Efficient multivariate entropy estimation and independence testing

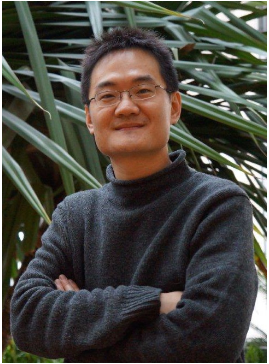
Tom Berrett
University of Cambridge

ECARES Statistics Seminar

November 22nd, 2018

Collaborators

Material in talk based on joint work with



Richard Samworth and Ming Yuan

Overview

Efficient entropy estimation

Entropy

For a random variable X with density f we define the *(differential) entropy* of X to be

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

Entropy

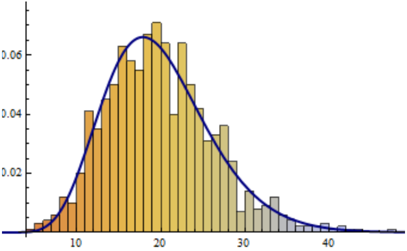
For a random variable X with density f we define the *(differential) entropy* of X to be

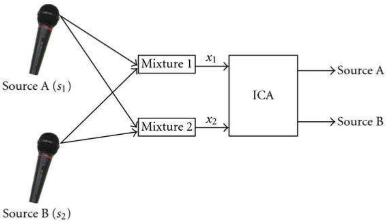
$$H(X) = H(f) = -\int f \log f = -\mathbb{E} \log f(X).$$

The quantity $-\log f(X)$ is often thought of as the information content of the observation X, and H as a measure of the 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) and estimation of information flows in deep neural networks (Goldfield, Greenewald and Polyanskiy, 2018).





The Kozachenko–Leonenko estimator

The *Kozachenko–Leonenko estimator* in particular has proved very popular in the nonparametric statistics literature 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

The *Kozachenko–Leonenko estimator* in particular has proved very popular in the nonparametric statistics literature 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).

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

where $\rho_{(k),i} = \|X_i - X_{(k),i}\|$ is the kth-nearest neighbour distance of X_i, $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$ denotes the volume of the unit d-dimensional Euclidean ball and $\Psi(k) \sim \log k$ denotes the digamma function.

Compact support

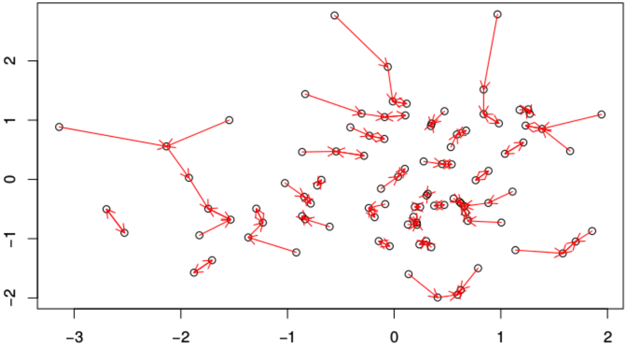
A Taylor expansion of $H(f)$ around a density estimator $\hat{f}$ yields

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

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

Nearest neighbours

Recall $\rho_{(k),i} = \|X_{(k),i} - X_i\|$. Write $h_x(r) = \mathbb{P}(\|X_1 - x\| \leq r) \approx V_d f(x) r^d$.



We have $h_{X_i}(\rho_{(k),i}) \overset{d}{=} U_{(k)} \sim \mathrm{Beta}(k, n-k)$, so that $V_d f(X_i) \rho^d_{(k),i} \approx k/n$.

Intuition on bias

Because $h_{X_i}(\rho_{(k),i}) \sim \mathrm{Beta}(k, n-k)$ and $V_d f(x) h_x^{-1}(s)^d \approx s$ we have

$$\begin{aligned}
\mathbb{E}\hat{H}_{n,(k)} &= \int_{\mathcal{X}} f(x) \int_0^1 \log\left(\frac{(n-1)V_d h_x^{-1}(s)^d}{e^{\Psi(k)}}\right) \mathrm{B}_{k,n-k}(s)\, ds\, dx \\
&\approx \int_{\mathcal{X}} f(x) \int_0^1 \log\left(\frac{(n-1)s}{e^{\Psi(k)} f(x)}\right) \mathrm{B}_{k,n-k}(s)\, ds\, dx \\
&= H(f) + \log(n-1) - \Psi(n).
\end{aligned}$$

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

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

Controlling smoothness

We now introduce our assumptions on f.

For $\theta = (\alpha, \beta, \rho, \nu) \in (0, \infty)^4$ and an $m = \lceil \beta \rceil - 1$ times differentiable density f set

$$M_{f,\theta}(x) := \max_{t=1,\ldots,m} \left(\frac{\|f^{(t)}(x)\|}{f(x)} \right)^{\frac{1}{t}} \vee \sup_{y \in B_x^\circ(r_0(x))} \left(\frac{\|f^{(m)}(y) - f^{(m)}(x)\|}{f(x)\|y - x\|^{\beta - m}} \right)^{\frac{1}{\beta}} \Bigg\},$$

where $r_0(x) = f(x)^\rho / (2\nu d^{1/2})$.

This provides a measure of the smoothness of f at x.

Classes of densities

As well as controlling the smoothness of f we also need to control the tails. Let $\mu_\alpha(f) := \int_{\mathbb{R}^d} \|z\|^\alpha f(z)\, dz$.

Classes of densities

As well as controlling the smoothness of f we also need to control the tails. Let $\mu_\alpha(f) := \int_{\mathbb{R}^d} \|z\|^\alpha f(z)\, dz$.

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

$$\mathcal{F}_{d,\theta} = \Big\{ f \in \mathcal{F}_d : \mu_\alpha(f) \leq \nu, \|f\|_\infty \leq \nu, \sup_{x : f(x) \geq \delta} M_{f,a,\beta}(x) \leq \frac{\nu}{\delta^\rho}\, \forall \delta > 0 \Big\}.$$

In comparison with a Hölder smoothness condition, when $\rho < 1$ we require f to vary less in the tails of the distribution and ρ controls the strength of this requirement.

Examples

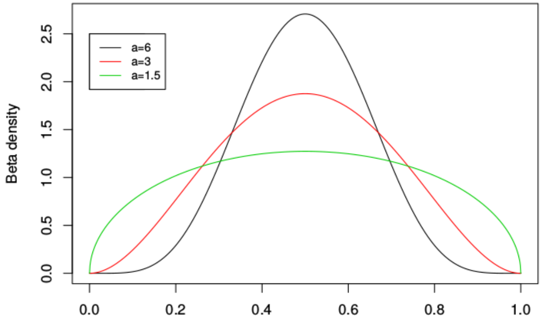
- The $N_d(0, I_d)$ density belongs to $\mathcal{F}_{d,\theta}$ for any $\alpha, \beta, \rho > 0$ and sufficiently large ν.

Examples

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

Examples

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



Bias of the weighted estimator

Fix $d \in \mathbb{N}$ and $\theta \in (0,\infty)^4$ and let $k^* = k_n^*$ be such that $k^* = O(n^{1-\epsilon})$ for some $\epsilon > 0$. If w is chosen suitably then

$$\sup_{f \in \mathcal{F}_{d,\theta}} \Big| \mathbb{E}_f\big(\hat{H}_n^w\big) - H(f) \Big| = O\left(\max\left\{ \frac{k^{\frac{\alpha}{(\alpha+d)(1+\rho d)}-\epsilon}}{n^{\frac{\alpha}{(\alpha+d)(1+\rho d)}-\epsilon}}, \frac{k^{\frac{2(\lfloor d/4 \rfloor+1)}{d}}}{n^{\frac{2(\lfloor d/4 \rfloor+1)}{d}}}, \frac{k^{\beta/d}}{n^{\beta/d}} \right\} \right)$$

uniformly for $k \in \{1,\ldots,k^*\}$.

Variance of the weighted estimator

Let $V(f) = \int f \log^2 f - H(f)^2$. When $\zeta = \frac{2\alpha}{(\alpha+d)(1+\rho d)} > 1$ we choose

$$\tau_1 < \min\left(\frac{\beta}{d+\beta}, \frac{\zeta}{\zeta+2}, \frac{\zeta-1}{\zeta}\right).$$

Let $k_0^* = k_{0,n}^*$ and $k_1^* = k_{1,n}^*$ satisfy $k_0^* \leq k_1^*$, $k_0^*/\log^5 n \to \infty$ and $k_1^* = O(n^{\tau_1})$. Then

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \left| n \mathrm{Var}_f \hat{H}_n^w - V(f) \right| \to 0.$$

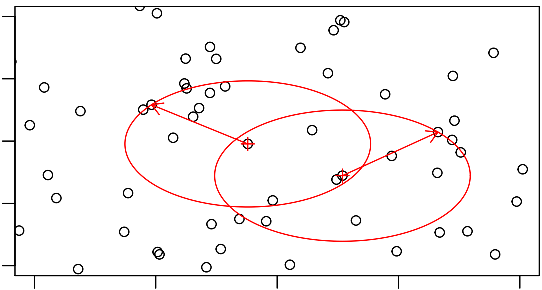
Variance of the weighted estimator

Let $V(f) = \int f \log^2 f - H(f)^2$. When $\zeta = \frac{2\alpha}{(\alpha+d)(1+\rho d)} > 1$ we choose

$$\tau_1 < \min\left(\frac{\beta}{d+\beta}, \frac{\zeta}{\zeta+2}, \frac{\zeta-1}{\zeta}\right).$$

Let $k_0^* = k_{0,n}^*$ and $k_1^* = k_{1,n}^*$ satisfy $k_0^* \leq k_1^*$, $k_0^* / \log^5 n \to \infty$ and $k_1^* = O(n^{\tau_1})$. Then

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \left| n \mathrm{Var}_f \hat{H}_n^w - V(f) \right| \to 0.$$



Efficiency of the weighted estimator

Recall the oracle estimator $H_n^* = -\frac{1}{n}\sum_{i=1}^n \log f(X_i)$.

Theorem

Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \rho, \nu) \in (0, \infty)^4$ with $\zeta > 1$ and $\beta > d/2$ and suppose that k_0^*, k_1^* satisfy the previous conditions and $k_1^* = o(n^{\tau_2})$, where

$$\tau_2 = \min\left(1 - \frac{d/4}{1 + \lfloor d/4 \rfloor}, 1 - \frac{d}{2\beta}\right).$$

Then

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} n\mathbb{E}_f\left\{(\hat{H}_n^w - H_n^*)^2\right\} \to 0,$$

Efficiency of the weighted estimator

Recall the oracle estimator $H_n^* = -\frac{1}{n}\sum_{i=1}^n \log f(X_i)$.

Theorem

Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \rho, \nu) \in (0, \infty)^4$ with $\zeta > 1$ and $\beta > d/2$ and suppose that k_0^*, k_1^* satisfy the previous conditions and $k_1^* = o(n^{\tau_2})$, where

$$\tau_2 = \min\left(1 - \frac{d/4}{1 + \lfloor d/4 \rfloor}, 1 - \frac{d}{2\beta}\right).$$

Then

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} n\mathbb{E}_f\{(\hat{H}_n^w - H_n^*)^2\} \to 0,$$

and in particular,

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \left| n\mathbb{E}_f\{(\hat{H}_n^w - H(f))^2\} - V(f) \right| \to 0.$$

Asymptotic Normality

Under the same conditions,

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} d_2\left(\mathcal{L}\left(\frac{n^{1/2}\{\hat{H}_n^w - H(f)\}}{V(f)^{1/2}}\right), N(0,1)\right) \to 0$$

as $n \to \infty$. Here d_2 is the 2nd Wasserstein distance.

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

Local asymptotic minimax lower bound

Fix $d \in \mathbb{N}$, $\theta = (\alpha, \beta, \rho, \nu) \in (0, \infty)^4$ and $f \in \mathcal{F}_{d,\theta}$. For $t \geq 0$ and a measurable $g : \mathbb{R}^d \to \mathbb{R}$, let

$$f_{t,g}(x) := \frac{2c(t)}{1 + e^{-2tg(x)}} f(x),$$

where $c(t)$ is a constant. For $\lambda \in \mathbb{R}$, let $g_\lambda := -\lambda\{\log f + H(f)\}$.

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

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

Moreover, if $t|\lambda| \leq 1 \wedge \{144 V(f)\}^{-1/2}$, then $f_{t,g_\lambda} \in \mathcal{F}_{d,\tilde{\theta}}$, where $\theta' = (\alpha, \beta, \rho, 4\nu)$.

Summary

- Kozachenko–Leonenko entropy 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 sufficient moments and smoothness.

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.

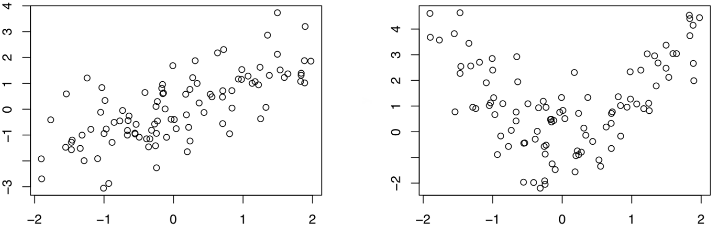
Independence testing

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.

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

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

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

When $X \perp\!\!\!\perp Y$ we have $V(X;Y) = 0$ so we don't have access to the null distribution, just that $\hat{I}_n = o_p(n^{-1/2})$.

Approximation to f_Y available

Suppose we have an approximation g_Y to f_Y that we can simulate from. For $B \in \mathbb{N}$ we may generate $\{Y_i^{(b)} : i = 1, \ldots, n, b = 1, \ldots, B\}$ and calculate

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

Approximation to f_Y available

Suppose we have an approximation g_Y to f_Y that we can simulate from. For $B \in \mathbb{N}$ we may 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

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

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)`.

Approximation to f_Y available

Suppose we have an approximation g_Y to f_Y that we can simulate from. For $B \in \mathbb{N}$ we may 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)`.

Since each $\hat{I}_n$ and $\hat{I}_n^{(b)}$ is shifted by $\hat{H}_n^X$ in the same way there is no need to calculate this quantity, and k_X and w_X do not need to be chosen.

Size of MINTknown

Lemma

For any $q \in (0,1)$ and $B \in \mathbb{N}$, the MINTknown(q) test satisfies

$$\sup_{k,k_Y \in \{1,\ldots,n-1\}} \sup_{(X,Y):I(X;Y)=0} \mathrm{pr}\big(\hat{I}_n > \hat{C}_q^{(n),B}\big) \leq q + d_{\mathrm{TV}}(f_Y^{\otimes n}, g_Y^{\otimes n}),$$

where the inner supremum is over all joint distributions of pairs (X, Y) with $I(X; Y) = 0$.

Size of `MINTknown`

Lemma

For any $q \in (0,1)$ and $B \in \mathbb{N}$, the `MINTknown`(q) *test satisfies*

$$\sup_{k,k_Y \in \{1,\ldots,n-1\}} \sup_{(X,Y):I(X;Y)=0} \mathrm{pr}\bigl(\hat{I}_n > \hat{C}_q^{(n),B}\bigr) \leq q + d_{\mathrm{TV}}(f_Y^{\otimes n}, g_Y^{\otimes n}),$$

where the inner supremum is over all joint distributions of pairs (X, Y) with $I(X;Y) = 0$.

Since $d_{\mathrm{TV}}^2(f_Y^{\otimes n}, g_Y^{\otimes n}) \leq 1 - \{1 - d_{\mathrm{H}}^2(f_Y, g_Y)\}^n$, if our approximation error is $o(n^{-1/2})$ then we have an approximately valid test.

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

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

and, for $b \geq 0$, let

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

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

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

and, for $b \geq 0$, let

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

Theorem

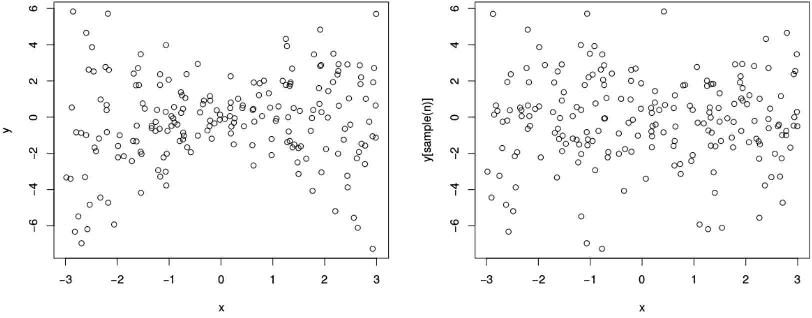
For suitable ϑ and choices of tuning parameters there exists a sequence (b_n) such that $b_n = o(n^{-1/2})$ and for each $q \in (0,1)$

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

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

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

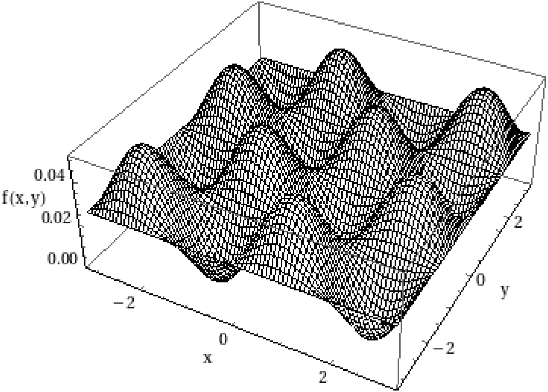


We refer to the resulting test as MINT(q).

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

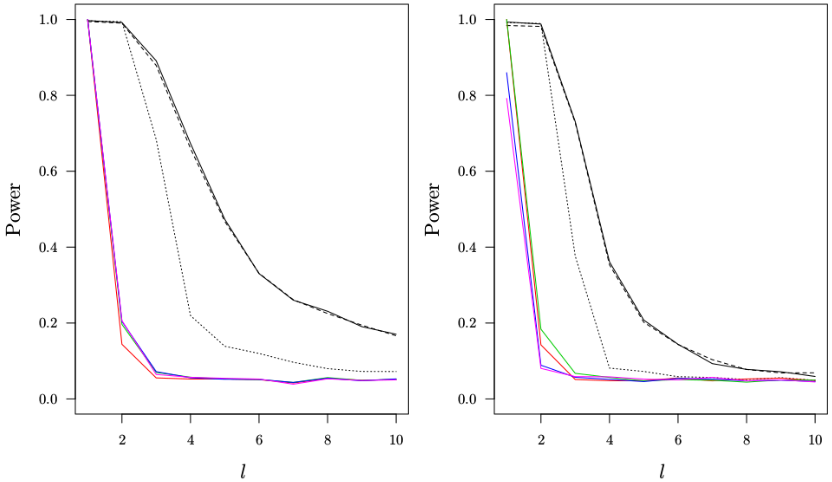
In the following we present power curves for `MINT` and `MINTknown` with oracle choices of k, k_Y, as well as power curves for `MINTav`, in which we average over $k \in \{1, \ldots, 20\}$ in `MINT`. In all cases we take $B = 100$.

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).

Regression setting

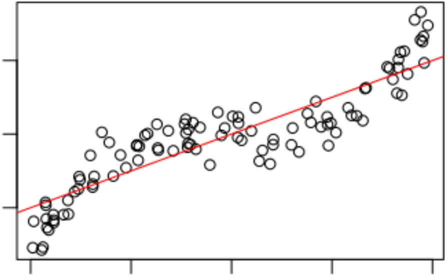
For any pair (X, Y) with $\mathbb{E}(Y^2) < \infty$ and $\mathbb{E}[XX^T]$ invertible we can define

$$\beta_0 := \mathrm{argmin}_{\beta \in \mathbb{R}^p} \{(Y - \beta^T X)^2\}$$

and

$$\epsilon = Y - \beta_0^T X.$$

If the model is correctly specified then we will have $X \perp\!\!\!\perp \epsilon$.



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.

Integral functional estimation

Two-sample estimation

Given samples $X_1, \ldots, X_m \sim f$ and $Y_1, \ldots, Y_n \sim g$ we may wish to estimate the general two-sample functional

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

Examples include Kullback–Leibler divergence, Rényi divergence, Hellinger distance etc.

It is natural to consider the estimator

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

Weights

As with entropy estimation, we can find suitable weight vectors $w \in \mathbb{R}^{k_X k_Y}$ such that

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

has bias $o(n^{-1/2})$.

Variance

Under regularity conditions this estimator achieves the local asymptotic minimax lower bound

$$\frac{1}{m}\mathrm{Var}(\phi_X + (f\phi_2)_X) + \frac{1}{n}\mathrm{Var}((f\phi_3)_Y),$$

where $\phi_2 = \frac{\partial\phi}{\partial f}$ and $\phi_3 = \frac{\partial\phi}{\partial g}$ and, e.g., $\phi_X = \phi(X, f(X), g(X))$.

Variance

Under regularity conditions this estimator achieves the local asymptotic minimax lower bound

$$\frac{1}{m}\mathrm{Var}(\phi_X + (f\phi_2)_X) + \frac{1}{n}\mathrm{Var}((f\phi_3)_Y),$$

where $\phi_2 = \frac{\partial \phi}{\partial f}$ and $\phi_3 = \frac{\partial \phi}{\partial g}$ and, e.g., $\phi_X = \phi(X, f(X), g(X))$.
For $\kappa \in (1/2, 1)$, consider

$$T_\kappa(f) := \int_{\mathcal{X}} f(x)^\kappa \, dx,$$

for which $\phi(f) = f^{\kappa-1}$ and

$$m\mathbb{E}\{(\hat{T}_m - T_\kappa(f))^2\} \to \mathrm{Var}(\kappa f(X)^{\kappa-1}).$$

Variance

Under regularity conditions this estimator achieves the local asymptotic minimax lower bound

$$\frac{1}{m}\mathrm{Var}(\phi_X + (f\phi_2)_X) + \frac{1}{n}\mathrm{Var}((f\phi_3)_Y),$$

where $\phi_2 = \frac{\partial \phi}{\partial f}$ and $\phi_3 = \frac{\partial \phi}{\partial g}$ and, e.g., $\phi_X = \phi(X, f(X), g(X))$.
For $\kappa \in (1/2, 1)$, consider

$$T_\kappa(f) := \int_{\mathcal{X}} f(x)^\kappa \, dx,$$

for which $\phi(f) = f^{\kappa-1}$ and

$$m\mathbb{E}\{(\hat{T}_m - T_\kappa(f))^2\} \to \mathrm{Var}(\kappa f(X)^{\kappa-1}).$$

Remarkably, this outperforms the natural oracle estimator

$$T_m^* = m^{-1}\sum_{i=1}^{m} f(X_i)^{\kappa-1}.$$

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. (2018) Efficient multivariate entropy estimation via k-nearest neighbour distances. *Ann. Statist., to appear*.

B. and Samworth, R. J. (2017) Nonparametric independence testing via mutual information. Available at `arXiv:1711.06642`.

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.

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.

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`.

Goria, M. N., Leonenko, N. N., Mergel, V. V. and Novi Inverardi, P. L. (2005) A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Stat.*, **17**, 277–297.

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.

References

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.

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`.

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

References

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

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.*

References

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`.

Polyanskiy, Y. and Wu, Y. (2016) Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Trans. Inf. Theory*, **62**, 3992–4002.

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.

References

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.

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.

References

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.

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`.