The Annals of Statistics
2013, Vol. 41, No. 5, 2263–2291
DOI: 10.1214/13-AOS1140

EQUIVALENCE OF DISTANCE-BASED AND RKHS-BASED STATISTICS IN HYPOTHESIS TESTING

BY DINO SEJDINOVIC, BHARATH SRIPERUMBUDUR, ARTHUR GRETTON AND KENJI FUKUMIZU

University College London, University of Cambridge, University College London and Max Planck Institute for Intelligent Systems, Tübingen, and Institute of Statistical Mathematics, Tokyo

We provide a unifying framework linking two classes of statistics used in two-sample and independence testing: on the one hand, the energy distances and distance covariances from the statistics literature; on the other, maximum mean discrepancies (MMD), that is, distances between embeddings of distributions to reproducing kernel Hilbert spaces (RKHS), as established in machine learning. In the case where the energy distance is computed with a semimetric of negative type, a positive definite kernel, termed distance kernel, may be defined such that the MMD corresponds exactly to the energy distance. Conversely, for any positive definite kernel, we can interpret the MMD as energy distance with respect to some negative-type semimetric. This equivalence readily extends to distance covariance using kernels on the product space. We determine the class of probability distributions for which the test statistics are consistent against all alternatives. Finally, we investigate the performance of the family of distance kernels in two-sample and independence tests: we show in particular that the energy distance most commonly employed in statistics is just one member of a parametric family of kernels, and that other choices from this family can yield more powerful tests.

1. Introduction. The problem of testing statistical hypotheses in high dimensional spaces is particularly challenging, and has been a recent focus of considerable work in both the statistics and the machine learning communities. On the statistical side, two-sample testing in Euclidean spaces (of whether two independent samples are from the same distribution, or from different distributions) can be accomplished using a so-called energy distance as a statistic [Székely and Rizzo (2004, 2005), Baringhaus and Franz (2004)]. Such tests are consistent against all alternatives as long as the random variables have finite first moments. A related dependence measure between vectors of high dimension is the distance covariance [Székely, Rizzo and Bakirov (2007), Székely and Rizzo (2009)], and the resulting test is again consistent for variables with bounded first moment. The distance covariance has had a major impact in the statistics community, with Székely

Received July 2012; revised March 2013.
MSC2010 subject classifications. Primary 62G10, 62H20, 68Q32; secondary 46E22.
Key words and phrases. Reproducing kernel Hilbert spaces, distance covariance, two-sample testing, independence testing.

and Rizzo (2009) being accompanied by an editorial introduction and discussion. A particular advantage of energy distance-based statistics is their compact representation in terms of certain expectations of pairwise Euclidean distances, which leads to straightforward empirical estimates. As a follow-up work, Lyons (2013) generalized the notion of distance covariance to metric spaces of negative type (of which Euclidean spaces are a special case).

On the machine learning side, two-sample tests have been formulated based on embeddings of probability distributions into reproducing kernel Hilbert spaces [Gretton et al. (2007, 2012a)], using as the test statistic the difference between these embeddings: this statistic is called the maximum mean discrepancy (MMD). This distance measure was also applied to the problem of testing for independence, with the associated test statistic being the Hilbert–Schmidt independence criterion (HSIC) [Gretton et al. (2005, 2008), Smola et al. (2007), Zhang et al. (2011)]. Both tests are shown to be consistent against all alternatives when a characteristic RKHS is used [Fukumizu et al. (2009), Sriperumbudur et al. (2010)].

Despite their striking similarity, the link between energy distance-based tests and kernel-based tests has been an open question. In the discussion of [Székely and Rizzo (2009), Gretton, Fukumizu and Sriperumbudur (2009), page 1289] first explored this link in the context of independence testing, and found that interpreting the distance-based independence statistic as a kernel statistic is not straightforward, since Bochner's theorem does not apply to the choice of weight function used in the definition of the distance covariance (we briefly review this argument in Section 5.3). Székely and Rizzo (2009), Rejoinder, page 1303, confirmed that the link between RKHS-based dependence measures and the distance covariance remained to be established, because the weight function is not integrable. Our contribution resolves this question, and shows that RKHS-based dependence measures are precisely the formal extensions of the distance covariance, where the problem of nonintegrability of weight functions is circumvented by using translation-variant kernels, that is, *distance-induced kernels*, introduced in Section 4.1.

In the case of two-sample testing, we demonstrate that energy distances are in fact maximum mean discrepancies arising from the same family of distance-induced kernels. A number of interesting consequences arise from this insight: first, as the energy distance (and distance covariance) derives from a particular choice of a kernel, we can consider analogous quantities arising from other kernels, and yielding more sensitive tests. Second, in relation to Lyons (2013), we obtain a new family of characteristic kernels arising from general semimetric spaces of negative type, which are quite unlike the characteristic kernels defined via Bochner's theorem [Sriperumbudur et al. (2010)]. Third, results from [Gretton et al. (2009), Zhang et al. (2011)] may be applied to obtain consistent two-sample and independence tests for the energy distance, without using bootstrap, which perform much better than the upper bound proposed by Székely, Rizzo and Bakirov (2007) as an alternative to the bootstrap.

In addition to the energy distance and maximum mean discrepancy, there are other well-known discrepancy measures between two probability distributions, such as the Kullback–Leibler divergence, Hellinger distance and total variation distance, which belong to the class of f-divergences. Another popular family of distance measures on probabilities is the integral probability metric [Müller (1997)], examples of which include the Wasserstein distance, Dudley metric and Fortet–Mourier metric. Sriperumbudur et al. (2012) showed that MMD is an integral probability metric and so is energy distance, owing to the equality (between energy distance and MMD) that we establish in this paper. On the other hand, Sriperumbudur et al. (2012) also showed that MMD (and therefore the energy distance) is not an f-divergence, by establishing the total variation distance as the only discrepancy measure that is both an IPM and f-divergence.

The equivalence established in this paper has two major implications for practitioners using the energy distance or distance covariance as test statistics. First, it shows that these quantities are members of a much broader class of statistics, and that by choosing an alternative semimetric/kernel to define a statistic from this larger family, one may obtain a more sensitive test than by using distances alone. Second, it shows that the principles of energy distance and distance covariance are readily generalized to random variables that take values in general topological spaces. Indeed, kernel tests are readily applied to structured and non-Euclidean domains, such as text strings, graphs and groups [Fukumizu et al. (2009)].

The structure of the paper is as follows: in Section 2, we introduce semimetrics of negative type, and extend the notions of energy distance and distance covariance to semimetric spaces of negative type. In Section 3, we provide the necessary definitions from RKHS theory and give a review of the maximum mean discrepancy (MMD) and the Hilbert–Schmidt independence criterion (HSIC), the RKHS-based statistics used for two-sample and independence testing, respectively. In Section 4, the correspondence between positive definite kernels and semimetrics of negative type is developed, and it is applied in Section 5 to show the equivalence between a (generalized) energy distance and MMD (Section 5.1), as well as between a (generalized) distance covariance and HSIC (Section 5.2). We give conditions for these quantities to distinguish between probability measures in Section 6, thus obtaining a new family of characteristic kernels. Empirical estimates of these quantities and associated two-sample and independence tests are described in Section 7. Finally, in Section 8, we investigate the performance of the test statistics on a variety of testing problems.

This paper extends the conference publication [Sejdinovic et al. (2012)], and gives a detailed technical discussion and proofs which were omitted in that work.

2. Distance-based approach. This section reviews the distance-based approach to two-sample and independence testing, in its general form. The generalized energy distance and distance covariance are defined.

2.1. *Semimetrics of negative type.* We will work with the notion of a semimetric of negative type on a nonempty set $\mathcal{Z}$, where the "distance" function need not satisfy the triangle inequality. Note that this notion of semimetric is different to that which arises from the seminorm (also called the pseudonorm), where the distance between two distinct points can be zero.

DEFINITION 1 (Semimetric). Let $\mathcal{Z}$ be a nonempty set and let $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ be a function such that $\forall z, z' \in \mathcal{Z}$,

1. $\rho(z, z') = 0$ if and only if $z = z'$, and
2. $\rho(z, z') = \rho(z', z)$.

Then $(\mathcal{Z}, \rho)$ is said to be a semimetric space and ρ is called a semimetric on $\mathcal{Z}$.

DEFINITION 2 (Negative type). The semimetric space $(\mathcal{Z}, \rho)$ is said to have negative type if $\forall n \geq 2$, $z_1, \dots, z_n \in \mathcal{Z}$, and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^n \alpha_i = 0$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0. \tag{2.1}$$

Note that in the terminology of Berg, Christensen and Ressel (1984), ρ satisfying (2.1) is said to be a *negative definite* function. The following proposition is derived from Berg, Christensen and Ressel (1984), Corollary 2.10, page 78, and Proposition 3.2, page 82.

PROPOSITION 3.

1. *If ρ satisfies* (2.1), *then so does ρ^q, for $0 < q < 1$.*
2. *ρ is a semimetric of negative type if and only if there exists a Hilbert space $\mathcal{H}$ and an injective map $\varphi : \mathcal{Z} \to \mathcal{H}$, such that*

$$\rho(z, z') = \|\varphi(z) - \varphi(z')\|_{\mathcal{H}}^2. \tag{2.2}$$

The second part of the proposition shows that $(\mathbb{R}^d, \|\cdot - \cdot\|^2)$ is of negative type, and by taking $q = 1/2$ in the first part, we conclude that all Euclidean spaces are of negative type. In addition, whenever ρ is a semimetric of negative type, $\rho^{1/2}$ is a metric of negative type, that is, even though ρ may not satisfy the triangle inequality, its square root must do if it obeys (2.1).

2.2. *Energy distance.* Unless stated otherwise, we will assume that $\mathcal{Z}$ is any topological space on which Borel measures can be defined. We will denote by $\mathcal{M}(\mathcal{Z})$ the set of all finite signed Borel measures on $\mathcal{Z}$, and by $\mathcal{M}_+^1(\mathcal{Z})$ the set of all Borel probability measures on $\mathcal{Z}$.

The energy distance was introduced by Székely and Rizzo (2004, 2005) and independently by Baringhaus and Franz (2004) as a measure of statistical distance between two probability measures P and Q on $\mathbb{R}^d$ *with finite first moments*, given by

$$D_E(P, Q) = 2\mathbb{E}_{ZW}\|Z - W\|_2 - \mathbb{E}_{ZZ'}\|Z - Z'\|_2 - \mathbb{E}_{WW'}\|W - W'\|_2, \tag{2.3}$$

where $Z, Z' \overset{\text{i.i.d.}}{\sim} P$ and $W, W' \overset{\text{i.i.d.}}{\sim} Q$. The moment condition is required to ensure that the expectations in (2.3) is finite. $D_E(P, Q)$ is always nonnegative, and is strictly positive if $P \neq Q$. In scalar case, it coincides with twice the Cramér–Von Mises distance.

Following Lyons (2013), the notion can be generalized to a metric space of negative type, which we further extend to semimetrics. Before we proceed, we need to first introduce a moment condition w.r.t. a semimetric ρ.

DEFINITION 4. For $\theta > 0$, we say that $\nu \in \mathcal{M}(\mathcal{Z})$ has a finite θ-moment with respect to a semimetric ρ of negative type if there exists $z_0 \in \mathcal{Z}$, such that $\int \rho^\theta(z, z_0)\, d|\nu|(z) < \infty$. We denote

$$\mathcal{M}_\rho^\theta(\mathcal{Z}) = \left\{\nu \in \mathcal{M}(\mathcal{Z}) : \exists z_0 \in \mathcal{Z} \text{ s.t. } \int \rho^\theta(z, z_0)\, d|\nu|(z) < \infty\right\}. \tag{2.4}$$

We are now ready to introduce a general energy distance $D_{E,\rho}$.

DEFINITION 5. Let $(\mathcal{Z}, \rho)$ be a semimetric space of negative type, and let $P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_\rho^1(\mathcal{Z})$. The energy distance between P and Q, w.r.t. ρ is

$$D_{E,\rho}(P, Q) = 2\mathbb{E}_{ZW}\rho(Z, W) - \mathbb{E}_{ZZ'}\rho(Z, Z') - \mathbb{E}_{WW'}\rho(W, W'), \tag{2.5}$$

where $Z, Z' \overset{\text{i.i.d.}}{\sim} P$ and $W, W' \overset{\text{i.i.d.}}{\sim} Q$.

If ρ is a metric, as in [Lyons (2013)], the moment condition $P, Q \in \mathcal{M}_\rho^1(\mathcal{Z})$ is easily seen to be sufficient for the existence of the expectations in (2.5). Namely, if we take $z_0, w_0 \in \mathcal{Z}$ such that $\mathbb{E}_Z\rho(Z, z_0) < \infty$, $\mathbb{E}_W\rho(W, w_0) < \infty$, then the triangle inequality implies:

$$\mathbb{E}_{ZW}\rho(Z, W) \leq \mathbb{E}_Z\rho(Z, z_0) + \mathbb{E}_W\rho(W, w_0) + \rho(z_0, w_0) < \infty.$$

If ρ is a general semimetric, however, a different line of reasoning is needed, and we will come back to this condition in Remark 21, where its sufficiency will become clear using the link between positive definite kernels and negative-type semimetrics established in Section 4.

Note that the energy distance can equivalently be represented in the integral form,

$$D_{E,\rho}(P, Q) = -\int \rho\, d([P - Q] \times [P - Q]), \tag{2.6}$$

whereby the negative type of ρ implies the nonnegativity of $D_{E,\rho}$, as discussed by Lyons [(2013), page 10].

2.3. *Distance covariance*. A related notion to the energy distance is that of distance covariance, which measures dependence between random variables. Let X be a random vector on $\mathbb{R}^p$ and Y a random vector on $\mathbb{R}^q$. The distance covariance was introduced by Székely, Rizzo and Bakirov (2007), Székely and Rizzo (2009) to address the problem of testing and measuring dependence between X and Y in terms of a weighted L_2-distance between characteristic functions of the joint distribution of X and Y and the product of their marginals. As a particular choice of weight function is used (we discuss this further in Section 5.3), it can be computed in terms of certain expectations of pairwise Euclidean distances,

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|X-X'\|_2\|Y-Y'\|_2 \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}\|X-X'\|_2\mathbb{E}_Y\mathbb{E}_{Y'}\|Y-Y'\|_2 \\ &\quad - 2\mathbb{E}_{XY}[\mathbb{E}_{X'}\|X-X'\|_2\mathbb{E}_{Y'}\|Y-Y'\|_2],\end{aligned}\tag{2.7}$$

where (X,Y) and (X',Y') are $\overset{\text{i.i.d.}}{\sim} P_{XY}$. As in the case of the energy distance, Lyons (2013) established that the generalization of the distance covariance is possible to metric spaces of negative type. We extend this notion to semimetric spaces of negative type.

DEFINITION 6. Let $(\mathcal{X},\rho_{\mathcal{X}})$ and $(\mathcal{Y},\rho_{\mathcal{Y}})$ be semimetric spaces of negative type, and let $X \sim P_X \in \mathcal{M}^2_{\rho_{\mathcal{X}}}(\mathcal{X})$ and $Y \sim P_Y \in \mathcal{M}^2_{\rho_{\mathcal{Y}}}(\mathcal{Y})$, having joint distribution P_{XY}. The generalized distance covariance of X and Y is

$$\begin{aligned}\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}(X,X')\rho_{\mathcal{Y}}(Y,Y') \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}\rho_{\mathcal{X}}(X,X')\mathbb{E}_Y\mathbb{E}_{Y'}\rho_{\mathcal{Y}}(Y,Y') \\ &\quad - 2\mathbb{E}_{XY}[\mathbb{E}_{X'}\rho_{\mathcal{X}}(X,X')\mathbb{E}_{Y'}\rho_{\mathcal{Y}}(Y,Y')].\end{aligned}\tag{2.8}$$

As with the energy distance, the moment conditions ensure that the expectations are finite (which can be seen using the Cauchy–Schwarz inequality). Equivalently, the generalized distance covariance can be represented in integral form,

$$\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) = \int \rho_{\mathcal{X}}\rho_{\mathcal{Y}}\, d([P_{XY}-P_XP_Y]\times[P_{XY}-P_XP_Y]),\tag{2.9}$$

where $\rho_{\mathcal{X}}\rho_{\mathcal{Y}}$ is viewed as a function on $(\mathcal{X}\times\mathcal{Y})\times(\mathcal{X}\times\mathcal{Y})$. Furthermore, Lyons (2013), Theorem 3.20, shows that distance covariance in a metric space characterizes independence [i.e., $\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y)=0$ if and only if X and Y are independent] if the metrics $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$ satisfy an additional property, termed *strong negative type*. The discussion of this property is relegated to Section 6.

REMARK 7. While the form of (2.6) and (2.9) suggests that the energy distance and the distance covariance are closely related, it is not clear whether $\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y)$ is simply $D_{E,\tilde{\rho}}(P_{XY},P_XP_Y)$ for some semimetric $\tilde{\rho}$ on $\mathcal{X}\times\mathcal{Y}$. In particular, $-\rho_{\mathcal{X}}\rho_{\mathcal{Y}}$ is certainly not a semimetric. This question will be resolved in Corollary 26.

3. Kernel-based approach. In this section, we introduce concepts and notation required to understand reproducing kernel Hilbert spaces (Section 3.1), and distribution embeddings into RKHS. We then introduce the maximum mean discrepancy (MMD) and Hilbert–Schmidt independence criterion (HSIC).

3.1. *RKHS and kernel embeddings.* We begin with the definition of a reproducing kernel Hilbert space (RKHS).

DEFINITION 8 (RKHS). Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall z \in \mathcal{Z}, k(\cdot, z) \in \mathcal{H}$, and
2. $\forall z \in \mathcal{Z}, \forall f \in \mathcal{H}, \langle f, k(\cdot, z)\rangle_{\mathcal{H}} = f(z)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space* (RKHS).

According to the Moore–Aronszajn theorem [Berlinet and Thomas-Agnan (2004), page 19], for every symmetric, positive definite function (henceforth *kernel*) $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, there is an associated RKHS $\mathcal{H}_k$ of real-valued functions on $\mathcal{Z}$ with reproducing kernel k. The map $\varphi : \mathcal{Z} \to \mathcal{H}_k$, $\varphi : z \mapsto k(\cdot, z)$ is called the canonical feature map or the Aronszajn map of k. We will say that k is a *nondegenerate kernel* if its Aronszajn map is injective. The notion of feature map can be extended to kernel embeddings of finite signed Borel measures on $\mathcal{Z}$ [Smola et al. (2007), Sriperumbudur et al. (2010), Berlinet and Thomas-Agnan (2004), Chapter 4].

DEFINITION 9 (Kernel embedding). Let k be a kernel on $\mathcal{Z}$, and $\nu \in \mathcal{M}(\mathcal{Z})$. The *kernel embedding* of ν into the RKHS $\mathcal{H}_k$ is $\mu_k(\nu) \in \mathcal{H}_k$ such that $\int f(z)\, d\nu(z) = \langle f, \mu_k(\nu)\rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k$.

Alternatively, the kernel embedding can be defined by the Bochner integral $\mu_k(\nu) = \int k(\cdot, z)\, d\nu(z)$. If a measurable kernel k is a bounded function, $\mu_k(\nu)$ exists for all $\nu \in \mathcal{M}(\mathcal{Z})$. On the other hand, if k is not bounded, there will always exist $\nu \in \mathcal{M}(\mathcal{Z})$, for which $\int k(\cdot, z)\, d\nu(z)$ diverges. The kernels we will consider in this paper will be continuous, and hence measurable, but unbounded, so kernel embeddings will not be defined for some finite signed measures. Thus, we need to restrict our attention to a particular class of measures for which kernel embeddings exist (this will be later shown to reflect the condition that random variables considered in distance covariance tests must have finite moments). Let k be a measurable kernel on $\mathcal{Z}$, and denote, for $\theta > 0$,

$$\mathcal{M}_k^{\theta}(\mathcal{Z}) = \left\{\nu \in \mathcal{M}(\mathcal{Z}) : \int k^{\theta}(z, z)\, d|\nu|(z) < \infty\right\}. \tag{3.1}$$

Clearly,

$$\theta_1 \le \theta_2 \quad \Rightarrow \quad \mathcal{M}_k^{\theta_2}(\mathcal{Z}) \subseteq \mathcal{M}_k^{\theta_1}(\mathcal{Z}). \tag{3.2}$$

Note that the kernel embedding $\mu_k(\nu)$ is well defined $\forall \nu \in \mathcal{M}_k^{1/2}(\mathcal{Z})$, by the Riesz representation theorem.

3.2. *Maximum mean discrepancy.* As we have seen, kernel embeddings of Borel probability measures in $\mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^{1/2}(\mathcal{Z})$ do exist, and we can introduce the notion of distance between Borel probability measures in this set using the Hilbert space distance between their embeddings.

DEFINITION 10 (Maximum mean discrepancy). Let k be a kernel on $\mathcal{Z}$, and let $P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^{1/2}(\mathcal{Z})$. The *maximum mean discrepancy* (*MMD*) γ_k between P and Q is given by Gretton et al. (2012a), Lemma 4,

$$\gamma_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}.$$

The following alternative representation of the squared MMD [from Gretton et al. (2012a), Lemma 6] will be useful

$$\begin{aligned}\gamma_k^2(P, Q) &= \mathbb{E}_{ZZ'}k(Z, Z') + \mathbb{E}_{WW'}k(W, W') - 2\mathbb{E}_{ZW}k(Z, W)\\ &= \int\int k\, d([P - Q] \times [P - Q]),\end{aligned} \tag{3.3}$$

where $Z, Z' \overset{\text{i.i.d.}}{\sim} P$ and $W, W' \overset{\text{i.i.d.}}{\sim} Q$. If the restriction of μ_k to some $\mathcal{P}(\mathcal{Z}) \subseteq \mathcal{M}_+^1(\mathcal{Z})$ is well defined and injective, then k is said to be characteristic to $\mathcal{P}(\mathcal{Z})$, and it is said to be characteristic (without further qualification) if it is characteristic to $\mathcal{M}_+^1(\mathcal{Z})$. When k is characteristic, γ_k is a metric on the entire $\mathcal{M}_+^1(\mathcal{Z})$, that is, $\gamma_k(P, Q) = 0$ iff $P = Q$, $\forall P, Q \in \mathcal{M}_+^1(\mathcal{Z})$. Conditions under which kernels are characteristic have been studied by Fukumizu et al. (2009), Sriperumbudur et al. (2008), Sriperumbudur et al. (2010). An alternative interpretation of (3.3) is as an integral probability metric [Müller (1997)],

$$\gamma_k(P, Q) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \le 1} [\mathbb{E}_{Z \sim P} f(Z) - \mathbb{E}_{W \sim Q} f(W)]. \tag{3.4}$$

See Gretton et al. (2012a) and Sriperumbudur et al. (2012) for details.

3.3. *Hilbert–Schmidt independence criterion* (*HSIC*). The MMD can be employed to measure statistical dependence between random variables [Gretton et al. (2005, 2008), Gretton and Györfi (2010), Smola et al. (2007), Zhang et al. (2011)]. Let $\mathcal{X}$ and $\mathcal{Y}$ be two nonempty topological spaces and let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be kernels on

$\mathcal{X}$ and $\mathcal{Y}$, with respective RKHSs $\mathcal{H}_{k_\mathcal{X}}$ and $\mathcal{H}_{k_\mathcal{Y}}$. Then, by applying Steinwart and Christmann [(2008), Lemma 4.6, page 114],

$$k\big((x,y),(x',y')\big)=k_\mathcal{X}(x,x')k_\mathcal{Y}(y,y') \tag{3.5}$$

is a kernel on the product space $\mathcal{X}\times\mathcal{Y}$ with RKHS $\mathcal{H}_k$ isometrically isomorphic to the tensor product $\mathcal{H}_{k_\mathcal{X}}\otimes\mathcal{H}_{k_\mathcal{Y}}$.

DEFINITION 11. Let $X\sim P_X$ and $Y\sim P_Y$ be random variables on $\mathcal{X}$ and $\mathcal{Y}$, respectively, having joint distribution P_{XY}. Furthermore, let k be a kernel on $\mathcal{X}\times\mathcal{Y}$, given in (3.5). The Hilbert–Schmidt independence criterion (HSIC) of X and Y is the MMD γ_k between the joint distribution P_{XY} and the product of its marginals P_XP_Y.

Following Smola et al. (2007), Section 2.3, we can expand HSIC as

$$\begin{aligned}
&\gamma_k^2(P_{XY},P_XP_Y)\\
&\quad=\big\|\mathbb{E}_{XY}\big[k_\mathcal{X}(\cdot,X)\otimes k_\mathcal{Y}(\cdot,Y)\big]-\mathbb{E}_Xk_\mathcal{X}(\cdot,X)\otimes\mathbb{E}_Yk_\mathcal{Y}(\cdot,Y)\big\|^2_{\mathcal{H}_{k_\mathcal{X}}\otimes\mathcal{H}_{k_\mathcal{Y}}}\\
&\quad=\mathbb{E}_{XY}\mathbb{E}_{X'Y'}k_\mathcal{X}(X,X')k_\mathcal{Y}(Y,Y')+\mathbb{E}_X\mathbb{E}_{X'}k_\mathcal{X}(X,X')\mathbb{E}_Y\mathbb{E}_{Y'}k_\mathcal{Y}(Y,Y')\\
&\qquad-2\mathbb{E}_{X'Y'}\big[\mathbb{E}_Xk_\mathcal{X}(X,X')\mathbb{E}_Yk_\mathcal{Y}(Y,Y')\big].
\end{aligned}\tag{3.6}$$

It can be shown that this quantity is equal to the squared Hilbert–Schmidt norm of the covariance operator between RKHSs [Gretton et al. (2005)]. We claim that $\gamma_k^2(P_{XY},P_XP_Y)$ is well defined as long as $P_X\in\mathcal{M}^1_{k_\mathcal{X}}(\mathcal{X})$ and $P_Y\in\mathcal{M}^1_{k_\mathcal{Y}}(\mathcal{Y})$. Indeed, this is a sufficient condition for $\mu_k(P_{XY})$ to exist, since it implies that $P_{XY}\in\mathcal{M}^{1/2}_k(\mathcal{X}\times\mathcal{Y})$, which can be seen from the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\int k^{1/2}\big((x,y),(x,y)\big)\,dP_{XY}(x,y)\\
&\quad=\int k^{1/2}_\mathcal{X}(x,x)k^{1/2}_\mathcal{Y}(y,y)\,dP_{XY}(x,y)\\
&\quad\le\left(\int k_\mathcal{X}(x,x)\,dP_X(x)\int k_\mathcal{Y}(y,y)\,dP_Y(y)\right)^{1/2}.
\end{aligned}$$

Furthermore, the embedding $\mu_k(P_XP_Y)$ of the product of marginals also exists, as it can be identified with the tensor product $\mu_{k_\mathcal{X}}(P_X)\otimes\mu_{k_\mathcal{Y}}(P_Y)$, where $\mu_{k_\mathcal{X}}(P_X)$ exists since $P_X\in\mathcal{M}^1_{k_\mathcal{X}}(\mathcal{X})\subset\mathcal{M}^{1/2}_{k_\mathcal{X}}(\mathcal{X})$, and $\mu_{k_\mathcal{Y}}(P_Y)$ exists since $P_Y\in\mathcal{M}^1_{k_\mathcal{Y}}(\mathcal{Y})\subset\mathcal{M}^{1/2}_{k_\mathcal{Y}}(\mathcal{Y})$.

4. Correspondence between kernels and semimetrics. In this section, we develop the correspondence of semimetrics of negative type (Section 2.1) to the RKHS theory, that is, to symmetric positive definite kernels. This correspondence will be key to proving the equivalence between the energy distance and MMD, and the equivalence between distance covariance and HSIC in Section 5.

4.1. *Distance-induced kernels*. Semimetrics of negative type and symmetric positive definite kernels are closely related, as summarized in the following lemma, adapted from Berg, Christensen and Ressel (1984), Lemma 2.1, page 74.

LEMMA 12. *Let $\mathcal{Z}$ be a nonempty set, and $\rho : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a semimetric on $\mathcal{Z}$. Let $z_0 \in \mathcal{Z}$, and denote $k(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z')$. Then k is positive definite if and only if ρ satisfies* (2.1).

As a consequence, $k(z, z')$ defined above is a valid kernel on $\mathcal{Z}$ whenever ρ is a semimetric of negative type. For convenience, we will work with such kernels scaled by $1/2$.

DEFINITION 13 (Distance-induced kernel). Let ρ be a semimetric of negative type on $\mathcal{Z}$ and let $z_0 \in \mathcal{Z}$. The kernel

$$k(z, z') = \tfrac{1}{2}\left[\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')\right] \tag{4.1}$$

is said to be the *distance-induced kernel* induced by ρ and *centred* at z_0.

For brevity, we will drop "induced" hereafter, and say that k is simply the *distance kernel* (with some abuse of terminology). Note that distance kernels are not strictly positive definite, that is, it is not true that $\forall n \in \mathbb{N}$, and for distinct $z_1, \dots, z_n \in \mathcal{Z}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(z_i, z_j) = 0 \quad \Rightarrow \quad \alpha_i = 0\ \forall i.$$

Indeed, if k were given by (4.1), it would suffice to take $n = 1$, since $k(z_0, z_0) = 0$. By varying the point at the center z_0, we obtain a family

$$\mathcal{K}_\rho = \left\{\tfrac{1}{2}\left[\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')\right]\right\}_{z_0 \in \mathcal{Z}}$$

of distance kernels induced by ρ. The following proposition follows readily from the definition of $\mathcal{K}_\rho$ and shows that one can always express (2.2) from Proposition 3 in terms of the canonical feature map for the RKHS $\mathcal{H}_k$.

PROPOSITION 14. *Let $(\mathcal{Z}, \rho)$ be a semimetric space of negative type, and $k \in \mathcal{K}_\rho$. Then*:

1. $\rho(z, z') = k(z, z) + k(z', z') - 2k(z, z') = \|k(\cdot, z) - k(\cdot, z')\|^2_{\mathcal{H}_k}$.
2. *k is nondegenerate, that is, the Aronszajn map $z \mapsto k(\cdot, z)$ is injective.*

EXAMPLE 15. Let $\mathcal{Z} \subseteq \mathbb{R}^d$ and write $\rho_q(z, z') = \|z - z'\|^q$. By Proposition 3, ρ_q is a valid semimetric of negative type for $0 < q \le 2$. The corresponding kernel centered at $z_0 = 0$ is given by the covariance function of the fractional Brownian motion,

$$k_q(z, z') = \tfrac{1}{2}\left(\|z\|^q + \|z'\|^q - \|z - z'\|^q\right). \tag{4.2}$$

Note that while Lyons [(2013), page 9] also uses the results in Proposition 3 to characterize metrics of negative type using embeddings to general Hilbert spaces, the relation with the theory of reproducing kernel Hilbert spaces is not exploited in his work.

4.2. *Semimetrics generated by kernels.* We now further develop the link between semimetrics of negative type and kernels. We start with a simple corollary of Proposition 3.

COROLLARY 16. *Let k be any nondegenerate kernel on $\mathcal{Z}$. Then,*

$$\rho(z, z') = k(z, z) + k(z', z') - 2k(z, z') \tag{4.3}$$

defines a valid semimetric ρ of negative type on $\mathcal{Z}$.

DEFINITION 17 (Equivalent kernels). Whenever the kernel k and semimetric ρ satisfy (4.3), we will say that k *generates* ρ. If two kernels generate the same semimetric, we will say that they are *equivalent kernels*.

It is clear that every distance kernel $\tilde{k} \in \mathcal{K}_\rho$ induced by ρ, also generates ρ. However, there are many other kernels that generate ρ. The following proposition is straightforward to show and gives a condition under which two kernels are equivalent.

PROPOSITION 18. *Let k and $\tilde{k}$ be two kernels on $\mathcal{Z}$. k and $\tilde{k}$ are equivalent if and only if $\tilde{k}(z, z') = k(z, z') + f(z) + f(z')$, for some shift function $f : \mathcal{Z} \to \mathbb{R}$.*

Not every choice of shift function f in Proposition 18 will be valid, as both k and $\tilde{k}$ are required to be positive definite. An important class of shift functions can be derived using RKHS functions, however. Namely, let k be a kernel on $\mathcal{Z}$ and let $f \in \mathcal{H}_k$, and define a kernel

$$\begin{aligned}\tilde{k}_f(z, z') &= \langle k(\cdot, z) - f, k(\cdot, z') - f\rangle_{\mathcal{H}_k} \\ &= k(z, z') - f(z) - f(z') + \|f\|^2_{\mathcal{H}_k}.\end{aligned}$$

Since it is representable as an inner product in a Hilbert space, $\tilde{k}_f$ is a valid kernel which is equivalent to k by Proposition 18. As a special case, if $f = \mu_k(P)$ for some $P \in \mathcal{M}^1_+(\mathcal{Z})$, we obtain the kernel centred at probability measure P:

$$\tilde{k}_P(z, z') := k(z, z') + \mathbb{E}_{WW'}k(W, W') - \mathbb{E}_W k(z, W) - \mathbb{E}_W k(z', W), \tag{4.4}$$

with $W, W' \overset{\text{i.i.d.}}{\sim} P$. Note that $\mathbb{E}_{ZZ' \overset{\text{i.i.d.}}{\sim} P} \tilde{k}_P(Z, Z') = 0$, that is, $\mu_{\tilde{k}_P}(P) = 0$. The kernels of form (4.4) that are centred at the point masses $P = \delta_{z_0}$ are precisely the distance kernels equivalent to k.

The relationship between positive definite kernels and semimetrics of negative type is illustrated in Figure 1.

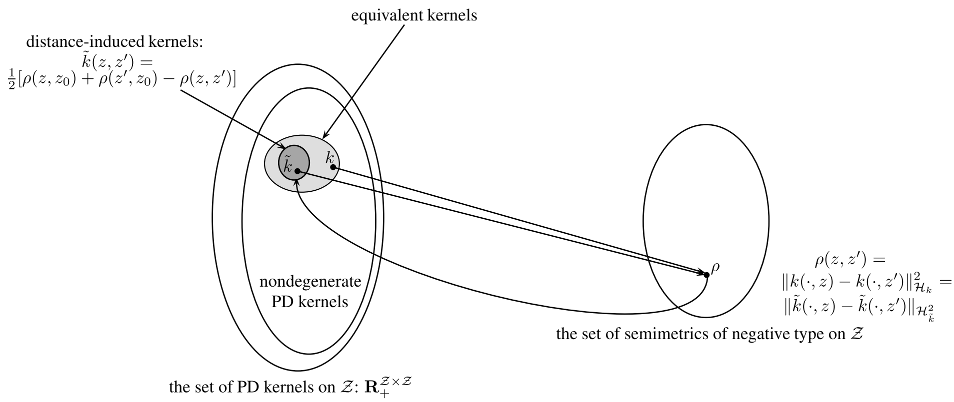

FIG. 1. *The relationship between kernels and semimetrics. An equivalence class of nondegenerate PD kernels is associated to a single semimetric of negative type, and distance kernels induced by that semimetric form only a subset of that class.*

REMARK 19. The requirement that kernels be characteristic (as introduced below Definition 10) is clearly important in hypothesis testing. A second family of kernels, widely used in the machine learning literature, are the universal kernels: universality can be used to guarantee consistency of learning algorithms [Steinwart and Christmann (2008)]. While these two notions are closely related, and in some cases coincide [Sriperumbudur, Fukumizu and Lanckriet (2011)], one can easily construct nonuniversal characteristic kernels as a consequence of Proposition 18. See Appendix B for details.

4.3. *Existence of kernel embedding through a semimetric.* In Section 3.1, we have seen that a sufficient condition for the kernel embedding $\mu_k(\nu)$ of $\nu \in \mathcal{M}(\mathcal{Z})$ to exist is that $\nu \in \mathcal{M}_k^{1/2}(\mathcal{Z})$. We will now interpret this condition in terms of the semimetric ρ generated by k, by relating $\mathcal{M}_k^{\theta}(\mathcal{Z})$ to the space $\mathcal{M}_\rho^{\theta}(\mathcal{Z})$ of measures with finite θ-moment w.r.t. ρ.

PROPOSITION 20. *Let k be a kernel that generates semimetric ρ, and let $n \in \mathbb{N}$. Then $\mathcal{M}_k^{n/2}(\mathcal{Z}) = \mathcal{M}_\rho^{n/2}(\mathcal{Z})$. In particular, if k_1 and k_2 generate the same semimetric ρ, then $\mathcal{M}_{k_1}^{n/2}(\mathcal{Z}) = \mathcal{M}_{k_2}^{n/2}(\mathcal{Z})$.*

PROOF. Let $\theta \geq \frac{1}{2}$. Suppose $\nu \in \mathcal{M}_k^{\theta}(\mathcal{Z})$. Then we have

$$\begin{aligned}\int \rho^{\theta}(z,z_0)\,d|\nu|(z) &= \int \|k(\cdot,z)-k(\cdot,z_0)\|^{2\theta}_{\mathcal{H}_k}\,d|\nu|(z)\\ &\leq \int \left(\|k(\cdot,z)\|_{\mathcal{H}_k}+\|k(\cdot,z_0)\|_{\mathcal{H}_k}\right)^{2\theta}\,d|\nu|(z)\\ &\leq 2^{2\theta-1}\left(\int \|k(\cdot,z)\|^{2\theta}_{\mathcal{H}_k}\,d|\nu|(z)+\int \|k(\cdot,z_0)\|^{2\theta}_{\mathcal{H}_k}\,d|\nu|(z)\right)\end{aligned}$$

$$
\begin{aligned}
&= 2^{2\theta-1}\left(\int k^{\theta}(z,z)\,d|\nu|(z) + k^{\theta}(z_0,z_0)|\nu|(\mathcal{Z})\right)\\
&< \infty,
\end{aligned}
$$

where we have used that $a^{2\theta}$ is a convex function of a. From the above it is clear that $\mathcal{M}_k^{\theta}(\mathcal{Z}) \subset \mathcal{M}_{\rho}^{\theta}(\mathcal{Z})$ for $\theta \geq 1/2$.

To prove the other direction, we show by induction that $\mathcal{M}_{\rho}^{\theta}(\mathcal{Z}) \subset \mathcal{M}_k^{n/2}(\mathcal{Z})$ for $\theta \geq \frac{n}{2}$, $n \in \mathbb{N}$. Let $n = 1$, $\theta \geq \frac{1}{2}$, and suppose that $\nu \in \mathcal{M}_{\rho}^{\theta}(\mathcal{X})$. Then, by invoking the reverse triangle and Jensen's inequalities, we have

$$
\begin{aligned}
\int \rho^{\theta}(z,z_0)\,d|\nu|(z) &= \int \|k(\cdot,z)-k(\cdot,z_0)\|_{\mathcal{H}_k}^{2\theta}\,d|\nu|(z)\\
&\geq \int \left|k^{1/2}(z,z)-k^{1/2}(z_0,z_0)\right|^{2\theta}\,d|\nu|(z)\\
&\geq \left|\int k^{1/2}(z,z)\,d|\nu|(z) - \|\nu\|_{\mathrm{TV}}k^{1/2}(z_0,z_0)\right|^{2\theta},
\end{aligned}
$$

which implies $\nu \in \mathcal{M}_k^{1/2}(\mathcal{Z})$, thereby satisfying the result for $n = 1$. Suppose the result holds for $\theta \geq \frac{n-1}{2}$, that is, $\mathcal{M}_{\rho}^{\theta}(\mathcal{Z}) \subset \mathcal{M}_k^{(n-1)/2}(\mathcal{Z})$ for $\theta \geq \frac{n-1}{2}$. Let $\nu \in \mathcal{M}_{\rho}^{\theta}(\mathcal{Z})$ for $\theta \geq \frac{n}{2}$. Then we have

$$
\begin{aligned}
&\int \rho^{\theta}(z,z_0)\,d|\nu|(z)\\
&\quad= \int \left(\|k(\cdot,z)-k(\cdot,z_0)\|_{\mathcal{H}_k}^{n}\right)^{2\theta/n}\,d|\nu|(z)\\
&\quad\geq \left|\int \left(\|k(\cdot,z)\|_{\mathcal{H}_k} - \|k(\cdot,z_0)\|_{\mathcal{H}_k}\right)^{n}\,d|\nu|(z)\right|^{2\theta/n}\\
&\quad= \left|\int \sum_{r=0}^{n}(-1)^r\binom{n}{r}\|k(\cdot,z)\|_{\mathcal{H}_k}^{n-r}\|k(\cdot,z_0)\|_{\mathcal{H}_k}^{r}\,d|\nu|(z)\right|^{2\theta/n}\\
&\quad= \Bigg|\underbrace{\int k^{n/2}(z,z)\,d|\nu|(z)}_{A}\\
&\qquad\quad \underbrace{+\sum_{r=1}^{n}(-1)^r\binom{n}{r}k^{r/2}(z_0,z_0)\int k^{(n-r)/2}(z,z)\,d|\nu|(z)}_{B}\Bigg|^{2\theta/n}.
\end{aligned}
$$

Note that the terms in B are finite since for $\theta \geq \frac{n}{2} \geq \frac{n-1}{2} \geq \cdots \geq \frac{1}{2}$, we have $\mathcal{M}_{\rho}^{\theta}(\mathcal{Z}) \subset \mathcal{M}_k^{(n-1)/2}(\mathcal{Z}) \subset \cdots \subset \mathcal{M}_k^{1}(\mathcal{Z}) \subset \mathcal{M}_k^{1/2}(\mathcal{Z})$ and therefore A is finite, which means $\nu \in \mathcal{M}_k^{n/2}(\mathcal{Z})$, that is, $\mathcal{M}_\rho^\theta(\mathcal{Z}) \subset \mathcal{M}_k^{n/2}(\mathcal{Z})$ for $\theta \geq \frac{n}{2}$. The result shows that $\mathcal{M}_\rho^\theta(\mathcal{Z}) = \mathcal{M}_k^\theta(\mathcal{Z})$ for all $\theta \in \{\frac{n}{2} : n \in \mathbb{N}\}$. $\square$

REMARK 21. We are now able to show that $P, Q \in \mathcal{M}_\rho^1(\mathcal{Z})$ is sufficient for the existence of $D_{E,\rho}(P, Q)$, that is, to show validity of Definition 5 for general semimetrics of negative type ρ. Namely, we let k be any kernel that generates ρ, whereby $P, Q \in \mathcal{M}_k^1(\mathcal{Z})$. Thus,

$$\mathbb{E}_{ZW}\rho(Z, W) = \mathbb{E}_Z k(Z, Z) + \mathbb{E}_W k(W, W) - 2\mathbb{E}_{ZW} k(Z, W) < \infty,$$

where the first term is finite as $P \in \mathcal{M}_k^1(\mathcal{Z})$, the second term is finite as $Q \in \mathcal{M}_k^1(\mathcal{Z})$, and the third term is finite by noticing that $|k(z, w)| \leq k^{1/2}(z, z) \times k^{1/2}(w, w)$ and $P, Q \in \mathcal{M}_k^1(\mathcal{Z}) \subset \mathcal{M}_k^{1/2}(\mathcal{Z})$.

Proposition 20 gives a natural interpretation of conditions on probability measures in terms of moments w.r.t. ρ. Namely, the kernel embedding $\mu_k(P)$, where kernel k generates the semimetric ρ, exists for every P with finite half-moment w.r.t. ρ, and thus the MMD, $\gamma_k(P, Q)$ between P and Q is well defined whenever both P and Q have finite half-moments w.r.t. ρ. Furthermore, HSIC between random variables X and Y is well defined whenever their marginals P_X and P_Y have finite first moments w.r.t. semimetric $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$ generated by kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ on their respective domains $\mathcal{X}$ and $\mathcal{Y}$.

5. Main results.

In this section, we establish the equivalence between the distance-based approach and the RKHS-based approach to two-sample and independence testing from Sections 2 and 3, respectively.

5.1. *Equivalence of MMD and energy distance.*

We show that for every ρ, the energy distance $D_{E,\rho}$ is related to the MMD associated to a kernel k that generates ρ.

THEOREM 22. *Let $(\mathcal{Z}, \rho)$ be a semimetric space of negative type and let k be any kernel that generates ρ. Then*

$$D_{E,\rho}(P, Q) = 2\gamma_k^2(P, Q) \qquad \forall P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_\rho^1(\mathcal{Z}).$$

In particular, equivalent kernels have the same maximum mean discrepancy.

PROOF. Since k generates ρ, we can write $\rho(z, w) = k(z, z) + k(w, w) - 2k(z, w)$. Denote $\nu = P - Q$. Then

$$\begin{aligned} D_{E,\rho}(P, Q) &= -\int\int \big[k(z, z) + k(w, w) - 2k(z, w)\big]\, d\nu(z)\, d\nu(w) \\ &= 2\int\int k(z, w)\, d\nu(z)\, d\nu(w) \\ &= 2\gamma_k^2(P, Q), \end{aligned}$$

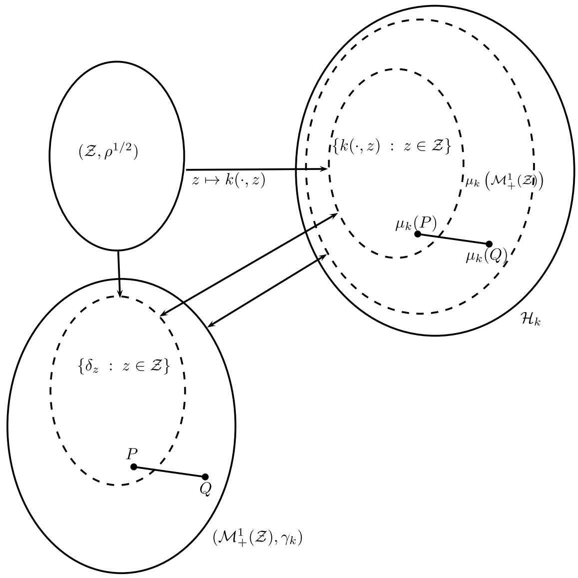

FIG. 2. *Isometries relating the semimetric* ρ *on* $\mathcal{Z}$ *with the RKHS corresponding to a kernel* k *that generates* ρ, *and with the set of probability measures on* $\mathcal{Z}$: (1) $z \mapsto k(\cdot, z)$ *embeds* $(\mathcal{Z}, \rho^{1/2})$ *into* $\mathcal{H}_k$, (2) $z \mapsto \delta_z$ *embeds* $(\mathcal{Z}, \rho^{1/2})$ *into* $(\mathcal{M}_+^1(\mathcal{Z}), \gamma_k)$, *and* (3) $P \mapsto \mu_k(P)$ *embeds* $(\mathcal{M}_+^1(\mathcal{Z}), \gamma_k)$ *into* $\mathcal{H}_k$.

where we used the fact that $\nu(\mathcal{Z}) = 0$. $\square$

This result may be compared with that of Lyons [(2013), page 11, equation (3.9)] for embeddings into general Hilbert spaces, where we have provided the link to RKHS-based statistics (and MMD in particular). Theorem 22 shows that all kernels that generate the same semimetric ρ on $\mathcal{Z}$ give rise to the same metric γ_k on (possibly a subset of) $\mathcal{M}_+^1(\mathcal{Z})$, whence γ_k is merely an extension of the metric induced by $\rho^{1/2}$ on point masses, since

$$\gamma_k(\delta_z, \delta_{z'}) = \|k(\cdot, z) - k(\cdot, z')\|_{\mathcal{H}_k} = \rho^{1/2}(z, z').$$

In other words, whenever kernel k generates ρ, $z \mapsto \delta_z$ is an isometry between $(\mathcal{Z}, \rho^{1/2})$ and $\{\delta_z : z \in \mathcal{Z}\} \subset \mathcal{M}_+^1(\mathcal{Z})$, endowed with the MMD metric $\gamma_k = \frac{1}{2} D_{E,\rho}^{1/2}$; and the Aronszajn map $z \mapsto k(\cdot, z)$ is an isometric embedding of a metric space $(\mathcal{Z}, \rho^{1/2})$ into $\mathcal{H}_k$. These isometries are depicted in Figure 2. For simplicity, we show the case of a bounded kernel, where kernel embeddings are well defined for all $P \in \mathcal{M}_+^1(\mathcal{Z})$, in which case $(\mathcal{M}_+^1(\mathcal{Z}), \gamma_k)$ and $\mu_k(\mathcal{M}_+^1(\mathcal{Z})) = \{\mu_k(P) : P \in \mathcal{M}_+^1(\mathcal{Z})\}$ endowed with the Hilbert-space metric inherited from $\mathcal{H}_k$ are also isometric (note that this implies that the subsets of RKHSs corresponding to equivalent kernels are also isometric).

REMARK 23. Theorem 22 requires that $P, Q \in \mathcal{M}_\rho^1(\mathcal{Z})$, that is, that P and Q have finite first moments w.r.t. ρ, as otherwise the energy distance between P and Q may be undefined; for example, each of the expectations $\mathbb{E}_{ZZ'}\rho(Z, Z')$, $\mathbb{E}_{WW'}\rho(W, W')$ and $\mathbb{E}_{ZW}\rho(Z, W)$ may be infinite. However, as long as a *weaker condition* $P, Q \in \mathcal{M}_\rho^{1/2}(\mathcal{Z})$ is satisfied, that is, P and Q have finite *half*-moments w.r.t. ρ, the maximum mean discrepancy γ_k will be well defined. If, in addition, $P, Q \in \mathcal{M}_\rho^1(\mathcal{Z})$, then the energy distance between P and Q is also well defined, and must be equal to γ_k. We will later invoke the same condition $P, Q \in \mathcal{M}_k^1(\mathcal{Z})$ when describing the asymptotic distribution of the empirical maximum mean discrepancy in Section 7.

5.2. *Equivalence between HSIC and distance covariance.* We now show that distance covariance is an instance of the Hilbert–Schmidt independence criterion.

THEOREM 24. *Let $(\mathcal{X}, \rho_\mathcal{X})$ and $(\mathcal{Y}, \rho_\mathcal{Y})$ be semimetric spaces of negative type, and let $X \sim P_X \in \mathcal{M}^2_{\rho_\mathcal{X}}(\mathcal{X})$ and $Y \sim P_Y \in \mathcal{M}^2_{\rho_\mathcal{Y}}(\mathcal{Y})$, having joint distribution P_{XY}. Let $k_\mathcal{X}$ and $k_\mathcal{Y}$ be any two kernels on $\mathcal{X}$ and $\mathcal{Y}$ that generate $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$, respectively, and denote*

$$k\big((x, y), (x', y')\big) = k_\mathcal{X}(x, x')k_\mathcal{Y}(y, y'). \tag{5.1}$$

Then, $\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X, Y) = 4\gamma_k^2(P_{XY}, P_X P_Y)$.

PROOF. Define $\nu = P_{XY} - P_X P_Y$. Then

$$\begin{aligned}
\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X, Y) &= \int\int \rho_\mathcal{X}(x, x')\rho_\mathcal{Y}(y, y')\,d\nu(x, y)\,d\nu(x', y') \\
&= \int\int \big(k_\mathcal{X}(x, x) + k_\mathcal{X}(x', x') - 2k_\mathcal{X}(x, x')\big) \\
&\qquad \times \big(k_\mathcal{Y}(y, y) + k_\mathcal{Y}(y', y') - 2k_\mathcal{Y}(y, y')\big)\,d\nu(x, y)\,d\nu(x', y') \\
&= 4\int\int k_\mathcal{X}(x, x')k_\mathcal{Y}(y, y')\,d\nu(x, y)\,d\nu(x', y') \\
&= 4\gamma_k^2(P_{XY}, P_X P_Y),
\end{aligned}$$

where we used that $\nu(\mathcal{X} \times \mathcal{Y}) = 0$, and that $\int g(x, y, x', y')\,d\nu(x, y)\,d\nu(x', y') = 0$ when g does not depend on one or more of its arguments, since ν also has zero marginal measures. Convergence of integrals of the form $\int k_\mathcal{X}(x, x) \times k_\mathcal{Y}(y, y)\,d\nu(x, y)$ is ensured by the moment conditions on the marginals. □

We remark that a similar result to Theorem 24 is given by Lyons [(2013), Proposition 3.16], but without making use of the link with kernel embeddings. Theorem 24 is a more general statement, in the sense that we allow ρ to be a semimetric of negative type, rather than metric. In addition, the kernel interpretation leads to a significantly simpler proof: the result is an immediate application of the HSIC expansion in (3.6).

REMARK 25. As in Remark 23, to ensure the existence of the distance covariance, we impose a stronger condition on the marginals: $P_X \in \mathcal{M}^2_{k_{\mathcal{X}}}(\mathcal{X})$ and $P_Y \in \mathcal{M}^2_{k_{\mathcal{Y}}}(\mathcal{Y})$, while $P_X \in \mathcal{M}^1_{k_{\mathcal{X}}}(\mathcal{X})$ and $P_Y \in \mathcal{M}^1_{k_{\mathcal{Y}}}(\mathcal{Y})$ are sufficient for the existence of the Hilbert–Schmidt independence criterion.

By combining the Theorems 22 and 24, we can establish the direct relation between energy distance and distance covariance, as discussed in Remark 7.

COROLLARY 26. *Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be semimetric spaces of negative type, and let $X \sim P_X \in \mathcal{M}^2_{\rho_{\mathcal{X}}}(\mathcal{X})$ and $Y \sim P_Y \in \mathcal{M}^2_{\rho_{\mathcal{Y}}}(\mathcal{Y})$, having joint distribution P_{XY}. Then $\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X, Y) = D_{E,\tilde{\rho}}(P_{XY}, P_X P_Y)$,* where *$\frac{1}{2}\tilde{\rho}$ is generated by the product kernel in (5.1).*

REMARK 27. As introduced by Székely, Rizzo and Bakirov (2007), the notion of distance covariance extends naturally to that of *distance variance* $\mathcal{V}^2(X) = \mathcal{V}^2(X, X)$ and of *distance correlation* (by analogy with the Pearson product-moment correlation coefficient),

$$\mathcal{R}^2(X, Y) = \begin{cases} \dfrac{\mathcal{V}^2(X, Y)}{\mathcal{V}(X)\mathcal{V}(Y)}, & \mathcal{V}(X)\mathcal{V}(Y) > 0, \\ 0, & \mathcal{V}(X)\mathcal{V}(Y) = 0. \end{cases}$$

The distance correlation can also be expressed in terms of associated kernels—see Appendix A for details.

5.3. *Characteristic function interpretation.* The distance covariance in (2.7) was defined by Székely, Rizzo and Bakirov (2007) in terms of a weighted distance between characteristic functions. We briefly review this interpretation here, and show that this approach *cannot* be used to derive a kernel-based measure of dependence [this result was first obtained by Gretton, Fukumizu and Sriperumbudur (2009), and is included here in the interest of completeness]. Let X be a random vector on $\mathcal{X} = \mathbb{R}^p$ and Y a random vector on $\mathcal{Y} = \mathbb{R}^q$. The characteristic functions of X and Y, respectively, will be denoted by f_X and f_Y, and their joint characteristic function by f_{XY}. The distance covariance $\mathcal{V}(X, Y)$ is defined via the norm of $f_{XY} - f_X f_Y$ in a weighted L_2 space on $\mathbb{R}^{p+q}$, that is,

$$\mathcal{V}^2(X, Y) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 w(t, s)\, dt\, ds \tag{5.2}$$

for a particular choice of weight function given by

$$w(t, s) = \frac{1}{c_p c_q} \cdot \frac{1}{\|t\|^{1+p}\|s\|^{1+q}}, \tag{5.3}$$

where $c_d = \pi^{(1+d)/2}/\Gamma((1 + d)/2)$, $d \geq 1$. An important property of distance covariance is that $\mathcal{V}(X, Y) = 0$ if and only if X and Y are independent. We next obtain a similar statistic in the kernel setting. Write $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and let $k(z, z') = \kappa(z - z')$ be a translation invariant RKHS kernel on $\mathcal{Z}$, where $\kappa : \mathcal{Z} \to \mathbb{R}$ is a bounded continuous function. Using Bochner's theorem, κ can be written as

$$\kappa(z) = \int e^{-z^\top u} \, d\Lambda(u)$$

for a finite nonnegative Borel measure Λ. It follows [Gretton, Fukumizu and Sriperumbudur (2009)] that

$$\gamma_k^2(P_{XY}, P_X P_Y) = \int_{\mathbb{R}^{p+q}} \left| f_{X,Y}(t, s) - f_X(t) f_Y(s) \right|^2 d\Lambda(t, s),$$

which is in clear correspondence with (5.2). The weight function in (5.3) is not integrable, however, so we cannot find a continuous translation invariant kernel for which γ_k coincides with the distance covariance. Indeed, the kernel in (5.1) is *not* translation invariant.

A further related family of statistics for two-sample tests has been studied by Alba Fernández, Jiménez Gamero and Muñoz García (2008), and the majority of results therein can be directly obtained via Bochner's theorem from the corresponding results on kernel two-sample testing, in the case of translation-invariant kernels on $\mathbb{R}^d$. That being said, we emphasise that the RKHS-based approach extends to general topological spaces and positive definite functions, and it is unclear whether every kernel two-sample/independence test has an interpretation in terms of characteristic functions.

6. Distinguishing probability distributions. Theorem 3.20 of Lyons (2013) shows that distance covariance in a metric space characterizes independence if the metrics satisfy an additional property, termed *strong negative type*. We review this notion and establish the interpretation of strong negative type in terms of RKHS kernel properties.

DEFINITION 28. The semimetric space $(\mathcal{Z}, \rho)$, where ρ is generated by kernel k, is said to have *a strong negative type* if $\forall P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^1(\mathcal{Z})$,

$$P \neq Q \Rightarrow \int \rho \, d\big([P - Q] \times [P - Q]\big) < 0. \tag{6.1}$$

Since the quantity in (6.1) is, by equation (2.6), exactly $-D_{E,\rho}(P, Q) = -2\gamma_k^2(P, Q)$, $\forall P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^1(\mathcal{Z})$, the following is immediate:

PROPOSITION 29. *Let kernel k generate ρ. Then $(\mathcal{Z}, \rho)$ has a strong negative type if and only if k is characteristic to $\mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^1(\mathcal{Z})$.*

Thus, the problem of checking whether a semimetric is of strong negative type is equivalent to checking whether its associated kernel is characteristic to an appropriate space of Borel probability measures. This conclusion has some overlap with

[Lyons (2013)]: in particular, Proposition 29 is stated in [Lyons (2013), Proposition 3.10], where the barycenter map β is a kernel embedding in our terminology, although Lyons does not consider distribution embeddings in an RKHS.

REMARK 30. From Lyons (2013), Theorem 3.25, every separable Hilbert space $\mathcal{Z}$ is of strong negative type, so a distance kernel k induced by the (inner product) metric on $\mathcal{Z}$ is characteristic to the appropriate space of probability measures.

REMARK 31. Consider the kernel in (5.1), and assume for simplicity that $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are bounded, so that we can consider embeddings of all probability measures. It turns out that k need not be characteristic—that is, it may not be able to distinguish between any two distributions on $\mathcal{X} \times \mathcal{Y}$, even if $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are characteristic. Namely, if $k_{\mathcal{X}}$ is the distance kernel induced by $\rho_{\mathcal{X}}$ and centred at x_0, then $k((x_0, y), (x_0, y')) = 0$ for all $y, y' \in \mathcal{Y}$. That means that for every two distinct $P_Y, Q_Y \in \mathcal{M}_+^1(\mathcal{Y})$, we have $\gamma_k^2(\delta_{x_0} P_Y, \delta_{x_0} Q_Y) = 0$. Thus, given that $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$ have strong negative type, the kernel in (5.1) characterizes independence, but not equality of probability measures on the product space. Informally speaking, distinguishing P_{XY} from $P_X P_Y$ is an easier problem than two-sample testing on the product space.

7. Empirical estimates and hypothesis tests. In this section, we outline the construction of tests based on the empirical counterparts of MMD/energy distance and HSIC/distance covariance.

7.1. *Two-sample testing.* So far, we have seen that the population expression of the MMD between P and Q is well defined as long as P and Q lie in the space $\mathcal{M}_k^{1/2}(\mathcal{Z})$, or, equivalently, have a finite half-moment w.r.t. semimetric ρ generated by k. However, this assumption will not suffice to establish a meaningful hypothesis test using empirical estimates of the MMD. We will require a stronger condition, that $P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^1(\mathcal{Z})$ (which is the same condition under which the energy distance is well defined). Note that, under this condition we also have $k \in L^2_{P \times P}(\mathcal{Z} \times \mathcal{Z})$, as $\int\int k^2(z, z')\, dP(z)\, dP(z') \leq (\int k(z, z)\, dP(z))^2$.

Given i.i.d. samples $\mathbf{z} = \{z_i\}_{i=1}^m \sim P$ and $\mathbf{w} = \{w_i\}_{i=1}^n \sim Q$, the empirical (biased) V-statistic estimate of (3.3) is given by

$$\begin{aligned}\hat{\gamma}_{k,V}^2(\mathbf{z}, \mathbf{w}) &= \gamma_k^2\left(\frac{1}{m}\sum_{i=1}^m \delta_{z_i}, \frac{1}{n}\sum_{j=1}^n \delta_{w_j}\right)\\ &= \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(z_i, z_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(w_i, w_j) \\ &\quad - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(z_i, w_j).\end{aligned} \tag{7.1}$$

Recall that if k generates ρ, this estimate involves only the pairwise ρ-distances between the sample points.

We now describe a two-sample test using this statistic. The kernel $\tilde{k}_P$ centred at P in (4.4) plays a key role in characterizing the null distribution of degenerate V-statistic. To $\tilde{k}_P$, we associate the *integral kernel operator* $S_{\tilde{k}_P} : L^2_P(\mathcal{Z}) \to L^2_P(\mathcal{Z})$ [cf., e.g., Steinwart and Christmann (2008), page 126–127], given by

$$S_{\tilde{k}_P} g(z) = \int_{\mathcal{Z}} \tilde{k}_P(z, w) g(w)\, dP(w). \tag{7.2}$$

The condition that $P \in \mathcal{M}^1_k(\mathcal{Z})$, and, as a consequence, that $\tilde{k}_P \in L^2_{P\times P}(\mathcal{Z} \times \mathcal{Z})$, is closely related to the desired properties of the integral operator. Namely, this implies that $S_{\tilde{k}_P}$ is a trace class operator, and, thus, a Hilbert–Schmidt operator [Reed and Simon (1980), Proposition VI.23]. The following theorem is a special case of Gretton et al. (2012a), Theorem 12, which extends Anderson, Hall and Titterington (1994), Section 2.3, to general RKHS kernels (as noted by Anderson et al., the form of the asymptotic distribution of the V-statistic requires $S_{\tilde{k}_P}$ to be trace-class, whereas the U-statistic has the weaker requirement that $S_{\tilde{k}_P}$ be Hilbert–Schmidt). For simplicity, we focus on the case where $m = n$.

THEOREM 32. *Let k be a kernel on $\mathcal{Z}$, and $\mathbf{Z} = \{Z_i\}_{i=1}^m$ and $\mathbf{W} = \{W_i\}_{i=1}^m$ be two i.i.d. samples from $P \in \mathcal{M}^1_+(\mathcal{Z}) \cap \mathcal{M}^1_k(\mathcal{Z})$. Assume $S_{\tilde{k}_P}$ is trace class. Then*

$$\frac{m}{2} \hat{\gamma}^2_{k,V}(\mathbf{Z}, \mathbf{W}) \rightsquigarrow \sum_{i=1}^{\infty} \lambda_i N_i^2, \tag{7.3}$$

where $N_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $i \in \mathbb{N}$, and $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues of the operator $S_{\tilde{k}_P}$.

Note that the limiting expression in (7.3) is a valid random variable precisely since $S_{\tilde{k}_P}$ is Hilbert–Schmidt, that is, since $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$.

7.2. *Independence testing.* In the case of independence testing, we are given i.i.d. samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \sim P_{XY}$, and the resulting V-statistic estimate (HSIC) is [Gretton et al. (2005, 2008)]

$$\mathrm{HSIC}(\mathbf{z}; k_{\mathcal{X}}, k_{\mathcal{Y}}) = \frac{1}{m^2} \mathrm{Tr}(K_{\mathcal{X}} H K_{\mathcal{Y}} H), \tag{7.4}$$

where $K_{\mathcal{X}}$, $K_{\mathcal{Y}}$ and H are $m \times m$ matrices given by $(K_{\mathcal{X}})_{ij} := k_{\mathcal{X}}(x_i, x_j)$, $(K_{\mathcal{Y}})_{ij} := k_{\mathcal{Y}}(y_i, y_j)$ and $H_{ij} = \delta_{ij} - \frac{1}{m}$ (centering matrix). The null distribution of HSIC takes an analogous form to (7.3) of a weighted sum of chi-squares, but with coefficients corresponding to the products of the eigenvalues of integral operators $S_{\tilde{k}_{P_X}} : L^2_{P_X}(\mathcal{X}) \to L^2_{P_X}(\mathcal{X})$ and $S_{\tilde{k}_{P_Y}} : L^2_{P_Y}(\mathcal{Y}) \to L^2_{P_Y}(\mathcal{Y})$. Similarly to the case of two-sample testing, we will require that $P_X \in \mathcal{M}^1_{k_{\mathcal{X}}}(\mathcal{X})$ and $P_Y \in \mathcal{M}^1_{k_{\mathcal{Y}}}(\mathcal{Y})$, implying that integral operators $S_{\tilde{k}_{P_X}}$ and $S_{\tilde{k}_{P_Y}}$ are trace class operators. The following theorem is from Zhang et al. (2011), Theorem 4. See also Lyons (2013), Remark 2.9.

THEOREM 33. *Let $\mathbf{Z} = \{(X_i, Y_i)\}_{i=1}^m$ be an i.i.d. sample from $P_{XY} = P_X P_Y$, with values in $\mathcal{X} \times \mathcal{Y}$, s.t. $P_X \in \mathcal{M}^1_{k_\mathcal{X}}(\mathcal{X})$ and $P_Y \in \mathcal{M}^1_{k_\mathcal{Y}}(\mathcal{Y})$. Then*

$$m\,\mathrm{HSIC}(\mathbf{Z}; k_\mathcal{X}, k_\mathcal{Y}) \rightsquigarrow \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \lambda_i \eta_j N_{i,j}^2, \tag{7.5}$$

where $N_{i,j} \sim \mathcal{N}(0, 1)$, $i, j \in \mathbb{N}$, are independent and $\{\lambda_i\}_{i=1}^\infty$ and $\{\eta_j\}_{j=1}^\infty$ are the eigenvalues of the operators $S_{\tilde{k}_{P_X}}$ and $S_{\tilde{k}_{P_Y}}$, respectively.

7.3. *Test designs.* We would like to design distance-based tests with an asymptotic Type I error of α, and thus we require an estimate of the $(1-\alpha)$-quantile of the null distribution. We investigate two approaches, both of which yield consistent tests: a bootstrap approach [Arcones and Giné (1992)] and a spectral approach [Gretton et al. (2009), Zhang et al. (2011)]. The latter requires empirical computation of eigenvalues of the integral kernel operators, a problem studied extensively in the context of kernel PCA [Schölkopf, Smola and Müller (1997)]. To estimate limiting distribution in (7.3), we compute the spectrum of the centred Gram matrix $\tilde{K} = HKH$ on the aggregated samples. Here, K is a $2m \times 2m$ matrix, with entries $K_{ij} = k(u_i, u_j)$, $\mathbf{u} = [\mathbf{zw}]$ is the concatenation of the two samples and H is the centering matrix. Gretton et al. (2009) show that the null distribution defined using the finite sample estimates of these eigenvalues converges to the population distribution, provided that the spectrum is square-root summable. As demonstrated empirically by Gretton et al. (2009), spectral estimation of the test threshold has a smaller computational cost than that of the bootstrap-based approach, while providing an indistinguishable performance. The same approach can be used in obtaining a consistent finite sample null distribution for HSIC, via computation of the empirical eigenvalues of $\tilde{K}_\mathcal{X} = HK_\mathcal{X}H$ and $\tilde{K}_\mathcal{Y} = HK_\mathcal{Y}H$; see Zhang et al. (2011).

Both Székely and Rizzo [(2004), page 14] and Székely, Rizzo and Bakirov [(2007), pages 2782–2783] establish that the energy distance and distance covariance statistics, respectively, converge to the weighted sums of chi-squares of forms similar to (7.3). Analogous results for the generalized distance covariance are presented in Lyons (2013), pages 7–8. These works do not propose test designs that attempt to estimate the coefficients λ_i, $i \in \mathbb{N}$, however. Besides the bootstrap, Székely, Rizzo and Bakirov [(2007), Theorem 6] also propose an independence test using a bound applicable to a general quadratic form Q of centered Gaussian random variables with $\mathbb{E}[Q] = 1 : \mathbb{P}\{Q \geq (\Phi^{-1}(1-\alpha/2)^2)\} \leq \alpha$, valid for $0 < \alpha \leq 0.215$. When applied to the distance covariance statistic, the upper bound of α is achieved if X and Y are independent Bernoulli variables. The authors remark that the resulting criterion might be over-conservative. Thus, more sensitive distance covariance tests are possible by computing the spectrum of the centred Gram matrices associated to distance kernels, which is the approach we apply in the next section.

8. Experiments. In this section, we assess the numerical performance of the distance-based and RKHS-based test statistics with some standard distance/kernel choices on a series of synthetic data examples.

8.1. *Two-sample experiments*. In the two-sample experiments, we investigate three different kinds of synthetic data. In the first, we compare two multivariate Gaussians, where the means differ in one dimension only, and all variances are equal. In the second, we again compare two multivariate Gaussians, but this time with identical means in all dimensions, and variance that differs in a single dimension. In our third experiment, we use the benchmark data of Sriperumbudur et al. (2009): one distribution is a univariate Gaussian, and the second is a univariate Gaussian with a sinusoidal perturbation of increasing frequency (where higher frequencies correspond to harder problems). All tests use a distance kernel induced by the Euclidean distance. As shown on the left-hand plots in Figure 3, the spectral and bootstrap test designs appear indistinguishable, and significantly outperform the test designed using the quadratic form bound, which appears to be far too conservative for the data sets considered. The average Type I errors are listed in Table 1, and are close to the desired test size of $\alpha = 0.05$ for the spectral and bootstrap tests.

We also compare the performance to that of the Gaussian kernel, commonly used in machine learning, with the bandwidth set to the median distance between points in the aggregation of samples. We see that when the means differ, both tests perform similarly. When the variances differ, it is clear that the Gaussian kernel has a major advantage over the distance-induced kernel, although this advantage decreases with increasing dimension (where both perform poorly). In the case of a sinusoidal perturbation, the performance is again very similar.

In addition, following Example 15, we investigate performance of kernels obtained using the semimetric $\rho(z, z') = \|z - z'\|^q$ for $0 < q \leq 2$. Results are presented in the right-hand plots of Figure 3. In the case of sinusoidal perturbation, we observe a dramatic improvement compared with the $q = 1$ case and the Gaussian kernel: values $q = 1/3$ (and smaller) offer virtually error-free performance even at high frequencies [note that $q = 1$ yields the energy distance described in Székely and Rizzo (2004, 2005)]. Small improvements over a wider q range are also observed in the cases of differing mean and variance.

We observe from the simulation results that distance-induced kernels with higher exponents are advantageous in cases where distributions differ in mean value along a single dimension (with noise in the remainder), whereas distance

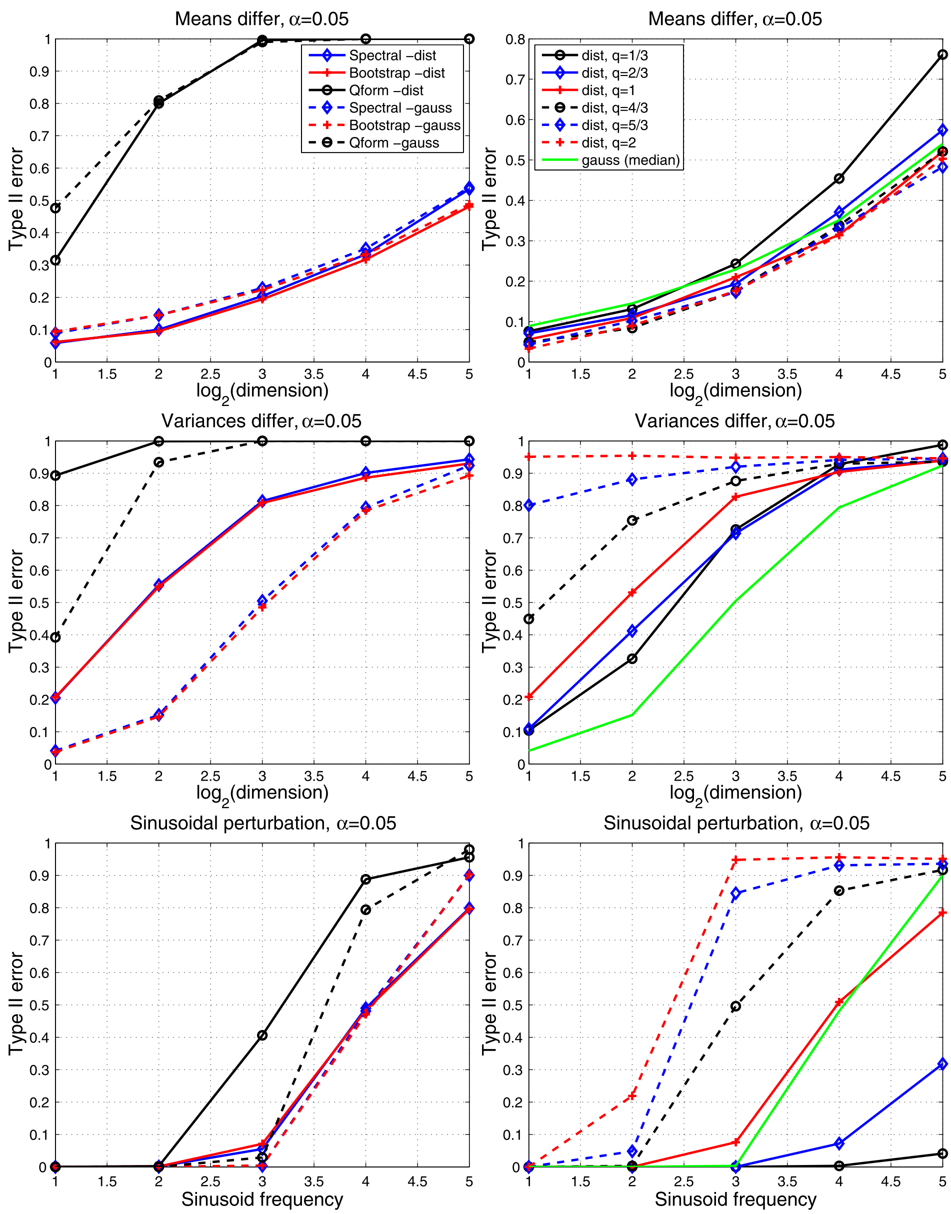

FIG. 3. (*Left*) *MMD using Gaussian and distance kernels for various tests*; (*right*) *Spectral MMD using distance kernels with various exponents.*

kernels with smaller exponents are more sensitive to differences in distributions at finer lengthscales (i.e., where the characteristic functions of the distributions differ at higher frequencies).

TABLE 1
Type I error (in %) for two-sample tests with distance-induced kernels

	mean	**var**	**sine**
Spec	4.66	4.72	5.10
Boot	5.02	5.16	5.20
Qform	0.02	0.05	0.98

8.2. *Independence experiments*. To assess independence tests, we used an artificial benchmark proposed by Gretton et al. (2008): we generated univariate random variables from the Independent Component Analysis (ICA) benchmark densities of Bach and Jordan (2002); rotated them in the product space by an angle between 0 and $\pi/4$ to introduce dependence; filled additional dimensions with independent Gaussian noise; and, finally, passed the resulting multivariate data through random and independent orthogonal transformations. The resulting random variables X and Y were dependent but uncorrelated. The case $m = 128$ (sample size) and $d = 2$ (dimension) is plotted in Figure 4 (left). As observed by Gretton, Fukumizu and Sriperumbudur (2009), the Gaussian kernel using the median inter-point distance as bandwidth does better than the distance-induced kernel with $q = 1$. By varying q, however, we are able to obtain a wide performance range: in particular, the values $q = 1/3$ (and smaller) have an advantage over the Gaussian kernel on this dataset. As for the two-sample case, bootstrap and spectral tests have indistinguishable performance, and are significantly more sensitive than the quadratic form-based test, which failed to detect any dependence on this dataset.

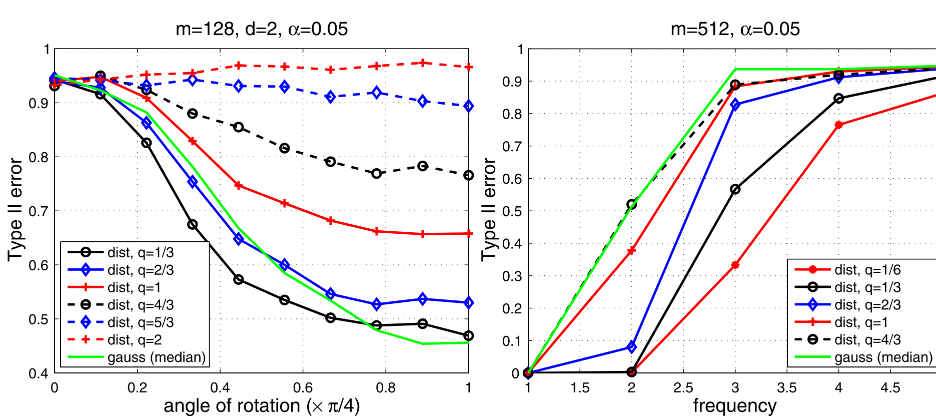

FIG. 4. *HSIC using distance kernels with various exponents and a Gaussian kernel as a function of* (*left*) *the angle of rotation for the dependence induced by rotation*; (*right*) *frequency* ℓ *in the sinusoidal dependence example*.

In addition, we assess the performance on sinusoidally dependent data. The sample of the random variable pair X, Y was drawn from $P_{XY} \propto 1 + \sin(\ell x) \times \sin(\ell y)$ for integer ℓ, on the support $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} := [-\pi, \pi]$ and $\mathcal{Y} := [-\pi, \pi]$. In this way, increasing ℓ causes the departure from a uniform (independent) distribution to occur at increasing frequencies, making this departure harder to detect given a small sample size. Results are in Figure 4 (right). The distance covariance outperforms the Gaussian kernel (median bandwidth) on this example, and smaller exponents result in better performance (lower Type II error when the departure from independence occurs at higher frequencies). Finally, we note that the setting $q = 1$, as described by Székely, Rizzo and Bakirov (2007), Székely and Rizzo (2009), is a reasonable heuristic in practice, but does not yield the most powerful tests on either dataset. Informally, the exponent in the distance-induced kernel plays a similar role as the bandwidth of the Gaussian kernel, and smaller exponents are able to detect dependencies at smaller lengthscales. Poor performance of the Gaussian kernel with median bandwidth in this example is a consequence of the mismatch between the overall lengthscale of the marginal distributions (captured by the median inter-point distances) and the lengthscales at which dependencies are present.

9. Conclusion. We have established an equivalence between the generalized notions of energy distance and distance covariance, computed with respect to semimetrics of negative type, and distances between embeddings of probability measures into certain reproducing kernel Hilbert spaces. As a consequence, we can view energy distance and distance covariance as members of a much larger class of discrepancy/dependence measures, and we can choose among this larger class to design more powerful tests. For instance, Gretton et al. (2012b) recently proposed a strategy of selecting from a candidate kernels so as to asymptotically optimize the relative efficiency of a two-sample test. Moreover, kernel-based tests can be performed on the data that do not lie in a Euclidean space. This opens the door to new and powerful tools for exploratory data analysis whenever an appropriate domain-specific notion of distance (negative type semimetric) or similarity (kernel) can be defined. Finally, the family of kernels that arises from the energy distance/distance covariance can be employed in many additional kernel-based applications in statistics and machine learning, such as conditional dependence testing and estimating the chi-squared distance [Fukumizu et al. (2008)], Bayesian inference [Fukumizu, Song and Gretton (2011)] and mixture density estimation [Sriperumbudur (2011)].

APPENDIX A: DISTANCE CORRELATION

As described by Székely, Rizzo and Bakirov (2007), the notion of distance covariance extends naturally to that of *distance variance* $\mathcal{V}^2(X) = \mathcal{V}^2(X, X)$ and of *distance correlation* (by analogy with the Pearson product-moment correlation coefficient),

$$\mathcal{R}^2(X,Y) = \begin{cases} \dfrac{\mathcal{V}^2(X,Y)}{\mathcal{V}(X)\mathcal{V}(Y)}, & \mathcal{V}(X)\mathcal{V}(Y) > 0, \\ 0, & \mathcal{V}(X)\mathcal{V}(Y) = 0. \end{cases}$$

Distance correlation also has a straightforward interpretation in terms of kernels,

$$\begin{aligned} \mathcal{R}^2(X,Y) &= \frac{\mathcal{V}^2(X,Y)}{\mathcal{V}(X)\mathcal{V}(Y)} \\ &= \frac{\gamma_k^2(P_{XY}, P_X P_Y)}{\gamma_k(P_{XX}, P_X P_X)\gamma_k(P_{YY}, P_Y P_Y)} \qquad \text{(A.1)} \\ &= \frac{\|\Sigma_{XY}\|_{\mathrm{HS}}^2}{\|\Sigma_{XX}\|_{\mathrm{HS}}\|\Sigma_{YY}\|_{\mathrm{HS}}}, \end{aligned}$$

where covariance operator $\Sigma_{XY} : \mathcal{H}_{k_\mathcal{X}} \to \mathcal{H}_{k_\mathcal{Y}}$ is a linear operator for which $\langle \Sigma_{XY} f, g \rangle_{\mathcal{H}_{k_\mathcal{Y}}} = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X f(X) \mathbb{E}_Y g(Y)$ for all $f \in \mathcal{H}_{k_\mathcal{X}}$ and $g \in \mathcal{H}_{k_\mathcal{Y}}$, and $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm [Gretton et al. (2005)]. It is clear that $\mathcal{R}$ is invariant to scaling $(X,Y) \mapsto (\varepsilon X, \varepsilon Y)$, $\varepsilon > 0$, whenever the corresponding semimetrics are homogeneous, that is, whenever $\rho_\mathcal{X}(\varepsilon x, \varepsilon x') = \varepsilon \rho_\mathcal{X}(x,x')$, and similarly for $\rho_\mathcal{Y}$. Moreover, $\mathcal{R}$ is invariant to translations, $(X,Y) \mapsto (X + x', Y + y')$, $x' \in \mathcal{X}$, $y' \in \mathcal{Y}$, whenever $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$ are translation invariant. Therefore, by varying the choice of kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$, we obtain in (A.1) a very broad class of dependence measures that generalize the distance correlation of Székely, Rizzo and Bakirov (2007) and can be used in exploratory data analysis as a measure of dependence between pairs of random variables that take values in multivariate or structured/non-Euclidean domains.

APPENDIX B: LINK WITH UNIVERSAL KERNELS

We briefly remark on how our results on equivalent kernels relate to the notion of universal kernels on compact metric spaces in the sense of Steinwart and Christmann (2008), Definition 4.52:

DEFINITION 34. A continuous kernel k on a compact metric space $\mathcal{Z}$ is said to be universal if its RKHS $\mathcal{H}_k$ is dense in the space $C(\mathcal{Z})$ of continuous functions on $\mathcal{Z}$, endowed with the uniform norm.

The family of universal kernels includes the most popular choices in machine learning literature, including the Gaussian and the Laplacian kernel. The following characterization of universal kernels is due to Sriperumbudur, Fukumizu and Lanckriet (2011):

PROPOSITION 35. *Let k be a continuous kernel on a compact metric space $\mathcal{Z}$. Then, k is universal if and only if $\mu_k : \mathcal{M}(\mathcal{Z}) \to \mathcal{H}_k$ is a vector space monomorphism, that is,*

$$\|\mu_k(\nu)\|^2_{\mathcal{H}_k} = \int\int k(z, z')\, d\nu(z)\, d\nu(z') > 0 \qquad \forall \nu \in \mathcal{M}(\mathcal{Z}) \setminus \{0\}.$$

As a direct consequence, every universal kernel k is also characteristic, as μ_k is, in particular, injective on the space of probability measures. Now, consider a kernel $\tilde{k}_f$ centered at $f = \mu_k(\nu)$ for some $\nu \in \mathcal{M}(\mathcal{Z})$, such that $\nu(\mathcal{Z}) = 1$. Then $\tilde{k}_f$ is no longer universal, since

$$\begin{aligned}
&\|\mu_{\tilde{k}_f}(\nu)\|^2_{\mathcal{H}_{\tilde{k}_f}} \\
&\quad = \int \tilde{k}_f(z, z')\, d\nu(z)\, d\nu(z') \\
&\quad = \int\int \biggl[k(z, z') - \int k(w, z)\, d\nu(w) - \int k(w, z')\, d\nu(w) \\
&\qquad\qquad + \int\int k(w, w')\, d\nu(w)\, d\nu(w')\biggr]\, d\nu(z)\, d\nu(z') \\
&\quad = (1 - \nu(\mathcal{Z}))^2 \|\mu_k(\nu)\|^2_{\mathcal{H}_k} \\
&\quad = 0.
\end{aligned}$$

However, $\tilde{k}_f$ is still characteristic, as it is equivalent to k. This means that all kernels of the form (4.4), including the distance kernels, are examples of nonuniversal characteristic kernels, provided that they generate a semimetric ρ of strong negative type. In particular, the kernel in (4.2) on a compact $\mathcal{Z} \subset \mathbb{R}^d$ is a characteristic nonuniversal kernel for $q < 2$. This result is of some interest to the machine learning community, as such kernels have typically been difficult to construct. For example, the two notions are known to be equivalent on the family of translation invariant kernels on $\mathbb{R}^d$ [Sriperumbudur, Fukumizu and Lanckriet (2011)].

Acknowledgments. D. Sejdinovic, B. Sriperumbudur and A. Gretton acknowledge support of the Gatsby Charitable Foundation. The work was carried out when B. Sriperumbudur was with Gatsby Unit, University College London. B. Sriperumbudur and A. Gretton contributed equally.

REFERENCES

ALBA FERNÁNDEZ, V., JIMÉNEZ GAMERO, M. D. and MUÑOZ GARCÍA, J. (2008). A test for the two-sample problem based on empirical characteristic functions. *Comput. Statist. Data Anal.* **52** 3730–3748. MR2427377

ANDERSON, N. H., HALL, P. and TITTERINGTON, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J. Multivariate Anal.* **50** 41–54. MR1292607

ARCONES, M. A. and GINÉ, E. (1992). On the bootstrap of U and V statistics. *Ann. Statist.* **20** 655–674. MR1165586

BACH, F. R. and JORDAN, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051

BARINGHAUS, L. and FRANZ, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.* **88** 190–206. MR2021870

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

BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, London.

FUKUMIZU, K., SONG, L. and GRETTON, A. (2011). Kernel Bayes' rule. In *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira and K. Q. Weinberger, eds.) **24** 1737–1745. Curran Associates, Red Hook, NY.

FUKUMIZU, K., GRETTON, A., SUN, X. and SCHÖLKOPF, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems* **20** 489–496. MIT Press, Cambridge, MA.

FUKUMIZU, K., SRIPERUMBUDUR, B., GRETTON, A. and SCHOELKOPF, B. (2009). Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems* **21** 473–480. Curran Associates, Red Hook, NY.

GRETTON, A., FUKUMIZU, K. and SRIPERUMBUDUR, B. K. (2009). Discussion of: Brownian distance covariance. *Ann. Appl. Stat.* **3** 1285–1294. MR2752132

GRETTON, A. and GYÖRFI, L. (2010). Consistent nonparametric tests of independence. *J. Mach. Learn. Res.* **11** 1391–1423. MR2645456

GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory* (S. Jain, H. U. Simon and E. Tomita, eds.). *Lecture Notes in Computer Science* **3734** 63–77. Springer, Berlin. MR2255909

GRETTON, A., BORGWARDT, K., RASCH, M., SCHÖLKOPF, B. and SMOLA, A. (2007). A kernel method for the two-sample problem. In *NIPS* 513–520. MIT Press, Cambridge, MA.

GRETTON, A., FUKUMIZU, K., TEO, C. H., SONG, L., SCHÖLKOPF, B. and SMOLA, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* **20** 585–592. MIT Press, Cambridge, MA.

GRETTON, A., FUKUMIZU, K., HARCHAOUI, Z. and SRIPERUMBUDUR, B. (2009). A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems* 22. Curran Associates, Red Hook, NY.

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

GRETTON, A., SRIPERUMBUDUR, B., SEJDINOVIC, D., STRATHMANN, H., BALAKRISHNAN, S., PONTIL, M. and FUKUMIZU, K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems* 25 1214–1222. Curran Associates, Red Hook, NY.

LYONS, R. (2013). Distance covariance in metric spaces. *Ann. Probab.* **41** 3284–3305.

MÜLLER, A. (1997). Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.* **29** 429–443. MR1450938

REED, M. and SIMON, B. (1980). *Methods of Modern Mathematical Physics. I: Functional Analysis*, 2nd ed. Academic Press, San Diego. MR0751959

SCHÖLKOPF, B., SMOLA, A. J. and MÜLLER, K. R. (1997). Kernel principal component analysis. In *ICANN* (W. Gerstner, A. Germond, M. Hasler and J. D. Nicoud, eds.). *Lecture Notes in Computer Science* **1327** 583–588. Springer, Berlin.

SEJDINOVIC, D., GRETTON, A., SRIPERUMBUDUR, B. and FUKUMIZU, K. (2012). Hypothesis testing using pairwise distances and associated kernels. In *Proceedings of the International Conference on Machine Learning* (*ICML*) 1111–1118. Omnipress, New York.

SMOLA, A. J., GRETTON, A., SONG, L. and SCHÖLKOPF, B. (2007). A Hilbert space embedding for distributions. In *Proceedings of the Conference on Algorithmic Learning Theory* (*ALT*) **4754** 13–31. Springer, Berlin.

SRIPERUMBUDUR, B. (2011). Mixture density estimation via Hilbert space embedding of measures. In *Proceedings of the International Symposium on Information Theory* 1027–1030. IEEE, Piscataway, NJ.

SRIPERUMBUDUR, B. K., FUKUMIZU, K. and LANCKRIET, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** 2389–2410. MR2825431

SRIPERUMBUDUR, B., GRETTON, A., FUKUMIZU, K., LANCKRIET, G. and SCHÖLKOPF, B. (2008). Injective Hilbert space embeddings of probability measures. In *Proceedings of the Conference on Learning Theory* (*COLT*) 111–122. Omnipress, New York.

SRIPERUMBUDUR, B., FUKUMIZU, K., GRETTON, A., LANCKRIET, G. and SCHOELKOPF, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems* 22. Curran Associates, Red Hook, NY.

SRIPERUMBUDUR, B. K., GRETTON, A., FUKUMIZU, K., SCHÖLKOPF, B. and LANCKRIET, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.* **11** 1517–1561. MR2645460

SRIPERUMBUDUR, B. K., FUKUMIZU, K., GRETTON, A., SCHÖLKOPF, B. and LANCKRIET, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electron. J. Stat.* **6** 1550–1599. MR2988458

STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York. MR2450103

SZÉKELY, G. and RIZZO, M. (2004). Testing for equal distributions in high dimension. *InterStat* **5**.

SZÉKELY, G. J. and RIZZO, M. L. (2005). A new test for multivariate normality. *J. Multivariate Anal.* **93** 58–80. MR2119764

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

SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127

ZHANG, K., PETERS, J., JANZING, D. and SCHOELKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (*UAI*) 804–813. AUAI Press, Corvallis, Oregon.

D. SEJDINOVIC
A. GRETTON
GATSBY COMPUTATIONAL NEUROSCIENCE UNIT
UNIVERSITY COLLEGE LONDON
17 QUEEN SQUARE
LONDON, WC1N 3AR
UNITED KINGDOM
E-MAIL: dino.sejdinovic@gmail.com
arthur.gretton@gmail.com

B. SRIPERUMBUDUR
STATISTICAL LABORATORY
CENTER FOR MATHEMATICAL SCIENCES
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE CB3 0WB
UNITED KINGDOM
E-MAIL: b.sriperumbudur@statslab.cam.ac.uk

K. FUKUMIZU
INSTITUTE OF STATISTICAL MATHEMATICS
10-3 MIDORICHO, TACHIKAWA
TOKYO 190-8562
JAPAN
E-MAIL: fukumizu@ism.ac.jp