The Energy of Data

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

[1]National Science Foundation, Arlington, VA 22230 and Rényi Institute of Mathematics, Hungarian Academy of Sciences, Hungary; email: gszekely@nsf.gov
[2]Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403; email: mrizzo@bgsu.edu

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–36

This article's doi:
10.1146/((please add article doi))

Copyright © YYYY by Annual Reviews.
All rights reserved

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 observations/data are complex objects, like functions or graphs, we can use their real valued distances for inference. Other advantages will be illustrated and discussed in this paper. Concrete examples include energy testing for normality, energy clustering, and distance correlation. Applications include genome, brain studies, and astrophysics. The direct connection between energy and mind/observations/data in this paper is a counterpart of the equivalence of energy and matter/mass in Einstein's $E = mc^2$.

Contents

1. INTRODUCTION 2
   1.1. Distances of observations – Rigid motion invariance 5
   1.2. Functions of distances – Jackknife invariance 5
   1.3. Energy inference – Scale invariance 7
2. Energy Distance 8
3. Why is Energy Distance Special? 10
4. One Sample Energy Statistics 11
   4.1. Energy goodness-of-fit 11
   4.2. Energy test of normality 12
   4.3. Energy test of multivariate normality 12
5. Kinetic Energy: the Schrödinger Equation of Goodness-of-fit 13
6. Multi-sample Energy Statistics 15
   6.1. Testing for equal distributions 15
   6.2. Testing for symmetry 16
   6.3. Distance components: A nonparametric extension of ANOVA 17
   6.4. $\mathcal{E}$-clustering: Extensions of Ward's minimum variance method and k-means 18
7. Distance Correlation: Measuring Dependence and the Energy Test of Independence 19
   7.1. Definitions of distance covariance and distance correlation 20
   7.2. An unbiased distance covariance statistic and its complexity 23
   7.3. Distance correlation for dissimilarity matrices; the additive constant invariance 23
   7.4. Partial distance correlation 24
   7.5. Alternative definitions of distance covariance; affine and monotone invariant versions 24
   7.6. Brownian covariance 25
8. Generalized Energy Distance 26
9. Energy in Metric Spaces and Machine Learning 27
10. Historical Background 29

1. INTRODUCTION

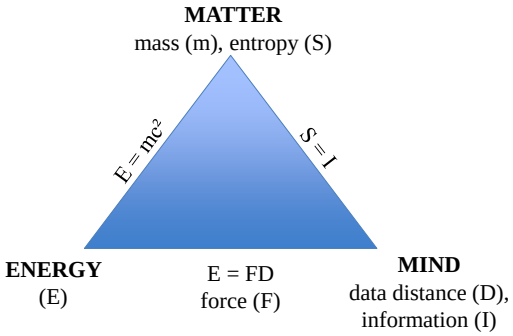
Energy statistics ($\mathcal{E}$-statistics) are functions of distances between statistical observations. The value of the energy statistics for a given data is the (potential) energy of the data. This concept is based on the notion of Newton's gravitational potential energy, which is a function of the 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 an 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 are mass (m), disorder (measured by entropy S), etc. Manifestations of the "immaterial" mind are the memory, information (I), observation, data, etc.; 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; namely, 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

Figure 1

Connections between Energy, Matter, and Mind



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 paper 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 magnitute F; that is, energy $E = F \cdot D$. This resembles the gravitational potential energy close to the Earth where the gravitational force is constant. 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 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 (6.1) where the sample sizes are not necessarily equal.

For statistical purposes we can also work with powers of distances like $|X_i - Y_m|^\alpha$ 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. Since Newton's gravitational potential is proportional to the reciprocal of the distance, while 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. We will see that in the world of data energy or statistical energy "harmonic" is replaced by "conditionally negative definite" (see Section 9).

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 as follows.

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 Section 9 we will see theorems stating that many important metric spaces share the following properties:

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

A necessary and sufficient condition for (i) is the conditional negative definiteness of δ. On this property see the classical papers of Schoenberg (1938a,b) and also Horn (1972); Steutel & van Harn (2004); Berg (2008). Metric spaces with properties (i) and (ii) include all Euclidean spaces, all separable Hilbert spaces, all hyperbolic spaces, and many graphs with geodesic distances. For more details see Propositions 1, 2, 3, and Section 9. These theorems make it possible to develop clustering methods that depend not only on cluster centers but 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 (i) holds but (ii) fails is the taxicab metric or Manhattan distance.

In this paper we present the foundational material, motivation, and unifying theory of energy statistics. Previously unpublished results as well as an overview of several published applications in inference and multivariate analysis 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, 2015), forecasting during geomagnetic storms (Lu et al. 2015), graphical models (Fan et al. 2015; Wang et al. 2015), chemometrics (Vaiciukynas et al. 2015), musicology (Zanoni et al. 2014), etc. 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, Fisher's F-statistics, etc.

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) and Székely (2003)). This idea became the central topic of the first author's NSA Grant "Singular Kernel Nonparametric Tests" submitted in 2000 (NSA grant # MDA 904-02-1-0091).

1.1. Distances of observations – Rigid motion invariance

Classical statistics operated with real valued data like height, weight, blood pressure, etc. 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 the central limit theorem 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, etc.? 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 will 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 like a 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 will have an energy counterpart 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 does not hold for Pearson's correlation, for example, for non-normal data.

1.2. Functions of distances – Jackknife invariance

Statistics are functions of data, while 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 $f^i_{n-1}(x_1, \ldots, x_n)$, $i = 1, 2, \ldots, n$, denote the statistic of a "jackknife sample" when we remove the i-th observation: $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$. ($f^i_0$

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*

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

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 Hou & Székely (2015).

In the following for simplicity we will discuss the case when $r = 2$. Energy statistics in this paper are U-statistics or V-statistics based on distances. That is, for a d-dimensional random sample $X_1, \ldots, X_n$, 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 Euclidean norm if the argument is real, and $|\cdot|$ denotes the complex norm when its argument is complex. The notation $\|\cdot\|$ is reserved for another type of norm in this paper.

An advantage of assuming all energy statistics are U-statistics or V-statistics is that we can apply their classical limit theory (Hoeffding 1948; Von Mises 1947) 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 applied (Koroljuk & Borovskich 1994, Ch. 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. See also Székely & Rizzo (2013b).

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 dist. is centrally symmetric)
Multi-sample 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 k-means clustering (sensitive to cluster centers)	Hierarchical energy clustering k-groups (energy) clustering (sensitive to cluster distributions)

1.3. Energy inference – Scale invariance

We will need an important restriction 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 e.g. 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.

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 then 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' 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 paper we will exploit Thales' simple but brilliant idea in many ways when applying it for statistical inferences. For an example see 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

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

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 dx in Cramér's distance by $dF(x)$, to obtain the Cramér-von Mises-Smirnov distance defined

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

There remains, however, another important disadvantage of both the Cramér L_2 distance (Cramér 1928) and 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 cdf's F and G, respectively, let X' denote an independent and identically distributed (iid) copy of X, and Y' denote an iid copy of Y. Then (see e.g. Székely (1989, 2003))

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

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

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

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 expression (2.3) is nonnegative and equals zero if and only if X and Y are identically distributed. See Proposition 1.

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

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

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

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

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

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

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

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

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

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

where

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

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

For a proof of this proposition see Székely & Rizzo (2005a). For its history see Section 10. For a generalization to metric spaces see Proposition 3.

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

In probabilistic terms, Gini's mean difference is $E|X-X'|$ where X' is an iid 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, on the other hand, 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,

i Consistent one-sample goodness-of-fit tests (Székely & Rizzo 2005a; Rizzo 2009; Yang 2012).

ii Consistent multi-sample tests of equality of distributions (Székely & Rizzo 2004; Rizzo 2002, 2003; Baringhaus & Franz 2004).

iii Hierarchical clustering algorithms (Székely & Rizzo 2005b) that extend and generalize the Ward's minimum variance algorithm.

iv Distance components (*DISCO*) (Rizzo & Székely 2010), a nonparametric extension of analysis of variance for structured data.

v Characterization and test for multivariate independence (Feuerverger 1993; Székely, Rizzo & Bakirov 2007; Székely & Rizzo 2009).

vi Change point analysis based on Székely & Rizzo (2004) (Kim et al. 2009; Matteson & James 2012)

vii 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 2014; R Core Team 2015).

3. Why is Energy Distance Special?

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

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

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

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

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

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

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

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

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

Interestingly, this weight function appears in Feuerverger (1993), where it is applied for testing bivariate dependence. Although this singular weight function is "special" from the equivariance point of view, other weight functions are also applied in tests based on characteristic functions. See e.g. Gurtler & Henze (2000), Henze & Zirkler (1990), 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 $x_1, \dots, x_n$ is drawn. The energy distance for the goodness-of-fit test $H_0 : F = F_0$ vs $H_1 : F \neq F_0$ is

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

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

$\mathcal{E}_n$ 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 (5.1), 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 (4.1) 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

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

where $X_{(k)}$ denotes the k-th 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

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

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 (4.1) and $n\mathcal{E}_n$.

4.2. Energy test of normality

The energy test for multivariate normality is 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 (4.1) 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 EDF 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

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

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 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 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 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,

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

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

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

where

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

See Székely & Rizzo (2005a) for details. One way to obtain the expression for $E|a - Z|_d$ follows from a result (Zacks 1981, p. 55), which states that if $Z \in R^d$ is standard normal, then the variable $|a - Z|_d^2$ has a noncentral chisquare distribution $\chi^2[\nu; \lambda]$ with degrees of freedom $\nu = d + 2\psi$, and noncentrality parameter $\lambda = |a|_d^2/2$, where ψ is a Poisson random variable with mean λ. For implementation, the sum in $E|a - Z|_d$ converges after 40 to 60 terms except for very large $|a|_d$, while if $|a|_d$ is large the limit $E|a - Z|_d \approxeq |a|_d$ can be applied. See the source code in "energy.c" of the *energy* package (Rizzo & Székely 2014) 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 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 to 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. Theory for the case of estimated parameters is derived in Rizzo (2002) and Székely & Rizzo (2005a).

The energy test of multivariate normality is applicable for arbitrary dimension (not constrained by sample size) and it 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); Székely & Rizzo (2005a) to illustrate the power of the test of multivariate normality against various alternatives, compared with competing tests like 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.

5. Kinetic Energy: the Schrödinger Equation of Goodness-of-fit

The one sample energy statistic $\mathcal{E}_n$ in (4.1) is a V-statistic whose kernel is a double centered potential function defined by

$$h(x, y) = E|x - X| + E|y - X| - E|X - X'| - |x - y|, \tag{5.1}$$

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 $n \to \infty$.

On the other hand, if H_0 is true, then the kernel h is degenerate; that is, $E[h(x, X)] = 0$ for almost all $x \in R^d$. Thus $n\mathcal{E}_n$ has a finite limit distribution under the extra condition $E[h^2(X, X')] < \infty$ (see Serfling (1980) or (Koroljuk & 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

$$Q = \sum_{k=1}^{\infty} \lambda_k Z_k^2 \tag{5.2}$$

of iid 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

$$\int_{R^d} h(x, y)\psi(y)\, dF(y) = \lambda\psi(x). \tag{5.3}$$

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

The kernel h is symmetric ($h(x, y) = h(y, x)$), hence the eigenvalues are real. Since $|x-y|$ is conditionally negative definite, one can easily see that $h(x, y)$ is positive semidefinite, and thus all eigenvalues in (5.3) are nonnegative. It is also known that their sum is finite and equal to $E|X - X'|$.

The kernel h is degenerate; that is,

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

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

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

For such a ψ in (5.3) the y-independent terms in $h(x, y)$ integrate to 0 and thus (5.3) simplifies to

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

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

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

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

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

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

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

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

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

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

$$-\frac{1}{2f}\psi'' = \mu\psi. \tag{5.4}$$

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 (5.4) corresponds to pure kinetic energy because the $V(x)\psi(x)$ term is missing in (5.4). We can thus call μ in (5.4) the *statistical kinetic energy level*.

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

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

where $\delta(\cdot)$ is the Dirac delta function, so in one dimension $|x-y|$ is a harmonic function.

6. Multi-sample 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

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

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 see Székely & Rizzo (2004); Rizzo (2003); 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 symmetry

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

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

and equality holds if and only if X is diagonally symmetric. We can thus introduce *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*

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

Distance skewness has the property that $0 \leq \text{dSkew}(X) \leq 1$, 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:

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

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

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

$$= \frac{\sum_{i,j=1}^n |X_i - Y_j| - \sum_{i,j=1}^n |X_i + Y_j|}{2 \sum_{i,j=1}^n |X_i|}. \tag{6.3}$$

Observe that T_n is a normalized two-sample energy statistic because the numerator of the fraction in (6.2) is exactly half the sample energy distance of X and $-X$. The numerator in (6.3) is proportional to (6.1) for samples X_i and $Y_i = -X_i'$, and its expected value is $2E|X|$.

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 in Mensheinin & Zubkov (2012).

Note that if C is an orthonormal matrix, that is, $CC^T = C^TC = I_d$ 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 multi-sample 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 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

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

for $0 < \alpha \leq 2$. ($g_\alpha(A, B)$ is a type of Gini mean distance.) The multi-sample 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

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

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

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

and the between-sample energy statistic

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

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 (6.7) 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 multi-sample test of equal distributions. It is important to note that we have a different result if $\alpha = 2$, because in that case 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 (2014). Examples and power comparisons are given in Rizzo & Székely (2010). An interesting application of DISCO analysis in behavioral biology is given in 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

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

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

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

where

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

Other types of distances can replace the Euclidean distance in (6.9) to obtain other clustering algorithms. In fact, if we substitute squared Euclidean distances for Euclidean distances in (6.9), with the same parameters, we obtain the updating formula for Ward's minimum variance method (6.10). 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(ii), the exponent $\alpha = 2$ only characterizes differences in cluster means. By Proposition 2(i) if we apply any exponent $0 < \alpha < 2$, the cluster distance measures differences between distributions. Thus, we can replace Euclidean distances in (6.9) 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 like centroid, median, or Ward, is the ability to separate and identify clusters with equal or nearly equal centers. Simulation studies in 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). While 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 dependence coefficients *distance covariance* and *distance correlation* introduced in Székely, Rizzo & Bakirov (2007) that measure all types of dependence between random vectors X and Y in arbitrary dimension. The corresponding energy statistics have simple computing formulae, and they apply to sample sizes $n \geq 2$ (n can be much smaller than dimension).

To quote Newton (2009)

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

The distance covariance statistic is computed as follows. First we compute all of the pairwise distances between sample observations of the X sample, to get a distance matrix. Similarly compute a distance matrix for the Y sample. Next, we center the entries of these distance matrices so that their row and column means are equal to zero. A very simple formula (7.9) accomplishes the centering. Now take the centered distances $A_{k\ell}$ and $B_{k\ell}$ and compute the sample distance covariance as the square root of

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

The statistic $\mathcal{V}_n$ converges almost surely to distance covariance (dCov), $\mathcal{V}(X, Y)$, to be defined below, which is always non-negative 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) and 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. For other recent consistent tests of bivariate or multivariate independence see e.g. Feuerverger (1993); 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 (2.7).

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

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|\hat{f}_{X,Y}(t,s) - \hat{f}_X(t)\hat{f}_Y(s)\|^2 \\ &= \frac{1}{c_p c_q}\int_{R^{p+q}} \frac{|\hat{f}_{X,Y}(t,s) - \hat{f}_X(t)\hat{f}_Y(s)|^2}{|t|_p^{1+p}\,|s|_q^{1+q}}\,dt\,ds. \end{aligned} \tag{7.1}$$

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

$$\begin{aligned}\mathcal{V}^2(X,Y) = E[|X-X'|_p|Y-Y'|_q] + E|X-X'|_p\,E|Y-Y''|_q \\ - 2E[|X-X'|_p|Y-Y''|_q],\end{aligned} \tag{7.2}$$

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

Distance variance is defined as the square root of

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

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

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

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\,\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0; \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases} \tag{7.3}$$

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 (7.2)) 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 2- 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 $\mathcal{V}(X + Y) \le \mathcal{V}(X) + \mathcal{V}(Y)$. 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, non-zero 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 $dCor^2(X, Y) \to 1/\sqrt{2} = 0.7071....$ We conjecture that this is the best upper bound. The fact that for uncorrelated random variables $dCor < 1$ is a very good property of dCor compared to some other measures of dependencies like 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 (7.10). 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, Rizzo & Bakirov 2007, Theorem 1) is the following:

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

$$\|\hat{f}_{X,Y}^n(t, s) - \hat{f}_X^n(t)\hat{f}_Y^n(s)\|^2 = S_1 + S_2 - 2S_3, \tag{7.4}$$

where

$$S_1 = \frac{1}{n^2}\sum_{k,\ell=1}^{n} |X_k - X_\ell|_p |Y_k - Y_\ell|_q, \tag{7.5}$$

$$S_2 = \frac{1}{n^2}\sum_{k,\ell=1}^{n} |X_k - X_\ell|_p \frac{1}{n^2}\sum_{k,\ell=1}^{n} |Y_k - Y_\ell|_q, \tag{7.6}$$

$$S_3 = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{\ell,m=1}^{n} |X_k - X_\ell|_p |Y_k - Y_m|_q, \tag{7.7}$$

and

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = S_1 + S_2 - 2S_3, \tag{7.8}$$

where $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$ is given by (7.10) defined below.

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

$$A_{k\ell} = a_{k\ell} - \bar{a}_{k\cdot} - \bar{a}_{\cdot\,\ell} + \bar{a}_{\cdot\cdot}\,, \qquad k, \ell = 1, \ldots, n, \tag{7.9}$$

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

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,\,\ell=1}^{n} A_{k\ell} B_{k\ell}\,. \tag{7.10}$$

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, Rizzo & Bakirov 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, Rizzo & Bakirov 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.

On properties of distance covariance, distance variance, and distance correlation of see Székely, Rizzo & Bakirov (2007).

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

The first paper on distance correlation, Székely, Rizzo & Bakirov (2007), contains its 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 see Rizzo (2009) and Yang (2012).

For time series, definitions of an "auto" distance correlation analogous to auto correlation have been considered by Matteson & Tsay (2011), Rémillard (2009), and Zhou (2012). On martingale difference sequences see Shao & Zhang (2014) and Park et al. (2015).

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 (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 due to the fact that 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 Hou & Székely (2015) 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 Hou & Székely (2015) include an implementation in Matlab.

On a bias corrected distance correlation and its application to a t-test of independence see Székely & Rizzo (2013a).

7.3. Distance correlation for dissimilarity matrices; 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. In Székely & Rizzo (2014) an algorithm is outlined, and this 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 is developed to define *partial distance correlation* analogous to (linear) partial correlation. There is a simple computing formula for the pdCor 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 in Székely & Rizzo (2014) and currently implemented in an R package *pdcor* Rizzo & Székely (2015).

7.5. Alternative definitions of distance covariance; affine and monotone invariant versions

The original distance covariance has been defined as the square root of $dCov^2(X, Y)$ rather than the squared coefficient itself. Thus $dCov(X, Y)$ 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: $dCov^2(X, Y)$. 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, finally 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 Gini's mean difference.

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

$$\mathcal{R}^*_n(X, Y; Z) = \begin{cases} \frac{\mathcal{R}_n(X,Y) - \mathcal{R}_n(X,Z)\mathcal{R}_n(Y,Z)}{\sqrt{1-\mathcal{R}^2_n(X,Z)}\sqrt{1-\mathcal{R}^2_n(Y,Z)}}, & \mathcal{R}_n(X, Z) \neq 1 \text{ and } \mathcal{R}_n(Y, Z) \neq 1; \\ 0, & \mathcal{R}_n(X, Z) = 1 \text{ or } \mathcal{R}_n(Y, Z) = 1. \end{cases} \tag{7.11}$$

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).

Also see Dueck et al. (2014) 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".* On a test of independence based on ranks of distances see 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 will see 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, define the U-centered version of X:

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

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

$$\begin{aligned} \mathcal{W}^2(X,Y) &= E[X_W X'_W Y_{W'} Y'_{W'}] \\ &= E|X - X'||Y - Y'| + E|X - X'|E|Y - Y'| \\ &\quad - E|X - X'||Y - Y''| - E|X - X''||Y - Y'|, \end{aligned} \tag{7.12}$$

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

If we compare Theorem 2 and (7.2), 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 non-random process, the identity function, while 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 $0 < H = \alpha/2 < 1$, resp (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, $dCor_\alpha$, 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 see Genovese (2009). For more details on these kinds of generalized energy inferences see Section 8.

8. Generalized Energy Distance

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

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

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

$$\begin{aligned} \mathcal{E}^{(\alpha)}(X,Y) &= 2E|X - Y|^\alpha - E|X - X'|^\alpha - E|Y - Y'|^\alpha \\ &= \frac{1}{C(d,\alpha)} \int_{R^d} \frac{|\hat{f}(t) - \hat{g}(t)|^2}{|t|^{d+\alpha}}\, dt, \end{aligned} \tag{8.1}$$

where

$$C(d,\alpha) = 2\pi^{d/2} \frac{\Gamma(1 - \alpha/2)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)}. \tag{8.2}$$

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

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

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, Rizzo & Bakirov 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*

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

where (t, x) represents inner product, $C(d, \alpha)$ is the constant (8.2) defined in Proposition 2, $t \in R^d$. (The integrals at $t = 0$ and $t = \infty$ are meant in the principal value sense: $\lim_{\varepsilon \to 0} \int_{R^d \setminus \{\varepsilon B + \varepsilon^{-1} \overline{B}\}}$, where B is the unit ball (centered at 0) in R^d and $\overline{B}$ is the complement of B.)

A proof of Lemma 1 is given in Székely & 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 in Schoenberg (1938a) and Schoenberg (1938b). 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 μ_i, $i = 1, 2$ on $\mathcal{X}$ have finite first moments; that is, $\int \delta(o, x) d\mu_i(x) < \infty$ 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:

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

On finite metric spaces of strictly negative type see 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 for a countably infinite metric space that is strictly negative definite but not strongly. 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, see 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, weighted trees have negative type (Meckes 2013, Theorem 3.6). Another open problem is the characterization of Riemannian spaces that have strong negative type. On this topic, 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 iid copy of X and Y' is an iid copy of Y.*

1. *A necessary and sufficient condition that*

$$2E\delta(X, Y) - E\delta(X, X') - E\delta(Y, Y') \geq 0 \tag{9.1}$$

   holds for all X, Y is that $(\mathcal{X}, \delta)$ has negative type.
2. *In (9.1), 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) proves 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. A disadvantage of these kernels is that if we apply them in high dimensions then the energy distance, which is the same as the "maximum mean discrepancy" in the world of machine learning, tends to 0 exponentially fast as $d \to \infty$ (Ramdas et al. 2014; Reddi et al. 2014). In contrast, the energy distance with our original conditionally negative definite distance $|x-y|$ typically tends to infinity at a rate of $\sqrt{d}$. This huge difference makes our original energy distance more appropriate for high dimensional problems. The following example illustrates the situation.

Let X, Y be d-dimensional independent Gaussian variables with iid coordinates whose expectations are 0, and whose variance is σ^2 in case of X, and τ^2 for Y. Then $2E|X-Y|_d - E|X-X'|_d - |Y-Y'|_d \to \infty$ at rate $\sqrt{d}$ because normalized $|X-Y|, |X-X'|$, and $|Y-Y'|$ are χ_d-distributed after we divide them by $\sqrt{\sigma^2+\tau^2}, \sqrt{2}\sigma$, and $\sqrt{2}\tau$, respectively. Thus,

$$2E|X-Y|_d - E|X-X'|_d - E|Y-Y'|_d = \left[2\sqrt{\sigma^2+\tau^2} - \sqrt{2}(\sigma+\tau)\right] \frac{\sqrt{2}\Gamma((d+1)/2)}{\Gamma(d/2)},$$

because the expected value of the χ_d-distribution is $\sqrt{2}\,\Gamma((d+1)/2)/\Gamma(d/2)$, and asymptotically this is a constant times $\sqrt{d}$.

Another advantage of the kernel $|x-y|$ and the corresponding energy distance is that they are scale equivariant, and this helps to make scale invariant statistical decisions.

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 and Rizzo (2007)). The energy perspective in this paper 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 and 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- 174)). 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 by Gauss a "beautiful theorem of probability theory" ("Schönes Theorem der Wahrscheinlichkeitrechnung" (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). On this inequality see more details 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. See also Mattner (1997) and Morgenstern (2001) from the German school. On early applications see the test of bivariate independence of Feuerverger (1993), and the test of homogeneity of Baringhaus & Franz (2004). See also historical comments on "Hoeffding-type inequalities" and their generalizations in (Gneiting & Raftery 2007, Sec. 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, like 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 like 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. These energy minimum estimators can be called $\mathcal{E}$-estimators. 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 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 is proposed by Chakraborty & Bhattacharjee (2015).
5. Instead of the data distance of this paper

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

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. While 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. On Wasserstein type distances see (Ajtai et al. 1984), (Talagrand 2005), and (Major 2013). For a first start see Ramdas et al. (2015).

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.

ACKNOWLEDGMENTS

Research of the first author supported by the National Science Foundation, while working at the Foundation.

LITERATURE CITED

Aaronson, J., Burton, R., Dehling, H., Gilat, D., Hill, T. and Weiss, B. 1996. Strong laws for L- and U-statistics. *Trans. Amer. Math. Soc.*, **348**, 2845–2865.

Ajtai, M., Komlós, J., and Tusnády, G. 1984. On optimal matchings, *Combinatorica* **4**/4, 259–264.

Baringhaus, L. and Franz, C. 2004. On a new multivariate two-sample test, *Journal of Multivariate Analysis* **88**, 190–206.

Baringhaus, L. and Franz, C. 2010. Rigid motion invariant two-sample tests. *Statistica Sinica*, **20**, 1333–1361.

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

Berg, C., Christensen, J. P. R. and Ressel, P. 1984. *Harmonic analysis on semigroups. Theory of positive definite and related functions*, Graduate Texts in Mathematics, vol. 100. Springer-Verlag, Berlin-Heidelberg-New York.

Brillouin, L. 2004. *Science and Information Theory*, 2nd ed. Dover.

Buja, A., Logan, B. F., Reeds, J. A., and Shepp, L. A. 1994. Inequalities and Positive Definite Functions Arising from a Problem in Multidimensional Scaling, *Ann. Stat.* **22**/1, 406–438.

Cailliez, F. 1983. The analytical solution of the additive constant problem. *Psychometrika*, **48**, 343–349.

Chakraborty, S. and Bhattacharjee, A. Distance Correlation Measures Applied to Analyze Relation between Variables in Liver Cirrhosis Marker Data *International Journal of Collaborative Research on Internal Medicine and Public Health*, **7**/11, 1–7.

Cramér, H. 1928. On the composition of elementary errors: II, Statistical applications. *Skand. Akt.*, **11**, 141–180.

Deza, M.M. and Laurent, M. 1997. *Geometry of Cuts and Metrics*, volume 15 of Algorithms and Combinatorics. Springer-Verlag, Berlin.

Dueck, J., Edelmann, D. Gneiting, T. and Richards, D. 2014. The affinely invariant distance correlation, *Bernoulli*, **20**/4, 2305–2330.

Einstein, A. 1905. Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig? *Annalen der Physik* **18**/13, 639–643.

Fan, J., Feng, Y., and Xia, L. 2015. A Conditional Dependence Measure with Applications to Undirected Graphical Models, arXiv preprint arXiv:1501.01617.

Feragen,A., Lauze, F., and Hauberg, S. 2015. Geodesic exponential kernels: When curvature and linearity conflict. *In IEEE Conference on Computer Vision and Pattern Recognition*, CVPR.

Ferenci, T., Körner, A., and Kovács, L. 2015. The interrelationship of HbA1c and real-time continuous glucose monitoring in children with type 1 diabetes, *Diabetes Research and Clinical Practice* **108**/1, 38–44.

Feuerverger, A. 1993. A consistent test for bivariate dependence. Int. Statist. Rev, **61**, 419–433.

Fischer, H. 2011. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*, Springer.

Gel'fand, I. M. and Shilov, G. E. 1964. *Generalized Functions*, Volume I: Properties and Operations, Academic Press, New York (translation by E. Salatan of the Russian edition of 1958).

Genovese, C. 2009. Discussion of: Brownian distance covariance, *Annals of Applied Statistics*, **3**(4), 1299–1302. DOI: 10.1214/09-AOAS312G

Gini, C. 1912. *Variabilità e Mutabilità* Bologna: Tipografia di Paolo Cuppini.

Gneiting, T. and Raftery, A. E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association*, **102** (477), 359–378.

Gretton, A. and Györfi, L. 2010. Consistent nonparametric tests of independence, *Journal of Machine Learning Research* **11**, 1391–1423.

Gretton, A. and Györfi, L. 2012. Strongly consistent nonparametric test of conditional independence, *Statistics and Probability Letters* **82**, 1145–1150.

Guo, X., Zhang, Y., Hu, W., Tan, H., and Wang, X. 2014. Inferring nonlinear gene regulatory network from gene expression data based on distance correlation, *Plos One* February 14.

Gurtler, N. and Henze, N. 2000. Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Annals of the Institute of Statistical Mathematics*, **52**(2), 267–286.

Heisenberg, W. 1927. Über den anschaulichen Inhalt der quantummmechanischen Kinematic und Mechanic. *Zeitschrift für Physik* **43**, 172–198.

Heller, R., Heller, Y, and Gorfine, M. 2013. A consistent multivariate test of association based on ranks of distances, *Biometrika* **100**/2, 503-510.

Henze, N., and Zirkler, B. 1990. A class of invariant and consistent tests for multivariate normality. *Commun. Statist. Theory Methods* **19**, 3595–3617.

Herbin, E., Merzbach, E. 2007. *The multiparameter fractional Brownian motion. Math Everywhere*, 93–101, Springer, Berlin.

Hjorth, P., Lisoněk, P., Markvorsen, S., and Thomassen, C. 1998. Finite metric spaces of strictly negative type. *Linear Algebra Appl.*, **270** 255–273.

Hoeffding, W. 1948. A class of statistics with asymptotic normal distribution, *Annals of Mathematic Statistics*, **19**/3 293–325.

Horn, H. R. 1972. On necessary and sufficcient conditions for an infinitely divisible distribution to be normal or degenerate. *Z. Wahrscheinlichkeitstheorie v. Geb* **21**, 179–187.

Horváth, L. and Kokoszka, P. 2012. *Inference for Functional data with Applications*, Springer, New York.

Hua, W., Nichols, T.E., and Ghosh, D. 2015. Multiple comparision procedures for neuroimaging genomewide association studies, *Biostatistics* **16**/1, 17–30.

Hou, X. and Székely, G.J. 2015. Fast Computing for Distance Covariance, *Technometrics*, *Techno-*

metrics http://dx.doi.org/10.1080/00401706.2015.1054435.

Jaroszewitz, Sz, Lukasz, Z. 2015. Székely regularization for uplift modeling *in Challenges in Computational Statistics and Data Mining*, Vol 605 of the series *Studies in Computational Intelligence*, 135–154, Springer.

Josse, J. and Holmes, S. 2014. Measures of dependence between random vectors and tests of independence, http://arxiv.org/abs/1307.7383.

Jovanovič, I. and Stanič, Z. 2012. Spectral distances of graphs, *Linear Algebra Appl.* **436**, 1425-1435.

Jurečková, J. 1972. Nonparametric estimate of regression coefficient, *Annals of Mathematical Statistics* **42**, 1328–1338.

Kim, A. Y., Marzban, C., Percival, D. B., and Stuetzle, W. 2009. Using labeled data to evaluate change detectors in a multivariate streaming environment, *Signal Processing*, **89**/12, 2529–2536.

Klebanov, L. 2005. $\mathcal{N}$*-Distances and Their Applications*. Charles University, Prague.

Kong, J., Klein, B. E. K., Klein, R., Lee, K., and Wahba, G. 2012. Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality, *Proc. of the National Acad. of Sciences*, **109**/50, 20352–20357. DOI:10.1073/pnas.1217269109

Koroljuk, V. S. and Borovskich, Yu. V. 1994. *Theory of U-statistics*, Mathematics and its Applications. 273 (Translated by P. V. Malyshev and D. V. Malyshev from the 1989 Russian original ed.). Dordrecht: Kluwer Academic Publishers Group.

Li, R., Zhong, W. and Zhu, L. 2012. Feature Screening via Distance Correlation Learning, *Journal of the American Statistical Asssociation*, **107**/499, 1129–1139. DOI: 10.1080/01621459.2012.695654

Li, S. 2015. k-groups: A Generalization of k-means by Energy Distance, Ph.D. thesis, Bowling Green State University.

Lovász, L. 2012. *Large Networks and Graph Limits* Colloquium Publications, Vol. 60, The American Mathematical Society.

Lu, J. Y., Peng, Y.X., Wang, M., Gu, S.J., and Zhao, M.X. 2015. Support vector machine combined with distance correlation learning for Dst forecasting during intense geomagnetic storms, *Planetary and Space Science* (to appear).

Lyons, R. 2013. Distance covariance in metric spaces, *Ann. Probab.* 41/5,3284–3305.

Lyons, R. 2015. Hyperbolic space has strong negative type, *Illinois J. Math* accepted for publication.

Major, P. 2013. *On estimation of multiple random integrals and U-statistics* Lectures Notes in Mathematics, Springer.

Mardia, K. V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530.

Matsui, M. and Takemura, A. 2005. Empirical characteristic function approach to goodness-of-fit tests for the Cauchy distribution with parameters estimated by MLE or EISE, *Annals of the Institute of Statistical Mathematics*, **57**/1, 183–199.

Matteson, D. S. and Tsay, R. S. 2011. Independent Component Analysis via Distance Covariance. http://www.stat.cornell.edu/~matteson/papers/MattesonTsay_dCovICA.pdf (preprint).

Matteson, D. S. and James, N. A. 2012. A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data, http://www.stat.cornell.edu/~matteson/papers/MattesonJames2012a.pdf (preprint).

Mattner, L. 1997. Strict negative definiteness of integrals via complete monotonicity of derivatives. *Transactions of the American Math. Soc.* **349**/8, 3321–3342.

Meckes, M. W. 2013. Positive definite metric spaces. Positivity, **17**/3, 733–757.

Mensheinin, D. O. and Zubkov, A. M. 2008. Properties of the Székely-Móri asymmetry criterion statistics in the case of binary vectors with independent components *Mathematical Notes* **91**/4, 62–72.

Morgenstern, D. 2001. Proof of a conjecture by Walter Deuber concerning the distance between points of two types in R^d. *Discrete Mathematics* **226**, 347–349.

Newton, M. A. 2009. Introducing the Discussion Paper by Székely and Rizzo, *Annals of Applied Statistics*, **3**(4), 1233–1235. DOI: 10.1214/09-AOAS34INTRO

Olkin, I. and Yitzhaki, S. 1992. Gini regression analysis, *International Statistical Review* **60**(2), 185–196.

Park, T., Shao, X., and Yao, S. 2015. Partial martingale difference correlation *E*lectron. J. Statist. **9**/1, 1492–1517.

Patra, R. K., Sen, B., and Székely, G. J. 2015. On a nonparametric notion of residual and its applications, *Statistics & Probability Letters* (to appear).

R Core Team 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Ramdas, A., Reddi, S.J., Póczos, B., Singh, A., and Wasserman, L. 2014. On the Decreasing Power of Kernel and Distance based Nonparametric Hypothesis Tests in High Dimensions, `http://repository.cmu.edu/cgi/viewcontent.cgi?article=1073&context=machine_learning`

Ramdas, A., Garcia, N., and Cuturi, M. 2015. On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests, eprint arXiv:1509.02237, 09/2015.

Reddi, S. J, Ramdas, A., Póczos, B., Singh, A., and Wasserman, L. 2014. On the High-dimensional Power of Linear-time Kernel Two-Sample Testing under Mean-difference Alternatives, `http://arxiv.org/abs/1411.6314`

Rémillard, B. 2009. Discussion of: 'Brownian distance covariance' *Annals of Applied Statistics*, Vol. 3, No. 4, 1295–1298. DOI: 10.1214/09-AOAS312F

Rizzo, M. L. 2002. A New Rotation Invariant Goodness-of-Fit Test, Ph.D. dissertation, Bowling Green State University.

Rizzo, M. L. 2003. A Test of Homogeneity for Two Multivariate Populations. 2002 Proceedings of the American Statistical Association, Physical and Engineering Sciences Section, Alexandria, VA: American Statistical Association.

Rizzo, M. L. 2009. New Goodness-of-Fit Tests for Pareto Distributions, *ASTIN Bulletin: Journal of the International Association of Actuaries*, **39**(2), 691–715.

Rizzo, M. L. and Székely, G. J. 2010. DISCO Analysis: A Nonparametric Extension of Analysis of Variance, *Annals of Applied Statistics*, **4**(2), 1034–1055.

Rizzo, M. L. and Székely, G. J. 2014. energy: E-statistics (energy statistics). R package version 1.6.1.

Rizzo, M. L. and Székely, G. J. 2015. pdcor: Partial distance correlation. R package version 1.0.2.

Rudas J., Guaje J., Demertzi A., Heine L., Tshibanda L., Soddu A., Laureys S., and Gomez F. 2014. A method for functional network connectivity using distance correlation. Conference Proceedings, Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society, 2793–2796.

Schilling, K., Oberdick, J, and Schilling, R. L. 2012. Toward an Efficient and Integrative Analysis of Limited-Choice Behavioral Experiments, *Journal of Neuroscience*, **32**(37), 12651–12656.

Schrödinger, E. 1926. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Physical Review* **28**(6), 1049–1070.

Schrödinger, E. 1944. *What is Life - the Physical Aspect of the Living Cell*, Cambridge University Press.

Schölkopf, B. and Smola, A. J. 2002. *Learning with Kernels*, MIT Press, Cambridge, MA.

Schoenberg, I. J. 1938. Metric spaces and completely monotone functions. *Ann. of Math.* (2), **39**/4, 811–841.

Schoenberg, I. J. 1938. Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, **44**/3, 522-536.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Sejdinovic, D, Sriperumbudur, B, Gretton, A., and Fukumizu, K. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing, *Ann. Statist.* **41**/5, 2263–2291.

Shamsuzzaman, Md., Satish, M., and Pintér, J. D. 2015. Distance correlation based nearly orthogonal space-filling experimental designs, *Int. J. Experimental Design and Process Optimization* **4**/3/4, 216–233.

Shao, X. and Zhang, J. 2014. Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association* 109/507, 1302–1318.

Sheng, W. and Yin, X. 2013. Direction estimation in single-index models via distance covariance *Journal of Multivariate Analysis* **122**, 148–161.

Sheng, W. and Yin, X. 2015. Sufficient dimension reduction via distance covariance *Journal of Computational and Graphical Statistics* (in press).

Simon, N. and Tibshirani, R.2011. Comment on "Detecting Novel Associations in Large Data Set" by Reshef et. al. `http://www-stat.stanford.edu/~tibs/reshef/comment.pdf`

Steutel, F, W. and van Harn, K. 2004. *Infinite Divisibility of Probability Distributions on the Real Line*, Marcel Dekker.

Székely, G. J. 1989. Potential and Kinetic Energy in Statistics, Lecture Notes, Budapest Institute of Technology (Budapest Technical University).

Székely, G. J. 1996. *Contests in Higher Mathematics*. Springer, New York.

Székely, G. J. (2003). E-statistics: Energy of Statistical Samples, Bowling Green State University, Department of Mathematics and Statistics Technical Report No. 03–05 (also technical reports by same title, from 2000–2003 and NSA grant # MDA 904-02-1-0091 (2000–2002).)

Székely, G. J. and Bakirov, N. K. 2003. Extremal probabilities for Gaussian quadratic forms. *Probab. Theory Related Fields* **126**, 184–202.

Székely, G. J., and Móri, T. F. 2001. A characteristic measure of asymmetry and its application for testing diagonal symmetry. *Communications in Statistics – Theory and Methods* **30** (8 & 9), 1633–1639.

Székely, G. J. and Rizzo, M. L. 2004. Testing for Equal Distributions in High Dimension, *InterStat*, November (5).

Székely, G. J. and Rizzo, M. L. 2005a. A new test for multivariate normality, *J. Multivariate. Anal.*, **93**/1, 58–80.

Székely, G. J. and Rizzo, M. L. 2005b. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method, *J. Classification*, **22**(2), 151–183.

Székely, G. J. and Rizzo, M. L. (2007). The uncertainty principle of game theory, *The American Mathematical Monthly* **114**/8, 688–702.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing independence by correlation of distances, *Ann. Statist.* **35**/6, 2769–2794.

Székely, G. J. and Rizzo, M. L. 2009. Brownian Distance Covariance, *Annals of Applied Statistics*, **3**/4, 1236–1265.

Székely, G. J. and Rizzo, M. L. 2012. On the uniqueness of distance covariance, *Statistics & Probability Letters*, **82**, 2278–2282.

Székely, G. J. and Rizzo, M. L. 2013. The distance correlation t-test of independence in high dimension, *Journal of Multivariate Analysis*, **117**, 193-213. DOI: 10.1016/j.jmva.2013.02.012

Székely, G. J. and Rizzo, M. L. 2013. Energy statistics: A class of statistics based on distances, *Journal of Statistical Planning and Inference* **143**/8, 1249–1272.

Székely, G. J. and Rizzo, M. L. 2014. Partial distance correlation with methods for dissimilarities, *Annals of Statistics*, **42**/6, 2382–2412.

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), *Zetschrift für Physik*, 53, 840–856. Based on Szilárd's 1922 Ph. D. dissertation in Berlin and on Szilard's 1925 paper in *Zetschrift für Physik*, **32**, 753.

Talagrand, M. 2005. *The general chaining* Springer Monographs in Mathematics, Springer.

Vaiciukynas, E, Verikas, A., Gelzinis, A., Bacauskiene, M. , and Olenina, I. 2015. Exploiting statistical energy test for comparison of multiple groups in morphometric and chemometric data, *Chemometrics and Intelligent Laboratory Systems* **146**, 10–23.

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

Varina, T., Bureaua, R., Muellerb, C. , and Willett, P. 2009. Clustering files of chemical structures

using the Székely–Rizzo generalization of Ward's method, *Journal of Molecular Graphics and Modelling* **28**/2, 187–195.

Villani, C. 2008. *Optimal transport: old and new* Springer.

Von Mises, R. 1947. On the Asymptotic Distributions of Differentiable Statistical Functionals, *Annals of Mathematical Statistics*, **2**, 209–348.

von Neumann, J. 1928. Theorie der Gesellschaftspiele. *Math. Annalen*, **100**, 295–320.

Wahba, G. 1990. *Spline Models for Observational Data*, volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

Wahba, G. 2014. Positive definite functions, Reproducing Kernel Hilbert Spaces and all that, The Fisher Lecture at JSM 2014. August 6, 2014, Boston, Massachusetts.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. 2015. Conditional distance correlation, *Journal of the American Statistical Association*, to appear.

Würtz, D., Chalabi, Y., Chen, W., and Ellis, A. 2009. *Portfolio Optimization with R/Rmetrics*, Rmetrics Association and Finance Online Publishing, Zurich, ISBN: 978-3-906041-01-8, `https://www.rmetrics.org/ebooks-portfolio`

Yang, G. 2012. The Energy Goodness-of-Fit Test for Univariate Stable Distributions, Ph.D. thesis, Bowling Green State University.

Yitzhaki, S. 2003. Gini's Mean difference: a superior measure of variability for non-normal distributions, *Metron* **61**, 285–316.

Zacks, S. 1981. *Parametric Statistical Inference: Basic Theory and Modern Approaches*. Pergamon, Oxford.

Zanoni, M, Setragno, F., and Sarti, A. 2014. The violin ontology, *Proc. of the 9th Conf. on Interdisciplinary Musicology – CIM14.* Berlin.

Zhou Z. 2012. Measuring Nonlinear Dependence in Time-Series, a Distance Correlation Approach, *Journal of Time Series Analysis*, **33**/3, 438–457.