ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)

Kernel-based tests for joint independence

Niklas Pfister and Peter Bühlmann,

Eidgenössiche Technische Hochschule Zürich, Switzerland

Bernhard Schölkopf

Max Planck Institute for Intelligent Systems, Tübingen, Germany

and Jonas Peters

Max Planck Institute for Intelligent Systems, Tübingen, Germany, and University of Copenhagen, Denmark

[Received April 2016. Final revision March 2017]

Summary. We investigate the problem of testing whether d possibly multivariate random variables, which may or may not be continuous, are jointly (or mutually) independent. Our method builds on ideas of the two-variable Hilbert–Schmidt independence criterion but allows for an arbitrary number of variables. We embed the joint distribution and the product of the marginals in a reproducing kernel Hilbert space and define the d-variable Hilbert–Schmidt independence criterion dHSIC as the squared distance between the embeddings. In the population case, the value of dHSIC is 0 if and only if the d variables are jointly independent, as long as the kernel is characteristic. On the basis of an empirical estimate of dHSIC, we investigate three non-parametric hypothesis tests: a permutation test, a bootstrap analogue and a procedure based on a gamma approximation. We apply non-parametric independence testing to a problem in causal discovery and illustrate the new methods on simulated and real data sets.

Keywords: Causal inference; Independence test; Kernel methods; V-statistics

1. Introduction

We consider the problem of non-parametric testing for joint or mutual independence of d random variables. This is a very different and more ambitious task than testing pairwise independence of a collection of random variables. Consistent pairwise non-parametric independence tests date back to Feuerverger (1993) and Romano (1986) and have more recently received considerable attention by using kernel-based methods (Gretton *et al.*, 2005, 2007), and other related approaches for estimating or testing pairwise (in)dependence including distance correlations (Székely and Rizzo, 2009, 2014), rank-based correlations (Bergsma and Dassios, 2014; Leung and Drton, 2016; Nandy *et al.*, 2016) or also non-parametric and semiparametric copula-based correlations (Liu *et al.*, 2012; Xue and Zou, 2012; Wegkamp and Zhao, 2016; Gaißer *et al.*, 2010).

One of our motivations to develop methods for non-parametric testing of joint independence originates from the area of causal inference, and we discuss this in Section 5.2: there, inferring pairwise independence is not sufficient as those models assume the existence of jointly independent noise variables. Our test can therefore be used as a goodness-of-fit test and for model selection; see Section 5.2. A further interesting application of joint independence tests is inde-

Address for correspondence: Niklas Pfister, Seminar für Statistik, Eidgenössiche Technische Hochschule Zürich, Rämistrasse 101, Zürich 8092, Switzerland.
E-mail: niklas.pfister@stat.math.ethz.ch

1369–7412/17/80000

pendent component analysis. Whereas many algorithms use a stepwise approach to construct the collection of independent features, a more direct option is to minimize a measure of mutual dependence explicitly (such as our d-variable Hilbert–Schmidt independence criterion (HSIC) dHSIC); for more details see Chen and Bickel (2006) or Matteson and Tsay (2016). (We thank a referee for pointing out this interesting application.)

For testing joint independence, consider the distribution $\mathbb{P}^{(X^1,\ldots,X^d)}$ of the random vector $\mathbf{X} = (X^1,\ldots,X^d)$. (Throughout the paper, a superscript on X always denotes an index rather than an exponent.) By definition, $(X^1,\ldots,X^d)$ are jointly or mutually independent if and only if $\mathbb{P}^{(X^1,\ldots,X^d)} = \mathbb{P}^{X^1} \otimes \ldots \otimes \mathbb{P}^{X^d}$. For a given positive definite kernel, we map both distributions into the reproducing kernel Hilbert space (RKHS) (see Section 2.1 for details) and consider their squared distance. Such a mapping can in fact be seen as a generalization of the L^2-distance between 'traditional' kernel density estimators; see the discussion on page 732 in Gretton, Borgwardt, Rasch, Schölkopf and Smola (2012). For characteristic kernels (e.g. the popular Gaussian kernel), the embedding of Borel probability measures is injective and the squared distance is 0 if and only if the variables are jointly independent. For the finite sample case, we compute a suitable estimator that can be used as a test statistic. We then construct three statistical tests: two tests are based on permutation and bootstrap procedures, and a third test approximates the distribution of the test statistic under independence with a gamma distribution. Our statistic extends the HSIC (Gretton *et al.*, 2005) and contains it as a special case. We therefore call the corresponding test procedure the d-variable HSIC dHSIC. We prove that the permutation-based approach has correct level and that the bootstrap approach has pointwise asymptotic level and is consistent in the sense that it has asymptotic power equal to 1 against any fixed alternative; see equation (3.5) in Section 3.

In the literature, other mutual independence tests have been proposed. Kankainen (1995) discussed a characteristic-function-based non-parametric mutual independence test; see Section 2.4. The dependence measure is a weighted integral over the difference between the characteristic functions of the joint and the product distribution. All weight functions result in special cases of dHSIC for an appropriate choice of kernel. We show that our results carry over to the characteristic function framework, whereas the opposite direction works for only a restrictive class of kernels. Moreover, although Kankainen (1995) did prove similar results about the asymptotic distribution of the test statistic as given in theorem 2 in Section 3.1, her proof cannot be directly extended to our more general framework. This is one of the reasons why we developed some of our general results about V-statistics. The test in Kankainen (1995) is shown to be consistent, but the word consistency there refers to the property that the asymptotic distribution of the test statistic under the alternative hypothesis diverges; instead, we employ the commonly used definition that a test is consistent if the testing procedure itself (in our case the bootstrap) is consistent in the sense that it has asymptotic power equal to 1; see equation (3.5). Our consistency results immediately carry over to the characteristic function framework, as it is contained as a special case of dHSIC.

Bakirov *et al.* (2006) used an independence coefficient as the measure of dependence, which is defined as the normalized distance between the characteristic functions of the product and marginal distributions and is hence strongly related to the approach by Kankainen (1995). They approximated the asymptotic test statistic, which is also a sum of χ^2-distributed random variables, using tail bounds. This results in a test that has (conservative) asymptotic level in the sense of inequality (3.4). However, because of the conservative bounds which are independent of the dependence strength, the resulting test is, in general, not able to detect all fixed dependences, even in the large sample limit.

One test for which a consistency result as in equation (3.5) has been shown is an older method based on Beran and Millar (1987) and Romano (1986), page 27; it does not seem to be used in practice very often. As a test statistic, it takes the maximal difference between the empirical distribution and the product of its marginals over a class of sets. One then chooses a distribution over sets and approximates this infinite class by $C < \infty$ randomly chosen sets; see Section 5.1. This makes the construction impractical with a rather *ad hoc* computational implementation. In our experiments, we found that this test has less power than dHSIC and is computationally more demanding, even for moderate values of C.

Both this test and the characteristic-function-based tests mentioned above are restricted to the Euclidean space; dHSIC allows more general kernels such as kernels on graphs or strings (see Gretton *et al.* (2007)).

Finally, it is possible to use the following alternative procedure that constructs a joint independence test from a bivariate test: joint independence holds if and only if for all $k \in \{2, \dots, d\}$ we have that X^k is independent of $(X^1, \dots, X^{k-1})$. We can therefore perform $d-1$ statistical tests and combine the results by using a Bonferroni correction. However, such a procedure is asymmetric in the d random variables and depends on the order of the random variables. Furthermore, it is known that the Bonferroni correction is often conservative and, because of performing $d-1$ tests, such a test is of order d times more computationally expensive than the direct dHSIC-approach; see Section 5.3.3.

1.1. Contribution

This work extends the two-variable HSIC (Gretton *et al.*, 2005, 2007; Smola *et al.*, 2007) to testing joint independence for an arbitrary number of variables. The resulting test, moreover, extends the work of Kankainen (1995) to the more flexible framework of kernel methods (see Section 2.4) and establishes consistency, as mentioned also in the previous section. Although the dHSIC test statistic was briefly mentioned by Sejdinovic *et al.* (2013), the derivation of the general results about asymptotic distributions (theorem 2 and theorem 3 in Section 3.1) as well as the mathematically rigorous treatment of the permutation test and the bootstrap test are novel: this concerns results for both types of test about their level (type I error) in proposition 3 in Section 3.2.1 and theorem 4 in Section 3.2.2 and the consistency (asymptotic power) of the bootstrap in theorem 5. In fact, the consistency result is quite remarkable, establishing asymptotic consistency for any fixed alternative. It is the first such result for kernel-based methods and maybe the first result for a practically feasible test for joint independence having asymptotic error control and asymptotic power equal to 1. We also prove that under the null hypothesis it holds that $\xi_2(h) > 0$, which has been implicitly assumed in for example Gretton *et al.* (2007), theorem 2.

For the gamma-approximation-based test, we compute general formulae for the mean and for the variance; see propositions 4 and 5 in Section 3.3. To make our tests accessible we have created an R package called `dHSIC`, which is available on the Comprehensive R Archive Network. Moreover, we have applied our dHSIC to real data in causality, showing its usefulness also in applied settings, in terms of both model selection and goodness-of-fit tests.

To establish these properties, we derive new results for V-statistics (see the on-line appendix C) that are of independent interest: lemma C.3 there (asymptotic difference between U- and V-statistics), theorem C.5 (asymptotic variance of a V-statistic), theorem C.6 (asymptotic bias of a V-statistic), theorem C.9 (asymptotic distribution of a degenerate V-statistic) and theorem C.13 (asymptotic distribution of a degenerate resampled V-statistic).

2. Hilbert–Schmidt independence criterion for *d*-variables

2.1. Reproducing kernel Hilbert spaces

We present here a brief introduction to RKHSs and the theory of mean embeddings. Given a set $\mathcal{X}$ we call a function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a positive semidefinite kernel if for any set of points $(x_1, \dots, x_n) \in \mathcal{X}^n$ the corresponding Gram matrix $(k(x_i, x_j))_{1 \leqslant i,j \leqslant n}$ is symmetric and positive semidefinite. Moreover, denote by $\mathcal{F}(\mathcal{X})$ the space of functions from $\mathcal{X}$ to $\mathbb{R}$. RKHSs on $\mathcal{X}$ are well behaved subclasses of $\mathcal{F}(\mathcal{X})$ defined as follows.

Definition 1 (RKHS). Let $\mathcal{X}$ be a set and let $\mathcal{H} \subseteq \mathcal{F}(\mathcal{X})$ be a Hilbert space. Then $\mathcal{H}$ is called an RKHS if there is a kernel k on $\mathcal{X}$ satisfying

(a) $\forall\, x \in \mathcal{X}$: $k(x, \cdot) \in \mathcal{H}$, and
(b) $\forall\, f \in \mathcal{H}, \forall\, x \in \mathcal{X} : \langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$. We then call k a *reproducing kernel* of $\mathcal{H}$.

For any positive semidefinite kernel k there is an RKHS with reproducing kernel k. A commonly used kernel on $\mathbb{R}^m$ is the Gaussian kernel, defined for all $x, y \in \mathbb{R}^m$ by

$$k(x, y) = \exp\{-\|x - y\|^2_{\mathbb{R}^m} / (2\sigma^2)\}. \tag{2.1}$$

It is possible to embed complicated objects into an RKHS and to analyse them by using the Hilbert space structure. Inner products can be expressed as function evaluations via the reproducing property, which simplifies computation within an RKHS. In this paper, we embed probability distributions in an RKHS. For this, we use the Bochner integral to define an embedding of $\mathcal{M}(\mathcal{X}) := \{\mu | \mu \text{ is a finite Borel measure on } \mathcal{X}\}$ into an RKHS.

Definition 2 (mean embedding function). Let $\mathcal{X}$ be a separable metric space, let k be a continuous bounded positive semidefinite kernel and let $\mathcal{H}$ be the RKHS with reproducing kernel k. Then, the *mean embedding* (associated with k) is defined as the function $\Pi : \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{H}$ with

$$\Pi(\mu) := \int_{\mathcal{X}} k(x, \cdot) \mu(\mathrm{d}x).$$

To infer that two distributions are equal given that their embeddings coincide, it is necessary that the mean embedding is injective. A kernel is called characteristic if the mean embedding Π is injective (see Fukumizu *et al.* (2007)). The Gaussian kernel (2.1) on $\mathbb{R}^m$, for example, is characteristic (e.g. Sriperumbudur *et al.* (2008), theorem 7).

2.2. Definition of dHSIC and independence property

Our goal is to develop a non-parametric hypothesis test to determine whether the components of a random vector $\mathbf{X} = (X^1, \dots, X^d)$ are mutually independent, based on an independently and identically distributed (IID) sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ of the vector $\mathbf{X}$. By definition, joint independence holds if and only if

$$\mathbb{P}^{X^1} \otimes \dots \otimes \mathbb{P}^{X^d} = \mathbb{P}^{(X^1, \dots, X^d)}.$$

The central idea is to embed both $\mathbb{P}^{X^1} \otimes \dots \otimes \mathbb{P}^{X^d}$ and $\mathbb{P}^{(X^1, \dots, X^d)}$ into an appropriate RKHS and then to check whether the embedded elements are equal. To keep an overview of all our assumptions, we summarize the setting that is used throughout the rest of this work.

2.2.1. Setting 1 (dHSIC)

For all $j \in \{1, \dots, d\}$, let $\mathcal{X}^j$ be a separable metric space and denote by $\boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \dots \times \mathcal{X}^d$ the product space. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and, for every $j \in \{1, \dots, d\}$, let $X^j : \Omega \rightarrow \boldsymbol{\mathcal{X}}^j$ be a random variable with law $\mathbb{P}^{X^j}$. Let $(\mathbf{X}_i)_{i \in \mathbb{N}}$ be a sequence of IID copies of $\mathbf{X} = (X^1, \dots, X^d)$.

For $j \in \{1, \ldots, d\}$, let $k^j : \mathcal{X}^j \times \mathcal{X}^j \rightarrow \mathbb{R}$ be a continuous, bounded, positive semidefinite kernel on $\mathcal{X}^j$ and denote by $\mathcal{H}^j$ the corresponding RKHS. Moreover, assume that the tensor product of the kernels k^j denoted by $\mathbf{k} = k^1 \otimes \ldots \otimes k^d$ is characteristic (Gretton (2015) argued (for $d = 2$) that this follows if the individual kernels are characteristic) and let $\boldsymbol{\mathcal{H}} = \mathcal{H}^1 \otimes \ldots \otimes \mathcal{H}^d$ be the (projective) tensor product of the RKHSs $\mathcal{H}^j$. Let $\Pi : \mathcal{M}(\mathcal{X}) \rightarrow \boldsymbol{\mathcal{H}}$ be the mean embedding function associated with $\mathbf{k}$.

It is straightforward to show that this setting ensures that $\boldsymbol{\mathcal{H}}$ is an RKHS with reproducing kernel $\mathbf{k}$, that $\mathbf{k}$ is continuous and bounded, that $\boldsymbol{\mathcal{H}}$ is separable and contains only continuous functions and that Π is injective. Using this setting we can extend the HSIC from two variables as described by Gretton *et al.* (2007) to the case of d variables. The extension is based on the HSIC characterization via the mean embedding described by Smola *et al.* (2007).

Definition 3 (dHSIC). Assume setting 1. Then, define the statistical functional

$$\mathrm{dHSIC}(\mathbb{P}^{(X^1,\ldots,X^d)}) := \|\Pi(\mathbb{P}^{X^1} \otimes \ldots \otimes \mathbb{P}^{X^d}) - \Pi(\mathbb{P}^{(X^1,\ldots,X^d)})\|^2_{\boldsymbol{\mathcal{H}}}$$

and call it dHSIC.

Therefore, dHSIC is the distance between the joint measure and the product measure after embedding them into an RKHS. Since the mean embedding Π is injective we obtain the following relationship between dHSIC and joint independence.

Proposition 1 (independence property of dHSIC). Assume setting 1. Then it holds that

$$\mathrm{dHSIC}(\mathbb{P}^{(X^1,\ldots,X^d)}) = 0 \quad \Leftrightarrow \quad \mathbb{P}^{X^1} \otimes \ldots \otimes \mathbb{P}^{X^d} = \mathbb{P}^{(X^1,\ldots,X^d)}.$$

This proposition implies that we can use dHSIC as a measure of joint dependence between variables. We express dHSIC in terms of the individual kernels $k^1, \ldots, k^d$, which will later be the basis of the estimator that is defined in Section 2.3. A proof is given in the on-line appendix D.6.

Proposition 2 (expansion of dHSIC). Assume setting 1. Then it holds that

$$\mathrm{dHSIC} = \mathbb{E}\left\{ \prod_{j=1}^{d} k^j(X_1^j, X_2^j) \right\} + \mathbb{E}\left\{ \prod_{j=1}^{d} k^j(X_{2j-1}^j, X_{2j}^j) \right\} - 2\,\mathbb{E}\left\{ \prod_{j=1}^{d} k^j(X_1^j, X_{j+1}^j) \right\}.$$

2.3. Estimating dHSIC

Our estimator will be constructed by using several V-statistics. We therefore start by summarizing a few well-known definitions and the most important results from the theory of V-statistics. Readers who are familiar with these topics may skip directly to definition 4.

Let $n \in \mathbb{N}$, $q \in \{1, \ldots, n\}$, let $\mathcal{X}$ be a metric space, $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, $X : \Omega \rightarrow \mathcal{X}$ a random variable with law $\mathbb{P}^X$ and let $(X_i)_{i \in \mathbb{N}}$ be IID copies of X, i.e. $(X_i)_{i \in \mathbb{N}} \sim^{\mathrm{IID}} \mathbb{P}^X$. Define $\mathbf{M}_q(n) := \{1, \ldots, n\}^q$ as the q-fold Cartesian product of the set $\{1, \ldots, n\}$. Consider a measurable and symmetric (i.e. invariant under any permutation of its input arguments) function $g : \mathcal{X}^q \rightarrow \mathbb{R}$, which we call the core function. The V-statistic

$$V_n(g) := \frac{1}{n^q} \sum_{\mathbf{M}_q(n)} g(X_{i_1}, \ldots, X_{i_q}) \tag{2.2}$$

then estimates the statistical functional $\theta_g := \theta_g(\mathbb{P}^X) := \mathbb{E}\{g(X_1, \ldots, X_q)\}$. As opposed to U-statistics, defined in expression (C.1) in the on-line appendix C, V-statistics are usually biased.

Here, we consider a V-statistic because it can be computed much faster than the corresponding U-statistic, especially if $q>2$. Whereas U-statistics have been extensively studied (e.g. Serfling (1980)), results for V-statistics are often restricted to $q=2$. For dHSIC, we require $q=2d$ (see lemma 1) and therefore develop general results for $q>2$ in the on-line appendix C.

The following notation appears throughout the paper in the context of V-statistics and is also commonly used for U-statistics; see Serfling (1980), section 5.1.5. Given the core function $g:\boldsymbol{\mathcal{X}}^q \to \mathbb{R}$ we define for every $c \in \{1,\ldots,q-1\}$ the function $g_c:\boldsymbol{\mathcal{X}}^c \to \mathbb{R}$ by

$$g_c(x_1,\ldots,x_c) := \mathbb{E}\{g(x_1,\ldots,x_c,X_{c+1},\ldots,X_q)\}$$

and $g_q \equiv g$. Then, g_c is again a symmetric core function such that for every $c \in \{1,\ldots,q-1\}$

$$\mathbb{E}\{g_c(X_1,\ldots,X_c)\} = \mathbb{E}\{g(X_1,\ldots,X_q)\} = \theta_g.$$

Further define $\tilde{g} \equiv g-\theta_g$ and for all $c \in \{1,\ldots,q\}$ define $\tilde{g}_c \equiv g_c - \theta_g$ to be the centred versions of the core functions. Moreover, define, for every $c \in \{1,\ldots,q\}$,

$$\xi_c := \mathrm{var}\{g_c(X_1,\ldots,X_c)\} = \mathbb{E}\{\tilde{g}_c(X_1,\ldots,X_c)^2\}. \tag{2.3}$$

We sometimes write $\xi_c(g)$ to make clear which core function we are talking about.

We now estimate each term in proposition 2 by a V-statistic.

Definition 4 ($\widehat{\mathrm{dHSIC}}$). Assume setting 1. For all $(\mathbf{x}_1,\ldots,\mathbf{x}_n) \in \boldsymbol{\mathcal{X}}^n$ we define $\widehat{\mathrm{dHSIC}} = (\widehat{\mathrm{dHSIC}}_n)_{n\in\mathbb{N}}$ as

$$\begin{aligned}\widehat{\mathrm{dHSIC}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n) := &\frac{1}{n^2}\sum_{\mathbf{M}_2(n)}\prod_{j=1}^{d} k^j(x^j_{i_1},x^j_{i_2}) + \frac{1}{n^{2d}}\sum_{\mathbf{M}_{2d}(n)}\prod_{j=1}^{d} k^j(x^j_{i_{2j-1}},x^j_{i_{2j}})\\ &-\frac{2}{n^{d+1}}\sum_{\mathbf{M}_{d+1}(n)}\prod_{j=1}^{d} k^j(x^j_{i_1},x^j_{i_{j+1}})\end{aligned}$$

if $n \in \{2d, 2d+1,\ldots\}$ and as $\widehat{\mathrm{dHSIC}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n) := 0$ if $n \in \{1,\ldots,2d-1\}$.

Whenever it is clear from the context, we drop the functional arguments and just write $\widehat{\mathrm{dHSIC}}_n$ instead of $\widehat{\mathrm{dHSIC}}_n(\mathbf{X}_1,\ldots,\mathbf{X}_n)$. To make this estimator more accessible for analysis we can express it as a V-estimator with a single core function. For this, define $h:\boldsymbol{\mathcal{X}}^{2d} \to \mathbb{R}$ to be the function satisfying for all $\mathbf{z}_1,\ldots,\mathbf{z}_{2d} \in \boldsymbol{\mathcal{X}}$ that

$$h(\mathbf{z}_1,\ldots,\mathbf{z}_{2d}) = \frac{1}{(2d)!}\sum_{\pi\in S_{2d}}\left\{\prod_{j=1}^{d} k^j(z^j_{\pi(1)},z^j_{\pi(2)}) + \prod_{j=1}^{d} k^j(z^j_{\pi(2j-1)},z^j_{\pi(2j)}) - 2\prod_{j=1}^{d} k^j(z^j_{\pi(1)},z^j_{\pi(j+1)})\right\}, \tag{2.4}$$

where S_{2d} is the set of permutations on $\{1,\ldots,2d\}$. The following lemma shows that $\widehat{\mathrm{dHSIC}}$ is a V-statistic with core function h. A proof is given in the on-line appendix D.7.

Lemma 1 (properties of the core function h). Assume setting 1. It holds that the function h that is defined in equation (2.4) is symmetric continuous, and there exists $C>0$ such that for all $\mathbf{z}_1,\ldots,\mathbf{z}_{2d} \in \boldsymbol{\mathcal{X}}$ we have $|h(\mathbf{z}_1,\ldots,\mathbf{z}_{2d})| < C$. Moreover, $V_n(h) = \widehat{\mathrm{dHSIC}}_n$ (see expression (2.2)), and $\theta_h = \mathbb{E}\{h(\mathbf{X}_1,\ldots,\mathbf{X}_{2d})\} = \mathrm{dHSIC}$.

2.4. *Characteristic function framework*

Kankainen (1995) considered a characteristic-function-based mutual independence test. She considered a weighted integral over the difference between the characteristic functions of the joint and the product distribution. For a weight function g, the resulting empirical test statistic (Kankainen (1995), page 25) is given by

$$T_n(\mathbf{x}_1,\ldots,\mathbf{x}_n) := n\Bigg\{\frac{1}{n^2}\sum_{i_1,i_2}\prod_{j=1}^{d}\int_{\mathbb{R}}\exp\{\mathrm{i}t^j(x^j_{i_1}-x^j_{i_2})\}g_j(t^j)\,\mathrm{d}t^j$$

$$+\frac{1}{n^{2d}}\prod_{j=1}^{d}\sum_{i_1,i_2}\int_{\mathbb{R}}\exp\{\mathrm{i}t^j(x^j_{i_1}-x^j_{i_2})\}g_j(t^j)\mathrm{d}t^j-\frac{2}{n^{d+1}}\sum_{i_1}\prod_{j=1}^{d}\sum_{i_2}\int_{\mathbb{R}}\exp\{\mathrm{i}t^j(x^j_{i_1}-x^j_{i_2})\}g_j(t^j)\,\mathrm{d}t^j\Bigg\}.$$

The characteristic function framework is contained in the dHSIC-framework as a special case. We recover our dHSIC test statistic by choosing

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \prod_{j=1}^{d}\int_{\mathbb{R}}\exp\{\mathrm{i}t^j(x^j-y^j)\}g_j(t^j)\,\mathrm{d}t^j. \tag{2.5}$$

This choice is justified by Bochner's theorem (e.g. Unser and Tafti (2014), theorem B.1).

Theorem 1 (Bochner's theorem). Let f be a bounded continuous function on $\mathbb{R}^d$. Then, f is positive semidefinite if and only if it is the (conjugate) Fourier transform of a non-negative and finite Borel measure μ, i.e.

$$f(\mathbf{x}) = \int_{\mathbb{R}^d}\exp(\mathrm{i}\langle\mathbf{x},\mathbf{t}\rangle)\,\mu(\mathrm{d}\mathbf{t}).$$

Given the characteristic function framework with a weight function g satisfying properties 1–5 in Kankainen (1995), page 25, it holds that the measure $\mu_g(\mathbf{B}) := \int_{\mathbf{B}}\Pi_{j=1}^{d}g_j(t^j)\,\mathrm{d}t^j$ is a non-negative finite Borel measure on $\mathbb{R}^d$ and hence $\mathbf{k}$ defined as in equation (2.5) is a positive semidefinite kernel. The setting that was given in Kankainen (1995) is thus entirely contained within our dHSIC-framework.

Furthermore, the dHSIC-framework is strictly more general. To see this, let $\mathbf{k}$ be a continuous bounded stationary positive semidefinite kernel on $\mathbb{R}^d$. Then, by stationarity there is a continuous bounded function f on $\mathbb{R}^d$ such that $\mathbf{k}(\mathbf{x},\mathbf{y}) = f(\mathbf{x}-\mathbf{y})$ and hence by Bochner's theorem there is a measure $\mu\in\mathcal{M}(\mathbb{R}^d)$ such that

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d}\exp(\mathrm{i}\langle\mathbf{x}-\mathbf{y},\mathbf{t}\rangle)\,\mu(\mathrm{d}\mathbf{t}).$$

This is, however, still more general than the setting in Kankainen (1995) as there it was additionally assumed that the measure μ is absolutely continuous with a density g satisfying properties 1–5, which in particular requires that g is a simple product, i.e. $g(t) = \Pi_{j=1}^{d}g_j(t^j)$ and that the components g_j are even. Both of these conditions are essential to the proofs that were given in Kankainen (1995). Therefore the results from the characteristic function framework in Kankainen (1995) cannot be transferred to our more general dHSIC-setting. Also note that the characteristic function framework is restricted to real-valued domains, whereas kernels are more flexible, e.g. kernels on graphs or strings (see Gretton *et al.* (2007)).

3. Statistical tests for joint independence

Assume setting 1 and denote by $\mathcal{P}(\boldsymbol{\mathcal{X}})$ the space of Borel probability measures. In this section we derive three statistical hypothesis tests for the null hypothesis

$$H_0 := \{\mu\in\mathcal{P}(\boldsymbol{\mathcal{X}})|\mathbf{X}\sim\mu = \mathbb{P}^{\mathbf{X}}, \mathbb{P}^{\mathbf{X}} = \mathbb{P}^{X^1}\otimes\ldots\otimes\mathbb{P}^{X^d}\} \tag{3.1}$$

against the alternative

Table 1. Properties of the three hypothesis tests

Hypothesis test	*Consistency*	*Level*	*Speed*
Permutation†	Unknown (remark 2, Section 3.2.1)	Valid (proposition 3, Section 3.2.1)	Slow
Bootstrap†	Pointwise (theorem 5, Section 3.2.2)	Pointwise asymptotic (theorem 4, Section 3.2.2)	Slow
Gamma approximation	No guarantee	No guarantee	Fast

†For implementation one can use the Monte Carlo approximation. This leads to a reasonably fast implementation, while conserving the (asymptotic) level and consistency results. Further details are given in Section 4.2.

$$H_{\mathrm{A}} := \{\mu \in \mathcal{P}(\boldsymbol{\mathcal{X}}) | \mathbf{X} \sim \mu = \mathbb{P}^{\mathbf{X}}, \mathbb{P}^{\mathbf{X}} \neq \mathbb{P}^{X^1} \otimes \ldots \otimes \mathbb{P}^{X^d}\}. \tag{3.2}$$

On the basis of the asymptotic behaviour given in theorem 2 in Section 3.1, we consider $n\,\widehat{\mathrm{dHSIC}}_n$ as test statistic and define a decision rule $\varphi = (\varphi_n)_{n\in\mathbb{N}}$ encoding rejection of H_0 if $\varphi_n = 1$ and no rejection of H_0 if $\varphi_n = 0$. For all $n \in \{1, \ldots, 2d-1\}$ we define $\varphi_n := 0$ and for all $n \in \{2d, 2d+1, \ldots\}$ and for all $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \boldsymbol{\mathcal{X}}^n$ we set

$$\varphi_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \mathbb{1}_{\{n\,\widehat{\mathrm{dHSIC}}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) > c_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)\}} \tag{3.3}$$

where the threshold $c = (c_n)_{n\in\mathbb{N}}$ remains to be chosen. Ideally, for fixed $\alpha \in (0,1)$ the hypothesis test should have (valid) level α, i.e., for every $\mu = \mathbb{P}^{\mathbf{X}} \in H_0$ and all n, $\mathbb{P}\{\varphi_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) = 1\} \leqslant \alpha$, where $\mathbf{X}_1, \mathbf{X}_2, \ldots \sim^{\mathrm{IID}} \mathbb{P}^X \in H_0$. A weaker condition states that the test respects the level in the large sample limit, i.e., for every $\mu = \mathbb{P}^X \in H_0$,

$$\limsup_{n\to\infty} \mathbb{P}\{\varphi_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) = 1\} \leqslant \alpha, \tag{3.4}$$

where $\mathbf{X}_1, \mathbf{X}_2, \ldots \sim^{\mathrm{IID}} \mathbb{P}^{\mathbf{X}} \in H_0$. Such a test is said to have pointwise asymptotic level. Additionally, the test is called pointwise consistent if for all fixed $\mathbb{P}^{\mathbf{X}} \in H_{\mathrm{A}}$ it holds that

$$\lim_{n\to\infty} \mathbb{P}\{\varphi_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) = 1\} = 1, \tag{3.5}$$

where $\mathbf{X}_1, \mathbf{X}_2, \ldots \sim^{\mathrm{IID}}$ (fixed) $\mathbb{P}^{\mathbf{X}} \in H_{\mathrm{A}}$. Table 1 summarizes the properties that our three tests satisfy.

In Section 3.1 we consider some of the asymptotic properties of the test statistic $n\,\widehat{\mathrm{dHSIC}}_n$. In particular, we show the existence of an asymptotic distribution under H_0. We then construct three hypothesis tests of the form (3.3). The first two are a permutation test and a bootstrap test which are discussed in Section 3.2. Both tests are based on resampling and hence do not rely on explicit knowledge of the asymptotic distribution under H_0. In Section 3.3 we consider a third test which is based on an approximation of the asymptotic distribution under H_0 by using a gamma distribution.

3.1. Asymptotic behaviour of the test statistic

We first determine the asymptotic distribution of $n\,\widehat{\mathrm{dHSIC}}_n$ under H_0, extending Gretton *et al.* (2007), theorem 2, from HSIC to dHSIC.

Theorem 2 (asymptotic distribution of $n\,\widehat{\mathrm{dHSIC}}_n$ under H_0). Assume setting 1. Let $(Z_i)_{i\in\mathbb{N}}$

be a sequence of independent standard normal random variables on $\mathbb{R}$, let

$$T_{h_2} \in L\{L^2 \mathbb{P}^{(X^1,\ldots,X^d)}, |\cdot|_{\mathbb{R}}\}$$

be such that for all $f \in L^2(\mathbb{P}^{(X^1,\ldots,X^d)}, |\cdot|_{\mathbb{R}})$ and for all $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ it holds that

$$(T_{h_2}(f))(\mathbf{x}) = \int_{\boldsymbol{\mathcal{X}}} h_2(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \mathbb{P}^{(X^1,\ldots,X^d)}(\mathrm{d}\mathbf{y}).$$

(Given a measure space $(\Omega, \mathcal{F}, \mu)$ the space $\mathcal{L}^r(\mu, |\cdot|_{\mathbb{R}})$ consists of all measurable functions $f: \Omega \rightarrow \mathbb{R}$ such that $\int_\Omega |f(\omega)|^r \mu(\mathrm{d}\omega) < \infty$. The corresponding space of equivalence classes of such functions is denoted by $L^r(\mu, |\cdot|_{\mathbb{R}})$. Moreover, we denote the space of all linear bounded operators from a Banach space $\mathcal{B}$ onto itself by $L(\mathcal{B})$.) Denote by $(\lambda_i)_{i \in \mathbb{N}}$ the eigenvalues of T_{h_2}. Then under H_0 it holds that

$$\xi_2(h) > 0$$

and

$$n\, \widehat{\mathrm{dHSIC}}_n \xrightarrow{\mathrm{d}} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2 \qquad \text{as } n \rightarrow \infty.$$

The proof of this result relies on the asymptotic distribution of degenerate V-statistics (see theorem C.9 in the on-line appendix). To show that the degenerate setting applies we need to prove that under H_0 it holds that $\xi_1(h) = 0$ and $\xi_2(h) > 0$. The latter statement is of interest in itself and has been for example implicitly assumed in Gretton *et al.* (2007), theorem 2. But, whereas $\xi_1(h) = 0$ follows more or less directly from the independence assumption under H_0 (see lemma D.9 in the on-line appendix), the condition $\xi_2(h) > 0$ is difficult to verify directly because of the complicated form of the core function h. We therefore circumvent direct verification by using empirical process theory to prove that the asymptotic distribution of $n\, \widehat{\mathrm{dHSIC}}_n$ has certain continuity properties (see theorem D.3 in the on-line appendix) that are not satisfied by the asymptotic distribution resulting from the theory of V-statistics if both $\xi_1(h)$ and $\xi_2(h)$ were 0. A full proof is given in the on-line appendix D.2.

Remark 1 (estimation of eigenvalues). It is possible to construct a test that estimates the eigenvalues of the integral operator in theorem 2 by first estimating the eigenvalues of the Gram matrix corresponding to h_2 and then computing the asymptotic distribution by using a bootstrap procedure; see Gretton *et al.* (2009). Given knowledge of the exact form of h_2 and under the assumption that h_2 is positive definite (which can be shown for $d = 2$, but is unknown for $d > 2$) one can prove consistency; see Pfister (2016). However, since h_2 is a complicated function (see lemma D.8 in the on-line appendix) depending on the unknown distribution $\mathbb{P}^{\mathbf{X}}$ (as opposed to Gretton *et al.* (2009)) one must estimate h_2, which means one would have to account additionally for that approximation. In simulations, the eigenvalue estimation generally performed worse than the gamma approximation in almost all our experiments. We have therefore decided not to include this approach in the paper. There is, however, an implementation in the `dHSIC` R package.

The following theorem (a proof is given in the on-line appendix D.2) is an important result required to establish consistency (of the bootstrap test), stating that $n\, \widehat{\mathrm{dHSIC}}_n$ diverges under H_A.

Theorem 3 (asymptotic distribution of $n\, \widehat{\mathrm{dHSIC}}_n$ under H_A). Assume setting 1. Then under H_A it holds for all $t \in \mathbb{R}$ that

$$\lim_{n \rightarrow \infty} \mathbb{P}(n\, \widehat{\mathrm{dHSIC}}_n \leqslant t) = 0.$$

3.2. Resampling tests

We first introduce the notation of a general resampling scheme which encompasses the permutation and bootstrap method that is used later. For every function $\boldsymbol{\psi}=(\psi^1,\ldots,\psi^d)$ satisfying for all $i\in\{1,\ldots,d\}$ that $\psi^i:\{1,\ldots,n\}\rightarrow\{1,\ldots,n\}$, define the function $g_{n,\boldsymbol{\psi}}:\boldsymbol{\mathcal{X}}^n\rightarrow\boldsymbol{\mathcal{X}}^n$

$$g_{n,\boldsymbol{\psi}}(\mathbf{x}_1,\ldots,\mathbf{x}_n):=(\mathbf{x}^{\boldsymbol{\psi}}_{n,1},\ldots,\mathbf{x}^{\boldsymbol{\psi}}_{n,n}),\ (\mathbf{x}_1,\ldots,\mathbf{x}_n)\in\boldsymbol{\mathcal{X}}^n, \tag{3.6}$$

where $\mathbf{x}^{\boldsymbol{\psi}}_{n,i}:=(x^1_{\psi^1(i)},\ldots,x^d_{\psi^d(i)})$. The diagram (3.7) illustrates the mapping $g_{n,\boldsymbol{\psi}}$:

$$\begin{array}{c|ccc} \mathbf{x}_1 & x^1_1 & \cdots & x^d_1 \\ \vdots & \vdots & & \vdots \\ \mathbf{x}_n & x^1_n & \cdots & x^d_n \end{array} \quad \xrightarrow{g_{n,\boldsymbol{\psi}}} \quad \begin{array}{c|ccc} \mathbf{x}^{\boldsymbol{\psi}}_{n,1} & x^1_{\psi^1(1)} & \cdots & x^d_{\psi^d(1)} \\ \vdots & \vdots & & \vdots \\ \mathbf{x}^{\boldsymbol{\psi}}_{n,n} & x^1_{\psi^1(n)} & \cdots & x^d_{\psi^d(n)} \end{array} \tag{3.7}$$

Define

$$B_n:=\{\psi:\{1,\ldots,n\}\rightarrow\{1,\ldots,n\}\,|\,\psi \text{ is a function}\}; \tag{3.8}$$

then for a subset $A_n\subseteq B^d_n$ we call the family of functions

$$g:=(g_{n,\boldsymbol{\psi}})_{\boldsymbol{\psi}\in A_n} \tag{3.9}$$

a resampling method. In the following two sections we formulate the bootstrap and permutation tests in terms of this resampling method.

3.2.1. Permutation test

The permutation test is based on the resampling (3.9) with $A_n=(S_n)^d$, where S_n is the set of permutations on $\{1,\ldots,n\}$. More precisely, we have the following definition.

Definition 5 (permutation test for dHSIC). Assume setting 1 and $\alpha\in(0,1)$. For all $\boldsymbol{\psi}\in(S_n)^d$, let $g_{n,\boldsymbol{\psi}}$ be defined as in expression (3.6). Moreover, for $n\in\{2d,2d+1,\ldots\}$, let $\hat{R}_n:\boldsymbol{\mathcal{X}}^n\times\mathbb{R}\rightarrow[0,1]$ be the resampling distribution functions defined for all $t\in\mathbb{R}$ by

$$\hat{R}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)(t):=\frac{1}{(n!)^d}\sum_{\boldsymbol{\psi}\in(S_n)^d}\mathbb{1}_{\{n\,\widehat{\text{dHSIC}}_n\{g_{n,\boldsymbol{\psi}}(\mathbf{x}_1,\ldots,\mathbf{x}_n)\}\leqslant t\}}. \tag{3.10}$$

Then the α-permutation test for dHSIC is defined by $\varphi_n:=0$ for $n\in\{1,\ldots,2d-1\}$, and for $n\in\{2d,2d+1,\ldots\}$ by

$$\varphi_n(\mathbf{x}_1,\ldots,\mathbf{x}_n):=\mathbb{1}_{\{n\,\widehat{\text{dHSIC}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)>\hat{R}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{-1}(1-\alpha)\}}.$$

Given that the resampling method has a group structure and additionally satisfies for all $\mathbf{X}$ with $\mathbb{P}^{\mathbf{X}}\in H_0$ that

$$g_{n,\boldsymbol{\psi}}(\mathbf{X}_1,\ldots,\mathbf{X}_n) \text{ is equal in distribution to } (\mathbf{X}_1,\ldots,\mathbf{X}_n),$$

where $\mathbf{X}_1,\mathbf{X}_2,\ldots\sim^{\text{IID}}\mathbb{P}^{\mathbf{X}}$, it can be shown that tests of this form have valid level. For the permutation test for dHSIC both these properties are satisfied; hence it has valid level.

Proposition 3 (permutation test for dHSIC has valid level). Assume setting 1 and let H_0 and H_A be defined as in expressions (3.1) and (3.2). Then for all $\alpha\in(0,1)$ the α-permutation test for dHSIC has valid level α when testing H_0 against H_A.

A proof is given in the on-line appendix D.3. It is important to note that the level property from proposition 3 is for the finite sample setting and does not depend on the asymptotic behaviour of the test statistics.

The size of the set $(S_n)^d$ is given by $(n!)^d$; therefore computing expression (3.10) quickly becomes infeasible. For implementation purposes we generally use a Monte Carlo approximated version, and the details are given in Section 4.2. Surprisingly, it can be shown that, whenever the probability distribution $\mathbb{P}^{\mathbf{X}}$ is continuous, the Monte Carlo approximated permutation test also has valid level; see proposition B.4 and the comments thereafter in the on-line appendix.

Remark 2 (pointwise consistency of the permutation test). Given the similarity between bootstrap and permutation tests, it seems likely that the permutation test for dHSIC is consistent, also. The proof of theorem 5 in Section 3.2.2, however, cannot be easily extended. A more promising approach would be to proceed similarly to Romano (1989), as the test statistics considered there are closely related to dHSIC; see expression (5.1), and expression (D.2) in the on-line appendix. The essential idea there is to use the theory of empirical processes (see the on-line appendix D.1) to prove the assumptions of Lehmann and Romano (2005), theorem 15.2.3. Unfortunately, we could not extend the results in Romano (1989) from Vapnik–Chervonenkis classes of sets to the required classes of functions. Although many results extend more or less directly (see the on-line appendix D.1), the difficulties lie in proving a similar representation for S_n to that given in the display of Romano (1989), proof of proposition 3.1, as well as a result similar to Romano (1989), lemma 5.1. As a side remark, extending the empirical process approach that was given in Romano (1988) to give an alternative proof of theorem 5 would require a uniform Donsker property for the unit ball of the RKHS.

3.2.2. *Bootstrap test*

The bootstrap test is based on the resampling (3.9) with $A_n = B_n^d$.

Definition 6 (bootstrap test for dHSIC). Assume setting 1 and $\alpha \in (0,1)$. For all $\psi \in B_n^d$ let the function $g_{n,\psi}$ be defined as in expression (3.6). Moreover, for $n \in \{2d, 2d+1, \ldots\}$, let $\hat{R}_n : \boldsymbol{\mathcal{X}}^n \times \mathbb{R} \to [0,1]$ be the resampling distribution functions defined for all $t \in \mathbb{R}$ by

$$\hat{R}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)(t) := \frac{1}{n^{nd}} \sum_{\psi \in B_n^d} \mathbb{1}_{\{n\, \widehat{\mathrm{dHSIC}}_n \{g_{n,\psi}(\mathbf{x}_1,\ldots,\mathbf{x}_n)\} \leqslant t\}}.$$

Then the α-bootstrap test for dHSIC is defined by $\varphi_n := 0$ for all $n \in \{1, \ldots, 2d-1\}$, and for $n \in \{2d, 2d+1, \ldots\}$ by

$$\varphi_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \mathbb{1}_{\{n\, \widehat{\mathrm{dHSIC}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n) > \hat{R}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{-1}(1-\alpha)\}}.$$

Unlike for the permutation test, the bootstrap resampling method no longer exhibits a group. We cannot therefore expect the bootstrap test to have valid level. However, it is possible to show that it has pointwise asymptotic level and even pointwise consistency. The reason that this can be done is that the resampling method in the bootstrap test is connected to the empirical product distribution $\hat{\mathbb{P}}_n^{X^1} \otimes \ldots \otimes \hat{\mathbb{P}}_n^{X^d}$. The following theorem proves that the bootstrap test for dHSIC has pointwise asymptotic level.

Theorem 4 (bootstrap test for dHSIC has pointwise asymptotic level). Assume setting 1 and let H_0 and H_{A} be defined as in expressions (3.1) and (3.2). Then for all $\alpha \in (0,1)$ the α-bootstrap test for dHSIC has pointwise asymptotic level α when testing H_0 against H_{A}.

A proof is given in the on-line appendix D.4. We now establish that the bootstrap test for dHSIC is consistent.

Theorem 5 (consistency of the bootstrap test for dHSIC). Assume setting 1 and let H_0 and H_{A} be defined as in expressions (3.1) and (3.2). Then for all $\alpha \in (0,1)$ the α-bootstrap test is pointwise consistent when testing H_0 against H_{A}.

A proof is given in the on-line appendix D.4. Similarly to the permutation test, the size of the set $(B_n)^d$ is n^{nd} which grows quickly. Again, we may use a Monte Carlo approximated version; see Section 4.2. In Chwialkowski *et al.* (2014) a similar consistency analysis has been performed for a wild bootstrap approach on time series.

3.3. Gamma approximation

We showed in theorem 2 that the asymptotic distribution of $n\,\widehat{\mathrm{dHSIC}}_n$ equals

$$\binom{2d}{2}\sum_{i=1}^{\infty}\lambda_i Z_i^2. \tag{3.11}$$

The essential idea behind the gamma approximation (see also Kankainen (1995) and Gretton *et al.* (2005)) is that a distribution of the form $\Sigma_{i=1}^{\infty}\lambda_i Z_i^2$ can be approximated fairly well by a gamma distribution with matched first and second moments (see Satterthwaite (1946) for basic empirical evidence). The intuition is that the gamma distribution would be correct if the sequence of eigenvalues λ_i from the integral operator contains only a finite number of non-zero values, which implies that it is a good approximation as long as the sequence of λ_i decays sufficiently fast. This has, however, been shown only empirically and no guarantees in the large sample limit are available. In fact, it is rather unlikely that such guarantees even exist as it is not difficult to find choices of λ_i for which expression (3.11) is not a gamma distribution. It is not as simple, however, to show that such values of λ_i can actually occur as solutions of the defining integral equation. Nevertheless, the approximation seems to work well for small d (see Section 5), and the test can be computed much faster than the other approaches.

The gamma distribution with parameters α and β is denoted by gamma(α,β) and corresponds to the distribution with density

$$f(x)=\frac{x^{\alpha-1}\exp(x/\beta)}{\beta^{\alpha}\Gamma(\alpha)},$$

where $\Gamma(t)=\int_0^{\infty}x^{t-1}\exp(-x)\,\mathrm{d}x$ is the gamma function. The first two moments of the gamma (α,β)-distributed random variable Y are given by $\mathbb{E}(Y)=\alpha\beta$ and $\mathrm{var}(Y)=\alpha\beta^2$. To match the first two moments we define for $\mathbf{X}_1,\mathbf{X}_2,\ldots\sim^{\mathrm{IID}}\mathbb{P}^{\mathbf{X}}\in H_0$ the two parameters

$$\alpha_n(\mathbb{P}^{\mathbf{X}}):=\mathbb{E}(\widehat{\mathrm{dHSIC}}_n)^2/\mathrm{var}(\widehat{\mathrm{dHSIC}}_n),$$
$$\beta_n(\mathbb{P}^{\mathbf{X}}):=n\mathrm{var}(\widehat{\mathrm{dHSIC}}_n)/\mathbb{E}(\widehat{\mathrm{dHSIC}}_n).$$

Then we use the approximation

$$n\,\widehat{\mathrm{dHSIC}}_n(\mathbf{X}_1,\ldots,\mathbf{X}_n)\sim\mathrm{gamma}\{\alpha_n(\mathbb{P}^{\mathbf{X}}),\beta_n(\mathbb{P}^{\mathbf{X}})\}. \tag{3.12}$$

The following two propositions give expansions of the involved moments in terms of the kernel.

Proposition 4 (mean of $\widehat{\mathrm{dHSIC}}$). Assume setting 1. Then under H_0 it holds that, as $n\to\infty$,

$$\mathbb{E}(\widehat{\mathrm{dHSIC}}_n)=\frac{1}{n}-\frac{1}{n}\sum_{r=1}^{d}\prod_{j\neq r}\mathbb{E}\{k^j(X_1^j,X_2^j)\}+\frac{d-1}{n}\prod_{j=1}^{d}\mathbb{E}\{k^j(X_1^j,X_2^j)\}+\mathcal{O}(n^{-2}).$$

A proof is given in the on-line appendix D.5

Proposition 5 (variance of $\widehat{\mathrm{dHSIC}}$). Assume setting 1. Then under H_0 it holds that

$$\mathrm{var}(\widehat{\mathrm{dHSIC}}_n)=2\frac{(n-2d)!}{n!}\frac{(n-2d)!}{(n-4d+2)!}\left\{\prod_{j=1}^{d}e_1(j)+(d-1)^2\prod_{j=1}^{d}e_0(j)^2+2(d-1)\prod_{j=1}^{d}e_2(j)\right.$$
$$+\sum_{j=1}^{d}e_1(j)\prod_{r\neq j}e_0(r)^2-2\sum_{j=1}^{d}e_1(j)\prod_{r\neq j}e_2(r)-2(d-1)\sum_{j=1}^{d}e_2(j)\prod_{r\neq j}e_0(r)^2$$
$$\left.+\sum_{j\neq l}e_2(j)e_2(l)\prod_{r\neq j,l}e_0(r)^2\right\}+\mathcal{O}(n^{-5/2})$$

as $n\to\infty$ and where, for all $j\in\{1,\ldots,d\}$,

$$e_0(j)=\mathbb{E}\{k^j(X_1^j,X_2^j)\},$$
$$e_1(j)=\mathbb{E}\{k^j(X_1^j,X_2^j)^2\},$$
$$e_2(j)=\mathbb{E}_{X_1^j}[\mathbb{E}_{X_2^j}\{k^j(X_1^j,X_2^j)\}^2].$$

A proof is given in the on-line Appendix D.5. On the basis of these two propositions we need only a method to estimate the terms $e_0(j)$, $e_1(j)$ and $e_2(j)$ for all $j\in\{1,\ldots,d\}$. One could use a U-statistic (C.1) in the on-line appendix for each expectation term as this would not add any bias. It turns out, however, that a computationally more efficient V-statistic also does not add any asymptotic bias in this particular case. This is due to theorem C.6 in the on-line appendix describing that the bias of a V-statistic is of order $\mathcal{O}(n^{-1})$ and hence is consumed by the error terms in propositions 4 and 5. The V-statistics for these terms are given for all $(\mathbf{x}_1,\ldots,\mathbf{x}_n)\in\boldsymbol{\mathcal{X}}^n$ by

$$\hat{e}_0(j)(\mathbf{x}_1,\ldots,\mathbf{x}_n):=\frac{1}{n^2}\sum_{i_1,i_2=1}^{n}k^j(x_{i_1}^j,x_{i_2}^j),$$
$$\hat{e}_1(j)(\mathbf{x}_1,\ldots,\mathbf{x}_n):=\frac{1}{n^2}\sum_{i_1,i_2=1}^{n}k^j(x_{i_1}^j,x_{i_2}^j)^2,$$
$$\hat{e}_2(j)(\mathbf{x}_1,\ldots,\mathbf{x}_n):=\frac{1}{n^3}\sum_{i_2=1}^{n}\left\{\sum_{i_1=1}^{n}k^j(x_{i_1}^j,x_{i_2}^j)\right\}^2.$$

On the basis of these terms we define the estimators $\hat{\mathbb{E}}_n$ and $\widehat{\mathrm{var}}_n$ for the mean and variance of $\widehat{\mathrm{dHSIC}}_n$ respectively by replacing all appearances of $e_0(j)$, $e_1(j)$ and $e_2(j)$ in propositions 4 and 5 by $\hat{e}_0(j)$, $\hat{e}_1(j)$ and $\hat{e}_2(j)$. We use the plug-in estimators

$$\hat{\alpha}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n):=\frac{\hat{\mathbb{E}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)^2}{\widehat{\mathrm{var}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)},$$
$$\hat{\beta}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n):=\frac{n\,\widehat{\mathrm{var}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)}{\hat{\mathbb{E}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)},\qquad(3.13)$$

and then define the following hypothesis test.

Definition 7 (gamma-approximation-based test for dHSIC). Assume setting 1 and $\alpha\in(0,1)$. Let $F_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ be the distribution function that is associated with the gamma$\{\hat{\alpha}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n),\hat{\beta}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)\}$ distribution, where $\hat{\alpha}_n$ and $\hat{\beta}_n$ are defined as in expression (3.13). Then the α–gamma approximation based test for dHSIC is defined by $\varphi_n:=0$ for all $n\in\{1,\ldots,2d-1\}$, and for $n\in\{2d,2d+1,\ldots\}$ by

$$\varphi_n(\mathbf{x}_1,\ldots,\mathbf{x}_n):=\mathbb{1}_{\{n\,\widehat{\mathrm{dHSIC}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)>F_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{-1}(1-\alpha)\}}.$$

4. Implementation

We now discuss an efficient implementation of the tests proposed and briefly comment on the choice of kernel. All methods are available in the R package dHSIC in the Comprehensive R Archive Network.

4.1. dHSIC-estimator

The dHSIC-estimator $\widehat{\text{dHSIC}}$ can be computed in quadratic time; see algorithm 1 in Table 2. Here, $\mathbf{1}_{k\times l}$ denotes a $k\times l$ matrix of 1s, the functions Sum and ColumnSum take the sums of all elements in a matrix and its columns respectively and an asterisk denotes the elementwise multiplication operator. The variables term1, term2 and term3 are related to the three components of the sum in definition 4, after changing the order of products and sums.

4.2. Resampling tests

From the definition of $\hat{R}_n$ we see that the permutation and bootstrap test involve $(n!)^d$ or n^{nd} evaluations of $\widehat{\text{dHSIC}}$ respectively. Instead of computing $\hat{R}_n$ explicitly we can use the Monte Carlo approximation; see definition B.1 in the on-line appendix. The p-value is then given by

$$\hat{p}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n) := \frac{1+|\{i\in\{1,\ldots,B\} : \widehat{\text{dHSIC}}\{g_{n,\psi_i}(\mathbf{x}_1,\ldots,\mathbf{x}_n)\} \geqslant \widehat{\text{dHSIC}}(\mathbf{x}_1,\ldots,\mathbf{x}_n)\}|}{1+B},$$

where $(\psi_i)_{i\in\mathbb{N}}$ is a sequence drawn from the uniform distribution on A_n (i.e. on $(S_n)^d$ for the permutation test and on B_n^d for the bootstrap test). The test then rejects the null hypothesis whenever $\hat{p}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)\leqslant\alpha$. Further details including critical values are provided in the on-line appendix B. Davison and Hinkley (1997) suggested the use of B between 99 and 999.

4.2.1. Permutation test

The Monte Carlo approximated version of the permutation test for dHSIC evaluates $\widehat{\text{dHSIC}}$ only B times (for B random permutations) instead of $(n!)^d$ times. The corresponding test (with the p-value as in Section 4.2 above) has valid level for any finite B: as in the proof of proposition 3 the resampling method g for the permutation test is a resampling group satisfying the invariance conditions (B.1) and (B.2) in the on-line appendix. Proposition B.4 there then shows that the Monte Carlo approximated permutation test has valid level for any finite B, given that we have continuous random variables as input. Algorithm 1 in the on-line appendix B implements the p-value and the critical value for the Monte Carlo approximated permutation test.

Table 2. Algorithm 1 computing the dHSIC V-estimator

```
Procedure dHSIC($\mathbf{x}_1,\ldots,\mathbf{x}_n$)
  for $j=1:d$ do
    $\mathbf{K}^j \leftarrow$ Gram matrix of kernel $k^j$ given $\mathbf{x}_1,\ldots,\mathbf{x}_n$
  term1 $\leftarrow \mathbf{1}_{n\times n}$; term2 $\leftarrow 1$; term3 $\leftarrow (2/n)\mathbb{1}_{1\times n}$
  for $j=1:d$ do
    term1 $\leftarrow$ term1 $*\,\mathbf{K}^j$
    term2 $\leftarrow (1/n^2)$term2 Sum($\mathbf{K}^j$)
    term3 $\leftarrow (1/n)$term3 $*$ ColumnSum($\mathbf{K}^j$)
  dHSIC $\leftarrow (1/n^2)$ Sum(term1) + term2 − Sum(term3)
  return dHSIC
```

4.2.2. Bootstrap test

Similarly, the Monte Carlo approximated version of the bootstrap test for dHSIC evaluates $\widehat{\text{dHSIC}}$ only B times (for B random draws with replacement) instead of n^{nd} times. One can show that the corresponding test (with the p-value as in Section 4.2 above) still has pointwise asymptotic level and is pointwise consistent if both n and B go to ∞. This follows from a standard concentration inequality argument (e.g. Lehmann and Romano (2005), theorem 11.2.18 and example 11.2.13). Algorithm 1 in the on-line appendix B implements the p-value and the critical value for the Monte Carlo approximated bootstrap test.

4.3. Gamma approximation test

Implementing the α–gamma approximation test consists of four steps (see Section 3.3).

Step 1: for all $j \in \{1,\ldots,d\}$ implement the estimators $\hat{e}_0(j), \hat{e}_1(j), \hat{e}_2(j)$.
Step 2: compute the estimates $\hat{\mathbb{E}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ and $\widehat{\text{var}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)$.
Step 3: using expression (3.13) compute the estimates $\hat{\alpha}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ and $\hat{\beta}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)$.
Step 4: compute the $(1-\alpha)$-quantile of the gamma$\{\hat{\alpha}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n), \hat{\beta}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)\}$ distribution.

The hypothesis test rejects H_0 if $n\,\widehat{\text{dHSIC}}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ is larger than the $(1-\alpha)$-quantile of the gamma$\{\hat{\alpha}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n), \hat{\beta}_n(\mathbf{x}_1,\ldots,\mathbf{x}_n)\}$ distribution calculated in the last step.

4.4. Choice of kernel

The choice of the kernel determines how well certain types of dependence can be detected and therefore influences the practical performance of dHSIC (see simulation 4 in Section 5.3). For continuous data a common choice is a Gaussian kernel as defined in expression (2.1). It is characteristic, which ensures that all the above results hold. In particular, any type of dependence can be detected in the large sample limit. We use the median heuristic for choosing the bandwidth σ by requiring that median$\{\|\mathbf{x}_i-\mathbf{x}_j\|^2_{\mathbb{R}^m} : i<j\} = 2\sigma^2$. This heuristic performs quite well in many practical applications. It may be possible, however, to extend alternative approaches from two-sample testing to independence testing (e.g. Gretton, Sejdinovic, Strathmann, Balakrishnan, Pontil, Fukumizu and Sriperumbudur (2012)). For discrete data, we choose a trivial kernel defined by $k(x,y) := \mathbb{1}_{\{x=y\}}$.

In practice, it is, moreover, possible and potentially beneficial also to consider other (potentially non-characteristic) kernels that are chosen in such a way that they are particularly powerful in detecting certain types of dependences.

5. Experiments

5.1. Competing methods

For comparisons we consider an approach, which has been suggested by Beran and Millar (1987) and Romano (1986), page 27. For testing the joint independence of d real-valued random variables $X^1,\ldots,X^d$, they considered the test statistic

$$\widehat{\text{BMR}}_n := \sup_{\mathbf{a}\in\mathbb{R}^d} |\hat{\mathbb{P}}_n(A_{\mathbf{a}}) - \hat{\mathbb{P}}_n^{\otimes}(A_{\mathbf{a}})|, \tag{5.1}$$

where $A_{\mathbf{a}} := (-\infty, a^1] \times (-\infty, a^2] \times \ldots \times (-\infty, a^d]$ is a subset of $\mathbb{R}^d$, $\hat{\mathbb{P}}_n := (1/n)\Sigma_{i=1}^n \delta_{\mathbf{X}_i}$ is the empirical joint measure and $\hat{\mathbb{P}}_n^{\otimes} := \Pi_{j=1}^d \{(1/n)\Sigma_{i=1}^n \delta_{X_i^j}\}$ is the empirical product measure. Usually, expression (5.1) cannot be computed exactly and must be approximated. We may choose a distribution μ with full support on $\mathbb{R}^d$, for example, and compute the supremum over $C<\infty$

randomly chosen $\mathbf{a}_1, \ldots, \mathbf{a}_C \sim^{\text{IID}} \mu$. In our experiments, we mainly choose $C = n$ since, for consistency, C must grow with n and since then the computational complexity is $\mathcal{O}(dn^2)$, which equals the computational complexity of dHSIC; see Section 4.1. As neither Beran and Millar (1987) nor Romano (1986) provided any other suggestion, we choose μ to be the d-dimensional Gaussian distribution with parameters estimated by maximum likelihood. The test itself is then based on a bootstrap procedure, which was described in Section 3.2.2. In the remainder of this section, we refer to this test as BMR-C.

Furthermore, we consider a multiple pairwise version of the two-variable HSIC test. To test for joint independence we use the following testing sequence.

Step 1: use HSIC to test whether X^d is independent of $[X^1, \ldots, X^{d-1}]$,
Step 2: use HSIC to test whether X^{d-1} is independent of $[X^1, \ldots, X^{d-2}]$,

$\vdots$

Step $d-1$: use HSIC to test whether X^2 is independent of X^1.

Finally, we account for the increased familywise error rate by using the Bonferroni correction, i.e. we perform all tests at level $\alpha/(d-1)$ and reject the null hypothesis if any of the individual tests rejects the null hypothesis. In what follows we simply denote this method as HSIC. We have mentioned in Section 1 that the Bonferroni correction is often conservative: this becomes particularly evident if this procedure is combined with a permutation-test-based HSIC. In that case it can be shown that the smallest possible p-value after the Bonferroni correction is given by $(d-1)/(B+1)$ and hence for $B = 100$ the test will not be able to reject the null hypothesis at a level of 5% if $d > 6$.

5.2. Causal inference

In causal discovery, one estimates the causal structure from an observed joint distribution. Here, we consider additive noise models (Peters *et al.*, 2014) with additive non-linear functions and Gaussian noise (Bühlmann *et al.*, 2014); these are special cases of structural equation models (Pearl, 2009). Assume that the distribution $\mathbb{P}^{\mathbf{X}} = \mathbb{P}^{(X^1, \ldots, X^d)}$ is induced by d structural equations

$$X^j := \sum_{k \in \mathrm{PA}^j} f^{j,k}(X^k) + N^j, \qquad j \in \{1, \ldots, d\}, \tag{5.2}$$

with PA^j being the parents of j in the associated directed acyclic graph (DAG) $\mathcal{G}_0$. The noise variables $N^1, \ldots, N^d$ are normally distributed and are assumed to be jointly independent. An important question in causality is whether the causal structure, in this case $\mathcal{G}_0$, can be inferred from the observational distribution $\mathbb{P}^{\mathbf{X}}$. Whereas this is impossible for general structural equation models (e.g. Peters *et al.* (2014), proposition 9), the additive noise structure renders the graph identifiable, i.e., if $f^{j,k}$ are assumed to be non-linear, any other additive noise model (5.2) with a structure that is different from $\mathcal{G}_0$ cannot induce the distribution $\mathbb{P}^{\mathbf{X}}$ (see Peters *et al.* (2014), corollary 31, for the full result). In other words, using conditional means as functions in the structural equation model, the corresponding residual variables will not be jointly independent.

We therefore propose the following *DAG verification method* for structure learning using generalized additive model regression (Wood and Augustin, 2002).

Given observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and a candidate DAG $\mathcal{G}$:

(a) use generalized additive model regression (Wood and Augustin, 2002) to regress each node X^j on all its parents PA^j and denote the resulting vector of residuals by res^j;
(b) perform a d-variable joint independence test (e.g. dHSIC) to test whether $(\mathrm{res}^1, \ldots, \mathrm{res}^d)$ is jointly independent;

(c) if $(\text{res}^1, \ldots, \text{res}^d)$ is jointly independent, then the DAG $\mathcal{G}$ is not rejected.

We can furthermore estimate the correct DAG by performing the verification method for all possible DAGs with the correct number of nodes. In practice, we expect this method to accept also supergraphs of the correct graph $\mathcal{G}_0$, which can be overcome by a variable selection method. Since this work concentrates on the dependence structure among the residuals, we instead consider only fully connected DAGs in the experiments (Section 5.3.4). In practice, we do not want to iterate over all possible graphs. A more efficient method, which is based on a similar idea, is the regression with subsequent independence test algorithm that was described in Peters *et al.* (2014), section 4.1. Also the computationally efficient causal additive model method (Bühlmann *et al.*, 2014) could be equipped with a joint independence test as a model check.

One issue deserves further attention. (We thank one of the referees for pointing this out.) In the regression step (a), we obtain only an approximation of the correct function, which results in estimated and thus dependent residuals rather than the true noise values. We show that this does not affect the asymptotic ordering of $\widehat{\text{dHSIC}}$; see theorem E.2 in the on-line appendix E. If we are interested in asymptotically valid p-values, we can perform sample splitting; see proposition E.3 in the on-line appendix E.

The DAG verification method described above can also be used to construct a statistical test for a more general causal hypothesis. For example, the causal hypothesis 'X is a causal ancestor of Y' can in principle be tested by applying the DAG verification method to all DAGs satisfying this ancestor relationship. One then reports the largest of the p-values appearing in step (b) of the DAG verification method. This test has, asymptotically, the correct size if there is indeed an underlying additive noise model that generated the data (again, using sample splitting, for example). Under a (minor) model misspecifcation, i.e., if the additive noise assumption does not hold, we might still find p-values that are much larger for the correct causal statement than for the reversed statement, e.g. 'Y is a causal ancestor of X' (see Peters *et al.* (2011)).

5.3. Results

We structure the experimental results into five parts: level analysis, power analysis, run time analysis and causal inference on simulated and a real data set.

5.3.1. *Level analysis*

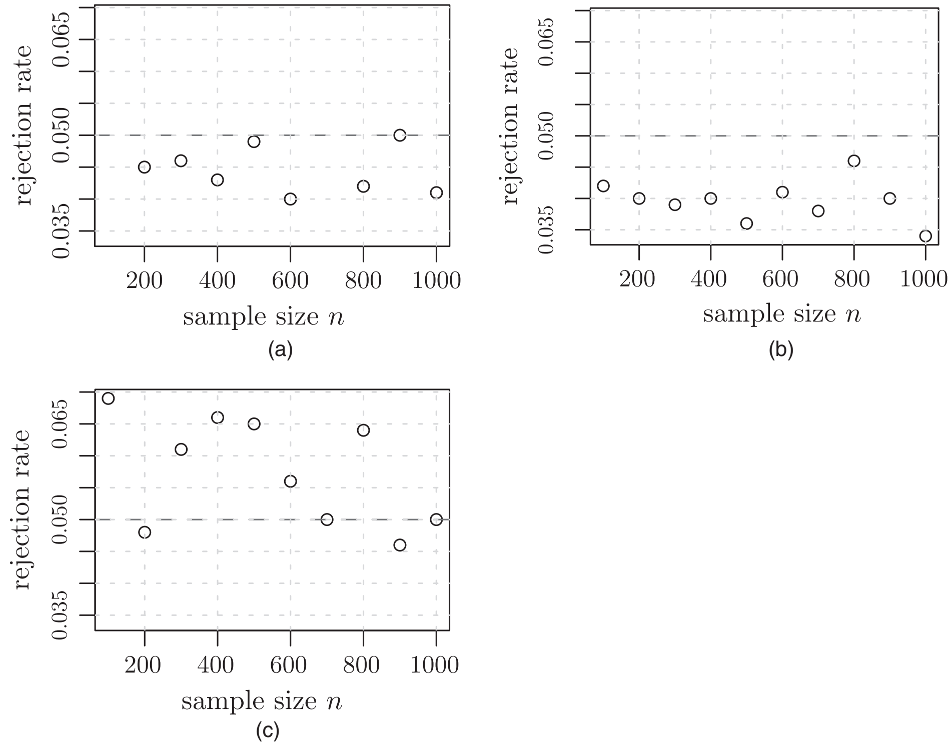
We consider an example with fixed $\mathbb{P}^{\mathbf{X}} \in H_0$ (simulation 1) and simulate $m = 1000$ independent realizations of $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim^{\text{IID}} \mathbb{P}^{\mathbf{X}}$ for various sample sizes n and check how often each of the three hypothesis tests reject the null hypothesis.

Simulation 1 (testing level—three continuous variables). Consider $X^1, X^2, X^3 \sim^{\text{IID}} \mathcal{N}(0, 1)$; then for $\mathbf{X} = (X^1, X^2, X^3)$ it holds that

$$\mathbb{P}^{\mathbf{X}} = \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2} \otimes \mathbb{P}^{X^3} \in H_0,$$

where H_0 is the null hypothesis defined in expression (3.1). Set $\alpha = 0.05$, $B = 25$ and $n \in \{100, 200, \ldots, 1000\}$. The rejection rates for the corresponding three hypothesis tests (permutation, bootstrap and gamma approximation) based on $m = 1000$ repetitions are plotted in Fig. 1.

A further simulation using discrete variables is given in the on-line appendix F.1. In both simulations we obtain similar results. We collect the most important observations.



(a) (b) (c)

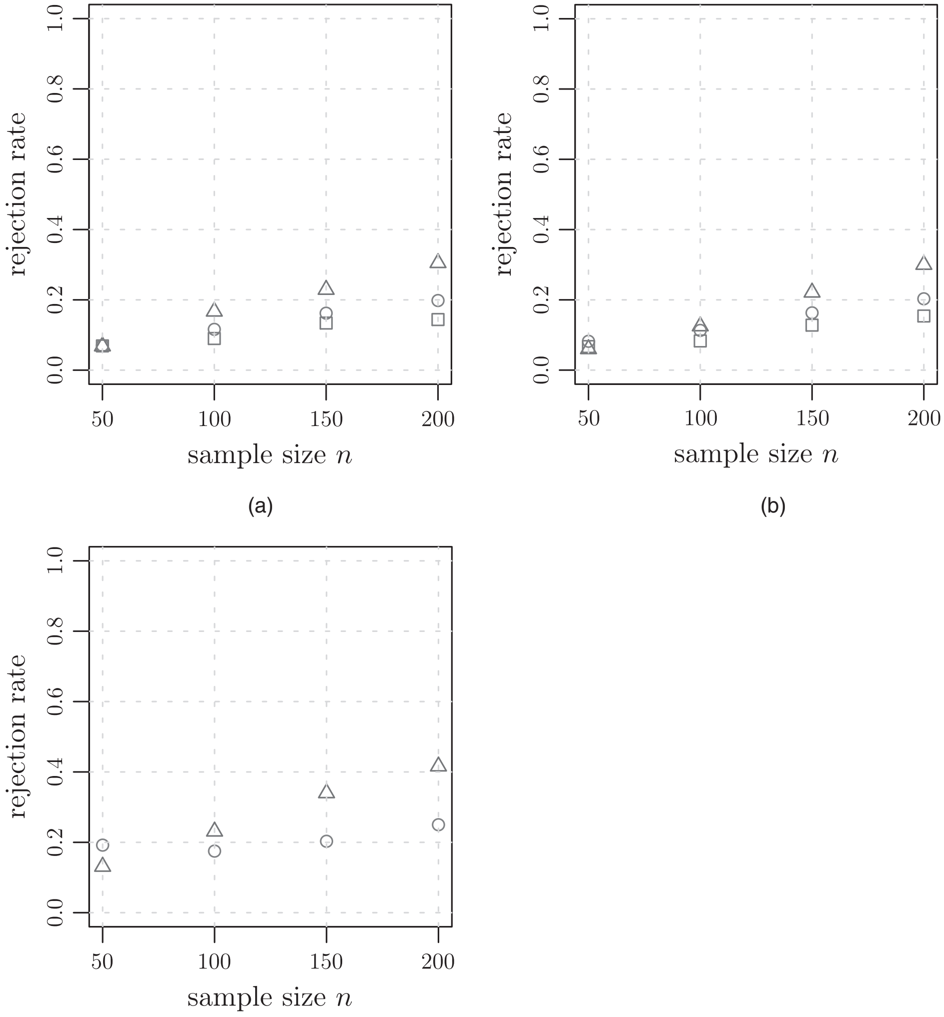
Fig. 1. Simulation 1 (testing level—three continuous variables)—rejection rates, based on $m = 1000$ repetitions, for each of the three hypothesis tests based on dHSIC; (the test has valid level if the rejection rate does not lie far above the dotted line at 0.05): (a) permutation; (b) bootstrap; (c) gamma

(a) The permutation test achieves level α. This corresponds to what has been proved in the previous section. As mentioned above, this result is rather surprising as it does not depend on the choice of B, which in simulation 1 is very small ($B = 25$).
(b) The bootstrap test achieves level α in most cases, even though we proved only that it has pointwise asymptotic level. This is due to the conservative choice of the p-value in the Monte Carlo approximation of the bootstrap test.
(c) The gamma-approximation-based test, at least in these two examples, has level close to α but often slightly exceeds the required level. For larger values of d the gamma approximation seems to break down. For instance, if we perform simulation 1 with 10 variables instead of three the rejection rate for a sample size of $n = 100$ is 0.40 and even for $n = 200$ it is still 0.21. The bootstrap test in contrast is not affected in this way (in the same setting we obtain 0.03 for $n = 100$ and 0.04 for $n = 200$).

5.3.2. Power analysis

Assessing the power of a test requires us to choose an alternative. In this section, we consider several generative models on $X^1, \ldots, X^d$ inducing different types of dependences and assess how well our method can detect that $\mathbb{P}^{\mathbf{X}} \neq \mathbb{P}^{X^1} \otimes \ldots \otimes \mathbb{P}^{X^d}$.

We begin by showing two examples: one favouring dHSIC (simulation 3); another favouring the multiple-testing approach using HSIC $d - 1$ times (simulation 2). In both simulations we use the BMR-C test with $C = n$ as reference. Using a BMR-C test with $C = 1000$ (which is not



(a)

(b)

(c)

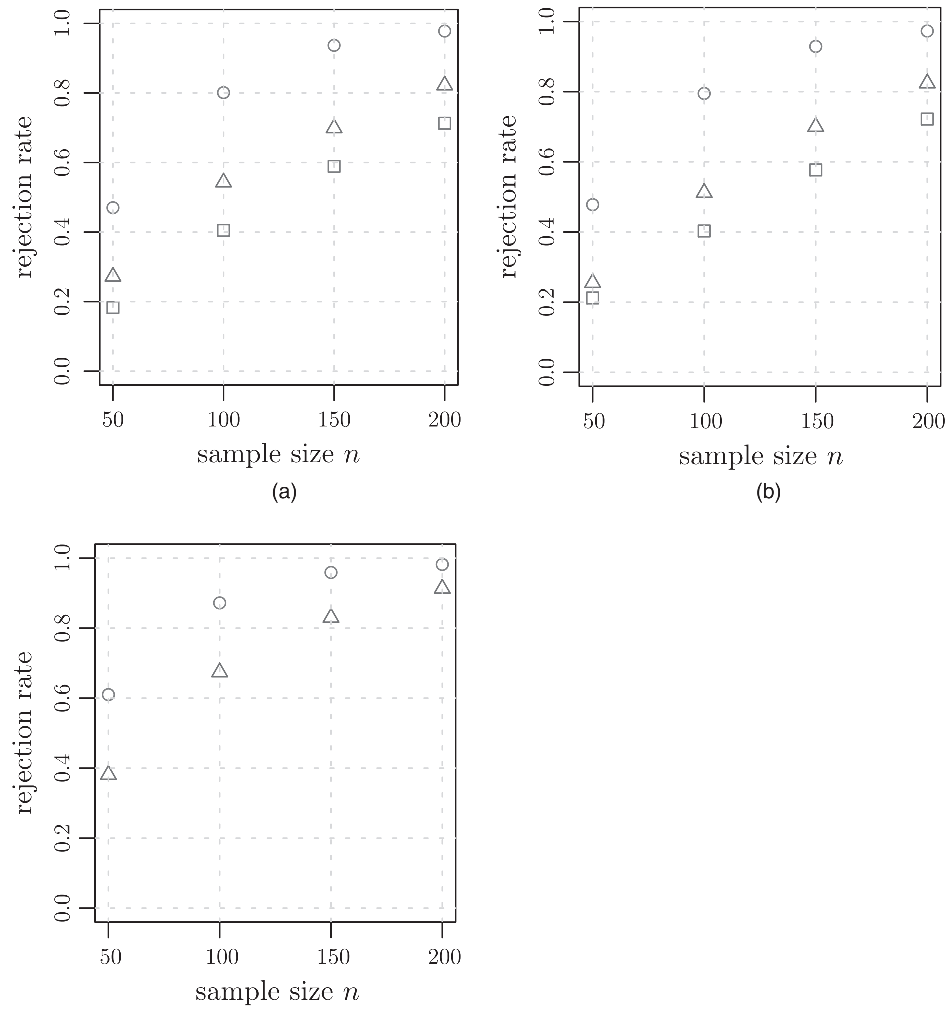
Fig. 2. Simulation 2 (comparing power—single edge)—rejection rates, based on $m = 1000$ repetitions, for each of the three hypothesis tests (the example (in particular the chosen order of variables) is constructed to favour the pairwise testing approach (HSIC); nevertheless, it performs only slightly better than dHSIC) (○, dHSIC; △, HSIC; □, BMR-n): (a) permutation; (b) bootstrap; (c) gamma

shown here) brings only marginal improvements which are not sufficient to beat HSIC in either simulation.

Simulation 2 (comparing power—single edge). For an additive noise model over random variables $X^1, \ldots, X^d$,

$$X^j := \sum_{k \in \mathrm{PA}^j} f^{j,k}(X^k) + N^j, \qquad j \in \{1, \ldots, d\},$$

with corresponding DAG $\mathcal{G}$, we sample data in the following way. The noise variables are Gaussian with a standard deviation sampled uniformly between $\sqrt{2}$ and 2. Nodes without parents follow a Gaussian distribution with standard deviation sampled uniformly between



(a)

(b)

(c)

Fig. 3. Simulation 3 (comparing power—full DAG)—rejection rates, based on $m = 1000$ repetitions, for each of the three hypothesis tests (as expected, dHSIC outperforms the competing method HSIC that is based on pairwise independence tests) (○, dHSIC; △, HSIC; □, BMR-n): (a) permutation; (b) bootstrap; (c) gamma

$5\sqrt{2}$ and 5×2. The functions $f^{j,k}$ are sampled from a Gaussian process with Gaussian kernel and bandwidth 1. Here we choose $d = 4$, let $\mathcal{G}$ be the graph that contains $1 \rightarrow 2$ as a single edge and use $m = 1000$ repetitions to compute rejection rates (Fig. 2). We expect this setting to favour the multiple-testing approach: because of the order of the variables, it tests X^1 against X^2.

Simulation 3 (comparing power—full DAG). We simulate the data as described in simulation 2 but this time using a (randomly chosen) full DAG $\mathcal{G}$ over $d = 4$ variables, i.e. every pair of two nodes is connected (Fig. 3). We expect that this setting favours dHSIC. Additionally, we fixed $n = 100$, varied d and used $m = 1000$ repetitions (Fig. 4).

We have restricted ourselves to comparing dHSIC with methods that are also capable of capturing all types of possible dependences. One should, however, keep in mind that although

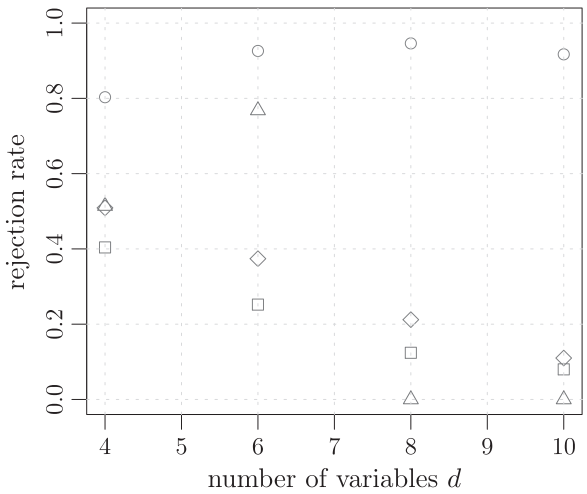


Fig. 4. Simulation 3 (comparing power—full DAG)—rejection rates, based on bootstrap ($B=100$; $n=100$) (BMR-C suffers from the curse of dimensionality and the pairwise HSIC approach cannot reject H_0 for $d>6$): ○, dHSIC; △, HSIC; □, BMR-n; ◇, BMR-1000

dHSIC is in general capable of capturing any type of dependence, it might not be the best method when additional information about the dependence structure is available. For example, consider a distribution that is Markov and faithful with respect to a known graphical chain $X^1 \to X^2 \to \ldots \to X^d$. A user might model the dependences by additive noise models but is not sure whether this is the correct model class. It might be useful (in terms of power) to use pairwise independence tests of the residuals instead of a joint independence test. (We thank one of the referees for pointing out the example of known orderings.) Another example is the three-variable interaction test by Sejdinovic *et al.* (2013), which has increased power for a specific type of dependence between three variables. The price to pay is that it is no longer possible to detect all potential dependences.

Finally, we analyse the influence of the choice of kernel on the empirical power (simulation 4). In this paper, we have mainly used the Gaussian kernel with median heuristic bandwidth. As mentioned in Section 4.4 this choice is not necessarily optimal. Using the Taylor expansion of the Gaussian kernel we obtain for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ that

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = 1 - \frac{1}{2\sigma^2} \sum_{j=1}^{d} (x^j - y^j)^2 + \frac{1}{4\sigma^4} \sum_{j,k=1}^{d} (x^j - y^j)^2 (x^k - y^k)^2 + \mathcal{O}(\sigma^{-6}),$$

as $\sigma \to \infty$. Therefore, it can be shown by using either the representation in definition 3 or that in proposition 2 that for large σ dHSIC using the Gaussian kernel is approximately given by dHSIC using the kernel

$$\tilde{\mathbf{k}}(\mathbf{x},\mathbf{y}) := \frac{1}{4\sigma^4} \sum_{j,k=1}^{d} (x^j - y^j)^2 (x^k - y^k)^2.$$

Such a kernel can, however, only detect pairwise dependence structures, and since the importance of this term becomes more prominent as the size of the bandwidth increases we expect the power of our dHSIC test to decrease when analysing dependences that have an additional dependence structure beyond a pairwise dependence. The following simulation illustrates this empirically on the basis of three dependences: a pairwise dependence, a more complex dependence due

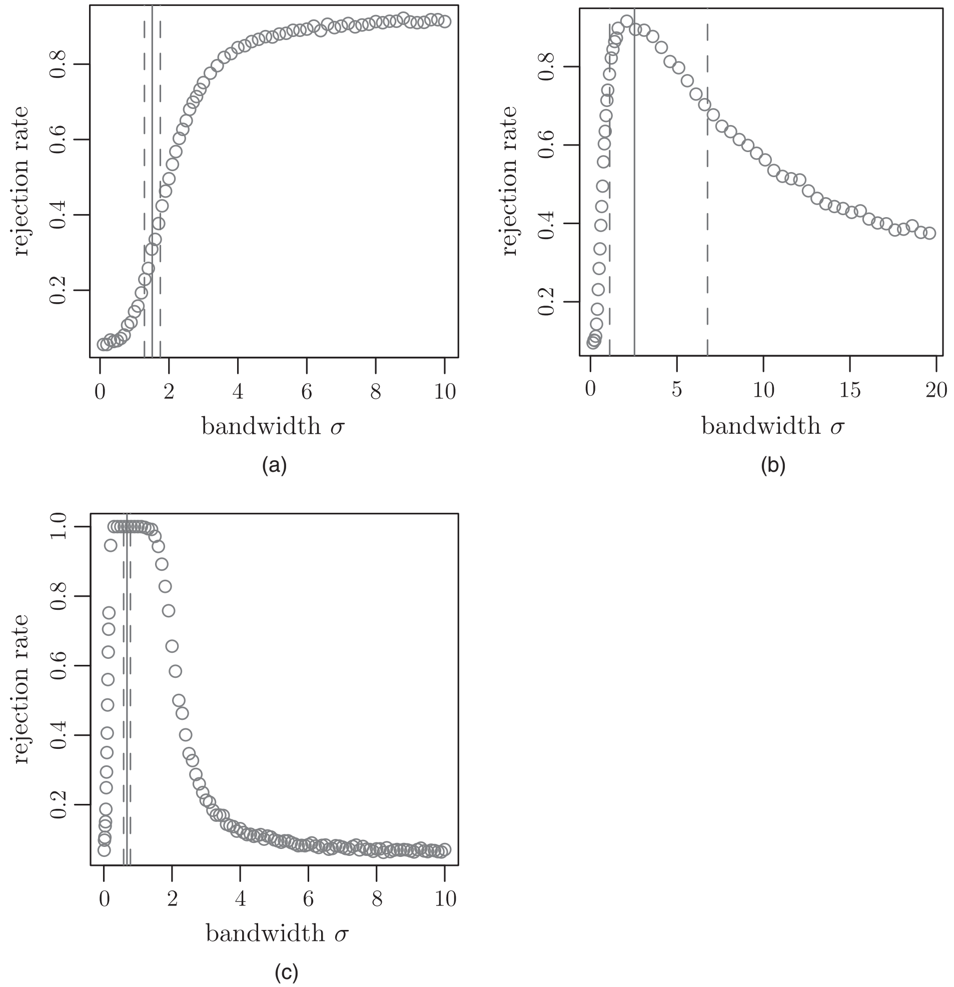


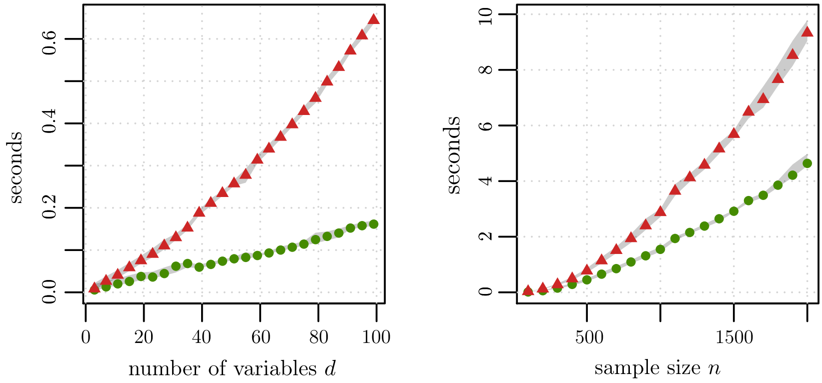
Fig. 5. Simulation 4 (comparing power—bandwidth)—rejection rates with $n=100$ for various bandwidths σ in the Gaussian kernel, based on $m=1000$ repetitions, of the permutation test ($B=100$) for data containing only a pairwise dependence, for data from a random non-linear Gaussian structural equation model and for dependent but pairwise independent data (the rejection rates resulting from the median heuristic are 0.30, 0.82 and 1 respectively) (¦, 95% confidence intervals for the bandwidth selected by using the median heuristic): (a) pairwise dependence; (b) mixed dependence; (c) pairwise independence

to a random non-linear Gaussian structural equation model and a dependence on three variables which is pairwise independent (Fig. 5). A further simulation analysing the differences in empirical power between sparse and dense alternatives is given in the on-line appendix F.2.

Simulation 4 (comparing power—bandwidth). We consider three dependences and analyse the behaviour of the empirical power of the dHSIC permutation test ($B=100$) based on different bandwidths for the Gaussian kernel. The first is generated by the linear Gaussian structural equation model

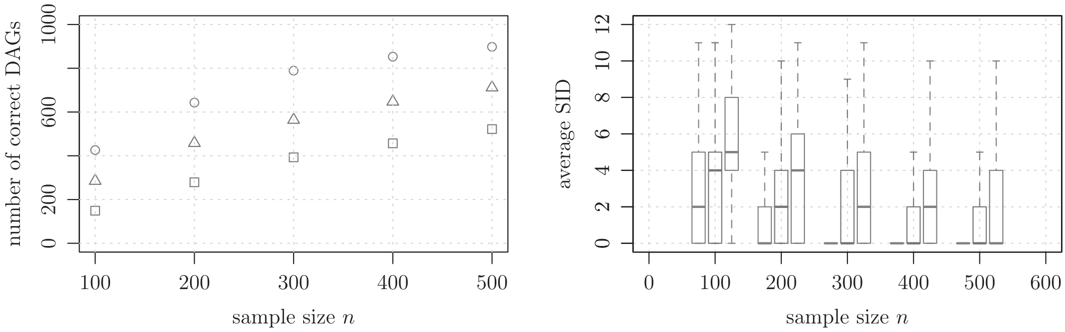
$$X^j = H + \varepsilon_j, \qquad H, \varepsilon_1, \ldots, \varepsilon_4 \sim^{\mathrm{IID}} \mathcal{N}(0,4),$$

and hence the only dependence is due to the confounder H. For the second dependence we use



(a) (b)

Fig. 6. Run time analysis: (a) varying number of variables and fixed sample size ($n=100$) and (b) varying sample size and fixed number of variables ($d=10$): ●, dHSIC; ▲, HSIC; ■, 95% error region



(a) (b)

Fig. 7. Causal inference ($m=1000$ repetitions): (a) how often the methods estimate the correct DAG (○, dHSIC; △, HSIC; □, BMR-n); (b) average structural intervention distance SID (small is good) between the correct and estimated DAG (Peters and Bühlmann, 2015) (from left to right, dHSIC, HSIC and BMR-n)

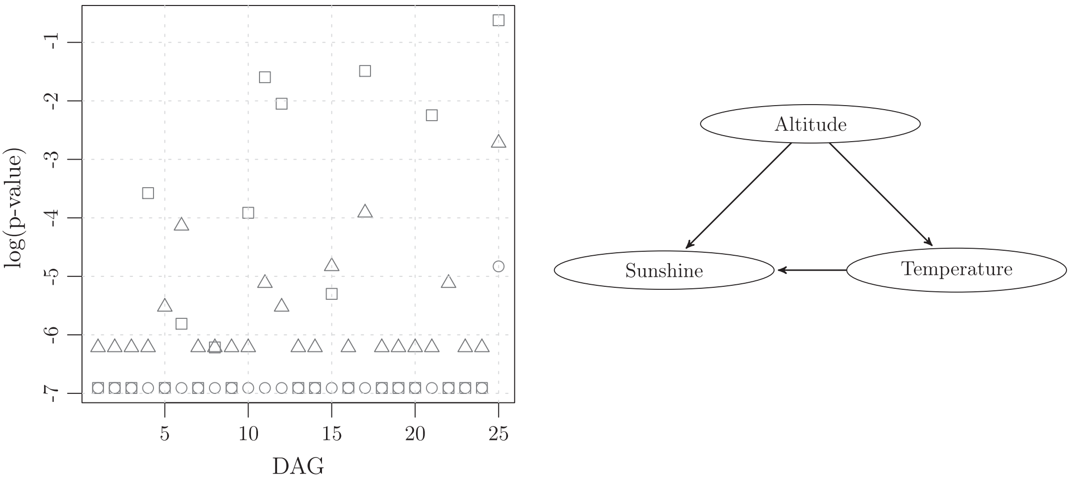
the same as in simulation 3, which has a more evolved dependence structure due to potential chains of ancestors. The third dependence has probability density

$$f(x^1,x^2,x^3)=\begin{cases} 2\varphi(x^1)\varphi(x^2)\varphi(x^3) & \text{if } x^1,x^2,x^3\geqslant 0, \text{ or } \exists ! j\in\{1,2,3\}: x^j\geqslant 0,\\ 0 & \text{otherwise,}\end{cases}$$

where φ is the standard normal density. The resulting distribution is, in particular, pairwise independent. For all examples we use a sample size of $n=100$ and $m=1000$ repetitions. The resulting plots are given in Fig. 5.

5.3.3. Run time analysis

The computational complexity for the dHSIC test statistic is $\mathcal{O}(dn^2)$ as can be seen from the considerations in Section 4.1. The multiple-testing approach for HSIC computes HSIC $d-1$ times, which appears to result in the same computational complexity. But since the dimension of the input variables for the HSIC tests generally depends on d, as well (at least in common



(a) (b)

Fig. 8. Real world data example: (a) p-values (on a log-scale), for each DAG over three nodes, from the DAG verification method (○, dHSIC; △, HSIC; □, BMR-1000) (even for small p-value thresholds, dHSIC can reject all incorrect models, whereas the competing HSIC method cannot); (b) graphical representation of DAG 25

settings such as for the Gaussian kernel), the overall complexity is $\mathcal{O}(d^2n^2)$. We numerically test these computational complexities by two simulations. In the first simulation we fix n and vary d; in the second simulation we fix d and vary n. The results are presented in Fig. 6. It might be possible to reduce computational complexity by using linear time approximation methods as described by Zhang *et al.* (2017) for the pairwise HSIC. (We thank one of the referees for pointing this out.)

5.3.4. *Causal inference (simulated data)*

We now apply both tests to the DAG verification method that was described in Section 5.2. As in simulation 2, we simulate data from an additive noise model. Here, we randomly choose a fully connected DAG $\mathcal{G}$ over $d=4$ nodes and choose Gaussian-distributed noise variables with standard deviation sampled uniformly between $1/5$ and $\sqrt{2}/5$ instead of $\sqrt{2}$ and 2. We then report how often (out of $m=1000$) the largest p-value leads to the correct DAG. Because of its computational advantage, we use the tests based on the gamma approximation for dHSIC and the pairwise HSIC, which work reasonably well for four nodes (strictly speaking, we use only the relative size of the p-values). Most of the time was spent computing the results for BMR-n as we were forced to use a bootstrap test ($B=100$) since no approximation is available for this test. The proposed dHSIC recovers the correct DAG in more cases than the pairwise approach and in even more cases than BMR-n (Fig. 7).

5.3.5. *Causal inference (real data example)*

We now apply the DAG verification method (see Section 5.2) to real world data. Given 349 measurements of the variables altitude, temperature and sunshine (the data set is taken from Mooij *et al.* (2016), pair0001.txt and pair0004.txt), we try to determine the correct causal structure out of 25 possible DAGs. We use permutation-based versions (with $B=1000$) of the dHSIC-test, the multiple-testing approach for HSIC and the BMR-1000 test and apply them to every possible DAG and compare the resulting p-values. The result is shown in Fig. 8(a).

Fig. 8(b) shows DAG 25: the DAG with the largest p-value. On the basis of physical background knowledge, we expect altitude to affect both sunshine and temperature. The effect of temperature on sunshine could be due to intermediate latent variables such as clouds or fog. Fig. 8(a) illustrates that the dHSIC-based test can reject all incorrect models, even for very low p-value thresholds. This is different for the competing HSIC and BMR-1000 methods. For example, DAG 12 has a p-value of about 0.01 but contains an edge from sunshine to altitude, which is clearly the wrong causal direction.

6. Summarizing remarks

We analyse a measure of joint dependence between d variables, called the d-variable HSIC. We propose an estimator of dHSIC based on a computationally attractive V-statistic and derive its asymptotic distribution. This enables us to construct three different hypothesis tests: a permutation test (definition 5), a bootstrap test (definition 6) and a test based on a gamma approximation (definition 7).

We prove several properties for these tests. First and foremost we establish that the bootstrap test achieves pointwise asymptotic level (theorem 4) and that it is consistent for detecting any fixed alternative with asymptotic power equal to 1 (theorem 5). For the permutation test, we show that it achieves a valid level (proposition 3) and, in particular, this property carries over to the Monte Carlo approximated version of the permutation test. Regarding the gamma-approximation-based test, we derive asymptotic expansions of the mean and variance of the dHSIC-estimator (proposition 4 and proposition 5) which serve as the main basis in the construction of the approximation. Although this test has no guarantees on level and consistency, it is computationally very fast and was found to perform well in numerical experiments.

Various simulations illustrate the advantages of dHSIC over a pairwise approach with HSIC and a traditional test that we call BMR-C. Notably, dHSIC is computationally less expensive than HSIC and also BMR-C if C grows larger than n. Moreover, when the dimension d is large the pairwise HSIC approach with Monte Carlo approximation (for fixed B) cannot reject the null hypothesis and BMR-C seems to suffer substantially from the curse of dimensionality. We also outline applications for model selection in causal inference which are based on joint independence testing of error terms in structural equation models. In our numerical experiments on real and simulated data, dHSIC outperforms both the other methods.

Acknowledgements

The authors thank the Joint Editor, Associate Editor and the reviewers for constructive comments, and Christoph Lampert, Arthur Gretton, Gilles Blanchard and Aaditya Ramdas for helpful discussions. JP received support from the Max Planck ETH Center for Learning Systems.

References

Bakirov, N. K., Rizzo, M. L. and Székely, G. J. (2006) A multivariate nonparametric test of independence. *J. Multiv. Anal.*, **97**, 1742–1756.

Beran, R. and Millar, P. W. (1987) Stochastic estimation and testing. *Ann. Statist.*, **15**, 1131–1154.

Bergsma, W. and Dassios, A. (2014) A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, **20**, 1006–1028.

Bühlmann, P., Peters, J. and Ernest, J. (2014) CAM: causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.*, **42**, 2526–2556.

Chen, A. and Bickel, P. J. (2006) Efficient independent component analysis. *Ann. Statist.*, **34**, 2825–2855.

Chwialkowski, K. P., Sejdinovic, D. and Gretton, A. (2014) A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*, vol. 27 (eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger), pp. 3608–3616. New York: Curran Associates.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Feuerverger, A. (1993) A consistent test for bivariate dependence. *Int. Statist. Rev.*, **61**, 419–433.

Fukumizu, K., Gretton, A., Sun, X. and Schölkopf, B. (2007) Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, vol. 20 (eds J. C. Platt, D. Koller, Y. Singer and S. T. Roweis), pp. 489–496. New York: Curran Associates.

Gaißer, S., Ruppert, M. and Schmid, F. (2010) A multivariate version of Hoeffding's phi-square. *J. Multiv. Anal.*, **101**, 2571–2586.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B. and Smola, A. (2012) A kernel two-sample test. *J. Mach. Learn. Res.*, **13**, 723–773.

Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005) Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory* (eds S. Jain, H. U. Simon and E. Tomita), pp. 63–77. Berlin: Springer.

Gretton, A., Fukumizu, K., Harchaoui, Z. and Sriperumbudur, B. K. (2009) A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, vol. 22 (eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta), pp. 673–681. New York: Curran Associates.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B. and Smola, A. J. (2007) A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, vol. 20 (eds J. C. Platt, D. Koller, Y. Singer and S. T. Roweis), pp. 585–592. New York: Curran Associates.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K. and Sriperumbudur, B. K. (2012) Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, vol. 25 (eds F. Pereira, C. J. C. Burges, L. Bottou and K. O. Weinberger), pp. 1205–1213. New York: Curran Associates.

Kankainen, A. (1995) *Consistent Testing of Total Independence based on the Empirical Characteristic Function*, vol. 29. Jyväskylä: University of Jyväskylä.

Lehmann, E. L. and Romano, J. P. (2005) *Testing Statistical Hypotheses*. New York: Springer.

Leung, D. and Drton, M. (2016) Testing independence in high dimensions with sums of squares of rank correlations. *Preprint arXiv1501.01732*. University of Washington, Seattle.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012) High-dimensional semi-parametric gaussian copula graphical models. *Ann. Statist.*, **40**, 2293–2326.

Matteson, D. S. and Tsay, R. S. (2016) Independent component analysis via distance covariance. *J. Am. Statist. Ass.*, to be published.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J. and Schölkopf, B. (2016) Distinguishing cause from effect using observational data: methods and benchmarks. *J. Mach. Learn. Res.*, **17**, 1–102.

Nandy, P., Weihs, L. and Drton, M. (2016) Large-sample theory for the Bergsma-Dassios sign covariance. *Electron. J. Statist.*, **10**, 2287–2311.

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*, 2nd edn. New York: Cambridge University Press.

Peters, J. and Bühlmann, P. (2015) Structural intervention distance (SID) for evaluating causal graphs. *Neurl Computn*, **27**, 771–799.

Peters, J., Janzing, D. and Schölkopf, B. (2011) Causal inference on discrete data using additive noise models. *IEEE Trans. Pattn Anal. Mach. Intell.*, **33**, 2436–2450.

Peters, J., Mooij, J. M., Janzing, D. and Schölkopf, B. (2014) Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, **15**, 2009–2053.

Pfister, N. (2016) Joint independence testing. *Master's Thesis*. Eidgenössiche Technische Hochschule Zürich, Zürich.

Romano, J. P. (1986) A bootstrap revival of some nonparametric tests. *Technical Report 254*. Department of Statistics, Stanford University, Stanford.

Romano, J. P. (1988) A bootstrap revival of some nonparametric distance tests. *J. Am. Statist. Ass.*, **83**, 698–708.

Romano, J. P. (1989) Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.*, **17**, 141–159.

Satterthwaite, F. E. (1946) An approximate distribution of estimates of variance components. *Biometr. Bull.*, **2**, 110–114.

Sejdinovic, D., Gretton, A. and Bergsma, W. (2013) A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems*, vol. 26 (eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger), pp. 1124–1132.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Chichester: Wiley.

Smola, A., Gretton, A., Song, L. and Schölkopf, B. (2007) A Hilbert space embedding for distributions. In *Algorithmic Learning Theory* (eds M. Hutter, R. A. Servedio and E. Takimoto), pp. 13–31. New York: Springer.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G. and Schölkopf, B. (2008) Injective Hilbert space embeddings of probability measures. In *Proc. Conf. Learning Theory* (eds R. Servedio and T. Zhang). Madison: Omnipress.

Székely, G. J. and Rizzo, M. L. (2009) Brownian distance covariance. *Ann. Appl. Statist.*, **3**, 1236–1265.

Székely, G. J. and Rizzo, M. L. (2014) Partial distance correlation with methods for dissimilarities. *Ann. Statist.*, **42**, 2382–2412.

Unser, M. and Tafti, P. D. (2014) *An Introduction to Sparse Stochastic Processes*. Cambridge: Cambridge University Press.

Wegkamp, M. and Zhao, Y. (2016) Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, **22**, 1184–1226.

Wood, S. N. and Augustin, N. H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol. Modllng*, **157**, 157–177.

Xue, L. and Zou, H. (2012) Regularized rank-based estimation of high-dimensional non-paranormal graphical models. *Ann. Statist.*, **40**, 2541–2571.

Zhang, Q., Filippi, S., Gretton, A. and Sejdinovic, D. (2017) Large-scale kernel methods for independence testing. *Statist. Comput.*, **27**, 1–18.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Kernel-based tests for joint independence’.