A regression perspective on generalized distance covariance and the Hilbert-Schmidt Independence Criterion

Edelmann, D.; Goeman, J.

Statistical Science
2022, Vol. 37, No. 4, 562–579
https://doi.org/10.1214/21-STS841

A Regression Perspective on Generalized Distance Covariance and the Hilbert–Schmidt Independence Criterion

Dominic Edelmann and Jelle Goeman

Abstract. In a seminal paper, Sejdinovic et al. (*Ann. Statist.* **41** (2013) 2263–2291) showed the equivalence of the Hilbert–Schmidt Independence Criterion (HSIC) and a generalization of distance covariance. In this paper, the two notions of dependence are unified with a third prominent concept for independence testing, the "global test" introduced in (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** (2006) 477–493). The new viewpoint provides novel insights into all three test traditions, as well as a unified overall view of the way all three tests contrast with classical association tests. As our main result, a regression perspective on HSIC and generalized distance covariance is obtained, allowing such tests to be used with nuisance covariates or for survival data. Several more examples of cross-fertilization of the three traditions are provided, involving theoretical results and novel methodology. To illustrate the difference between classical statistical tests and the unified HSIC/distance covariance/global tests we investigate the case of association between two categorical variables in depth.

Key words and phrases: Distance covariance, distance correlation, Hilbert–Schmidt Independence Criterion, global test, equivalence, locally most powerful.

1. INTRODUCTION

During the last three decades, we have witnessed major developments in statistical methods for high-dimensional datasets. While the most prominent developments may concern statistical modeling and variable selection [13, 65, 71], there have also been important contributions to hypothesis testing. Three concepts for independence testing that enjoy great popularity are the distance covariance [59, 64], the Hilbert–Schmidt Independence Criterion (HSIC) [21, 23] and a family of locally most powerful score tests known as "global tests" [16, 18].

Distance covariance, proposed by Gabór Székely et al. [59, 64], is a concept for testing and quantifying general dependencies. Its most salient property is that—in contrast to classical covariance—its population version equals zero if and only if the random variables under consideration are independent, implying that distance covariance is able to detect arbitrary association between datasets. It features a strikingly simple sample version, making it simple to use in applications [32, 33, 39, 44]. There is already a large literature on the theory of distance covariance [7, 8, 62, 66]; see [9] for an overview. Particularly important is the extension to generalized distance covariance [30, 38], that enables independence testing in general metric spaces.

Around the same time that distance covariance was first introduced, a different notion of independence—the Hilbert–Schmidt Independence Criterion—gained popularity in the machine learning community [21, 23]. At its inception, HSIC was already formulated for general measurable spaces, as it is based on kernel functions rather than distances. HSIC and distance covariance have some obvious similarities, which generated early conjectures [22] of some conceptual equivalence. The exact equivalence of HSIC with an extension of the generalized distance covariance proposed in [38] was eventually proven in [51], opening up a vivid exchange between the two communities [4, 28, 53].

We will link these two testing traditions to a third tradition: the family of global tests [16–18, 24], developed

Dominic Edelmann is Professor, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany (e-mail: dominic.edelmann@dkfz.de). Jelle Goeman is Postdoctoral Researcher, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, Netherlands (e-mail: J.J.Goeman@lumc.nl).

for high-dimensional datasets in biomedical applications from a tradition of goodness-of-fit tests [35, 55]. This type of test includes popular tests for genetic data such as the Sequence Kernel Association test (SKAT) [68] and other genotype-phenotype tests [41]. Global tests are closely connected to the classical RV coefficient [47]. In contrast to distance covariance or HSIC, global tests are formulated as locally most powerful tests on specific statistical models, restricting them to the detection of prespecified types of association. However, global tests differ markedly from the classical tests for the models they are defined in (such as for example the F-test in linear regression): unlike classical tests, global tests are not invariant under affine transformations of the data. While this may seem like a severe drawback on first sight, it has been shown that it is necessary to give up affine invariance in order to obtain a test with nontrivial power in high-dimensional data [18].

Interestingly, also distance covariance and HSIC are noninvariant under affine transformations, and both can be applied for testing independence in high-dimensional data. Moreover, the linear global test statistic shows a clear similarity to the sample versions of distance covariance and HSIC, suggesting possible relations between the three measures.

In this article, we investigate analogies between generalized distance covariance, HSIC and global tests. We unify the concepts of generalized distance covariance, HSIC and global tests to derive interesting theoretical insights as well as novel statistical methodology. A large part of the paper is devoted to practical implications arising from the cross-fertilization between the three different traditions. In particular, global tests may profit from reformulating them as HSIC or generalized distance covariance test to enable detection of nonlinear associations; conversely, the formulation of generalized distance covariance or HSIC as global tests gives these tests a foundation in an underlying model. This may provide a better understanding against which alternatives the tests possess good power, and allows well-motivated extensions of these tests. We provide examples of extensions to models with nuisance covariates, goodness-of-fit testing, survival analysis, and heavy-tailed data. Finally, the unification of the three theories allows the transfer of important results and concepts, such as ways to approximate the distribution of the test statistic, or the notion of distance correlation as an effect size measure.

The first main result of the paper is in Section 5, in which we show how HSIC and distance covariance tests may be seen as global tests, and global tests as HSIC or distance covariance tests. In particular, HSIC and distance covariance sample measures can be expressed as global test statistics involving the induced feature maps of their corresponding kernels or premetrics; the linear global test is simply a special case of generalized distance covariance and HSIC using squared Euclidean distance or the linear kernels, respectively. We also formulate the equivalence of generalized distance covariance and HSIC in a slightly more general way than given in [51], strengthening this connection. Section 5 further explores the nature of the equivalence by showing how special cases of generalized distance covariance and HSIC can be formulated as locally most powerful tests in Gaussian regression models. This provides insights into the statistical properties of these tests. We also investigate connections to kernel principal component analysis (kPCA) and kernel partial least squares (kPLS). Practical implications of the equivalence of the three theories, in Section 6, include a generalization of the concept of the distance correlation coefficient, a celebrated concept for quantifying dependence between random vectors [59, 64]. Section 7 shows the applicability of the results of the paper, demonstrating how the model-based formulation of generalized distance covariance/HSIC can be used to develop novel tests for complex associations in various statistical models, such as in models involving nuisance covariates, in survival models and in heavy-tailed data. Finally in Section 8 we derive and investigate a test for association between two categorical variables that can be equivalently derived from all three traditions. This test contrasts markedly from the classical chi-squared test and provides insight into the common philosophy of HSIC, distance covariance and global tests.

All proofs for novel results are provided in the Supplementary Material [10].

2. PROBLEM DESCRIPTION

Consider two jointly distributed random variables X and Y defined on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, where we suppress the σ-algebras for simplicity. The joint distribution of X and Y will be denoted by P_{XY}, while P_X and P_Y denote the marginal distributions. For $n \in \mathbb{N}$, we will further denote i.i.d. samples of (X, Y) by $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$. In this work, we consider various test statistics for testing the null hypothesis of independence between X and Y, that is,

$$H_0 : P_{XY} = P_X\, P_Y$$

given samples $\boldsymbol{X}$ and $\boldsymbol{Y}$. To emphasize their purpose, the test statistics will carry a hat, whereas their corresponding sample measure/almost sure limit is denoted by the respective expression without hat (i.e., $\widehat{T}$ denotes a test statistic, whereas T denotes its respective population measure).

All considered test statistics $\widehat{T}$ have in common that their respective population measure T equals zero under the null hypothesis of independence, that is,

$$P_{XY} = P_X\, P_Y \quad \Longrightarrow \quad T = 0.$$

For some of the test statistics (such as for example, standard distance covariance or certain instances of HSIC), the reverse implication

$$T = 0 \implies P_{XY} \neq P_X P_Y,$$

also holds, implying that T characterizes independence between X and Y.

The following notation will be used throughout the manuscript. For each $p \in \mathbb{N}$, $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^p$ and $\| \cdot \|$ the corresponding Euclidean norm. $\mathbf{1} = (1, \dots, 1)$ denotes a vector of ones, where the length of the vector will be clear from the context and I_p denotes the identity matrix in $\mathbb{R}^p$. For any matrix $\mathbf{M}$, $\mathbf{M}^t$ will refer to its transpose.

3. THE HILBERT–SCHMIDT INDEPENDENCE CRITERION

3.1 Definition

The *Hilbert–Schmidt Independence Criterion* (HSIC) is a statistic for independence testing established in [21, 23]. To define HSIC, consider a measurable space $\mathcal{Z}$, where we suppress the σ-algebra in the notation for reasons of simplicity. We call a function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a *kernel* if it is symmetric in its arguments and satisfies the property of positive definiteness, that is, for all $n \geq 1$, $z_1, \dots, z_n \in \mathcal{Z}$ and $a_1, \dots, a_n \in \mathbb{R}$, we have

$$\sum_{i,j=1}^{n} a_i a_j k(z_i, z_j) \geq 0.$$

Given kernels $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, for which $\mathbb{E}|k_\mathcal{X}(X, X')| + |k_\mathcal{Y}(Y, Y')| < \infty$ the HSIC statistic is defined as [51]

$$\begin{aligned} &\mathrm{HSIC}_{k_\mathcal{X}, k_\mathcal{Y}}(X, Y) \\ &\quad = \mathbb{E}[k_\mathcal{X}(X, X')\{k_\mathcal{Y}(Y, Y') - k_\mathcal{Y}(Y, Y'') \\ &\qquad - k_\mathcal{Y}(Y', Y'') + k_\mathcal{Y}(Y'', Y''')\}], \end{aligned} \tag{1}$$

where (X, Y), (X', Y'), (X'', Y''), (X''', Y''') are i.i.d. copies of (X, Y).

It is easy to show that $\mathrm{HSIC}_{k_\mathcal{X}, k_\mathcal{Y}}(X, Y) \geq 0$ and that independence of X and Y induces $\mathrm{HSIC}_{k_\mathcal{X}, k_\mathcal{Y}}(X, Y) = 0$. However, the reverse holds true only under a stronger condition on the kernels: $\mathrm{HSIC}_{k_\mathcal{X}, k_\mathcal{Y}}(X, Y) = 0$ implies the independence of X and Y if and only if the kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ are characteristic. A kernel $k_\mathcal{X}$ is characteristic if

$$P_X = P_Z \quad \Leftrightarrow \quad \mathbb{E}[k_\mathcal{X}(x, X)] = E[k_\mathcal{X}(x, Z)]$$

for all $x \in \mathcal{X}$ and random variables X and Z on $\mathcal{X}$. We easily see that the linear kernel $k(x, x') = \langle x, x' \rangle$ is characteristic if and only if $\mathcal{X}$ contains no more than two elements. On the other hand, the Gaussian kernel, cf. Section 3.2, is characteristic on $\mathbb{R}^p$, $p \in \mathbb{N}$.

Remembering that we denote i.i.d. joint samples of (X, Y) by $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$, we define the (i, j)th element of the kernel matrices $\boldsymbol{K}^{\boldsymbol{X}}$ and $\boldsymbol{K}^{\boldsymbol{Y}}$ by $\boldsymbol{K}^{\boldsymbol{X}}_{i,j} = k_\mathcal{X}(X_i, X_j)$ and $\boldsymbol{K}^{\boldsymbol{Y}}_{i,j} = k_\mathcal{Y}(Y_i, Y_j)$, respectively, and denote the double-centered versions by

$$\begin{aligned} \tilde{\boldsymbol{K}}^{\boldsymbol{X}} &= (I - H)\boldsymbol{K}^{\boldsymbol{X}}(I - H), \\ \tilde{\boldsymbol{K}}^{\boldsymbol{Y}} &= (I - H)\boldsymbol{K}^{\boldsymbol{Y}}(I - H), \end{aligned}$$

where $H = \frac{1}{n}\mathbf{1}\mathbf{1}^t$. Then a consistent sample version for $\mathrm{HSIC}_{k_\mathcal{X}, k_\mathcal{Y}}(X, Y)$ is given by

$$\widehat{\mathrm{HSIC}}_{k_\mathcal{X}, k_\mathcal{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{\boldsymbol{K}}^{\boldsymbol{X}}_{i,j} \tilde{\boldsymbol{K}}^{\boldsymbol{Y}}_{i,j}. \tag{2}$$

Since both $\tilde{\boldsymbol{K}}^{\boldsymbol{X}}$ and $\tilde{\boldsymbol{K}}^{\boldsymbol{Y}}$ are symmetric, it is easily shown that an alternative representation for the right-hand side of equation (10) is $\frac{1}{n^2}\operatorname{tr}(\tilde{\boldsymbol{K}}^{\boldsymbol{X}} \tilde{\boldsymbol{K}}^{\boldsymbol{Y}})$. Moreover, since $(I - H)$ is idempotent it is sufficient to center only one of the two distance matrices, removing the tilde above one of the two matrices in equation (10).

3.2 Feature Maps

It is useful to decompose kernel functions into *feature maps*. A feature map for a kernel k on $\mathcal{Z}$ is a function $\boldsymbol{\Phi}^{(k)} = (\Phi_1^{(k)}, \dots, \Phi_d^{(k)}) : \mathcal{Z} \to \mathbb{R}^d$, such that $k(z, z') = \langle \boldsymbol{\Phi}^{(k)}(z), \boldsymbol{\Phi}^{(k)}(z') \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^d$; we here allow for $d = \infty$ (countable). There are many prominent kernels for which such a feature map can be derived, such as the Gaussian kernel, the discrete kernel or the polynomial kernel; conditions under which feature maps exist are provided by Mercer's theorem [40] and its numerous extensions [57, 58].

If feature maps for kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ exist, $\widehat{\mathrm{HSIC}}_{k_\mathcal{X}, k_\mathcal{Y}}(X, Y)$ can be written as a linear association test between the features,

$$\begin{aligned} &\frac{1}{n^2} \sum_{l=1}^{d_\mathcal{X}} \sum_{m=1}^{d_\mathcal{Y}} \sum_{i=1}^{n} [(\Phi_l^{(k_\mathcal{X})}(X_i) \\ &\quad - \overline{\Phi_l^{(k_\mathcal{X})}(\boldsymbol{X})})(\Phi_m^{(k_\mathcal{Y})}(Y_i) - \overline{\Phi_m^{(k_\mathcal{Y})}(\boldsymbol{Y})})]^2, \end{aligned} \tag{3}$$

where $\overline{\Phi_l^{(k_\mathcal{X})}(\boldsymbol{X})} = \frac{1}{n}\sum_{i=1}^{n} \Phi_l^{(k_\mathcal{X})}(X_i)$, and $\overline{\Phi_m^{(k_\mathcal{Y})}(\boldsymbol{Y})} = \frac{1}{n}\sum_{i=1}^{n} \Phi_m^{(k_\mathcal{Y})}(Y_i)$. Denoting the usual covariance by Cov, we directly obtain the following decomposition of the population measure:

$$\begin{aligned} &\mathrm{HSIC}_{k_\mathcal{X}, k_\mathcal{Y}}(X, Y) \\ &\quad = \sum_{l=1}^{d_\mathcal{X}} \sum_{m=1}^{d_\mathcal{Y}} \mathrm{Cov}^2(\Phi_l^{(k_\mathcal{X})}(X), \Phi_m^{(k_\mathcal{Y})}(Y)). \end{aligned} \tag{4}$$

In the following, we state some prominent kernel functions together with their corresponding feature maps, where we restrict ourselves on kernels on $\mathbb{R}^p$ and discrete sets.

The Gaussian kernel is arguably the most popular kernel used for HSIC. For $x, x' \in \mathbb{R}^p$, it is defined as

$$k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2),$$

where $\|\cdot\|$ denotes the Euclidean distance. Denoting $x = (x_1, \dots, x_p)$, a feature map for the Gaussian kernel is given by

$$\boldsymbol{\Phi}(x) = \left(\exp\left(\frac{-\|x\|^2}{2i\sigma^2}\right)\sqrt{\binom{i}{n_1, \dots, n_p}} \times \frac{x_1^{n_1} \cdots x_p^{n_p}}{(\sqrt{i!\sigma^2})^{1/i}}\right)_{i \in \{0,\dots,\infty\}, \sum n_j = i}.$$

For $p = 1$, this reduces to

$$\boldsymbol{\Phi}(x) = \left(\sqrt{\frac{1}{i!\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)x^i\right)_{i \in \{0,\dots,\infty\}}.$$

The discrete kernel function is useful for testing independence on unordered finite sets (i.e., categorical data). Notably, let $\mathcal{X} = \{x_1, \dots, x_m\}$ be a finite set. Then the discrete kernel $k : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ is defined as

$$k(x, x') = 1_{\{x = x'\}}.$$

The corresponding premetric is the discrete metric

$$\rho(x, x') = 1_{\{x \neq x'\}},$$

and a feature map is given by

$$\boldsymbol{\Phi}(x) = (1_{\{x = x_i\}})_{i \in \{1,\dots,m\}}.$$

Tests based on discrete kernels will be studied in more detail in Section 8.

The polynomial kernel for $x, x' \in \mathbb{R}^p$ is defined as

$$k(x, x') = (\langle x, x' \rangle + c)^b. \tag{5}$$

Using the binomial theorem, equation (5) reads

$$k(x, x') = \sum_{k=0}^{b} \binom{b}{i} \langle x, x' \rangle^i c^{b-i}.$$

A feature map can be obtained by expanding $\langle x, x' \rangle^i$. For example, for $b = 2$, $p = 2$ and $x = (x_1, x_2)$, a feature map is given by

$$\boldsymbol{\Phi}(x) = (c, \sqrt{c}x_1, \sqrt{c}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2).$$

The linear kernel arises as a special case of the polynomial kernel with $b = 1$, that is,

$$k(x, x') = \langle x, x' \rangle, \tag{6}$$

the feature map of the linear kernel is the identity function $\boldsymbol{\Phi}(x) = x$.

There are a large number of other kernels that have proven to be useful in applications such as spline kernels, ANOVA kernels and kernels tailored for applications in text analyses; for a (nonextensive) overview we refer to [26].

4. GENERALIZED DISTANCE COVARIANCE

4.1 Definition

We start this section by introducing the standard version of distance covariance as defined by Gábor Székely and his coauthors [59, 64] and extend this definition to general premetric spaces thereafter.

Distance covariance and distance correlation have been proposed by [59, 64] as alternatives to the classical covariance and Pearson correlation. Consider jointly distributed random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with finite first moments and denote by

$$f_{X,Y}(s, t) = \exp\{\sqrt{-1}(\langle X, s \rangle + \langle Y, t \rangle)\}$$

the joint characteristic function of (X, Y). The distance covariance is defined as a weighted L^2-norm of the difference between $f_{X,Y}(s, t)$ and the product of the corresponding marginal characteristic functions $f_X(s) = f_{X,Y}(s, 0)$ and $f_Y(t) = f_{X,Y}(0, t)$. More precisely, denoting by $|\cdot|_{\mathbb{C}}$ the modulus in $\mathbb{C}$, and by $\|s\|$ the Euclidean norm in $\mathbb{R}^p$, and defining the constants

$$c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)},$$

$p \in \mathbb{N}$, the *distance covariance* is defined for random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with finite first moments as

$$\mathcal{V}^2(X, Y) = \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} |f_{X,Y}(s, t) - f_X(s) f_Y(t)|_{\mathbb{C}}^2 \times \frac{\mathrm{d}s\mathrm{d}t}{\|s\|^{p+1}\|t\|^{q+1}}. \tag{7}$$

Since the integrand is nonnegative, it is obvious that $\mathcal{V}^2(X, Y) \geq 0$. Moreover, since $f_{X,Y}(s, t)$ equals the product of its marginals only in the case of independence between X and Y, one immediately recognizes that $\mathcal{V}^2(X, Y) = 0$ if and only if X and Y are independent.

For the purpose of establishing empirical versions of the distance covariance, an alternative representation of $\mathcal{V}^2(X, Y)$ for X, Y with finite second moments was derived in [59, 64],

$$\begin{aligned} \mathcal{V}^2(X, Y) = {} & \mathbb{E}[\|X - X'\|\|Y - Y'\|] \\ & + \mathbb{E}[\|X - X'\|]\mathbb{E}[\|Y - Y'\|] \\ & - 2\mathbb{E}[(\|X - X'\|)(\|Y - Y''\|)], \end{aligned} \tag{8}$$

where (X', Y') and (X'', Y'') denote i.i.d. copies of (X, Y).

Several authors have considered generalizations of the concept of distance covariance, for example, by replacing the inverse of $c_p c_q \|s\|^{p+1}\|t\|^{q+1}$ in equation (7) with alternative weights [4] or by applying metrics different from the Euclidean distance $\rho(x, x') = \|x - x'\|$ in equation (8) [30, 38, 51].

In this paper, we will consider a slight but novel extension of the version of *generalized distance covariance* defined by Dino Sejdinovic and his coauthors [51]. In the following, we will call a function $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ on a set $\mathcal{Z}$ a premetric if it is symmetric in its arguments and satisfies

$$x = y \quad \Rightarrow \quad \rho(x, y) = 0.$$

Then $(\mathcal{Z}, \rho)$ is called a premetric space. A premetric space $(\mathcal{Z}, \rho)$ is said to have negative type [51], Definition 2, if for all $n \geq 2$, $z_1, \dots, z_n \in \mathcal{Z}$ and $a_1, \dots, a_n \in \mathbb{R}$ with $\sum_{i=1}^n a_i = 0$,

$$\sum_{i,j=1}^n a_i a_j \rho(z_i, z_j) \leq 0.$$

Now let $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$ denote premetrics of negative type on $\mathcal{X}$ and $\mathcal{Y}$. Then, extending the definition of [51] to random variables with finite first moments of the corresponding premetrics, we define the *generalized distance covariance* as follows.

Definition 1. For any two premetric spaces of negative type $(\mathcal{X}, \rho_\mathcal{X})$ and $(\mathcal{Y}, \rho_\mathcal{Y})$ and random variables X, Y with values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, define the *squared generalized distance covariance* as

$$\begin{aligned}\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X, Y) &\\ = \mathbb{E}(\rho_\mathcal{X}(X, X')&(\rho_\mathcal{Y}(Y, Y') - \rho_\mathcal{Y}(Y, Y'') \\ &- \rho_\mathcal{Y}(Y', Y'') + \rho_\mathcal{Y}(Y'', Y'''))),\end{aligned} \tag{9}$$

where (X, Y), (X', Y'), (X'', Y''), (X''', Y''') are i.i.d. copies of (X, Y).

Proposition 1. *If* $\mathbb{E}|\rho_\mathcal{X}(X, X') + \rho_\mathcal{Y}(Y, Y')| < \infty$, *then* $\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X, Y) < \infty$.

It is easy to see that $\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X, Y) \geq 0$ and that independence of X and Y implies $\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X, Y) = 0$. However, the reverse implication holds only true if the premetrics $(\mathcal{X}, \rho_\mathcal{X})$ and $(\mathcal{Y}, \rho_\mathcal{Y})$ are of *strong negative type*, that is, if

$$\begin{aligned}&P_X = P_Z \\ \Leftrightarrow \quad &\int_\mathcal{X}\int_\mathcal{X} \rho_\mathcal{X}(x, x')\,\mathrm{d}(P_X(x) \\ &\quad - P_Z(x))\,\mathrm{d}(P_X(x') - P_Z(x')) = 0\end{aligned}$$

for all distributions P_X and P_Z; similarly for $\rho_\mathcal{Y}$.

Consider now i.i.d. joint samples $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ of X and Y, and define the (i, j)th element of the distance matrices $\boldsymbol{D}^{\boldsymbol{X}}$ and $\boldsymbol{D}^{\boldsymbol{Y}}$ by $\boldsymbol{D}^{\boldsymbol{X}}_{i,j} = \rho_\mathcal{X}(X_i, X_j)$ and $\boldsymbol{D}^{\boldsymbol{Y}}_{i,j} = \rho_\mathcal{Y}(Y_i, Y_j)$, respectively. Then defining the double-centered versions

$$\tilde{\boldsymbol{D}}^{\boldsymbol{X}} = (I - H)\boldsymbol{D}^{\boldsymbol{X}}(I - H),$$

$$\tilde{\boldsymbol{D}}^{\boldsymbol{Y}} = (I - H)\boldsymbol{D}^{\boldsymbol{Y}}(I - H),$$

where $H = \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^t$, a consistent empirical estimator for (9) is given by [64],

$$\begin{aligned}\widehat{\mathcal{V}}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(\boldsymbol{X}, \boldsymbol{Y}) &= \frac{1}{n^2}\sum_{i,j=1}^n \tilde{\boldsymbol{D}}^{\boldsymbol{X}}_{i,j}\tilde{\boldsymbol{D}}^{\boldsymbol{Y}}_{i,j} \\ &= \frac{1}{n^2}\operatorname{tr}(\tilde{\boldsymbol{D}}^{\boldsymbol{X}}\tilde{\boldsymbol{D}}^{\boldsymbol{Y}}).\end{aligned} \tag{10}$$

As with HSIC, the tilde above one of the two matrices in (10) may be removed.

4.2 Equivalence of Generalized Distance Covariance and HSIC

The similarity of the definitions and properties of Sections 3 and 4 have raised conjectures about the equivalence of generalized distance covariance and HSIC [22], which was indeed derived in [51]. Extended to the definition of generalized distance covariance in Definition 1, this equivalence is expressed by the following theorem.

Theorem 1. $\mathrm{HSIC}_{k_\mathcal{X},k_\mathcal{Y}}$ *is equivalent to the generalized distance covariance* $\mathcal{V}^2_{\rho(\cdot,\cdot;k_\mathcal{X}),\rho(\cdot,\cdot;k_\mathcal{Y})}$, *where the premetric* $\rho(z_1, z_2; k)$ *induced by a kernel* k *is defined via*

$$\begin{aligned}&\rho(z_1, z_2; k) \\ &\quad = \frac{1}{2}\big(k(z_1, z_1) + k(z_2, z_2) - 2k(z_1, z_2)\big).\end{aligned} \tag{11}$$

Conversely, $\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}$ *is equivalent to the Hilbert–Schmidt Independence Criterion* $\mathrm{HSIC}_{k(\cdot,\cdot;\rho_\mathcal{X},x_0),k(\cdot,\cdot;\rho_\mathcal{Y},y_0)}$, *where the kernel* $k(z_1, z_2; \rho, z_0)$ *induced by a premetric* ρ *is defined via*

$$\begin{aligned}&k(z_1, z_2; \rho, z_0) \\ &\quad = \rho(z_1, z_0) + \rho(z_2, z_0) - \rho(z_1, z_2),\end{aligned} \tag{12}$$

and $x_0 \in \mathcal{X}$, $y_0 \in \mathcal{Y}$ *are arbitrary.*

The proof of Theorem 1 follows directly from [2], p. 74, Lemma 2.1.

Moreover, it is straightforward to show the equality of the corresponding empirical versions given in equations (10) and (2); see [53] for details.

By means of Theorem 1, we obtain that the Gaussian kernel in the HSIC formulation corresponds to the metric

$$\rho(x, x') = 1 - \exp(-\|x - x'\|^2/2\sigma^2)$$

in the generalized distance covariance formulation. The discrete kernel corresponds to the discrete metric

$$\rho(x, x') = 1_{\{x \neq x'\}}.$$

The linear kernel $k(x, x) = \langle x, x'\rangle$ in $\mathbb{R}^p$ corresponds to one-half of the corresponding squared Euclidean distance

$$p(x, x') = \frac{1}{2}\|x - x'\|^2.$$

On the other hand, Theorem 1 also allows us to represent the classical squared distance covariance as Hilbert–Schmidt Independence Criterion.

COROLLARY 1. *The classical squared distance covariance can be expressed as*

$$\mathcal{V}^2(X,Y) = \mathrm{HSIC}_{k_p,k_q}(X,Y),$$

where k_p is given by

$$k_p(x,x') = \|x\| + \|x'\| - \|x-x'\|, \tag{13}$$

and k_q is defined analogously on $\mathbb{R}^q$.

This formulation enables the derivation of a feature map for the classical distance covariance, which we present in the following section.

4.3 A Feature Map for the Classical Distance Covariance

For establishing feature maps for the classical distance covariance $\mathcal{V}^2(X,Y)$ (cf. equation (7) or (8)), we focus on the bivariate case, that is, real-valued random variables X and Y. Moreover, we will need to assume that both X and Y are bounded and we will further assume w.l.o.g. ($\mathcal{V}^2$ is linear in its arguments) that $X, Y \in [0,1]$. In that case, we have

$$k_p(x,x') = 2\min(x,x'). \tag{14}$$

While this kernel does not play a major role in machine learning applications, it is well-known in the fields of probability theory and stochastic processes, where it equals two times the covariance function of a Brownian motion. In order to obtain series representations for Brownian motion, numerous expansions of the kernel in (14) have been derived [37]. Using these classical expansions, we obtain several feature maps that are valid for $x, x' \in [0,1]$. We mention $\mathbf{\Phi}(x) = \{\Phi_k(x)\}_{k=1}^{\infty}$ with

$$\Phi_k(x) = 2\frac{\sin(\pi(k-\frac{1}{2})x)}{(k-\frac{1}{2})\pi},$$

and $\mathbf{\Phi}(x) = \{\Phi_k(x)\}_{k=0}^{\infty}$ with

$$\begin{aligned}&\Phi_0(x) = \sqrt{2}x,\\ &\Phi_k(x) = 2\frac{\sin(\pi k x)}{\pi k}, \quad k = 1, \dots,\end{aligned} \tag{15}$$

similar expressions can be derived for $x, x' \in [a,b]$.

Combining equations (15) and (4) yields the following proposition.

PROPOSITION 2. *For jointly distributed random variables X, Y on* $[0,1]$, *it holds that*

$$\begin{aligned}\mathcal{V}^2(X,Y) = {} & 4\,\mathrm{Cov}^2(X,Y) + 8\sum_{k=1}^{\infty}\biggl(\mathrm{Cov}^2\biggl(X, \frac{\sin(\pi k Y)}{\pi k}\biggr)\\ & + \mathrm{Cov}^2\biggl(\frac{\sin(\pi k X)}{\pi k}, Y\biggr)\biggr)\\ & + 16\sum_{j,k=1}^{\infty}\mathrm{Cov}^2\biggl(\frac{\sin(\pi j X)}{\pi j}, \frac{\sin(\pi k Y)}{\pi k}\biggr).\end{aligned}$$

Applying the bilinearity of squared distance covariance and regular covariance, we find a lower bound to the distance covariance in terms of the usual covariance.

COROLLARY 2. *For jointly distributed random variables $X \in [a,b]$ and $Y \in [c,d]$, it holds*

$$\mathcal{V}^2(X,Y) \geq \frac{4}{(b-a)(c-d)}\mathrm{Cov}^2(X,Y).$$

5. A REGRESSION PERSPECTIVE ON HSIC AND GENERALIZED DISTANCE COVARIANCE

5.1 The Global Test

In the same period that HSIC and distance covariance were developed, a novel type of tests was proposed for applications in genomics. These *global tests* were defined as locally most powerful tests for the global null hypothesis in generalized linear regression models [15, 17, 18].

We will show in this section that the global test for the linear model arises as a special case of HSIC and generalized distance covariance, by applying the linear kernel on both variables under consideration. Conversely, we will show that HSIC and generalized distance covariance can be written in the form of a linear global test statistic provided that corresponding feature maps exist. As one of the main results of this article, we finally obtain Theorem 3, stating that in certain instances, HSIC and generalized distance covariance can be derived from locally most powerful tests.

Consider an empirical Bayes linear model with fixed and known intercept $\mu \in \mathbb{R}$ and error variance σ^2,

$$y_i|\beta \sim \mathcal{N}(\mu + \beta^t X_i, \sigma^2), \tag{16}$$

where $\boldsymbol{\beta}$ is a random variable with $\boldsymbol{\beta} = \tau\boldsymbol{b}$, $\mathbb{E}[\boldsymbol{b}] = 0$ and $\mathbb{E}[\boldsymbol{b}\boldsymbol{b}^t] = I_p$. Here $\tau \in \mathbb{R}$ is left as a fixed and unknown parameter, and no assumptions are made on the distribution of the marginals of $\boldsymbol{b}$. We denote by $\ell(\beta)$ the likelihood of β in (16). In this model we consider the problem of testing the null hypothesis

$$H_0 : \tau^2 = 0.$$

Obviously, $\tau^2 = 0$ if and only if $\boldsymbol{\beta} = 0$ a.s. Hence, H_0 is equivalent to

$$H_0 : \beta = 0.$$

Testing H_0 is now performed by integrating out β. Letting $\mathbb{E}_{\boldsymbol{\beta}|\tau^2}[\cdot]$ denote the expectation over the distribution of $\boldsymbol{\beta}$ for fixed τ^2, we obtain the marginal likelihood of τ^2 as [18]

$$\bar{\ell}(\tau^2) = \mathbb{E}_{\boldsymbol{\beta}|\tau^2}[\ell(\boldsymbol{\beta})]. \tag{17}$$

Since the marginal model is based on a prior distribution on the regression parameters, but the parameter τ^2 of the prior is treated as fixed and unknown, one may regard this approach as empirical Bayesian. This could be seen as a slight abuse of terminology, since the subsequent use of the model is different from the empirical Bayes methodology introduced by Robbins [46]. We refer to [18] for a discussion of this issue.

In [18], the locally most powerful test statistic corresponding to the likelihood $\overline{\ell}(\tau^2)$ is shown to be equivalent to

$$\frac{1}{n^2}\sum_{i,j=1}^{n}\langle X_i, X_j\rangle(y_i-\mu)(y_j-\mu). \tag{18}$$

In most practical circumstances μ will be unknown. To circumvent this problem, we switch to the profile likelihood [17], substituting μ by its maximum likelihood estimate $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$. We obtain the pivot

$$\widehat{GT}(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n^2}\sum_{i,j=1}^{n}\langle X_i, X_j\rangle\,(y_i-\widehat{\mu})(y_j-\widehat{\mu}). \tag{19}$$

This is what is introduced in [17] as the global test for the Gaussian linear model. In [6], it was shown that the argument similarly extends to multivariate responses $Y \in \mathbb{R}^q$, deriving that

$$\widehat{GT}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i,j=1}^{n}\langle X_i, X_j\rangle\,\langle Y_i-\widehat{\boldsymbol{\mu}}, Y_j-\widehat{\boldsymbol{\mu}}\rangle, \tag{20}$$

where $\widehat{\boldsymbol{\mu}}$ is the vector of coefficient-wise means, can be regarded as a pivot to the locally most powerful test in certain multivariate Gaussian linear models.

Comparing equations (20), (2) and (10), we obtain the following theorem.

THEOREM 2. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be samples of jointly distributed random vectors* $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$. *Then,*

$$\widehat{GT}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\mathrm{HSIC}}_{k_p,k_q}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\mathcal{V}}^2_{\rho_p\rho_q}(\boldsymbol{X}, \boldsymbol{Y}),$$

where $k_p(x,x') = \langle x, x'\rangle$ *and* $k_q(y,y') = \langle y, y'\rangle$ *are the corresponding linear kernels and* $\rho_p(x,x') = \frac{1}{2}\|x-x'\|^2$, $\rho_q(y,y') = \frac{1}{2}\|y-y'\|^2$ *are one half of the squared Euclidean distances. On the other hand, whenever feature maps for the kernels* $k_{\mathcal{X}}$ *and* $k_{\mathcal{Y}}$ *exist, then*

$$\begin{aligned}\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},k_{\mathcal{Y}}}(\boldsymbol{X}, \boldsymbol{Y}) &= \widehat{\mathcal{V}}^2_{\rho(\cdot,\cdot;k_{\mathcal{X}}),\rho(\cdot,\cdot;k_{\mathcal{Y}})}(\boldsymbol{X}, \boldsymbol{Y})\\ &= \widehat{GT}\big(\boldsymbol{X}^{\boldsymbol{\Phi}^{(k_{\mathcal{X}})}}, \boldsymbol{Y}^{\boldsymbol{\Phi}^{(k_{\mathcal{Y}})}}\big),\end{aligned}$$

where $\boldsymbol{X}^{\boldsymbol{\Phi}^{(k_{\mathcal{X}})}} \in \mathbb{R}^{n\times d_{\mathcal{X}}}$ *is the matrix with entries*

$$\boldsymbol{X}_{ij}^{\boldsymbol{\Phi}^{(k_{\mathcal{X}})}} = \Phi_j^{(k_{\mathcal{X}})}(X_i), \tag{21}$$

and $d_{\mathcal{X}}$ *is the dimension of the feature map of* $k_{\mathcal{X}}$; $d_{\mathcal{Y}}$ *and* $\boldsymbol{Y}^{\boldsymbol{\Phi}^{(k_{\mathcal{Y}})}} \in \mathbb{R}^{n\times d_{\mathcal{Y}}}$ *are defined analogously.*

5.2 HSIC and Generalized Distance Covariance as Locally Most Powerful Tests in Gaussian Regression Models

By the construction of HSIC and distance covariance as global tests, they inherit the property of being locally most powerful for specific models. This is given as Lemma 1. This lemma gives insight into the power properties of these methods, as it specifies exactly against which alternative the tests are optimal. In this section, we derive such consequences of Theorem 2, focusing on the case of a univariate response y with linear kernel, but general X.

LEMMA 1 (Lemma 4, [18]). *Consider the setting of the Gaussian linear model* (16) *with* $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta} = \tau\boldsymbol{b}$, $\mathbb{E}[\boldsymbol{b}] = 0$ *and* $\mathbb{E}[\boldsymbol{b}\boldsymbol{b}^t] = I_p$. *Let* $\overline{\omega}(\boldsymbol{\beta}) = P_{\boldsymbol{y}|\boldsymbol{\beta}}(\widehat{GT}(\boldsymbol{X}, \boldsymbol{y}) \geq k)$ *denote the power function of the global test. Let* $\omega(\boldsymbol{\beta}) = P_{\boldsymbol{y}|\boldsymbol{\beta}}(A)$ *be the power function of any test for* $H_0 : \boldsymbol{\beta} = 0$. *Then either of*

(i) $\omega(\mathbf{0}) = \overline{\omega}(\mathbf{0})$
(ii) $\omega(\mathbf{0}) \leq \overline{\omega}(\mathbf{0})$ *and* $k \geq 0$

implies

$$\mathbb{E}_{\xi}\left[\frac{d}{d\tau^2}\omega_{\xi}(0)\right] \leq \mathbb{E}_{\xi}\left[\frac{d}{d\tau^2}\overline{\omega}_{\xi}(0)\right],$$

where $\omega_{\xi}(\tau) = \omega(\tau\xi)$, $\overline{\omega}_{\xi}(\tau) = \overline{\omega}(\tau\xi)$ *and* ξ *has a uniform distribution on the unit p-ball. The same result holds when* ξ *has any other distribution on the unit p-ball such that* $\mathbb{E}[\xi] = 0$ *and* $\mathbb{E}[\xi\xi^t] = I_p$.

Gaussian process regression [43] is an approach for modeling nonlinear relations that has recently gained popularity in the area of machine learning. In Gaussian regression, one assumes that the response data $\boldsymbol{y}$ is generated by a zero-mean Gaussian process $V(\cdot)$ on the predictors $\boldsymbol{X}$, say $y_i = V(X_i) + \mu$. In most cases, it seems more appropriate to assume that we are given noisy observations of the Gaussian process [43], p. 16, say

$$y_i = \mu + V(X_i) + \epsilon_i, \tag{22}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. A direct consequence of Lemma 1 is that HSIC and distance covariance tests can be interpreted as locally most powerful tests in Gaussian process regression models of the form (22). In fact, $V(\cdot)$ is not required to be Gaussian, but can be replaced by any mean zero stochastic process with the same covariance function. The connection between stochastic process regression, HSIC and distance covariance has been independently explored by [28], however, without making the link to locally most powerful tests.

THEOREM 3. *Let* $V : \mathcal{X} \to \mathbb{R}$ *be a stochastic process with* $E[V(s)] \equiv 0$ *and* $\mathbb{E}[V(s)\,V(t)] = k(s,t)$ *for some kernel function* $k(\cdot,\cdot)$. *For* $i = 1,\ldots,n$, *consider the model*

$$y_i \sim \mathcal{N}(\mu + r_i, \sigma^2),$$

where $r_i = \tau V(X_i)$, $\tau \in \mathbb{R}$, further denote its likelihood by $g(r_i)$. Then the locally most powerful test statistic for testing $H_0 : \tau^2 = 0$ against $H_1 : \tau^2 > 0$ in the marginal model

$$\overline{\ell}(\tau^2) = \mathbb{E}_{V(\cdot)|\tau^2}\left[\prod_{i=1}^{n} g(r_i)\right], \tag{23}$$

is (up to translation and multiplication by constants)

$$\frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i, X_j)(y_i - \mu)(y_j - \mu). \tag{24}$$

REMARK 1. Plugging in $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$, a pivot for the locally most powerful test is given by

$$\widehat{\mathrm{HSIC}}_{k,l}(\boldsymbol{X}, \boldsymbol{y}) = \widehat{\mathcal{V}}^2_{\rho,\mathrm{euc}}(\boldsymbol{X}, \boldsymbol{y}),$$

where $l(y, y') = yy'$, $\mathrm{euc}(y, y') = \frac{1}{2}|y - y'|^2$ and $\rho(\cdot, \cdot)$ is obtained from $k(\cdot, \cdot)$ via equation (11).

We note that Theorem 3 does not require the existence of feature maps for the kernel $k(\cdot, \cdot)$. While the Theorem is based on a fixed sample $\boldsymbol{X}$, it can still be interpreted as a result for testing the null hypothesis of independence between X and y. In particular, $\tau^2 = 0$ implies that $P_{Y|X} = P_Y$, which is well known to characterize independence.

The central message of Theorem 3 is that it gives an alternative model-based interpretation of generalized distance covariance and HSIC, that provides a deeper understanding of these measures. Notably, such tests can be interpreted as locally most powerful test statistics in stochastic process regression models. The stochastic process model gives an indication which alternatives the test is targeted against, since the test is equivalent to a test that is optimally focused to detect that alternative.

We note that finite sample local optimality of the test statistic can not be guaranteed for the test statistic based on the profile likelihood and not the original likelihood since the locally most powerful test is the unavailable oracle test that knows the value of μ. Yet, from the above considerations, one may expect that these statistics will have comparably (relative to other tests) high power for alternatives with small τ^2, since it differs from the locally most powerful test only by a single, easily estimable parameter.

The results of this section rely on results from the global test literature that use a single simple response y. It is yet an open problem to generalize them to a general-dimensional Y with general kernel. That is, it is not yet clear what model would lead to HSIC as the locally most powerful test.

In the setting with general kernels $k_{\mathcal{X}}$ on X and $k_{\mathcal{Y}}$ on Y we obtain the following result for the population measures (see also [4]). This result has been shown for standard distance covariance and the linear multivariate global test without using kernels in [6, 59].

PROPOSITION 3. *Let $V_{\mathcal{X}}$ and $V_{\mathcal{Y}}$ be stochastic processes with mean 0 and covariance functions $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, respectively. Then*

$$\begin{aligned}\mathrm{HSIC}_{k_{\mathcal{X}},k_{\mathcal{Y}}}&(X, Y)\\ &= \mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X, Y)\\ &= \mathbb{E}_{(V_{\mathcal{X}},V_{\mathcal{Y}})}\left[\mathrm{Cov}^2\left(V_{\mathcal{X}}(X), V_{\mathcal{Y}}(Y)|V_{\mathcal{X}}, V_{\mathcal{Y}}\right)\right],\end{aligned} \tag{25}$$

where $\rho_{\mathcal{X}}(\cdot, \cdot)$ and $\rho_{\mathcal{Y}}(\cdot, \cdot)$ are obtained from $k_{\mathcal{X}}(\cdot, \cdot)$ and $k_{\mathcal{Y}}(\cdot, \cdot)$ via equation (11).

5.3 Eigenvalue Decompositions

We have shown that in the special case of a univariate y with a linear kernel HSIC/generalized distance covariance arise from locally most powerful tests in certain Gaussian regression models. We will now provide additional insight how this property arises. A deeper understanding can be obtained by decomposing the data matrices using kernel principal component analysis (kPCA) [50]. What's more, since this decomposition holds for all instances of HSIC/generalized distance covariance and not only for settings in which a linear kernel is used on the response, we see that the local optimality property at least partly translates to this more general setting.

PROPOSITION 4. *Let $\tilde{\boldsymbol{K}}_X$ and $\tilde{\boldsymbol{K}}_Y$ denote the double-centered kernel matrix corresponding to the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, and consider the decompositions*

$$\tilde{\boldsymbol{K}}_X = \sum_{i=1}^{n} \widehat{\lambda}_i^{\boldsymbol{X}} Q_i, \qquad \tilde{\boldsymbol{K}}_Y = \sum_{i=1}^{n} \widehat{\lambda}_i^{\boldsymbol{Y}} P_i,$$

where $\widehat{\lambda}_i^{\boldsymbol{X}}$ is the ith largest eigenvalue of $\tilde{\boldsymbol{K}}_X$ and $Q_i = v_i v_i^t$, where v_i is the corresponding eigenvector of $\tilde{\boldsymbol{K}}_X$. Similarly for $\tilde{\boldsymbol{K}}_Y$, $\widehat{\lambda}_i^{\boldsymbol{Y}}$ is the ith largest eigenvalue of $\tilde{\boldsymbol{K}}_Y$ and $P_i = w_i w_i^t$, where w_i is the corresponding eigenvector of $\tilde{\boldsymbol{K}}_Y$. Further denote by $\widehat{\mathrm{Cor}}$ the empirical correlation. Then

$$\begin{aligned}\widehat{\mathrm{HSIC}}&_{k_X,k_Y}(\boldsymbol{X}, \boldsymbol{Y})\\ &= \widehat{\mathcal{V}}^2_{\rho_X,\rho_Y}(\boldsymbol{X}, \boldsymbol{Y})\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \widehat{\lambda}_i^{\boldsymbol{X}}\widehat{\lambda}_j^{\boldsymbol{Y}}\widehat{\mathrm{Cor}}^2(w_j, v_i).\end{aligned} \tag{26}$$

The vector v_i is the projection of the data matrix $\boldsymbol{X}$ on its ith kernel principal component, centered and normalized to an empirical variance of $\frac{1}{n-1}$; the vector w_j is the projection of the data matrix $\boldsymbol{Y}$ on its jth kernel principal component, similarly standardized as v_i. Since $\widehat{\lambda}_1^{\boldsymbol{X}} \geq \cdots \geq \widehat{\lambda}_n^{\boldsymbol{X}}$ and $\widehat{\lambda}_1^{\boldsymbol{Y}} \geq \cdots \geq \widehat{\lambda}_n^{\boldsymbol{Y}}$, this decomposition shows that HSIC and generalized distance covariance put more weight on the directions of the first kernel principal components and hence focus on the main axes of variation in feature space.

The decomposition of Proposition 4 as a weighted average gives insight into the power properties of the test. The test statistic is large whenever the large variance kernel principle components of $\boldsymbol{X}$ are strongly correlated with the large variance kernel principle components of $\boldsymbol{Y}$. In settings for which the the maximal correlation is small, there is little hope to detect possible associations linked with directions of small variation. In this case, it is obviously preferable to focus on alternatives that are associated with larger (kernel) principal components. Decomposition (26) gives an alternative insight into the local optimality property of Theorem 3; see also [18], Section 7, for a detailed discussion of this issue for the global test case of the linear model with univariate response.

Proposition 4 also points to the most important contrast between the tests discussed in this article, and classical tests, such as the F-test in linear regression. Such tests typically put the same weight on all independent directions of the data, regardless of the eigenvalues. While this is feasible if the number of nonzero eigenvalues p_0 satisfies $p_0 \ll n$, it is not possible to have power into all directions for $p_0 > n$; this was a motivation for introducing the global test ([18], Section 1). For kernel tests based on characteristic kernels, the dimension of feature space is infinite; consequently we cannot achieve power for all directions in feature space. The three tests handle this problem in precisely the same way: by preferring alternatives that are associated with changes in axes of high variation in the original space (global test) or feature space (HSIC/generalized distance covariance).

As in [18], we can contrast HSIC/distance covariance/global test with an F-like test, constructed by putting different weights on the kernel principal components in decomposition (26). In general, we may always obtain an alternative pre-weighted test statistic of the form $\sum_{i=1}^n \sum_{j=1}^n a_i b_j (w_j^t v_i)^2$ where $\boldsymbol{a} = (a_1 \dots, a_n) \in \mathbb{R}^n$ and $\boldsymbol{b} = (b_1, \dots b_n) \in \mathbb{R}^n$. One could example choose the sequence $a_i = 1_{\{i \le \gamma(n)\}} 1_{\{\widehat{\lambda_i^X} > 0\}}$, $b_i = 1_{\{i \le \delta(n)\}} 1_{\{\widehat{\lambda_j^Y} > 0\}}$, where $\gamma : \mathbb{N} \to \mathbb{N}$ and $\delta : \mathbb{N} \to \mathbb{N}$ are monotonously increasing functions. This choice would imply equally weighting the correlation between the first $\gamma(n)$ kernel PCs of $\boldsymbol{X}$ and the first $\delta(n)$ kernel PCs of $\boldsymbol{Y}$. Note that an F-test-type choice would be achieved by setting $\gamma(n) = n$ and $\delta(n) = n$. This would lead to a test that is affine invariant in X and Y. However, such a test is not feasible if there is variation in all n kernel PCs, such as in case of high-dimensional data or for continuous data with characteristic kernels; it is easy to see that the test statistic in equation (26) is then constant.

The decomposition (26) also suggests an approximation of the distribution of the statistic $n\,\widehat{\mathrm{HSIC}}_{k_X,k_Y}(\boldsymbol{X}, \boldsymbol{Y})$, via

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\widehat{\lambda}_i^{\boldsymbol{X}}}{n} \frac{\widehat{\lambda}_j^{\boldsymbol{Y}}}{n} Z_{ij}^2, \tag{27}$$

with $Z_{ij}^2 \sim \chi_1^2$, for the purpose of calculating p-values for corresponding independence tests. It is straightforward to show (see [17]) that this approximation is exact for the Gaussian model with known intercept and known σ^2, that is, in the setting of Theorem 3, and adjustments for estimation of μ and σ^2 are straightforward [17]. For HSIC and distance covariance, similar procedures have been proposed in [29, 42, 69]; however, in these traditions, it is more common to use two-moment gamma approximations [23, 29, 42] or permutation tests [16, 23, 64] Recently, a procedure based on matching the first three moments of the statistic to a Pearson type III distribution [3] has been shown to markedly outperform two-moment gamma approximations in practice.

Finally, for a univariate response $\boldsymbol{y}$ with a linear kernel on $\boldsymbol{y}$, we establish an interesting link to kernel partial least squares (kPLS) regression [48, 49]. For this purpose, we first give a connection of the global test with classical PLS that has been first stated in [15].

LEMMA 2. *Define the first component of PLS regression* [67] *as the linear combination* $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$ ($\|\boldsymbol{w}\| = 1$) *satisfying*

$$\widehat{\mathrm{Cov}}(\boldsymbol{t}, \boldsymbol{y}) = \max_{\|\boldsymbol{u}\|=1} \widehat{\mathrm{Cov}}(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{y}).$$

Then

$$\widehat{GT}(\boldsymbol{X}, \boldsymbol{y}) = \frac{(n-1)^2}{n^2} \widehat{\mathrm{Cov}}^2(\boldsymbol{t}, \boldsymbol{y}),$$

where $\widehat{\mathrm{Cov}}$ *is the standard sample covariance.*

The result of Lemma 2 directly carries over to HSIC and generalized distance covariance, respectively.

THEOREM 4. *Let* $k(\cdot, \cdot)$ *be a kernel with corresponding feature map* $\Phi(\cdot)$. *Now define the first component of kernel PLS regression as the linear combination* $\boldsymbol{t} = X^{\Phi(k)}\boldsymbol{w}$ ($\|\boldsymbol{w}\| = 1$) *satisfying*

$$\widehat{\mathrm{Cov}}(\boldsymbol{t}, \boldsymbol{y}) = \max_{\|\boldsymbol{u}\|=1} \widehat{\mathrm{Cov}}\big(X^{\Phi(k)}\boldsymbol{u}, \boldsymbol{y}\big),$$

where $X^{\Phi(k)}$ *is defined via equation* (21). *Then*

$$\widehat{\mathrm{HSIC}}_{k,l}(\boldsymbol{X}, \boldsymbol{y}) = \widehat{\mathcal{V}}^2_{\rho,euc}(\boldsymbol{X}, \boldsymbol{y}) = \frac{(n-1)^2}{n^2} \widehat{\mathrm{Cov}}^2(\boldsymbol{t}, \boldsymbol{y}),$$

where $l(y, y') = yy'$, $euc(y, y') = \frac{1}{2}|y - y'|^2$ *and* $\rho(\cdot, \cdot)$ *is obtained from* $k(\cdot, \cdot)$ *via equation* (11).

6. GENERALIZED CORRELATION COEFFICIENTS

Since distance covariance shares many commonalities with classical covariance [11, 59, 64], it is natural to standardize distance covariance to a *distance correlation*, a measure quantifying the strength of association between datasets. We will extend the concept of distance correlation to general premetrics, so that it applies to HSIC and global test as well, through Theorem 2. We define the *generalized distance correlation* (see also [51], Appendix A) and a corresponding sample measure as

$$\begin{aligned}\mathcal{R}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y) &= \frac{\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y)}{\mathcal{V}_{\rho_\mathcal{X},\rho_\mathcal{X}}(X,X)\,\mathcal{V}_{\rho_\mathcal{Y},\rho_\mathcal{Y}}(Y,Y)},\\ \widehat{\mathcal{R}}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(\boldsymbol{X},\boldsymbol{Y}) &= \frac{\widehat{\mathcal{V}}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(\boldsymbol{X},\boldsymbol{Y})}{\widehat{\mathcal{V}}_{\rho_\mathcal{X},\rho_\mathcal{X}}(\boldsymbol{X},\boldsymbol{X})\,\widehat{\mathcal{V}}_{\rho_\mathcal{Y},\rho_\mathcal{Y}}(\boldsymbol{Y},\boldsymbol{Y})},\end{aligned} \tag{28}$$

where $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$ are premetrics on the spaces $\mathcal{X}$ and $\mathcal{Y}$. In the case that $\widehat{\mathcal{V}}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y)$ is the global test statistic, that is, if $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$ are one half the squared Euclidean distance, this leads to an improved version of the global test correlation coefficient proposed in [6], Section 2.3, in the sense that it satisfies a set of axiomatic properties; see Proposition 5.

Proposition 5 gives some important properties of generalized distance correlation, several of which were already stated in [51], Appendix A. These properties suggest that generalized distance correlation may be used for interpreting the strength of dependence between random variables.

PROPOSITION 5.

(i) $0 \le \mathcal{R}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y) \le 1$.

(ii) *If $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$ are of strong negative type, then*

$$\mathcal{R}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y) = 0 \iff X \textit{ and } Y \textit{ are independent.}$$

(iii) *If $\mathcal{X} = \mathcal{Y}$ and $\rho_\mathcal{X} = \rho_\mathcal{Y}$, then*

$$X = Y \Longrightarrow \mathcal{R}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y) = 1.$$

Moreover, if $\mathcal{X}$ and $\mathcal{Y}$ are vector spaces over a field F, then:

(iv) *If $\rho_\mathcal{X}(x_1,x_2) = \rho_\mathcal{X}(x_1+c,x_2+c)$ and $\rho_\mathcal{Y}(y_1,y_2) = \rho_\mathcal{Y}(y_1+d,y_2+d)$ for all $x_1,x_2,c \in \mathcal{X}$ and all $y_1,y_2,d \in \mathcal{Y}$, then*

$$\mathcal{R}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X+c,Y+d) = \mathcal{R}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y),$$

for all $c \in \mathcal{X}$, $d \in \mathcal{Y}$.

(v) *If there exist functions $g: F\backslash\{0\} \to \mathbb{R}_+$, $h: F\backslash\{0\} \to \mathbb{R}_+$, such that $\rho_\mathcal{X}(ax_1,ax_2) = g(a)\rho_\mathcal{X}(x_1,x_2)$ and $\rho_\mathcal{Y}(by_1,by_2) = h(b)\rho_\mathcal{Y}(y_1,y_2)$ for all $x_1,x_2 \in \mathcal{X}$, $y_1,y_2 \in \mathcal{Y}$ and nonzero scalars $a,b \in F$, then*

$$\mathcal{R}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(aX,bY) = \mathcal{R}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y),$$

for nonzero scalars $a,b \in F$.

Properties analogous to (iv) and (v) also hold when the corresponding properties are satisfied for the premetric-induced kernels instead of the premetrics themselves. Moreover, analogous properties may also be derived for the sample measure $\widehat{\mathcal{R}}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(\boldsymbol{X},\boldsymbol{Y})$. We remark that the invariance properties (iv) and (v) are satisfied for standard distance correlation and the (new) correlation measure based on the linear global test statistic, while correlation measures based on higher-order polynomial kernels satisfy neither of (iv) and (v), and the Gaussian kernel does not satisfy (v).

The generalized distance correlation can be seen as an alternative to classical R squared measures "with a different interpretation and different properties" [6]. Given a data matrix of predictors $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and a response $\boldsymbol{y} \in \mathbb{R}^n$, the classical R squared may be retained as

$$\widehat{\mathcal{R}}^2_{\rho_p,\rho_1}(\boldsymbol{X}\,\Sigma_{\boldsymbol{X}}^{-1/2}, \boldsymbol{y}),$$

where for $x_1,x_2 \in \mathbb{R}^q$, $\rho_q(x_1,x_2) = \frac{1}{2}\|x_1-x_2\|^2$ and $\Sigma_{\boldsymbol{X}}$ is the empirical covariance matrix of $\boldsymbol{X}$. A similar measure based on distance correlation is the *affinely invariant distance correlation* [7, 64]; for multivariate Gaussian data the affinely invariant distance correlation is just a deterministic function of the classical R squared [7, 8, 64].

A drawback when using the empirical version of generalized distance correlation for measuring independence is that these measures are often severely biased. Notably, when fixing the sample size n while letting the dimensions p of X and q of Y go to infinity, it may happen that $\widehat{\mathcal{R}}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(\boldsymbol{X},\boldsymbol{Y})$ converges to 1 even under independence of X and Y [60]. Even when only one of the dimensions is high, large difficulties in the interpretation of distance correlation may arise ([11], Section 4).

To obtain better interpretable results for the classical distance correlation, it has been proposed [11, 60] to use U-statistic versions of the corresponding squared distance covariances and distance standard deviations. A U-statistic estimator of the squared generalized distance covariance is given by

$$\begin{aligned}\widehat{\Omega}_{\rho_\mathcal{X},\rho_\mathcal{Y}}(\boldsymbol{X},\boldsymbol{Y}) &= \frac{1}{n(n-3)}\Bigg[\sum_{i,j=1}^n \rho_\mathcal{X}(X_i,X_j)\rho_\mathcal{Y}(Y_i,Y_j)\\ &\quad + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^n \rho_\mathcal{X}(X_i,X_j)\\ &\quad \cdot \sum_{i,j=1}^n \rho_\mathcal{Y}(Y_i,Y_j)\\ &\quad - \frac{2}{(n-2)}\sum_{i,j,k=1}^n \rho_\mathcal{X}(X_i,X_j)\rho_\mathcal{Y}(Y_i,Y_k)\Bigg],\end{aligned} \tag{29}$$

Analogous U-statistic estimators for HSIC and the global test can be straightforwardly established using the equivalence results derived in this work.

For the classical distance correlation, it has been shown that an empirical distance correlation based on U-statistic estimates of the distance covariance and distance standard deviations gives more meaningful results than the estimator in (28) [11], Section 4; this phenomenon similarly arises for the generalized distance correlation. It is hence recommended to use U-statistic estimators also for generalized distance correlations based on HSIC or the global test.

7. MODEL-BASED EXTENSIONS OF HSIC AND GENERALIZED DISTANCE COVARIANCE

We have seen in Section 5 that, if the linear kernel or quadratic Euclidean distance, respectively, are applied on the univariate response data $\boldsymbol{y}$, the corresponding HSIC and generalized distance covariance arise from locally most powerful test statistics in certain stochastic process regression models. Global tests arising from locally most powerful tests have been derived for Empirical Bayes versions of general statistical models in [18] with a special focus on generalized linear models in [17]. We now extend these results using stochastic processes to obtain useful model-based extensions of HSIC and generalized distance covariance.

THEOREM 5. *Let $V : \mathcal{X} \to \mathbb{R}$ be a stochastic process with $E[V(s)] \equiv 0$ and $\mathbb{E}[V(s)V(t)] = k(s,t)$ for some kernel function $k(\cdot,\cdot)$. For $i \in \{1,\dots,n\}$, let $X_i \in \mathcal{X}$ and consider a statistical model with likelihood*

$$g_i(r_i) = \exp\big(f_i(r_i)\big), \tag{30}$$

where $r_i = \tau V(X_i)$ with $\tau \in \mathbb{R}$. We further assume that, for all i, the first two derivatives of f_i exist almost everywhere and are bounded in a neighborhood of 0. Then the locally most powerful test statistic for testing $H_0 : \tau^2 = 0$ against $H_1 : \tau^2 > 0$ in the marginal model

$$\bar{\ell}(\tau^2) = \mathbb{E}_{V(\cdot)|\tau^2}\Bigg[\prod_{i=1}^{n} g_i(r_i)\Bigg], \tag{31}$$

is (up to translation and multiplication by constants)

$$\begin{aligned}&\frac{1}{2}\Bigg(\sum_{i=1}^{n} k(X_i,X_i)\frac{\partial^2 f_i(0)}{(\partial r_i)^2}\\ &\quad + \sum_{i,j=1}^{n} k(X_i,X_j)\frac{\partial f_i(0)}{\partial r_i}\frac{\partial f_j(0)}{\partial r_j}\Bigg).\end{aligned} \tag{32}$$

If f_i does not depend on i, the statistic in (32) can easily be recognized as a V-statistic of order two, so that testing can be carried out using standard theory [34]. In many applications, however, either f_i depends on i, or some of the model parameters are not known and pivot statistics must be found. In such case, tailored procedures for calculating p-values must be developed for different models, cf. [17] for the generalized linear model, and [12, 15] for survival analysis models.

In the remainder of this section, we present several useful extensions of HSIC and generalized distance covariance emerging from statistics of the form (32). We emphasize that this list is not exhaustive and many more interesting test statistics can be established using the same theory. As an example, the results for generalized linear models [17] may be extended to Bayesian generalized kernel models [70] in the future.

7.1 Linearly Adjusting for Nuisance Covariates

In applications it is often necessary to correct for nuisance covariates, for example, confounders. Consider for example the problem that motivated the original global test [16] of testing whether the expression of certain groups of genes (defined, e.g., via pathway or gene ontology terms) is associated with some clinical response, say response to treatment or development of a certain disease. In this case, a practitioner will often need to adjust for clinical covariates that act as confounders, having an impact on both the response and the gene expression. A direct application of Theorem 5 yields the following corollary.

COROLLARY 3. *Let $V : \mathcal{X} \to \mathbb{R}$ be a stochastic process with $E[V(s)] \equiv 0$ and $\mathbb{E}[V(s)V(t)] = k(s,t)$ for some kernel function $k(\cdot,\cdot)$. For $i \in \{1,\dots,n\}$, let $X_i \in \mathcal{X}$ and $Z_i = (1, Z_1,\dots,Z_{p_0})^t \in \mathbb{R}^{(p_0+1)}$. Consider a statistical model with likelihood*

$$g_i(r_i) = \exp\big(f_i(r_i + c_i)\big),$$

where $r_i = \tau V(X_i)$, $\tau \in \mathbb{R}$ and $c_i = \boldsymbol{\gamma}^t Z_i$ with known regression coefficients of the nuisance covariates $\boldsymbol{\gamma} = (\mu, \gamma_1,\dots,\gamma_{p_0})$. We further assume that, for all i, the first two derivatives of f_i exist almost everywhere and are bounded in a neighborhood of 0. Then the locally most powerful test statistic for testing $H_0 : \tau^2 = 0$ against $H_1 : \tau^2 > 0$ in the marginal model

$$\bar{\ell}(\tau^2) = \mathbb{E}_{V(\cdot)|\tau^2}\Bigg[\prod_{i=1}^{n} g_i(r_i)\Bigg], \tag{33}$$

is (up to translation and multiplication by constants),

$$\begin{aligned}&\frac{1}{2}\Bigg(\sum_{i=1}^{n} k(X_i,X_i)\frac{\partial^2 f_i(c_i)}{(\partial r_i)^2}\\ &\quad + \sum_{i,j=1}^{n} k(X_i,X_j)\frac{\partial f_i(c_i)}{\partial r_i}\frac{\partial f_j(c_j)}{\partial r_j}\Bigg).\end{aligned} \tag{34}$$

Consider, for example, the model $\boldsymbol{y} = (y_1,\dots,y_n)^t$,

$$y_i \sim \mathcal{N}(c_i + r_i, \sigma^2), \tag{35}$$

where $r_i = \tau V(X_i)$ and V is a Gaussian process with covariance function k and $\sigma^2 > 0$. Then the locally most powerful test for testing $H_0 : \tau^2 = 0$ is

$$\sum_{i,j=1}^{n} k(X_i,X_j)(y_i - c_i)(y_j - c_j). \tag{36}$$

Plugging in the ML estimates for the regression coefficients $\boldsymbol{\gamma}$ and simplifying the outcome, we obtain as pivot statistic

$$\sum_{i,j=1}^{n} \tilde{\boldsymbol{K}}_{ij}^{\boldsymbol{Z}} \tilde{\boldsymbol{L}}_{ij}^{\boldsymbol{Z}}, \tag{37}$$

where $\tilde{\boldsymbol{L}}^{\boldsymbol{Z}} = (I - H_{\boldsymbol{Z}})\boldsymbol{Y}\boldsymbol{Y}^t(I - H_{\boldsymbol{Z}})$, $\tilde{\boldsymbol{K}}^{\boldsymbol{Z}} = (I - H_{\boldsymbol{Z}}) \times \boldsymbol{K}(I - H_{\boldsymbol{Z}})$, $H_{\boldsymbol{Z}} = \boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\boldsymbol{Z}^t$ and $\boldsymbol{K}$ is the matrix of which the (i, j)th entry is given by $k(X_i, X_j)$. If there are no nuisance covariates, that is, $\boldsymbol{\gamma} = \mu$, (37) simplifies to the standard HSIC/generalized distance covariance with a linear kernel on $\boldsymbol{y}$. With nuisance covariates, we obtain a novel version of HSIC/generalized distance, linearly adjusted for the influence of $\boldsymbol{Z}$. This test statistic differs from the standard HSIC test statistic only in the hat matrix $H_{\boldsymbol{Z}}$ used, that orthogonalizes the kernel matrix to all columns of $\boldsymbol{Z} = (Z_1, \dots, Z_n)^t$ rather than only the intercept, as H in the usual HSIC did.

In the presence of nuisance parameters, corresponding permutation tests (and bootstrap tests) are only valid under the highly restrictive assumption that the nuisance covariates are independent of the covariates X. A valid alternative in this setting is given by the sign-flipping test [25], approximating the distribution of (37) under the null by

$$\sum_{i,j=1}^{n} \tilde{\boldsymbol{K}}_{ij}^{\boldsymbol{Z}} \tilde{\boldsymbol{L}}_{ij}^{\boldsymbol{Z}} g_j g_j^t, \tag{38}$$

where, for $j \in \{1, \dots, B\}$, $g_j = (g_{j1}, \dots, g_{jn})^t$ are random functions in $\{-1, 1\}^n$. This, however, may lead to conservative tests. An asymptotically exact test based on (37) can be established by replacing the score contributions in the corresponding tests by the *efficient score contributions*; see [25] for details. An alternative way to obtain asymptotically exact tests in Gaussian regression models with nuisance covariates is to apply a rotation-based test [54]; however, this method relies on the assumption of normally distributed errors.

7.2 Goodness-of-Fit Testing

Goodness-of-fit tests may be established in a similar way as the tests in Section 7.1, and also require the adjustment for nuisance covariates. As an example consider the null model

$$y_i \sim \mathcal{N}(\boldsymbol{\beta}^t \tilde{X}_i, \sigma^2), \quad \boldsymbol{\beta} \in \mathbb{R}^{p+1}, \sigma^2 > 0, \tag{39}$$

where $\tilde{X}_i = (1, \boldsymbol{X}_{i1}, \dots, \boldsymbol{X}_{ip}) \in \mathbb{R}^{p+1}$ is the ith observation of $\boldsymbol{X}$, expanded to include an intercept, and $\boldsymbol{\beta} = (\mu, \beta_1, \dots, \beta_p)$ is the vector of regression coefficients. Also, we denote by $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \dots, \tilde{X}_n)^t$ the data matrix expanded by a column of ones. This model may be embedded into the more general model

$$\begin{gathered} y_i \sim \mathcal{N}(\boldsymbol{\beta}^t \tilde{X}_i + h(X_i), \sigma^2), \\ \boldsymbol{\beta} \in \mathbb{R}^{p+1}, \sigma^2 > 0, h \in \mathcal{H}, \end{gathered} \tag{40}$$

where $\mathcal{H} = \{h : \mathbb{R}^p \mapsto \mathbb{R} | h(\boldsymbol{x}) = \sum_{j=1}^{d} \Phi_j(\boldsymbol{x})\}$ and $\boldsymbol{\Phi} = (\Phi_1, \dots, \Phi_d)$ is a feature map of a kernel k (possibly $d = \infty$).

We could then be interested in testing whether the model (39) is adequate for expressing the relation between y_i and X_i, compared to the more general model (40), that is, testing $h(\cdot) = 0$. Along the lines of Section 7.1, we may derive a suitable test statistic for this purpose, given by

$$\sum_{i,j=1}^{n} \tilde{\boldsymbol{K}}_{ij}^{\tilde{\boldsymbol{X}}} \tilde{\boldsymbol{L}}_{ij}^{\tilde{\boldsymbol{X}}}, \tag{41}$$

where $\tilde{\boldsymbol{L}}^{\tilde{\boldsymbol{X}}} = (I - H_{\tilde{\boldsymbol{X}}})\boldsymbol{Y}\boldsymbol{Y}^t(I - H_{\tilde{\boldsymbol{X}}})$, $\tilde{\boldsymbol{K}}^{\tilde{\boldsymbol{X}}} = (I - H_{\tilde{\boldsymbol{X}}}) \times \boldsymbol{K}(I - H_{\tilde{\boldsymbol{X}}})$, $H_{\tilde{\boldsymbol{X}}} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^t$ and $\boldsymbol{K}$ is the matrix of which the (i, j)th entry is given by $k(X_i, X_j)$. Similar tests may be constructed for other statistical models, such as proportional hazards regression. For complementary details about goodness-of-fit tests derived from global tests, we refer the reader to [55].

7.3 Heavy-Tailed Data

In Section 5, we have shown that certain instances of HSIC and generalized distance covariance arise as locally most powerful models of the form (16), assuming a Gaussian distribution of the errors. In settings with heavy-tailed responses such an assumption is sub-optimal and may lead to low power of the corresponding tests. The following corollary presents a test for t-distributed errors.

COROLLARY 4. *Let $V : \mathcal{X} \to \mathbb{R}$ be a stochastic process with $E[V(s)] \equiv 0$ and $\mathbb{E}[V(s)V(t)] = k(s, t)$ for some kernel function $k(\cdot, \cdot)$. For $i \in \{1, \dots, n\}$, let $X_i \in \mathcal{X}$ and*

$$y_i = \mu + V(X_i) + \sigma\epsilon_\nu, \tag{42}$$

where $r_i = \tau V(X_i)$, $\tau \in \mathbb{R}$ and ϵ_ν follows a t-distribution with ν degrees of freedom; denote its likelihood by $g(r_i)$. Then the locally most powerful test statistic for testing $H_0 : \tau^2 = 0$ against $H_1 : \tau^2 > 0$ in the marginal model

$$\bar{\ell}(\tau^2) = \mathbb{E}_{V(\cdot)|\tau^2}\left[\prod_{i=1}^{n} g(r_i)\right], \tag{43}$$

is given by (up to translation and multiplication by constants)

$$\begin{aligned} &\sum_{i,j=1}^{n} k(x_i, x_j) \\ &\quad \times \frac{y_i - \mu}{(y_i - \mu)^2 + \sigma^2\nu} \frac{y_j - \mu}{(y_j - \mu)^2 + \sigma^2\nu} \\ &\quad + \sum_{i=1}^{n} k(x_i, x_i) \frac{(y_i - \mu)^2 - \sigma^2\nu}{((y_i - \mu)^2 + \sigma^2\nu)^2}. \end{aligned}$$

In practice, typically neither of the parameters σ^2, ν or μ will be known, which makes it necessary to plug in ML estimates of the parameters, which may be obtained using the Expectation Maximization (EM) algorithm [36].

7.4 Survival Models

Consider a proportional hazards survival model, where the hazard of individual i at time t is given by

$$h_i(t) = h(t)\exp(r_i),$$

with $r_i = \tau V(X_i)$, where V is a Gaussian process with covariance function k and $h(t)$ is the baseline hazard. We allow the follow-up times to be right-censored, assuming that the censoring times and survival times are independent given the covariates. This is essentially a classical Cox model allowing for nonlinear relations between the predictor X and the log-hazard ratio. Given that the hazard functions are known, the full likelihood in this model can be written in the form stated in equation (30) with

$$f_i(r_i) = \Delta_i\big(\log\big(h(t_i)\big) + r_i\big) - H(t_i)\exp(r_i),$$

where $H(\cdot)$ is the cumulative hazard function and t_i and Δ_i denote the follow-up time and censoring indicator of the ith individual respectively.

Applying Theorem 5 yields that the locally most powerful test statistic in this model is given by

$$\begin{aligned}\frac{1}{2}\Bigg(&-\sum_{i=1}^{n} k(X_i, X_i)H(t_i)\\ &+\sum_{i,j=1}^{n} k(X_i, X_j)\big(\Delta_i - H(t_i)\big)\big(\Delta_j - H(t_i)\big)\Bigg);\end{aligned} \tag{44}$$

see [15] for details on the corresponding linear test. In practice the baseline hazard will be unknown. To solve this problem, we plug in estimates for $H(t_i)$, $i = 1, \dots, n$,

$$\widehat{H}(t_i) = \sum_{t_j \le t_i} \frac{\Delta_j}{R(t_j)},$$

where $R(t)$ is the number of individuals at risk at time t. In contrast to the global test for the standard Cox model in [15], the resulting statistics is also able to detect alternatives in which the log-hazard ratio is a nonlinear function of the predictors. We note that a very similar test has been proposed in [5]. The test statistic (44) can be extended to a stratified proportional hazards model of the form

$$h_i(t) = h^{(i)}(t)\exp(r_i),$$

where $h^{(i)}(\cdot)$ is the baseline hazard of the ith stratum, see [12] for the derivation in the case where k is the linear kernel. As special cases of the stratified model, we may obtain test statistics for conditional logistic regression, competing risks and multistate models [12].

8. INDEPENDENCE TESTING USING THE DISCRETE KERNEL

A particularly interesting choice for a kernel function in HSIC, which has gained surprisingly little attention in literature, is the discrete kernel, which corresponds to the discrete distance in generalized distance covariance, cf. Section 3.2. The discrete kernel is invaluable for tests involving categorical data, which are ubiquitous in many applications. In this section, we present discrete tests following from the considerations in Sections 3.2 and 7.

Section 8.1 considers that only one of the two kernels is discrete, showing up important connections between independence testing and location testing of m samples; we note that some of these results have been independently derived in [52]. This subsection establishes a link between HSIC and distance covariance tests and (multinomial) logistic regression.

Next, Section 8.2 discusses a test between two sets of categorical data. This second test is particularly interesting since it arises naturally and identically from all three traditions discussed in this work. For HSIC and distance covariance, it results from applying the discrete kernel or discrete distance, respectively on both datasets. Within the global test framework, it arises a a locally most powerful test in a multinomial logistic regression model with a categorical predictor. By contrasting this test to the classical chi-square test for $m \times r$ contingency tables, we provide additional insight into commonality between the three traditions considered in this article, and their contrast to standard tests.

8.1 Testing Independence Between Categorical and Continuous or Ordinal Data

If the discrete kernel is applied to one variable and a different kernel is applied to another variable, we typically obtain tests that can be used for testing independence between categorical and continuous or ordinal data. Let d denote the discrete kernel on $\mathcal{Y}$.

We first consider that $\mathcal{Y}$ contains only two elements. Without loss of generality we assume that $\mathcal{Y} = \{0, 1\}$, while $\mathcal{X}$ may be any arbitrary set. For observations $\boldsymbol{X} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \{0, 1\}^n$, we now study $\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d}$, where $k_{\mathcal{X}}$ is an arbitrary kernel on $\mathcal{X}$.

From Theorem 5 and the considerations in [14, 17], it follows directly that $\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d}$ is a pivot for the locally most powerful test for the null hypothesis $H_0 : \tau^2 = 0$ in a binomial regression model with logit link,

$$y_i \sim \mathcal{B}\bigg(1, \frac{\exp^{\mu+V(X_i)}}{1+\exp^{\mu+V(X_i)}}\bigg), \tag{45}$$

where $\mathcal{B}(N, p)$ denotes the binomial distribution and $V(\cdot)$ is a stochastic process with covariance function $\tau^2 k_{\mathcal{X}}$.

From another viewpoint, $\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d}$ can be regarded as measure of distance between the distributions of the X_i

for which $y_i = 0$ and the X_i for which $y_i = 1$. Proposition 6 states that $\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d}$ is just a multiple of the squared empirical *maximum mean discrepancy* (MMD) with kernel function $k_{\mathcal{X}}$, a well-known distance between probability distributions used in machine learning [19, 20].

PROPOSITION 6. *Let $k_{\mathcal{X}}$ denote an arbitrary kernel on some set $\mathcal{X}$ and let d denote the discrete kernel on $\{0, 1\}$. Also, for any $S = \{s_1, \dots, s_l\} \subset \{1, \dots, n\}$ with $s_1 < s_2 \dots < s_l$, let $X_j^S = X_{s_j}$ for $j \in \{1, \dots, l\}$ and $\boldsymbol{X}^S = (X_1^S, \dots, X_l^S)$. Moreover, for $j = 0, 1$, denote $S_j = \{i \in \{1, \dots, n\} | y_i = j\}$ and $n_j = |S_j|$. Then it holds that*

$$\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d}(\boldsymbol{X}, \boldsymbol{y}) = \frac{2 n_0^2 n_1^2}{n^4} \widehat{\mathrm{MMD}}_{k_{\mathcal{X}}}^2 (\boldsymbol{X}^{S_0}, \boldsymbol{X}^{S_1}), \tag{46}$$

where $\widehat{\mathrm{MMD}}_{k_{\mathcal{X}}}$ is the empirical version of the maximum mean discrepancy [19], *Eq.* (6).

Choosing for $k_{\mathcal{X}}$ the kernel function corresponding to distance covariance, the expression in (46) reduces to a multiple of the squared classical energy distance [61, 63]. Using the linear kernel leads to a linear two-sample test that is applicable in high dimensions and was previously studied in [1, 56].

When $\mathcal{Y}$ consists of more than two elements, say $\mathcal{Y} = \{0, 1 \dots, m\}$, we can use the results in [14] and proceed along the lines of Section 7 to see that $\mathrm{HSIC}_{k_{\mathcal{X}},d}$ is a pivot for the locally most powerful test for $\tau^2 = 0$ in an overparametrized multinomial model of the form

$$\begin{aligned} y_i \sim \mathcal{M}\Bigg(1, &\frac{\exp^{\mu + V_0(X_i)}}{\sum_{g=0}^m \exp^{\mu + V_g(X_i)}}, \dots, \\ &\times \frac{\exp^{\mu + V_m(X_i)}}{\sum_{g=0}^m \exp^{\mu + V_g(X_i)}}\Bigg), \end{aligned} \tag{47}$$

where $\mathcal{M}(N, p_0, \dots, p_m)$ denotes the multinomial distribution with parameters $N, p_0, \dots, p_m$ and $V_1(\cdot), \dots, V_m(\cdot)$ are independent Gaussian processes with covariance function $\tau^2 k_{\mathcal{X}}$.

Useful representations for the discrete kernel with more than two elements can be traced back to the two-elements case. In particular, as noted in [14], the feature map representation $d(y_1, y_2) = (1_{\{y_1=0\}}, \dots, 1_{\{y_1=m\}})^t (1_{\{y_2=0\}}, \dots, 1_{\{y_2=m\}})$, directly implies that

$$\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d}(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{2} \sum_{j=0}^m \widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d^{(j)}}(\boldsymbol{X}, \boldsymbol{y}), \tag{48}$$

where $d^{(j)} = 1_{\{y_1=j, y_2=j\}} + 1_{\{y_1 \neq j, y_2 \neq j\}}$ is the discrete kernel on the set $\{\{j\}, \mathcal{Y} \backslash \{j\}\}$. This leads to the following corollary of Proposition 6.

COROLLARY 5. *Let $k_{\mathcal{X}}$ denote an arbitrary kernel on some set $\mathcal{X}$ and let d denote the discrete kernel on $\{0, \dots, m\}$. Also, for any $S = \{s_1, \dots, s_l\} \subset \{1, \dots, n\}$ with $s_1 < s_2 \dots < s_l$, let $X_j^S = X_{s_j}$ for $j \in \{1, \dots, l\}$ and $\boldsymbol{X}^S = (X_1^S, \dots, X_l^S)$. Moreover, for $j = 0, \dots, m$, denote $S_j = \{i \in \{1, \dots, n\} | y_i = j\}$ and $n_j = |S_j|$. Then we have*

$$\begin{aligned} &\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d}(\boldsymbol{X}, \boldsymbol{y}) \\ &\quad = \sum_{j=0}^m \frac{n_j^2 (n - n_j)^2}{n^4} \widehat{\mathrm{MMD}}_{k_{\mathcal{X}}}^2 (\boldsymbol{X}^{S_j}, \boldsymbol{X}^{\mathcal{Y} \backslash S_j}). \end{aligned} \tag{49}$$

If $n_0 = \dots = n_m$ and $k_{\mathcal{X}}$ is the distance covariance kernel, representation (49) reduces to a multiple of the DISCO statistic [45], equation (2.3). Also, if $n_0 = \dots = n_m$ and $k_{\mathcal{X}}$ is the linear kernel, $\widehat{\mathrm{HSIC}}_{k_{\mathcal{X}},d}(\boldsymbol{X}, \boldsymbol{y})$ is a multiple of the SST (sum of squared error due to treatments) in a classical ANOVA [45], p. 4. With unequal $n_0, \dots, n_m$, however, the test statistic can be markedly different from these classical tests. In particular, as can be understood from the decomposition in Section 5.3, the test puts larger weight on the outcome categories i with large sample size n_i. This is a sensible property, since there is more power to be had for these outcome categories [18]. Since the test puts weight zero to outcome categories i with $n_i = 0$, the test adapts more naturally to sparse contexts than its classical counterparts.

8.2 An Alternative to the Chi-Square Independence Test for Categorical Data

All three traditions of testing can be used to construct a test between two sets of categorical data, and the resulting test is the same. For HSIC and distance covariance, the test is derived by applying discrete kernels or distances, respectively, on both X and Y. From the global test perspective, the same test arises as a locally most powerful test in a multinomial regression model with categorical predictors, cf. the model in (47). In the remainder of this section, we will derive and analyze this test, and contrast its properties with those of the classical chi-square test.

THEOREM 6. *Let $\mathcal{X} = \{0, \dots, m\}$ and $\mathcal{Y} = \{0, \dots, r\}$ and denote the discrete kernels on $\mathcal{X}$ and $\mathcal{Y}$ by d_m and d_r, respectively. We further consider the $m \times r$ contingency table of the sample $(\boldsymbol{X}, \boldsymbol{Y})$. The entry in the (j, l)th cell will be denoted by n_{jl}, that is,*

$$n_{jl} = \sum_{i=1}^n 1_{\{X_i = j, Y_i = l\}},$$

moreover let $n_{k\cdot}$ and $n_{\cdot l}$ denote the respective row and column sums, that is,

$$n_{j\cdot} = \sum_{l=1}^n n_{jl} = \sum_{i=1}^n 1_{\{X_i = j\}},$$

$$n_{\cdot l} = \sum_{k=1}^n n_{jl} = \sum_{i=1}^n 1_{\{Y_i = l\}}.$$

Finally, let

$$n_{jl}^* = \frac{1}{n} n_{j\cdot} n_{\cdot l}$$

be the expected value of n_{jl} *under independence of* X *and* Y (*and with fixed marginal frequencies*). *Then*

$$\mathrm{HSIC}_{d_m,d_r} = \frac{1}{n^2} \sum_{j=0}^{m} \sum_{l=0}^{r} (n_{jl} - n_{jl}^*)^2. \tag{50}$$

The asymptotic distribution of the statistic in (50) is a Gaussian quadratic form with a finite number of nonzero weights, as stated by the following theorem.

THEOREM 7. *Let* $p_j = \mathbb{P}(X = j)$, $j = 0, \dots, m$ *and* $q_l = \mathbb{P}(Y = l)$, $l = 0, \dots, r$. *Also define the matrices* $L^X \in \mathbb{R}^{(m+1)\times(m+1)}$ *and* $L^Y \in \mathbb{R}^{(r+1)\times(r+1)}$ *via*

$$L_{jl}^X = p_l \left(1_{\{j=l\}} - p_j - p_l + \sum_{k=0}^{m} p_k^2 \right),$$

$$L_{jl}^Y = q_l \left(1_{\{j=l\}} - q_j - q_l + \sum_{k=0}^{r} q_k^2 \right),$$

and denote by $\lambda_0^X \geq \cdots \geq \lambda_{m-1}^X$ *and* $\lambda_0^Y \geq \cdots \geq \lambda_{r-1}^Y$ *the nonzero eigenvalues of* L^X *and* L^Y, *respectively. Then, for* $n \to \infty$,

$$n\,\widehat{\mathrm{HSIC}}_{d_m,d_r} \xrightarrow{\mathcal{D}} \sum_{j=0}^{m-1} \sum_{l=0}^{r-1} \lambda_j^X \lambda_l^Y Z_{jl}^2,$$

where the Z_{jl} *are independent standard normally distributed random variables*

We note that while a more technical formulation of Theorem 6 has been given in [38], Remark 4.12, the statement of Theorem 7 is new, to the best knowledge of the authors.

REMARK 2. We remark that the result of Theorem 7 yields a procedure for approximating p-values requiring only the calculation of the eigenvalues of two square matrices with $m + 1$ and $r + 1$ rows, respectively, instead of the calculation of the eigenvalues of two quadratic matrices with n rows. To the best knowledge of the authors, no general closed form expressions for matrices of the form of L^X (or L^Y, respectively) are known. However, assuming w.l.o.g. that $p_0 \geq \cdots \geq p_m$, it follows directly from the interlacing property in [27], Theorem 3, that, for $j = 0, \dots, m - 1$,

$$p_{j+1} \leq \lambda_j^X \leq p_j,$$

and similarly for the eigenvalues of L^Y.

Since the discussed discrete test arises naturally from all three traditions discussed in this article, it is insightful to contrast the test statistic in Theorem 7,

$$n\,\widehat{\mathrm{HSIC}}_{d_m,d_r} = \frac{1}{n} \sum_{j=0}^{m} \sum_{l=0}^{r} (n_{jl} - n_{jl}^*)^2, \tag{51}$$

with the standard chi-square test statistic for testing independence in $(m + 1) \times (r + 1)$ contingency tables, which is given by

$$\begin{aligned} \widehat{S} &= \sum_{j=0}^{m} \sum_{l=0}^{r} \frac{(n_{jl} - n_{jl}^*)^2}{n_{jl}^*} \\ &= \frac{1}{n} \sum_{j=0}^{m} \sum_{l=0}^{r} \frac{(n_{jl} - n_{jl}^*)^2}{p_{jl}^*}, \end{aligned} \tag{52}$$

where $p_{jl}^* = \frac{n_{jl}^*}{n}$ is the estimated probability of observing $(X_i, Y_i) = (j, l)$ under the null hypothesis.

A comparison between the test statistics (51) and (52) will point to general differences between the unified testing concept in this article and standard tests. The only arithmetic difference between (51) and (52) is the additional normalizing factor $(p_{jl}^*)^{-1}$ in each summand of the chi-square test statistic. The missing normalizing factor in equation (51) implies that this test gives relatively large weight to deviations in categories in which the expected count is large. More precisely, is does not down-weight such deviations like the chi-square test. The test, therefore, is relatively more directed against alternatives that are associated with deviations from the expected number of counts n_{jl}^* in cells with large n_{jl}^*, which is closely connected with the *local optimality* property underlying HSIC/generalized distance covariance/global tests, cf. also Section 5.3.

For further investigation of the alternatives against which the presented test is directed, it is useful to consider the decomposition in equation (26),

$$n\,\widehat{\mathrm{HSIC}}_{d_m,d_r} = n \sum_{j=0}^{m} \sum_{l=0}^{r} \frac{\widehat{\lambda}_j^{\boldsymbol{X}}}{n} \frac{\widehat{\lambda}_l^{\boldsymbol{Y}}}{n} (w_l^t v_j)^2, \tag{53}$$

and comparing it with the corresponding representation of the chi-square statistic,

$$\widehat{S} = n \sum_{j=0}^{m} \sum_{l=0}^{r} (w_l^t v_j)^2. \tag{54}$$

While $n\,\widehat{\mathrm{HSIC}}_{d_m,d_r}$ is specifically directed against alternatives associated with large variation in feature space, the chi-square statistic assigns equal weights to all directions. This implies lower power of the chi-square statistic for the alternatives associated with large variation in feature space, which correspond to frequent values of X and Y, respectively, and higher power for the alternatives associated with small variation in feature space, corresponding to rare values of X and Y. Comparison of equations (53) and (54) also explains the different asymptotic distributions of $\widehat{\mathrm{HSIC}}_{d_m,d_r}$ and $\widehat{S}$. Notably, the $\chi_{m\,r}^2$ limit distribution of the chi-square statistic is simply a consequence of the standardizing factors $(p_{kl}^*)^{-1}$ of the summands in (52).

The comparison above is highly analogous to the difference between the global test for linear regression and the classical F-test [18]. In that comparison, the striking property of the F-test was the invariance under affine transformations. While this is a classic (and usually considered desirable) property, it is the noninvariance of the global test, which enables the functionality in high dimensions. When contrasting now the discrete global test in equation (51) with the chi-square test, the noninvariance property in feature space of the latter does not seem particularly meaningful. The main advantage, and presumably one of the main reasons to originally construct the chi-square test in this way, is the simple asymptotic distribution of the chi-square test statistic. The discrete test in equation (51) on the other hand features the local optimality property; by focusing on cells with a high expected count, it prefers to focus on alternatives for which we can have considerable power even if the dependence between X and Y is weak. Moreover, since the test statistic does not feature the standardizing factor $(p^*_{kl})^{-1}$ in each summand, it may be expected that parametric versions of the test also perform well when some expected cell counts are low or even zero; this is not the case for the chi-square test.

In view of these advantages, the discrete global test should be considered a serious competitor for the classical chi-square test. Generalizations to ordinal data can be derived starting from [31].

9. DISCUSSION

We have established a unification of generalized distance covariance, HSIC and locally most powerful global tests that, we think, leads to a better understanding of all three theories. Most importantly, we have provided model-based interpretations of HSIC and generalized distance covariance, see Section 5. As implied by this central observation, a large new family of tests arises (Section 7) which may be very useful for applications. This family can be regarded as either an extension of generalized distance covariance and HSIC to generalized linear regression models or as a kernel version of the global test. A way of obtaining further generalizations of all three measures is to apply different weight functions in equation (26), as we discussed in Section 5.3.

Besides exploring the connection of generalized distance covariance, HSIC and global tests and extending their unified concept, we have contrasted them to classical tests such as the F-test for linear regression or the chi-square test. We found that the main difference between the two approaches is that that the classical tests satisfy an invariance property in feature space, which does not hold for generalized distance covariance, HSIC or global tests. We have also pointed out that it is precisely this lack of this invariance property that provides nontrivial power when testing in high dimensions or for general associations.

In many applications, variation is closely related to signal strength; in this context the property of affine invariance looks rather unnatural. In many applications, axes of large variation will usually be related to signal, while axes of small variations may often be explained by noise [18]. Following this argumentation, we conclude that noninvariant tests represent serious competitors to classical tests for problems where affinely invariant standard tests are routinely applied, such as the contingency table case of Section 8.2 or survival analysis [12]. We see global tests, HSIC tests and distance covariance tests as examples of a broader class of noninvariant tests; we believe that viewing them in this way provides a starting point for further exploration of their properties, both in their own right and in contrast to classical tests.

SUPPLEMENTARY MATERIAL

Supplement to "A Regression Perspective on Generalized Distance Covariance and the Hilbert–Schmidt Independence Criterion" (DOI: 10.1214/21-STS841SUPP; .pdf). Supplementary information.

REFERENCES

[1] BARINGHAUS, L. and FRANZ, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.* **88** 190–206. MR2021870 https://doi.org/10.1016/S0047-259X(03)00079-4

[2] BERG, C., CHRISTENSEN, J. P. R. and RESSEL, P. (1984). *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions. Graduate Texts in Mathematics* **100**. Springer, New York. MR0747302 https://doi.org/10.1007/978-1-4612-1128-0

[3] BERSCHNEIDER, G. and BÖTTCHER, B. (2019). On complex Gaussian random fields, Gaussian quadratic forms and sample distance multivariance. Preprint. arXiv:1808.07280.

[4] BÖTTCHER, B., KELLER-RESSEL, M. and SCHILLING, R. L. (2018). Detecting independence of random vectors: Generalized distance covariance and Gaussian covariance. *Modern Stoch. Theory Appl.* **5** 353–383. MR3868546 https://doi.org/10.15559/18-vmsta116

[5] CAI, T., TONINI, G. and LIN, X. (2011). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics* **67** 975–986. MR2829272 https://doi.org/10.1111/j.1541-0420.2010.01544.x

[6] CHATURVEDI, N., DE MENEZES, R. X. and GOEMAN, J. J. (2017). A global × global test for testing associations between two large sets of variables. *Biom. J.* **59** 145–158. MR3593726 https://doi.org/10.1002/bimj.201500106

[7] DUECK, J., EDELMANN, D., GNEITING, T. and RICHARDS, D. (2014). The affinely invariant distance correlation. *Bernoulli* **20** 2305–2330. MR3263106 https://doi.org/10.3150/13-BEJ558

[8] DUECK, J., EDELMANN, D. and RICHARDS, D. (2017). Distance correlation coefficients for Lancaster distributions. *J. Multivariate Anal.* **154** 19–39. MR3588555 https://doi.org/10.1016/j.jmva.2016.10.012

[9] EDELMANN, D., FOKIANOS, K. and PITSILLOU, M. (2019). An updated literature review of distance correlation and its applications to time series. *Int. Stat. Rev.* **87** 237–262. MR3994758 https://doi.org/10.1111/insr.12294

[10] EDELMANN, D. and GOEMAN, J. (2022). Supplement to "A Regression Perspective on Generalized Distance Covariance and the Hilbert–Schmidt Independence Criterion." https://doi.org/10.1214/21-STS841SUPP

[11] EDELMANN, D., RICHARDS, D. and VOGEL, D. (2020). The distance standard deviation. *Ann. Statist.* **48** 3395–3416. MR4185813 https://doi.org/10.1214/19-AOS1935

[12] EDELMANN, D., SAADATI, M., PUTTER, H. and GOEMAN, J. (2020). A global test for competing risks survival analysis. *Stat. Methods Med. Res.* **29** 3666–3683. MR4159219 https://doi.org/10.1177/0962280220938402

[13] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067

[14] GOEMAN, J. J. and LE CESSIE, S. (2006). A goodness-of-fit test for multinomial logistic regression. *Biometrics* **62** 980–985. MR2297668 https://doi.org/10.1111/j.1541-0420.2006.00581.x

[15] GOEMAN, J. J., OOSTING, J., CLETON-JANSEN, A.-M., ANNINGA, J. K. and VAN HOUWELINGEN, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics* **21** 1950–1957.

[16] GOEMAN, J. J., VAN DE GEER, S. A., DE KORT, F. and VAN HOUWELINGEN, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics* **20** 93–99.

[17] GOEMAN, J. J., VAN HOUWELINGEN, H. C. and FINOS, L. (2011). Testing against a high-dimensional alternative in the generalized linear model: Asymptotic type I error control. *Biometrika* **98** 381–390. MR2806435 https://doi.org/10.1093/biomet/asr016

[18] GOEMAN, J. J., VAN DE GEER, S. A. and VAN HOUWELINGEN, H. C. (2006). Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 477–493. MR2278336 https://doi.org/10.1111/j.1467-9868.2006.00551.x

[19] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2007). A kernel method for the two-sample problem. *Adv. Neural Inf. Process. Syst.* **20** 513–520.

[20] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[21] GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory. Lecture Notes in Computer Science* **3734** 63–77. Springer, Berlin. MR2255909 https://doi.org/10.1007/11564089_7

[22] GRETTON, A., FUKUMIZU, K. and SRIPERUMBUDUR, B. K. (2009). Discussion of: Brownian distance covariance [MR2752127]. *Ann. Appl. Stat.* **3** 1285–1294. MR2752132 https://doi.org/10.1214/09-AOAS312E

[23] GRETTON, A., FUKUMIZU, K., TEO, C. H., SONG, L., SCHÖLKOPF, B. and SMOLA, A. J. (2008). A kernel statistical test of independence. In *Adv. Neur. Inf. Proc. Sys.* **21** 585–592.

[24] GUO, B. and CHEN, S. X. (2016). Tests for high dimensional generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1079–1102. MR3557190 https://doi.org/10.1111/rssb.12152

[25] HEMERIK, J., GOEMAN, J. J. and FINOS, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 841–864. MR4112787 https://doi.org/10.1111/rssb.12369

[26] HOFMANN, T., SCHÖLKOPF, B. and SMOLA, A. J. (2008). Kernel methods in machine learning. *Ann. Statist.* **36** 1171–1220. MR2418654 https://doi.org/10.1214/009053607000000677

[27] HONEINE, P. (2014). An eigenanalysis of data centering in machine learning. Preprint. arXiv:1407.2904.

[28] HUA, W.-Y. and GHOSH, D. (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics* **71** 812–820. MR3402617 https://doi.org/10.1111/biom.12314

[29] HUANG, C. and HUO, X. (2017). A statistically and numerically efficient independence test based on random projections and distance covariance. Preprint. arXiv:1701.06054.

[30] JAKOBSEN, M. E. (2017). Distance Covariance in Metric Spaces: Non-Parametric Independence Testing in Metric Spaces. Master's thesis. University of Copenhagen. Available on arXiv:1706.03490.

[31] JELIZAROW, M., MANSMANN, U. and GOEMAN, J. J. (2016). A Cochran-Armitage-type and a score-free global test for multivariate ordinal data. *Stat. Med.* **35** 2754–2769. MR3513716 https://doi.org/10.1002/sim.6898

[32] KONG, J., KLEIN, B. E., KLEIN, R., LEE, K. E. and WAHBA, G. (2012). Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proc. Natl. Acad. Sci. USA* **109** 20352–20357.

[33] KONG, J., KLEIN, B. E., KLEIN, R., LEE, K. E. and WAHBA, G. (2013). Correction for Kong et al., Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proc. Natl. Acad. Sci. USA* **110** 13691–13691.

[34] LEE, A. J. (2019). *U-Statistics: Theory and Practice*. Routledge, London.

[35] LE CESSIE, S. and VAN HOUWELINGEN, H. C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics* 600–614.

[36] LIU, C. and RUBIN, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statist. Sinica* **5** 19–39. MR1329287

[37] LOÈVE, M. (1978). *Probability Theory. II*, 4th ed. *Graduate Texts in Mathematics*, *Vol.* 46. Springer, New York. MR0651018

[38] LYONS, R. (2013). Distance covariance in metric spaces. *Ann. Probab.* **41** 3284–3305. MR3127883 https://doi.org/10.1214/12-AOP803

[39] MARTÍNEZ-GÓMEZ, E., RICHARDS, M. T. and RICHARDS, D. S. P. (2014). Distance correlation methods for discovering associations in large astrophysical databases. *Astrophys. J.* **781** 39 (11pp).

[40] MINH, H. Q., NIYOGI, P. and YAO, Y. (2006). Mercer's theorem, feature maps, and smoothing. In *Learning Theory. Lecture Notes in Computer Science* **4005** 154–168. Springer, Berlin. MR2280604 https://doi.org/10.1007/11776420_14

[41] PAN, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* **33** 497–507.

[42] PFISTER, N., BÜHLMANN, P., SCHÖLKOPF, B. and PETERS, J. (2018). Kernel-based tests for joint independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 5–31. MR3744710 https://doi.org/10.1111/rssb.12235

[43] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

[44] RICHARDS, M. T., RICHARDS, D. S. P. and MARTÍNEZ-GÓMEZ, E. (2014). Interpreting the distance correlation results for the COMBO-17 survey. *Astrophys. J. Lett.* **784** L34 (5 pp.).

[45] Rizzo, M. L. and Székely, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.* **4** 1034–1055. MR2758432 https://doi.org/10.1214/09-AOAS245

[46] Robbins, H. and Pitman, E. J. G. (1949). Application of the method of mixtures to quadratic forms in normal variates. *Ann. Math. Stat.* **20** 552–560. MR0032151 https://doi.org/10.1214/aoms/1177729947

[47] Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **25** 257–265. MR0440801 https://doi.org/10.2307/2347233

[48] Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection. SLSFS* 2005 (C. Saunders, M. Grobelnik, S. Gunn and J. Shawe-Taylor, eds.). *Lecture Notes in Computer Science* **3940** 34–51. Springer, Berlin.

[49] Rosipal, R. and Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Mach. Learn. Res.* **2** 97–123.

[50] Schölkopf, B., Smola, A. and Müller, K.-R. (1997). Kernel principal component analysis. In *Artificial Neural Networks – ICANN'97. ICANN* 1997 (W. Gerstner, A. Germond, M. Hasler and J. D. Nicoud, eds.). *Lecture Notes in Computer Science* **1327** 583–588. Springer, Berlin.

[51] Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41** 2263–2291. MR3127866 https://doi.org/10.1214/13-AOS1140

[52] Shen, C., Priebe, C. E. and Vogelstein, J. T. (2019). The exact equivalence of independence testing and two-sample testing. Preprint. arXiv:1910.08883.

[53] Shen, C. and Vogelstein, J. T. (2021). The exact equivalence of distance and kernel methods in hypothesis testing. *AStA Adv. Stat. Anal.* **105** 385–403. MR4304315 https://doi.org/10.1007/s10182-020-00378-1

[54] Solari, A., Finos, L. and Goeman, J. J. (2014). Rotation-based multiple testing in the multivariate linear model. *Biometrics* **70** 954–961. MR3295756 https://doi.org/10.1111/biom.12238

[55] Solari, A., le Cessie, S. and Goeman, J. J. (2012). Testing goodness of fit in regression: A general approach for specified alternatives. *Stat. Med.* **31** 3656–3666. MR3041837 https://doi.org/10.1002/sim.5417

[56] Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99** 386–402. MR2396970 https://doi.org/10.1016/j.jmva.2006.11.002

[57] Steinwart, I. and Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constr. Approx.* **35** 363–417. MR2914365 https://doi.org/10.1007/s00365-012-9153-3

[58] Sun, H. (2005). Mercer theorem for RKHS on noncompact sets. *J. Complexity* **21** 337–349. MR2138444 https://doi.org/10.1016/j.jco.2004.09.002

[59] Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127 https://doi.org/10.1214/09-AOAS312

[60] Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Multivariate Anal.* **117** 193–213. MR3053543 https://doi.org/10.1016/j.jmva.2013.02.012

[61] Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference* **143** 1249–1272. MR3055745 https://doi.org/10.1016/j.jspi.2013.03.018

[62] Székely, G. J. and Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities. *Ann. Statist.* **42** 2382–2412. MR3269983 https://doi.org/10.1214/14-AOS1255

[63] Szekely, G. J. and Rizzo, M. L. (2017). The energy of data. *Annu. Rev. Stat. Appl.* **4** 447–479.

[64] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665 https://doi.org/10.1214/009053607000000505

[65] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[66] Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H. (2015). Conditional distance correlation. *J. Amer. Statist. Assoc.* **110** 1726–1734. MR3449068 https://doi.org/10.1080/01621459.2014.993081

[67] Wold, S., Sjöström, M. and Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58** 109–130.

[68] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

[69] Zhang, K., Peters, J., Janzing, D. and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. UAI'11* 804–813. AUAI Press, Arlington, VA, USA.

[70] Zhang, Z., Dai, G., Wang, D. and Jordan, M. I. (2010). Bayesian generalized kernel models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 972–979.

[71] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x