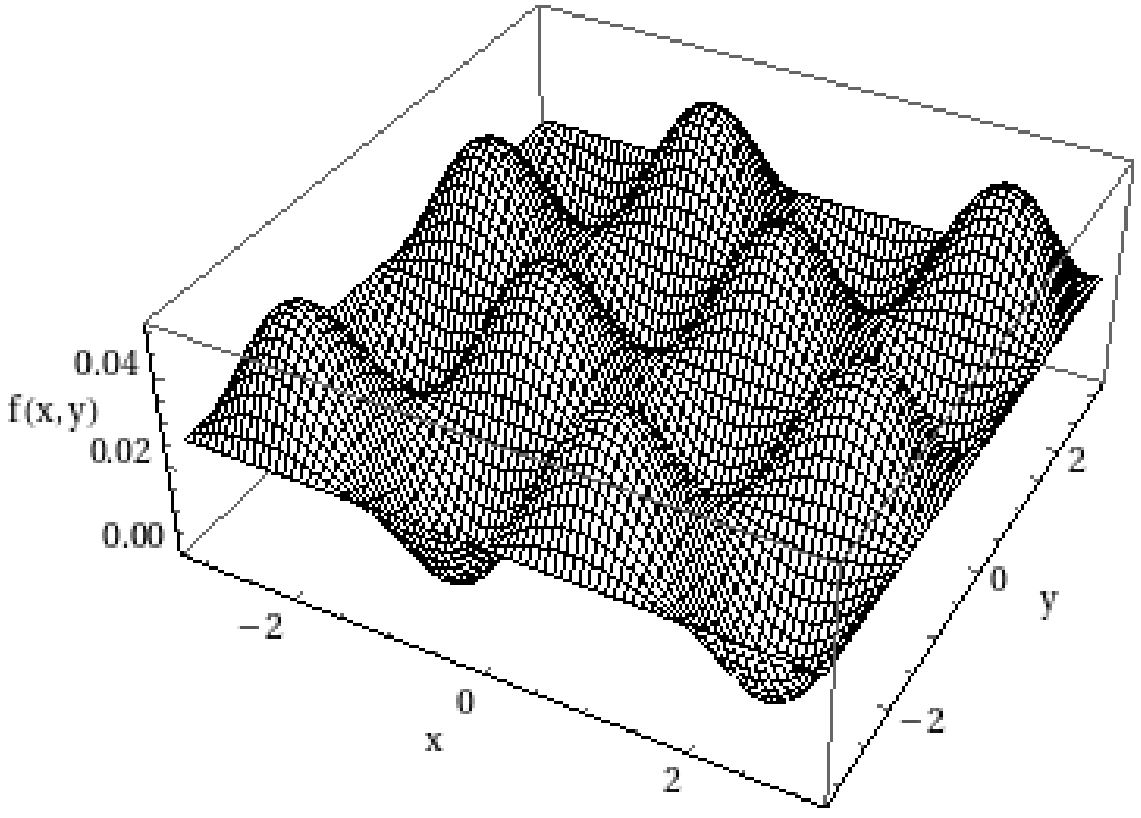




Practical performance

Due to the local nature of our test statistics, we find that `MINT` tends to perform well in settings in which the dependence is local, or in which the scale of the dependence is different to the scale of the marginal distributions.

Sinusoidal data



$$f_l(x,y) = \frac{1}{4\pi^2}\{1 + \sin(lx)\sin(ly)\} \quad \text{for } l = 1, 2, \ldots$$

This example was identified by Sejdinovic et al. (2013) as challenging for independence testing.

Simulation study

In the following we present power curves for `MINT` and `MINTknown` with oracle choices of k, k_Y, as well as power curves for `MINTav`, in which we average over $k \in \{1, \ldots, 20\}$ in `MINT`. In all cases we take $B = 100$.

Simulation study

In the following we present power curves for `MINT` and `MINTknown` with oracle choices of k, k_Y, as well as power curves for `MINTav`, in which we average over $k \in \{1, \ldots, 20\}$ in `MINT`. In all cases we take $B = 100$.

For comparison we present the power curves for tests based on:

- Empirical copula processes in the `R` package `copula` (Hofert et al., 2017);
- RKHS methods in the `R` package `dHSIC` (Pfister and Peters, 2017);
- Distance covariance in the `R` package `energy` (Rizzo and Szekely, 2017);
- a multivariate extension of Hoeffding's D in the `R` package `SymRC` (Weihs et al., 2017).

Simulation study

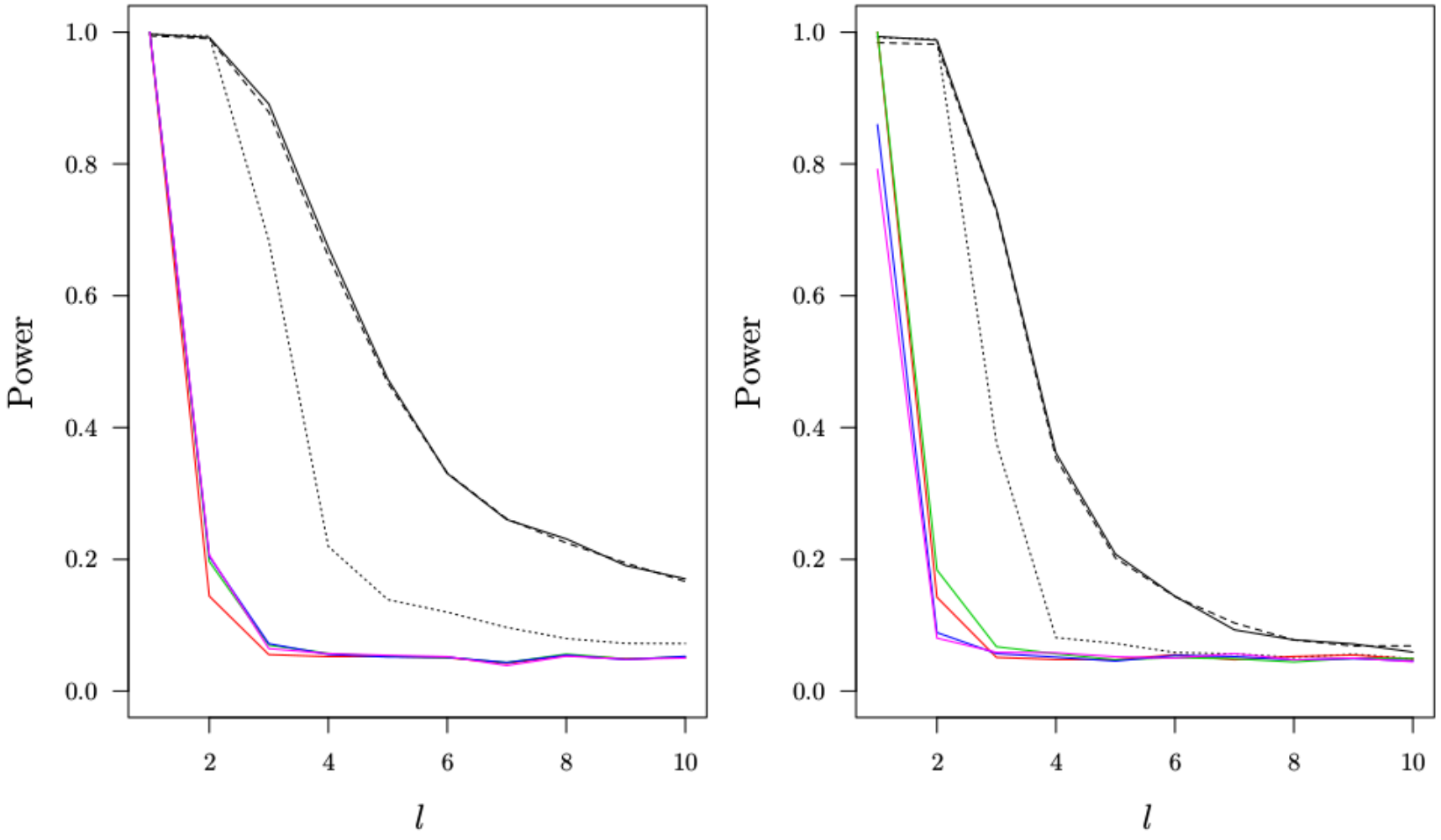
In the following we present power curves for `MINT` and `MINTknown` with oracle choices of k, k_Y, as well as power curves for `MINTav`, in which we average over $k \in \{1, \ldots, 20\}$ in `MINT`. In all cases we take $B = 100$.

For comparison we present the power curves for tests based on:

- Empirical copula processes in the `R` package `copula` (Hofert et al., 2017);
- RKHS methods in the `R` package `dHSIC` (Pfister and Peters, 2017);
- Distance covariance in the `R` package `energy` (Rizzo and Szekely, 2017);
- a multivariate extension of Hoeffding's D in the `R` package `SymRC` (Weihs et al., 2017).

We present settings in which (X, Y) have sinusoidal distributions, as well as a multivariate setting (X_1, X_2, Y_1, Y_2) in which (X_1, Y_1) have the sinusoidal distributions and $X_2, Y_2 \in \mathrm{U}[0, 1]$ are independent.

Results



Power curves as functions of the respective shape parameters for `MINT` (——), `MINTknown` (- - -), `MINTav` (· · · · ·), HSIC (——), Distance covariance (——), Copula (——), Hoeffding's D (——). The marginals are univariate (left) and bivariate (right).

Summary

Using recently-developed efficient entropy estimators we have constructed an independence test based on mutual information. This test has good theoretical properties in arbitrary dimensions and we have shown that it can perform well in practice.

Summary

Using recently-developed efficient entropy estimators we have constructed an independence test based on mutual information. This test has good theoretical properties in arbitrary dimensions and we have shown that it can perform well in practice.

The ideas easily generalise to the estimation of conditional mutual information $I(X; Y|W)$ and $I(X_1; X_2; \ldots; X_p)$, and to the testing of conditional independence and mutual independence between p random vectors.

Summary

- Nearest neighbour methods offer very intuitive, computationally feasible approaches for many nonparametric problems
- Our understanding of their theoretical properties is improving rapidly, but there is still more to be done!

References

B., Samworth, R. J. and Yuan, M. (2019) Efficient multivariate entropy estimation via k-nearest neighbour distances. *Ann. Statist.*, **47**, 288–318.

B. and Samworth, R. J. (2019) Nonparametric independence testing via mutual information. *Biometrika, to appear*.

B., Grose, D. J. and Samworth, R. J. (2018) **IndepTest**: nonparametric independence tests based on entropy estimation. Available at https://cran.r-project.org/web/packages/IndepTest/index.html.

B. and Samworth, R. J. (2019) Efficient functional estimation and the super-oracle phenomenon. *In preparation.*

Thank you!

References

Amiri, F., Yousefi, M. R., Lucas, C., Shakery, A. and Yazdani, N. (2011) Mutual information-based feature selection for intrusion detection systems. *J. Netw. Comput. Appl.*, **34**, 1184–1199.

Bach, F. R. and Jordan, M. I. (2002) Kernel independent component analysis. *J. Mach. Learn. Res.*, **3**, 1–48.

Biau, G. and Devroye, L. (2015) *Lectures on the Nearest Neighbor Method*. Springer, New York.

Carter, K. M., Raich, R. and Hero, A. O. (2010) On local intrinsic dimension estimation and its applications. *IEEE Trans. Signal Process.*, **58**, 650–663.

Comon, P. (1994). Independent component analysis, a new concept?. *Signal Process.*, **36**, 287–314.

Delattre, S. and Fournier, N. (2017) On the Kozachenko–Leonenko entropy estimator. *J. Statist. Plann. Inf.*, **185**, 69–93.

Donges, J. F., Zou, Y., Marwan, N. and Kurths, J. (2009) Complex networks in climate dynamics. *Eur. Phys. J. Special Topics*, **174**, 157–179.

References

Gao, W., Oh, S. and Viswanath, P. (2016) Demystifying fixed k-nearest neighbor information estimators. *IEEE Trans. Inf. Theory.*, **64**, 5629–5661

Goldfield, Z., Greenewald, K. and Polyanskiy, Y. (2018). Estimating differential entropy under Gaussian convolutions. Available at `arXiv:1810.11589`.

Gretton A., Bousquet O., Smola A. and Schölkopf B. (2005) Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Algorithmic Learning Theory*, 63–77.

Gretton, A. and Györfi, L. (2010). Consistent nonparametric tests of independence. *J. Mach. Learn. Res.*, **11**, 1391–423.

Heller, R., Heller, Y., Kaufman, S., Brill, B. and Gorfine, M. (2016) Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.*, **17**, 1–54.

Hoeffding, W. (1948) A non-parametric test of independence. *Ann. Math. Statist.*, **19**, 546–57.

References

Hofert, M., Kojadinovic, I., Mächler, M. & Yan, J. (2017) copula: Multivariate Dependence with Copulas. *R Package version 0.999-18*. https://cran.r-project.org/web/packages/copula/index.html.

Huber, P. (1985) Projection pursuit. *Ann. of Statist.*, **13**, 435–475.

Jiao, J., Gao, W. and Han, Y. (2017) The nearest neighbor information estimator is adaptively near minimax rate-optimal. Available at `arXiv:1711.08824`.

Kanamori, T., Suzuki, T. and Sugiyama, M. (2012) f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE. Trans. Inf. Theory*, **58**, 708–720.

Kendall, M. G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.

Kinney, J. B. & Atwal, G. S. (2014) Equitability, mutual information, and the maximal information coefficient. *Proc. Nat. Acad. Sci.*, **111**, 3354–9.

References

Kojadinovic, I. and Holmes, M. (2009) Tests of independence among continuous random vectors based on Cramér–von Mises functionals of the empirical copula process. *J. Multivariate Anal.*, **100**, 1137–54.

Kozachenko, L. F. and Leonenko, N. N. (1987) Sample estimate of the entropy of a random vector. *Probl. Inform. Transm.*, **23**, 95–101.

Laurent, B. (1996) Efficient estimation of integral functionals of a density. *Ann. Statist.*, **24**, 659–681.

Nguyen, D. and Eisenstein, J. (2017) A kernel independence test for geographical language variation. *Comput. Ling.*, **43**, 567–592.

Pearson, K. (1920) Notes on the history of correlation. *Biometrika*, **13**, 25–45.

Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2017) Kernel-based tests for joint independence. *J. Roy. Statist. Soc., Ser. B*.

Pfister, N. and Peters, J. (2017). `dHSIC`: Independence Testing via Hilbert Schmidt Independence Criterion. `R` package version 2.0, `https://cran.r-project.org/web/packages/dHSIC`.

References

Póczos, B., Xiong, L. and Schneider, J. (2012) Nonparametric divergence estimation with applications to machine learning on distributions. Available at `arXiv:1202.3758`.

Rizzo, M. L. and Szekely, G. J. (2017). energy: E-Statistics: Multivariate Inference via the Energy of Data. *R Package version 1.7-2*. https://cran.r-project.org/web/packages/energy/index.html.

Singh, S. and Póczos, B. (2016) Analysis of k nearest neighbor distances with application to entropy estimation. *NIPS*, **29**, 1217–1225.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, **41**, 2263–2291.

Shannon, C. E. (1948) A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

Steuer, R., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**, 231–40.

References

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.

Székely, G. J. and Rizzo, M. L. (2013) The distance correlation t-test of independence in high dimension. *J. Multivariate Anal.*, **117**, 193–213.

Tsybakov, A. B. and Van der Meulen, E. C. (1996) Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Stat.*, **23**, 75–83.

Vasicek, O. (1976) A test for normality based on sample entropy. *J. Roy. Statist. Soc., Ser. B.*, **38**, 54–59.

Vicente, R., Wibral, M., Lindner, M. and Pipa, G. (2011) Transfer entropy – a model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.*, **30**, 45–67.

Viola, P. and Wells, W. M. (1997) Alignment by maximization of mutual information (1997). *Int. J. Comput. Vis.*, **24**, 137–154.

References

Weihs, L., Drton, M. and Meinshausen, N. (2018) Symmetric rank covariances: a generalised framework for nonparametric measures of dependence. *Biometrika*, **105**, 547–562.

Weihs, L., Drton, M. & Meinshausen, N. (2017) SymRC: Estimating symmetric rank covariances. `https://github.com/Lucaweihs/SymRC`.