The Annals of Probability
2013, Vol. 41, No. 5, 3284–3305
DOI: 10.1214/12-AOP803
© Institute of Mathematical Statistics, 2013

DISTANCE COVARIANCE IN METRIC SPACES[1]

BY RUSSELL LYONS

Indiana University

We extend the theory of distance (Brownian) covariance from Euclidean spaces, where it was introduced by Székely, Rizzo and Bakirov, to general metric spaces. We show that for testing independence, it is necessary and sufficient that the metric space be of strong negative type. In particular, we show that this holds for separable Hilbert spaces, which answers a question of Kosorok. Instead of the manipulations of Fourier transforms used in the original work, we use elementary inequalities for metric spaces and embeddings in Hilbert spaces.

1. Introduction. Székely, Rizzo and Bakirov (2007) introduced a new statistical test for the following problem: given IID samples of a pair of random variables (X, Y), where X and Y have finite first moments, are X and Y independent? Among the virtues of their test is that it is extremely simple to compute, based merely on a quadratic polynomial of the distances between points in the sample, and that it is consistent against all alternatives (with finite first moments). The test statistic is based on a new notion called "distance covariance" or "distance correlation." The paper by Székely and Rizzo (2009) introduced another new notion, "Brownian covariance," and showed it to be the same as distance covariance. That paper also gave more examples of its use. This latter paper elicited such interest that it was accompanied by a 3-page editorial introduction and 42 pages of comments.

Although the theory presented in those papers is very beautiful, it also gives the impression of being rather technical, relying on various manipulations with Fourier transforms and arcane integrals. Answering a question from Székely (personal communication, 2010), we show that almost the entire theory can be developed for general metric spaces, where it necessarily becomes much more elementary and transparent. A crucial point of the theory is that the distance covariance of (X, Y) is 0 iff X and Y are independent. This does not hold for general metric spaces, but we characterize those for which it does hold. Namely, they are the metric spaces that have what we term "strong negative type."

In fact, negative type (defined in the next paragraph) had arisen already in the work of Székely, Rizzo and Bakirov (2007), hereinafter referred to as SRB. It was

Received August 2011; revised March 2012.

[1] Supported by NSF Grant DMS-10-07244 and Microsoft Research.

MSC2010 subject classifications. Primary 62H20, 62G20, 51K99; secondary 62H15, 30L05.

Key words and phrases. Negative type, hypothesis testing, independence, distance correlation, Brownian covariance.

especially prominent in its predecessors, Székely and Rizzo (2005a, 2005b). The notion of strict negative type is standard, but we need a strengthening of it that we term "strong negative type." [These notions were conflated in SRB and Székely and Rizzo (2005a, 2005b).] The notion of strong negative type was also defined by Klebanov (2005).

The concept of negative type is old, but has enjoyed a resurgence of interest recently due to its uses in theoretical computer science, where embeddings of metric spaces, such as graphs, play a useful role in algorithms; see, for example, Naor (2010) and Deza and Laurent (1997). The fact that Euclidean space has negative type is behind the following charming and venerable puzzle: given n red points x_i and n blue points x'_i in $\mathbb{R}^p$, show that the sum $2\sum_{i,j}\|x_i - x'_j\|$ of the distances between the $2n^2$ ordered pairs of points of opposite color is at least the sum $\sum_{i,j}(\|x_i - x_j\| + \|x'_i - x'_j\|)$ of the distances between the $2n^2$ ordered pairs of points of the same color. The reason the solution is not obvious is that it requires a special property of Euclidean space. In fact, a metric space is defined to have negative type precisely when the preceding inequality holds (points are allowed to be repeated in the lists of n red and n blue points, and n is allowed to be arbitrary). The connection to embeddings is that, as Schoenberg (1937, 1938) showed, negative type is equivalent to a certain property of embeddability into Hilbert space. Indeed, if distance in the puzzle were replaced by squared distance, it would be easy.

If we replace the sums of distances in the puzzle by averages, and then replace the two finite sets of points by two probability distributions (with finite first moments), we arrive at an equivalent condition for a metric space to have negative type. The condition that equality holds only when the two distributions are equal is called "strong negative type." It means that a simple computation involving average distances allows one to distinguish any two probability distributions. Many statistical tests are aimed at distinguishing two probability distributions, or distinguishing two families of distributions. This is what lies directly behind the tests in Székely and Rizzo (2005a, 2005b). It is also what lies behind the papers Bakirov, Rizzo and Székely (2006), SRB, and Székely and Rizzo (2009), but there it is somewhat hidden. We bring this out more clearly in showing how distance covariance allows a test for independence precisely when the two marginal distributions lie in metric spaces of strong negative type. See Székely and Rizzo (2013) for an invited review paper on statistics that are functions of distances between observations.

In Section 2 we define distance covariance and prove its basic properties for general metric spaces. This includes a statistical test for independence, but the test statistic cannot distinguish between independence and the alternative in all metric spaces. In Section 3 we specialize to metric spaces of negative type and show that the test statistic distinguishes between independence and the alternative precisely in the case of spaces of strong negative type. In Section 3 we also sketch short proofs of Schoenberg's theorem and short solutions of the above puzzle (none being original). It turns out that various embeddings into Hilbert space, though necessarily equivalent at the abstract level, are useful for different specific purposes. In both sections, we separate needed results from other interesting results by putting the latter in explicit remarks. We show that the full theory extends to separable-Hilbert-space-valued random variables, which resolves a question of Kosorok (2009). We remark at the end of the paper that if $(\mathscr{X}, d)$ is a metric space of negative type, then $(\mathscr{X}, d^r)$ has strong negative type for all $r \in (0, 1)$; this means that if in a given application one has negative type but not strong negative type (e.g., in an L^1 metric space), then a simple modification of the metric allows the full theory to apply.

2. General metric spaces. Let $(\mathscr{X}, d)$ be a metric space. Let $M(\mathscr{X})$ denote the finite signed Borel measures on $\mathscr{X}$ and $M_1(\mathscr{X})$ be the subset of probability measures. We say that $\mu \in M(\mathscr{X})$ has a *finite first moment* if $\int_{\mathscr{X}} d(o, x)\, d|\mu|(x) < \infty$ for some $o \in \mathscr{X}$. The choice of $o \in \mathscr{X}$ does not matter by virtue of the triangle inequality. If $\mu, \mu' \in M(\mathscr{X})$ both have finite first moments, then $\int d(x, x')\, d(|\mu| \times |\mu'|)(x, x') < \infty$ since $d(x, x') \le d(o, x) + d(o, x')$. Therefore, $\int d(x, x')\, d\mu(x)\, d\mu'(x')$ is defined and finite. In particular, we may define

$$a_\mu(x) := \int d(x, x')\, d\mu(x')$$

and

$$D(\mu) := \int d(x, x')\, d\mu^2(x, x')$$

as finite numbers when $\mu \in M(\mathscr{X})$ has a finite first moment. Also, write

$$d_\mu(x, x') := d(x, x') - a_\mu(x) - a_\mu(x') + D(\mu).$$

The function d_μ is better behaved than d in the following sense:

LEMMA 2.1. *Let $\mathscr{X}$ be any metric space. If $\mu \in M_1(\mathscr{X})$ has a finite first moment, that is, $d(x, x') \in L^1(\mu \times \mu)$, then $d_\mu(x, x') \in L^2(\mu \times \mu)$.*

PROOF. For simplicity, write $a(x) := a_\mu(x)$ and $a := D(\mu)$. Let $X, X' \sim \mu$ be independent. By the triangle inequality, we have

$$|d(x, x') - a(x)| \le a(x'), \tag{2.1}$$

whence

$$\int d_\mu(x, x')\, d\mu^2(x, x') = \mathbf{E}\big[\big(d(X, X') - a(X) - a(X') + a\big)^2\big] \le \mathbf{E}[X_1 X_2],$$

where $X_1 := \max\{|a - 2a(X')|, a\}$ and $X_2 := \max\{|a - 2a(X)|, a\}$. Since X_1 and X_2 are integrable and independent, $X_1 X_2$ is also integrable, with $\mathbf{E}[X_1 X_2] \le 4a^2$. □

The proof of Lemma 2.1 shows that $\|d_\mu\|_2 \le 2D(\mu) = 2\|d\|_1$, but the factor of 2 will be removed in Proposition 2.3.

We call $\mu \in M(\mathscr{X})$ *degenerate* if its support consists of only a single point.

REMARK 2.2. Let $\mu \in M_1(\mathscr{X})$ have finite first moment and be nondegenerate. Although $d_\mu(x, x') < d(x, x')$ for all $x, x' \in \mathscr{X}$, it is not true that $|d_\mu(x, x')| \le d(x, x')$ for all x, x' in the support of μ. To see these, we prove first that

$$a_\mu(x) > D(\mu)/2 \tag{2.2}$$

for all $x \in \mathscr{X}$. Indeed,

$$D(\mu) = \int d(x', x'')\, d\mu^2(x', x'') \le \int \big[d(x', x) + d(x, x'')\big]\, d\mu^2(x', x'') = 2a_\mu(x).$$

Furthermore, if equality holds, then $d(x', x'') = d(x', x) + d(x, x'')$ for all x', x'' in the support of μ. Put $x' = x''$ to get that $x = x'$, contradicting that μ is not degenerate. This proves (2.2). Using (2.2) twice in the definition of d_μ gives $d_\mu < d$. On the other hand, (2.2) also shows that $d_\mu(x, x) < 0 = -d(x, x)$ for all x.

Now let $(\mathscr{Y}, d)$ be another metric space. Let $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ have finite first moments for each of its marginals μ on $\mathscr{X}$ and ν on $\mathscr{Y}$. Define

$$\delta_\theta\big((x, y), (x', y')\big) := d_\mu(x, x')\, d_\nu(y, y').$$

By Lemma 2.1 and the Cauchy–Schwarz inequality, we may define

$$\operatorname{dcov}(\theta) := \int \delta_\theta\big((x, y), (x', y')\big)\, d\theta^2\big((x, y), (x', y')\big).$$

It is immediate from the definition that if θ is a product measure, then $\operatorname{dcov}(\theta) = 0$; the converse statement is not always true and is the key topic of the theory. Metric spaces that satisfy this are characterized in Section 3 as those of strong negative type. Similarly, spaces for which $\operatorname{dcov} \ge 0$ are characterized in Section 3 as those of negative type. SRB call the *square root* of $\operatorname{dcov}(\theta)$ the *distance covariance* of θ, but they work only in the context of Euclidean spaces, where $\operatorname{dcov} \ge 0$. They denote that square root by $\operatorname{dCov}(\theta)$.

When (X, Y) are random variables with distribution $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$, we also write $\operatorname{dcov}(X, Y) := \operatorname{dcov}(\theta)$. If (X, Y) and (X', Y') are independent, both with distribution θ having marginals μ and ν, then

$$\begin{aligned}\operatorname{dcov}(\theta) = \mathbf{E}\big[&\big(d(X, X') - a_\mu(X) - a_\mu(X') + D(\mu)\big)\\ &\times \big(d(Y, Y') - a_\nu(Y) - a_\nu(Y') + D(\nu)\big)\big].\end{aligned}$$

The following generalizes (2.5) of SRB.

PROPOSITION 2.3. *Let $\mathscr{X}$ and $\mathscr{Y}$ be any metric spaces. Let $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ have finite first moments for each of its marginals μ on $\mathscr{X}$ and ν on $\mathscr{Y}$. Let $(X, Y) \sim \theta$. Then*

$$\begin{aligned}|\mathrm{dcov}(X, Y)| &\le \sqrt{\mathrm{dcov}(X, X)\,\mathrm{dcov}(Y, Y)} \\ &\le D(\mu)D(\nu).\end{aligned} \tag{2.3}$$

Furthermore, $\mathrm{dcov}(X, X) = D(\mu)^2$ *iff μ is concentrated on at most two points.*

PROOF. The Cauchy–Schwarz inequality shows (2.3). It remains to show that

$$\mathrm{dcov}(X, X) \le D(\mu)^2 \tag{2.4}$$

and to analyze the case of equality. As before, write $a(x) := a_\mu(x)$ and $a := D(\mu)$. By (2.1), we have

$$\mathbf{E}\big[|d(X, X') - a(X)|a(X)\big] \le \mathbf{E}\big[a(X')a(X)\big] = a^2 < \infty,$$

whence $\mathbf{E}[[d(X, X') - a(X)]a(X)] = 0$ by Fubini's theorem (i.e., condition on X). Similarly, $\mathbf{E}[[d(X, X') - a(X')]a(X')] = 0$. Thus, expanding the square in $\mathrm{dcov}(X, X) = \mathbf{E}[(d(X, X') - a(X) - a(X') + a)^2]$ and replacing $d(X, X')^2$ there by the larger quantity $d(X, X')[a(X) + a(X')]$ yields $[d(X, X') - a(X)]a(X) + [d(X, X') - a(X')]a(X')$ plus other terms that are individually integrable with integrals summing to a^2. This shows the inequality (2.4). Furthermore, it shows that equality holds iff for all points x, x' in the support of μ, if $d(x, x') \ne 0$, then $d(x, x') = a(x) + a(x')$. Since the right-hand side equals $\int[d(x, o) + d(o, x')]\,d\mu(o)$, it follows that $d(x, x') = d(x, o) + d(o, x')$ for all o in the support of μ. If there is an $o \ne x, x'$ in the support of μ, then we similarly have that $d(x, o) = d(x, x') + d(x', o)$. Adding these equations together shows that $d(o, x') = 0$, a contradiction. That is, if $\mathrm{dcov}(X, X) = D(\mu)^2$, then the support of μ has size 1 or 2. The converse is clear. □

The next proposition generalizes Theorem 4(i) of SRB.

PROPOSITION 2.4. *If* $\mathrm{dcov}(X, X) = 0$, *then X is degenerate.*

PROOF. As before, write $a(x) := a_\mu(x)$ and $a := D(\mu)$, where $X \sim \mu$. The hypothesis implies that $d(X, X') - a(X) - a(X') + a = 0$ a.s. Since all functions here are continuous, we have $d(x, x') - a(x) - a(x') + a = 0$ for all x, x' in the support of μ. Put $x = x'$ to deduce that for all x in the support of μ, we have $a(x) = a/2$. Therefore, $d(X, X') = 0$ a.s. □

Assume that μ and ν are nondegenerate. Then the right-hand side of (2.3) is not 0; the quotient $\mathrm{dcov}(\theta)/[D(\mu)D(\nu)]$ is the square of what is called the *distance correlation* of θ in SRB. In SRB, this quotient is always nonnegative.

This next proposition extends Theorem 3(iii) of SRB.

PROPOSITION 2.5. *If μ and ν are nondegenerate and equality holds in* (2.3), *then for some $c > 0$, there is a continuous map $f : \mathscr{X} \to \mathscr{Y}$ such that for all x, x' in the support of μ, we have $d(x, x') = cd(f(x), f(x'))$ and $y = f(x)$ for θ-a.e. (x, y).*

PROOF. Write $a(x) := a_\mu(x)$, $a := D(\mu)$, $b(y) := a_\nu(y)$ and $b := D(\nu)$. Equality holds in (2.3) iff there is some constant c such that

$$d(x, x') - a(x) - a(x') + a = c\bigl(d(y, y') - b(y) - b(y') + b\bigr)$$

for θ^2-a.e. (x, y), (x', y'), that is,

$$d(x, x') - cd(y, y') = a(x) - cb(y) + a(x') - cb(y') + cb - a.$$

Since all functions here are continuous, the same holds for all (x, y), (x', y') in the support of θ. Put $(x, y) = (x', y')$ to deduce that for all (x, y) in the support of θ, we have $a(x) - cb(y) = (a - cb)/2$. This means that $d(x, x') = cd(y, y')$ θ^2-a.s. The conclusion follows. □

We now extend Theorem 2 of SRB.

PROPOSITION 2.6. *Let $\mathscr{X}$ and $\mathscr{Y}$ be metric spaces. Let $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ have marginals with finite first moment. Let θ_n be the* (*random*) *empirical measure of the first n samples from an infinite sequence of IID samples of θ. Then* $\operatorname{dcov}(\theta_n) \to \operatorname{dcov}(\theta)$ *a.s.*

PROOF. Let $(X^i, Y^i) \sim \theta$ be independent for $1 \le i \le 6$. Write

$$f(z_1, z_2, z_3, z_4) := d(z_1, z_2) - d(z_1, z_3) - d(z_2, z_4) + d(z_3, z_4).$$

Here, $z_i \in \mathscr{X}$ or $z_i \in \mathscr{Y}$. The triangle inequality gives that

$$\bigl|f(z_1, z_2, z_3, z_4)\bigr| \le g(z_1, z_3, z_4) := 2\max\bigl\{d(z_3, z_4), d(z_1, z_3)\bigr\}$$

and

$$\bigl|f(z_1, z_2, z_3, z_4)\bigr| \le g(z_2, z_4, z_3) = 2\max\bigl\{d(z_3, z_4), d(z_2, z_4)\bigr\}.$$

Since $g(X^1, X^3, X^4)$ and $g(Y^2, Y^6, Y^5)$ are integrable and independent, it follows that

$$h\bigl((X^1, Y^1), \dots, (X^6, Y^6)\bigr) := f(X^1, X^2, X^3, X^4) f(Y^1, Y^2, Y^5, Y^6)$$

is integrable. Fubini's theorem thus shows that its expectation equals $\operatorname{dcov}(\theta)$. Similarly, $\operatorname{dcov}(\theta_n)$ are the V-statistics for the kernel h of degree 6. Hence, the result follows. □

The proof of Proposition 2.6 for general metric spaces is more straightforward if second moments are finite, as in Remark 3 of SRB.

We next extend Theorem 5 of SRB.

THEOREM 2.7. *Let $\mathscr{X}$, $\mathscr{Y}$ be metric spaces. Let $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ have marginals μ, ν with finite first moment. Let θ_n be the empirical measure of the first n samples from an infinite sequence of IID samples of θ. Let λ_i be the eigenvalues (with multiplicity) of the map that sends $F \in L^2(\theta)$ to the function*

$$(x, y) \mapsto \int \delta_\theta\big((x, y), (x', y')\big) F(x', y')\, d\theta(x', y').$$

If $\theta = \mu \times \nu$, then $n\,\mathrm{dcov}(\theta_n) \Rightarrow \sum_i \lambda_i Z_i^2$, where Z_i are IID standard normal random variables and $\sum_i \lambda_i = D(\mu)D(\nu)$.

PROOF. We use the same notation as in the proof of Proposition 2.6. That proof shows that h is integrable when μ and ν have finite first moments; the case $X^i = Y^i$ shows then that $f(X^1, X^2, X^3, X^4)$ has finite second moment. Therefore, when $\theta = \mu \times \nu$, $h((X^1, Y^1), \dots, (X^6, Y^6))$ has finite second moment.

Assume now that $\theta = \mu \times \nu$. Then kernel h is degenerate of order 1. Let $\bar{h}$ be the symmetrized version of h. That is, for $z_i \in \mathscr{X} \times \mathscr{Y}$, we let $\bar{h}(z_1, \dots, z_6)$ be the average of $h(z_{\sigma(1)}, \dots, z_{\sigma(6)})$ over all permutations σ of $\{1, \dots, 6\}$. Then since $\theta = \mu \times \nu$,

$$\begin{aligned}\bar{h}_2\big((x, y), (x', y')\big) &:= \mathbf{E}\big[\bar{h}\big((x, y), (x', y'), (X^3, Y^3), \dots, (X^6, Y^6)\big)\big]\\ &= \delta_\theta\big((x, y), (x', y')\big)/15.\end{aligned}$$

Hence, the result follows from the theory of degenerate V-statistics [compare Theorem 5.5.2 in Serfling (1980) or Example 12.11 in van der Vaart (1998) for the case of U-statistics]. Finally, we have $\sum \lambda_i = \int \delta_\theta((x, y), (x, y))\, d\theta(x, y) = D(\mu)D(\nu)$ since $\theta = \mu \times \nu$. □

COROLLARY 2.8. *Let $\mathscr{X}$, $\mathscr{Y}$ be metric spaces. Let $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ have nondegenerate marginals μ, ν with finite first moment. Let θ_n be the empirical measure of the first n samples from an infinite sequence of IID samples of θ. Let μ_n, ν_n be the marginals of θ_n. If $\theta = \mu \times \nu$, then*

$$\frac{n\,\mathrm{dcov}(\theta_n)}{D(\mu_n)D(\nu_n)} \Rightarrow \frac{\sum_i \lambda_i Z_i^2}{D(\mu)D(\nu)}, \tag{2.5}$$

where λ_i and Z_i are as in Theorem 2.7 and the right-hand side has expectation 1. *If* $\mathrm{dcov}(\theta) \neq 0$, *then the left-hand side of* (2.5) *tends to* $\pm\infty$ *a.s.*

PROOF. Since $D(\mu_n)$ and $D(\nu_n)$ are V-statistics, we have $D(\mu_n) \to D(\mu)$ and $D(\nu_n) \to D(\nu)$ a.s. Thus, the first case follows from Theorem 2.7. The second case follows from Proposition 2.6. □

REMARK 2.9. Since $\theta = \mu \times \nu$, the map in Theorem 2.7 is the tensor product of the maps

$$L^2(\mu) \ni F \mapsto \left(x \mapsto \int d_\mu(x, x') F(x')\, d\mu(x')\right)$$

and

$$L^2(\nu) \ni F \mapsto \left(y \mapsto \int d_\nu(y, y')F(y')\,d\nu(y')\right).$$

Therefore, the eigenvalues λ_i are the products of the eigenvalues of these two maps.

3. Spaces of negative type.

Corollary 2.8 is incomplete in that it does not specify what happens when $\mathrm{dcov}(\theta) = 0$ and θ is not a product measure. In order for the statistics $\mathrm{dcov}(\theta_n)$ to give a test for independence that is consistent against all alternatives, it suffices to rule out this missing case. In this section, we show that this case never arises for metric spaces of strong negative type, but otherwise it does. This will require the development of several other theorems of independent interest. We intersperse these theorems with their specializations to Euclidean space.

The puzzle we recalled in the Introduction can be stated the following way for a metric space $(\mathscr{X}, d)$: let $n \geq 1$ and $x_1, \dots, x_{2n} \in \mathscr{X}$. Write α_i for the indicator that x_i is red minus the indicator that x_i is blue. Then $\sum_{i=1}^{2n} \alpha_i = 0$ and

$$\sum_{i,j \leq 2n} \alpha_i \alpha_j d(x_i, x_j) \leq 0.$$

By considering repetitions of x_i and taking limits, we arrive at a superficially more general property: for all $n \geq 1$, $x_1, \dots, x_n \in \mathscr{X}$, and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ with $\sum_{i=1}^{n} \alpha_i = 0$, we have

$$\sum_{i,j \leq n} \alpha_i \alpha_j d(x_i, x_j) \leq 0. \tag{3.1}$$

We say that $(\mathscr{X}, d)$ has *negative type* if this property holds. A list of metric spaces of negative type appears as Theorem 3.6 of Meckes (2013); in particular, this includes all L^p spaces for $1 \leq p \leq 2$. On the other hand, $\mathbb{R}^n$ with the ℓ^p-metric is not of negative type whenever $3 \leq n \leq \infty$ and $2 < p \leq \infty$, as proved by Dor (1976) combined with Theorem 2 of Bretagnolle, Dacunha-Castelle and Krivine (1965/1966); see Koldobsky and Lonke (1999) for an extension to spaces that include some Orlicz spaces, among others.

If we define the $n \times n$ distance matrix K whose (i, j) entry is $d(x_i, x_j)$, then (3.1) says, by definition, that K is conditionally negative semidefinite. This explains the name "negative type." We can construct another matrix $\bar{K}$ from K that is negative semidefinite as follows: let P be the orthogonal projection of $\mathbb{R}^n$ onto the orthocomplement of the constant vectors. Then as operators, $\bar{K} := PKP$. Let μ_n be the empirical measure of $x_1, \dots, x_n$. The (i, j) entry of $\bar{K}$ is easily verified to be $d_{\mu_n}(x_i, x_j)$, which begins to explain the appearance of d_μ in Section 2. We write $\bar{K} \leq 0$ to mean that $\bar{K}$ is negative semidefinite.

If $\mathscr{X}$ and $\mathscr{Y}$ are both metric spaces of negative type and $(x_i, y_i) \in \mathscr{X} \times \mathscr{Y}$, then let K and L be the distance matrices for x_i and y_i, respectively. Let θ_n be the empirical measure of the sequence $\langle (x_i, y_i); 1 \le i \le n\rangle$. We have $\bar{K} \le 0$ and $\bar{L} \le 0$, whence $\operatorname{tr}(\bar{K}\bar{L}) = \operatorname{tr}(\sqrt{-\bar{K}}\sqrt{-\bar{L}}\sqrt{-\bar{L}}\sqrt{-\bar{K}}) \ge 0$, that is,

$$0 \le \operatorname{tr}(\bar{K}\bar{L}) = n^2 \operatorname{dcov}(\theta_n).$$

This begins to explain the origin of dcov. To go further, we use embeddings into Hilbert space.

Now $\mathscr{X}$ is of negative type iff there is a Hilbert space H and a map $\phi : \mathscr{X} \to H$ such that $\forall x, x' \in \mathscr{X}$ $d(x, x') = \|\phi(x) - \phi(x')\|^2$, as shown by Schoenberg (1937, 1938). We sketch two proofs of Schoenberg's theorem: given such a ϕ, (3.1) is easy to verify; see (3.4) below. For the converse, consider $x_1, \dots, x_n \in \mathscr{X}$. Since $\bar{K} \le 0$, there are vectors $v_i \in \mathbb{R}^n$ such that $\langle v_i, v_j\rangle$ is the (i, j)-entry of $-\bar{K}$ for all i, j (the matrix $\sqrt{-\bar{K}}$ has v_i for its ith column). Computing $\|v_i - v_j\|^2$ then yields $\|v_i/\sqrt{2} - v_j/\sqrt{2}\|^2 = d(x_i, x_j)$. This provides a map ϕ defined on the points $x_1, \dots, x_n$. When we increase the domain of such a ϕ, the distances of the images already defined are preserved, whence we may embed all these images in a fixed Hilbert space. If $\mathscr{X}$ is separable, we may thus define ϕ on a countable dense subset by induction, and then extend by continuity. In general and alternatively, define

$$d_o(x, x') := [d(x, o) + d(o, x') - d(x, x')]/2$$

for some fixed $o \in \mathscr{X}$. Let V be the finitely supported functions on $\mathscr{X}$. The fact that $\mathscr{X}$ is of negative type implies that $\langle f, g\rangle := \sum_{x,x' \in \mathscr{X}} f(x)g(x')\, d_o(x, x')$ is a semi-inner product on V. The Cauchy–Schwarz inequality implies that $V_0 := \{f \in V; \langle f, f\rangle = 0\}$ is a subspace of V. Let H be the completion of V/V_0. Then the map $\phi : x \mapsto \mathbf{1}_{\{x\}} + V_0$ has the property desired. Note that H is separable when $\mathscr{X}$ is.

Of course, any two isometric embeddings $\phi_1, \phi_2 : (\mathscr{X}, d^{1/2}) \to H$ are equivalent in the sense that there exists an isometry $g : H_1 \to H_2$ such that $\phi_2 = g \circ \phi_1$, where H_i is the closed affine span of the image of ϕ_i. To see this, define $g(\phi_1(x)) := \phi_2(x)$ for $x \in \mathscr{X}$, extend by affine linearity (which is well defined by a property of Euclidean space), and then extend by continuity. We shall call an isometric embedding $\phi : (\mathscr{X}, d^{1/2}) \to H$ simply an *embedding*.

A direct proof that $\mathbb{R}^n$ is of negative type is the following. When $n = 1$, define $\phi(x)$ to be the function $\mathbf{1}_{[0,\infty)} - \mathbf{1}_{[x,\infty)}$ in $L^2(\mathbb{R}, \lambda)$, where λ is the Lebesgue measure. This is easily seen to have the desired property. When $n \ge 2$, define $f_x(s) := \|x - s\|^{-(n-1)/2}$ and $g_x := f_x - f_{\mathbf{0}}$ for $x \in \mathbb{R}^n$. Then $g_x \in L^2(\mathbb{R}^n, \lambda^n)$, as calculus shows [for large s, we have $g_x(s) = O(\|s\|^{-(n+1)/2})$]. Furthermore, there is a constant c such that $\|g_x\|_2 = c\|x\|^{1/2}$ by homogeneity, whence translation invariance gives $\|g_x - g_{x'}\|_2 = \|g_{x-x'}\|_2 = c\|x - x'\|^{1/2}$, so that $\phi(x) := g_x/c$ has the desired property. Call this embedding the *Riesz embedding* since $f_x(s)$ is a Riesz kernel.

Another embedding ϕ for $\mathbb{R}^n$ is as follows: $\phi(x)$ is the function $s \mapsto c(1 - e^{-is\cdot x})$ in $L^2(F\lambda^n)$ for some constant c, where $F(s) := \|s\|^{-(n+1)}$. See Lemma 1 of Székely and Rizzo (2005a) for a proof. This is the Fourier transform of the Riesz embedding, in other words, the composition of the Riesz embedding with the Fourier isometry. We shall refer to this embedding as the *Fourier embedding*.

Other important embeddings use Brownian motion. When $n = 1$, let B_x be Brownian motion defined for $x \in \mathbb{R}$ with $B_0 = \mathbf{0}$. We may then define $\phi(x) := B_x$, thought of as a function in $L^2(\mathbf{P})$ for some probability measure $\mathbf{P}$. Likewise, the case $n \geq 2$ can be accomplished by using Lévy's multiparameter Brownian motion. We shall refer to these embeddings as the *Brownian embeddings*. Sample-path continuity of these Brownian motions plays no role for us; only their Gaussian structure matters. In fact, their existence depends only on the fact that $\mathbb{R}^n$ has negative type.

An embedding that does not rely on calculation goes as follows: let σ be the (infinite) Borel measure on half-spaces $S \subset \mathbb{R}^n$ that is invariant under translations and rotations, normalized so that

$$\sigma(\{\mathbf{0} \in S, x \notin S\}) = \|x\|/2 \tag{3.2}$$

for $\|x\| = 1$. If we parametrize half-spaces as $S = \{x \in \mathbb{R}^n; z \cdot x \leq s\}$ with $z \in \mathbb{S}^{n-1}$ and $s \in \mathbb{R}$, then $\sigma = c_n \Omega_n \times \lambda$ for some constant c_n, where Ω_n is volume measure on $\mathbb{S}^{n-1}$. Scaling shows that (3.2) holds for all x. Now let $\phi(x)$ be the function $S \mapsto \mathbf{1}_S(\mathbf{0}) - \mathbf{1}_S(x)$ in $L^2(\sigma)$. We call this the *Crofton embedding*, as Crofton (1868) was the first to give a formula for the distance of points in the plane in terms of lines intersecting the segment joining them.

We return now to general metric spaces of negative type. Suppose that $\mu_1, \mu_2 \in M_1(\mathscr{X})$ have finite first moments. By approximating μ_i by probability measures of finite support (e.g., IID samples give V-statistics), we see that when $\mathscr{X}$ has negative type,

$$D(\mu_1 - \mu_2) \leq 0. \tag{3.3}$$

We say that $(\mathscr{X}, d)$ has *strong negative type* if it has negative type and equality holds in (3.3) only when $\mu_1 = \mu_2$. When μ_i are restricted to measures of finite support, then this is the condition that $(\mathscr{X}, d)$ be of *strict negative type*. A simple example of a metric space of nonstrict negative type is ℓ^1 on a 2-point space, that is, $\mathbb{R}^2$ with the ℓ^1-metric. See Remark 3.3 below for an example of a metric space of strict but not strong negative type.

Consider an embedding ϕ as above. Define the (linear) barycenter map $\beta = \beta_\phi : \mu \mapsto \int \phi(x)\, d\mu(x)$ on the set of measures $\mu \in M(\mathscr{X})$ with finite first moment. [Although it suffices that $\int d(o,x)^{1/2}\, d|\mu|(x) < \infty$ to define $\beta(\mu)$, this will not suffice for our purposes.] Note that

$$\iint d(x_1, x_2)\, d\mu_1(x_1)\, d\mu_2(x_2) = -2\langle \beta(\mu_1), \beta(\mu_2) \rangle$$

when $\mu_i \in M(\mathscr{X})$ satisfy $\mu_i(\mathscr{X}) = 0$. In particular,

$$D(\mu) = -2\|\beta(\mu)\|^2 \tag{3.4}$$

when $\mu \in M(\mathscr{X})$ satisfies $\mu(\mathscr{X}) = 0$. Thus, we have the following:

PROPOSITION 3.1. *Let $\mathscr{X}$ have negative type as witnessed by the embedding ϕ. Then $\mathscr{X}$ is of strong negative type iff the barycenter map β_ϕ is injective on the set of probability measures on $\mathscr{X}$ with finite first moment.*

For example, Euclidean spaces have strong negative type; this is most directly seen via the Fourier embedding, since then $\beta(\mu)$ is the function $s \mapsto c(1 - \widehat{\mu}(s))$, where $\widehat{\mu}$ is the Fourier transform of $\mu \in M_1(\mathbb{R}^n)$. The fact that μ is determined by its Fourier transform then implies that Euclidean space has strong negative type. Alternatively, one can see that Euclidean spaces have strong negative type via the Crofton embedding and the Cramér–Wold device, but the only decent proof of that device uses Fourier transforms. (Of course, in one dimension, the Crofton embedding is simple and easily shows that $\mathbb{R}$ has strong negative type without the use of Fourier transforms.) The barycenter of μ for the Riesz embedding is essentially the Riesz potential of μ; more precisely, if μ and μ' are probability measures with finite first moment, then up to a constant factor, $\beta(\mu - \mu')$ is the Riesz potential of $\mu - \mu'$ for the exponent $(n-1)/2$. However, Riesz potentials will not concern us here.

REMARK 3.2. Another way of saying Proposition 3.1 is that a metric space $(\mathscr{X}, d)$ has strong negative type iff the map $(\mu_1, \mu_2) \mapsto \sqrt{-D(\mu_1 - \mu_2)/2}$ is a metric on the set of probability measures on $\mathscr{X}$ with finite first moment, in which case it extends the metric on $(\mathscr{X}, d^{1/2})$ when we identify $x \in \mathscr{X}$ with the point mass at x. This metric is referred to by Klebanov (2005) as an "$\mathfrak{N}$-distance."

REMARK 3.3. Here we give an example of a metric space of strict negative type that is not of strong negative type. In fact, it fails the condition for probability measures with countable support. The question amounts to whether, given a subset of a Hilbert space in which no 3 points form an obtuse triangle and such that the barycenter of every finitely supported probability measure determines the measure uniquely, the barycenter of every probability measure determines the measure uniquely. The answer is no. For example, let $\langle e_i \rangle$ be an orthonormal basis of a Hilbert space. The desired subset consists of the vectors

$$e_1,$$

$$e_1 + e_2/2,$$

$$e_2 + e_3,$$

$$e_3 + e_4/2,$$

$$e_4 + e_5,$$

$$e_5 + e_6/2,$$

etc. It is obvious that finite convex combinations are unique and that there are no obtuse angles. But if v_n denotes the nth vector, then

$$v_1/2 + v_3/4 + v_5/8 + \cdots = v_2/2 + v_4/4 + v_6/8 + \cdots.$$

REMARK 3.4. If $\mathscr{X}$ is a metric space of negative type, then $\alpha : \mu \mapsto a_\mu$ is injective on $\mu \in M_1(\mathscr{X})$ with finite first moment iff $\mathscr{X}$ has strong negative type. Part of this statement is contained in Theorem 4.1 of Klebanov (2005); this same part occurs later in Theorem 3.6 of Nickolas and Wolf (2009). To prove the equivalence, let ϕ be an embedding of $\mathscr{X}$ such that $\mathbf{0}$ lies in the image of ϕ, which we may achieve by translation. Then

$$a_\mu(x) = \|\phi(x)\|^2 - 2\langle \phi(x), \beta(\mu)\rangle + \int \|\phi(x')\|^2 \, d\mu(x'),$$

whence $a_\mu = a_{\mu'}$ iff $\langle \phi(x), \beta(\mu)\rangle = \langle \phi(x), \beta(\mu')\rangle$ for all x [first use x so that $\phi(x) = \mathbf{0}$] iff $\langle z, \beta(\mu)\rangle = \langle z, \beta(\mu')\rangle$ for all z in the closed linear span of the image of ϕ iff $\beta(\mu) = \beta(\mu')$. Now apply Proposition 3.1. On the other hand, there are metric spaces not of negative type for which α is injective on the probability measures; for example, take a finite metric space in which the distances to a fixed point are linearly independent. The map α is injective also for all separable L^p spaces $(1 < p < \infty)$; see Linde (1986b) or Gorin and Koldobskiĭ (1987).

Given an H-valued random variable Z with finite first moment, we define its *variance* to be $\mathrm{Var}(Z) := \mathbf{E}[\|Z - \mathbf{E}[Z]\|^2]$.

PROPOSITION 3.5. *If $\mathscr{X}$ has negative type as witnessed by the embedding ϕ and $\mu \in M_1(\mathscr{X})$ has finite first moment, then for all $x, x' \in \mathscr{X}$,*

$$a_\mu(x) = \|\phi(x) - \beta_\phi(\mu)\|^2 + D(\mu)/2,$$

$D(\mu) = 2\,\mathrm{Var}(\phi(X))$ *if $X \sim \mu$, and*

$$d_\mu(x, x') = -2\langle \phi(x) - \beta_\phi(\mu), \phi(x') - \beta_\phi(\mu)\rangle.$$

PROOF. Let $X \sim \mu$. We have

$$\begin{aligned} a_\mu(x) = \mathbf{E}[d(x, X)] &= \mathbf{E}[\|\phi(x) - \phi(X)\|^2] \\ &= \mathbf{E}[\|(\phi(x) - \beta(\mu)) - (\phi(X) - \beta(\mu))\|^2] \\ &= \|\phi(x) - \beta_\phi(\mu)\|^2 + \mathrm{Var}(\phi(X)). \end{aligned}$$

Integrating over x gives the first two identities. Substituting the first identity into the definition of d_μ gives the last identity. □

For simplicity, we may, without loss of generality, work only with real Hilbert spaces. Let $\mathscr{X}$ and $\mathscr{Y}$ be metric spaces of negative type, witnessed by the embeddings ϕ and ψ, respectively. Consider the tensor embedding $(x, y) \mapsto \phi(x) \otimes \psi(y)$ of $\mathscr{X} \times \mathscr{Y} \to H \otimes H$. This will be the key to analyzing when $\operatorname{dcov}(\theta) = 0$. Recall that the inner product on $H \otimes H$ satisfies $\langle h_1 \otimes h_1', h_2 \otimes h_2' \rangle := \langle h_1, h_2 \rangle \langle h_1', h_2' \rangle$.

REMARK 3.6. Although we shall not need it, we may give $\mathscr{X} \times \mathscr{Y}$ the associated "metric"

$$d_{\phi\otimes\psi}\big((x, y), (x', y')\big) := \|\phi(x) \otimes \psi(y) - \phi(x') \otimes \psi(y')\|^2,$$

so necessarily it is of negative type. Actually, one can check that this need not satisfy the triangle inequality, but, following a suggestion of ours, Leonard Schulman (personal communication, 2010) showed that it is indeed a metric when the images of ϕ and ψ both contain the origin. Since we may translate ϕ and ψ so that this holds, we may take this to be a metric if we wish. In this case, one can also express $d_{\phi\otimes\psi}$ in terms of the original metrics on $\mathscr{X}$ and $\mathscr{Y}$. However, we shall not use $d_{\phi\otimes\psi}$ anywhere other than in Remarks 3.10 and 3.14.

PROPOSITION 3.7. *Let $\mathscr{X}$, $\mathscr{Y}$ have negative type as witnessed by the embeddings ϕ, ψ. Let $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ have marginals $\mu \in M_1(\mathscr{X})$ and $\nu \in M_1(\mathscr{Y})$, both with finite first moment. Then $\theta \circ (\phi \otimes \psi)^{-1}$ has finite first moment, so that $\beta_{\phi\otimes\psi}(\theta)$ is defined, and we have that*

$$\operatorname{dcov}(\theta) = 4\|\beta_{\phi\otimes\psi}(\theta - \mu \times \nu)\|^2.$$

PROOF. Write $\widehat{\phi} := \phi - \beta_\phi(\mu)$ and $\widehat{\psi} := \psi - \beta_\psi(\nu)$. By Proposition 3.5, we have

$$\begin{aligned}\operatorname{dcov}(\theta) &= 4\int \langle \widehat{\phi}(x), \widehat{\phi}(x') \rangle \langle \widehat{\psi}(y), \widehat{\psi}(y') \rangle \, d\theta^2\big((x, y), (x', y')\big) \\ &= 4\int \langle \widehat{\phi}(x) \otimes \widehat{\psi}(y), \widehat{\phi}(x') \otimes \widehat{\psi}(y') \rangle \, d\theta^2\big((x, y), (x', y')\big) \\ &= 4\|\beta_{\widehat{\phi}\otimes\widehat{\psi}}(\theta)\|^2.\end{aligned}$$

In addition, since $\|\phi(x)\| \in L^2(\mu)$ and $\|\psi(y)\| \in L^2(\nu)$, we have $\|\phi(x) \otimes \psi(y)\| \in L^1(\theta)$ by the Cauchy–Schwarz inequality, whence $\beta_{\phi\otimes\psi}(\theta)$ is defined and

$$\begin{aligned}\beta_{\widehat{\phi}\otimes\widehat{\psi}}(\theta) &= \int \widehat{\phi}(x) \otimes \widehat{\psi}(y) \, d\theta(x, y) \\ &= \int \big(\phi(x) - \beta_\phi(\mu)\big) \otimes \big(\psi(y) - \beta_\psi(\nu)\big) \, d\theta(x, y) \\ &= \int \phi(x) \otimes \psi(y) \, d\theta(x, y) - \beta_\phi(\mu) \otimes \beta_\psi(\nu) \\ &= \beta_{\phi\otimes\psi}(\theta - \mu \times \nu).\end{aligned}$$

□

In the special case where $\mathscr{X}$ and $\mathscr{Y}$ are Euclidean spaces and the embeddings ϕ, ψ are the Fourier embeddings, Proposition 3.7 shows that dcov coincides with (the square of) the original definition of distance covariance in SRB [see (2.6) there], while if the embeddings are the Brownian embeddings, then Proposition 3.7 shows that distance covariance is the same as Brownian covariance [Theorem 8 of Székely and Rizzo (2009); the condition there that X and Y have finite second moments is thus seen to be superfluous]. The Crofton embedding gives

$$\beta_{\phi\otimes\psi}(\theta - \mu \times \nu) : (z, s, w, t) \\ \mapsto c_p c_q\big[\theta(z \cdot x \le s, w \cdot y \le t) - \mu(z \cdot x \le s)\nu(w \cdot y \le t)\big]$$

for $\theta \in M_1(\mathbb{R}^p \times \mathbb{R}^q)$ with marginals μ, ν having finite first moments, whence for $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$, Proposition 3.7 shows that

$$\begin{aligned}&\operatorname{dcov}(X, Y)\\ &\quad = 4c_p c_q \iint \big|\mathbf{P}[z \cdot X \le s, w \cdot Y \le t] \\ &\qquad\qquad - \mathbf{P}[z \cdot X \le s]\mathbf{P}[w \cdot Y \le t]\big|^2 \, d(\Omega_p \times \Omega_q)(z, w)\, d\lambda^2(s, t).\end{aligned}$$

When $p = q = 1$, this formula was shown to us by Gábor Székely (personal communication, 2010).

Write $M^1(\mathscr{X})$ for the subset of $\mu \in M(\mathscr{X})$ such that $|\mu|$ has a finite first moment. Write $M^{1,1}(\mathscr{X} \times \mathscr{Y})$ for the subset of $\theta \in M(\mathscr{X} \times \mathscr{Y})$ such that both marginals of $|\theta|$ have finite first moment.

LEMMA 3.8. *Let $\mathscr{X}$, $\mathscr{Y}$ have negative type as witnessed by the embeddings ϕ, ψ. If ϕ and ψ have the property that β_ϕ and β_ψ are injective on both $M^1(\mathscr{X})$ and $M^1(\mathscr{Y})$* (*not merely on the probability measures*), *then $\beta_{\phi\otimes\psi}$ is injective on $M^{1,1}(\mathscr{X} \times \mathscr{Y})$.*

PROOF. Let $\theta \in M^{1,1}(\mathscr{X} \times \mathscr{Y})$ satisfy $\beta_{\phi\otimes\psi}(\theta) = 0$. For $k \in H$, define the bounded linear map $T_k : H \otimes H \to H$ by linearity, continuity and

$$T_k(u \otimes v) := \langle u, k\rangle v.$$

More precisely, one uses the above definition on $e_i \otimes e_j$ for an orthonormal basis $\{e_i\}$ of H and then extends. Also, define

$$\nu_k(B) := \int \langle \phi(x), k\rangle \mathbf{1}_B(y)\, d\theta(x, y) \qquad (B \subseteq \mathscr{Y} \text{ Borel}),$$

so that

$$\begin{aligned}\beta_\psi(\nu_k) &= \int \langle \phi(x), k\rangle \psi(y)\, d\theta(x, y) = \int T_k\big(\phi(x) \otimes \psi(y)\big)\, d\theta(x, y)\\ &= T_k\big(\beta_{\phi\otimes\psi}(\theta)\big) = 0.\end{aligned}$$

This implies that $\nu_k = 0$ by injectivity of β_ψ. As this is valid for each $k \in H$, we obtain that for every Borel $B \subseteq \mathscr{Y}$,

$$\int \phi(x)\mathbf{1}_B(y)\, d\theta(x, y) = 0.$$

Defining

$$\mu_B(A) := \theta(A \times B) \qquad (A \subseteq \mathscr{X} \text{ Borel}),$$

we have $\beta_\phi(\mu_B) = \int \phi(x)\mathbf{1}_B(y)\, d\theta(x, y) = 0$, whence $\mu_B = 0$ by injectivity of β_ϕ. In other words, $\theta(A \times B) = 0$ for every pair of Borel sets A and B. Since such product sets generate the product σ-field on $\mathscr{X} \times \mathscr{Y}$, it follows that $\theta = 0$. □

LEMMA 3.9. *Let $\mathscr{X}$ have strong negative type. There exists an embedding ϕ so that β_ϕ is injective on $M^1(\mathscr{X})$ (not merely on the probability measures).*

PROOF. If $\phi : \mathscr{X} \to H$ is an embedding that induces an injective barycenter map on $M_1^1(\mathscr{X})$, then the map $x \mapsto (\phi(x), 1) \in H \times \mathbb{R}$ is an embedding that induces an injective barycenter map on $M^1(\mathscr{X})$. □

REMARK 3.10. We may choose the embeddings so that $d_{\phi\otimes\psi}$ is a metric and $\beta_{\phi\otimes\psi}$ is injective on $M^1(\mathscr{X} \times \mathscr{Y})$, which yields that $d_{\phi\otimes\psi}$ is of strong negative type by Proposition 3.1. Indeed, first translate ϕ and ψ so that each contains $\mathbf{0}$ in its image. This makes $d_{\phi\otimes\psi}$ a metric by Remark 3.6. Then use the embedding $x \mapsto (\phi(x), 1)$ and likewise for ψ. This does not change the metric.

As we observed in Section 2, it is immediate from the definition that if θ is a product measure, then $\operatorname{dcov}(\theta) = 0$. A converse and the key result of the theory holds for metric spaces of strong negative type:

THEOREM 3.11. *Suppose that both $\mathscr{X}$ and $\mathscr{Y}$ have strong negative type and θ is a probability measure on $\mathscr{X} \times \mathscr{Y}$ whose marginals have finite first moment. If* $\operatorname{dcov}(\theta) = 0$, *then θ is a product measure.*

This is an immediate corollary of Proposition 3.7 and Lemmas 3.8 and 3.9. Therefore, Corollary 2.8 gives a test for independence that is consistent against all alternatives when $\mathscr{X}$ and $\mathscr{Y}$ both have strong negative type. See Theorem 6 of SRB for the significance levels of the test.

For the Fourier embedding of Euclidean space, Theorem 3.11 amounts to the fact that $\theta = \mu \times \nu$ if the Fourier transform of θ is the (tensor) product of the Fourier transforms of μ and ν. This was the motivation presented in SRB for dCov.

REMARK 3.12. In the case of categorical data, we may embed each data space as a simplex with edges of unit length. Let the corresponding Hilbert-space vectors be $e_x/\sqrt{2}$ and $f_y/\sqrt{2}$, where e_x are orthonormal and f_y are orthonormal. The product space then embeds as a simplex on the orthogonal vectors $e_x \otimes f_y/2$ and the barycenter of θ is $\sum_{x,y} \theta(x,y) e_x \otimes f_y/2$. Let θ_n, μ_n and ν_n be the empirical measures as in Corollary 2.8. Proposition 3.7 yields

$$\operatorname{dcov}(\theta_n) = \sum_{x,y} [\theta_n(x,y) - \mu_n(x)\nu_n(y)]^2.$$

The test statistic in (2.5) is thus

$$\frac{n \sum_{x,y} [\theta_n(x,y) - \mu_n(x)\nu_n(y)]^2}{\sum_x \mu_n(x)[1-\mu_n(x)] \sum_y \nu_n(y)[1-\nu_n(y)]}.$$

For comparison, Pearson's χ^2-statistic is

$$n \sum_{x,y} \frac{[\theta_n(x,y) - \mu_n(x)\nu_n(y)]^2}{\mu_n(x)\nu_n(y)}.$$

REMARK 3.13. As Gábor Székely has remarked (personal communication, 2010), there is a two-dimensional random variable (X, Y) such that X and Y are not independent, yet if (X', Y') is an independent copy of (X, Y), then $|X - X'|$ and $|Y - Y'|$ are uncorrelated. Indeed, consider the density function $p(x, y) := (1/4 - q(x)q(y))\mathbf{1}_{[-1,1]^2}(x, y)$ with $q(x) := -(c/2)\mathbf{1}_{[-1,0]} + (1/2)\mathbf{1}_{(0,c)}$, where $c := \sqrt{2} - 1$. Then it is not hard to check that this gives such an example.

REMARK 3.14. According to Proposition 3.7, $\operatorname{dcov}(\theta) = -2D(\theta - \mu \times \nu)$ for the metric space $(\mathscr{X} \times \mathscr{Y}, d_{\phi\otimes\psi})$. Since this metric space has strong negative type when $\mathscr{X}$ and $\mathscr{Y}$ do, we can view the fact that $\operatorname{dcov}(\theta) = 0$ only for product measures as a special case of the fact that $D(\theta_1 - \theta_2) = 0$ only when $\theta_1 = \theta_2$ for $\theta_i \in M_1^1(\mathscr{X} \times \mathscr{Y})$. Similarly, any other metric on $\mathscr{X} \times \mathscr{Y}$ of strong negative type could be used to give a test of independence via $D(\theta - \mu \times \nu)$; indeed, when $\mathscr{X} = \mathbb{R}^p$ and $\mathscr{Y} = \mathbb{R}^q$, the Euclidean metric on $\mathbb{R}^{p+q}$ was used by Bakirov, Rizzo and Székely (2006) for precisely such a test.

No such result as Theorem 3.11 holds if either $\mathscr{X}$ or $\mathscr{Y}$ is not of strong negative type:

PROPOSITION 3.15. *If $\mathscr{X}$ is not of negative type, then for every metric space $\mathscr{Y}$ with at least two points, there exists $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ whose marginals have finite first moments and such that* $\operatorname{dcov}(\theta) < 0$. *If $\mathscr{X}$ is not of strong negative type, then for every metric space $\mathscr{Y}$ with at least two points, there exists $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ whose marginals have finite first moments and such that* $\operatorname{dcov}(\theta) = 0$, *yet θ is not a product measure.*

PROOF. Choose two distinct points $y_1, y_2 \in \mathscr{Y}$. Let $\mu_1 \neq \mu_2 \in M_1(\mathscr{X})$ have finite first moments and satisfy $D(\mu_1 - \mu_2) \geq 0$, where > 0 applies if $\mathscr{X}$ does not have negative type. In this latter case, set $\theta := (\mu_1 \times \delta(y_1) + \mu_2 \times \delta(y_2))/2$. Then a little algebra reveals that

$$\operatorname{dcov}(\theta) = -d(y_1, y_2)D(\mu_1 - \mu_2)/8 < 0.$$

In general, note that if $x_1 \neq x_2$, then $D(\delta(x_1) - \delta(x_2)) < 0$, whence there is some $\gamma \in (0, 1]$ such that if $\tau_i := \gamma\mu_i + (1 - \gamma)\delta(x_i)$, then $D(\tau_1 - \tau_2) = 0$. Set $\theta := (\tau_1 \times \delta(y_1) + \tau_2 \times \delta(y_2))/2$. Then

$$\operatorname{dcov}(\theta) = -d(y_1, y_2)D(\tau_1 - \tau_2)/8 = 0,$$

yet θ is not a product measure. □

There remains the possibility that the kernel h in the proof of Proposition 2.6 is degenerate of order 1 only when θ is a product measure. If that is true, then Corollary 2.8 gives a consistent test for independence even in metric spaces not of negative type, since when h is not degenerate and $\operatorname{dcov}(\theta) = 0$, $\sqrt{n}\operatorname{dcov}(\theta_n)$ has a nontrivial normal limit in distribution, whence $n\operatorname{dcov}(\theta_n) \to \pm\infty$ a.s. We have not investigated this possibility.

Since every Euclidean space is of strict negative type, so is every Hilbert space. Separable Hilbert spaces are even of strong negative type, though this is considerably more subtle. Therefore, $\operatorname{dcov}(\theta) = 0$ implies that $\theta \in M_1(\mathscr{X} \times \mathscr{Y})$ is a product measure when $\mathscr{X}$ and $\mathscr{Y}$ are separable Hilbert spaces, which resolves a question of Kosorok (2009).

THEOREM 3.16. *Every separable Hilbert space is of strong negative type.*

PROOF. This follows from Remark 3.4 and Theorem 6 of Linde (1986a) or Theorem 1 of Koldobskiĭ (1982), who prove more. Likewise, separable L^p spaces with $1 < p < 2$ are of strong negative type. However, we give a direct proof that is shorter, which keeps our paper self-contained.

Our proof relies on a known Gaussian variant of the Crofton embedding. Let Z_n ($n \geq 1$) be IID standard normal random variables with law ρ on $\mathbb{R}^\infty$. Given $u = \langle u_n; n \in \mathbb{Z}^+\rangle \in \ell^2(\mathbb{Z}^+)$, define the random variable $Z(u) := \sum_{n\geq 1} u_n Z_n$. Then $Z(u)$ is a centered normal random variable with standard deviation equal to $\|u\|_2$. Therefore, $\mathbf{E}[|Z(u)|] = c\|u\|_2$ with $c := \mathbf{E}[|Z_1|]$.

Let λ be the Lebesgue measure on $\mathbb{R}$. For $w, u \in \mathbb{R}^\infty$, write $w(u) := \limsup_N \sum_{n=1}^N u_n w_n$. We choose $\ell^2(\mathbb{Z}^+)$ as our separable Hilbert space, which we embed into another Hilbert space, $L^2(\mathbb{R}^\infty \times \mathbb{R}, \rho \times \lambda)$, by

$$\phi(u) : (w, s) \mapsto \mathbf{1}_{[w(u)/c,\infty)}(s) - \mathbf{1}_{[0,\infty)}(s).$$

Then $\|\phi(u) - \phi(u')\|_2^2 = \|\phi(u) - \phi(u')\|_1 = \|u - u'\|_2$ for all $u, u' \in \ell^2(\mathbb{Z}^+)$. Let $\mu_1, \mu_2 \in M_1(\ell^2(\mathbb{Z}^+))$ have finite first moments. Set $\mu := \mu_1 - \mu_2$. Because $\int d\mu = 0$, we have

$$\beta_\phi(\mu) : (w, s) \mapsto \mu\{u; w(u) \le cs\}.$$

Note that since for every $u \in \ell^2(\mathbb{Z}^+)$, the series $w(u)$ converges ρ-a.s., Fubini's theorem tells us that for ρ-a.e. w, $w(u)$ converges for μ_i-a.e. u. We need to show that if $\beta_\phi(\mu) = 0$ $\rho \times \lambda$-a.s., then $\mu = 0$. So assume that $\beta_\phi(\mu) = 0$ $\rho \times \lambda$-a.s. It suffices to show that $\mu\{u; \langle u, v\rangle \le s\} = 0$ for every finitely supported $v \in \mathbb{R}^\infty$ and every $s \in \mathbb{R}$, since that implies that the finite-dimensional marginals of μ are 0 by the Cramér–Wold device.

Let $K \ge 1$. For $w \in \mathbb{R}^\infty$, write $w_{\le K}$ for the vector $(w_1, \dots, w_K) \in \mathbb{R}^K$ and $w_{>K}$ for $(w_{K+1}, w_{K+2}, \dots) \in \mathbb{R}^\infty$. Since the law ρ of $w = (w_{\le K}, w_{>K})$ is a product measure, with λ^K absolutely continuous with respect to the first factor and with the second factor equal to ρ, Fubini's theorem gives that for ρ-a.e. w, for λ^K-a.e. $v \in \mathbb{R}^K$, and for λ-a.e. $s \in \mathbb{R}$, we have $\beta(\mu)((v, w), s) = 0$. Since $(v, s) \mapsto \beta(\mu)((v, w), s)$ possesses sufficient continuity properties, we have that for ρ-a.e. w, for all $v \in \mathbb{R}^K$ and all $s \in \mathbb{R}$, $\beta(\mu)((v, w), s) = 0$.

Let $\varepsilon > 0$. Choose K so large that $c \int \|u_{>K}\|_2 \, d\mu_i(u) < \varepsilon^2$ for $i = 1, 2$, which is possible by Lebesgue's dominated convergence theorem and the fact that μ_i has finite first moment. Let

$$A(\varepsilon) := \{(u, w) \in \ell^2(\mathbb{Z}^+) \times \mathbb{R}^\infty; |w(u_{>K})| \ge \varepsilon\}.$$

Markov's inequality yields that

$$(\mu_i \times \rho)A(\varepsilon) \le \varepsilon^{-1} \|w(u_{>K})\|_{L^1(\mu_i \times \rho)} = \varepsilon^{-1} c \int \|u_{>K}\|_2 \, d\mu_i(u) < \varepsilon,$$

where the equality arises from Fubini's theorem. Therefore, there is some w such that denoting $A(w, \varepsilon) := \{u; |w(u_{>K})| \ge \varepsilon\}$, we have $\beta(\mu)((v, w), s) = 0$ for all $v \in \mathbb{R}^K$, $s \in \mathbb{R}$ and

$$\mu_i A(w, \varepsilon) < \varepsilon.$$

For such a w, we have for all v, s that

$$\begin{aligned} \mu_i\{u; \langle u_{\le K}, v\rangle \le s - \varepsilon\} - \varepsilon &< \mu_i\{u; \langle u_{\le K}, v\rangle + w(u_{>K}) \le s\} \\ &< \mu_i\{u; \langle u_{\le K}, v\rangle \le s + \varepsilon\} + \varepsilon. \end{aligned}$$

The middle quantity is the same for $i = 1$ as for $i = 2$ by choice of w. Therefore, for all $v \in \mathbb{R}^K$ and $s \in \mathbb{R}$,

$$\mu_1[\langle u_{\le K}, v\rangle \le s - \varepsilon] - \varepsilon < \mu_2[\langle u_{\le K}, v\rangle \le s + \varepsilon] + \varepsilon$$

and

$$\mu_2\big[\langle u_{\le K}, v\rangle \le s-\varepsilon\big]-\varepsilon < \mu_1\big[\langle u_{\le K}, v\rangle \le s+\varepsilon\big]+\varepsilon.$$

Although K depends on ε, it follows that for all $L \le K$ and all $v \in \mathbb{R}^L$, $s \in \mathbb{R}$,

$$\mu_1\big[\langle u_{\le L}, v\rangle \le s-\varepsilon\big]-\varepsilon < \mu_2\big[\langle u_{\le L}, v\rangle \le s+\varepsilon\big]+\varepsilon$$

and

$$\mu_2\big[\langle u_{\le L}, v\rangle \le s-\varepsilon\big]-\varepsilon < \mu_1\big[\langle u_{\le L}, v\rangle \le s+\varepsilon\big]+\varepsilon.$$

Thus, if we fix L, the above inequalities hold for all ε, which implies that

$$\mu_1\big[\langle u_{\le L}, v\rangle \le s\big] = \mu_2\big[\langle u_{\le L}, v\rangle \le s\big].$$

This is what we needed to show. □

REMARK 3.17. Nonseparable Hilbert spaces H are of strong negative type iff their dimension is a cardinal of measure zero. [Whether there exist cardinals not of measure zero is a subtle question that involves foundational issues; see Chapter 23 of Just and Weese (1997).] To see this equivalence, note first that if every Borel probability measure on H is carried by a separable subset, then H has strong negative type by the preceding theorem. Now a theorem of Marczewski and Sikorski (1948) [or see Theorem 2 of Appendix III in Billingsley (1968)] implies that this separable-carrier condition holds if (and only if) the dimension of H is a cardinal of measure zero. Conversely, if the dimension of H is not a cardinal of measure zero, then let I be an orthonormal basis of H. By definition, there exists a probability measure μ on the subsets of I that vanishes on singletons. Write $I = I_1 \cup I_2$, where I_1 and I_2 are disjoint and equinumerous with I. Define μ_j $(j = 1, 2)$ on I_j by pushing forward μ via a bijection from I to I_j. Extend μ_j to H in the obvious way (all subsets of I are Borel in H since they are G_δ-sets). Then $\mu_1 \neq \mu_2$, yet $D(\mu_1 - \mu_2) = 0$.

COROLLARY 3.18. *If $(\mathscr{X}, d)$ is a separable metric space of negative type, then $(\mathscr{X}, d^{1/2})$ is a metric space of strong negative type.*

PROOF. Let $\phi : (\mathscr{X}, d^{1/2}) \to H$ be an isometric embedding to a separable Hilbert space. Let $\psi : (H, \|\cdot\|^{1/2}) \to H'$ be an isometric embedding to another separable Hilbert space such that β_ψ is injective on $M_1^1(H)$, which exists by Theorem 3.16. Then $\psi \circ \phi : (\mathscr{X}, d^{1/4}) \to H'$ is an isometric embedding to a Hilbert space whose barycenter map is injective on $M_1^1(\mathscr{X}, d^{1/2})$. □

This means that we can apply a distance covariance test of independence to any pair of metric spaces of negative type provided we use square roots of distances in place of distances. This even has the small advantage that the probability measures in question need have only finite half-moments.

REMARK 3.19. We claim that if $(\mathscr{X}, d)$ has negative type, then $(\mathscr{X}, d^r)$ has strong negative type when $0 < r < 1$. When $\mathscr{X}$ is finite, and so strong negative type is the same as strict negative type, this result is due to Li and Weston (2010), Theorem 5.4. To prove our claim, we use the result of Linde (1986a) that the map $\alpha : \mu \mapsto a_\mu$ of Remark 3.4 is injective on $M_1^1(H, \|\cdot\|^r)$ for all $r \in \mathbb{R}^+ \setminus 2\mathbb{N}$. Let $\phi : (\mathscr{X}, d^{1/2}) \to H$ be an isometric embedding. By Linde's result, the map

$$\mu \mapsto \left(x \mapsto \int d(x, x')^r \, d\mu(x') = \int \|\phi(x) - \phi(x')\|^{2r} \, d\mu(x')\right)$$

is injective. Since $(\mathscr{X}, d^r)$ has negative type by a theorem of Schoenberg (1938), the claim follows from Remark 3.4.

COROLLARY 3.20. *If $(\mathscr{X}, d_{\mathscr{X}})$ and $(\mathscr{Y}, d_{\mathscr{Y}})$ are metric spaces of negative type, then $(\mathscr{X} \times \mathscr{Y}, (d_{\mathscr{X}} + d_{\mathscr{Y}})^{1/2})$ is a metric space of strong negative type.*

PROOF. It is easy to see that $(\mathscr{X} \times \mathscr{Y}, d_{\mathscr{X}} + d_{\mathscr{Y}})$ is of negative type, whence the result follows from Corollary 3.18. □

Thus, another way to test independence for metric spaces $(\mathscr{X}, d_{\mathscr{X}})$ and $(\mathscr{Y}, d_{\mathscr{Y}})$ of negative type (not necessarily strong) uses not $\operatorname{dcov}(\theta)$, but $D(\theta - \mu \times \nu)$ with respect to the metric $(d_{\mathscr{X}} + d_{\mathscr{Y}})^{1/2}$ on $\mathscr{X} \times \mathscr{Y}$; compare Remark 3.14. By Remark 3.19, the same holds for $(\mathscr{X} \times \mathscr{Y}, (d_{\mathscr{X}} + d_{\mathscr{Y}})^r)$ with any $r \in (0, 1)$.

We remark finally that for separable metric spaces of negative type, the proofs of Proposition 2.6, Theorem 2.7 and Corollary 2.8 are more straightforward, as they can rely on the strong law of large numbers and the central limit theorem in Hilbert space.

Acknowledgments. I am grateful to Gábor Székely for asking in what metric spaces a theory of distance covariance applies and for other encouraging correspondence. I thank Ori Gurel-Gurevich for a proof of a version of Lemma 3.8, as well as Omer Angel and Assaf Naor for discussions. I thank Svante Janson for corrections on an earlier draft and a simplification of the example in Remark 3.3.

REFERENCES

BAKIROV, N. K., RIZZO, M. L. and SZÉKELY, G. J. (2006). A multivariate nonparametric test of independence. *J. Multivariate Anal.* **97** 1742–1756. MR2298886

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. MR0233396

BRETAGNOLLE, J., DACUNHA-CASTELLE, D. and KRIVINE, J.-L. (1965/1966). Lois stables et espaces L^p. *Ann. Inst. H. Poincaré Sect. B* (*N.S.*) **2** 231–259. MR0203757

CROFTON, M. W. (1868). On the theory of local probability, applied to straight lines drawn at random in a plane; the methods used being also extended to the proof of certain new theorems in the integral calculus. *Philos. Trans. Royal Soc. London* **158** 181–199.

DEZA, M. M. and LAURENT, M. (1997). *Geometry of Cuts and Metrics. Algorithms and Combinatorics* **15**. Springer, Berlin. MR1460488

DOR, L. E. (1976). Potentials and isometric embeddings in L_1. *Israel J. Math.* **24** 260–268. MR0417756

GORIN, E. A. and KOLDOBSKIĬ, A. L. (1987). On potentials of measures in Banach spaces. *Sibirsk. Mat. Zh.* **28** 65–80, 225. MR0886854

JUST, W. and WEESE, M. (1997). *Discovering Modern Set Theory. II. Set-Theoretic Tools for Every Mathematician. Graduate Studies in Mathematics* **18**. Amer. Math. Soc., Providence, RI. MR1474727

KLEBANOV, L. B. (2005). *$\mathfrak{N}$-Distances and Their Applications*. The Karolinum Press, Charles University in Prague.

KOLDOBSKIĬ, A. L. (1982). On isometric operators in vector-valued L^p-spaces. Investigations on linear operators and the theory of functions, X. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **107** 198–203, 233. (Russian with English summary.) MR0676160

KOLDOBSKY, A. and LONKE, Y. (1999). A short proof of Schoenberg's conjecture on positive definite functions. *Bull. Lond. Math. Soc.* **31** 693–699. MR1711028

KOSOROK, M. R. (2009). Discussion of: Brownian distance covariance [MR2752127]. *Ann. Appl. Stat.* **3** 1270–1278. MR2752129

LI, H. and WESTON, A. (2010). Strict p-negative type of a metric space. *Positivity* **14** 529–545. MR2680513

LINDE, W. (1986a). On Rudin's equimeasurability theorem for infinite-dimensional Hilbert spaces. *Indiana Univ. Math. J.* **35** 235–243. MR0833392

LINDE, W. (1986b). Uniqueness theorems for measures in L_r and $C_0(\Omega)$. *Math. Ann.* **274** 617–626.

MARCZEWSKI, E. and SIKORSKI, R. (1948). Measures in non-separable metric spaces. *Colloq. Math.* **1** 133–139. MR0025548

MECKES, M. W. (2013). Positive definite metric spaces. *Positivity*. **17** 733–757. MR3090690

NAOR, A. (2010). L_1 embeddings of the Heisenberg group and fast estimation of graph isoperimetry. In *Proceedings of the International Congress of Mathematicians. Volume III* 1549–1575. Hindustan Book Agency, New Delhi. MR2827855

NICKOLAS, P. and WOLF, R. (2009). Distance geometry in quasihypermetric spaces. I. *Bull. Aust. Math. Soc.* **80** 1–25. MR2520521

SCHOENBERG, I. J. (1937). On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space. *Ann. of Math.* (2) **38** 787–793. MR1503370

SCHOENBERG, I. J. (1938). Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.* **44** 522–536. MR1501980

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

SZÉKELY, G. J. and RIZZO, M. L. (2005a). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *J. Classification* **22** 151–183. MR2231170

SZÉKELY, G. J. and RIZZO, M. L. (2005b). 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

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

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

DEPARTMENT OF MATHEMATICS
INDIANA UNIVERSITY
831 EAST 3RD ST.
BLOOMINGTON, INDIANA 47405-7106
USA
E-MAIL: rdlyons@indiana.edu