Journal of Statistical Planning and Inference 143 (2013) 1249–1272

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Review

Energy statistics: A class of statistics based on distances

Gábor J. Székely [a,b,1], Maria L. Rizzo [c,*]

[a] National Science Foundation, 4201 Wilson Blvd. #1025, Arlington, VA 22230, United States
[b] Rényi Institute of Mathematics, Hungarian Academy of Sciences, Hungary
[c] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, United States

ARTICLE INFO

Article history:
Received 31 May 2012
Received in revised form
1 March 2013
Accepted 4 March 2013
Available online 20 March 2013

Keywords:
Energy distance
Goodness-of-fit
Multivariate independence
Distance covariance
Distance correlation

ABSTRACT

Energy distance is a statistical distance between the distributions of random vectors, which characterizes equality of distributions. The name energy derives from Newton's gravitational potential energy, and there is an elegant relation to the notion of potential energy between statistical observations. Energy statistics are functions of distances between statistical observations in metric spaces. Thus even if the observations are complex objects, like functions, one can use their real valued nonnegative distances for inference. Theory and application of energy statistics are discussed and illustrated. Finally, we explore the notion of potential and kinetic energy of goodness-of-fit.

Contents

* Corresponding author. Tel.: +1 419 372 7474; fax: +1 419 372 6092.
E-mail addresses: gszekely@nsf.gov (G.J. Székely), mrizzo7@gmail.com, mrizzo@bgsu.edu (M.L. Rizzo).
[1] Part of this research was based on work supported by the National Science Foundation, while working at the Foundation.

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.03.018

1. Introduction

Energy statistics ($\mathcal{E}$-statistics) are functions of distances between statistical observations. This concept is based on the notion of Newton's gravitational potential energy which is a function of the distance between two bodies. The idea of energy statistics is to consider statistical observations as heavenly bodies governed by a statistical potential energy, which is zero if and only if an underlying statistical null hypothesis is true.

In this paper we present the foundational material, motivation, and unifying theory of energy statistics. Previously unpublished results as well as an overview of several published applications in inference and multivariate analysis that illustrate the power of this concept are discussed. We will see that energy statistics are extremely useful and are typically more general and often more powerful against general alternatives than classical (non-energy type) statistics such as correlation, *F*-statistics, etc. For historical background, see Section 9.

If the observations play a symmetric role, then it makes sense to suppose that energy statistics are symmetric functions of distances between observations. Energy statistics in this paper are *U*-statistics or *V*-statistics based on distances; that is, for a *d*-dimensional random sample $X_1,\ldots,X_n$, and kernel function $h : R^d \times R^d \to R$,

$$U_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} h(X_i, X_j)$$

or

$$V_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(X_i, X_j),$$

where $h(X_i,X_j) = h(X_j,X_i)$ is a symmetric function of Euclidean distances $|X_i - X_j|$ between sample elements. Here we use $|\cdot|_d$ or $|\cdot|$ (if dimension d is clear in context) to denote Euclidean norm if the argument is real, and $|\cdot|$ denotes the complex norm when its argument is complex. The notation $\|\cdot\|$ is reserved for another type of norm in this paper.

Since energy statistics are *U*-statistics or *V*-statistics, we can apply their classical limit theory (Hoeffding, 1948; Von Mises, 1947) to obtain the limiting behavior of these statistics. See Serfling (1980) or Koroljuk and Borovskich (1994) for details.

A familiar example is the *U*-statistic for dispersion, Gini's mean difference (Yitzhaki, 2003),

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} |X_i - X_j|.$$

Alternately one can apply the *V*-statistic

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|.$$

In some of the applications discussed below, a generalized *V*-statistic is applied, see e.g. Koroljuk and Borovskich (1994, Chapter 4). Although energy statistics can be defined in terms of either *U* or *V*-statistics, we apply *V*-statistics throughout. One of the reasons for applying *V*-statistics instead of *U*-statistics is that our energy statistics will be nonnegative and more easily interpreted as a statistical distance.

This paper summarizes many of the most interesting results and some of the applications of energy statistics. The content is organized in sections as follows.

1. Introduction.
2. Energy distance.
3. Why is energy distance special?
4. One sample energy statistics: goodness-of-fit, multivariate normality.
5. Generalized energy distance.
6. Two-sample and multi-sample energy statistics, distance components for structured data, $\mathcal{E}$-clustering, symmetry.
7. Distance correlation, energy test of independence, Brownian covariance.
8. Statistical potential and kinetic energy.
9. Historical background.

2. Energy distance

There are many types of distances that can be defined between statistical objects. One of the best known and most applied is the L_2-distance. If F is the cumulative distribution function (cdf) of a random variable and F_n is the empirical cdf, then their L_2 distance,

$$\int_{-\infty}^{\infty}(F_n(x)-F(x))^2\,dx \tag{2.1}$$

was introduced in Cramér (1928). This distance has the disadvantage that it is not distribution-free; thus if we want to apply this distance for testing goodness-of-fit, then the critical values depend on F. This problem was easily rectified via replacing dx by $dF(x)$ which leads to the Cramér–von Mises Smirnov distance:

$$\int_{-\infty}^{\infty}(F_n(x)-F(x))^2\,dF(x). \tag{2.2}$$

There is, however, another important disadvantage of Cramér's distance (2.1) and also of the Cramér–von Mises–Smirnov distance (2.2) that remains. If the sample comes from a d-dimensional space where $d>1$, then neither of them are rotation invariant. This is a very important problem if e.g. we want to test multivariate normality. Here is how to overcome this difficulty.

Suppose that X,Y are real-valued independent random variables with cumulative distribution function F and G, respectively. It is easy to show (see e.g. Székely, 1989, 2003) that if X' is an independent and identically distributed (iid) copy of X, and Y' is an iid copy of Y, then

$$2\int_{-\infty}^{\infty}(F(x)-G(x))^2\,dx = 2E|X-Y|-E|X-X'|-E|Y-Y'|.$$

A rotation invariant natural extension for higher dimension is

$$2E|X-Y|_d-E|X-X'|_d-E|Y-Y'|_d, \tag{2.3}$$

where $X,Y\in R^d$ are independent. Rotational invariance of this expression is trivial, but it is *not trivial at all* that this quantity (2.3) is nonnegative and equals zero if and only if X and Y are identically distributed. See Proposition 1.

Definition 1 (*Energy distance*). The energy distance between the d-dimensional independent random variables X and Y is defined as

$$\mathcal{E}(X,Y) = 2E|X-Y|_d-E|X-X'|_d-E|Y-Y'|_d, \tag{2.4}$$

where $E|X|_d<\infty$, $E|Y|_d<\infty$, X' is an iid copy of X, and Y' is an iid copy of Y. We omit the subscript d whenever it is clear in context.

Denote the Fourier-transform (characteristic function) of the probability density functions f and g by $\hat{f}$ and $\hat{g}$, respectively. Then, according to the Parseval–Plancherel formula,

$$2\pi\int_{-\infty}^{\infty}(f(x)-g(x))^2\,dx = \int_{-\infty}^{\infty}|\hat{f}(t)-\hat{g}(t)|^2\,dt.$$

Since the Fourier transform of the cdf $F(x)=\int_{-\infty}^{x}f(u)\,du$ is $\hat{f}(t)/(it)$, where $i=\sqrt{-1}$, we have

$$2\pi\int_{-\infty}^{\infty}(F(x)-G(x))^2\,dx = \int_{-\infty}^{\infty}\frac{|\hat{f}(t)-\hat{g}(t)|^2}{t^2}\,dt. \tag{2.5}$$

The pleasant surprise is that the natural multivariate generalization of the right-hand side of (2.5) is rotation invariant and it is exactly a constant multiple of (2.4).

Proposition 1. *If the d-dimensional random variables X and Y are independent with $E|X|_d+E|Y|_d<\infty$, and $\hat{f},\hat{g}$ denote their respective characteristic functions, then their energy distance*

$$2E|X-Y|_d-E|X-X'|_d-E|Y-Y'|_d = \frac{1}{c_d}\int_{R^d}\frac{|\hat{f}(t)-\hat{g}(t)|^2}{|t|_d^{d+1}}\,dt, \tag{2.6}$$

where

$$c_d=\frac{\pi^{(d+1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}, \tag{2.7}$$

and $\Gamma(\cdot)$ is the complete gamma function. Thus $\mathcal{E}(X,Y)\geq 0$ with equality to zero if and only if X and Y are identically distributed.

For a proof of this proposition see Székely and Rizzo (2005a) or see the proof of its generalization, Proposition 2 below. For the prehistory of this inequality see Section 9.

In view of (2.6), the square root of energy distance $\mathcal{E}(X,Y)^{1/2}$ is a metric on the set of d-variate distribution functions.

It is easy to define $\mathcal{E}$ for all pairs of random variables X,Y that take their values in a metric space with distance function δ:

$$\mathcal{E}(X,Y) = 2E[\delta(X,Y)] - E[\delta(X,X')] - E[\delta(Y,Y')],$$

provided that these expectations exist; but if we replace Euclidean distance with the metric δ in an arbitrary metric space, then the claim of Proposition 1 that "$\mathcal{E}(X,Y) \geq 0$ *with equality to zero if and only if X and Y are identically distributed*" does not necessarily hold. It does hold in separable Hilbert spaces (see Lyons, to appear), which is an important result for applications.

Energy distance $\mathcal{E}(F,G)$ provides a characterization of equality of distributions F and G. The applications in dimensions $d \geq 1$ include:

(i) Consistent one-sample goodness-of-fit tests (Székely and Rizzo, 2005a; Rizzo, 2009; Yang, 2012).
(ii) Consistent multi-sample tests of equality of distributions (Székely and Rizzo, 2004; Rizzo, 2002, 2003; Baringhaus and Franz, 2004).
(iii) Hierarchical clustering algorithms (Székely and Rizzo, 2005b) that extend and generalize the Ward's minimum variance algorithm.
(iv) Distance components (*DISCO*) (Rizzo and Székely, 2010), a nonparametric extension of analysis of variance for structured data.
(v) Characterization and test for multivariate independence (Feuerverger, 1993; Székely et al., 2007; Székely and Rizzo, 2009).
(vi) Change point analysis based on Székely and Rizzo (2004) is applied in Kim et al. (2009) and Matteson and James (2012).

Several of these applications are discussed below. Software for energy statistics applications is available under General Public License in the *energy* (Rizzo and Székely, 2011) package for R (R Development Core Team, 2012).

3. Why is energy distance special?

We see that energy distance (2.6) is a weighted L_2 distance between characteristic functions, with weight function $w(t) = |t|_d^{-(d+1)}$. Suppose that the following three technical conditions on the weight function hold: $w(t) > 0$, $w(t)$ is continuous, and

$$\int |\hat{f}(t) - \hat{g}(t)|^2 w(t)\, dt < \infty. \tag{3.1}$$

We claim that under these conditions if the weighted L_2 distance between $\hat{f}$ and $\hat{g}$ is rotation invariant and scale equivariant, then $w(t) = \text{const}/|t|^{d+1}$. In other words, rotation invariance and scale equivariance (under some technical conditions) imply that the weighted L_2 distance between characteristic functions is the energy distance.

Why do we have this characterization? One can show that if two weighted L_2 distances of the type (3.1) are equal for all characteristic functions $\hat{f}$ and $\hat{g}$, then the (continuous) weight functions are also equal (for proof of a similar claim see Székely and Rizzo, 2012).

Scale equivariance and rotation invariance imply that for all real numbers a

$$\int |\hat{f}(at) - \hat{g}(at)|^2 w(t)\, dt = |a| \times \int |\hat{f}(t) - \hat{g}(t)|^2 w(t)\, dt.$$

Introduce $s = at$. We can see that if $a \neq 0$ then

$$\int |\hat{f}(s) - \hat{g}(s)|^2 \frac{w(s/a)}{|a|}\, ds = |a| \times \int |\hat{f}(t) - \hat{g}(t)|^2 w(t)\, dt.$$

Thus $w(s/a)/|a| = |a| w(s)$. That is, if $c := w(1)$ then $w(1/a) = ca^2$, implying that $w(t) = \text{const}/|t|^{d+1}$.

Interestingly, this weight function appears in Feuerverger (1993), where it is applied for testing bivariate dependence. Although this singular weight function is "special" from the equivariance point of view, other weight functions are also applied in tests based on characteristic functions; see e.g. Gurtler and Henze (2000), Henze and Zirkler (1990), or Matsui and Takemura (2005).

4. One sample energy statistics

4.1. Energy goodness-of-fit statistics

Let $X_1,\ldots,X_n$ be a random sample (iid) from a d-variate population with distribution F, and let $x_1,\ldots,x_n$ be the observed values of the random sample. The one sample version of energy distance for testing the goodness-of-fit hypothesis $H_0 : F = F_0$ vs $H_1 : F \neq F_0$ is

$$\mathcal{E}_n(X,F_0) = \frac{2}{n}\sum_{i=1}^{n} E|x_i - X| - E|X - X'| - \frac{1}{n^2}\sum_{\ell=1}^{n}\sum_{m=1}^{n} |x_\ell - x_m|, \tag{4.1}$$

where X and X' are independent and identically distributed (iid) with distribution F_0, and the expectations are taken with respect to the null distribution F_0. The energy goodness-of-fit statistic is $n\mathcal{E}_n = n\mathcal{E}_n(X,F_0)$.

Notice that $\mathcal{E}_n$ is a V-statistic, and its unbiased versions are U-statistics. (The kernel function for the energy goodness-of-fit statistic is (8.1), which is discussed in Section 8.) Under the null hypothesis, the test statistic $n\mathcal{E}_n$ tends to a nondegenerate limit distribution as $n\to\infty$ (see Section 8), while under an alternative hypothesis $n\mathcal{E}_n$ tends to infinity. Thus a goodness-of-fit test that rejects the null for large values of $n\mathcal{E}_n$ is consistent against general alternatives.

Fig. 1 illustrates the sampling distribution of an energy goodness-of-fit statistic. The sampling distributions of all energy statistics have similar shapes under the null hypothesis, with rejection region in the upper tail.

For goodness-of-fit statistics, if parameters of the null distribution are estimated, in many applications the energy statistic is a degenerate kernel V-statistic, thus the distribution has a similar shape with rejection region in the upper tail. Here the energy goodness-of-fit test is developed for a specified parametric family of distributions $F_0(\theta)$ using estimated parameter(s) $\hat{\theta}$, so the asymptotic distribution would depend on θ and the distribution of $\hat{\theta}$. In this case, Monte Carlo methods can typically be applied to obtain a test decision. For a discussion of goodness-of-fit tests with estimated parameters see DasGupta (2008, pp. 451–455) Section 28.1, Section 28.2 on the special case of EDF tests, and the references at the end of Chapter 28. For further details on energy goodness-of-fit tests, see Section 8 of this paper and Rizzo (2002, Theorem 5) for proof of consistency in the case of estimated parameters for the multivariate normal distribution.

Energy tests based on (4.1) have been implemented for testing the composite hypothesis of multivariate normality (Székely and Rizzo, 2005a), Pareto family (Rizzo, 2009), stable (Yang, 2012), and other distributions. Let us first introduce energy goodness-of-fit tests with a few univariate examples.

4.1.1. *Two-parameter exponential distribution*

Suppose for example, that we wish to test whether T has a two-parameter exponential distribution, with density

$$f_T(t) = \lambda e^{-\lambda(t-\mu)}, \quad t \geq \mu.$$

Then we apply (4.1) using

$$E|t-T| = t-\mu+\frac{1}{\lambda}(1-2F_T(t)), \quad t\geq\mu;$$

$$E|T-T'| = \frac{1}{\lambda}.$$

A computing formula for the corresponding test statistic $n\mathcal{E}_n$ is easily derived.

4.1.2. *Energy statistic for uniform distribution*

The energy test for the continuous uniform distribution is particularly simple. If $X\sim$ Uniform (a,b), then

$$E|x-X| = \frac{(x-a)^2}{b-a} - x + \frac{b-a}{2}, \quad E|X-X'| = \frac{b-a}{3}.$$

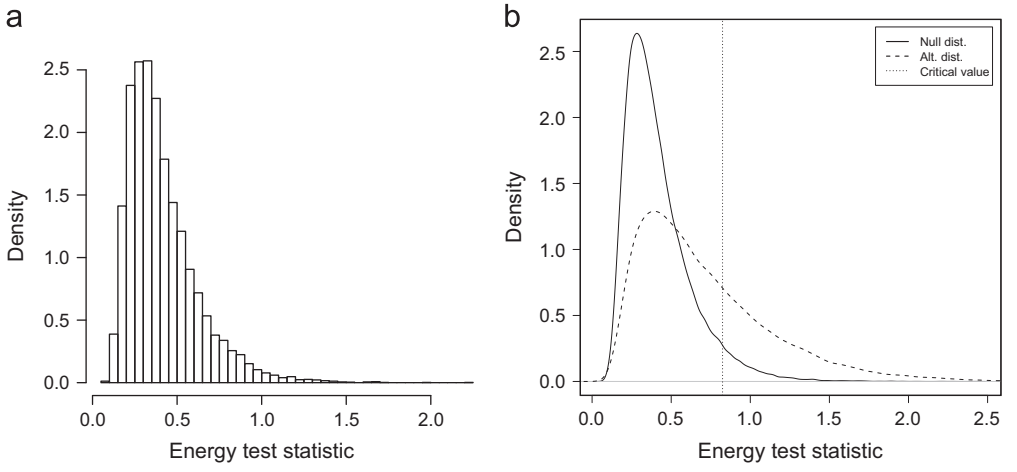

Fig. 1. Sampling distribution of the energy goodness-of-fit statistic for univariate normality, sample size $n=30$, in Example 1. Subfigure (a) displays simulated replicates of the statistic under the null hypothesis. Subfigure (b) compares the density of $n\hat{\mathcal{E}}_n$ under the null and alternative hypotheses, for a normal location mixture. The rejection region is in the upper tail.

In particular, the energy test statistic for a goodness-of-fit test of $H_0 : X \sim$ Uniform(0,1) is given by

$$n\mathcal{E}_n = n\left(\frac{2}{n}\sum_{i=1}^{n}\left(X_i^2 - X_i + \frac{1}{2}\right) - \frac{1}{3} - \frac{2}{n^2}\sum_{k=1}^{n}(2k-1-n)X_{(k)}\right),$$

where $X_{(k)}$ denotes the k-th order statistic of the sample. The linearization in the last sum simplifies the statistic for any univariate test, reducing the computational complexity to $O(n \log n)$ in the univariate case. The statistic can be simplified further.

4.2. Energy test of normality

The energy statistic for testing whether a sample $X_1,\ldots,X_n$ is from a multivariate normal distribution $N(\mu,\Sigma)$ is developed by Székely and Rizzo (2005a). Let $x_1,\ldots,x_n$ denote an observed random sample.

4.2.1. Univariate normality

In the special case of testing univariate normality, the test statistic is $n\mathcal{E}_n$, where $\mathcal{E}_n$ is given by (4.1) with

$$E|x_i - X| = 2(x_i-\mu)F(x_i) + 2\sigma^2 f(x_i) - (x_i-\mu), \quad E|X-X'| = \frac{2\sigma}{\sqrt{\pi}},$$

where F,f are respectively the cdf and density of the hypothesized $N(\mu,\sigma^2)$ distribution.

Example 1 (*Distribution of energy goodness-of-fit statistics*). The sampling distribution of the energy test statistic for testing normality is illustrated by Fig. 1(a) and (b) for sample size $n=30$. Fig. 1(b) displays a histogram of replicates of the test statistic for normal samples. In Fig. 1(b) the densities of the sampling distribution under the null and alternative are compared. The test is implemented in the *energy* package (Rizzo and Székely, 2011) using estimated parameters for μ and σ. The alternative in this example is a 90–10% normal location mixture of N(0,1) and N(3,1) data. The approximate critical value 0.82 for a test at 5% significance is marked with a vertical line in Fig. 1(b). As n increases, the distribution under the alternative shifts farther to the right.

Note that the shape of the sampling distribution under the null hypothesis illustrated in Example 1 is typical of all energy statistics discussed in this paper.

4.2.2. Relation to quadratic EDF statistics

The quadratic empirical distribution function (EDF) statistics are based on weighted L_2 distances of the type (2.2):

$$\int_{-\infty}^{\infty}(F_n(x)-F(x))^2 w(x)\, dF(x), \tag{4.2}$$

where $w(\cdot)$ is a suitable weight function. When $w(x)$ is the identity function, the test is called the Cramér–von Mises test. The Anderson–Darling test is obtained using a weight function $w(x) = [F(x)(1-F(x))]^{-1}$. In case of standard normal null F, the shape of the curve $w(x) = F(x)(1-F(x))$ is similar to the shape of the standard normal density; their ratio is close to a constant c (empirically 0.67). That is, in the univariate case the distribution of the energy statistic for standard normal distribution hardly differs from the powerful Anderson–Darling test of normality. The energy test of normality can thus be viewed as a computationally simple way to lift the Anderson–Darling test to arbitrarily high dimension. The energy test of multivariate normality is rigid motion invariant and consistent against all fixed alternatives with $E|X| < \infty$. When the test is applied to standardized samples, it is affine invariant.

4.3. Energy test of multivariate normality

For the test of multivariate normality, first the sample is standardized by a linear transformation. For standard multivariate normal $Z \in R^d$ with mean vector 0 and identity covariance matrix,

$$E|Z-Z'|_d = \sqrt{2}E|Z|_d = 2\,\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}.$$

If $y_1,\ldots,y_n$ denote the standardized sample elements, the computing formula for the d-variate normality test statistic is given by

$$n\mathcal{E}_{n,d} = n\left(\frac{2}{n}\sum_{j=1}^{n}E|y_j-Z|_d - 2\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} - \frac{1}{n^2}\sum_{j,k=1}^{n}|y_j-y_k|_d\right)$$

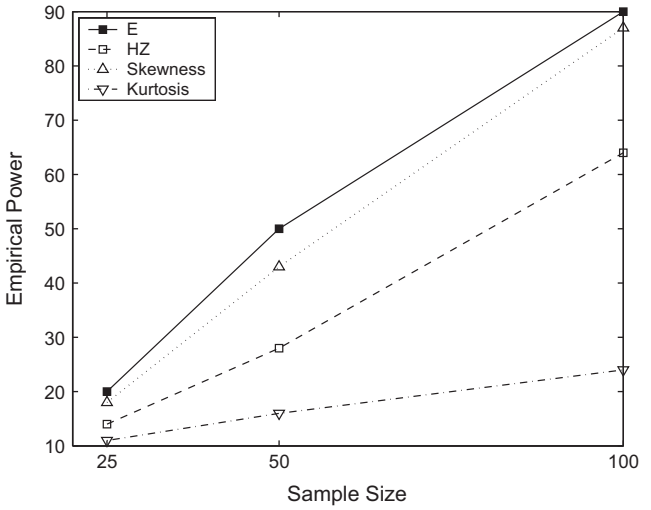

Fig. 2. Empirical power of tests of multivariate normality ($d=5$, $n=25$, 50, 100) against normal location mixture $0.9N_5(0,I)+0.1N_5(2,I)$: percent of significant tests of 2000 Monte Carlo samples at $\alpha=0.05$. E denotes the energy test and HZ denotes the Henze–Zirkler test.

where

$$E|a-Z|_d = \frac{\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} + \sqrt{\frac{2}{\pi}}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!2^k}\frac{|a|_d^{2k+2}}{(2k+1)(2k+2)}\frac{\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(k+\frac{3}{2}\right)}{\Gamma\left(k+\frac{d}{2}+1\right)}.$$

The expression for $E|a-Z|_d$ follows from the fact (see e.g. Zacks, 1981, p. 55) that if Z is a d-variate standard normal random vector, $|a-Z|_d^2$ has a noncentral chisquare distribution $\chi^2[\nu;\lambda]$ with degrees of freedom $\nu=d+2\psi$, and noncentrality parameter $\lambda=|a|_d^2/2$, where ψ is a Poisson random variable with mean λ. Typically the sum in $E|a-Z|_d$ converges after 40–60 terms, but may require more terms if $|a|_d$ is large; however, when $|a|_d$ is large the limit $E|a-Z|_d\cong|a|_d$ can be applied. See the source code in "energy.c" of the *energy* package (Rizzo and Székely, 2011) for an implementation.

If the mean vector μ and covariance matrix Σ are not specified, then the test is modified by transforming the observed sample using the sample mean vector and the sample covariance matrix. This is the method of implementation in the following simulations, using `mvnorm.etest` in the *energy* package for R (Rizzo and Székely, 2011). The modified test statistic $n\hat{\mathcal{E}}_{n,d}$ has the same type of limit distribution as $n\mathcal{E}_{n,d}$, but with different critical values. Theory for estimated parameters is derived in Rizzo (2002) and Székely and Rizzo (2005a). In the *energy* package (Rizzo and Székely, 2011) the sample is standardized using estimated parameters and a decision is obtained by parametric bootstrap. Alternately, for large sample sizes (100 or more observations), one can obtain critical values by numerical solution to the eigenvalue problem (8.5), or see Rizzo (2002) for tabulated critical values of $n\hat{\mathcal{E}}_{n,d}$.

This new test of multivariate normality is practical to apply for arbitrary dimension and sample size ($d>n$ is not a problem.) Monte Carlo power comparisons suggest that it is a powerful competitor to other affine invariant tests of multivariate normality. Overall, the energy test is a powerful omnibus test of multivariate normality, consistent against all alternatives with relatively good power compared with other commonly applied tests.

Example 2 (*Power of energy test of multivariate normality*). Several examples appear in Rizzo (2002) and Székely and Rizzo (2005a) to illustrate the power of the test of multivariate normality against various alternatives, compared with competing tests. As an example, we summarize a comparison for a 90–10% normal location mixture in dimension $d=5$ against Mardia's (1970) skewness and kurtosis tests, and the Henze–Zirkler (1990) test. The latter is another type of test that is based on the characteristic function. Results for 2000 tests at significance level 5% are summarized in Fig. 2. Here we see that the energy test dominates the other tests, and we can also observe that the test is consistent with power increasing to 1 as sample size increases. □

5. Generalized energy distance

Since many important distributions do not have finite expected values we need the following generalization of Proposition 1.

Proposition 2. *Let X and Y be independent d-dimensional random variables with characteristic functions $\hat{f}, \hat{g}$. If $E|X|^\alpha < \infty$ and $E|Y|^\alpha < \infty$ for some $0 < \alpha \le 2$, then*

(i) *For $0 < \alpha < 2$,*

$$\mathcal{E}^{(\alpha)}(X,Y) = 2E|X-Y|^\alpha - E|X-X'|^\alpha - E|Y-Y'|^\alpha = \frac{1}{C(d,\alpha)} \int_{R^d} \frac{|\hat{f}(t)-\hat{g}(t)|^2}{|t|^{d+\alpha}}\, dt, \tag{5.1}$$

where

$$C(d,\alpha) = 2\pi^{d/2} \frac{\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)}. \tag{5.2}$$

(ii) $\mathcal{E}^{(2)}(X,Y) = 2|E(X)-E(Y)|^2$.

Statements (i) and (ii) show that for all $0 < \alpha < 2$, we have $\mathcal{E}^{(\alpha)}(X,Y) \ge 0$ with equality to zero if and only if X and Y are identically distributed; but this characterization does not hold for $\alpha = 2$ since we have equality to zero in (ii) whenever $E(X) = E(Y)$.

Applications of Proposition 2 include:

(i) Goodness-of-fit tests for heavy tailed distributions such as stable distributions (Yang, 2012) and Pareto distributions (Rizzo, 2009).
(ii) Generalization of Ward's minimum variance criterion in hierarchical cluster analysis (see Section 6.4).
(iii) Generalization of distance covariance for heavy tailed distributions (see Section 7.3).
(iv) The *energy score* (see Gneiting and Raftery, 2007).

On the historical background of Proposition 2 see Section 9. For the sake of easy reference we provide a proof.

Proof of Proposition 2. Statement (ii) is obvious. For (i), let $\overline{f(\cdot)}$ denote the complex conjugate of $f(\cdot)$. Notice that

$$\begin{aligned}
|\hat{f}(t)-\hat{g}(t)|^2 &= [\hat{f}(t)-\hat{g}(t)]\overline{[\hat{f}(t)-\hat{g}(t)]} \\
&= [1-\hat{f}(t)\overline{\hat{g}(t)}]+[1-\overline{\hat{f}(t)}\hat{g}(t)]-[1-\hat{f}(t)\overline{\hat{f}(t)}]-[1-\hat{g}(t)\overline{\hat{g}(t)}] \\
&= E\{[2-\exp\{i(t,X-Y)\}-\exp\{i(t,Y-X)\}]-[1-\exp\{i(t,X-X')\}]-[1-\exp\{i(t,Y-Y')\}]\} \\
&= E\{2[1-\cos(t,X-Y)]-[1-\cos(t,X-X')]-[1-\cos(t,Y-Y')]\},
\end{aligned}$$

thus

$$\int_{R^d} \frac{|\hat{f}(t)-\hat{g}(t)|^2}{|t|^{d+\alpha}}\, dt = E\left[\int_{R^d} \frac{2[1-\cos(t,X-Y)]-[1-\cos(t,X-X')]-[1-\cos(t,Y-Y')]}{|t|^{d+\alpha}}\, dt\right].$$

Therefore for (i), all we need to prove is the following lemma.

Lemma 1. *For all $x \in R^d$, if $0 < \alpha < 2$, then*

$$\int_{R^d} \frac{1-\cos(t,x)}{|t|_d^{d+\alpha}}\, dt = C(d,\alpha)|x|_d^\alpha,$$

where (t,x) represents inner product, $C(d,\alpha)$ is the constant (5.2) defined in Proposition 2, $t \in R^d$. (The integrals at $t=0$ and $t=\infty$ are meant in the principal value sense: $\lim_{\varepsilon \to 0} \int_{R^d \setminus \{\varepsilon B + \varepsilon^{-1}\overline{B}\}}$, where B is the unit ball (centered at 0) in R^d and $\overline{B}$ is the complement of B.)

A proof of Lemma 1 is given in Székely and Rizzo (2005b). Because of the importance of this lemma, we reproduce the proof in the Appendix of this paper. □

What class of functions can replace $|x-y|_d^\alpha$ in Proposition 2? That is, for which functions ϕ does the statement

$$2E\phi(X-Y)-E\phi(X-X')-E\phi(Y-Y') \ge 0 \tag{5.3}$$

hold, with equality to zero if and only if X and Y are identically distributed? A necessary and sufficient condition is established in Proposition 3 below. For this result, we need the definition of conditionally negative definite functions.

A function ϕ from R^d to the complex numbers is called *conditionally negative definite* if for all choices of $z_{i,j} = x_i - y_j$, $i,j = 1,\dots,n$, for all complex numbers $c_1,\dots,c_n$, and all natural numbers n we have

$$\sum_{i=1}^n \sum_{j=1}^n c_i \overline{c}_j \phi(z_{i,j}) \le 0 \tag{5.4}$$

whenever $c_1+c_2+\cdots+c_n=0$ (see e.g. Berg, 2008). The function ϕ is *strictly negative definite* if it is negative definite and equality holds in (5.4) only if $c_1=\cdots=c_n=0$.

If ϕ is a symmetric real valued function, then it is enough to consider real numbers $c_1,\ldots,c_n$ in (5.4).

Proposition 3. *Let ϕ be a continuous symmetric function from R^d to R, and let $X\in R^d$, $Y\in R^d$ be independent.*
(i) *A necessary and sufficient condition that*

$$2E\phi(X-Y)-E\phi(X-X')-E\phi(Y-Y')\geq 0 \tag{5.5}$$

holds for all X,Y such that $E[\phi(X-X')+E\phi(Y-Y')]<\infty$ is that ϕ is conditionally negative definite.
(ii) *In (5.5), a necessary and sufficient condition that*

$$2E\phi(X-Y)-E\phi(X-X')-E\phi(Y-Y')=0$$

if and only if X and Y are identically distributed is that ϕ is strictly negative definite.

According to a characterization theorem of Schoenberg (Berg et al., 1984, Theorem 3.2.2), a function is conditionally negative definite, continuous, symmetric, and takes the value 0 at 0 if and only if it is the negative logarithm of an infinitely divisible characteristic function. The functions

$$\phi(z)=|z|^\alpha,\quad 0<\alpha\leq 2$$

correspond to infinitely divisible characteristic functions that are symmetric stable with parameter α. Other examples include $\log(1+|z|^2)$ which corresponds to the characteristic function of the Laplace distribution. Note that in case $\phi(z)=|z|^2$ (case $\alpha=2$) we have conditional negative definiteness, but not strict conditional negative definiteness. This is an intrinsic limitation of classical inference procedures applied for least squares estimates or Pearson's correlation. For application of other negative definite kernels $\phi(z)$ see Baringhaus and Franz (2010).

6. Multi-sample energy statistics

6.1. Testing for equal distributions

The two sample energy statistic corresponding to the energy distance $\mathcal{E}(X,Y)$, for independent random samples $\mathbf{X}=X_1,\ldots,X_{n_1}$ and $\mathbf{Y}=Y_1,\ldots,Y_{n_2}$, is

$$\mathcal{E}_{n_1,n_2}(\mathbf{X},\mathbf{Y})=\frac{2}{n_1n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}|X_i-Y_m|-\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|X_i-X_j|-\frac{1}{n_2^2}\sum_{\ell=1}^{n_2}\sum_{m=1}^{n_2}|Y_\ell-Y_m|. \tag{6.1}$$

The statistic $T_{n_1,n_2}=(n_1n_2/(n_1+n_2))\mathcal{E}_{n_1,n_2}$ can be applied for testing homogeneity (equality of distributions of X and Y). As the null distribution of T_{n_1,n_2} depends on the distributions of X and Y, the test is implemented as a permutation test in the *energy* package. The hypothesis of equal distributions is rejected for large T_{n_1,n_2}. For details, application and power comparisons see Székely and Rizzo (2004), Rizzo (2003), and Baringhaus and Franz (2004).

Several applications and extensions of the two-sample energy statistic follow.

6.2. Testing for symmetry

A test for diagonal symmetry is a special case of the two-sample energy test in Section 6.1. Diagonal symmetry holds if the distributions of X and $-X$ coincide. It was shown in Buja et al. (1994) and also in Székely and Móri (2001) that if X, X' are iid R^d valued random variables then

$$E|X+X'|\geq E|X-X'|,$$

and equality holds if and only if X is diagonally symmetric. We can thus introduce a measure of asymmetry, the distance skewness.

Definition 2. If $X\in R^d$ and $E|X|<\infty$, the *distance skewness* coefficient of a random vector is defined as

$$\mathrm{dSkew}(X)=\begin{cases}1-\dfrac{E|X-X'|}{E|X+X'|}, & E|X+X'|>0;\\ 1, & E|X+X'|=0.\end{cases}$$

Distance skewness is a good measure of symmetry because $0\leq\mathrm{dSkew}(X)\leq 1$, with equality to zero if and only if X is diagonally symmetric.

If $\mathbf{X}=X_1,\ldots,X_n$ is a random sample from the distribution of X, the sample distance skewness can be defined as follows:

$$\mathrm{dSkew}_n(\mathbf{X}):=1-\frac{\sum_{i,j=1}^n|X_i-X_j|}{\sum_{i,j=1}^n|X_i+X_j|}.$$

A consistent test against general alternatives can be based on the statistic

$$T_n(\mathbf{X}) := 1 + \sum_{1\le i<j\le n} \frac{|X_i+X_j|-|X_i-X_j|}{\sum_{1\le i\le n}|X_i|} \tag{6.2}$$

$$T_n(\mathbf{X}) = \frac{\sum_{i,j=1}^n |X_i-Y_j| - \sum_{i,j=1}^n |X_i+Y_j|}{2\sum_{i,j=1}^n |X_i|}, \tag{6.3}$$

where $Y_i = -X'_i, i = 1,\dots,n$, is the reflected $\mathbf{X}$ sample in a randomized order. Here the numerator of the fraction in (6.2) is exactly half the sample energy distance of X and $-X$. The numerator in (6.3) is proportional to (6.1) for samples X_i and $Y_i = -X'_i$, and its expected value is $2E|X|$; thus T_n is a (normalized) two-sample energy statistic.

With this standardized statistic, one can apply the chi-squared test criterion in Székely and Bakirov (2003). That is, reject the null hypothesis at significance level α if $T_n \ge (\Phi^{-1}(1-\alpha/2))^2$, where Φ is the standard normal cdf. This criterion is valid for any significance level less than 0.215. (This test criterion tends to be conservative in general.)

For an interesting discussion of a special case of the inequality $E|X+X'| \ge E|X-X'|$ see Mensheinin and Zubkov (2012).

6.3. Distance components: a nonparametric extension of ANOVA

A multi-sample test of equal distributions is a type of generalization of the hypothesis of equal means. Analogous to the ANOVA decomposition of variance we can obtain a decomposition of distances called distance components (DISCO) and a test statistic for the K-sample hypothesis $H_0 : F_1 = \cdots = F_K$, $K \ge 2$. To simplify subsequent notation, for two samples $A = \{a_1,\dots,a_{n_1}\}$, $B = \{b_1,\dots,b_{n_2}\}$, let

$$g_\alpha(A,B) := \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2} |a_i-b_m|^\alpha, \tag{6.4}$$

for $0 < \alpha \le 2$. The multi-sample statistics are defined as follows. If $A_1,\dots,A_K$ are the samples of sizes $n_1,n_2,\dots,n_K$, respectively, and $N = \sum_{j=1}^K n_j$, we define the total dispersion of the observed response by

$$T_\alpha = T_\alpha(A_1,\dots,A_K) = \frac{N}{2} g_\alpha(A,A), \tag{6.5}$$

where A is the pooled sample (size N) and g_α is given by (6.4). The within-sample dispersion statistic is defined by

$$W_\alpha = W_\alpha(A_1,\dots,A_K) = \sum_{j=1}^K \frac{n_j}{2} g_\alpha(A_j,A_j). \tag{6.6}$$

The between-sample energy statistic is

$$S_{n,\alpha} = \sum_{1\le j<k\le K} \left(\frac{n_j+n_k}{2N}\right)\left[\frac{n_j n_k}{n_j+n_k}\mathcal{E}^{(\alpha)}_{n_j,n_k}(A_j,A_k)\right] = \sum_{1\le j<k\le K}\left\{\frac{n_j n_k}{2N}(2g_\alpha(A_j,A_k) - g_\alpha(A_j,A_j) - g_\alpha(A_k,A_k))\right\}, \tag{6.7}$$

Then if $0 < \alpha \le 2$ we have the decomposition $T_\alpha = S_\alpha + W_\alpha$, where both S_α and W_α are nonnegative.

If $0 < \alpha < 2$, the statistic (6.7) determines a statistically consistent test of the hypothesis that the distributions are identical (Rizzo and Székely, 2010). If $\alpha = 2$ the corresponding energy distance can be zero if the means of the distributions are identical. In fact, in the case where F_j are univariate distributions and $\alpha = 2$, the statistic $S_{n,2}$ is the ANOVA between sample sum of squared error (sum of squares for treatments) and the decomposition $T_2 = S_2 + W_2$ is the ANOVA decomposition. By choosing $\alpha = 1$ or any $0 < \alpha < 2$ as the exponent on Euclidean distance, we obtain a test of equality of distributions that is consistent against all alternatives with finite α moments.

The power of the DISCO test of the multisample hypothesis $H_0 : F_1 = \cdots = F_K$ is illustrated in the following power comparison. The test has been implement by permutation bootstrap in the `disco` function of the *energy* package for R (Rizzo and Székely, 2011). For more examples, see Rizzo and Székely (2010).

Example 3. For this simulated data, the multivariate response is in R^p and the DISCO test is compared with MANOVA tests based on Wilks Lambda and Pillai statistics. There are four groups each with sample size $n = 30$. Here groups 2–4 have iid marginal Gamma (shape = 2, rate = 0.1) distributions, while group 1 is Gamma (shape = 2, rate = 0.1) with multiplicative errors distributed as Lognormal ($\mu = 0, \sigma$). (The natural logarithm of the group 1 response has an additive normally distributed error with mean 0 and variance σ^2.)

Empirical power performance is compared at significance level 10%, which is summarized in Fig. 3(a) and (b). In Fig. 3(a) at $\sigma = 0$ one can check that each test achieves approximately the correct significance level 10% under the null. The standard error of the power estimate is at most 0.005, based on 10,000 tests.

In Fig. 3(a) the parameter σ is varied while dimension is fixed at $p = 10$. Each test has power increasing with σ, but the DISCO test is clearly more powerful than the MANOVA tests in this example. In Fig. 3(b) the dimension p is varied while $\sigma = 0.4$ remains fixed. In this simulation one can see that the difference in power between the DISCO and MANOVA tests is increasing with dimension.

For an interesting application of distance components analysis in behavioral biology see Schilling et al. (2012).

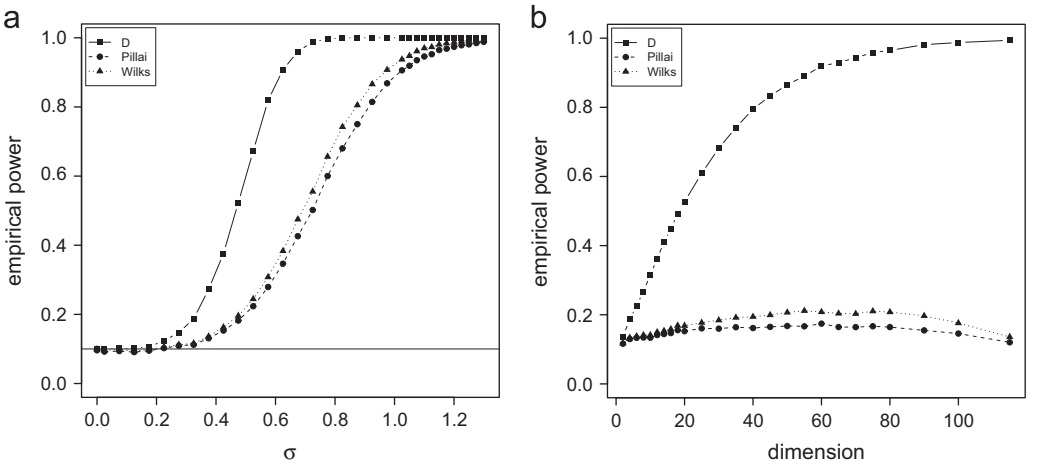

Fig. 3. Monte Carlo results for Example 3. Power of the DISCO (D) and MANOVA (Wilks, Pillai) tests for Gamma(shape$=$2, rate$=$0.1) data, in four groups with $n=30$ per group. Group 1 is Gamma(shape$=$2, rate$=$0.1) with multiplicative errors distributed as Lognormal($\mu=0,\sigma$). In (a) dimension $p=10$ and σ varies and in (b) p varies and $\sigma=0.4$.

6.4. $\mathcal{E}$-clustering: an extension of Ward's minimum variance method

Energy distance can be applied in hierarchical cluster analysis. In agglomerative hierarchical clustering algorithms, at each step, we seek to merge clusters that are homogeneous, as measured by the algorithm's *cluster distance*, while individual clusters are well-separated according to this cluster distance. In $\mathcal{E}$-clustering, we seek to merge clusters with minimum energy distance.

For a fairly general class of hierarchical clustering algorithms including Ward's minimum variance, an algorithm is uniquely determined by its recursive formula for updating cluster distances (see Székely and Rizzo, 2005b for details). The energy clustering algorithm is also identified by the same type of recursive formula. Suppose at the current step in the hierarchical clustering, the disjoint clusters C_i,C_j would be merged. Then the $\mathcal{E}$-distance between the new cluster $C_i\cup C_j$ and a disjoint cluster C_k is given by the following recursive formula:

$$d(C_i\cup C_j,C_k)=\frac{n_i+n_k}{n_i+n_j+n_k}\,d(C_i,C_k)+\frac{n_j+n_k}{n_i+n_j+n_k}\,d(C_j,C_k)-\frac{n_k}{n_i+n_j+n_k}\,d(C_i,C_j), \tag{6.8}$$

where $d(C_i,C_j)=\mathcal{E}_{n_i,n_j}(C_i,C_j)$, and n_i,n_j,n_k are the number of elements in clusters C_i,C_j,C_k, respectively. From formula (6.8) above, if $d_{ij}:=\mathcal{E}(C_i,C_j)$ is given by (6.1), $\mathcal{E}$-distance can be computed recursively by

$$d_{(ij)k}:=d(C_i\cup C_j,C_k)=\alpha_i d(C_i,C_k)+\alpha_j d(C_j,C_k)+\beta d(C_i,C_j)=\alpha_i d_{ik}+\alpha_j d_{jk}+\beta d_{ij}+\gamma|d_{ik}-d_{jk}|, \tag{6.9}$$

where

$$\alpha_i=\frac{n_i+n_k}{n_i+n_j+n_k},\quad \beta=\frac{-n_k}{n_i+n_j+n_k},\quad \gamma=0. \tag{6.10}$$

In the recursive equation (6.9) if we substitute squared Euclidean distances for Euclidean distances, with the same parameters (6.10), we have the updating formula for Ward's minimum variance method. Applying Proposition 2, we can replace Euclidean distances in (6.9) with $|x-y|^\alpha$ for any $0<\alpha\le 2$ to obtain a class of clustering algorithms that contain Ward's minimum variance method as a special case. By Proposition 2(ii) we know that Ward's method ($\alpha=2$) is a geometrical method that separates and identifies clusters by their centers; consistency does not hold for Ward's method ($\alpha=2$), because cluster distance is zero when groups have equal means, while the underlying populations could have different distributions. On the other hand, Proposition 2(i) implies that for every $0<\alpha<2$ the energy clustering algorithm separates clusters that differ in distribution (in any way).

The ability of $\mathcal{E}$ to separate and identify clusters with equal or nearly equal centers is potentially an important practical advantage over geometric or cluster center methods such as centroid, median, or Ward's minimum variance methods. In Székely and Rizzo (2005b), simulations showed that $\mathcal{E}$-clustering effectively recovers the underlying hierarchical structure in several different scenarios, including high dimensional data, and data with attributes on different scales. Moreover, in an example clustering simulated normal data with different covariance but nearly equal means, $\mathcal{E}$ outperformed six standard methods compared. Overall in our empirical results the theoretical properties of $\mathcal{E}$ are indeed an advantage for certain clustering problems, without sacrificing the good properties of Ward's minimum variance method for separating spherical clusters.

7. Distance correlation: measuring dependence and the energy test of independence

In this section we focus on dependence coefficients *distance covariance* and *distance correlation* introduced in Székely et al. (2007) that measure all types of dependence between random vectors X and Y in arbitrary dimension. The corresponding energy statistics have simple computing formulae, and they apply to sample sizes $n \geq 2$ (n can be much smaller than dimension).

To quote Newton (2009)

> Distance covariance not only provides a bona fide dependence measure, but it does so with a simplicity to satisfy Don Geman's elevator test (i.e., a method must be sufficiently simple that it can be explained to a colleague in the time it takes to go between floors on an elevator!).

The distance covariance statistic is computed as follows. First we compute all the pairwise distances between sample observations of the X sample, to get a distance matrix. Similarly compute a distance matrix for the Y sample. Next, we center the entries of these distance matrices so that their row and column means are equal to zero. A very simple formula (7.10) accomplishes the centering. Now take the centered distances $A_{k\ell}$ and $B_{k\ell}$ and compute the sample distance covariance as the square root of

$$\mathcal{V}_n^2 = \frac{1}{n^2} \sum_{k,\ell=1}^{n} A_{k\ell} B_{k\ell}.$$

The statistic $\mathcal{V}_n$ converges almost surely to distance covariance (dCov), $\mathcal{V}(X,Y)$, to be defined below, which is always nonnegative and equals zero if and only if X and Y are independent. Once we have dCov, we can define distance variance (dVar), and distance correlation (dCor) is computed as the normalized coefficient analogous to Pearson's correlation ρ.

Pearson's product-moment covariance measures linear dependence and in the bivariate normal case $\rho = 0$ is equivalent to independence. For the multivariate normal distribution, diagonal covariance matrix implies independence, but the converse does not hold in general. More generally, in the case of quadrant dependent random variables (including the multivariate normal), zero correlation(s) are equivalent to (mutual) independence (Lehmann, 1966). In general, however, Pearson's correlation and covariance do not characterize independence. Distance covariance and distance correlation are more general measures of independence as they do characterize independence of random vectors.

Classical inference based on normal theory tests the hypothesis of multivariate independence via a likelihood ratio statistic based on the covariance matrix of (X,Y) or their marginal ranks. These tests are not consistent against general alternatives, because like correlation measures, the statistics measure linear or monotone association. The distance covariance energy test is based on measuring the difference between the joint and marginal characteristic functions, thus it characterizes independence. For other recent consistent tests of bivariate or multivariate independence see e.g. Feuerverger (1993), Gretton and Györfi (2010), and Gretton and Györfi (2012).

In the special case when (X,Y) are jointly distributed as bivariate normal, distance correlation $(\mathcal{R})$ is a deterministic function of Pearson's correlation $\rho = \rho(X,Y)$ (Székely et al., 2007, Theorem 7):

$$\mathcal{R}^2(X,Y) = \frac{\rho \arcsin \rho + \sqrt{1-\rho^2} - \rho \arcsin \frac{\rho}{2} - \sqrt{4-\rho^2} + 1}{1 + \pi/3 - \sqrt{3}}.$$

Note that $\mathcal{R}(X,Y) \leq |\rho(X,Y)|$ with equality when $\rho = 0$ or $\rho = \pm 1$.

7.1. Definitions of distance covariance and distance correlation

In this section, we suppose that X in R^p and Y in R^q are random vectors, where p and q are positive integers. If $\hat{f}_X$ and $\hat{f}_Y$ denote the characteristic functions of X and Y, respectively, and their joint characteristic function is denoted $\hat{f}_{X,Y}$, then X and Y are independent if and only if $\hat{f}_{X,Y} = \hat{f}_X \hat{f}_Y$.

We define the $\|\cdot\|_w$-norm for complex functions γ defined on $\gamma : R^p \times R^q \mapsto R$ in the weighted L_2 space of functions on R^{p+q} by

$$\|\gamma(t,s)\|_w^2 = \int_{R^{p+q}} |\gamma(t,s)|^2 w(t,s)\, dt\, ds, \tag{7.1}$$

where $w(t,s)$ is an arbitrary positive weight function for which the integral above exists.

Distance covariance is defined as a measure of the distance between $\hat{f}_{X,Y}$ and $\hat{f}_X \hat{f}_Y$: it is the nonnegative square root of

$$\mathcal{V}^2(X,Y;w) = \|\hat{f}_{X,Y}(t,s) - \hat{f}_X(t)\hat{f}_Y(s)\|_w^2 = \int_{R^{p+q}} |\hat{f}_{X,Y}(t,s) - \hat{f}_X(t)\hat{f}_Y(s)|^2 w(t,s)\, dt\, ds,$$

where w is a suitable weight function. The choice of weight function is critical to obtaining a distance covariance with the properties one would require for a useful measure of dependence.

We also define (analogous to variance)

$$\mathcal{V}^2(X;w)=\mathcal{V}^2(X,X;w)=\|\hat{f}_{X,X}(t,s)-\hat{f}_X(t)\hat{f}_X(s)\|_w^2=\int_{R^{2p}}|\hat{f}_{X,X}(t,s)-\hat{f}_X(t)\hat{f}_X(s)|^2w(t,s)\,dt\,ds.$$

A standardized version of $\mathcal{V}(X,Y;w)$ is

$$\mathcal{R}_w=\frac{\mathcal{V}(X,Y;w)}{\sqrt{\mathcal{V}(X;w)\mathcal{V}(Y;w)}},$$

which is an unsigned "correlation" coefficient.

Our definitions above are rotation invariant. We further require that $\mathcal{R}_w$ be scale invariant. This will follow if we require that $\mathcal{V}_w$ be scale equivariant. This property holds if the weight function w is proportional to the reciprocal of $|t|_p^{1+p}|s|_q^{1+q}$ (see (7.2)–(7.3)). In Székely and Rizzo (2012) it was proved that the distance covariance weight function is unique. That is, scale invariance and rigid motion invariance, along with natural technical conditions such as integrability of the weight function, imply the uniqueness of w. Thus, in the following definitions, the weight function $w(t,s)=(c_pc_q|t|_p^{1+p}|s|_q^{1+q})^{-1}$ is applied, where c_p,c_q are given by (2.7).

Definition 3. The *distance covariance* between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{V}(X,Y)$ defined by

$$\mathcal{V}^2(X,Y)=\|\hat{f}_{X,Y}(t,s)-\hat{f}_X(t)\hat{f}_Y(s)\|^2=\frac{1}{c_pc_q}\int_{R^{p+q}}\frac{|\hat{f}_{X,Y}(t,s)-\hat{f}_X(t)\hat{f}_Y(s)|^2}{|t|_p^{1+p}\,|s|_q^{1+q}}\,dt\,ds. \tag{7.2}$$

If $E|X|_p<\infty$ and $E|Y|_q<\infty$ then by Lemma 1 and by Fubini's theorem we can evaluate

$$\mathcal{V}^2(X,Y)=E[|X-X'|_p|Y-Y'|_q]+E|X-X'|_p\,E|Y-Y''|_q-2E[|X-X'|_p|Y-Y''|_q], \tag{7.3}$$

where (X,Y), (X',Y'), and (X'',Y'') are iid.

Distance variance is defined as the square root of

$$\mathcal{V}^2(X)=\mathcal{V}^2(X,X)=\|\hat{f}_{X,X}(t,s)-\hat{f}_X(t)\hat{f}_X(s)\|^2.$$

By definition of the norm $\|\cdot\|$, it is clear that $\mathcal{V}(X,Y)\geq 0$ and $\mathcal{V}(X,Y)=0$ if and only if X and Y are independent.

Definition 4. The *distance correlation* (*dCor*) between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{R}(X,Y)$ defined by

$$\mathcal{R}^2(X,Y)=\begin{cases}\dfrac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y)>0,\\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y)=0.\end{cases} \tag{7.4}$$

The energy dependence statistics are defined as follows. The sample distance covariance statistic $\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})$ introduced at the beginning of this section has a simple form (7.11). It is equivalent to the following definition.

Let $\hat{f}_X^n(t)$, $\hat{f}_Y^n(s)$, and $\hat{f}_{X,Y}^n(t,s)$ denote the empirical characteristic functions of the samples $\mathbf{X}$, $\mathbf{Y}$, and $(\mathbf{X},\mathbf{Y})$, respectively. It is natural to consider a statistic based on the L_2 norm of the difference between the empirical characteristic functions; that is, to substitute the empirical characteristic functions for the characteristic functions in the definition of the norm (7.1).

A key result (Székely et al., 2007, Theorem 1) is the following:

If $(\mathbf{X},\mathbf{Y})$ is a random sample from the joint distribution of (X,Y), then

$$\|\hat{f}_{X,Y}^n(t,s)-\hat{f}_X^n(t)\hat{f}_Y^n(s)\|^2=S_1+S_2-2S_3, \tag{7.5}$$

where

$$S_1=\frac{1}{n^2}\sum_{k,\ell=1}^n|X_k-X_\ell|_p|Y_k-Y_\ell|_q, \tag{7.6}$$

$$S_2=\frac{1}{n^2}\sum_{k,\ell=1}^n|X_k-X_\ell|_p\frac{1}{n^2}\sum_{k,\ell=1}^n|Y_k-Y_\ell|_q, \tag{7.7}$$

$$S_3=\frac{1}{n^3}\sum_{k=1}^n\sum_{\ell,m=1}^n|X_k-X_\ell|_p|Y_k-Y_m|_q, \tag{7.8}$$

and

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = S_1 + S_2 - 2S_3, \tag{7.9}$$

where $\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})$ is given by (7.11) defined below.

For a random sample $(\mathbf{X},\mathbf{Y}) = \{(X_k,Y_k) : k = 1,\dots,n\}$ of n iid random vectors (X,Y) from the joint distribution of random vectors X in R^p and Y in R^q, compute the Euclidean distance matrices $(a_{k\ell}) = (|X_k - X_\ell|_p)$ and $(b_{k\ell}) = (|Y_k - Y_\ell|_q)$. Define the centered distances

$$A_{k\ell} = a_{k\ell} - \overline{a}_{k\cdot} - \overline{a}_{\cdot\ell} + \overline{a}_{\cdot\cdot}, \quad k,\ell = 1,\dots,n, \tag{7.10}$$

where

$$\overline{a}_{k\cdot} = \frac{1}{n}\sum_{\ell=1}^{n} a_{k\ell}, \quad \overline{a}_{\cdot\ell}, = \frac{1}{n}\sum_{k=1}^{n} a_{k\ell}, \quad \overline{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,\ell=1}^{n} a_{k\ell}.$$

Similarly, define $B_{k\ell} = b_{k\ell} - \overline{b}_{k\cdot} - \overline{b}_{\cdot\ell} + \overline{b}_{\cdot\cdot}$, for $k,\ell = 1,\dots,n$.

The sample distance covariance $\mathcal{V}_n(\mathbf{X},\mathbf{Y})$ and sample distance correlation $\mathcal{R}_n(\mathbf{X},\mathbf{Y})$ are defined by

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2}\sum_{k,\ell=1}^{n} A_{k\ell}B_{k\ell}, \tag{7.11}$$

and

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

respectively, where the sample distance variance is defined by

$$\mathcal{V}_n^2(\mathbf{X}) = \mathcal{V}_n^2(\mathbf{X},\mathbf{X}) = \frac{1}{n^2}\sum_{k,\ell=1}^{n} A_{k\ell}^2.$$

As a corollary we have that $\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) \geq 0$, $\mathcal{V}_n^2(\mathbf{X}) \geq 0$.

One can also show (Székely et al., 2007, Theorem 2) that we have the almost sure convergence:

$$\lim_{n\to\infty} \mathcal{V}_n(\mathbf{X},\mathbf{Y}) = \mathcal{V}(X,Y);$$

$$\lim_{n\to\infty} \mathcal{R}_n^2(\mathbf{X},\mathbf{Y}) = \mathcal{R}^2(X,Y).$$

Under independence $n\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})$ converges in distribution to a quadratic form $Q \overset{\mathcal{D}}{=} \sum_{j=1}^{\infty}\lambda_j Z_j^2$, where Z_j are independent standard normal random variables, and $\{\lambda_j\}$ are nonnegative constants that depend on the distribution of (X,Y) (Székely et al., 2007, Theorem 5). (For more details on λ_j see Section 8.) Under dependence of (X,Y), $n\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) \to \infty$ as $n \to \infty$, hence a test that rejects independence for large $n\mathcal{V}_n^2$ is consistent against dependent alternatives.

On the definition of population distance covariance: Population distance covariance of random vectors X and Y with finite expectations can alternately be defined along the same lines as the sample distance covariance. For a pair of vectors $x,x' \in R^p$, and a random vector $X \in R^p$, introduce the notation

$$m_X(x) := E[|X - x|_p], \quad \overline{m}_X := E[m_X(X)],$$

and

$$d_X(x,x') := |x - x'|_p - m_X(x) - m_X(x') + \overline{m}_X,$$

which is a doubly centered distance analogous to (7.10). Then an equivalent definition of population distance covariance is

$$\mathcal{V}^2(X,Y) = E[d_X(X,X')d_Y(Y,Y')], \tag{7.12}$$

which may be heuristically easier to understand than (7.3).

Distance covariance and distance correlation can be extended via (7.2), (7.11) and (7.12) to measure dependence of several random variables. For example, dCor(X,Y,Z) measures the mutual dependence of X,Y,Z. The definition and theory of *partial distance correlation*, however, is more complex; it will be addressed in a forthcoming paper.

Some other important properties of distance covariance are

(i) $\mathcal{V}(a_1 + b_1C_1X, a_2 + b_2C_2Y) = \sqrt{|b_1b_2|}\,\mathcal{V}(X,Y)$, for all constant vectors $a_1 \in R^p$, $a_2 \in R^q$, scalars b_1, b_2 and orthonormal matrices C_1, C_2 in R^p and R^q, respectively.

(ii) Distance covariance is not covariance of distances, but (applying (7.3)) it can be expressed in terms of Pearson's covariance of distances as

$$\mathcal{V}^2(X,Y) = \mathrm{Cov}(|X - X'|_p, |Y - Y'|_q) - 2\,\mathrm{Cov}(|X - X'|_p, |Y - Y''|_q).$$

It is interesting to note that $\mathrm{Cov}(|X-X'|p,|Y-Y'|q)=0$ does not imply independence of X and Y. Indeed, there is a simple two-dimensional random variable (X,Y) such that X and Y are not independent, but $|X-X'|$ and $|Y-Y'|$ are uncorrelated.

Some additional properties of distance variance are:

(i) $\mathcal{V}(X)=0$ implies that $X=E[X]$, almost surely.
(ii) $\mathcal{V}_n(\mathbf{X})=0$ if and only if every sample observation is identical.
(iii) If X and Y are independent, then $\mathcal{V}(X+Y)\le\mathcal{V}(X)+\mathcal{V}(Y)$. Equality holds if and only if one of the random vectors X or Y is constant.
(iv) $\mathcal{V}(a+bCX)=|b|\mathcal{V}(X)$, for all constant vectors a in R^p, scalars b, and $p\times p$ orthonormal matrices C.

In addition to the properties stated above for distance correlation, we have

(i) $0\le\mathcal{R}_n(\mathbf{X},\mathbf{Y})\le 1$.
(ii) $\mathcal{R}_n(\mathbf{X},\mathbf{Y})=1$ implies that the dimensions of the linear subspaces spanned by $\mathbf{X}$ and $\mathbf{Y}$ respectively are almost surely equal, and if we assume that these subspaces are equal then in this subspace

$$\mathbf{Y}=a+b\mathbf{X}C$$

for some vector a, nonzero real number b and orthogonal matrix C.

See Székely et al. (2007) and Székely and Rizzo (2009) for proofs of these results and other properties. Also see Dueck et al. (submitted for publication) on affine invariant distance correlation.

Expected value of distance covariance: The statistic $\mathcal{V}_n^2(X,Y)$ is asymptotically unbiased for $\mathcal{V}^2(X,Y)$. Under independence of X and Y,

$$E[\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})]=\frac{n-1}{n^2}E|X-X'||Y-Y'|.$$

In general, one can derive that

$$E[\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})]=\frac{(n-1)(n-2)^2}{n^3}\mathcal{V}^2(X,Y)+\frac{2(n-1)^2}{n^3}\gamma-\frac{(n-1)(n-2)}{n^3}\alpha\beta,$$

where $\gamma=E|X-X'||Y-Y'|$, $\alpha=E|X-X'|$, and $\beta=E|Y-Y'|$. Here

$$\hat{\gamma}=\frac{1}{n(n-1)}\sum_{i,j=1}^{n}a_{ij}b_{ij},\quad \widehat{\alpha\beta}=\frac{1}{n(n-1)(n-2)(n-3)}\sum_{i,j=1}^{n}\sum_{k,m\notin\{i,j\}}a_{ij}b_{km},$$

are unbiased estimators of γ and $\alpha\beta$, respectively, and the statistic

$$U_n^2(X,Y)=\frac{1}{(n-1)(n-2)^2}\left(n^3\mathcal{V}_n^2(X,Y)-\frac{2}{n}\sum_{i,j=1}^{n}a_{ij}b_{ij}+\frac{1}{n(n-3)}\sum_{i,j=1}^{n}\sum_{k,m\notin\{i,j\}}a_{ij}b_{km}\right)$$

is unbiased for $\mathcal{V}^2(X,Y)$. A simpler and faster computing formula for an unbiased estimator of $\mathcal{V}^2(X,Y)$ is

$$\mathcal{U}_n^2(X,Y)=\frac{1}{n(n-1)}\sum_{i,j=1}^{n}a_{ij}b_{ij}+\sum_{i,j=1}^{n}\frac{a_{ij}(b_{..}-2b_{i.}-2b_{.j}+2b_{ij})}{n(n-1)(n-2)(n-3)}-2\sum_{i,j=1}^{n}\frac{a_{ij}(b_{i.}-b_{ij})}{n(n-1)(n-2)},$$

where $b_{i.},b_{.j},b_{..}$ are the row i sum, column j sum, and grand sum, respectively of the distance matrix (b_{ij}).

These results apply in arbitrary dimensions. In addition, for high dimensional problems, one can apply an alternate type of correction for bias and obtain a distance correlation t-test of independence (Székely and Rizzo, 2013).

Example 4 (*Measuring nonlinear dependence*). In this example, which originally appeared in Székely and Rizzo (2009), we illustrate how to isolate the nonlinear dependence between random vectors to test for nonlinearity. The bivariate data follow the Gumbel bivariate exponential distribution, which has density function

$$f(x,y;\theta)=[(1+\theta x)(1+\theta y)]\exp(-x-y-\theta xy),\quad x,y>0;\ 0\le\theta\le 1.$$

Here the marginal distributions are standard exponential, so there is a nonlinear, but monotone dependence relation between X and Y. The conditional densities have the form

$$f(y|x)=e^{-(1+\theta x)y}[(1+\theta x)(1+\theta y)-\theta],\quad y>0.$$

The correlation between X and Y depends on θ. If $\theta=0$ then $\rho=\rho(X,Y)=0$, and it is easy to check that independence holds. If $\theta=1$ then $\rho=-0.40365$.

Random samples from the Gumbel bivariate exponential were generated with $\theta=0.5$. Empirical power of dCov and correlation tests are compared in Fig. 4 for sample sizes varying from 10 to 300 (10,000 replications for each $n=10,20,\ldots$). Fig. 4(a) reveals that the distance correlation test and correlation test are comparable in power against this alternative. This

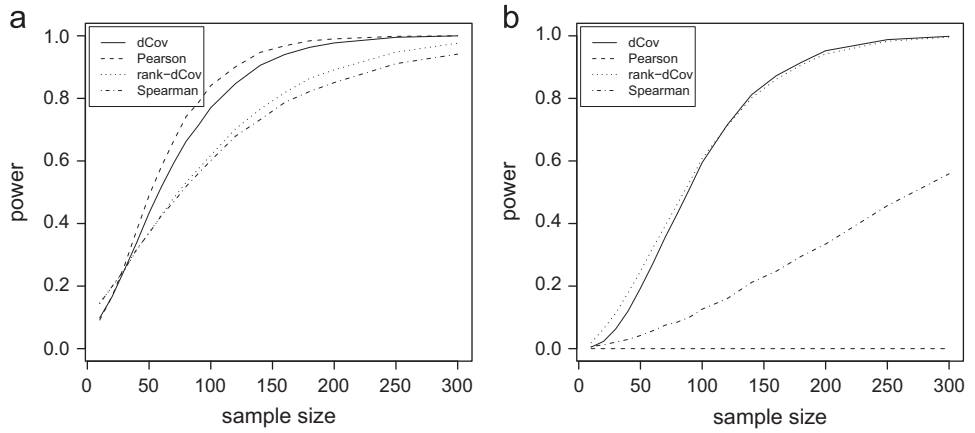

Fig. 4. Power comparison of distance covariance and correlation tests at 10% significance level for Gumbel's bivariate exponential distribution. In subfigure (a), the null is $H: X,Y$ are independent; in (b) the hypothesis tested is independence of X and $\hat{\varepsilon}$, where $\hat{\varepsilon}$ are residuals of the regression of Y on X.

is not surprising because $E[Y|X=x]=(1+\theta+x\theta)/(1+x\theta)^2$ is monotone. In this example the rank tests (rank-dCov and Spearman correlation test) were also applied after first transforming the **X** and **Y** samples to ranks.

One of the great advantages of distance correlation is that it can detect all types of dependence, including any nonlinear or non-monotone association. We can effectively use this property to measure the lack-of-fit of a linear model, which is illustrated next.

Although we cannot split the dCor or dCov coefficient into linear and nonlinear components, we can extract the linear component from the data first by regressing Y on X, and then compute dCor on the residuals. Then we can separately analyze the linear and nonlinear components of bivariate or multivariate dependence relations.

To extract the linear component of dependence, first we fit a linear model $Y=X\beta+\varepsilon$ to the sample $(\mathbf{X},\mathbf{Y})$ by ordinary least squares. We can then apply the dCov test of independence to $(\mathbf{X},\hat{\varepsilon})$.

The results are illustrated for our Gumbel bivariate exponential ($\theta=0.5$) data in Fig. 4(b). The power of dCov tests is increasing to 1 with sample size, which demonstrates consistency against the nonlinear dependence that remains in the residuals of the linear model.

This "lack-of-fit" procedure for a linear model is easily applied in arbitrary dimension. One can fit a linear multiple regression model to extract the linear component of dependence. One could also apply this method to a model with a multivariate response. This has important practical application for evaluating models in higher dimension.

In addition to testing independence or nonlinearity, there are several other important applications. In Székely and Rizzo (2009), dCov statistics were applied to identify influential observations. Distance covariance has been applied by Matteson and Tsay (2011) for independent component analysis. Li et al. (2012) applied distance correlation for feature screening for ultra-high dimensional data. For another interesting application of distance correlation see Kong et al. (2012).

7.2. Generalization of distance covariance for heavy tailed distributions

Based on Lemma 1 one can easily generalize the definition of distance covariance to α-distance covariance as follows:

Definition 5. For $0<\alpha<2$, and random vectors X and Y such that $E(|X|^\alpha+|Y|^\alpha)<\infty$, define the *$\alpha$-distance covariance* $\mathrm{d}Cov_\alpha(X,Y)$ as the nonnegative square root of $\mathcal{V}^{2^{(\alpha)}}(X,Y)$ defined by

$$\mathcal{V}^{2^{(\alpha)}}(X,Y)=\frac{1}{C(p,\alpha)C(q,\alpha)}\int_{R^{p+q}}\frac{|\hat{f}_{X,Y}(t,s)-\hat{f}_X(t)\hat{f}_Y(s)|^2}{|t|_p^{\alpha+p}|s|_q^{\alpha+q}}\,dt\,ds \tag{7.13}$$

If $E|X|_p^\alpha<\infty$ and $E|Y|_q^\alpha<\infty$ then

$$\mathcal{V}^{2^{(\alpha)}}(X,Y)=E[|X-X'|_p^\alpha|Y-Y'|_q^\alpha]+E|X-X'|_p^\alpha E|Y-Y'|_q^\alpha-E[|X-X'|_p^\alpha|Y-Y''|_q^\alpha]-E[|X-X''|_p^\alpha|Y-Y'|_q^\alpha]. \tag{7.14}$$

Based on this generalized distance covariance, we can generalize distance correlation. It is clear that $\alpha=1$ corresponds to distance correlation (7.4).

Interestingly, if the α-distance correlation is computed from the α-distance covariance when $\alpha=2$, and $p=q=1$, then the distance covariance in (7.14) is

$$E|X-X'|^2|Y-Y'|^2+E|X-X'|^2E|Y-Y'|^2-E|X-X'|^2E|Y-Y''|^2-E|X-X''|^2E|Y-Y'|^2.$$

From this formula one can easily show that the ($\alpha=2$)-distance correlation is exactly the same as the absolute value of Pearson's correlation $\rho(X,Y)$ in the bivariate case. This correlation ρ cannot be expressed in terms of characteristic functions, like in (7.13), and in fact we know that $\rho=0$ does not imply independence. Despite this fact, both distance correlation and the absolute value of Pearson's correlation are special cases of generalized α-distance correlation.

By this generalization, note that one can apply distance covariance to test independence of distributions that do not have finite variance. For example, financial data are often modeled by stable distributions; in this case one can choose α such that $E|X-X'|^{2\alpha}<\infty$ (see Yang, 2012 for a proof) and an energy test can be applied.

We have applied the corresponding generalization in energy goodness-of-fit for Pareto family (Rizzo, 2009), Cauchy and stable distributions (Yang, 2012). For the case of Pareto with $\alpha<1$ or non-Gaussian stable distribution, our statistic can be applied using a smaller value of α. (If variance is finite, then we do not find that different choices of α in (0, 2) have significantly different performance in terms of power.)

7.3. Brownian covariance

There is a very interesting duality between distance covariance and a covariance with respect to a stochastic process, defined below. We will see that when the stochastic process is Brownian motion (Wiener process) the *Brownian covariance* coincides with distance covariance ($\alpha=1$). For more details see Székely and Rizzo (2009). See also the discussion of Genovese (2009). One can show that fractional Brownian motion-covariance with Hurst parameter H corresponds to generalized distance covariance (7.13) with exponent $\alpha=2H$. (On fractional Brownian motion see e.g. Herbin and Merzbach, 2007.) When the stochastic process is identity, we obtain classical product-moment covariance.

To motivate Definition 6, first, consider two real-valued random variables X,Y. The square of their ordinary covariance can be written as

$$E^2[(X-E(X))(Y-E(Y))]=E[(X-E(X))(X'-E(X'))(Y-E(Y))(Y'-E(Y'))].$$

Now define the square of conditional covariance, given two real-valued stochastic processes $U(\cdot)$ and $V(\cdot)$. If $X\in R$ and $\{U(t):t\in R\}$ is a real-valued stochastic process, independent of X, define the U-centered version of X:

$$X_U=U(X)-\int_{-\infty}^{\infty}U(t)\,dF_X(t)=U(X)-E[U(X)|U],$$

whenever the conditional expectation exists. Notice that $X_{id}=X-E[X]$, where *id* is the identity.

Next consider a two-sided, one-dimensional Brownian motion (Wiener process) W with expectation zero and covariance function

$$|s|+|t|-|s-t|=2\min(s,t),\quad t,s\geq 0.$$

(This is twice the covariance of the standard Brownian motion.)

Definition 6. The *Brownian covariance* of two real-valued random variables X and Y with finite first moments is a nonnegative number defined by its square

$$\mathcal{W}^2(X,Y)=\mathrm{Cov}_W^2(X,Y)=E[X_WX'_WY_{W'}Y'_{W'}],$$

where (W,W') does not depend on (X,Y,X',Y').

If W in Cov_W is replaced by the (non-random) identity function *id*, then $\mathrm{Cov}_{id}(X,Y)=|\mathrm{Cov}(X,Y)|$ is the absolute value of product-moment covariance.

Definition 6 can be extended to random processes and random vectors in higher dimension; see Székely and Rizzo (2009) for details. The *Brownian variance* is defined by

$$\mathcal{W}(X)=\mathrm{Var}_W(X)=\mathrm{Cov}_W(X,X),$$

and Brownian correlation is

$$\mathrm{Cor}_W(X,Y)=\frac{\mathcal{W}(X,Y)}{\sqrt{\mathcal{W}(X)\mathcal{W}(Y)}}$$

whenever the denominator is not zero; otherwise $\mathrm{Cor}_W(X,Y)=0$.

It was proved (Székely and Rizzo, 2009, Theorem 7) that $\mathrm{Cov}_W(X,Y)$ exists for random vectors X and Y with finite second moments.

We collect generalizations of distance covariance to Brownian covariance and distance covariance of Gaussian processes in the following theorem that is split into three parts for readability.

Theorem 1 (i). *If X is an R^p-valued random variable, Y is an R^q-valued random variable, and $E(|X|+|Y|)<\infty$, then $E[X_WX'_WY_{W'}Y'_{W'}]$ is nonnegative and finite, and*

$$\mathcal{W}^2(X,Y)=E[X_WX'_WY_{W'}Y'_{W'}]=E|X-X'||Y-Y'|+E|X-X'|E|Y-Y'|-E|X-X'||Y-Y''|-E|X-X''||Y-Y'|, \tag{7.15}$$

where (X,Y), (X',Y'), and (X'',Y'') are iid.

If we compare Theorem 1(i) and (7.3), there is a surprising coincidence: *Brownian covariance is equal to distance covariance*; that is, $\mathcal{W}(X,Y)=\mathcal{V}(X,Y)$ in arbitrary dimension. See Székely and Rizzo (2009, Theorem 8) for the proof of Theorem 1(i).

Now consider the Lévy fractional Brownian motion $\{W_H^d(t), t\in\mathbb{R}^d\}$ with Hurst index $H\in(0,1)$, which is a centered Gaussian random process with covariance function

$$E[W_H^d(t)W_H^d(s)]=|t|^{2H}+|s|^{2H}-|t-s|^{2H},\quad t,s\in R^d.$$

The following generalization of Theorem 1 follows by application of Lemma 1 (see Székely and Rizzo, 2009).

Theorem 1 (*ii*). *Let W_H and W'_{H^*} denote independent fractional Brownian motion processes with Hurst index H, and suppose that $X\in R^p, Y\in R^q$, $E|X|_p^h<\infty$, $E|Y|_q^{h^*}<\infty$, and $(X,Y),(X',Y'),(X'',Y'')$ are independent. Then for $0<H, H^*\le 1$, $h:=2H$, and $h^*:=2H^*$, we have*

$$\begin{aligned}\mathrm{Cov}^2_{W_H^p,W'^q_{H^*}}(X,Y)&=\frac{1}{C(p,h)C(q,h^*)}\int_{R^q}\int_{R^p}\frac{|f(t,s)-f(t)g(s)|^2\,dt\,ds}{|t|_p^{p+h}|s|_q^{q+h^*}}\\&=E|X-X'|_p^h|Y-Y'|_q^{h^*}+E|X-X'|_p^hE|Y-Y'|_q^{h^*}-E|X-X'|_p^h|Y-Y''|_q^{h^*}-E|X-X''|_p^h|Y-Y'|_q^{h^*}.\end{aligned}\tag{7.16}$$

Observe that when $h=h^*=1$, (7.16) is Eq. (7.15) of Theorem 1(i). That is, we have Brownian motion and Brownian covariance when $H=1/2$ and $\alpha=1$.

One can also generalize the notion of Brownian covariance to define a distance covariance for more general Gaussian processes.

Definition 7. Let ϕ_i, $i=1,2$ be two conditionally negative definite continuous symmetric functions from R to R, such that $E\phi_i(|X-X'|)<\infty$ and $E\phi_i(|Y-Y'|)<\infty$. Consider the zero mean Gaussian processes (Gaussian fields) G_i with covariance functions $k_i(s,t)=\phi_i(|s|)+\phi_i(|t|)-\phi_i(|s-t|)$. The (squared) distance covariance of X and Y with respect to G_1 and G_2 is defined as

$$\mathrm{Cov}^2_{G_1,G_2}(X,Y):=E[X_{G_1}X'_{G_1}Y_{G_2}Y'_{G_2}],\tag{7.17}$$

where (X,Y), (X',Y'), and (X'',Y'') are iid.

Theorem 1 (*iii*). *In Definition 7,*

$$\mathrm{Cov}^2_{G_1,G_2}(X,Y)=E[X_{G_1}X'_{G_1}Y_{G_2}Y'_{G_2}]\tag{7.18}$$

is nonnegative, finite, and

$$\mathrm{Cov}^2_{G_1,G_2}(X,Y)=E\phi_1(|X-X'|)\phi_2(|Y-Y'|)+E\phi_1(|X-X'|)E\phi_2(|Y-Y'|)-E\phi_1(|X-X'|)\phi_2(|Y-Y''|)-E\phi_1(|X-X''|)\phi_2(|Y-Y'|).$$

In addition to the functions $k(s,t)=|t|^{2H}+|s|^{2H}-|t-s|^{2H}$ (fractional Brownian motion), important examples of conditionally negative definite functions include the negative logarithm of the symmetric Laplace characteristic function, $\log(1+|t|^2)$. The covariance function of the corresponding Gaussian process is

$$k(s,t)=\text{constant}\times\log\left(\frac{(1+|s|^2)(1+|t|^2)}{1+|s-t|^2}\right).$$

This process (field) can be called a Laplace–Gaussian process. Another example is the negative logarithm of the characteristic function of the difference of two iid Poisson variables. The corresponding $\phi_i(t)=\cos t-1$ and

$$k(s,t)=\cos s+\cos t-\cos(s-t)-1,$$

and this process can be called a Poisson–Gaussian process. Details on proof and application of Theorem 1(iii) will be published elsewhere.

Remark 1. If Φ_i, $i=1,2$, denotes the centered version of ϕ_i such that the conditional expectations $E(\Phi_1(X-X')|X)$ and $E(\Phi_2(Y-Y')|Y)$ are equal to zero with probability one, then the Gaussian distance covariance is

$$\mathrm{Cov}^2_{G_1,G_2}(X,Y)=E[\Phi_1(X-X')\Phi_2(Y-Y')].\tag{7.19}$$

For the proof of a similar claim see Lyons (to appear).

7.4. Dependent observations

So far in this paper we have assumed that the observations are iid, but this is a stronger assumption than necessary. In fact it is sufficient to suppose that the sample $\{X_i : i=0,\pm1,\pm2,\ldots\}$ is strongly stationary and ergodic. To see this, observe

that the conditional negative definiteness of $\phi : R^d \to R$ implies that the kernel

$$h(x,y) = h_\phi(x,y) := E\phi(x-X) + E\phi(y-X) - E\phi(X-X') - \phi(x-y)$$

is positive semidefinite. If $Eh(X,X') < +\infty$, then under the null hypothesis that in the definition of the kernel h the observations have the same distribution as X, we have that

$$E[h(x,X_n)|X_1,X_2,\ldots,X_{n-1}] = 0, \quad n = 1,2,\ldots$$

almost surely (a.s.) with respect to the probability distribution of X. This implies that $E[h(x,X)] = 0$ a.s. Thus the kernel h is degenerate, but in fact we have more: a martingale difference type property. If we suppose that ϕ is also symmetric and continuous as in Proposition 3, then by Theorem 1 of Leucht and Neumann (2013) we have that the asymptotic distribution of the V-statistic with kernel h is of the same type as if the sample elements were iid: a quadratic form Q (8.2) of iid standard normal random variables.

It is also true that the strong law of large numbers applies to the corresponding energy statistics (Aaronson et al., 1996, Theorem U). This means that we can apply energy tests even if the sample elements are not iid, but only strongly stationary and ergodic, which is a standard regularity condition for time series, stochastic processes and random fields.

On the other hand, if we know that the observations are increments of a stochastic process (or random field) $X(t)$ with stationary increments, such that $E[X(t)] = 0$ and $\phi(t-s) := \mathrm{Var}(X(t)-X(s))$, $t,s \in R^d$, then ϕ is symmetric and conditionally negative definite (because $\mathrm{Var}(\sum_{1\le i\le n} c_i X(t_i)) \ge 0$). If ϕ is also continuous, then by Proposition 3 it is natural to apply the corresponding generalized energy kernel, $h_\phi(x,y)$, for statistical tests. In case of Brownian distance covariance we had iid observations and $X(t)$ was Brownian motion. For general Gaussian distance covariance (Definition 7), it makes sense to choose the Gaussian processes with $\phi(t) = \mathrm{Var}(X_1 + \cdots + X_t)$, which implies that for the corresponding covariance function $k(s,t)$ we have

$$\begin{aligned} 2k(s,t) &= 2E[X(t)X(s)] \\ &= E[X(t)^2] + E[X(s)^2] - E[X(t)-X(s)]^2 = \phi(t) + \phi(s) - \phi(t-s). \end{aligned}$$

We also have the converse identity for computing ϕ from k:

$$\phi(t-s) = k(t,t) + k(s,s) - 2k(s,t).$$

Theoretical foundation for degenerate kernel V-statistics (directly relevant to energy statistics for dependent observations) is found in the recent work of Leucht and Neumann (2013). For time series, definitions of an "auto" distance correlation analogous to auto correlation have been considered by Matteson and Tsay (2011), Rémillard (2009), and Zhou (2012).

8. Statistical potential and kinetic energy

Energy is one of the most fundamental concepts in science. Energy is the capacity of an item to do work, the capacity of acting. Potential energy is the general name of energy which has to do with the location of an object relative to something else, as in Newton's potential energy. If μ is a mass distribution or a probability distribution in R^3, $x \in R^3$, and $y \in R^3$, the Newton potential function of μ is defined by the formula

$$u(x) = \int_{R^3} \frac{d\mu(y)}{|x-y|},$$

where $|\cdot|$ denotes Euclidean distance. The potential energy function measures the energy necessary to move one unit mass from the location x to infinity in a gravitational space with mass distribution μ. The same formula describes Coulomb's electrostatic potential if μ denotes the charge distribution. As a result, according to classical ideas, the orbits of electrons in atoms are similar to the gravitational orbits in the solar system. The kernel $\gamma(x,y) = |x-y|^{-1} = r^{-1}$, where $r = |x-y|$ is traditionally called Green's function. If the exponent of r is not -1 but 2, then we get the law of elasticity (Hooke's law) where the force (the negative gradient of the potential) is proportional to the stretching (r) of a solid body, e.g. of a spiral spring. Thus in physics both positive and negative exponents are applicable.

For statistical applications in R^d we need the following extension:

$$u^{(\alpha)}(x) = \int_{R^d} |x-y|^\alpha \, d\mu(y) = E|x-Y|^\alpha,$$

where $\alpha \in R$, E is the expected value, and Y is a random variable with distribution μ and finite absolute α-moment.

Now let us revisit the goodness-of-fit problem based on the statistical energy $\mathcal{E}_n$ of samples of size n in the 'field' of a given probability distribution (null distribution), the distribution of a random vector X.

To apply $\mathcal{E}$ for univariate or multivariate goodness-of-fit tests of the null hypothesis $H_0 : X \sim F_0$, the energy statistic $\mathcal{E}_n$ (4.1) is a V-statistic with kernel $h : R^d \times R^d \to R$ defined by

$$h(x,y) = E|x-X| + E|y-X| - E|X-X'| - |x-y|. \tag{8.1}$$

By the law of large numbers for V-statistics (see e.g. Serfling, 1980 or Koroljuk and Borovskich, 1994), we have

$$\lim_{n\to\infty} \mathcal{E}_n = E[h(X,X')]$$

with probability one. Applying Proposition 1 we see that $\mathcal{E}(F,F_0)>0$ whenever $H_0: F=F_0$ is false. Hence under an alternative hypothesis $n\mathcal{E}_n\to\infty$ with probability one, as $n\to\infty$.

On the other hand, if H_0 is true, then the kernel h is degenerate; that is, $E[h(x,X)]=0$ for almost all $x\in R^d$. Thus $n\mathcal{E}_n$ has a finite limit distribution under the extra condition $E[h^2(X,X')]<\infty$ (see Serfling, 1980 or Koroljuk and Borovskich, 1994, Theorem 5.3.1). This result combined with the property that $n\mathcal{E}_n\to\infty$ under the alternative shows that tests can be constructed based on $\mathcal{E}_n$ that are consistent against general alternatives.

Under the null hypothesis, if $E[h^2(X,X')]<\infty$, the limit distribution of $n\mathcal{E}_n$ is a quadratic form

$$Q=\sum_{k=1}^{\infty}\lambda_k Z_k^2 \tag{8.2}$$

of iid standard normal random variables Z_k, $k=1,2,\ldots$ (Koroljuk and Borovskich, 1994, Theorem 5.3.1). The nonnegative coefficients λ_k are eigenvalues of the integral operator with kernel $h(x,y)$, satisfying the Hilbert–Schmidt eigenvalue equation

$$\int_{R^d} h(x,y)\psi(y)\,dF(y)=\lambda\psi(x). \tag{8.3}$$

We will call the eigenvalues λ the *statistical potential energy levels*.

The kernel h is symmetric ($h(x,y)=h(y,x)$), hence the eigenvalues are real. Since $|x-y|$ is conditionally negative definite, one can easily see that $h(x,y)$ is positive semidefinite, and thus all eigenvalues in (8.3) are nonnegative. It is also known that their sum is finite and equal to $E|X-X'|$.

The kernel h is degenerate; that is,

$$\int h(x,y)\,dF(y)=0.$$

Thus $\psi_0=1$ is an eigenfunction with eigenvalue 0. Since eigenfunctions with different eigenvalues are orthogonal we have for any ψ corresponding to a nonzero λ that

$$\int \psi(y)\,dF(y)=0.$$

For such a ψ in (8.3) the y-independent terms in $h(x,y)$ integrate to 0 and thus (8.3) simplifies to

$$\int (E|y-X|-|x-y|)\psi(y)\,dF(y)=\lambda\psi(x).$$

In the one dimensional case, if we differentiate with respect to x we get

$$-\int_a^b \operatorname{sign}(x-y)\psi(y)\,dF(y)=\lambda\psi'(x),$$

where (a,b) is the support of dF. (Note that a,b can be infinite). Now letting $x\to a$,

$$\lambda\psi'(a)=-\int_a^b \operatorname{sign}(a-y)\psi(y)\,dF(y)=\int_a^b \psi(y)\,dF(y)=0.$$

We get a similar equation for b. Therefore the boundary conditions are

$$\psi'(a)=\psi'(b)=0,$$

and this also holds for $\psi_0=1$. One more differentiation leads to

$$-2f\psi=\lambda\psi'',$$

where $f=F'$. This means that for $f\neq 0$ and for $\mu:=1/\lambda$ we have the simple eigenvalue equation

$$-\frac{1}{2f}\psi''=\mu\psi. \tag{8.4}$$

Now let us recall the time independent (stationary) Schrödinger (1926) equation of quantum physics:

$$-\frac{\psi''(x)}{2m}+V(x)\psi(x)=\mathcal{E}\psi(x).$$

Here ψ is the standing wave function, m is the mass of a particle, $V(x)$ is the potential function, and $\mathcal{E}$ denotes the energy level. The left hand side of (8.4) corresponds to pure kinetic energy because the $V(x)\psi(x)$ term is missing in (8.4). We can thus call μ in (8.4) the *statistical kinetic energy level*.

We have just proved that in one dimension the statistical potential energy level λ is the exact reciprocal of the statistical kinetic energy level μ.

The derivation of this nice property relies on the fact that $(1/2)|x-y|$ is the fundamental solution of the one-dimensional Laplace equation

$$\frac{d^2}{dx^2}\frac{1}{2}|x-y| = -\delta(x-y),$$

where $\delta(\cdot)$ is the Dirac delta function.

In dimension d, the fundamental solution of the d-dimensional Laplace equation, i.e. the solution of $\Delta\phi(x) = -\delta(x-y)$ where Δ is the Laplace operator, ϕ is a scalar function on R^d and δ is the Dirac delta, is $(4\pi)^{-1}|x-y|_d^{2-d}$ for $d>2$, and $(2\pi)^{-1}\log|x-y|_d$ for $d=2$. Therefore the relationship between the statistical potential energy levels and the kinetic energy level is more complex if $d\geq 2$.

We can also compute the statistical potential energy levels by applying Proposition 1. If the characteristic function of F is $\hat{f}$ and the empirical characteristic function of F is $\hat{f}_n$, then provided that the variance of F exists, we have that under the null $\sqrt{n}(\hat{f}_n(t)-\hat{f}(t))$ tends to a (complex) Gaussian process with zero expected value and covariance function $\hat{f}(s-t)-\hat{f}(s)\overline{\hat{f}(t)}$.

Now from the Karhunen–Loève expansion for Gaussian processes, we have the following equation for the eigenvalues λ:

$$\frac{1}{c_d}\int_{R^d}\frac{\hat{f}(s-t)-\hat{f}(s)\overline{\hat{f}(t)}}{|s|_d^{(d+1)/2}|t|_d^{(d+1)/2}}\Psi(s)\,ds = \lambda\Psi(t),$$

where

$$c_d = C(d,1) = \frac{\pi^{(d+1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}.$$

It is known that this eigenvalue equation has a countable spectrum; that is, we have a discrete set of solutions $\{\lambda_k : k = 1,2,\ldots\}$.

For example, if $\hat{f}$ is the standard normal characteristic function, then

$$\hat{f}(s-t)-\hat{f}(s)\overline{\hat{f}(t)} = e^{-(s^2+t^2)/2}[e^{st}-1],$$

thus

$$\frac{1}{c_d}\int_{R^d}\frac{e^{-(s^2+t^2)/2}[e^{st}-1]}{|s|_d^{(d+1)/2}|t|_d^{(d+1)/2}}\Psi(s)\,ds = \lambda\Psi(t),$$

where st, s^2, and t^2 denote inner products if $d>1$.

Typically we do not know the parameters of the normal distribution but we can standardize the sample using the sample mean and sample standard deviation. If the characteristic function of the standardized sample is $\hat{g}_n(t)$ and $\hat{f}(t)$ denotes the standard normal characteristic function, then it is still true that $\sqrt{n}(\hat{g}_n(t)-\hat{f}(t))$ tends to a (complex) Gaussian process with zero expected value, but one can show that the covariance function now changes to

$$e^{-(s^2+t^2)/2}(e^{st}-1-st-(st)^2/2),$$

and the eigenvalue equation becomes

$$\frac{1}{c_d}\int_{R^d}\frac{e^{-(s^2+t^2)/2}(e^{st}-1-st-(st)^2/2)}{|s|_d^{(d+1)/2}|t|_d^{(d+1)/2}}\Psi(s)\,ds = \lambda\Psi(t). \tag{8.5}$$

After discretization of the integral one can compute approximate values of the first few eigenvalues in (8.5). For $d=1$ the approximate eigenvalues in decreasing order are: 0.11311, 0.08357, 0.03911, 0.03182, 0.01990, 0.01698, 0.01207, 0.01060, 0.00810, …(computation by V.N. Rokhlin). See e.g. Rokhlin (1983) or Atkinson (1997, Chapter 4) on numerical solutions of equations of this type.

If only the mean is estimated then we just need to delete the term $(st)^2/2$ in (8.5). For more details see Móri and Székely (2011).

9. Historical background

The notion of $\mathcal{E}$-statistic or energy statistic was introduced at least as early as the mid-1980s, in several lectures given in Budapest, Hungary, in the Soviet Union, and at MIT, Yale, and Columbia (for lecture notes and Technical Reports see Székely, 1989, 2003). This topic was also the central topic of the first author's NSA Grant "Singular Kernel Nonparametric Tests" submitted in 2000 (NSA Grant # MDA 904-02-1-0091).

The prehistory of Proposition 2 goes back to Gel'fand and Shilov, who showed that in the world of generalized functions, the Fourier transform of a power of a Euclidean distance is also a (constant multiple of a) power of the same Euclidean distance (see Eqs. (12) and (19) for the Fourier transform of $|x|^\alpha$, Gel'fand and Shilov, 1964, pp. 173–174). Thus, one can extend the validity of Lemma 1 using generalized functions, but the Proposition itself is not in Gel'fand and Shilov (1964). The

duality between powers of distances and their Fourier transforms (characteristic functions) is similar to the duality between probability density functions of random variables and their characteristic functions (especially of normal distributions whose probability density functions have the same form as their characteristic functions). This duality was called by Gauss a "beautiful theorem of probability theory" ("Schönes Theorem der Wahrscheinlichkeitrechnung", Hans, 2011, p. 46). The proof of Propositions 1 and 2 in the univariate case, appeared as early as 1989 (Székely, 1989).

An important special case of Proposition 1, namely $E|X+X'|\geq E|X-X'|$ for all real valued X and X' with finite expectations, was a college level contest problem in Hungary in 1990 (Székely, 1996, p. 458). Russian mathematicians also published proofs of these propositions and their generalizations to metric spaces, and for more general functions than Euclidean distances; see e.g. Klebanov (2005). See also Mattner (1997) and Morgenstern (2001) of the Austrian-German school. It seems that by now many versions of the energy distance of Definition 1 and Proposition 1 and its proofs have become international folklore. On applications see the test of bivariate independence of Feuerverger (1993), and the test of homogeneity of Baringhaus and Franz (2004). See also historical comments on "Hoeffding-type inequalities" and their generalizations in Gneiting and Raftery (2007, Section 5.2).

In the 2000s many applications of energy statistics have appeared in the statistics literature as well as the literature of many other disciplines. The applications and extensions discussed in this paper represent only a subset of recent work that has appeared.

Acknowledgments

The first named co-author is thankful for much helpful advice over the past 25 years from Nail K. Bakirov, Peter Bickel, Lev Klebanov, Tamás F. Móri, Michael Newton, Vladimir N. Rokhlin, and Victor Roytburd. The authors would like to acknowledge the editor and two referees for their very careful reading and many suggestions that have greatly improved the article.

Appendix A. Proof of Lemma 1

First let us see the proof for $\alpha=1$. Apply an orthogonal transformation $t\mapsto z=(z_1,z_2,\ldots,z_d)$ with $z_1=(t,x)/|x|$ followed by a change of variables: $s=|x|\cdot z$ to get

$$\int_{R^d}\frac{1-\cos(z_1|x|)}{|z|^{d+1}}\,dz=|x|\int_{R^d}\frac{1-\cos s_1}{|s|^{d+1}}\,ds,$$

where $s=(s_1,s_2,\ldots,s_d)$. Then for $s=(s_1,s_2,\ldots,s_d)$,

$$c_d:=C(d,1)=\int_{R^d}\frac{1-\cos s_1}{|s|^{d+1}}\,ds=\frac{\pi^{(d+1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}.$$

For $d=1$ see Prudnikov et al. (1986, p. 442). Note that $2c_d=\omega_{d+1}$ is the area of the unit sphere in R^{d+1}. In the general case when both d and α can differ from 1, more technical steps are needed. Applying formulas 3.3.2.1, p. 585, 2.2.4.24 p. 298 and 2.5.3.13 p. 387 of Prudnikov et al. (1986), we obtain

$$A:=\int_{R^{d-1}}\frac{dz_2dz_3\ldots dz_d}{(1+z_2^2+z_3^2+\cdots+z_d^2)^{(d+\alpha)/2}}=\frac{2\pi^{(d-1)/2}}{\Gamma\left(\frac{d-1}{2}\right)}\int_0^\infty\frac{x^{d-2}\,dx}{(1+x^2)^{(d+\alpha)/2}}=\frac{\pi^{(d-1)/2}\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{d+\alpha}{2}\right)};$$

$$\frac{d}{da}\left(\int_0^\infty\frac{1-\cos au}{u^{1+\alpha}}\,du\right)=a^{\alpha-1}\int_0^\infty\frac{\sin v}{v^\alpha}\,dv=a^{\alpha-1}\frac{\sqrt{\pi}\Gamma\left(1-\frac{\alpha}{2}\right)}{2^\alpha\Gamma\left(\frac{\alpha+1}{2}\right)}.$$

Introduce new variables $s_1:=z_1$, and $s_k:=s_1z_k$ for $k=2,\ldots,d$. Then

$$C(d,\alpha)=A\times\int_{-\infty}^\infty\frac{1-\cos z_1}{|z_1|^{1+\alpha}}\,dz_1=\frac{\pi^{(d-1)/2}\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{d+\alpha}{2}\right)}\times\frac{2\sqrt{\pi}\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha2^\alpha\Gamma\left(\frac{\alpha+1}{2}\right)}=\frac{2\pi^{d/2}\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha2^\alpha\Gamma\left(\frac{d+\alpha}{2}\right)},$$

and this was to be proved. □

References

Aaronson, J., Burton, R., Dehling, H., Gilat, D., Hill, T., Weiss, B., 1996. Strong laws for L- and U-statistics. Transactions of the American Mathematical Society 348, 2845–2865.

Atkinson, K.E., 1997. The Numerical Solution of Integral of the Second Kind. Cambridge University Press ISBN 0-521-58391-8.
Baringhaus, L., Franz, C., 2004. On a new multivariate two-sample test. Journal of Multivariate Analysis 88, 190–206.
Baringhaus, L., Franz, C., 2010. Rigid motion invariant two-sample tests. Statistica Sinica 20, 1333–1361.
Berg, C., 2008. Stieltjes–Pick–Bernstein–Schoenberg and their connection to complete monotonicity. In: Mateu, J., Porcu, E. (Eds.), Positive Definite Functions: From Schoenberg to Space-Time Challenges. Department of Mathematics, University Jaume I, Castellon, Spain.
Berg, C., Christensen, J.P.R., Ressel, P., 1984. Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions. Graduate Texts in Mathematics, vol. 100. Springer-Verlag, Berlin, Heidelberg, New York.
Buja, A., Logan, B.F., Reeds, J.A., Shepp, L.A., 1994. Inequalities and positive definite functions arising from a problem in multidimensional scaling. Annals of Statistics 22 (1), 406–438.
Cramér, H., 1928. On the composition of elementary errors: II. Statistical applications. Skandinavisk Aktuarietidskrift 11, 141–180.
DasGupta, A., 2008. Asymptotic Theory of Statistics and Probability. Springer, New York.
Dueck, J., Edelmann, D., Gneiting, T., Richards, D. The affinely invariant distance correlation, submitted for publication. ⟨http://arxiv.org/abs/1210.2482⟩.
Feuerverger, A., 1993. A consistent test for bivariate dependence. International Statistical Review 61, 419–433.
Fischer, Hans, 2011. A History of the Central Limit Theorem: From Classical to Modern Probability Theory. Springer.
Gel'fand, I.M., Shilov, G.E., 1964. Generalized Functions, Volume I: Properties and Operations. Academic Press, New York (translation by E. Salatan of the Russian edition of 1958).
Genovese, C., 2009. Discussion of: Brownian distance covariance. Annals of Applied Statistics 3 (4), 1299–1302, http://dx.doi.org/10.1214/09-AOAS312G.
Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.
Gretton, A., Györfi, L., 2010. Consistent nonparametric tests of independence. Journal of Machine Learning Research 11, 1391–1423.
Gretton, A., Györfi, L., 2012. Strongly consistent nonparametric test of conditional independence. Statistics and Probability Letters 82, 1145–1150.
Gurtler, N., Henze, N., 2000. Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. Annals of the Institute of Statistical Mathematics 52 (2), 267–286.
Henze, N., Zirkler, B., 1990. A class of invariant and consistent tests for multivariate normality. Communications in Statistics—Theory and Methods 19, 3595–3617.
Herbin, E., Merzbach, E., 2007. The multiparameter fractional Brownian motion. In: Math Everywhere. Springer, Berlin, pp. 93–101.
Hoeffding, W., 1948. A class of statistics with asymptotic normal distribution. Annals of Mathematic Statistics 19 (3), 293–325.
Kim, A.Y., Marzban, C., Percival, D.B., Stuetzle, W., 2009. Using labeled data to evaluate change detectors in a multivariate streaming environment. signal Processing 89 (12), 2529–2536. ISSN 0165-1684.
Klebanov, L., 2005. $\mathcal{N}$-Distances and Their Applications. Charles University, Prague.
Kong, J., Klein, B.E.K., Klein, R., Lee, K., Wahba, G., 2012. Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. Proceedings of the National Academy of Sciences 109 (50), 20352–20357, http://dx.doi.org/10.1073/pnas.1217269109.
Koroljuk, V.S., Borovskich, Yu.V., 1994. Theory of U-statistics, Mathematics and its Applications, vol. 273. Kluwer Academic Publishers Group, Dordrecht (Translated by P.V. Malyshev, D.V. Malyshev from the 1989 Russian original ed.).
Lehmann, E.L., 1966. Some concepts of dependence. Annals of Mathematical Statistics 37, 1137–1153.
Leucht, A., Neumann, M.H., 2013. Degenerate *U*- and *V*-statistics under ergodicity: asymptotics, bootstrap and applications in statistics. Annals of the Institute of Statistical Mathematics 65 (2), 349–386.
Li, R., Zhong, W., Zhu, L., 2012. Feature screening via distance correlation learning. Journal of the American Statistical Association 107 (499), 1129–1139, http://dx.doi.org/10.1080/01621459.2012.695654.
Lyons, R. Distance covariance in metric spaces, Annals of Probability, to appear. ⟨http://mypage.iu.edu/~rdlyons/pdf/dcov.pdf⟩.
Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–530.
Matsui, M., Takemura, A., 2005. Empirical characteristic function approach to goodness-of-fit tests for the Cauchy distribution with parameters estimated by MLE or EISE. Annals of the Institute of Statistical Mathematics 57 (1), 183–199.
Matteson, D.S., Tsay, R.S., 2011. Independent Component Analysis via Distance Covariance. ⟨http://www.stat.cornell.edu/matteson/papers/MattesonTsay_dCovICA.pdf⟩, preprint.
Matteson, D.S., James, N.A., 2012. A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data ⟨http://www.stat.cornell.edu/~matteson/papers/MattesonJames2012a.pdf⟩, preprint.
Mattner, L., 1997. Strict negative definiteness of integrals via complete monotonicity of derivatives. Transactions of the American Mathematical Society 349 (8), 3321–3342.
Mensheinin, D.O., Zubkov, A.M., 2012. Properties of the Székely–Móri symmetry criterion statistics in the case of binary vectors. Mathematical Notes 91 (4), 62–72.
Móri, T.F., Székely, G.J., 2011. On the covariance function of Gaussian processes, preprint.
Morgenstern, D., 2001. Proof of a conjecture by Walter Deuber concerning the distance between points of two types in R^d. Discrete Mathematics 226, 347–349.
Newton, M.A., 2009. Introducing the discussion paper by Székely and Rizzo. Annals of Applied Statistics 3 (4), 1233–1235, http://dx.doi.org/10.1214/09-AOAS34INTRO.
Prudnikov, A.P., Brychkov, A., Marichev, O.I., 1986. Integrals and Series. Gordon and Breach Science Publishers, New York.
R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, ⟨http://www.R-project.org⟩.
Rémillard, B., 2009. Discussion of: 'Brownian distance covariance'. Annals of Applied Statistics 3 (4), 1295–1298, http://dx.doi.org/10.1214/09-AOAS312F.
Rizzo, M.L., 2002. A New Rotation Invariant Goodness-of-Fit Test. Ph.D. Dissertation. Bowling Green State University.
Rizzo, M.L., 2003. A test of homogeneity for two multivariate populations. In: 2002 Proceedings of the American Statistical Association, Physical and Engineering Sciences Section. American Statistical Association, Alexandria, VA.
Rizzo, M.L., 2009. New goodness-of-Fit tests for Pareto distributions. ASTIN Bulletin: Journal of the International Association of Actuaries 39 (2), 691–715.
Rizzo, M.L., Székely, G.J., 2010. DISCO analysis: a nonparametric extension of analysis of variance. Annals of Applied Statistics 4 (2), 1034–1055.
Rizzo, M.L., Székely, G.J., 2011. Energy: E-statistics (Energy Statistics). R package version 1.4-0.
Rokhlin, V., 1983. Rapid solution of integral equations of classical potential theory. Journal of Computational Physics 60, 187–207.
Schilling, K., Oberdick, J., Schilling, R.L., 2012. Toward an efficient and integrative analysis of limited-choice behavioral experiments. Journal of Neuroscience 32 (37), 12651–12656.
Schrödinger, E., 1926. An undulatory theory of the mechanics of atoms and molecules. Physical Review 28 (6), 1049–1070.
Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.
Székely, G.J., 1989. Potential and Kinetic Energy in Statistics. Lecture Notes. Budapest Institute of Technology (Technical University).
Székely, G.J., 1996. Contests in Higher Mathematics. Springer, New York.
Székely, G.J., 2003. E-statistics: Energy of Statistical Samples. Bowling Green State University, Department of Mathematics and Statistics Technical Report No. 03-05 (also technical reports by same title, from 2000–2003 and NSA Grant # MDA 904-02-1-0091 (2000–2002).).
Székely, G.J., Bakirov, N.K., 2003. Extremal probabilities for Gaussian quadratic forms. Probability Theory and Related Fields 126, 184–202.
Székely, G.J., Móri, T.F., 2001. A characteristic measure of asymmetry and its application for testing diagonal symmetry. Communications in Statistics—Theory and Methods 30 (8&9), 1633–1639.
Székely, G.J., Rizzo, M.L., 2004. Testing for Equal Distributions in High Dimension. InterStat, November (5).
Székely, G.J., Rizzo, M.L., 2005a. A new test for multivariate normality. Journal of Multivariate Analysis 93 (1), 58–80.

Székely, G.J., Rizzo, M.L., 2005b. Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method. Journal of Classification 22 (2), 151–183.
Székely, G.J., Rizzo, M.L., 2009. Brownian distance covariance. Annals of Applied Statistics 3 (4), 1236–1265.
Székely, G.J., Rizzo, M.L., 2012. On the uniqueness of distance covariance. Statistics & Probability Letters 82, 2278–2282.
Székely, G.J., Rizzo, M.L., 2013. The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis 117, 193–213.
Székely, G.J., Rizzo, M.L., Bakirov, N.K., 2007. Measuring and testing independence by correlation of distances. Annals of Statistics 35 (6), 2769–2794.
Von Mises, R., 1947. On the asymptotic distributions of differentiable statistical functionals. Annals of Mathematical Statistics 2, 209–348.
Yang, G., 2012. The Energy Goodness-of-Fit Test for Univariate Stable Distributions. Ph.D. Thesis. Bowling Green State University.
Yitzhaki, S., 2003. Gini's Mean difference: a superior measure of variability for non-normal distributions. Metron 61, 285–316.
Zacks, S., 1981. Parametric Statistical Inference: Basic Theory and Modern Approaches. Pergamon, Oxford.
Zhou, Z., 2012. Measuring nonlinear dependence in time-series, a distance correlation approach. Journal of Time Series Analysis 33 (3), 438–457.