The Energy of Data

Gábor J. Székely[1,2] and Maria L. Rizzo[3]

[1]National Science Foundation, Arlington, Virginia 22230; email: gszekely@nsf.gov

[2]Rényi Institute of Mathematics, Hungarian Academy of Sciences, 1053 Budapest, Hungary

[3]Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio 43403; email: mrizzo@bgsu.edu

Annu. Rev. Stat. Appl. 2017. 4:447–79

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-060116-054026

Keywords

energy distance, statistical energy, goodness-of-fit, multivariate independence, distance covariance, distance correlation, *U*-statistic, *V*-statistic, data distance, graph distance, statistical tests, clustering

Abstract

The energy of data is the value of a real function of distances between data in metric spaces. The name energy derives from Newton's gravitational potential energy, which is also a function of distances between physical objects. One of the advantages of working with energy functions (energy statistics) is that even if the data are complex objects, such as functions or graphs, we can use their real-valued distances for inference. Other advantages are illustrated and discussed in this review. Concrete examples include energy testing for normality, energy clustering, and distance correlation. Applications include genome studies, brain studies, and astrophysics. The direct connection between energy and mind/observations/data in this review is a counterpart of the equivalence of energy and matter/mass in Einstein's $E = mc^2$.

1. INTRODUCTION

Energy statistics ($\mathcal{E}$-statistics) are functions of distances between statistical observations. The value of the energy statistics for a given data set is the (potential) energy of the data. 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.

Following Székely (1989), we can represent the three pairs of dualities between energy, matter, and mind with the help of a triangle (**Figure 1**) whose vertices are energy, matter, and mind, and the connecting sides represent the equivalence/duality/dichotomy between these notions. Manifestations of matter include mass (m) and disorder (measured by entropy S). Manifestations of the immaterial mind are memory, information (I), observation, data, and inputs passed on by the sensory organs.

The duality between E and m is Einstein's famous $E = mc^2$ (Einstein 1905). The duality between matter and mind is Szilárd's idea, which first appeared in his 1922 dissertation, that in a closed material system, the decrease of uncertainty/entropy (S) corresponds to the increase of information (I) in our mind (Szilárd 1929, Schrödinger 1944, Brillouin 2004). To use Szilárd's words, it is possible to reduce the entropy of a thermodynamic system by the intervention of intelligent beings, for example a "Maxwell's demon." Thus Szilárd eradicated the ancient dichotomy of matter and mind just as Einstein eradicated the dichotomy of energy and matter. This review is about the third side of the triangle, the connection between energy and mind in terms of data distance D defined below. Our mind regulates the flow of information (I) and data distance (D), the source of statistical energy, to help achieve mental harmony.

For data $\mathbf{X} = X_1, \ldots, X_n$ and $\mathbf{Y} = Y_1, \ldots, Y_n$ in Euclidean spaces, define the data distance of $\mathbf{X}$ and $\mathbf{Y}$ as

$$D := 2\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - Y_j| - \sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j| - \sum_{i=1}^{n}\sum_{j=1}^{n}|Y_i - Y_j|.$$

We will see that D is always nonnegative. Moreover, D is always the square of a metric in the space of samples of size n. The source of statistical energy is D. This can also be viewed as energy in physics if we multiply D by a constant force of magnitude F; that is, energy $E = F \cdot D$. This resembles the gravitational potential energy close to the Earth where the gravitational force is constant.

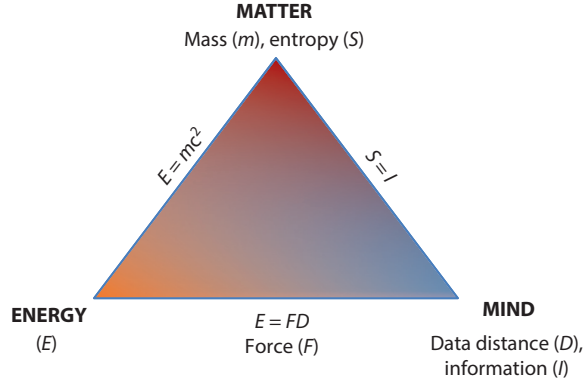

Figure 1

Connections between energy, matter, and mind.

The signs in D resemble the computation of electrostatic potential where the signs depend on the charges. For statistical inferences we do not need to know the magnitude F of the force. This would play an important role if we wanted to free the statistical energy as a counterpart of nuclear energy. If we choose $F = n^{-2}$, then we get a special case of the formula for the two-sample energy statistic $\mathcal{E}_{n_1,n_2}(\mathbf{X}, \mathbf{Y})$ in Equation 16 (Section 6.1) where the sample sizes are not necessarily equal.

For statistical purposes we can also work with powers of distances such as $|X_i - Y_m|^\alpha$ where $0 < \alpha < 2$, because in this range of exponents,

$$D_\alpha := 2\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - Y_j|^\alpha - \sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|^\alpha - \sum_{i=1}^{n}\sum_{j=1}^{n}|Y_i - Y_j|^\alpha$$

remains the square of a metric (see Section 8 on generalized energy distance). The square root of D_α is a metric and $(D_\alpha)^{1/\alpha}$ is measured in the same units as $|X_i - Y_j|$, and thus $E_\alpha := F(D_\alpha)^{1/\alpha}$ can be considered a generalized statistical energy. Because Newton's gravitational potential is proportional to the reciprocal of the distance, whereas in elasticity the energy is proportional to the second power of the extension/displacement (Hooke's law), we have that the exponents relevant for statistics are between the exponents relevant in gravitation and in elasticity. The potential energy of conservative vector fields in physics is always a harmonic function. In Section 9, we show that in the world of data energy or statistical energy, "harmonic" is replaced by "conditionally negative definite."

In what follows, X' denotes an independent copy of the random variable X, Y' is an independent copy of the random variable Y, and E denotes expected value. The (potential) energy of X with respect to Y, or vice versa, or simply the energy of (X, Y), is defined in the sidebar Energy $\mathcal{E}(X, Y)$.

In Section 9, we show theorems stating that many important metric spaces share the following properties:

1. $\mathcal{E}(X, Y) \geq 0$ and
2. $\mathcal{E}(X, Y) = 0$ if and only if X and Y are identically distributed.

A necessary and sufficient condition for property 1 is the conditional negative definiteness of δ. For more information on this property, the reader is directed to the classical papers of Schoenberg (1938a,b) and also Horn (1972), Steutel & van Harn (2004), and Berg (2008). Metric spaces with properties 1 and 2 include all Euclidean spaces, all separable Hilbert spaces, all hyperbolic spaces, and many graphs with geodesic distances. For more details, see Propositions 1–3 and Section 9. These theorems make it possible to develop clustering methods that depend not only on cluster centers but also on cluster distributions (Section 6) and to develop a simple new dependence measure, the distance correlation, that is zero if and only if the random variables are independent (Section 7). A simple example where property 1 holds but property 2 fails is the taxicab metric or Manhattan distance. By the way, $E[\delta(X, X')]$ can be viewed as the δ-energy of X or of the distribution of X. For the special case $\delta = |X - X'|^\alpha$, see Riesz (1938, 1949).

ENERGY $\mathcal{E}(X, Y)$

The (potential) energy of the independent random variables X, Y that take their values in a metric space with distance function δ is defined as

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

provided that these expectations are finite.

In this review 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 are discussed that illustrate the power of this concept in diverse applications including gene regulatory networks (Guo et al. 2014), neuroimaging (Hua et al. 2015), fMRI studies (Rudas et al. 2014), clustering chemical structures (Varina et al. 2009), experimental design (Shamsuzzaman et al. 2015), glucose monitoring (Ferenci et al. 2015), community ecology (Székely & Rizzo 2014), portfolio optimization (Würtz et al. 2009), dimension reduction (Sheng & Yin 2013, 2016), forecasting during geomagnetic storms (Lu et al. 2015), graphical models (Fan et al. 2015, Wang et al. 2015), chemometrics (Vaiciukynas et al. 2015), and musicology (Zanoni et al. 2014). We will see that energy statistics are extremely useful, and tests based on them are typically more general and often more powerful against general alternatives than corresponding tests based on classical (non-energy type) statistics such as Pearson's correlation and Fisher's F-statistics.

The notion of statistical energy was introduced in 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, 2002). This idea became the central topic of the first author's NSA grant "Singular Kernel Nonparametric Tests," submitted in 2000.

1.1. Distances of Observations—Rigid Motion Invariance

Classical statistics operated with real-valued data such as height, weight, or blood pressure, and it was supposed that typical data are (approximately) normally distributed so that one can apply the theory of normal (Gaussian) distributions for inference. Even if the observations were not Gaussian, in case of big data when the number of observations, n, was large, one could often refer to central limit theorems to claim that for large n, the normal approximation is valid and classical methods are applicable.

What happens if the data are not real numbers but vectors, functions, graphs, and so on? In this case even addition or multiplication of data might be a problem if, for example, the observed vectors have different dimensions. We can overcome this difficulty if the observations are elements of a metric space. In this case, instead of working with the observations themselves, we can work with their (nonnegative) real-valued distances. This brings us back to the real world where we can work with real numbers. We call this type of inference energy inference.

The energy approach to testing a hypothesis is built on comparing distributions for equality, in contrast to normal theory methods such as the two-sample t-test for comparing means. Energy tests can detect any difference between distributions. An attractive property of working with distances is that energy statistics have invariance with respect to any distance-preserving transformation of the full data set to be analyzed (rigid motion invariance), which includes translation, reflection, and angle-preserving rotation of coordinate axes. We have energy counterparts of the variance, covariance, and correlation, called respectively distance variance, distance covariance, and distance correlation. An important advantage of applying them is that the distance correlation coefficient equals zero if and only if the variables are independent. This characterization of independence typically does not hold for Pearson's correlation, for example, for nonnormal data.

1.2. Functions of Distances—Jackknife Invariance

Statistics are functions of data, and energy statistics are functions of distances between data. If the observations play a symmetric role, it makes sense to suppose that our statistics are symmetric functions of data. Suppose that the sample size is n, so that a statistic is an n-variate function of

the sample. What is the connection between the n-variate functions for different values of n? A natural principle is jackknife invariance, which is as follows.

Let $r \geq 1$ be an arbitrary integer and $f_n(x_1, \ldots, x_n)$ a function $R^n \to R$ of a sample $x_1, \ldots, x_n$ defined for all $n \geq r$. Let $f_{n-1}^i(x_1, \ldots, x_n)$, $i = 1, 2, \ldots, n$, denote the statistic of a jackknife sample when we remove the ith observation: $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$. [$f_0^i$ is defined as $f_1(x_1)$.] Jackknife invariance of order r means that

$$n \cdot f_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} f_{n-1}^i(x_1, \ldots, x_n)$$

holds for all $n \geq r$ and r is the smallest positive integer with this property.

It is easy to see that all U-statistics defined below are jackknife invariant of order r. For a real-valued random sample $X_1, \ldots, X_n$, and a symmetric kernel function $h : R^r \to R$, the U-statistic

$$U_n = \frac{1}{n(n-1)\cdots(n-r+1)} \sum_{i_1 < i_2 < \ldots < i_r} h(X_{i_1}, \ldots, X_{i_r})$$

has degree r if we cannot have this representation with a kernel having fewer arguments than r. A familiar example for a U-statistic of degree 2 is Gini's mean difference (Gini 1912, Yitzhaki 2003), a U-statistic for dispersion:

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

Alternately one can apply the corresponding V-statistic for dispersion:

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

Theorem 1. A necessary and sufficient condition for $f_n(X_1, \ldots, X_n)$, $n \geq r$, to be a U-statistic of degree r is the jackknife invariance of order r of $f_n(x_1, \ldots, x_n)$.

For a proof see Huo & Székely (2016).

In the following, for simplicity, we will discuss the case when $r = 2$. Energy statistics in this review are U-statistics or V-statistics based on distances. That is, for a d-dimensional random sample $X_1, \ldots, X_n$, an energy statistic is defined by a kernel function $h : R^d \times R^d \to R$, and

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

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 the 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 review.

An advantage of assuming that all energy statistics are U-statistics or V-statistics is that we can apply their classical limit theory (Von Mises 1947, Hoeffding 1948) to obtain the asymptotic distributions of these statistics (see, e.g., Serfling 1980, Koroljuk & Borovskich 1994, or Aaronson et al. 1996 for details). In some of the applications discussed below, a generalized V-statistic is

Table 1 Many methods of classical statistical inferences have their counterparts in the energy world

	CLASSICAL APPROACH	ENERGY APPROACH
Dependence	Pearson's correlation (can be 0 for dependent variables)	Distance correlation (equals 0 iff the variables are independent)
Goodness-of-fit & homogeneity	EDF tests	Energy goodness-of-fit & homogeneity characterizes equality of distributions
	Likelihood ratio tests	Nonparametric energy tests
Tests for normality	Shapiro-Wilk, Anderson-Darling (lack natural multivariate extensions)	Energy test for normality (arbitrary dimension)
Multivariate normality	Multivariate skewness & kurtosis (test is not consistent)	Multivariate energy test of normality (test is consistent)
Skewness	Third central moment (can be 0 for nonsymmetric distributions)	Distance skewness (equals 0 iff the probability distance is centrally symmetric)
Multisample problems	ANOVA (tests if normally distributed variables have equal means)	DISCO (tests if random variables have the same distribution)
Cluster analysis	Ward's minimum variance method and k-means clustering (sensitive to cluster centers)	Hierarchical energy clustering and k-groups (energy) clustering (sensitive to cluster distributions)

Abbreviations: ANOVA, analysis of variance; DISCO, distance components; EDF, empirical distribution function.

applied (Koroljuk & Borovskich 1994, chapter 11). Although energy statistics can be defined in terms of either U- or V-statistics, in our development we apply V-statistics throughout. The main reason for applying V-statistics instead of U-statistics is that an energy V-statistic is always nonnegative and thus it can more easily be interpreted as a (statistical) distance.

Many classical statistical methods have their energy counterparts, as shown in **Table 1**. The reader is also directed to Székely & Rizzo (2013a).

1.3. Energy Inference—Scale Invariance

An important restriction is necessary on energy statistics and thus on the kernel h of the corresponding U-statistics and V-statistics. If we change the measurement unit of the data then this must not change the statistical inference, for example, in case of testing for normality or testing for independence.

The kernel h is called scale equivariant if, for all real numbers $a > 0$ and for some real function $g(a) \neq 0$, we have

$$h(ax_1, ax_2) = g(a)h(x_1, x_2).$$

The ratio of two U-statistics or V-statistics with scale equivariant kernels and the same function $g(.)$ is clearly scale invariant. For example, the kernel $h(x_1, x_2) = |x_1 - x_2|^\alpha$ is scale equivariant with $g(a) = a^\alpha$.

The special case $\alpha = 1$ is the classical geometric similarity in Euclidean spaces. Roughly, they can be described as transformations preserving shapes, but changing scales: magnifying ($a > 1$) or contracting ($a < 1$). Ratios of distances do not change when the transformation is similarity: when we multiply the distances by the same number a, their ratios and the corresponding angles remain invariant. This was already known to Thales around 600 BC. He understood similar triangles and used that knowledge in practical ways, for example, when he measured the height of the pyramids by their shadows at the moment when his own shadow was equal to his height. In geometry, Thales's unifying notion was the angle (*gonia*). With the help of angles he introduced a new geometry, which

is based on similarity, not on coincidence. In this review we exploit Thales's simple but brilliant idea in many ways when applying it for statistical inferences. An example can be found in Section 7.2.

Typical inference problems that we will work with satisfy our three basic invariances:

1. Rigid motion invariance
2. Jackknife invariance
3. Scale invariance

Our main tool is the rigid motion invariant and scale equivariant energy distance introduced in the next section.

2. ENERGY DISTANCE

Many types of distances can be defined between statistical objects. The L_2 distance is one well-known and widely applied distance. Let F be the cumulative distribution function (CDF) of a random variable, and let F_n denote the empirical CDF of a sample of size n. The L_2 distance between F and F_n

$$\int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, \mathrm{d}x \tag{1}$$

was introduced in Cramér (1928). However, Cramér's distance is not distribution-free; thus, to apply this distance for testing goodness of fit, the critical values must depend on F. This disadvantage was easily rectified by replacing $\mathrm{d}x$ in Cramér's distance by $\mathrm{d}F(x)$, to obtain the Cramér–von Mises–Smirnov distance defined

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

There remains, however, another important disadvantage of both the Cramér L_2 distance (Cramér 1928) and the Cramér–von Mises–Smirnov distance. For samples from a d-dimensional space where $d > 1$, neither distance is rotation invariant. For important problems such as testing for multivariate normality, we require rotation invariance. Below we discuss how to overcome this difficulty.

Let X and Y be independent real-valued random variables with CDFs F and G, respectively, let X' denote an independent and identically distributed (i.i.d.) copy of X, and Y' denote an i.i.d. copy of Y. Then (see e.g., Székely 1989, 2002),

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

A rotation-invariant natural extension of the above identity for higher dimensions $d \geq 1$ is

$$2E|X - Y|_d - E|X - X'|_d - E|Y - Y'|_d, \tag{3}$$

where $X, Y \in R^d$ are independent. Proof of the rotational invariance of this expression is straightforward, but it is not trivial at all that the expression in Equation 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

$$\mathcal{E}(X, Y) = 2E|X - Y|_d - E|X - X'|_d - E|Y - Y'|_d, \tag{4}$$

where $E|X|_d < \infty$, $E|Y|_d < \infty$, X' is an i.i.d. copy of X, and Y' is an i.i.d. 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,

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

The Fourier transform of the CDF $F(x) = \int_{-\infty}^{x} f(u) \, \mathrm{d}u$ is $\hat{f}(t)/(\imath t)$, where $\imath = \sqrt{-1}$, thus we have

$$2\pi \int_{-\infty}^{\infty} (F(x) - G(x))^2 \, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{|\hat{f}(t) - \hat{g}(t)|^2}{t^2} \, \mathrm{d}t. \tag{5}$$

The pleasant surprise is that the natural multivariate generalization of the right-hand side of Equation 5 is rotation invariant and it is exactly a constant multiple of Equation 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

$$2E|X - Y|_d - E|X - X'|_d - E|Y - Y'|_d = \frac{1}{c_d} \int_{R^d} \frac{|\hat{f}(t) - \hat{g}(t)|^2}{|t|_d^{d+1}} \, \mathrm{d}t, \tag{6}$$

where

$$c_d = \frac{\pi^{(d+1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}, \tag{7}$$

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 & Rizzo (2005a). For its history, see Section 10. For a generalization to metric spaces, see Proposition 3. In view of Equation 6, the square root of energy distance $\mathcal{E}(X, Y)^{1/2}$ is a metric on the set of d-variate distribution functions.

In probabilistic terms, Gini's mean difference is $E|X - X'|$, where X' is an i.i.d. copy of X. Thus Gini's mean difference is an L_1 distance $E|X - Y|$ between X and $Y = X'$. It is clear that $E|X - Y| = 0$ if and only if $Y = X$ with probability 1. Energy distance is a double-centered L_1 distance, which is zero if and only if X and Y are identically distributed. It is this dramatic difference that makes it so powerful for statistical applications. These applications include, for example,

1. Consistent one-sample goodness-of-fit tests (Székely & Rizzo 2005a, Rizzo 2009, Yang 2012)
2. Consistent multisample tests of equality of distributions (Rizzo 2002, 2003; Székely & Rizzo 2004; Baringhaus & Franz 2004)
3. Hierarchical clustering algorithms (Székely & Rizzo 2005b) that extend and generalize the Ward's minimum variance algorithm
4. Distance components (DISCO) (Rizzo & Székely 2010), a nonparametric extension of analysis of variance for structured data
5. Characterization and test for multivariate independence (Feuerverger 1993, Székely et al. 2007, Székely & Rizzo 2009)
6. Change point analysis based on Székely & Rizzo (2004) (Kim et al. 2009, Matteson & James 2013)
7. Uplift modeling (Jaroszewitz & Lukasz 2015)

Several of these applications are discussed below. Software for energy statistics applications is available under General Public License in the `energy` package for R (Rizzo & Székely 2016, R Core Team 2016).

3. WHY IS ENERGY DISTANCE SPECIAL?

We see that energy distance (Equation 6) is a weighted L_2 distance between characteristic functions, with weight function $w(t) = \text{const}/|t|^{d+1}$. Suppose that the following three technical conditions on the weight function hold: $w(t) > 0$, $w(t)$ is continuous, and

$$\int |\hat{f}(t) - \hat{g}(t)|^2 w(t)\, \mathrm{d}t < \infty. \tag{8}$$

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 shown in Equation 8 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 & Rizzo 2012).

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

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

If we introduce $s = at$, we can see that if $a \neq 0$ then

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

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., Henze & Zirkler 1990, Gurtler & Henze 2000, or Matsui & Takemura 2005).

If in the definition of equivariance above we replace $|a|$ by $|a|^\alpha$ where $0 < \alpha < 2$, then we get a more general weight function, which is $w(t) = \text{const}/|t|^{d+\alpha}$. Ratios of the corresponding statistics remain scale invariant (see Section 8).

4. ONE SAMPLE ENERGY STATISTICS

4.1. Energy Goodness-of-Fit

A one-sample goodness-of-fit statistic is designed to measure the distance between a hypothesized distribution F_0 and the distribution F from which an observed random sample $x_1, \ldots, x_n$ is drawn. The energy distance for the goodness-of-fit test $H_0 : F = F_0$ versus $H_1 : F \neq F_0$ is

$$\mathcal{E}_n(\mathbf{X}, F_0) = \frac{2}{n} \sum_{i=1}^{n} E|x_i - X| - E|X - X'| - \frac{1}{n^2} \sum_{\ell=1}^{n} \sum_{m=1}^{n} |x_\ell - x_m|, \tag{9}$$

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

$\mathcal{E}_n$ is a V-statistic, and the corresponding unbiased statistics are U-statistics. The kernel function for the energy goodness-of-fit V- or U-statistic is shown in Equation 12, discussed in Section 5.

Under the null hypothesis $H_0 : F = F_0$, the test statistic $n\mathcal{E}_n$ has a nondegenerate asymptotic distribution as $n \to \infty$ (see Section 5). Under an alternative hypothesis $H_1 : F \neq F_0$, the statistic $n\mathcal{E}_n$ tends to infinity stochastically. Hence, the energy goodness-of-fit test that rejects the null hypothesis for large values of $n\mathcal{E}_n$ is consistent against general alternatives.

Energy goodness-of-fit tests based on Equation 9 have been implemented for testing the composite hypothesis of multivariate normality (Székely & Rizzo 2005a), Pareto family (Rizzo 2009), stable (Yang 2012), and other distributions.

Let us begin by illustrating the application of energy goodness-of-fit tests with a few univariate examples.

4.1.1. Energy statistic for continuous uniform distribution. If a random variable has the continuous uniform distribution over an interval (a, b), then for any fixed x, we have

$$E|x - X| = \frac{(x-a)^2}{b-a} - x + \frac{b-a}{2}$$

and $E|X - X'| = \frac{b-a}{3}$. The energy test statistic for a goodness-of-fit test of standard uniform distribution, that is, $H_0 : X \sim \text{Uniform}(0,1)$, is therefore 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 kth order statistic of the sample.

Note that the linearization in the last sum,

$$\frac{1}{n^2}\sum_{i,j=1}^{n}|X_i - X_j| = \frac{2}{n^2}\sum_{k=1}^{n}(2k - 1 - n)X_{(k)},$$

simplifies the energy statistic for any univariate sample, reducing the computational complexity to $O(n \log n)$ in the one-dimensional case.

4.1.2. Two-parameter exponential distribution. Suppose that one wishes to test whether a variable T has a two-parameter exponential distribution, with density

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

where $\lambda > 0$ and $\mu \in \mathbb{R}$ are constants. It is easy to verify that

$$E|t - T| = t - \mu + \frac{1}{\lambda}(1 - 2F_T(t)), \qquad t \geq \mu;$$
$$E|T - T'| = \frac{1}{\lambda}.$$

These expectations immediately determine a computing formula for the corresponding test statistic (Equation 9) and $n\mathcal{E}_n$.

4.2. Energy Test of Normality

The energy test for multivariate normality was developed by Székely & Rizzo (2005a). Let $x_1, \ldots, x_n$ denote an observed random sample.

4.2.1. Univariate normality. For testing univariate normality, the energy test statistic is $n\mathcal{E}_n$, where $\mathcal{E}_n$ is given by Equation 9 with

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

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

4.2.2. Relation to quadratic empirical distribution function statistics. The class of quadratic empirical distribution function (EDF) statistics are based on weighted L_2 distances of the type discussed above, essentially a weighted Cramér distance

$$\int_{-\infty}^{\infty} (F_n(x) - F(x))^2 w(x)\, dF(x), \tag{10}$$

where $w(\cdot)$ is a suitable weight function. In the Cramér–von Mises test, $w(x)$ is the identity function, and the Anderson-Darling test applies a weight function $w(x) = [F(x)(1 - F(x))]^{-1}$. In the case of testing for standard normal distribution, the shape of the curve $w(x) = F(x)(1 - F(x))$ is in fact similar to the shape of the standard normal density, and their ratio is close to a constant c (empirically $c \approxeq 0.67$). Thus, although the energy test of standard normal distribution and the Anderson-Darling test are different statistics and different tests, the tests have quite similar performance. In this sense, the multivariate energy test of normality can also be viewed as a computationally simple way to lift the univariate Anderson-Darling test of normality to arbitrarily high dimension.

4.3. Energy Test of Multivariate Normality

The energy test of multivariate normality is rigid motion invariant and consistent. When we apply the test to standardized samples, it is affine invariant.

First we develop the test for the case when the parameters are specified. In that case, the data can be transformed to standard multivariate normal $N_d(0, I_d)$, where I_d is the $d \times d$ identity matrix. For standard multivariate normal $Z \in R^d$ with mean vector 0 and identity covariance matrix, we obtain

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

Then, if $y_1, \ldots, y_n$ denote the standardized sample elements, the energy test statistic for standard d-variate normal distribution is

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

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 reader is directed to Székely & Rizzo (2005a) for details. One way to obtain the expression for $E|a - Z|_d$ follows from a result (Zacks 1981, p. 55) stating that if $Z \in R^d$ is standard normal, then the variable $|a - Z|_d^2$ has a noncentral chi-squared 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. For implementation, the sum in $E|a - Z|_d$ converges after 40–60 terms except for very large $|a|_d$, and if $|a|_d$ 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 & Székely 2016) for an implementation.

For the composite hypothesis of normality, where the mean vector μ and covariance matrix Σ are not specified, the test is modified as follows. Apply a linear transformation of the observed sample to standardize it, using the sample mean vector and the sample covariance matrix. The modified test statistic $n\hat{\mathcal{E}}_{n,d}$ has the same type of limit distribution as $n\mathcal{E}_{n,d}$, with the rejection region in the upper tail, but with different critical values. In both cases a Monte Carlo test implementation can be applied. The statistic $n\hat{\mathcal{E}}_{n,d}$ is compared with replicates of the energy statistic for standardized normal samples of equal dimension. This is the method of implementation in function `mvnorm.etest` in the `energy` package for R. A theory for the case of estimated parameters is derived by Rizzo (2002) and Székely & Rizzo (2005a).

The energy test of multivariate normality is applicable for arbitrary dimension (not constrained by sample size) and is practical to apply. Monte Carlo power comparisons in the references suggest that energy is a powerful competitor to other tests of multivariate normality. Several examples appear in Rizzo (2002) and Székely & Rizzo (2005a) to illustrate the power of the test of multivariate normality against various alternatives, compared with competing tests such as Mardia's test and the Henze-Zirkler test (Mardia 1970, Henze & Zirkler 1990). 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.

4.4. Testing for Spherical Symmetry

Spherical symmetry around the origin (around 0) holds if the projection of the variable X onto the surface of the unit sphere is uniform. That is, for $X \neq 0$, $X/|X|$ is uniformly distributed on the surface of the unit sphere. With the following result, we can test for uniform distribution on the sphere.

Suppose that we want to test if a random sample $\{x_1, \ldots, x_n\}$ of points on the surface of a d-dimensional unit sphere (whose radius is 1) is from a uniform distribution on the d-dimensional unit sphere. Under the null hypothesis, the rotational invariance implies that $E|X - x| = E|X - X'|$ for every x on the unit sphere. If $d = 2$ (unit circle) then $E|X - X'| = 4/\pi$, thus the energy goodness-of-fit test statistic (Equation 9) is

$$n\mathcal{E}_n = \frac{4n}{\pi} - \frac{1}{n}\sum_{\ell=1}^{n}\sum_{m=1}^{n}|x_\ell - x_m|.$$

For $d = 3$ we have $E|X - X'| = 4/3$. For arbitrary dimension d, the energy test statistic for uniform distribution of points on the d-dimensional unit sphere is

$$n\mathcal{E}_n = nE|X - X'| - \frac{1}{n}\sum_{\ell=1}^{n}\sum_{m=1}^{n}|x_\ell - x_m|,$$

where

$$E|X - X'| = \frac{\left(\Gamma\left(\frac{d}{2}\right)\right)^2 2^{d-1}}{\Gamma\left(d - \frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)}. \tag{11}$$

The sequence (Equation 11) is monotone increasing with limit $\sqrt{2}$. Because energy is always nonnegative, we have that $E|X - X'|$ is maximum if (and, in fact, only if) X and X' have uniform distributions on the sphere. If the center of symmetry is unknown, then apply the same test for the

centered sample when the sample mean is subtracted from all sample elements and then projected onto the surface of the unit sphere.

5. KINETIC ENERGY: THE SCHRÖDINGER EQUATION OF GOODNESS-OF-FIT

The one-sample energy statistic $\mathcal{E}_n$ in Equation 9 is a V-statistic whose kernel is a double-centered potential function defined by

$$h(x, y) = E|x - X| + E|y - X| - E|X - X'| - |x - y|, \tag{12}$$

where $h : R^d \times R^d \to R$. By the law of large numbers for V-statistics (see, e.g., Serfling 1980 or Koroljuk & 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 $n \to \infty$.

Alternatively, 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 & 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

$$Q = \sum_{k=1}^{\infty} \lambda_k Z_k^2 \tag{13}$$

of i.i.d. standard normal random variables Z_k, $k = 1, 2, \ldots$ (Koroljuk & Borovskich 1994, theorem 5.3.1). The nonnegative coefficients $\{\lambda_k\}$ are eigenvalues of the integral operator with kernel $h(x, y)$, satisfying the Hilbert-Schmidt eigenvalue equation

$$\int_{R^d} h(x, y)\psi(y)\, \mathrm{d}F(y) = \lambda\psi(x). \tag{14}$$

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. Because $|x - y|$ is conditionally negative definite, one can easily see that $h(x, y)$ is positive semidefinite, and thus all eigenvalues in Equation 14 are nonnegative. It is also known that their sum is finite and equal to $E|X - X'|$.

The kernel h is degenerate, that is,

$$\int h(x, y)\, \mathrm{d}F(y) = 0.$$

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

$$\int \psi(y)\, \mathrm{d}F(y) = 0.$$

For such a ψ in Equation 14, the y-independent terms in $h(x, y)$ integrate to 0 and thus Equation 14 simplifies to

$$\int (E|y - X| - |x - y|)\psi(y)\, \mathrm{d}F(y) = \lambda\psi(x).$$

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

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

where (a,b) is the support of $\mathrm{d}F$. (Note that a, b can be infinite.) Now, letting $x \to a$,

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

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

$$-\frac{1}{2f}\psi'' = \mu\psi. \tag{15}$$

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

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

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 Equation 15 corresponds to pure kinetic energy because the $V(x)\psi(x)$ term is missing in Equation 15. We can thus call μ in Equation 15 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. This can be called a counterpart of the law of conservation of energy in physics.

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

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

where $\delta(\cdot)$ is the Dirac delta function, so in one dimension $|x-y|$ is a harmonic function. In higher dimensions $|x-y|$ is not harmonic but is at least subharmonic.

6. MULTISAMPLE ENERGY STATISTICS

6.1. Testing for Equal Distributions

Suppose that $\mathbf{X} = X_1, \ldots, X_{n_1}$ and $\mathbf{Y} = Y_1, \ldots, Y_{n_2}$ are independent random samples from the distributions of X and Y, respectively. $\mathcal{E}(X,Y)$ is the energy distance between the distributions of the variables X and Y. The two-sample energy statistic corresponding to the energy distance $\mathcal{E}(X,Y)$ is

$$\mathcal{E}_{n_1,n_2}(\mathbf{X},\mathbf{Y}) = \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}|X_i - Y_m| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|X_i - X_j| - \frac{1}{n_2^2}\sum_{\ell=1}^{n_2}\sum_{m=1}^{n_2}|Y_\ell - Y_m|. \tag{16}$$

The statistic $T_{n_1,n_2} = \frac{n_1 n_2}{n_1+n_2}\mathcal{E}_{n_1,n_2}$ can be applied for testing equality of distributions of X and Y. The null distribution of T_{n_1,n_2} depends on the distributions of X and Y, so the test is implemented nonparametrically as a permutation test in the `energy` package. The null hypothesis of equal distributions is rejected for large T_{n_1,n_2}. For details, applications, and power comparisons, the reader is directed to Rizzo (2003), Székely & Rizzo (2004), and Baringhaus & Franz (2004, 2010).

The two-sample energy distance can be applied to other problems in inference. Some of these applications and extensions of the two-sample energy statistic follow.

6.2. Testing for Diagonal 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 by Buja et al. (1994) and also by Székely & Móri (2001) that if X, X' are i.i.d. 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 distance skewness, a measure of asymmetry.

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

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

Distance skewness has the property that $0 \leq \mathrm{dSkew}(X) \leq 1$, 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 coefficient is defined as

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

Let $Y_i = -X'_i, i = 1, \ldots, n$ be the reflected $\mathbf{X}$ sample in randomized order. A consistent test of diagonal symmetry against general alternatives can be based on the statistic

$$T_n(\mathbf{X}) := 1 + \sum_{1 \leq i < j \leq n} \frac{|X_i + X_j| - |X_i - X_j|}{\sum_{1 \leq i \leq n} |X_i|} \tag{17}$$

$$= \frac{\sum_{i,j=1}^n |X_i - Y_j| - \sum_{i,j=1}^n |X_i + Y_j|}{2\sum_{i,j=1}^n |X_i|}. \tag{18}$$

Observe that T_n is a normalized two-sample energy statistic because the numerator of the fraction in Equation 17 is exactly half the sample energy distance of X and $-X$. The numerator in Equation 18 is proportional to Equation 16 for samples X_i and $Y_i = -X'_i$, and its expected value is $2E|X|$.

For test implementation, one can apply a randomization test. Alternately one can apply the chi-squared test criterion in Székely & Bakirov (2003). That is, reject the null hypothesis at significance level α if $T_n \geq (\Phi^{-1}(1-\alpha/2))^2$, where Φ is the standard normal CDF. This criterion is valid for any significance level less than 0.215. However, the chi-squared test criterion tends to be quite conservative based on empirical studies. An interesting special case is the inequality $E|X + X'| \geq E|X - X'|$, which is discussed by Mensheinin & Zubkov (2012).

Note that if C is an orthonormal matrix, that is, $CC^T = C^T C = I_d$ where I_d is the $d \times d$ identity matrix and T denotes the transpose, then $E|X - CX'| \geq E|X - X'|$ with equality if and only if the distribution of X is invariant with respect to the group generated by C. Based on this observation we can construct a test for spherical symmetry when the distribution of X is invariant with respect to the group of all orthogonal matrices. With a minor modification we can also test if X is elliptically symmetric.

6.3. Distance Components: A Nonparametric Extension of ANOVA

A multisample test of equal distributions is a type of generalization of the hypothesis of equal means. Distance components (DISCO) refers to a decomposition of total dispersion measured by distances that is analogous to the ANOVA (analysis of variance) decomposition of variance. The resulting energy statistic determines a consistent test for the K-sample hypothesis $H_0 : F_1 = \cdots = F_K$, $K \geq 2$. For two samples $A = \{a_1, \ldots, a_{n_1}\}$ and $B = \{b_1, \ldots, b_{n_2}\}$, let

$$g_\alpha(A, B) := \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{m=1}^{n_2} |a_i - b_m|^\alpha, \tag{19}$$

for $0 < \alpha \leq 2$. [$g_\alpha(A, B)$ is a type of Gini mean distance.] The multisample statistics are defined as follows. For samples $A_1, \ldots, A_K$ of sizes $n_1, n_2, \ldots, n_K$, respectively, we define the total dispersion of the observed response by

$$T_\alpha = T_\alpha(A_1, \ldots, A_K) = \frac{N}{2} g_\alpha(A, A), \tag{20}$$

where A is the pooled sample of size $N = \sum_{j=1}^K n_j$. We also define the within-sample dispersion statistic as

$$W_\alpha = W_\alpha(A_1, \ldots, A_K) = \sum_{j=1}^K \frac{n_j}{2} g_\alpha(A_j, A_j), \tag{21}$$

and the between-sample energy statistic as

$$\begin{aligned} S_{n,\alpha} &= \sum_{1 \leq j < k \leq K} \left(\frac{n_j + n_k}{2N} \right) \left[\frac{n_j n_k}{n_j + n_k} \mathcal{E}^{(\alpha)}_{n_j, n_k}(A_j, A_k) \right] \\ &= \sum_{1 \leq j < k \leq K} \left\{ \frac{n_j n_k}{2N} \left(2g_\alpha(A_j, A_k) - g_\alpha(A_j, A_j) - g_\alpha(A_k, A_k) \right) \right\}. \end{aligned} \tag{22}$$

If $0 < \alpha \leq 2$ we have the decomposition of total dispersion $T_\alpha = S_\alpha + W_\alpha$, where both S_α and W_α are nonnegative.

For every $0 < \alpha < 2$, the statistic given in Equation 22 determines a statistically consistent test of the null hypothesis that all of the samples are drawn from identical distributions (Rizzo & Székely 2010), that is, the multisample test of equal distributions. It is important to note that we have a different result if $\alpha = 2$, because in that case the corresponding distance can be zero if the means of the distributions are identical. Indeed, if we consider the case $\alpha = 2$ when F_j are univariate distributions, the statistic $S_{n,2}$ is the ANOVA between-sample sum of squared error (or sum of squares for treatments) and the decomposition $T_2 = S_2 + W_2$ is the ANOVA decomposition. We do not get the characterization of equal distributions for $\alpha = 2$, but 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 DISCO test for equal distributions has been implemented by permutation bootstrap in the `disco` function of the `energy` package for R

(Rizzo & Székely 2016). Examples and power comparisons are given by Rizzo & Székely (2010). An interesting application of DISCO analysis in behavioral biology is given by Schilling et al. (2012).

6.4. $\mathcal{E}$-Clustering: Extensions of Ward's Minimum Variance Method and k-Means

In cluster analysis, Ward's minimum variance method refers to a widely applied agglomerative hierarchical clustering algorithm based on squared distances of observations to cluster centers. In an agglomerative hierarchical clustering algorithm, at each step, the objective is to merge clusters that are homogeneous, as measured by a given cluster distance. Ward's criterion minimizes the within-cluster variance. When we apply $\mathcal{E}$-clustering, we seek to merge clusters with minimum energy distance.

A general class of hierarchical clustering algorithms is determined by a type of recursive formula for updating cluster distances. The energy clustering algorithm defined below and Ward's minimum variance have the same type of recursive formula (see Székely & Rizzo 2005b) and we will see that $\mathcal{E}$-clustering generalizes Ward's minimum variance method.

At a given step in the hierarchical clustering, suppose that the disjoint clusters C_i, C_j would be merged next. Then the updated $\mathcal{E}$-distance between the new cluster $C_i \cup C_j$ and any disjoint cluster C_k is given by

$$d(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k}\, d(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k}\, d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k}\, d(C_i, C_j), \tag{23}$$

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. Using Equation 23, if $d_{ij} := \mathcal{E}(C_i, C_j)$ is given by Equation 16, the updated $\mathcal{E}$-distance can be computed recursively by

$$\begin{aligned} d_{(ij)k} &:= d(C_i \cup C_j, C_k) \\ &= \alpha_i\, d(C_i, C_k) + \alpha_j\, d(C_j, C_k) + \beta\, d(C_i, C_j) \\ &= \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|, \end{aligned} \tag{24}$$

where

$$\alpha_i = \frac{n_i + n_k}{n_i + n_j + n_k}; \qquad \beta = \frac{-n_k}{n_i + n_j + n_k}; \qquad \gamma = 0. \tag{25}$$

Other types of distances can replace the Euclidean distance in Equation 24 to obtain other clustering algorithms. In fact, if we substitute squared Euclidean distances for Euclidean distances in Equation 24, with the same parameters, we obtain the updating formula for Ward's minimum variance method (Equation 25). That is, if we change the exponent on Euclidean distance to $\alpha = 2$, we obtain Ward's method. However, as we have shown in Proposition 2*b* (Section 8), the exponent $\alpha = 2$ only characterizes differences in cluster means. By Proposition 2*a*, if we apply any exponent $0 < \alpha < 2$, the cluster distance measures differences between distributions. Thus, we can replace Euclidean distances in Equation 24 with $|x - y|^\alpha$ for any $0 < \alpha \leq 2$ to obtain a class of clustering algorithms that contain Ward's minimum variance method as a special case. Energy clustering for $0 < \alpha < 2$ generalizes Ward's minimum variance method, providing a class of algorithms that separates clusters that differ in distribution (in any way).

One practical advantage of $\mathcal{E}$-clustering over geometric or cluster-center methods, such as centroid, median, or Ward, is the ability to separate and identify clusters with equal or nearly equal centers. Simulation studies by Székely & Rizzo (2005b) show that $\mathcal{E}$-clustering effectively

recovers the underlying hierarchical structure of data in a variety of scenarios, including high-dimensional data and data with attributes on different scales. In an example with simulated normal data with different covariance structures but nearly equal means, $\mathcal{E}$ outperformed six commonly applied hierarchical clustering methods. Results suggest that the theoretical advantages of energy distance correspond to a performance advantage for certain clustering problems, and that benefit is achieved without sacrificing the good properties of Ward's minimum variance method for separating spherical clusters.

Another energy method in cluster analysis is k-groups, which generalizes the popular k-means clustering method (Li 2015). Whereas k-means finds clusters with well-separated cluster means, k-groups separates clusters with different distributions.

7. DISTANCE CORRELATION: MEASURING DEPENDENCE AND THE ENERGY TEST OF INDEPENDENCE

In this section we focus on the dependence coefficients, distance covariance, and distance correlation introduced by 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 the dimension). To quote Newton (2009, p. 1),

> 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 of the pairwise distances between sample observations of the X sample, to get a distance matrix. Similarly, we 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 (Equation 34) accomplishes the centering. Now, we 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), $\mathcal{V}(X, Y)$, 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. Distance correlation is a very effective tool to detect novel associations in large data sets (see Simon & Tibshirani 2011, Wahba 2014).

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. Other recent consistent tests of bivariate or multivariate independence have been proposed, for example by Feuerverger (1993) and Gretton & Györfi (2010, 2012).

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 $\hat{f}_{X,Y} = \hat{f}_X \hat{f}_Y$. In the following definition, c_p, c_q are given by Equation 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

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|\hat{f}_{X,Y}(t,s) - \hat{f}_X(t)\hat{f}_Y(s)\|^2 \\ &= \frac{1}{c_p c_q}\int_{R^{p+q}} \frac{|\hat{f}_{X,Y}(t,s) - \hat{f}_X(t)\hat{f}_Y(s)|^2}{|t|_p^{1+p}\,|s|_q^{1+q}}\,\mathrm{d}t\,\mathrm{d}s. \end{aligned} \tag{26}$$

If $E|X|_p < \infty$ and $E|Y|_q < \infty$ then by Lemma 1 (Section 8) and by Fubini's theorem, we can evaluate

$$\begin{aligned}\mathcal{V}^2(X,Y) = E[|X - X'|_p|Y - Y'|_q] + E|X - X'|_p\, E|Y - Y''|_q \\ - 2E[|X - X'|_p|Y - Y''|_q],\end{aligned} \tag{27}$$

where (X,Y), (X',Y'), and (X'',Y'') are i.i.d..

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

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0; \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases} \tag{28}$$

Some properties of distance covariance are

1. $\mathcal{V}(a_1 + b_1C_1X,\, a_2 + b_2C_2Y) = \sqrt{|b_1b_2|}\,\mathcal{V}(X,Y)$, 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.
2. Distance covariance is not covariance of distances, but (applying Equation 27) 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 properties of distance variance are:

1. $\mathcal{V}(X) = 0$ implies that $X = E[X]$, almost surely.
2. If X and Y are independent, then $\mathcal{V}(X + Y) \leq \mathcal{V}(X) + \mathcal{V}(Y)$. Equality holds if and only if one of the random vectors X or Y is constant. [It would be interesting to find sharp upper and lower bounds on $\mathcal{V}(X + Y)$ in terms of $\mathcal{V}(X), \mathcal{V}(Y), \mathcal{V}(X,Y)$.]

3. $\mathcal{V}(a + bCX) = |b|\mathcal{V}(X)$, 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

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

$$Y = a + bXC$$

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

The last property shows that for uncorrelated (nondegenerate) random variables X, Y, the distance correlation cannot be 1. If $P(X = 0) = p$ and $P(X = -1) = P(X = 1) = (1-p)/2$, $Y = |X|$ then X, Y are uncorrelated, and as $p \to 1$ we have that $\mathrm{dCor}^2(X, Y) \to 1/\sqrt{2} = 0.7071\ldots$. We conjecture that this is the best upper bound. The fact that for uncorrelated random variables $\mathrm{dCor} < 1$ is a very good property of dCor compared with some other measures of dependencies such as maximal correlation that can easily be 1 for uncorrelated variables.

The sample distance covariance statistic $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$ introduced at the beginning of this section has a simple form (Equation 35). 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. 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

$$\|\hat{f}_{X,Y}^n(t, s) - \hat{f}_X^n(t)\hat{f}_Y^n(s)\|^2 = S_1 + S_2 - 2S_3, \tag{29}$$

where

$$S_1 = \frac{1}{n^2}\sum_{k,\ell=1}^{n} |X_k - X_\ell|_p |Y_k - Y_\ell|_q, \tag{30}$$

$$S_2 = \frac{1}{n^2}\sum_{k,\ell=1}^{n} |X_k - X_\ell|_p \frac{1}{n^2}\sum_{k,\ell=1}^{n} |Y_k - Y_\ell|_q, \tag{31}$$

$$S_3 = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{\ell,m=1}^{n} |X_k - X_\ell|_p |Y_k - Y_m|_q, \tag{32}$$

and

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = S_1 + S_2 - 2S_3, \tag{33}$$

where $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$ is given by Equation 35 (defined below).

For a random sample $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \ldots, n\}$ of n i.i.d. 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

$$A_{k\ell} = a_{k\ell} - \bar{a}_{k\cdot} - \bar{a}_{\cdot\ell} + \bar{a}_{\cdot\cdot}, \qquad k, \ell = 1, \ldots, n, \tag{34}$$

where

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

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

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

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,\ell=1}^{n} A_{k\ell} B_{k\ell}, \tag{35}$$

and

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

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). 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. For more details on the properties of distance covariance, distance variance, and distance correlation, see Székely et al. (2007).

In addition to testing independence, there are several other important applications of distance covariance and distance correlation. Székely & Rizzo (2009) applied dCov statistics to identify influential observations. Distance covariance has been applied by Matteson & Tsay (2016) 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).

The first paper on distance correlation, Székely et al. (2007), contains a generalization for heavy-tailed distributions whose first moment does not exist. For applications of this generalization to energy goodness-of-fit for Pareto family, Cauchy, and stable distributions, the reader is directed to Rizzo (2009) and Yang (2012).

For time series, definitions of an "auto" distance correlation analogous to autocorrelation have been considered by Matteson & Tsay (2016), Rémillard (2009), and Zhou (2012). For Martingale difference sequences, see Shao & Zhang (2014) and Park et al. (2015), and for a literature review of dependence measures, see Josse & Holmes (2014).

7.2. An Unbiased Distance Covariance Statistic and Its Complexity

The following is from Székely & Rizzo (2014).

Let $A = (a_{ij})$ be a symmetric, real-valued $n \times n$ matrix with zero diagonal (not necessarily Euclidean distances). Define the U-centered matrix $\tilde{A}$ by the (i, j)th entry

of $\tilde{A}$:

$$\tilde{A}_{i,j} = \begin{cases} a_{i,j} - \frac{1}{n-2}\sum_{i=1}^{n} a_{i,j} - \frac{1}{n-2}\sum_{j=1}^{n} a_{i,j} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^{n} a_{i,j}, & i \neq j; \\ 0, & i = j. \end{cases}$$

Here, "U-centered" refers to the result that the corresponding squared distance covariance statistic is an unbiased estimator of the population coefficient.

As the first step in computing the unbiased statistic, we replace the double centering operation with U-centering, to obtain U-centered distance matrices $\tilde{A}$ and $\tilde{B}$. Then

$$(\tilde{A} \cdot \tilde{B}) := \frac{1}{n(n-3)} \sum_{i \neq j} \tilde{A}_{i,j} \tilde{B}_{i,j}$$

is an unbiased estimator of squared population distance covariance $\mathcal{V}^2(X, Y)$. The inner product notation is used because this statistic is an inner product in the Hilbert space of U-centered distance matrices (Székely & Rizzo 2014).

A bias-corrected $\mathcal{R}_n^2$ is defined by normalizing the inner product statistic with the bias-corrected dVar statistics. The bias-corrected dCor statistic is implemented in the R `energy` package by the `bcdcor` function. The inner product notation itself suggests that the bias corrected dCor statistic can be interpreted as the cosine of the angle between two U-centered matrices in the Hilbert space of these matrices.

It is clear that both the sample dCov and the sample dCor can be computed in $O(n^2)$ steps. Recently, Huo & Székely (2016) proved that for real-valued samples the above-described unbiased estimator of the squared population distance covariance can be computed by an $O(n \log n)$ algorithm. The supplementary files to Huo & Székely (2016) include an implementation in Matlab.

For more information on a bias-corrected distance correlation and its application to a t-test of independence see Székely & Rizzo (2013b).

7.3. Distance Correlation for Dissimilarity Matrices and the Additive Constant Invariance

It is important to notice that $\tilde{A}$ does not change if we add the same constant to all off-diagonal entries and U-center the result. This additive constant invariance is crucial for applications when, instead of distance matrices, we only have dissimilarity matrices (zero-diagonal, symmetric matrices). It is known (Cailliez 1983) that if we add a big enough constant to all off-diagonal entries of a dissimilarity matrix then we get a distance matrix; thus, we can apply our distance correlation approach methods because additive constants do not change distance correlation. Székely & Rizzo (2014) outline an algorithm, and their approach is a strong competitor of the Mantel test for association between dissimilarity matrices. This makes dCov tests and energy statistics ready to apply to problems in community ecology, where one must often work with data in the form of non-Euclidean dissimilarity matrices.

7.4. Partial Distance Correlation

Based on the inner product dCov statistic (the unbiased estimator of $\mathcal{V}^2$) theory was developed to define partial distance correlation analogous to (linear) partial correlation. There is a simple computing formula for the pdCor (partial distance correlation) statistic and there is a test for the hypothesis of zero pdCor based on the inner product. Energy statistics are defined for random

vectors, so $\mathrm{pdCor}(X, Y\,; Z)$ is a scalar coefficient defined for random vectors X, Y, and Z in arbitrary dimension. The statistics and tests are described in detail by Székely & Rizzo (2014) and currently implemented in the R package `energy` (Rizzo & Székely 2016).

7.5. Alternative Definitions of Distance Covariance: Affine and Monotone Invariant Versions

The original distance covariance was defined as the square root of $\mathrm{dCov}^2(X, Y)$ rather than the squared coefficient itself. Thus, $\mathrm{dCov}(X, Y)$ has the property that it is the energy distance between the joint distribution of X, Y and the product of its marginals. Under this definition, however, the distance variance, rather than the distance standard deviation, is measured in the same units as the pairwise X distances.

Alternately, one could define distance covariance to be the square of the energy distance: $\mathrm{dCov}^2(X, Y)$. In this case, the distance standard deviation of X is measured in the same units as the X distances. Standard deviation, Gini's mean difference, and distance standard deviation are measures of dispersion: Standard deviation works with deviations from the center measured by the arithmetic average of data, Gini's mean difference works with data distance without centering them, and distance standard deviation works with centered (doubly centered) data distances. It is easy to see that for real-valued random variables with finite variance, the distance standard deviation cannot be bigger than the standard deviation, nor can it be bigger than Gini's mean difference.

Under these alternate definitions, the distance correlation is also defined as the square $\mathrm{dCor}^2(X, Y)$, rather than the square root. Using this alternate definition, the formula for sample partial distance correlation in Székely & Rizzo (2014) becomes

$$\mathcal{R}^*_n(X, Y\,; Z) = \begin{cases} \frac{\mathcal{R}_n(X,Y) - \mathcal{R}_n(X,Z)\mathcal{R}_n(Y,Z)}{\sqrt{1-\mathcal{R}^2_n(X,Z)}\sqrt{1-\mathcal{R}^2_n(Y,Z)}}, & \mathcal{R}_n(X, Z) \neq 1 \text{ and } \mathcal{R}_n(Y, Z) \neq 1; \\ 0, & \mathcal{R}_n(X, Z) = 1 \text{ or } \mathcal{R}_n(Y, Z) = 1, \end{cases} \tag{36}$$

where $\mathcal{R}_n$ denotes the sample (partial) distance correlation.

This formula is similar to a computing formula for partial correlation where X, Y, Z are real-valued random variables. For partial distance correlation in the formula above, we do not need this restriction: X, Y, Z can have arbitrary, not necessarily equal dimensions. A related approach is via a nonparametric notion of residuals (Patra et al. 2015).

The reader is also directed to Dueck et al. (2014) for more information on affine invariant distance correlation. Here we apply the distance correlation formulae for the standardized sample. One can also apply the distance correlation formula to copula-transformed random variables or to ranks of observations; this results in a version of distance correlation that is invariant with respect to monotone transformations. The copula version of distance correlation can be considered more equitable. For more information on a test of independence based on ranks of distances, the reader is directed to Heller et al. (2013).

7.6. Brownian Covariance

There is a very interesting duality between distance covariance and a covariance with respect to a stochastic process, defined below. We show in this section that when the stochastic process is Brownian motion (Wiener process) the Brownian covariance coincides with distance covariance (for more details, see Székely & Rizzo 2009).

To motivate Definition 5, first, consider two real-valued random variables X, Y. The square of their ordinary covariance can be written

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

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, we define the U-centered version of X:

$$X_U = U(X) - \int_{-\infty}^{\infty} U(t)\, \mathrm{d}F_X(t) = U(X) - E[U(X) \mid U],$$

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

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

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

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

Definition 5. The Brownian covariance of two real-valued random variables X and Y with finite first moments is a nonnegative number defined by its square

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

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

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

Definition 5 can be extended to random processes and random vectors in higher dimension (see Székely & 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 & Rizzo 2009, theorem 7) that $\mathrm{Cov}_W(X, Y)$ exists for random vectors X and Y with finite second moments:

Theorem 2. If X is an R^p-valued random variable, Y is an R^q-valued random variable, and $E(|X| + |Y|) < \infty$, then $E[X_W X'_W Y_{W'} Y'_{W'}]$ is nonnegative and finite, and

$$\begin{aligned} \mathcal{W}^2(X, Y) &= E[X_W X'_W Y_{W'} Y'_{W'}] \\ &= E|X - X'||Y - Y'| + E|X - X'|E|Y - Y'| \\ &\quad - E|X - X'||Y - Y''| - E|X - X''||Y - Y'|, \end{aligned} \tag{37}$$

where (X, Y), (X', Y'), and (X'', Y'') are i.i.d..

If we compare Theorem 2 and Equation 27, 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 & Rizzo 2009, theorem 8 for the proof of Theorem 2). Thus, Pearson's classical correlation corresponds to the simplest nonrandom process, the identity function, whereas the distance correlation, dCor, corresponds to the correlation with respect to Brownian motion. This is an interesting duality.

We can broaden the applicability of distance correlation and Brownian correlation if we work with $0 < \alpha < 2$ powers of distances and with fractional Brownian motions with Hurst parameter $0 < H = \alpha/2 < 1$, respectively (Herbin & Merzbach 2007). Here, we do not need to suppose that X, Y have finite expectations, we just need finite α moments for some positive α. We have the same duality as before: the distance correlation computed from α powers of distances, dCor_α, equals the correlation with respect to fractional Brownian motion with Hurst parameter $0 < H = \alpha/2 < 1$. We can go even further and work with conditionally negative definite distances δ in metric spaces. For a related interesting result, the reader is directed to Genovese (2009). More details on these kinds of generalized energy inferences can be found in Section 8.

Interestingly, a statistic for the special case $\alpha = 2$ (dCor_2) was introduced by Escoufier (1973) and by Robert & Escoufier (1976), where it is called the RV coefficient. (It does not characterize independence, but generalizes Pearson's coefficient to higher dimensions.)

8. GENERALIZED ENERGY DISTANCE

Because 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

(a) For $0 < \alpha < 2$,

$$\mathcal{E}^{(\alpha)}(X,Y) = 2E|X-Y|^\alpha - E|X-X'|^\alpha - E|Y-Y'|^\alpha = \frac{1}{C(d,\alpha)} \int_{R^d} \frac{|\hat{f}(t)-\hat{g}(t)|^2}{|t|^{d+\alpha}}\,\mathrm{d}t, \tag{38}$$

where

$$C(d,\alpha) = 2\pi^{d/2}\frac{\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)}. \tag{39}$$

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

Statements *a* and *b* 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 $\alpha = 2$ because we have equality to zero in *b* whenever $E(X) = E(Y)$.

Some applications of Proposition 2 include

1. Goodness-of-fit tests for heavy tailed distributions such as stable distributions (Yang 2012) and Pareto distributions (Rizzo 2009)
2. Generalization of Ward's minimum variance criterion in hierarchical cluster analysis (see Section 6.4)
3. Generalization of distance covariance for heavy tailed distributions (Székely et al. 2007)
4. The energy score (Gneiting & Raftery 2007)

A proof of Proposition 2 is based on the following lemma.

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

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

where (t, x) represents inner product, $C(d, \alpha)$ is the constant (Equation 39) 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 by Székely & Rizzo (2005b).

9. ENERGY IN METRIC SPACES AND MACHINE LEARNING

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; however, if we replace Euclidean distance with a metric δ in an arbitrary metric space, then the claim 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.

A necessary and sufficient condition that the above characterization of equal distributions holds is established in Proposition 3 below. A metric space $(\mathcal{X}, \delta)$ has negative type if for all $n \geq 1$ and all sets of n red points x_i and n blue points x_i' in $\mathcal{X}$, we have

$$2 \sum_{i,j} \delta(x_i, x_j') - \sum_{i,j} \delta(x_i, x_j) - \sum_{i,j} \delta(x_i', x_j') \geq 0.$$

If we take repetitions of x_i and take limits then we get a seemingly more general property called conditional negative definiteness of δ: For all $n \leq 1$, $x_1, x_2, \ldots, x_n \in \mathcal{X}$, and $a_1, a_2, \ldots, a_n$ real numbers with $\sum_{i=1}^n a_i = 0$, we have

$$\sum_{i,j} a_i a_j \delta(x_i, x_j) \leq 0.$$

Some classical results on conditional negative definiteness are given by Schoenberg (1938a,b). The metric space $(\mathcal{X}, \delta)$ has strict negative type if for every $n \geq 1$ and for all distinct points $x_1, x_2, \ldots, x_n$, equality holds only if $a_i = 0$ for all i. Now suppose that the Borel probability measures $\mu_i, i = 1, 2$ on $\mathcal{X}$ have finite first moments—that is, $\int \delta(o, x) \mathrm{d}\mu_i(x) < \infty$ for some (and thus for all) $o \in \mathcal{X}$. If we approximate μ_i by probability measures of finite support, we arrive at an even more general version of the energy inequality:

$$\int \delta(x_1, x_2) \mathrm{d}(\mu_1 - \mu_2)^2(x_1, x_2) \leq 0.$$

For more information on finite metric spaces of strictly negative type, the reader is directed to Hjorth et al. (1998). The metric space $(\mathcal{X}, \delta)$ has strong negative type if it has negative type and equality in the last inequality holds if and only if $\mu_1 = \mu_2$. Lyons (2013, Remark 3.3) gives an example of a countably infinite metric space that is strictly, but not strongly, negative definite. More information on the application of the concept of negative definiteness to metric spaces of connected graphs where the distance between vertices is the length of the shortest path can be found in Deza & Laurent (1997) and Meckes (2013). So far, we do not have a characterization

of graphs whose metric space has strict negative type but, for example, whose weighted trees have negative type (Meckes 2013, theorem 3.6). Another open problem is the characterization of Riemannian spaces that have strong negative type (see, e.g., Feragen et al. 2015). It is easy to see that spheres do not have strict negative type. For a counterexample, it is enough to check four points of the sphere: the north pole, the south pole, and the middle points of the semicircles that connect them on the sphere.

The following proposition is easy to see.

Proposition 3. Let X, Y be independent random variables with distributions μ_1, μ_2, respectively, X' is an i.i.d. copy of X and Y' is an i.i.d. copy of Y.

(a) A necessary and sufficient condition that

$$2E\delta(X,Y) - E\delta(X,X') - E\delta(Y,Y') \geq 0 \tag{40}$$

holds for all X, Y is that $(\mathcal{X}, \delta)$ has negative type.

(b) In Equation 40, a necessary and sufficient condition that

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

holds if and only if X and Y are identically distributed ($\mu_1 = \mu_2$) is that the metric space has strong negative type.

Proposition 1 shows that Euclidean spaces have strong negative type. The same holds for hyperbolic spaces (Lyons 2015) and for all separable Hilbert spaces (Lyons 2013). This is very important in applications to function valued data (Horváth & Kokoszka 2012). Lyons (2013) proved that if $(\mathcal{X}, \delta)$ has negative type then $(\mathcal{X}, \delta^r)$ with $0 < r < 1$ has strong negative type. Thus, Proposition 2 follows from the observation that in Euclidean spaces $|x-y|^2$ has negative type because $\sum_{i=1}^n a_i = 0$ implies that $\sum_{i,j} a_i a_j |x_i - x_j|^2 = -2|\sum_{i,j} a_i x_i|^2 \leq 0$.

In the theory of machine learning, conditionally negative definite distances $\delta(x,y)$ and positive definite kernels $k(x,y)$ play crucial roles (see, e.g., Vapnik 1995, Wahba 1990, Schölkopf & Smola 2002, Sejdinovic et al. 2013). According to the kernel trick, if $k(x,y)$ is a positive definite kernel, then $k(x,x) + k(y,y) - 2k(x,y)$ is the square of a conditionally negative definite distance (Berg et al. 1984). The most frequently applied positive definite kernels are the Gaussian $\exp(-c|x-y|^2)$ or the Laplacian $\exp(-c|x-y|)$, where x, y are in a d-dimensional Euclidean space.

10. HISTORICAL BACKGROUND

In 1927–28 two fundamental papers changed our views of matter and mind. Heisenberg's uncertainty principle (Heisenberg 1927) led to a probabilistic description of matter (unlike Newton's deterministic laws), and von Neumann's minimax theorem of zero-sum games (von Neumann 1928) led to a probabilistic description of our mind (see also the uncertainty principle of game theory in Székely & Rizzo 2007). The energy perspective in this review reflects another type of duality between matter and mind. The terms "statistical energy," "$\mathcal{E}$-statistic," and "energy of data" were coined in Székely's lectures at MIT, Yale, and Columbia University in 1985.

The prehistory of statistical energy can be traced back to Riesz (1938, 1949), who defined the α-energy of a measure μ as $E|X - X'|^\alpha$ where X, X' are i.i.d. with probability measure μ. The history 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 Equations 12 and 19 for the

Fourier transform of $|x|^{\alpha}$ (Gel'fand & Shilov 1964, pp. 173–74)]. Thus, one can extend the validity of Lemma 1 using generalized functions, but the Proposition itself is not in Gel'fand & Shilov (1964). The duality between powers of distances and their Fourier transforms 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 a "beautiful theorem of probability theory (*Schönes Theorem der Wahrscheinlichkeitrechnung*)" by Gauss (Fischer 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). More details on this inequality can be found in Buja et al. (1994). Russian mathematicians also published proofs of these propositions and their generalizations to metric spaces (see Klebanov 2005 and the references therein). Mattner (1997) and Morgenstern (2001) present details from the German school. For more information on early applications, the reader is directed to the test of bivariate independence of Feuerverger (1993) and the test of homogeneity of Baringhaus & Franz (2004). Historical comments on Hoeffding-type inequalities and their generalizations are presented by Gneiting & Raftery (2007, section 5.2). In the past decade, the energy of data has become a widely applicable tool in many areas of science and technology.

SUMMARY POINTS

1. Energy statistics are nonnegative-valued functions of distances between observations, based on statistical potential energy.
2. Sample energy distance is the square of a metric on the space of samples of size n.
3. Energy tests of independence, equality of distributions, goodness-of-fit, and skewness, as well as hierarchical clustering and distance components analysis of structured data, are some of the important applications of energy statistics.
4. Energy methods are powerful, consistent, and practical to apply.

FUTURE ISSUES

1. For energy inference on data better represented in Riemannian manifolds than in Euclidean spaces (e.g., diffusion MRI imaging of the brain), characterize Riemannian manifolds that have strong negative type.
2. For energy inference on graph-structured data, for example tree-structured data, characterize graphs with geodesic distance that is conditionally (strictly) negative definite. Concerning important distances between graphs, such as sampling distance, cut distance (Lovász 2012), and spectral distance (Jovanovič & Stanič 2012), characterize their (strict) conditional negative definiteness.
3. When we have flexibility to choose the metric δ, find a good/optimal choice of δ for statistical inferences such as clustering or testing independence. A special case of this open problem is a good/optimal choice of the exponent α in Euclidean spaces.

4. Work out details of parameter estimation based on minimization of the energy distance between data and important parametric families of distributions. They are clearly special cases of M-estimators that minimize $\sum_{i=1}^{n} g(x_i, \theta)$ with respect to the parameter θ where g is some function of an observation x_i and a parameter. A related topic is energy regression or distance regression, which is based on the minimization of the energy distance between the data y_i and its estimator $\hat{y}_i$, $i = 1, 2, \ldots, n$, for example, in the case of simple linear regression $\hat{y}_i = a + bx_i$ where a, b are real numbers. This topic is of course related to nonparametric regression (Jurečková 1972) and to Gini's regression (Olkin & Yitzhaki 1992). A Bayesian approach was proposed by Chakraborty & Bhattacharjee (2015).

5. Instead of the data distance of this review

$$D_\alpha := 2\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - Y_j|^\alpha - \sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|^\alpha - \sum_{i=1}^{n}\sum_{j=1}^{n}|Y_i - Y_j|^\alpha$$

we can work with Wasserstein-type distances (Villani 2008)

$$W_\alpha := \min_{\pi}\sum_{i=1}^{n}|X_i - Y_{\pi(i)}|^\alpha$$

where π is a permutation of the integers $1, 2, \ldots, n$. Work out the details of the corresponding statistical inferences. Whereas D_α includes all possible distances between observations, W_α includes the minimum distances only. Compare D_α and W_α from statistical points of view. For more information on Wasserstein type distances, the reader is directed to the work of Ajtai et al. (1984), Talagrand (2005), and Major (2013). The reader is directed to Ramdas et al. (2015) for some initial results.

6. Combine energy inference with sparse inference.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENT

The research of G.S. is supported by the National Science Foundation while working at the Foundation.

LITERATURE CITED

Aaronson J, Burton R, Dehling H, Gilat D, Hill T, Weiss B. 1996. Strong laws for L- and U-statistics. *Trans. Am. Math. Soc.* 348:2845–65

Ajtai M, Komlós J, Tusnády G. 1984. On optimal matchings. *Combinatorica* 4:259–64

Baringhaus L, Franz C. 2004. On a new multivariate two-sample test. *J. Multivar. Anal.* 88:190–206

Baringhaus L, Franz C. 2010. Rigid motion invariant two-sample tests. *Stat. Sin.* 20:1333–61

Berg C. 2008. Stieltjes-Pick-Bernstein-Schoenberg and their connection to complete monotonicity. In *Positive Definite Functions: From Schoenberg to Space-Time Challenges*, ed. J Mateu, E Porcu. Castellon, Spain: Dept. Math. University Jaume I

Berg C, Christensen JPR, Ressel P. 1984. *Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions*. New York: Springer-Verlag
Brillouin L. 2004. *Science and Information Theory*. Mineola, New York: Dover. 2nd ed.
Buja A, Logan BF, Reeds JA, Shepp LA. 1994. Inequalities and positive definite functions arising from a problem in multidimensional scaling. *Ann. Stat.* 22:406–38
Cailliez F. 1983. The analytical solution of the additive constant problem. *Psychometrika* 48:343–49
Chakraborty S, Bhattacharjee A. 2015. Distance correlation measures applied to analyze relation between variables in liver cirrhosis marker data. *Int. J. Collab. Res. Intern. Med. Public Health* 7:1–7
Cramér H. 1928. On the composition of elementary errors. Second paper: statistical applications. *Scand. Actuar. J.* 11:141–80
Deza MM, Laurent M. 1997. *Geometry of Cuts and Metrics*. Berlin: Springer-Verlag
Dueck J, Edelmann D, Gneiting T, Richards D. 2014. The affinely invariant distance correlation. *Bernoulli* 20:2305–30
Einstein A. 1905. Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig? *Ann. Phys.* 18:639–43
Escoufier Y. 1973. Le traitement des variables vectorielles. *Biometrics* 29:751–60
Fan J, Feng Y, Xia L. 2015. A conditional dependence measure with applications to undirected graphical models. arXiv:1501.01617 [stat.ME]
Feragen A, Lauze F, Hauberg S. 2015. Geodesic exponential kernels: when curvature and linearity conflict. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3032–42. New York: IEEE
Ferenci T, Körner A, Kovács L. 2015. The interrelationship of HbA1c and real-time continuous glucose monitoring in children with type 1 diabetes. *Diabetes Res. Clin. Pract.* 108:38–44
Feuerverger A. 1993. A consistent test for bivariate dependence. *Int. Stat. Rev.* 61:419–33
Fischer H. 2011. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. New York: Springer
Gel'fand IM, Shilov GE. 1964. *Generalized Functions, Volume I: Properties and Operations*, transl. E Salatan. New York: Academic
Genovese C. 2009. Discussion of: Brownian distance covariance. *Ann. Appl. Stat.* 3:1299–302
Gini C. 1912. *Variabilità e Mutabilità*. Bologna: Tipografia di Paolo Cuppini
Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* 102:359–78
Gretton A, Györfi L. 2010. Consistent nonparametric tests of independence. *J. Mach. Learn. Res.* 11:1391–423
Gretton A, Györfi L. 2012. Strongly consistent nonparametric test of conditional independence. *Stat. Probab. Lett.* 82:1145–50
Guo X, Zhang Y, Hu W, Tan H, Wang X. 2014. Inferring nonlinear gene regulatory network from gene expression data based on distance correlation. *PLOS ONE* 9(2):e87446
Gurtler N, Henze N. 2000. Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Ann. Inst. Stat. Math.* 52:267–86
Heisenberg W. 1927. Über den anschaulichen Inhalt der quantummmechanischen Kinematic und Mechanic. *Z. Phys.* 43:172–98
Heller R, Heller Y, Gorfine M. 2013. A consistent multivariate test of association based on ranks of distances. *Biometrika* 100:503–10
Henze N, Zirkler B. 1990. A class of invariant and consistent tests for multivariate normality. *Commun. Stat. Theory Methods* 19:3595–617
Herbin E, Merzbach E. 2007. The multiparameter fractional Brownian motion. In *Math Everywhere*, ed. G Aletti, A Micheletti, D Morale, M Burger, pp. 93–101. Berlin: Springer
Hjorth P, Lisoněk P, Markvorsen S, Thomassen C. 1998. Finite metric spaces of strictly negative type. *Linear Algebra Appl.* 270:255–73
Hoeffding W. 1948. A class of statistics with asymptotic normal distribution. *Ann. Math. Stat.* 19:293–325
Horn HR. 1972. On necessary and sufficient conditions for an infinitely divisible distribution to be normal or degenerate. *Z. Wahrscheinlichkeitstheorie Verwandte Geb.* 21:179–87
Horváth L, Kokoszka P. 2012. *Inference for Functional Data with Applications*. New York: Springer
Hua W, Nichols TE, Ghosh D. 2015. Multiple comparison procedures for neuroimaging genomewide association studies. *Biostatistics* 16:17–30

Huo X, Székely GJ. 2016. Fast computing for distance covariance. *Technometrics* 58:435–47
Jaroszewitz SZ, Lukasz Z. 2015. Székely regularization for uplift modeling. In *Challenges in Computational Statistics and Data Mining*, pp. 135–54. New York: Springer
Josse J, Holmes S. 2014. Measures of dependence between random vectors and tests of independence. arXiv:1307.7383 [stat.ME]
Jovanovič I, Stanič Z. 2012. Spectral distances of graphs. *Linear Algebra Appl.* 436:1425–35
Jurečková J. 1972. Nonparametric estimate of regression coefficient. *Ann. Math. Stat.* 42:1328–38
Kim AY, Marzban C, Percival DB, Stuetzle W. 2009. Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Process.* 89:2529–36
Klebanov L. 2005. *N-Distances and Their Applications*. Prague: Charles Univ.
Kong J, Klein BEK, Klein R, Lee K, Wahba G. 2012. Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *PNAS* 109:20352–57
Koroljuk VS, Borovskich YuV. 1994. *Theory of U-statistics*, transl. PV Malyshev. Dordrecht: Kluwer
Li R, Zhong W, Zhu L. 2012. Feature screening via distance correlation learning. *J. Am. Stat. Asssoc.* 107:1129–39
Li S. 2015. *k-groups: a generalization of k-means by energy distance*. PhD thesis, Bowling Green State Univ.
Lovász L. 2012. *Large Networks and Graph Limits.* Providence, RI: Am. Math. Soc.
Lu JY, Peng YX, Wang M, Gu SJ, Zhao MX. 2015. Support vector machine combined with distance correlation learning for *Dst* forecasting during intense geomagnetic storms. *Planet. Space Sci.* 120:48–55
Lyons R. 2013. Distance covariance in metric spaces. *Ann. Probab.* 41:3284–305
Lyons R. 2015. Hyperbolic space has strong negative type. *Ill. J. Math.* 58:1009–13
Major P. 2013. *On Estimation of Multiple Random Integrals and U-statistics*. New York:Springer
Mardia KV. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57:519–30
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. *Ann. Inst. Stat. Math.* 57:183–99
Matteson DS, James NA. 2013. A nonparametric approach for multiple change point analysis of multivariate data. arXiv:1306.4933 [stat.ME]
Matteson DS, Tsay RS. 2016. Independent component analysis via distance covariance. *J. Am. Stat. Assoc.* **http://dx.doi.org/10.1080/01621459.2016.1150851**
Mattner L. 1997. Strict negative definiteness of integrals via complete monotonicity of derivatives. *Trans. Am. Math. Soc.* 349:3321–42
Meckes MW. 2013. Positive definite metric spaces. *Positivity* 17:733–57
Mensheinin DO, Zubkov AM. 2012. Properties of the Székely-Móri asymmetry criterion statistics in the case of binary vectors with independent components. *Math. Notes* 91:62–72
Morgenstern D. 2001. Proof of a conjecture by Walter Deuber concerning the distance between points of two types in R^d. *Discret. Math.* 226:347–49
Newton MA. 2009. Introducing the discussion paper by Székely and Rizzo. *Ann. Appl. Stat.* 3:1233–35
Olkin I, Yitzhaki S. 1992. Gini regression analysis. *Int. Stat. Rev.* 60:185–96
Park T, Shao X, Yao S. 2015. Partial Martingale difference correlation. *Electron. J. Stat.* 9:1492–517
Patra RK, Sen B, Székely GJ. 2015. On a nonparametric notion of residual and its applications. *Stat. Probability Lett.* 109:208–13
R Core Team. 2016. R: A language and environment for statistical computing. R Found. Stat. Comput., Vienna, Austria. **https://www.R-project.org/**
Ramdas A, Garcia N, Cuturi M. 2015. On Wasserstein two sample testing and related families of nonparametric tests. arXiv:1509.02237 [math.ST]
Rémillard B. 2009. Discussion of: Brownian distance covariance. *Ann. Appl. Stat.* 3(4):1295–98
Riesz M. 1938. Intégrales de Riemann–Liouville et potentiels. *Acta Sci. Math. Szeged* 9:1–42
Riesz M. 1949. L'intégrale de Riemann–Liouville et le problème de Cauchy. *Acta Math.* 81:1–223
Rizzo ML. 2002. *A new rotation invariant goodness-of-fit test*. PhD thesis, Bowling Green State Univ.
Rizzo ML. 2003. A test of homogeneity for two multivariate populations. *2002 Proc. Am. Stat. Assoc. Spring Res. Conf., SPES* Alexandria, VA: Am. Stat. Assoc.
Rizzo ML. 2009. New goodness-of-fit tests for Pareto distributions. *ASTIN Bull.* 39:691–715

Rizzo ML, Székely GJ. 2010. DISCO analysis: a nonparametric extension of analysis of variance. *Ann. Appl. Stat.* 4:1034–55

Rizzo ML, Székely GJ. 2016. `energy`: *E*-statistics: multivariate inference via the energy of data. R package version 1.7-0. **https://CRAN.R-project.org**

Robert P, Escoufier Y. 1976. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Appl. Stat.* 25:257–65

Rudas J, Guaje J, Demertzi A, Heine L, Tshibanda L, et al. 2014. A method for functional network connectivity using distance correlation. *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 2793–2796. New York: IEEE

Schilling K, Oberdick J, Schilling RL. 2012. Toward an efficient and integrative analysis of limited-choice behavioral experiments. *J. Neurosci.* 32:12651–56

Schrödinger E. 1926. An undulatory theory of the mechanics of atoms and molecules. *Phys. Rev.* 28:1049–70

Schrödinger E. 1944. *What Is Life—the Physical Aspect of the Living Cell*. Cambridge, UK: Cambridge Univ. Press

Schölkopf B, Smola AJ. 2002. *Learning with Kernels*. Cambridge, MA: MIT Press

Schoenberg IJ. 1938a. Metric spaces and completely monotone functions. *Ann. Math.* 39:811–41

Schoenberg IJ. 1938b. Metric spaces and positive definite functions. *Trans. Am. Math. Soc.* 44:522–36

Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.* 41:2263–91

Serfling RJ. 1980. *Approximation Theorems of Mathematical Statistics*. New York: Wiley

Shamsuzzaman MD, Satish M, Pintér JD. 2015. Distance correlation based nearly orthogonal space-filling experimental designs. *Int. J. Exp. Des. Process Optim.* 4:216–33

Shao X, Zhang J. 2014. Martingale difference correlation and its use in high-dimensional variable screening. *J. Am. Stat. Assoc.* 109:1302–18

Sheng W, Yin X. 2013. Direction estimation in single-index models via distance covariance. *J. Multivar. Anal.* 122:148–61

Sheng W, Yin X. 2016. Sufficient dimension reduction via distance covariance. *J. Comput. Gr. Stat.* 25:91–104

Simon N, Tibshirani R. 2011. Comment on "Detecting Novel Associations in Large Data Set" by Reshef et al. arXiv:1401.7645 [stat.ME]

Steutel FW, van Harn K. 2004. *Infinite Divisibility of Probability Distributions on the Real Line*. Boca Raton: CRC

Székely GJ. 1989. *Potential and kinetic energy in statistics*. Lect. notes, Budapest Inst. of Technol.

Székely GJ. 1996. *Contests in Higher Mathematics*. New York: Springer

Székely GJ. 2002. E-statistics: energy of statistical samples. Tech. Rep. No. 02–16, Bowling Green State Univ., Dep. Math. Stat. doi: **http://dx.doi.org/10.13140/RG.2.1.5063.9761**

Székely GJ, Bakirov NK. 2003. Extremal probabilities for Gaussian quadratic forms. *Probab. Theory Related Fields* 126:184–202

Székely GJ, Móri TF. 2001. A characteristic measure of asymmetry and its application for testing diagonal symmetry. *Commun. Stat. Theory Methods* 30:1633–39

Székely GJ, Rizzo ML. 2004. Testing for equal distributions in high dimension. *InterStat*, Nov. **http://interstat.statjournals.net/YEAR/2004/abstracts/0411005.php**

Székely GJ, Rizzo ML. 2005a. A new test for multivariate normality. *J. Multivar. Anal.* 93:58–80

Székely GJ, Rizzo ML. 2005b. Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method. *J. Classif.* 22(2):151–83

Székely GJ, Rizzo ML. 2007. The uncertainty principle of game theory. *Am. Math. Mon.* 114:688–702

Székely GJ, Rizzo ML. 2009. Brownian distance covariance. *Ann. Appl. Stat.* 3:1236–65

Székely GJ, Rizzo ML. 2012. On the uniqueness of distance covariance. *Stat. Probab. Lett.* 82:2278–82

Székely GJ, Rizzo ML. 2013a. Energy statistics: a class of statistics based on distances. *J. Stat. Plann. Inference* 143:1249–72

Székely GJ, Rizzo ML. 2013b. The distance correlation *t*-test of independence in high dimension. *J. Multivar. Anal.* 117:193–213

Székely GJ, Rizzo ML. 2014. Partial distance correlation with methods for dissimilarities. *Ann. Stat.* 42:2382–412

Székely GJ, Rizzo ML, Bakirov NK. 2007. Measuring and testing independence by correlation of distances. *Ann. Stat.* 35:2769–94

Szilárd L. 1929. Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen [On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings]. *Z. Phys.* 53:840–56

Talagrand M. 2005. *The General Chaining*. New York: Springer

Vaiciukynas E, Verikas A, Gelzinis A, Bacauskiene M, Olenina I. 2015. Exploiting statistical energy test for comparison of multiple groups in morphometric and chemometric data. *Chemom. Intell. Lab. Syst.* 146:10–23

Vapnik V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer

Varina T, Bureaua R, Muellerb C, Willett P. 2009. Clustering files of chemical structures using the Székely–Rizzo generalization of Ward's method. *J. Mol. Graph. Model.* 28:187–95

Villani C. 2008. *Optimal Transport: Old and New*. New York: Springer

Von Mises R. 1947. On the asymptotic distributions of differentiable statistical functionals. *Ann. Math. Stat.* 2:209–348

von Neumann J. 1928. Theorie der Gesellschaftspiele. *Math. Ann.* 100:295–320

Wahba G. 1990. *Spline Models for Observational Data*. Philadelphia: SIAM

Wahba G. 2014. *Positive definite functions, reproducing kernel Hilbert spaces and all that.* Presented at Joint Stat. Meet., Aug. 6, Boston, MA. **http://www.stat.wisc.edu/~wahba/talks1/fisher.14/wahba.fisher.7.11.pdf**

Wang X, Pan W, Hu W, Tian Y, Zhang H. 2015. Conditional distance correlation. *J. Am. Stat. Assoc.* 110:1726–34

Würtz D, Chalabi Y, Chen W, Ellis A. 2009. *Portfolio Optimization with R/Rmetrics*. Zurich: Rmetrics Assoc. and Finance Online Publ.

Yang G. 2012. *The energy goodness-of-fit test for univariate stable distributions*. PhD thesis, Bowling Green State Univ.

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*. Oxford, UK: Pergamon

Zanoni M, Setragno F, Sarti A. 2014. The violin ontology. *Proc. 9th Conf. Interdiscip. Musicol.—CIM14, Berlin.* In press

Zhou Z. 2012. Measuring nonlinear dependence in time-series, a distance correlation approach. *J. Time Ser. Anal.* 33:438–57

Annual Review of Statistics and Its Application

Volume 4, 2017

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics