Biometrika (2018), **xx**, x, *pp*. 1–20
Printed in Great Britain

Nonparametric independence testing via mutual information

BY T. B. BERRETT AND R. J. SAMWORTH
Statistical Laboratory, University of Cambridge, CB3 0WB, U.K.
t.berrett@statslab.cam.ac.uk r.samworth@statslab.cam.ac.uk

SUMMARY

We propose a test of independence of two multivariate random vectors, given a sample from the underlying population. Our approach, which we call `MINT`, is based on the estimation of mutual information, whose decomposition into joint and marginal entropies facilitates the use of recently-developed efficient entropy estimators derived from nearest neighbour distances. The proposed critical values, which may be obtained by simulation in the case where an approximation to one marginal is available or by permuting the data otherwise, facilitate size guarantees, and we provide local power analyses, uniformly over classes of densities whose mutual information satisfies a lower bound. Our ideas may be extended to provide new goodness-of-fit tests of normal linear models based on assessing the independence of our vector of covariates and an appropriately-defined notion of an error vector. The theory is supported by numerical studies on both simulated and real data.

Some key words: Independence test; Mutual information; Nearest neighbours; Entropy estimation.

1. INTRODUCTION

Independence is a fundamental concept in statistics and many related fields, underpinning the way practitioners frequently think about model building, as well as much of statistical theory. Often we would like to assess whether or not the assumption of independence is reasonable, for instance as a method of exploratory data analysis (Steuer et al., 2002; Albert et al., 2015; Nguyen and Eisenstein, 2017), or as a way of evaluating the goodness-of-fit of a statistical model (Einmahl and van Keilegom, 2008). Testing independence and estimating dependence are well-established areas of statistics, with the related idea of the correlation between two random variables dating back to Francis Galton's work at the end of the 19th century (Stigler, 1989). This was subsequently expanded upon by Karl Pearson (e.g. Pearson, 1920). Since then many new measures of dependence have been developed and studied, each with its own advantages and disadvantages, and there is no universally accepted measure. For surveys of several measures, see, for example, Schweizer (1981), Joe (1989), Mari and Kotz (2001) and the references therein. We give an overview of more recently-introduced quantities below; see also Josse and Holmes (2016).

In addition to the applications mentioned above, dependence measures play an important role in independent component analysis, a special case of blind source separation, in which a linear transformation of the data is sought so that the transformed data is maximally independent; see e.g. Comon (1994), Bach and Jordan (2002), Miller and Fisher (2003) and Samworth and Yuan (2012). Here, independence tests may be carried out to check the convergence of an algorithm and to validate the results (e.g. Wu et al., 2009). Further examples include feature selection, where one seeks a set of features which contains the maximum possible information about a

response (Torkkola, 2003; Song et al., 2012), and the evaluation of the quality of a clustering in cluster analysis (Vinh et al., 2010).

When dealing with discrete data, often presented in a contingency table, the independence testing problem is typically reduced to testing the equality of two discrete distributions via a chi-squared test. Here we will focus on the case of distributions that are absolutely continuous with respect to the relevant Lebesgue measure. Classical nonparametric approaches to measuring dependence and independence testing in such cases include Pearson's correlation coefficient, Kendall's tau, and Spearman's rank correlation coefficient. Though these approaches are widely used, they suffer from a lack of power against many alternatives; indeed Pearson's correlation only measures linear relationships between variables, while Kendall's tau and Spearman's rank correlation coefficient measure monotonic relationships. Thus, for example, if X has a symmetric distribution on the real line, and $Y = X^2$, then the population quantities corresponding to these test statistics are zero in all three cases. Hoeffding's test of independence (Hoeffding, 1948) is able to detect a wider class of departures from independence and is distribution-free under the null hypothesis but, as with these other classical methods, was only designed for univariate variables. Recent work of Weihs et al. (2018) has aimed to address some of computational challenges involved in extending these ideas to multivariate settings.

Other recent research has focused on constructing tests that can be used for more complex data and that are consistent against wider classes of alternatives. The concept of distance covariance was introduced in Székely et al. (2007) and can be expressed as a weighted L_2 norm of the difference between the characteristic function of the joint distribution and the product of the marginal characteristic functions. This concept has also been studied in high dimensions (Székely and Rizzo, 2013; Yao et al., 2017), and for testing independence of several random vectors (Fan et al., 2017). In Sejdinovic et al. (2013) tests based on distance covariance were shown to be equivalent to a reproducing kernel Hilbert space test for a specific choice of kernel. Such tests have been widely studied in the machine learning community, with early understanding of the subject given by Bach and Jordan (2002) and Gretton et al. (2005), in which the Hilbert–Schmidt independence criterion was proposed. These tests are based on embedding the joint distribution and product of the marginal distributions into a Hilbert space and considering the norm of their difference in this space. One drawback of the kernel paradigm here is the computational complexity, though Jitkrittum et al. (2016) and Zhang et al. (2017) have recently attempted to address this issue. The performance of these methods may also be strongly affected by the choice of kernel. In another line of work, there is a large literature on testing independence based on an empirical copula process; see for example Kojadinovic and Holmes (2009) and the references therein. Other test statistics include those based on partitioning the sample space (e.g. Gretton and Györfi, 2010; Heller et al., 2016). These have the advantage of being distribution-free under the null hypothesis, though their performance depends on the particular partition chosen.

We also remark that the basic independence testing problem has spawned many variants. For instance, Pfister et al. (2017) extend kernel tests to tests of mutual independence between a group of random vectors. Another important extension is to testing conditional independence, which is central to graphical modelling (Lauritzen, 1996) and also relevant to causal inference (Pearl, 2009). Existing tests of conditional independence include the proposals of Su and White (2008), Zhang et al. (2011), Fan et al. (2017), Shah and Peters (2018) and Berrett et al. (2018c).

To formalize the problem, suppose that $d \in \mathbb{N}$ can be written as $d = d_X + d_Y$ for some $d_X, d_Y \in \mathbb{N}$, that X and Y are random vectors taking values in $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$ respectively, and that $Z = (X, Y)$ has density f with respect to Lebesgue measure on $\mathbb{R}^d$. We write f_X and f_Y for the marginal Lebesgue densities of X and Y. Given independent and identically distributed copies $Z_1, \ldots, Z_n$ of Z, we wish to test the null hypothesis H_0 that X and Y are independent

against the alternative that X and Y are not independent. Our approach is based on constructing an estimator $\hat{I}_n = \hat{I}_n(Z_1, \ldots, Z_n)$ of the mutual information $I(X;Y)$ between X and Y. Mutual information turns out to be a very attractive measure of dependence in this context; we review its definition and basic properties in Section 2 below. In fact, its decomposition into joint and marginal entropies (see (1) below) facilitates the use of recently-developed efficient entropy estimators derived from nearest neighbour distances (Berrett et al., 2018b), though we emphasize that our results on power in Sections 4 and 5 in particular require several new ideas in the analysis of nearest neighbour methods and permutation tests. We remark that nearest neighbour approaches are now known to enjoy many attractive properties for nonparametric statistical problems; see Biau and Devroye (2015) for several recent developments.

In the simpler setting where an approximation to either of the marginals f_X and f_Y is available, a simulation-based approach can be employed to yield a critical value for a test having at most its nominal size $q \in (0,1)$, up to an additional term that reflects the quality of the approximation to the marginal. This latter term vanishes when the approximation is exact. Our main result in this setting is to provide regularity conditions under which the power of our test converges to 1 as $n \to \infty$, uniformly over classes of alternatives with $I(X;Y) \geq b_n$, where we may even take $b_n = o(n^{-1/2})$. To the best of our knowledge this is the first time that such a local power analysis has been carried out for an independence test for multivariate data. When neither marginal is known, we obtain our critical value via a permutation approach, yielding a test of at most nominal size. Here, the test has power converging uniformly to 1 as $n \to \infty$ over the subset of densities in our classes whose mutual information is bounded below by a positive constant. Again, we believe that such uniform power analyses have not previously been provided for permutation tests of independence. We call our test `MINT`, short for Mutual Information Test; it is implemented in the `R` package `IndepTest` (Berrett et al., 2018a).

As an application of these ideas, we are able to introduce new goodness-of-fit tests of normal linear models based on assessing the independence of our vector of covariates and an appropriately-defined notion of an error vector. Such tests do not follow immediately from our earlier work as we do not observe realisations of the error vector directly; instead, we only have access to residuals from a fitted model. Nevertheless, we are able to provide rigorous justification, again in the form of a local analysis, for our approach. It seems that, when fitting normal linear models, current standard practice in the applied statistics community for assessing goodness-of-fit is based on visual inspection of diagnostic plots such as those provided by the `plot` command in `R` when applied to an object of class `lm`. Our aim, then, is to augment the standard toolkit by providing a formal basis for inference regarding the validity of the model. We mention that Sen and Sen (2014) propose an alternative test based on the Hilbert–Schmidt independence criterion, where a residual bootstrap approach is used to obtain the critical value. Under regularity conditions, they provide an asymptotic size guarantee, and prove that their test has asymptotic power 1 against any fixed alternative. Further related work here includes Neumeyer (2009), Neumeyer and Van Keilegom (2010), Müller et al. (2012) and Shah and Bühlmann (2017).

The following notation is used throughout. For $D \in \mathbb{N}$, let λ_D and $\|\cdot\|$ denote Lebesgue measure and the Euclidean norm on $\mathbb{R}^D$ respectively. If $f = d\mu/d\lambda_D$ and $g = d\nu/d\lambda_D$ are densities on $\mathbb{R}^D$ with respect to λ_D, we write $f \ll g$ if $\mu \ll \nu$. We also write $d_{\mathrm{TV}}(f,g) = 2^{-1}\int_{\mathbb{R}^D} |f-g|\, d\lambda_D$ and $d_{\mathrm{H}}(f,g) = \big\{\int_{\mathbb{R}^D}(f^{1/2} - g^{1/2})^2\, d\lambda_D\big\}^{1/2}$ for the total variation and Hellinger distances between f and g respectively, and $f^{\otimes n}$ for the n-fold product of f, given by $f^{\otimes n}(x) = \prod_{i=1}^n f(x_i)$ for $x = (x_1, \ldots, x_n) \in \mathbb{R}^D$. For $z \in \mathbb{R}^D$ and $r \in [0,\infty)$, we write $B_z(r) = \{w \in \mathbb{R}^D : \|w - z\| \leq r\}$ and $B_z^\circ(r) = B_z(r) \setminus \{z\}$. Write $\lambda_{\min}(A)$ for the smallest eigenvalue of a positive definite matrix A, and $\|B\|_{\mathrm{F}}$ for the Frobenius norm of a matrix B.

2. MUTUAL INFORMATION

2·1. *Definition and basic properties*

Retaining our notation from the introduction, let $\mathcal{Z} = \{(x, y) : f(x, y) > 0\}$. A very natural measure of dependence is the mutual information between X and Y, defined to be

$$I(X;Y) = I(f) = \int_{\mathcal{Z}} f(x,y) \log \frac{f(x,y)}{f_X(x) f_Y(y)} \, d\lambda_d(x,y).$$

This is the Kullback–Leibler divergence between the joint distribution of (X, Y) and the product of the marginal distributions, so is non-negative, and equal to zero if and only if X and Y are independent. Another attractive feature of mutual information as a measure of dependence is that it is invariant to invertible transformations of X and Y. In fact, a consequence of the data processing inequality for mutual information (e.g. Cover and Thomas, 2006, Theorem 2.8.1) is that $I\big(\phi(X); Y\big) = I(X; Y)$ whenever Y and X are conditionally independent given $\phi(X)$ (Kinney and Atwal, 2014). This means that mutual information is *self-equitable* in the terminology of Kinney and Atwal (2014) as well as *nonparametric* in the sense of Weihs et al. (2018), whereas several other measures of dependence, including distance covariance, reproducing kernel Hilbert space measures and correlation-based notions are not in general. Under a mild assumption, the mutual information between X and Y can be expressed in terms of their joint and marginal entropies; more precisely, writing $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = \{y : f_Y(y) > 0\}$, and provided that each of $H(X, Y)$, $H(X)$ and $H(Y)$ are finite,

$$\begin{aligned} I(X;Y) &= \int_{\mathcal{Z}} f \log f \, d\lambda_d - \int_{\mathcal{X}} f_X \log f_X \, d\mu_{d_X} - \int_{\mathcal{Y}} f_Y \log f_Y \, d\mu_{d_Y} \\ &= -H(X,Y) + H(X) + H(Y). \end{aligned} \tag{1}$$

Thus, mutual information estimators can be constructed from entropy estimators.

Moreover, the concept of mutual information is easily generalized to more complex situations. For instance, suppose now that (X, Y, U) has joint density f on $\mathbb{R}^{d+d_U}$, and let $f_{(X,Y)|U}(\cdot \mid u)$, $f_{X|U}(\cdot \mid u)$ and $f_{Y|U}(\cdot \mid u)$ denote the (joint) conditional densities of (X, Y), X, and Y given $U = u$ respectively. The conditional mutual information between X and Y given U is defined as

$$I(X;Y \mid U) = \int_{\mathcal{U}} f(x,y,u) \log \frac{f_{(X,Y)|U}(x,y \mid u)}{f_{X|U}(x \mid u) f_{Y|U}(y \mid u)} \, d\lambda_{d+d_U}(x,y,u),$$

where $\mathcal{U} = \{(x, y, u) : f(x, y, u) > 0\}$. This can similarly be written as

$$I(X;Y \mid U) = H(X,U) + H(Y,U) - H(X,Y,U) - H(U),$$

provided each of the summands is finite. Another extension is to situations with p random vectors. In particular, suppose that $X_1, \ldots, X_p$ have joint density f on $\mathbb{R}^d$, where $d = d_1 + \cdots + d_p$ and that X_j has marginal density f_j on $\mathbb{R}^{d_j}$. Then, writing $\mathcal{X}_p = \{(x_1, \ldots, x_p) \in \mathbb{R}^d : f(x_1, \ldots, x_p) > 0\}$, we can define

$$\begin{aligned} I(X_1; \ldots; X_p) &= \int_{\mathcal{X}_p} f(x_1, \ldots, x_p) \log \frac{f(x_1, \ldots, x_p)}{f_1(x_1) \cdots f_p(x_p)} \, d\lambda_d(x_1, \ldots, x_p) \\ &= \sum_{j=1}^{p} H(X_j) - H(X_1, \ldots, X_p), \end{aligned}$$

with the second equality holding provided that each of the entropies is finite. The tests we introduce in Sections 3 and 4 therefore extend in a straightforward manner to tests of independence of several random vectors.

2·2. *Estimation of mutual information*

For $i = 1, \dots, n$ with $n \geq 2$, let $Z_{(1),i}, \dots, Z_{(n-1),i}$ denote a permutation of $\{Z_1, \dots, Z_n\} \setminus \{Z_i\}$ such that $\|Z_{(1),i} - Z_i\| \leq \cdots \leq \|Z_{(n-1),i} - Z_i\|$. For conciseness, we let

$$\rho_{(k),i} = \|Z_{(k),i} - Z_i\|$$

denote the distance between Z_i and the kth nearest neighbour of Z_i. To estimate the unknown entropies, we will use a weighted version of the Kozachenko–Leonenko estimator (Kozachenko and Leonenko, 1987). For $k = k_Z \in \{1, \dots, n-1\}$ and weights $w_1^Z, \dots, w_k^Z$ satisfying $\sum_{j=1}^k w_j^Z = 1$, this is defined as

$$\hat{H}_n^Z = \hat{H}_{n,k}^{d,w^Z}(Z_1, \dots, Z_n) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k w_j^Z \log\left(\frac{\rho_{(j),i}^d V_d (n-1)}{e^{\Psi(j)}} \right),$$

where $V_d = \pi^{d/2}/\Gamma(1 + d/2)$ denotes the volume of the unit d-dimensional Euclidean ball and where Ψ denotes the digamma function. Berrett et al. (2018b) provided conditions on k, $w_1^Z, \dots, w_k^Z$ and the underlying data generating mechanism under which $\hat{H}_n^Z$ is an efficient estimator of $H(Z)$ (in the sense that its asymptotic normalized squared error risk achieves the local asymptotic minimax lower bound) in arbitrary dimensions. With estimators $\hat{H}_n^X$ and $\hat{H}_n^Y$ of $H(X)$ and $H(Y)$ defined analogously as functions of $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ respectively, we can use (1) to define an estimator of mutual information by

$$\hat{I}_n = \hat{I}_n(Z_1, \dots, Z_n) = \hat{H}_n^X + \hat{H}_n^Y - \hat{H}_n^Z. \tag{2}$$

Thus our mutual information estimator $\hat{I}_n$ depends on k, k_X, k_Y, w, w^X, w^Y, though we suppress this dependence for notational simplicity. Kraskov et al. (2004) have proposed an alternative, popular estimator of mutual information. For our purposes, however, (2) is more analytically tractable, as well as significantly quicker to compute. Having identified an appropriate mutual information estimator, we turn our attention in the next two sections to obtaining appropriate critical values for our independence tests.

3. THE CASE WHERE AN APPROXIMATION TO ONE MARGINAL IS AVAILABLE

In this section, we consider the case where an approximation to at least one of f_X and f_Y, constructed independently of the data $Z_1, \dots, Z_n$, is available and remark that in our experience, little is gained by having an approximation to the second marginal density. To give examples of practical situations where this may be realistic, observe that in the nonparametric regression model

$$Y = g(X) + \epsilon,$$

where X and ϵ are independent, testing the independence of X and Y is equivalent to testing the null hypothesis that $g = 0$. In particular, one can imagine for example medical studies where the question of interest is whether a health outcome, such as blood pressure, is associated with age. Here, age distributions can be well approximated from independent demographic information. In Section 5 below, we also consider the robustness of goodness-of-fit tests of regression models where an approximation, up to a scale factor, to the error distribution is available.

Without loss of generality, we assume that the marginal density f_Y is approximated by g_Y, and further assume that we can generate independent and identically distributed random variables, denoted $\{Y_i^{(b)} : i = 1, \ldots, n, b = 1, \ldots, B\}$, from the density g_Y, independently of $Z_1, \ldots, Z_n$. Our test in this setting, which we refer to as `MINTknown`, or `MINTknown`(q) when the nominal size $q \in (0,1)$ needs to be made explicit, will reject H_0 for large values of $\hat{I}_n$. The ideal critical value, if both marginal densities were known, would therefore be

$$C_q^{(n)} = \inf\{r \in \mathbb{R} : \mathrm{pr}_{f_X f_Y}(\hat{I}_n > r) \leq q\}.$$

Using our pseudo-data $\{Y_i^{(b)} : i = 1, \ldots, n, b = 1, \ldots, B\}$, generated as described above, we define the statistics

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

for $b = 1, \ldots, B$. Motivated by the thought that these statistics have approximately the same distribution as $\hat{I}_n$ under H_0, we can estimate the critical value $C_q^{(n)}$ by

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

the $(1-q)$th quantile of $\{\hat{I}_n, \hat{I}_n^{(1)}, \ldots, \hat{I}_n^{(B)}\}$. An interesting feature of `MINTknown`, which is apparent from the proof of Lemma 1 below, is that there is no need to calculate $\hat{H}_n^X$ in (1), either on the original data, or on the pseudo-data sets $\{(X_1, Y_1^{(b)}), \ldots, (X_n, Y_n^{(b)}) : b = 1, \ldots, B\}$. This is because in the decomposition of the event $\{\hat{I}_n^{(b)} \geq \hat{I}_n\}$ into entropy estimates, $\hat{H}_n^X$ appears on both sides of the inequality, so it cancels. This observation somewhat simplifies our assumptions and analysis, as well as reducing the number of tuning parameters that need to be chosen. The following lemma justifies this critical value estimate.

LEMMA 1. *For any $q \in (0,1)$ and $B \in \mathbb{N}$, the* `MINTknown`(q) *test that rejects H_0 if and only if $\hat{I}_n > \hat{C}_q^{(n),B}$ 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 $I(X;Y) = 0$.

In particular, we see from Lemma 1 that if our approximation to f_Y is exact, in the sense that $g_Y = f_Y$, then `MINTknown` has at most its nominal size. More generally, since $d_{\mathrm{TV}}^2(f_Y^{\otimes n}, g_Y^{\otimes n}) \leq 1 - \{1 - d_{\mathrm{H}}^2(f_Y, g_Y)\}^n$, we see that whenever the approximation error $d_{\mathrm{H}}(f_Y, g_Y)$ is small by comparison with $n^{-1/2}$, the test will have approximately its nominal size. As an example, suppose that f_Y and g_Y are the $N(0,1)$ and $N(0, \hat{\sigma}_m^2)$ densities respectively, where $\hat{\sigma}_m^2$ is the sample variance of an independent sample of size m, independent of $Z_1, \ldots, Z_n$, from f_Y. Then, if $m/n \to \infty$ as $n \to \infty$, we have that

$$E d_{\mathrm{TV}}(f_Y^{\otimes n}, g_Y^{\otimes n}) = \Big(\frac{n}{4\pi}\Big)^{1/2} E|\hat{\sigma}_m^2 - 1| + o(n^{1/2} m^{-1/2}) = \Big(\frac{n}{m\pi^2}\Big)^{1/2} \{1 + o(1)\},$$

so the size of `MINTknown` is controlled asymptotically.

The remainder of this section is devoted to a rigorous study of the power of `MINTknown` that is compatible with a sequence of local alternatives (f_n) having mutual information $I_n \to 0$. To this end, we first define the classes of alternatives that we consider. Let $\mathcal{F}_d$ denote the class of all

density functions with respect to Lebesgue measure on $\mathbb{R}^d$. For $f \in \mathcal{F}_d$ and $\alpha > 0$, let

$$\mu_\alpha(f) = \int_{\mathbb{R}^d} \|z\|^\alpha f(z)\, dz.$$

Now let $\mathcal{A}$ denote the class of decreasing functions $a : (0,\infty) \to [1,\infty)$ satisfying $a(\delta) = o(\delta^{-\epsilon})$ as $\delta \searrow 0$, for every $\epsilon > 0$. If $a \in \mathcal{A}$, $\beta > 0$, $f \in \mathcal{F}_d$ is $m = (\lceil \beta \rceil - 1)$-times differentiable and $z \in \mathcal{Z}$, we define $r_a(z) = \{8d^{1/2} a(f(z))\}^{-1/(\beta \wedge 1)}$ and

$$M_{f,a,\beta}(z) = \max\left\{ \max_{t=1,\ldots,m} \frac{\|f^{(t)}(z)\|}{f(z)}\,,\ \sup_{w \in B_z^\circ(r_a(z))} \frac{\|f^{(m)}(w) - f^{(m)}(z)\|}{f(z)\|w-z\|^{\beta-m}} \right\}.$$

The quantity $M_{f,a,\beta}(z)$ measures the smoothness of derivatives of f in neighbourhoods of z, relative to $f(z)$ itself, and these neighbourhoods of z are allowed to become smaller when $f(z)$ is small. Finally, for $\Theta = (0,\infty)^4 \times \mathcal{A}$, and $\theta = (\alpha, \beta, \nu, \gamma, a) \in \Theta$, let

$$\mathcal{F}_{d,\theta} = \left\{ f \in \mathcal{F}_d : \mu_\alpha(f) \le \nu, \|f\|_\infty \le \gamma, \sup_{z : f(z) \ge \delta} M_{f,a,\beta}(z) \le a(\delta)\ \forall \delta > 0 \right\}.$$

Berrett et al. (2018b) show that all Gaussian and multivariate-t densities, amongst others, belong to $\mathcal{F}_{d,\theta}$ for appropriate $\theta \in \Theta$. However, the classes do rule out severe oscillations in the tails: for instance, the density $f(x) = \{1 - \cos(x^2)\}\mathbb{1}_{\{x \neq 0\}} / \{(2\pi)^{1/2} x^2\}$ does not belong to $\mathcal{F}_{1,\theta}$ for any $\theta \in \Theta$.

Now, for $d_X, d_Y \in \mathbb{N}$ and $\vartheta = (\theta, \theta_Y) \in \Theta^2$, 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 $b \ge 0$, let

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

Thus, $\mathcal{F}_{d_X,d_Y,\vartheta}(b)$ consists of pairs (f, g_Y) where the mutual information of f is greater than b. In Theorem 1 below, we will show that for a suitable choice of $b = b_n$ and for certain $\vartheta \in \Theta^2$, the power of the test defined in Lemma 1 converges to 1, uniformly over $\mathcal{F}_{d_X,d_Y,\vartheta}(b)$.

Before we can state this result, however, we must define the allowable choices of k, k_Y and the weight vectors. Given $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ let

$$\tau_1(d,\theta) = \min\left\{ \frac{2\alpha}{5\alpha + 3d}\,,\ \frac{\alpha - d}{2\alpha}\,,\ \frac{4(\beta \wedge 1)}{4(\beta \wedge 1) + 3d} \right\}$$

and

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

We have that $\min_{i=1,2} \tau_i(d,\theta) > 0$ if and only if both $\alpha > d$ and $\beta > d/2$. Finally, for $k \in \mathbb{N}$, let

$$\mathcal{W}^{(k)} = \left\{ w = (w_1, \ldots, w_k) \in \mathbb{R}^k : \sum_{j=1}^k w_j \frac{\Gamma(j + 2\ell/d)}{\Gamma(j)} = 0 \quad \text{for } \ell = 1, \ldots, \lfloor d/4 \rfloor \right.$$
$$\left. \sum_{j=1}^k w_j = 1 \,\text{and}\, w_j = 0 \,\text{if}\, j \notin \{\lfloor k/d \rfloor, \lfloor 2k/d \rfloor, \ldots, k\} \right\}.$$

Thus, our weights sum to 1; the other constraints ensure that the dominant contributions to the bias of the unweighted Kozachenko–Leonenko estimator cancel out to sufficiently high order, and that the corresponding jth nearest neighbour distances are not too highly correlated. In practice, we recommend choosing $w \in \mathcal{W}^{(k)}$ to minimize $\|w\|$; this can be obtained in closed form using Lagrangian methods.

THEOREM 1. *Fix $d_X, d_Y \in \mathbb{N}$, set $d = d_X + d_Y$ and fix $\vartheta = (\theta, \theta_Y) \in \Theta^2$ with*

$$\min\Big\{\tau_1(d,\theta)\,,\,\tau_1(d_Y,\theta_Y)\,,\,\tau_2(d,\theta)\,,\,\tau_2(d_Y,\theta_Y)\Big\} > 0.$$

Let $k_0^* = k_{0,n}^*, k_Y^* = k_{Y,n}^*$ and $k^* = k_n^*$ denote any deterministic sequences of positive integers with $k_0^* \le \min\{k_Y^*, k^*\}$, with $k_0^*/\log^5 n \to \infty$ and with

$$\max\left\{\frac{k^*}{n^{\tau_1(d,\theta)-\epsilon}}\,,\,\frac{k_Y^*}{n^{\tau_1(d_Y,\theta_Y)-\epsilon}}\,,\,\frac{k^*}{n^{\tau_2(d,\theta)}}\,,\,\frac{k_Y^*}{n^{\tau_2(d_Y,\theta_Y)}}\right\} \to 0$$

for some $\epsilon > 0$. Also suppose that $w_Y = w_Y^{(k_Y)} \in \mathcal{W}^{(k_Y)}$ and $w = w^{(k)} \in \mathcal{W}^{(k)}$, and that $\limsup_n \max(\|w\|, \|w_Y\|) < \infty$. Then there exists a sequence (b_n) such that $b_n = o(n^{-1/2})$ and with the property that for each $q \in (0,1)$ and any sequence (B_n^*) with $B_n^* \to \infty$,

$$\inf_{B_n \ge B_n^*} \inf_{\substack{k_Y \in \{k_0^*,\dots,k_Y^*\} \\ k \in \{k_0^*,\dots,k^*\}}} \inf_{(f,g_Y) \in \mathcal{F}_{d_X,d_Y,\vartheta}(b_n)} \mathrm{pr}_{f,g_Y}(\hat{I}_n > \hat{C}_q^{(n),B_n}) \to 1.$$

Theorem 1 provides a strong guarantee on the ability of `MINTknown` to reject H_0, uniformly over classes whose mutual information is at least b_n, where we may even have $b_n = o(n^{-1/2})$. An interesting feature of this result is that we make no assumptions about how well g_Y approximates f_Y, because we are able to show that $\hat{I}_n^{(b)} = o_p(n^{-1/2})$ for $b = 1, \dots, B$. One choice of k_Y and k that satisfies the conditions of Theorem 1 without knowledge of the parameter $\vartheta \in \Theta^2$ is to set $k_Y = k = \min(\log^6 n, n-1)$.

4. THE CASE OF UNKNOWN MARGINAL DISTRIBUTIONS

We now consider the setting in which the marginal distributions of both X and Y are unknown. Our test statistic remains the same, but now we estimate the critical value by permuting our sample in an attempt to mimic the behaviour of the test statistic under H_0. More explicitly, for some $B \in \mathbb{N}$, we propose independently of $(X_1, Y_1), \dots, (X_n, Y_n)$ to simulate independent random variables $\tau_1, \dots, \tau_B$ uniformly from S_n, the permutation group of $\{1, \dots, n\}$, and for $b = 1, \dots, B$, set $Z_i^{(b)} = (X_i, Y_{\tau_b(i)})$ and $\tilde{I}_n^{(b)} = \hat{I}_n(Z_1^{(b)}, \dots, Z_n^{(b)})$. For $q \in (0,1)$, we can now estimate $C_q^{(n)}$ by

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

and refer to the test that rejects H_0 if and only if $\hat{I}_n > \tilde{C}_q^{(n),B}$ as `MINTunknown`(q). Now $\hat{I}_n > \tilde{C}_q^{(n),B}$ if and only if

$$1 + \sum_{b=1}^{B} \mathbb{1}_{\{\tilde{I}_n^{(b)} \ge \hat{I}_n\}} \le (B+1)q. \tag{3}$$

This shows that estimating either of the marginal entropies is unnecessary to carry out the test, since $\tilde{I}_n^{(b)} - \hat{I}_n = \hat{H}_n^Z - \tilde{H}_n^{(b)}$, where $\tilde{H}_n^{(b)} = \hat{H}_{n,k}^{d,w^Z}(Z_1^{(b)}, \ldots, Z_n^{(b)})$ is the weighted Kozachenko–Leonenko joint entropy estimator based on the permuted data.

LEMMA 2. *For any $q \in (0,1)$ and $B \in \mathbb{N}$, the* `MINTunknown`(q) *test has size at most q:*

$$\sup_{k \in \{1,\ldots,n-1\}} \sup_{(X,Y): I(X;Y)=0} \mathrm{pr}\big(\hat{I}_n > \tilde{C}_q^{(n),B}\big) \leq q.$$

We now study the power of `MINTunknown`, and begin by introducing the classes of marginal densities that we consider. To define an appropriate notion of smoothness, for $z \in \{v : g(v) > 0\} = \mathcal{V}$, $g \in \mathcal{F}_d$ and $\delta > 0$, let

$$r_{z,g,\delta} = \left\{ \frac{\delta e^{\Psi(k)}}{V_d (n-1) g(z)} \right\}^{1/d}. \tag{4}$$

Now, for A belonging to the class of Borel subsets of $\mathcal{V}$, denoted $\mathcal{B}(\mathcal{V})$, define

$$M_g(A) = \sup_{\delta \in (0,2]} \sup_{z \in A} \left| \frac{1}{V_d r_{z,g,\delta}^d g(z)} \int_{B_z(r_{z,g,\delta})} g \, d\lambda_d - 1 \right|.$$

Both $r_{z,g,\delta}$ and $M_g(\cdot)$ depend on n and k, but for simplicity we suppress this in our notation. Let $\phi = (\alpha, \mu, \nu, (c_n), (p_n)) \in (0,\infty)^3 \times [0,\infty)^{\mathbb{N}} \times [0,\infty)^{\mathbb{N}} = \Phi$ and define

$$\mathcal{G}_{d_X,d_Y,\phi} = \Big\{ f \in \mathcal{F}_{d_X+d_Y} : \max\{\mu_\alpha(f_X), \mu_\alpha(f_Y)\} \leq \mu, \max\{\|f_X\|_\infty, \|f_Y\|_\infty\} \leq \nu,$$
$$\exists \mathcal{V}_n \in \mathcal{B}(\mathcal{X} \times \mathcal{Y}) \text{ s.t. } M_{f_X f_Y}(\mathcal{V}_n) \leq c_n, \int_{\mathcal{V}_n^c} f_X f_Y \, d\lambda_d \leq p_n \; \forall n \in \mathbb{N} \Big\}.$$

In addition to controlling the αth moment and uniform norms of the marginals f_X and f_Y, the class $\mathcal{G}_{d,\phi}$ asks for there to be a (large) set $\mathcal{V}_n$ on which this product of marginal densities is uniformly well approximated by a constant over small balls. This latter condition is satisfied by products of many standard parametric families of marginal densities, including normal, Weibull, Gumbel, logistic, gamma, beta, and t densities, and is what ensures that nearest neighbour methods are effective in this context.

The corresponding class of joint densities we consider, for $\phi = (\alpha, \mu, \nu, (c_n), (p_n)) \in \Phi$, is

$$\mathcal{H}_{d,\phi} = \Big\{ f \in \mathcal{F}_d : \mu_\alpha(f) \leq \mu, \|f\|_\infty \leq \nu,$$
$$\exists \mathcal{Z}_n \in \mathcal{B}(\mathcal{Z}) \text{ s.t. } M_f(\mathcal{Z}_n) \leq c_n, \int_{\mathcal{Z}_n^c} f \, d\lambda_d \leq p_n \; \forall n \in \mathbb{N} \Big\}.$$

In many cases, we may take $\mathcal{Z}_n = \{z : f(z) \geq \delta_n\}$, for some appropriately chosen sequence (δ_n) with $\delta_n \to 0$ as $n \to \infty$. For instance, suppose we fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \nu, \gamma, a) \in \Theta$. Then, by Berrett et al. (2018b, Lemma 12), there exists such a sequence (δ_n), as well as sequences (c_n) and (p_n), where

$$\delta_n = \frac{ka(k/(n-1))^{\frac{d}{\beta \wedge 1}}}{n-1} \log(n-1), \quad c_n = \frac{15}{7} \frac{2^{\frac{\beta \wedge 1}{d}} d^{3/2}}{d + (\beta \wedge 1)} \log^{-\frac{\beta \wedge 1}{d}}(n-1),$$

for large n and $p_n = o((k/n)^{\alpha/(\alpha+d)-\epsilon})$ for every $\epsilon > 0$, such that $\mathcal{F}_{d,\theta} \subseteq \mathcal{H}_{d,\phi}$ with $\phi = (\alpha, \mu, \nu, (c_n), (p_n)) \in \Phi$. We may now state our main result on the power of `MINTunknown`.

THEOREM 2. *Let $d_X, d_Y \in \mathbb{N}$, let $d = d_X + d_Y$ and fix $\phi = (\alpha, \mu, \nu, (c_n), (p_n)) \in \Phi$ with $c_n \to 0$ and $p_n = o(1/\log n)$ as $n \to \infty$. Let $k_0^* = k_{0,n}^*$ and $k_1^* = k_{1,n}^*$ denote two deterministic sequences of positive integers satisfying $k_0^* \leq k_1^*$, $k_0^*/\log^2 n \to \infty$ and $(k_1^* \log^2 n)/n \to 0$. Then for any $b > 0$, $q \in (0,1)$ and any sequence (B_n^*) with $B_n^* \to \infty$ as $n \to \infty$,*

$$\inf_{B_n \geq B_n^*} \inf_{k \in \{k_0^*, \ldots, k_1^*\}} \inf_{\substack{f \in \mathcal{G}_{d_X, d_Y, \phi} \cap \mathcal{H}_{d,\phi}: \\ I(f) \geq b}} \mathrm{pr}_f(\hat{I}_n > \tilde{C}_q^{(n), B_n}) \to 1$$

as $n \to \infty$.

Theorem 2 shows that `MINTunknown` is uniformly consistent against a wide class of alternatives.

5. REGRESSION SETTING

In this section we aim to extend the ideas developed above to the problem of goodness-of-fit testing in linear models. Suppose we have independent and identically distributed pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ taking values in $\mathbb{R}^p \times \mathbb{R}$, with $E(Y_1^2) < \infty$ and $E(X_1 X_1^T)$ finite and positive definite. Then

$$\beta_0 = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} E\{(Y_1 - X_1^T \beta)^2\}$$

is well-defined, and we can further define $\epsilon_i = Y_i - X_i^T \beta_0$ for $i = 1, \ldots, n$. We show in the proof of Theorem 3 in the supplement that $E(\epsilon_1 X_1) = 0$, but for the purposes of interpretability and inference, it is often convenient if the random design linear model

$$Y_i = X_i^T \beta_0 + \epsilon_i, \quad i = 1, \ldots, n,$$

holds with X_i and ϵ_i independent. A goodness-of-fit test of this property amounts to a test of the independence of X_1 and ϵ_1. The main difficulty here is that $\epsilon_1, \ldots, \epsilon_n$ are not observed directly. Given an estimator $\hat{\beta}$ of β_0, the standard approach for dealing with this problem is to compute residuals $\hat{\epsilon}_i = Y_i - X_i^T \hat{\beta}$ for $i = 1, \ldots, n$, and use these as a proxy for $\epsilon_1, \ldots, \epsilon_n$. Many introductory statistics textbooks, e.g. Dobson (2002, Section 2.3.4), Dalgaard (2002, Section 5.2) suggest examining for patterns plots of residuals against fitted values, as well as plots of residuals against each covariate in turn, as a diagnostic, though it is difficult to formalize this procedure. It is also interesting that when applying the `plot` function in `R` to an object of type `lm`, these latter plots of residuals against each covariate in turn are not produced, presumably because it may be prohibitively time-consuming to check them all in the case of many covariates; they are, however, available in the package `car`.

The naive approach based on our work so far is simply to use the permutation test of Section 4 on the data $(X_1, \hat{\epsilon}_1), \ldots, (X_n, \hat{\epsilon}_n)$. Unfortunately, calculating the test statistic $\hat{I}_n$ on permuted data sets does not result in an exchangeable sequence, which makes it difficult to ensure that this test has the nominal size q. To circumvent this issue, we assume that the marginal distribution of ϵ_1 under H_0 has $E(\epsilon_1) = 0$, that $\sigma^2 = E(\epsilon_1^2)$ is finite, and that the density f_η of $\eta_1 = \epsilon_1/\sigma$ can be approximated by g_η, say, where $\int_{-\infty}^{\infty} x g_\eta(x)\, dx = 0$ and $\int_{-\infty}^{\infty} x^2 g_\eta(x)\, dx = 1$. In practice, it will often be the case that we take g_η to be the $N(0,1)$ density. We also assume that we can sample from g_η; of course, this is straightforward in the normal distribution case above. Let $X = (X_1 \cdots X_n)^T$, $Y = (Y_1, \ldots, Y_n)$, and suppose the vector of residuals $\hat{\epsilon} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^T$ is computed from the least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$. We then define standardized residuals by $\hat{\eta}_i = \hat{\epsilon}_i/\hat{\sigma}$, for $i = 1, \ldots, n$, where $\hat{\sigma}^2 = n^{-1}\|\hat{\epsilon}\|^2$; these standardized residuals are

invariant under changes of scale of $\epsilon = (\epsilon_1, \dots, \epsilon_n)$. Suppressing the dependence of our entropy estimators on k and the weights $w_1, \dots, w_k$ for notational simplicity, our test statistic is now given by

$$\check{I}_n = \hat{H}_n^p(X_1, \dots, X_n) + \hat{H}_n^1(\hat{\eta}_1, \dots, \hat{\eta}_n) - \hat{H}_n^{p+1}\big((X_1, \hat{\eta}_1), \dots, (X_n, \hat{\eta}_n)\big).$$

Writing $\eta = \epsilon/\sigma$, we have

$$\hat{\eta}_i = \frac{1}{\hat{\sigma}}(Y_i - X_i^T\hat{\beta}) = \frac{1}{\hat{\sigma}}\big\{\epsilon_i - X_i^T(\hat{\beta} - \beta_0)\big\} = \frac{n^{1/2}\big\{\eta_i - X_i^T(X^TX)^{-1}X^T\eta\big\}}{\|\eta - X^T(X^TX)^{-1}X^T\eta\|},$$

whose distribution does not depend on the unknown β_0 or σ^2. Let $\{\eta^{(b)} = (\eta_1^{(b)}, \dots, \eta_n^{(b)}) : b = 1, \dots, B\}$ denote independent random vectors, whose components are generated independently from g_η. For $b = 1, \dots, B$ we then set $\hat{s}^{(b)} = n^{-1/2}\|(I - X(X^TX)^{-1}X^T)\eta^{(b)}\|$ and, for $i = 1, \dots, n$, let

$$\hat{\eta}_i^{(b)} = \frac{1}{\hat{s}^{(b)}}\big\{\eta_i^{(b)} - X_i^T(X^TX)^{-1}X^T\eta^{(b)}\big\}.$$

We finally compute

$$\check{I}_n^{(b)} = \hat{H}_n^p(X_1, \dots, X_n) + \hat{H}_n^1(\hat{\eta}_1^{(b)}, \dots, \hat{\eta}_n^{(b)}) - \hat{H}_n^{p+1}\big((X_1, \hat{\eta}_1^{(b)}), \dots, (X_n, \hat{\eta}_n^{(b)})\big), \quad (5)$$

where the second and third terms in (5) are weighted Kozachenko–Leonenko estimates with tuning parameters k_η and k respectively. Analogously to our development in Sections 3 and 4, we can then define a critical value by

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

The following lemma controls the size of the resulting test.

LEMMA 3. *For each $q \in (0,1)$ and $B \in \mathbb{N}$, the* `MINTregression`*(q) test that rejects H_0 if and only if $\check{I}_n > \check{C}_q^{(n),B}$ satisfies*

$$\sup_{k,k_\eta \in \{1,\dots,n-1\}} \sup_{(X,Y): I(X;Y)=0} \mathrm{pr}\big(\check{I}_n > \check{C}_q^{(n),B}\big) \leq q + d_{\mathrm{TV}}(f_\eta^{\otimes n}, g_\eta^{\otimes n}).$$

As in previous sections, we are only interested in the differences $\check{I}_n - \check{I}_n^{(b)}$ for $b = 1, \dots, B$, and in these differences, the $\hat{H}_n^p(X_1, \dots, X_n)$ terms cancel out, so these marginal entropy estimators need not be computed.

In fact, to simplify our power analysis, it is more convenient to define a slightly modified test, which also has at most the nominal size. Specifically, we assume for simplicity that $m = n/2$ is an integer, and consider a test in which the sample is split in half, with the second half of the sample used to calculate the estimators $\hat{\beta}_{(2)}$ and $\hat{\sigma}^2_{(2)}$ of β_0 and σ^2 respectively. On the first half of the sample, we calculate

$$\hat{\eta}_{i,(1)} = \frac{Y_i - X_i^T\hat{\beta}_{(2)}}{\hat{\sigma}_{(2)}}$$

for $i = 1, \dots, m$ and the test statistic

$$\breve{I}_n = \hat{H}_m^p(X_1, \dots, X_m) + \hat{H}_m^1(\hat{\eta}_{1,(1)}, \dots, \hat{\eta}_{m,(1)}) - \hat{H}_m^{p+1}\big((X_1, \hat{\eta}_{1,(1)}), \dots, (X_m, \hat{\eta}_{m,(1)})\big).$$

Corresponding estimators $\{\breve{I}_n^{(b)} : b = 1, \ldots, B\}$ based on the simulated data may also be computed using the same sample-splitting procedure, and we then obtain the critical value $\breve{C}_q^{(n),B}$ in the same way as above. The advantage from a theoretical perspective of this approach is that, conditional on $\hat{\beta}_{(2)}$ and $\hat{\sigma}^2_{(2)}$, the random variables $\hat{\eta}_{1,(1)}, \ldots, \hat{\eta}_{m,(1)}$ are independent and identically distributed.

To describe the power properties of `MINTregression`, we first define several densities: for $\gamma \in \mathbb{R}^p$ and $s > 0$, let $f^{\gamma,s}_{\hat{\eta}}$ and $f^{\gamma,s}_{\hat{\eta}^{(1)}}$ denote the densities of $\hat{\eta}_1^{\gamma,s} = (\eta_1 - X_1^T\gamma)/s$ and $\hat{\eta}_1^{(1),\gamma,s} = (\eta_1^{(1)} - X_1^T\gamma)/s$ respectively; further, let $f^{\gamma,s}_{X,\hat{\eta}}$ and $f^{\gamma,s}_{X,\hat{\eta}^{(1)}}$ be the densities of $(X_1, \hat{\eta}_1^{\gamma,s})$ and $(X_1, \hat{\eta}_1^{(1),\gamma,s})$ respectively. Imposing assumptions on these densities amounts to imposing assumptions on the joint density f of (X, ϵ), and on the approximating density g_η for f_η. For $\Omega = \Theta^2 \times (0,\infty) \times (0,1) \times (0,\infty)^2$, and $\omega = (\theta_1, \theta_2, r_0, s_0, \Lambda, \lambda_0)$, we therefore let $\mathcal{F}^*_{p+1,\omega}$ denote the class of pairs of densities (f, g_η) satisfying the following three properties: first we ask that

$$\Big\{f^{\gamma,s}_{\hat{\eta}} : \gamma \in B_0(r_0), s \in [s_0, 1/s_0]\Big\} \cup \Big\{f^{\gamma,s}_{\hat{\eta}^{(1)}} : \gamma \in B_0(r_0), s \in [s_0, 1/s_0]\Big\} \subseteq \mathcal{F}_{1,\theta_1}$$

and

$$\Big\{f^{\gamma,s}_{X,\hat{\eta}} : \gamma \in B_0(r_0), s \in [s_0, 1/s_0]\Big\} \cup \Big\{f^{\gamma,s}_{X,\hat{\eta}^{(1)}} : \gamma \in B_0(r_0), s \in [s_0, 1/s_0]\Big\} \subseteq \mathcal{F}_{p+1,\theta_2}.$$

Next, we require the following moment bounds:

$$\sup_{\gamma \in B_0(r_0)} \max\Big\{E \log^2 f^{\gamma,1}_{\hat{\eta}}(\eta_1)\,,\, E \log^2 f^{\gamma,1}_{\hat{\eta}^{(1)}}(\eta_1)\Big\} \leq \Lambda, \tag{6}$$

and

$$\sup_{\gamma \in B_0(r_0)} \max\Big\{E \log^2 f_\eta\big(\hat{\eta}_1^{\gamma,1}\big)\,,\, E \log^2 f_\eta\big(\hat{\eta}_1^{(1),\gamma,1}\big)\Big\} \leq \Lambda. \tag{7}$$

Finally, writing $\Sigma = E(X_1 X_1^T)$, we ask that $\lambda_{\min}(\Sigma) \geq \lambda_0$.

The first of these requirements ensures that we can estimate efficiently the marginal entropy of our scaled residuals, as well as the joint entropy of these scaled residuals and our covariates. The second condition is a moment condition that allows us to control $|H(\eta_1 - X_1^T\gamma) - H(\eta_1)|$, and similar quantities, in terms of $\|\gamma\|$, when γ belongs to a small ball around the origin. To illustrate the second part of this condition, it is satisfied, for instance, if f_η is a standard normal density and $E(\|X_1\|^4) < \infty$, or if f_η is a t density and $E(\|X_1\|^\alpha) < \infty$ for some $\alpha > 0$; the first part of the condition is a little more complicated but similar. The final condition is very natural for random design regression problems.

By the same observation on the sequence $(\breve{I}_n, \breve{I}_n^{(1)}, \ldots, \breve{I}_n^{(B)})$ as was made regarding the sequence $(\check{I}_n, \check{I}_n^{(1)}, \ldots, \check{I}_n^{(B)})$ just before Lemma 3, we see that the sample-splitting version of the `MINTregression`(q) test has size at most q.

THEOREM 3. *Fix $p \in \mathbb{N}$ and $\omega = (\theta_1, \theta_2, r_0, s_0, \Lambda, \lambda_0) \in \Omega$, where the first component of θ_2 is $\alpha_2 \geq 4$ and the second component of θ_1 is $\beta_1 \geq 1$. Assume that*

$$\min\Big\{\tau_1(1,\theta_1)\,,\, \tau_1(p+1,\theta_2)\,,\, \tau_2(1,\theta_1)\,,\, \tau_2(p+1,\theta_2)\Big\} > 0.$$

Let $k_0^* = k_{0,n}^*, k_\eta^* = k_{\eta,n}^*$ and $k^* = k_n^*$ denote any deterministic sequences of positive integers with $k_0^* \leq \min\{k_\eta^*, k^*\}$, with $k_0^*/\log^5 n \to \infty$ and with

$$\max\left\{ \frac{k^*}{n^{\tau_1(p+1,\theta_2)-\epsilon}} \, , \, \frac{k_\eta^*}{n^{\tau_1(1,\theta_1)-\epsilon}} \, , \, \frac{k^*}{n^{\tau_2(p+1,\theta_2)}} \, , \, \frac{k_\eta^*}{n^{\tau_2(1,\theta_1)}} \right\} \to 0$$

for some $\epsilon > 0$. Also suppose that $w_\eta = w_\eta^{(k_\eta)} \in \mathcal{W}^{(k_\eta)}$ and $w = w^{(k)} \in \mathcal{W}^{(k)}$, and that $\limsup_n \max(\|w\|, \|w_\eta\|) < \infty$. Then for any sequence (b_n) such that $n^{1/2} b_n \to \infty$, any $q \in (0,1)$ and any sequence (B_n^*) with $B_n^* \to \infty$,

$$\inf_{B_n \geq B_n^*} \inf_{\substack{k_\eta \in \{k_0^*, \ldots, k_\eta^*\} \\ k \in \{k_0^*, \ldots, k^*\}}} \inf_{(f,g_\eta) \in \mathcal{F}_{1,p+1}^* : I(f) \geq b_n} \mathrm{pr}_{f,g_\eta}(\check{I}_n > \check{C}_q^{(n),B_n}) \to 1.$$

Finally in this section, we consider partitioning our design matrix as $X = (X^* \, X^{**}) \in \mathbb{R}^{n \times (p_0+p_1)}$, with $p_0 + p_1 = p$, and describe an extension of `MINTregression` to cases where we are interested in testing the independence between ϵ and X^*. For instance, X^{**} may consist of an intercept term, or transformations of variables in X^*, as in the real data example presented in Section 6·3 below. Our method for simulating standardized residual vectors $\{\hat{\eta}^{(b)} : b = 1, \ldots, B\}$ remains unchanged, but our test statistic and corresponding null statistics become

$$\begin{aligned} \bar{I}_n &= \hat{H}_n^p(X_1^*, \ldots, X_n^*) + \hat{H}_n^1(\hat{\eta}_1, \ldots, \hat{\eta}_n) - \hat{H}_n^{p+1}\big((X_1^*, \hat{\eta}_1), \ldots, (X_n^*, \hat{\eta}_n)\big) \\ \bar{I}_n^{(b)} &= \hat{H}_n^p(X_1^*, \ldots, X_n^*) + \hat{H}_n^1(\hat{\eta}_1^{(b)}, \ldots, \hat{\eta}_n^{(b)}) - \hat{H}_n^{p+1}\big((X_1^*, \hat{\eta}_1^{(b)}), \ldots, (X_n^*, \hat{\eta}_n^{(b)})\big), \end{aligned}$$

for $b = 1, \ldots, B$. Similar arguments to those employed in Lemma 3 show that the test that rejects the null hypothesis of independence between ϵ and X^* when

$$1 + \sum_{b=1}^{B} \mathbb{1}_{\{\bar{I}_n^{(b)} \geq \bar{I}_n\}} \leq (B+1)q$$

has size bounded above by $q + d_{\mathrm{TV}}(f_\eta^{\otimes n}, g_\eta^{\otimes n})$.

6. Numerical studies

6·1. *Practical considerations*

Choice of k: For practical implementation of the `MINTunknown` test, we consider a multiscale approach that averages over a range of values of k. To describe this approach, let $\mathcal{K} \subseteq \{1, \ldots, n-1\}$ and, for $k \in \mathcal{K}$, let $\hat{h}(k) = \hat{H}_{n,k}$ denote the unweighted Kozachenko–Leonenko entropy estimate with tuning parameter k based on the original data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Now, for $b = 1, \ldots, B$ and $k \in \mathcal{K}$, we let $\hat{h}^{(b)}(k) = \tilde{H}_{n,(k)}^{(b)}$ denote the Kozachenko–Leonenko entropy estimate with tuning parameter k based on the permuted data $Z_1^{(b)}, \ldots, Z_n^{(b)}$. Writing $\bar{h} = |\mathcal{K}|^{-1} \sum_{k \in \mathcal{K}} \hat{h}(k)$ and $\bar{h}^{(b)} = |\mathcal{K}|^{-1} \sum_{k \in \mathcal{K}} \hat{h}^{(b)}(k)$ for $b = 1, \ldots, B$, we then define the p-value for our test to be

$$\frac{1 + \sum_{b=0}^{B} \mathbb{1}_{\{\bar{h}^{(0)} \geq \bar{h}^{(b)}\}}}{B+1}.$$

By the exchangeability of $(\bar{h}, \bar{h}^{(1)}, \ldots, \bar{h}^{(B)})$ under H_0, the corresponding test has at most its nominal size. We refer to this test as `MINTav`, and remark that if $\mathcal{K}$ is taken to be a singleton

set then we recover `MINTunknown`. In our simulations below, we took $\mathcal{K} = \{1, \ldots, 20\}$ and $B = 100$.

Running time: Here, as in our study of the statistical properties of `MINT`, we consider d to be fixed. The slowest step in the computation of our test statistic is the computation of the nearest neighbour distances $\{\rho_{(1),i}, \ldots, \rho_{(k),i} : i = 1, \ldots, n\}$, which takes $O(kn \log n)$ operations (e.g. Vaidya, 1989). Since, in common with many other independence tests, we use a permutation approach with B permutations to obtain the critical value for our test, the overall complexity of our algorithm is $O(Bkn \log n)$. This compares favourably with several of the other methods listed in the Introduction, e.g. distance covariance, reproducing kernel Hilbert space methods and copula methods, which typically have complexity $O(Bn^2)$.

6·2. *Simulated data*

To study the empirical performance of our methods, we first compare our tests to existing approaches through their performance on simulated data. For comparison, we present corresponding results for a test based on the empirical copula process decribed by Kojadinovic and Holmes (2009) and implemented in the `R` package `copula` (Hofert et al., 2017), a test based on the HSIC implemented in the `R` package `dHSIC` (Pfister and Peters, 2017), a test based on the distance covariance implemented in the `R` package `energy` (Rizzo and Szekely, 2017) and the improvement of Hoeffding's test, known as Hoeffding's D, introduced in Weihs et al. (2018) and implemented in the `R` package `SymRC` (Weihs et al., 2017). We also present results for an oracle version of our tests, denoted simply as `MINT`, which for each parameter value in each setting, uses the most powerful choice of k. Throughout, we took $q = 0.05$ and $n = 200$, ran 5000 repetitions for each parameter setting, and for our comparison methods, used the default tuning parameter values recommended by the corresponding authors. We consider three classes of data generating mechanisms, designed to illustrate different possible types of dependence:

Setting 1: For $l \in \mathbb{N}$ and $(x, y) \in [-\pi, \pi]^2$, define the density function

$$f_l(x, y) = \frac{1}{4\pi^2}\{1 + \sin(lx)\sin(ly)\}.$$

This class of densities, which we refer to as sinusoidal, form a particularly interesting class. On the one hand, by the periodicity of the sine function, we have that the mutual information does not depend on l: indeed,

$$\begin{aligned} I(f_l) &= \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\{1 + \sin(lx)\sin(ly)\}\log\bigl(1 + \sin(lx)\sin(ly)\bigr)\,dx\,dy \\ &= \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}(1 + \sin u \sin v)\log(1 + \sin u \sin v)\,du\,dv \approx 0{\cdot}143. \end{aligned}$$

On the other hand, the class is identified by Sejdinovic et al. (2013) as challenging to detect dependence; intuitively, this is because as l increases, the dependence becomes increasingly localized, while the marginal densities are uniform on $[-\pi, \pi]$ for each l. To explain this intuition more formally, suppose we have any, potentially randomized, test of at most its nominal size q, with the additional randomness encoded via a random variable taking values in a space $\mathcal{T}$, say. Thus, for each $n \in \mathbb{N}$, we have a Borel measurable function $\phi_n : ([-\pi, \pi] \times [-\pi, \pi])^n \times \mathcal{T} \to \{0, 1\}$. Suppose further that $(X_1^{(l)}, Y_1^{(l)}), \ldots, (X_n^{(l)}, Y_n^{(l)})$ are independent and identically distributed pairs with density f_l, that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed pairs with the uniform density on $[-\pi, \pi] \times [-\pi, \pi]$ and that T takes values in $\mathcal{T}$ and is independent of our other data. Then it can be shown, e.g. using the Riemann–Lebesgue lemma,

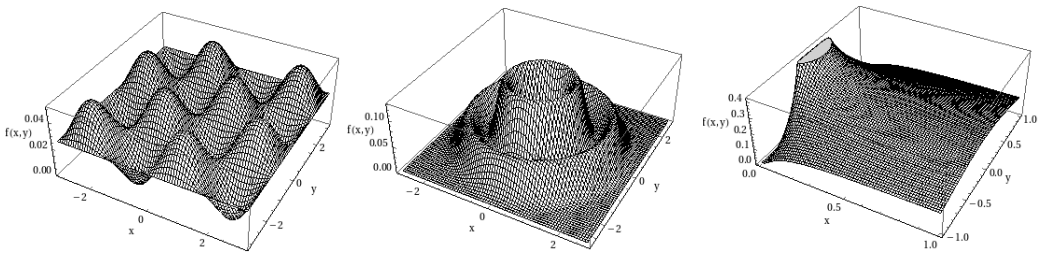


Fig. 1: Plots of the joint densities in our simulation study for particular values of the respective shape parameters. Left: Setting 1 with $l = 2$; middle: Setting 2 with $l = 2$; right: Setting 3 with $\rho = 0.3$.

that the distribution of $(X_1^{(l)}, Y_1^{(l)}, \ldots, X_n^{(l)}, Y_n^{(l)}, T)$ converges *strongly* as $l \to \infty$ to the distribution of $(X_1, Y_1, \ldots, X_n, Y_n, T)$ in the sense that for any Borel measurable subset B of $([-\pi, \pi] \times [-\pi, \pi])^n \times \mathcal{T}$, we have

$$\mathrm{pr}\big\{(X_1^{(l)}, Y_1^{(l)}, \ldots, X_n^{(l)}, Y_n^{(l)}, T) \in B\big\} \to \mathrm{pr}\big\{(X_1, Y_1, \ldots, X_n, Y_n, T) \in B\big\}$$

as $l \to \infty$. In particular,

$$\mathrm{pr}\big\{\phi_n(X_1^{(l)}, Y_1^{(l)}, \ldots, X_n^{(l)}, Y_n^{(l)}, T) = 1\big\} \to \mathrm{pr}\big\{\phi_n(X_1, Y_1, \ldots, X_n, Y_n, T) = 1\big\} \leq q,$$

and we conclude that, asymptotically as $l \to \infty$, no randomized test of at most nominal size can have better than trivial power.

Setting 2: Let $L, \Theta, \epsilon_1, \epsilon_2$ be independent with $L \sim U(\{1, \ldots, l\})$ for some $l \in \mathbb{N}$, $\Theta \sim U[0, 2\pi]$, and $\epsilon_1, \epsilon_2 \sim N(0, 1)$. Set $X = L\cos\Theta + \epsilon_1/4$ and $Y = L\sin\Theta + \epsilon_2/4$. For large values of l, the distribution of $(X/l, Y/l)$ approaches the uniform distribution on the unit disc. The distribution of (X, Y) is spherically symmetric with density given by

$$f(r\cos\theta, r\sin\theta) = \frac{8}{\pi l} e^{-8r^2} \sum_{s=1}^{l} e^{-8s^2} I_0(16sr),$$

for $r \geq 0$ and $\theta \in [0, 2\pi)$, where $I_0(z) = \pi^{-1} \int_0^\pi e^{z\cos t}\, dt$ is a modified Bessel function of the first kind.

Setting 3: Let X, ϵ be independent with $X \sim U[-1, 1]$, $\epsilon \sim N(0, 1)$, and for a parameter $\rho \in [0, \infty)$, let $Y = |X|^\rho \epsilon$.

Figure 1 gives plots of the joint densities for particular values of l, in Settings 1 and 2, and ρ, in Setting 3. For each of these three classes of data generating mechanisms, we also consider a corresponding multivariate setting in which we wish to test the independence of X and Y when $X = (X_1, X_2), Y = (Y_1, Y_2)$. Here, $(X_1, Y_1), X_2, Y_2$ are independent, with X_1 and Y_1 having the dependence structures described above, and $X_2, Y_2 \sim U(0, 1)$.

The results are presented in Figure 2. Naturally, given the huge range of different possible types of dependence, there is no uniformly most powerful test, and if the nature of the dependence were known in advance, it may well be possible to design a tailor-made test with good power. For instance, if it were known that the data were bivariate normal with non-negative correlation, Pearson's correlation coefficient would be expected to do well, and indeed this is verified empirically in the supplementary material. However, as mentioned in the Introduction, this test statistic has no power against certain other, non-linear, dependence structures. The aim of our simulation study, then, is to demonstrate the types of dependence structure for which `MINT` provides good power. In this regard, Figure 2 shows that, especially in Settings 1 and 2, the `MINT`

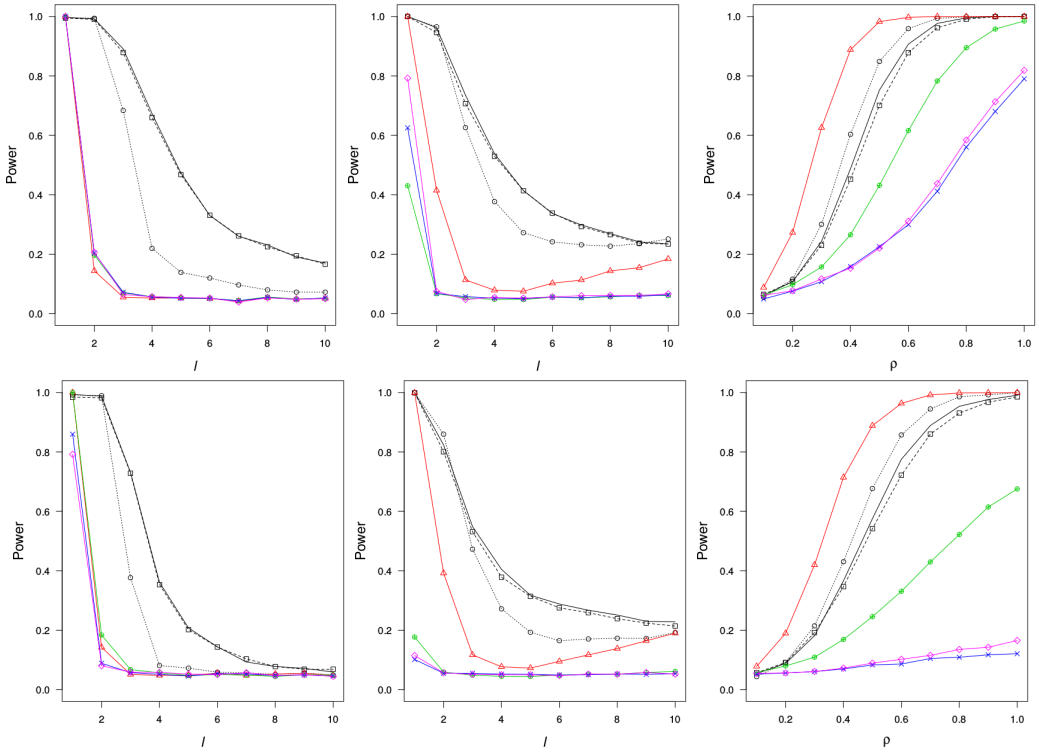
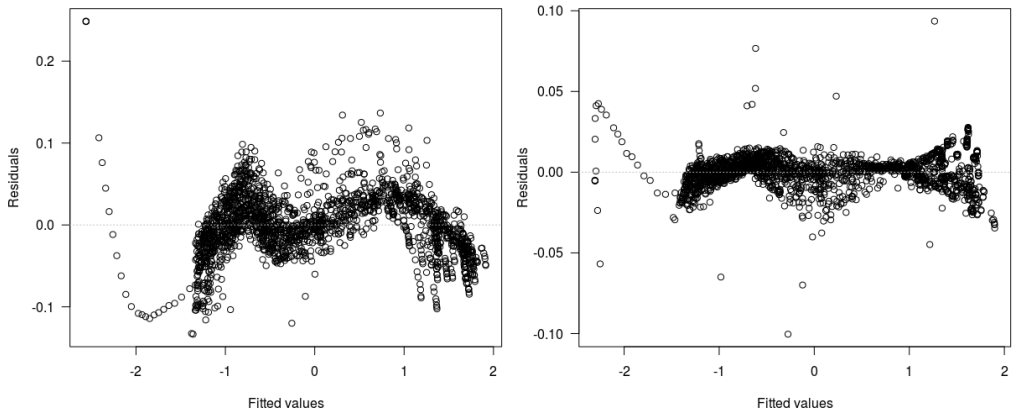


Fig. 2: Power curves as functions of the respective shape parameters for `MINT` (——), `MINTknown` (-□-), `MINTav` (·○·), HSIC (—△—), Distance covariance (—⊕—), Copula (—✳—), Hoeffding's D (—◇—) for Settings 1 (left), 2 (middle) and 3 (right). The marginals are univariate (top) and bivariate (bottom).

and `MINTav` approaches have very strong performance. In these examples, the dependence becomes increasingly localized as l increases, and the flexibility to choose a smaller value of k in such settings means that `MINT` approaches are particularly effective. Where the dependence is more global in nature, such as in Setting 3, other approaches may be better suited, though even here, `MINT` is competitive. `MINTav` appears to be a good method for tuning parameter selection; indeed, in Setting 3, it even outperforms the oracle choice of k.

6·3. *Real data*

In this section we illustrate the use of `MINTregression` on the CalCOFI oceanographic dataset, which is available from `https://www.kaggle.com/sohier/calcofi`, and which comprises various readings on water conditions off the coast of California from 1949 to 2016. To avoid time heterogeneity and dependence, we only consider those readings collected in November 2016, which are the most recent readings. We consider water temperature as the response variable of interest, with four predictor variables, namely depth below the surface, salinity, specific-volume anomaly and dynamic height, all of which were centred. After removing observations for which any of these variables were missing, there were 1989 observations. Our initial linear model, fit to the whole dataset, yielded the plot of residuals against fitted values shown in Figure 3a, indicating that this model is not a good fit to the data. To improve the model we next added quadratic terms in each of the covariates, but from the corresponding residual plot in Figure 3b, we see that this model is also not a good fit. Now, as a demonstration of `MINTregression`, we randomly subsampled this data with sample sizes



(a) (b)

Fig. 3: Plots of the residuals against fitted values for the linear models regressing water temperatue on depth below the surface, salinity, specific-volume anomaly and dynamic height without quadratic terms (left) and with quadratic terms (right).

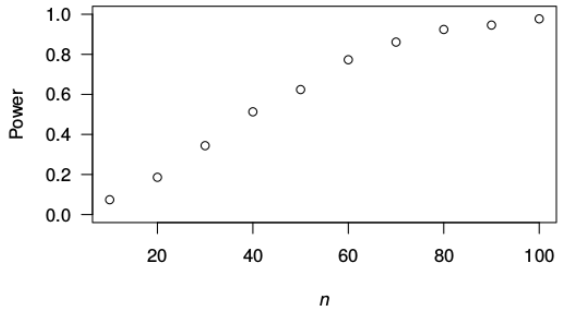


Fig. 4: Plot of the power of optimally tuned `MINTregression` against sample size.

$n \in \{10, 20, \ldots, 100\}$, fitted the linear model with quadratic terms and ran our procedure as described at the end of Section 5 to test for the goodness-of-fit. A range of values of k and k_η were used with $B = 100$ and $q = 0.05$ and estimates of the power of our tests were found by averaging over 1000 repetitions of the subsampling. For each value of n, the most powerful choices of k and k_η were selected, and the power of these tests is shown in Figure 4. It can be seen from this figure that, even though we fit a linear model with 9 covariates, including the intercept, and estimate entropies of five-dimensional random vectors, `MINTregression` achieves good power with relatively small sample sizes.

ACKNOWLEDGEMENTS

T.B.B. and R.J.S. were supported by an EPSRC Programme grant. T.B.B. was supported by a PhD scholarship from the SIMS fund; R.J.S. was supported by an EPSRC Fellowship and the Leverhulme Trust. We would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme 'Statistical Scalability' when work on this paper was undertaken. This work was supported by EPSRC grant numbers EP/K032208/1 and EP/R014604/1. We thank the anonymous reviewers for their constructive comments.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of all stated results and additional numerical studies.

REFERENCES

ALBERT, M., BOURET, Y., FROMONT, M. & REYNAUD-BOURET, P. (2015). Bootstrap and permutation tests of independence for point processes. *Ann. Statist.*, **43**, 2537–64.

BACH, F. R. & JORDAN, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.*, **3**, 1–48.

BERRETT, T. B., GROSE, D. J. & SAMWORTH, R. J. (2018a). **IndepTest**: nonparametric independence tests based on entropy estimation. Available at https://cran.r-project.org/web/packages/IndepTest/index.html.

BERRETT, T. B., SAMWORTH, R. J. & YUAN, M. (2018b). Efficient multivariate entropy estimation via k-nearest neighbour distances. *Ann. Statist.*, to appear.

BERRETT, T. B., WANG, Y., BARBER, R. F. AND SAMWORTH, R. J. (2018c) The conditional permutation test. *arXiv*:1807.05405.

BIAU, G. & 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.

COVER, T. M. & THOMAS, J. A. (2006). *Elements of Information Theory* (2nd edition). Wiley, Hoboken, New Jersey.

DALGAARD, P. (2002). *Introductory Statistics with R*. Springer-Verlag, New York.

DOBSON, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall, London.

EINMAHL, J. H. J. & VAN KEILEGOM, I. (2008). Tests for independence in nonparametric regression. *Statistica Sinica*, **18**, 601–15.

FAN, J., FENG, Y. & XIA, L. (2017). A projection based conditional dependence measure with applications to high-dimensional undirected graphical models. *arXiv*:1501.01617.

FAN, Y., LAFAYE DE MICHEAUX, P., PENEV, S. & SALOPEK, D. (2017). Multivariate nonparametric test of independence. *J. Multivariate Anal.*, **153**, 189–210.

GIBBS, A. L. & SU, F. E. (2002). On choosing and bounding probability metrics. *Int. Statist. Review*, **70**, 419–35.

GRETTON A., BOUSQUET O., SMOLA A. & SCHÖLKOPF B. (2005). Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Algorithmic Learning Theory*, 63–77.

GRETTON, A. & GYÖRFI, L. (2010). Consistent nonparametric tests of independence. *J. Mach. Learn. Res.*, **11**, 1391–423.

HELLER, R., HELLER, Y., KAUFMAN, S., BRILL, B. & 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.

JITKRITTUM, W., SZABÓ, Z. & GRETTON, A. (2016). An adaptive test of independence with analytic kernel embeddings. *arXiv*:1610.04782.

JOE, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.*, **84**, 157–64.

JOSSE, J. & HOLMES, S. (2016). Measuring multivariate association and beyond. *Statist. Surveys*, **10**, 132–67.

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

KOJADINOVIC, I. & 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. & LEONENKO, N. N. (1987). Sample estimate of the entropy of a random vector. *Probl. Inform. Transm.*, **23**, 95–101.

KRASKOV, A., STÖGBAUER H. & GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E*, **69**, 066138.

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.

MARI, D. D. & KOTZ, S. (2001). *Correlation and Dependence*. Imperial College Press, London.

MILLER, E. G. & FISHER, J. W. (2003). ICA using spacings estimates of entropy. *J. Mach. Learn. Res.*, **4**, 1271–95.

MÜLLER, U. U., SCHICK, A. & WEFELMEYER, W. (2012). Estimating the error distribution function in semiparametric additive regression models. *J. Stat. Plan. Inference*, **142**, 552–66.

NEUMEYER, N. (2009). Testing independence in nonparametric regression. *J. Multivariate Anal.*, **100**, 1551–66.

NEUMEYER, N. & VAN KEILEGOM, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. Multivariate Anal.*, **101**, 1067–78.

NGUYEN, D. & EISENSTEIN, J. (2017). A kernel independence test for geographical language variation. *Comput. Ling., to appear*.

PEARL, J. (2009). *Causality*. Cambridge University Press, Cambridge.

PEARSON, K. (1920). Notes on the history of correlation. *Biometrika*, **13**, 25–45.
PFISTER, N., BÜHLMANN, P., SCHÖLKOPF, B. & PETERS, J. (2017). Kernel-based tests for joint independence. *J. Roy. Statist. Soc., Ser. B*, to appear.
PFISTER, N. & PETERS, J. (2017). dHSIC: Independence Testing via Hilbert Schmidt Independence Criterion. *R Package version 2.0*. https://cran.r-project.org/web/packages/dHSIC/index.html.
RIZZO, M. L. & 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.
SAMWORTH, R. J. & YUAN, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.*, **40**, 2973–3002.
SCHWEIZER, B. & WOLFF, E. F. (1981). On nonparametric measures of dependence for random variables. *Ann. Statist.*, **9**, 879–85.
SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. & FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, **41**, 2263–91.
SEN, A. & SEN, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika*, **101**, 927–42.
SHAH, R. D. & BÜHLMANN, P. (2017). Goodness of fit tests for high-dimensional linear models. *J. Roy. Statist. Soc., Ser. B*, to appear.
SHAH, R. D. AND PETERS, J. (2018). The hardness of conditional independence and the generalised covariance measure. *arXiv*:1804.07203.
SONG, L., SMOLA, A., GRETTON, A., BEDO, J. & BORGWARDT, K. (2012). Feature selection via dependence maximization. *J. Mach. Learn. Res.*, **13**, 1393–434.
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.
STIGLER, S. M. (1989). Francis Galton's account of the invention of correlation. *Stat. Sci.*, **4**, 73–86.
SU, L. & WHITE, H. (2008). A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, **24**, 829–64.
SZÉKELY, G. J., RIZZO, M. L. & BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–94.
SZÉKELY, G. J. & RIZZO, M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Multivariate Anal.*, **117**, 193–213.
TORKKOLA, K. (2003). Feature extraction by non-parametric mutual information maximization. *J. Mach. Learn. Res.*, **3**, 1415–38.
VAIDYA, P. M. (1989). An $O(n \log n)$ algorithm for the all-nearest-neighbors problem. *Discrete Comput. Geom.*, **4**, 101–15.
VINH, N. X., EPPS, J. & BAILEY, J. (2010). Information theoretic measures for clusterings comparison: variants, properties, normalisation and correction for chance. *J. Mach. Learn. Res.*, **11**, 2837–54.
WEIHS, L., DRTON, M. & LEUNG, D. (2016). Efficient computation of the Bergsma–Dassios sign covariance. *Comput. Stat.*, **31**, 315–28.
WEIHS, L., DRTON, M. & 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`.
WU, E. H. C., YU, P. L. H. & LI, W. K. (2009). A smoothed bootstrap test for independence based on mutual information. *Comput. Stat. Data Anal.*, **53**, 2524–36.
YAO, S., ZHANG, X. & SHAO, X. (2017). Testing mutual independence in high dimension via distance covariance. *J. Roy. Statist. Soc., Set. B*, to appear.
ZHANG, K., PETERS, J., JANZING, D. & SCHÖLKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, AUAI Press, USA, 804-813.
ZHANG, Q., FILIPPI, S., GRETTON, A. & SEJDINOVIC, D. (2017). Large-scale kernel methods for independence testing. *Stat. Comput.*, **27**, 1–18.