Advanced Review

Energy distance

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

Energy distance is a metric that measures the distance between the distributions of random vectors. Energy distance is zero if and only if the distributions are identical, thus it characterizes equality of distributions and provides a theoretical foundation for statistical inference and analysis. Energy statistics are functions of distances between observations in metric spaces. As a statistic, energy distance can be applied to measure the difference between a sample and a hypothesized distribution or the difference between two or more samples in arbitrary, not necessarily equal dimensions. The name energy is inspired by the close analogy with Newton's gravitational potential energy. Applications include testing independence by distance covariance, goodness-of-fit, nonparametric tests for equality of distributions and extension of analysis of variance, generalizations of clustering algorithms, change point analysis, feature selection, and more. © 2015 Wiley Periodicals, Inc.

How to cite this article:
WIREs Comput Stat 2016, 8:27–38. doi: 10.1002/wics.1375

Keywords: Multivariate, goodness-of-fit, distance correlation, DISCO, independence

INTRODUCTION

Energy distance is a distance between probability distributions. The name 'energy' is motivated by analogy to the potential energy between objects in a gravitational space. The potential energy is zero if and only if the location (the gravitational center) of the two objects coincide, and increases as their distance in space increases. One can apply the notion of potential energy to data as follows. Let X and Y be independent random vectors in $\mathbb{R}^d$, with cumulative distribution function (CDF) F and G, respectively. In what follows, $\| \cdot \|$ denotes the Euclidean norm (length) of its argument, $\mathbb{E}$ denotes expected value, and a primed random variable X' denotes an independent and identically distributed (iid) copy of X; that is, X and X' are iid. Similarly, Y and Y' are iid. The squared energy distance can be defined in terms of expected distances between the random vectors

$$D^2(F,G) := 2\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| \\ - \mathbb{E}\|Y-Y'\| \geq 0,$$

and the energy distance between distributions F and G is defined as the square root of $D^2(F,G)$.

It can be shown that energy distance $D(F,G)$ satisfies all axioms of a metric, and in particular $D(F,G) = 0$ if and only if $F = G$. Therefore, energy distance provides a characterization of equality of distributions, and a theoretical basis for the development of statistical inference and multivariate analysis based on Euclidean distances. In this review we discuss several of the important applications and illustrate their implementation.

*Correspondence to: mrizzo@bgsu.edu

[1] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, USA

[2] National Science Foundation, Arlington, VA, USA

[3] Rényi Institute of Mathematics, Hungarian Academy of Sciences, Hungary

Conflict of interest: The authors have declared no conflicts of interest for this article.

BACKGROUND AND APPLICATION OF ENERGY DISTANCE

The notion of 'energy statistics'[1–3] was introduced by Székely in 1984–1985 in a series of lectures given in Budapest, Hungary, and at MIT, Yale, and Columbia. As mentioned above, Székely's main idea and the name derive from the concept of Newton's potential energy. Statistical observations can be considered as objects in a metric space that are governed by a statistical potential energy that is zero if and only if an underlying statistical null hypothesis is true. Energy statistics (E-statistics) are a class of functions of distances between statistical observations. Several examples of one-sample, two-sample, and multi-sample energy statistics will be illustrated below.

Cramér's distance[4] is closely related, but only in the univariate (real valued) case. For two real-valued random variables with CDFs F and G, the squared energy distance is exactly twice the distance proposed by Harald Cramér:

$$D^2(F,G) = 2\int_{-\infty}^{\infty} (F(x) - G(x))^2 \mathrm{d}x.$$

However, the equivalence of energy distance with Cramer's distance cannot extend to higher dimensions, because while energy distance is rotation invariant, Cramér's distance does not have this property.

A proof of the basic energy inequality, $D(F,G) \geq 0$ with equality if and only if $F = G$ follows from Ref 5 and also from Mattner's result.[6] An alternate proof related to a result of Morgenstern[7] appears in Refs 8,9.

Application to testing for equality of two distributions appeared in Refs 8,10–12 as well as a multi-sample test for equality of distributions 'distance components' (DISCO).[13] Goodness-of-fit tests have been developed for multivariate normality,[8,9] stable distribution,[14] Pareto distribution,[15] and multivariate Dirichlet distribution.[16] Hierarchical clustering and a generalization of k-means clustering based on energy distance are developed in Refs 17,18.

Generalizations and interesting special cases of the energy distance have appeared in the recent literature; see Refs 19–22. A similar idea related to energy distance and E-statistics were considered as N-distances and N-statistics in Ref 23; see also Ref 5. Measures of the energy distance type have also been studied in the machine learning literature.[21]

TESTING FOR EQUAL DISTRIBUTIONS

Consider the null hypothesis that two random variables, X and Y, have the same cumulative distribution functions: $F = G$. For samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ from X and Y, respectively, the E-statistic for testing this null hypothesis is

$$\mathcal{E}_{n,m}(X,Y) := 2A - B - C,$$

where A, B, and C are simply averages of pairwise distances:

$$A = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \| x_i - y_j \|, \; B = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \| x_i - x_j \|,$$
$$C = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} \| y_i - y_j \|.$$

One can prove[2,9] that $\mathcal{E}(X, Y) := D^2(F, G)$ is zero if and only if X and Y have the same distribution ($F = G$). It is also true that the statistic $\mathcal{E}_{n,m}$ is always non-negative. When the null hypothesis of equal distributions is true, the test statistic

$$T = \frac{nm}{n+m}\mathcal{E}_{n,m}(X,Y).$$

converges in distribution to a quadratic form of independent standard normal random variables. Under an alternative hypothesis the statistic T tends to infinity stochastically as sample sizes tends to infinity, so the energy test for equal distributions that rejects the null for large values of T is consistent.[12]

Because the null distribution of T depends on the distributions of X and Y, the test is implemented as a permutation test in the *energy* package,[24] which is available for R[25] on the Comprehensive R Archive Network (CRAN) under general public license. The test is implemented in the function `eqdist.etest`. The data argument can be the data matrix or distance matrix of the pooled sample, with the default being the data matrix. The second argument is a vector of sample sizes. Here we test whether two species of iris data differ in the distribution of their four dimensional measurements. In this example, $n = m = 50$ and we use 999 permutation replicates for the test decision.

```
> library(energy)
> eqdist.etest(iris[1:100, 1:4], c(50, 50), R = 999)

  Multivariate 2-sample E-test of equal distributions

data:  sample sizes 50 50, replicates 999
E-statistic = 123.5538, p-value = 0.001
```

To compute the energy statistic only:

```
> eqdist.e(iris[1:100, 1:4], c(50,50))
E-statistic
   123.5538
```

The *E*-statistic is not standardized, so it is reported only for reference. To interpret the value, one should normalize the statistic. One way of doing this is to divide by an estimate of $\mathbb{E} \| X - Y \|$. Note that if

$$H := \frac{D^2(F_X, F_Y)}{2\mathbb{E} \| X - Y \|} = \frac{2\mathbb{E} \| X - Y \| - \mathbb{E} \| X - X' \| - \mathbb{E} \| Y - Y' \|}{2\mathbb{E} \| X - Y \|},$$

then $0 \le H \le 1$ with $H = 0$ if and only if X and Y are identically distributed.

For background on permutation tests, see Ref 26 or Ref 27.

For more details, applications, and power comparisons see Refs 11,12 and the documentation included with the *energy* package. The same functions are generalized to handle multi-sample problems, discussed below.

MULTI-SAMPLE ENERGY STATISTICS

Distance Components: A Nonparametric Extension of ANOVA

Analogous to the ANOVA decomposition of variance, we partition the total dispersion of the pooled samples into between and within components, called distance components (DISCO). For two samples, the between-sample component is the two-sample energy statistic above. For several samples, the between component is a weighted combination of pairwise two-sample energy statistics.

To test the K-sample hypothesis $H_0 : F_1 = \ldots = F_K$, $K \ge 2$, one can apply the between-sample test statistic, or alternately a ratio statistic similar to the familiar F statistic of ANOVA, both implemented as a randomization test.

First, let us introduce a generalization of the energy distance. The characterization of equality of distributions by energy distance also holds if we replace Euclidean distance by $\| X - Y \|^{\alpha}$, where $0 < \alpha < 2$. The characterization does not hold if $\alpha = 2$ because $2\mathbb{E} \| X - Y \|^2 - \mathbb{E} \| X - X' \|^2 - \mathbb{E} \| Y - Y' \|^2 = 0$ whenever $\mathbb{E}X = \mathbb{E}Y$. We denote the corresponding two-sample energy statistic by $\mathcal{E}_{n,m}^{(\alpha)}$.

Let $A = \{a_1, \ldots, a_{n_1}\}$, $B = \{b_1, \ldots, b_{n_2}\}$ be two samples, and define

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

for $0 < \alpha \le 2$. If $A_1, \ldots, A_K$ are samples of sizes $n_1, n_2, \ldots, n_K$, respectively, and $N = \sum_{j=1}^{K} n_j$, the within-sample dispersion statistic is

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

and the total dispersion of the observed response is

$$T_\alpha = T_\alpha(A_1, \ldots, A_K) = \frac{N}{2} g_\alpha(A, A), \qquad (3)$$

where A is the pooled sample. The between-sample energy statistic is

$$\begin{aligned} S_{n,\alpha} &= \sum_{1 \le j < k \le K} \left(\frac{n_j + n_k}{2N} \right) \left[\frac{n_j n_k}{n_j + n_k} \mathcal{E}_{n_j, n_k}^{(\alpha)}(A_j, A_k) \right] \\ &= \sum_{1 \le j < k \le K} \left\{ \frac{n_j n_k}{2N} \left(2 g_\alpha(A_j, A_k) - g_\alpha(A_j, A_j) - g_\alpha(A_k, A_k) \right) \right\}. \end{aligned} \qquad (4)$$

Note that $S_{n,\alpha}$ weights each pairwise E-statistic based on the proportion of data in the two samples. Analogous to the decomposition of between-sample variance and within-sample variance of ANOVA, we have the decomposition of distances

$$T_\alpha = S_\alpha + W_\alpha,$$

where both S_α and W_α are nonnegative.

For every $0 < \alpha < 2$, the statistic Eq. (4) determines a consistent test of the multi-sample hypothesis of equal distributions.[13] In the special case where all F_j are univariate distributions and $\alpha = 2$, $S_{n,2}$ is the ANOVA between sample sum of squared error and the decomposition $T_2 = S_2 + W_2$ is the ANOVA decomposition. The ANOVA test statistic measures differences in means, not distributions. However, if we apply $\alpha = 1$ (Euclidean distance) or any $0 < \alpha < 2$ as the exponent on Euclidean distance, the corresponding energy test is consistent against all alternatives with finite α moments. If any of the underlying distributions may have non-finite first moment, a suitable choice of α extends the energy test to this situation.

Returning to the iris data example, we can easily apply the test for equality of the three species' distributions using a choice of methods, and the relevant options here are `method="discoB"` or `method="discoF"`. The first uses the between-sample statistic, and the second uses an 'F' ratio like ANOVA. Both are implemented as permutation tests.

```
> eqdist.etest(iris[, 1:4], c(50,50,50), method="discoB", R = 999)

  DISCO (Between-sample)

data:  x
DISCO between statistic = 119.2373, p-value < 2.2e-16
```

$$F_{n,\alpha} = \frac{S_{n,\alpha}/(K-1)}{W_{n,\alpha}/(N-K)},$$

and the decomposition details are displayed in a table similar to an ANOVA table. Although it has the same form as an F statistic, it does not have an F distribution and a permutation test is applied.

```
> eqdist.etest(iris[, 1:4], c(50,50,50), method="discoF", R = 999)
disco(x = x, factors = g, distance = distance, index = 1, R = R,
    method = method)

Distance Components: index  1.00
Source                 Df   Sum Dist   Mean Dist    F-ratio    p-value
factors                 2  119.23731    59.61865    124.597    0.001
Within                147   70.33848     0.47849
Total                 149  189.57579
```

One can obtain a table of pairwise energy statistics for any number of samples in one step using the `edist` function in the energy package. It can be used, for example, to display E-statistics for the result of a cluster analysis.

E-clustering

Energy distance has been applied in hierarchical cluster analysis.[17] It generalizes the well-known Ward's minimum variance method in a similar way that DISCO generalizes ANOVA. The DISCO decomposition has recently been applied to generalize k-means clustering.[18]

In an agglomerative hierarchical clustering algorithm, starting with single observations, at each step we merge clusters that have minimum cluster distance. In the energy distance algorithm, the cluster distance is the two sample energy statistic.

There is a general class of hierarchical clustering algorithms uniquely determined by their respective recursive formula for updating all cluster distances following each merge of two clusters. One can show[17] that the energy clustering algorithm is also a member of this class and its recursive formula shows that it it is formally similar to Ward's method.

Suppose that the disjoint clusters C_i, C_j are to be merged at the current step. If C_k is a disjoint cluster, then the new cluster distances can be computed by the following recursive formula:

$$\begin{aligned} d(C_i \cup C_j, C_k) = & \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) \\ & + \frac{n_j + n_k}{n_i + n_j + n_k} d(C_j, C_k) \\ & - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j), \end{aligned} \tag{5}$$

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 sizes of clusters C_i, C_j, C_k, respectively. Let $d_{ij} := d(C_i, C_j)$. Then

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

where

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

If we substitute squared Euclidean distances for Euclidean distances in this recursive formula, keeping the same parameters (α_i, α_j, β, γ), then we obtain the updating formula for Ward's minimum variance method. However, we know that Ward's method (with exponent $\alpha = 2$ on distances) is a geometrical method that separates clusters by their centers, not by their distributions. $\mathcal{E}$-clustering generalizes Ward because for every $0 < \alpha < 2$, the energy clustering algorithm separates clusters that differ in distribution.

Overall in simulations and real data examples[17] the characterization property of $\mathcal{E}$ is a clear advantage for certain clustering problems, without sacrificing the good properties of Ward's minimum variance method for separating spherical clusters.

The `hclust` hierarchical clustering function provided in R[25] implements exactly the above recursive formula Eq. (5), and therefore one can use `hclust` to apply either the $\mathcal{E}$-clustering solution or Ward's method by specifying method '`ward.D`' (energy) or '`ward.D2`' (Ward).

TESTING INDEPENDENCE

An important application for two samples applies to testing independence of random vectors. In this case, we test whether the joint distribution of X and Y is equal to the product of their marginal distributions. Interestingly, the statistics can be expressed in a product–moment expression involving the double-centered distance matrices of the X and Y samples. The statistics based on distances are analogous to, but more general than, product–moment covariance and correlation. This suggests the names *distance covariance* (dCov) and *distance correlation* (dCor), defined below.

Distance Covariance

The simplest formula for the *distance covariance* statistic is the square root of

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} \widehat{A}_{ij} \widehat{B}_{ij},$$

where $\hat{A}$ and $\widehat{B}$ are the double-centered distance matrices of the X sample and the Y sample, respectively, and the subscript ij denotes the entry in the i-th row and j-th column. The double-centered distance matrices are computed as in classical multidimensional scaling. Given a random sample $(x, y) = \{(x_i, y_i) : i = 1, \ldots, n\}$ from the joint distribution of random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$, compute the Euclidean distance matrix $(a_{ij}) = (\|x_i - x_j\|)$ for the X sample and $(b_{ij}) = (\|y_i - y_j\|)$ for the Y sample. The ij-th entry of $\hat{A}$ is

$$\widehat{A}_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}, \quad i, j = 1, \ldots, n,$$

where

$$\bar{a}_{i.} = \frac{1}{n} \sum_{j=1}^{n} a_{ij}, \quad \bar{a}_{.j}, = \frac{1}{n} \sum_{i=1}^{n} a_{ij}, \quad \bar{a}_{..} = \frac{1}{n^2} \sum_{i,j=1}^{n} a_{ij}.$$

Similarly, the ij-th entry of $\widehat{B}$ is

$$\widehat{B}_{ij} = b_{ij} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..}, \quad i, j = 1, \ldots, n.$$

The sample *distance variance* is

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X, X) = \frac{1}{n^2} \sum_{i,j=1}^{n} \widehat{A}_{ij}^2 .$$

The distance covariance statistic is always non-negative, and $\mathcal{V}_n^2(X) = 0$ only if all of the sample observations are identical (see Ref 28).

Distance Correlation

Distance correlation is the standardized distance covariance. We have defined the squared distance covariance statistic

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} \widehat{A}_{ij} \widehat{B}_{ij} ,$$

and the squared distance correlation is defined by

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

Distance correlation satisfies

1. $0 \le \mathcal{R}_n(X,\ Y) \le 1$.
2. If $\mathcal{R}_n(X,\ Y) = 1$ then there exists a vector a, a non-zero real number b and an orthogonal matrix R such that $Y = a + bXR$, for the data matrices X and Y.

Remark 1. One could also define dCov as $\mathcal{V}_n^2$ and dCor as $\mathcal{R}_n^2$ rather than by their respective square roots. There are reasons to prefer each definition, but historically[28,29] the above definitions were used. When we deal with unbiased statistics, we no longer have the non-negativity property, so we cannot take the square root and need to work with the square.

Population Coefficients

Suppose that $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, $\mathbb{E}\,\|X\| < \infty$, and $\mathbb{E}\,\|Y\| < \infty$. The squared population distance covariance coefficient can be written in terms of expected distances:

$$\mathcal{V}^2(X,Y) = \mathbb{E}\,\|X - X'\|\|\,Y - Y'\| + \mathbb{E}\,\|X - X'\| \cdot \mathbb{E}\,\|Y - Y'\| - 2\mathbb{E}\,\|X - X'\|\|\,Y - Y''\|,$$

where (X,Y), $(X',\ Y')$, and $(X'',\ Y'')$ are iid.[29] Here $\mathcal{V}^2(X,Y)$ is an energy distance between the joint distribution of (X,Y) and the product of the marginal distributions of X and Y.

We have a characterization of independence: $\mathcal{V}^2(X,Y) \ge 0$ with equality to 0 if and only if X and Y are independent. Population distance correlation $\mathcal{R}(X,\ Y)$ is the square root of the standardized coefficient:

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

The empirical coefficients $\mathcal{V}_n(X,Y)$ and $\mathcal{R}_n(X,\ Y)$ converge almost surely to the population coefficients $\mathcal{V}(X,Y)$ and $\mathcal{R}(X,\ Y)$, as $n \to \infty$. The distribution of $Q_n = n\mathcal{V}_n^2(X,Y)$ converges to a quadratic form $\sum_{i=1}^{\infty} \lambda_i Z_i^2$, where Z_i are iid standard normal and λ_i are non-negative coefficients that depend on the distributions of the underlying random variables. Values of Q_n near zero are consistent with the null hypothesis of independence, while large Q_n support the alternative. Thus, a consistent test of multivariate independence is based on the distance covariance $Q_n = n\mathcal{V}_n^2$, and it can be implemented as a nonparametric permutation test. The dCov test applies to random vectors in arbitrary, not necessarily equal dimensions, for any sample size $n \ge 4$. For high dimensional X and Y, there is also a distance correlation t-test of independence introduced in Ref 30 is applicable when dimension exceeds sample size.

For the permutation test in this case, the null distribution is sampled by permuting the indices of one of the two variables each time a sample is drawn. The replicates then provide a reference distribution for estimating the tail probability to the right of the observed test statistic Q_n.

Remark 2. Note that the permutation test is only applicable if the observations are exchangeable under the null, which would not be true e.g., for time series data.

The statistics and tests are implemented in the *energy* package for R.[24,25] Functions `dcor, dcov, dcov.test` and `dcor.ttest` compute the statistics and tests of independence. The following example uses the `crabs` data in the MASS package. After converting the binary factors to integers, we test if the two-dimensional categorical variable (species, sex) is independent of the vector of body measurements.

```
> library(MASS)
> sp <- as.integer(unlist(crabs$sp)) - 1   #species
> sx <- as.integer(unlist(crabs$sex)) - 1  #sex
> ss <- cbind(sp, sx)                      #species and sex
> x <- crabs[, 4:8]  #five measurements
```

The data arguments to dcov.test can be data matrices or distance objects returned by the R `dist` function. Here the arguments are data matrices.

```
> dcov.test(x, ss)

  dCov test of independence

data:  index 1, replicates 199
nV^2 = 71.6769, p-value = 0.005
sample estimates:
     dCov
0.5986521
```

The test is significant and we reject independence. One may also want to compute the statistics using `dcor` or `dcov`:

```
> dcor(x, ss)
[1] 0.2996931
```

To recover all of the statistics from dCor, a utility function is provided:

```
> unlist(DCOR(x, ss))
     dCov      dCor     dVarX     dVarY
0.5986521 0.2996931 7.6586679 0.5210054
```

An Unbiased Distance Covariance Statistic

The population distance covariance is zero under independence, but the dCov statistic is non-negative, hence its expected value is positive except in degenerate cases. Clearly the dCov statistic in its original formulation is biased for the population coefficient. The bias is in fact increasing with dimension.

An unbiased estimator of $\mathcal{V}^2(X, Y)$ was given in Ref. 30 and an equivalent unbiased statistic was given in Ref 31. Although the latter one looks simpler, the original may be faster to compute. The following is from Ref 31.

Let $A = (a_{ij})$ be a symmetric, real valued $n \times n$ matrix with zero diagonal (not necessarily Euclidean distances). Instead of the classical method of double centering A, we introduce the U-centered matrix $\tilde{A}$. The (i,j)-th entry of $\tilde{A}$ is

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

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 $\widetilde{B}$. Then

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

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.[31]

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. Returning to the example, here is a comparison of biased and bias-corrected $\mathcal{R}_n^2$:

```
> dcor(x, ss)^2
[1] 0.08981596
> bcdcor(x, ss)
[1] 0.07627263
```

The above unbiased inner product dCov statistic is easy to compute, but since it can take negative values, we cannot define the bias corrected statistic to be the square root of it. Thus we avoid the 'squared' notation and use the inner product operator or $\mathcal{V}_n^*$ and $\mathcal{R}_n^*$. One could have defined 'dCov' from the start to be the square of the energy distance. Historically, the rationale for choosing the square root definition is that in this case distance covariance is the energy distance between the joint distribution of the variables and the product of their marginals. A disadvantage is that the distance variance, rather than the distance standard deviation, is measured in the same units as the distances.

It is clear that both the sample dCov and the sample dCor can be computed in $O(n^2)$ steps. Recently Huo and Székely[32] proved that for real valued samples the unbiased estimator of the squared population distance covariance can be computed by an $O(n \log n)$ algorithm. The supplementary files to Ref. 32 include an implementation in Matlab.

Distance Correlation for Dissimilarity Matrices

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. In Ref 31, it is shown that the inner product version of dCov can be applied to any U-

centered dissimilarity matrix (zero-diagonal symmetic matrix). The algorithm is outlined in Ref 31, and it is a strong competitor of the Mantel test for association between dissimilarity matrices based on comparisons in the paper. This makes dCov tests and energy statistics ready to apply to problems in e.g., community ecology, where one must often work with data in the form of non-Euclidean dissimilarity matrices.

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 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 Ref. 31 and currently implemented in an R package *pdcor*,[33] which will become part of the *energy* package.

GOODNESS-OF-FIT

Goodness-of-fit is a one-sample problem, but there are two distributions to consider: one is the hypothesized distribution and the other is the underlying distribution from which the observed sample has been drawn. Energy distance applies to compare these two distributions with a variation of the two sample energy distance.

The energy distance for this problem must be the same as $\mathcal{E}(X, Y)$ where one of the variables now represents the unknown sampled distribution. Suppose that a random sample $x_1, \ldots, x_n$ is observed and the problem is to test whether the sampled distribution F is equal to the hypothesized distribution F_X. The energy goodness-of-fit statistic is

$$\mathcal{E}_n = n\left(\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\,\|x_i - X\|^{\alpha} - \mathbb{E}\,\|X - X'\|^{\alpha} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\|^{\alpha}\right), \tag{6}$$

where X and X' are iid with distribution F_X, and $0 < \alpha < 2$. The statistic is defined in arbitrary dimension and is not restricted by sample size. The only required condition is that ||X|| has finite α moment under the null hypothesis. Under the null hypothesis $\mathbb{E}\mathcal{E}_n = \mathbb{E}\,\|X - X'\|^{\alpha}$, and the asymptotic distribution of $\mathcal{E}_n$ is a quadratic form of centered Gaussian random variables. The rejection region is in the upper tail. Under an alternative hypothesis, $\mathcal{E}_n$ tends to infinity stochastically, and therefore $\mathcal{E}_n$ determines a consistent goodness-of-fit test.

For most applications the exponent $\alpha = 1$ (Euclidean distance) can be applied, but smaller exponents have been applied for testing distributions with heavy tails including Pareto, Cauchy, and stable distributions.[14,15] The important special case of testing multivariate normality[8,9] is fully implemented in the *energy*[24] package for R.

A detailed introduction to the energy goodness-of-fit tests with several simple examples can be found in Ref 3. Here we will focus on the important applications of testing for multivariate normality, starting with the special case of univariate normality.

The energy statistic for testing whether a sample $X_1, \ldots, X_n$ is from a multivariate normal distribution $N(\mu, \Sigma)$ is developed by Székely and Rizzo.[9] Let $x_1, \ldots, x_n$ denote an observed random sample.

Univariate Normality

For a test of univariate normality, we apply the statistic Eq. (6). Suppose that the null hypothesis is that the sample has a Normal distribution with mean μ and variance σ^2. Then it can be derived that

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

where F and f denote the cdf and density of the hypothesized Normal(μ, σ^2) distribution. The last sum in the statistic $\mathcal{E}_n$ can be linearized in terms of the ordered sample, which allows computation in $O(n \log n)$ time.

Generally, the parameters μ and σ are unknown. In that case we first standardize the sample using the sample mean and the sample standard deviation, then test the fit to the standard normal distribution. The estimated parameters change the critical values of the distribution but not the general shape, and the rejection region is in the upper tail. See Ref 8 for a detailed proof that the test with estimated parameters is statistically consistent (also for the multivariate case) against all alternatives with finite moments.

Multivariate Normality

The statistic $\mathcal{E}_n$ for testing multivariate normality is considerably more difficult to derive. We first standardize the sample using the sample mean vector and the sample covariance matrix to estimate parameters. Then we test the sample for fit to standard multivariate normal. If Z and Z' are iid standard normal in dimension d, we have

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

where $\Gamma(\cdot)$ is the complete gamma function. If $y_1, \ldots, y_n$ are the standardized sample elements, the computing formula for the test of multivariate normality in $\mathbb{R}^d$ is

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

where

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

The expression for $E \parallel a - Z \parallel_d$ follows from the fact that if Z is a d-variate standard normal random vector, then $\|a-Z\|_d^2$ has a noncentral chisquare distribution $\chi^2[\nu;\lambda]$ with noncentrality parameter $\lambda = |a|_d^2/2$, and degrees of freedom $\nu = d + 2\psi$, where ψ is a Poisson random variable with mean λ. Typically the sum in $E \parallel a - Z \parallel_d$ converges to within a small tolerance after 40–60 terms, except when a is a true outlier of the standard multivariate normal distribution (when $\parallel a \parallel$ is very large). However, $\mathbb{E}\|\mathrm{a}-Z\|$ converges to $\parallel a \parallel$ as $\parallel a \parallel \to \infty$, so we can evaluate $\mathbb{E}\|a-Z\|\approxeq\|\mathrm{a}\|$ in that case. See the source code in 'energy.c' of the *energy* package[24] for an implementation.

The test is implemented for the multivariate normal distribution as follows, in the *energy* package[24] with the `mvnorm.etest` function. If the observed sample size is n, it is standardized using the sample mean vector and sample covariance, and the observed test statistic $\mathcal{E}_{n,d}$ computed. A large number M of standard multivariate normal samples of size n are generated, $j = 1, \ldots, M$, each standardized using the j-th sample mean vector and sample covariance, and $\mathcal{E}_{n,d}^{(j)}$ is computed for each of these samples to obtain a reference distribution. An estimated p-value is obtained by finding the proportion of replicates $\mathcal{E}_{n,d}^{(j)}$ that exceed the observed $\mathcal{E}_{n,d}$ statistic. An example illustrating the test of normality for the four-dimensional *iris setosa* data (included in the R distribution) follows. In this example, the hypothesis of normality is rejected at significance level 0.05.

```
> mvnorm.etest(iris[1:50, 1:4], R=999)

  Energy test of multivariate normality: estimated parameters

data:  x, sample size 50, dimension 4, replicates 999
E-statistic = 1.2034, p-value = 0.02503
```

Under the null hypothesis, $n\mathcal{E}_{n,d}$ converges in distribution to a quadratic form $Q_d = \sum_{i=1}^{\infty}\lambda_i Z_i^2$, as $n \to \infty$, where Z_i are iid standard normal random variables and λ_i are non-negative constants that depend on the parameters of the null distribution. Figure 1 displays replicates for testing the iris data with the observed $\mathcal{E}_{n,d}$ identified by the large black dot. The density curve overlaid on the plot is an approximation ($n = 50$) of the density of the asymptotic distribution Q_d.

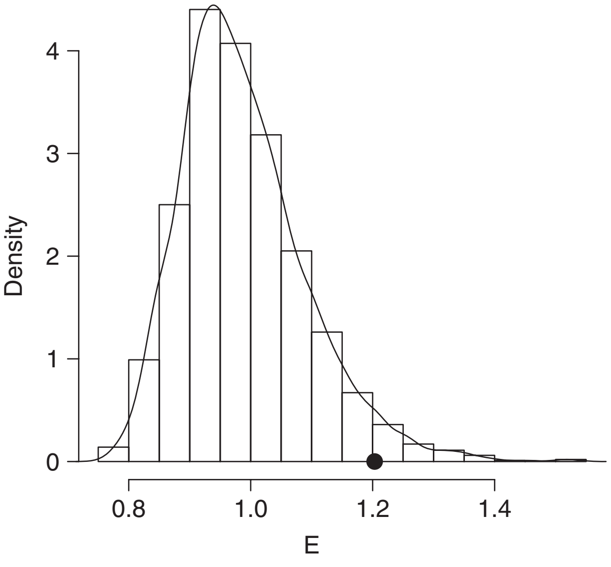

FIGURE 1 | Replicates generated under the null hypothesis in a test of multivariate normality for the *iris setosa* data. The test statistic $\mathcal{E}_{n,d}$ of the observed *iris* sample is located by the black dot.

The energy goodness-of-fit test could alternately be implemented by evaluating the constants λ_i, but that is a difficult problem except in special cases. Some analytical results for univariate normality are given in Ref 3 and a numerical approach investigated in Ref 14. See also Ref 8 for tabulated critical values of $n\widehat{\mathcal{E}}_{n,d}$ for several (n, d) obtained by large scale simulation.

The energy test of multivariate normality is practical to apply via parametric bootstrap as illustrated above for arbitrary dimension, and sample size $n < d$ is not a problem. Monte Carlo power comparisons[9] suggest that the energy test is a powerful competitor to other tests of multivariate normality. Indeed, there are very few other tests in the literature for multivariate normality like energy that are consistent, powerful, omnibus tests with practical implementation; the BHEP tests,[34,35] which also apply a characterization of equality between distributions, share these properties, and have recently been implemented in an R package *MVN*. See Ref 9 for comparisons.

GENERALIZATIONS

One might wonder what makes the functions $|\cdot|^\alpha$, $0 < \alpha < 2$ special in the definition above. One can show that the key property is that $|\cdot|^\alpha$, $0 < \alpha < 2$ is *strongly negative definite* (see Lyons[20]). In this case, the generalized distance function remains a metric for measuring the distance between probability distributions F and G, it is nonnegative and equals zero if and only if $F = G$. What makes the distance function special in the class of strongly negative definite functions is that the distance function is scale equivariant. If we change the scale by replacing x by cx and y by cy, then the squared distance $D^2(F, G)$ is multiplied by c, and therefore the ratio of these functions does not depend on the constant c.

This property of invariance also holds for $0 < \alpha < 2$, because if we change the measurement units in $D^{2(\alpha)}(F, G)$, replace X by cX and Y by cY, then $D^{2(\alpha)}(F, G)$ is multiplied by c^α and the ratio of two statistics of this type is again invariant with respect to c. Hence the statistical decisions based on these ratios do not depend on the choice of measurement units. This invariance property is essential.

One can further generalize energy distance to probability distributions on metric spaces. Consider a metric space (M, d) which has Borel sigma algebra $\mathcal{B}(M)$, so $(M, \mathcal{B}(M))$ is a measurable space. Let $\mathcal{P}(M)$ denote the set of probability measures on $(M, \mathcal{B}(M))$. Then for any pair of probability measures μ and ν in $\mathcal{P}(M)$, we can define the energy distance in terms of the associated random variables X and Y and the metric d as the square root of

$$D^2(\mu,\nu) = 2\mathbb{E}[d(X,Y)] - \mathbb{E}[d(X,X')] - \mathbb{E}[d(Y,Y')],$$

provided that $D^2(\mu, \nu) \geq 0$. However, in general $D^2(\mu, \nu)$ can be negative. In order that D is a metric, it is necessary and sufficient that (M,d) is strongly negative definite (see Lyons[20]). When (M,d) is strongly negative definite, then the energy distance $D(\mu, \nu)$ equals zero if and only if the distributions are equal. A commonly applied metric that is negative definite but not strongly negative definite is the taxicab metric in $\mathbb{R}^2$. Lyons[20] showed that all separable Hilbert spaces (and in particular Euclidean spaces) have strong negative type.

CONCLUSION

Energy distance is a powerful tool for multivariate analysis. It applies to random vectors in arbitrary dimensions, and the methodology requires only the mild assumption of finite first moments or at least finite $\alpha > 0$ moments for some positive α. Computing formulas are simple and the tests have been implemented by nonparametric methods using resampling or Monte Carlo methods. We have illustrated the use of the functions in the *energy* package[24] for several of the methods. The package is open source and distributed under general public license. To scale up to big data problems, for problems such as cluster analysis, one could apply a 'divide and recombine' (D&R) analysis.

Readers may also refer to several interesting applications in a variety of disciplines, under Further Reading. Our review of the background and applications of energy distance is not an exhaustive bibliography, but intended as a starting point.

FURTHER READING

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

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

Székely GJ, Rizzo ML. On the uniqueness of distance covariance. *Stat Probab Lett* 2012, 82:2278–2282. doi:10.1016/j.spl.2012.08.007.

Wahba G. Positive definite functions, reproducing kernel Hilbert spaces and all that. *The Fisher Lecture at JSM 2014*, 2014. Available at: http://www.stat.wisc.edu/wahba/talks1/fisher.14/wahba.fisher.7.11.pdf.

Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 2007, 102:359–378. doi:10.1198/016214506000001437.

Gretton A. A simpler condition for consistency of a kernel independence test, 2015. Available at: http://arxiv.org/abs/1501.06103v1.

Gretton A, Györfi L. Consistent nonparametric tests of independence. *J Mach Learn Res* 2010, 11:1391–1423.

Kong J, Wang S, Wahba G. Using distance covariance for improved variable selection with application to learning genetic risk models. *Stat Med* 2015, 34/10:1097–1258. doi:10.1002/sim.6441.

Kong J, Klein BEK, Klein R, Lee KE, Wahba G. Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proc Natl Acad Sci USA* 2012, 109:20352–20357. doi:10.1073/pnas.1217269109.

Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. *J Am Stat Assoc* 2012, 107:1129–1139. doi:10.1080/01621459.2012.695654.

Martinez-Gomez E, Richards MT, Richards DSP. Distance correlation methods for discovering associations in large astrophysical databases. *Astrophys J* 2014, 781:39.

Kim AY, Marzban C, Percival DB, Stuetzle W. Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Process* 2009, 89/12:2529–2536. doi:10.1016/j.sigpro.2009.04.011.

Menshenin DD, Zubkov AM. Properties of the Szekely-Mori symmetry criterion statistics in the case of binary vectors. *Math Notes* 2012, 91:62–72.

Székely GJ, Móri TF. A characteristic measure of asymmetry and its application for testing diagonal symmetry. *Commun Stat Theory Methods* 2001, 30:1633–1639.

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

Varin T, Bureau R, Mueller C, Willett P. Clustering files of chemical structures using the Szekely-Rizzo generalization of Ward's method. *J Mol Graphics Modell* 2009, 28/2:187–195. doi:10.1016/j.jmgm.2009.06.006.

Zhou Z. Measuring nonlinear dependence in time-series, a distance correlation approach. *J Time Ser Anal* 2012, 33:438–457.

REFERENCES

1. Székely GJ. *Potential and Kinetic Energy in Statistics*, Lecture Notes, Budapest Institute of Technology (Technical University), 1989.
2. Székely GJ. E-statistics: The Energy of Statistical Samples, Technical Report, Bowling Green State University, Department of Mathematics and Statistics No. 02–16 and Technical Reports by the same title from 2000–2003, e.g. No.03-05 and NSA grant # MDA 904-02-1-s0091 (2000–2002), 2002.
3. Székely GJ, Rizzo ML. Energy statistics: statistics based on distances. *J Stat Plann Infer* 2013, 143:1249–1272. doi:10.1016/j.jspi.2013.03.018.
4. Cramér H. On the composition of elementary errors. *Skand Aktuar* 1928, 11:141–180.
5. Zinger AA, Kakosyan AV, Klebanov LB. Characterization of distributions by means of mean values of some statistics in connection with some probability metrics, In: *Stability Problems for Stochastic Models*. Moscow, VNIISI, 1989, 47–55. (in Russian), English Translation: A characterization of distributions by mean values of statistics and certain probabilistic metrics, *Journal of Soviet Mathematics* (1992), 2012.
6. Mattner L. Strict negative definiteness of integrals via complete monotonicity of derivatives. *Trans Am Math Soc* 1997, 349:3321–3342.
7. Morgenstern D. Proof of a conjecture by Walter Deubner concerning the distance between points of two types in R^d. *Discrete Math* 2001, 226:347–349.
8. Rizzo ML. A new rotation invariant goodness-of-fit test, PhD dissertation, Bowling Green State University, 2002.

9. Székely GJ, Rizzo ML. A new test for multivariate normality. *J Multivar Anal* 2005, 93:58–80.
10. Baringhaus L, Franz C. On a new multivariate two-sample test. *J Multivar Anal* 2004, 88:190–206.
11. Rizzo ML. A test of homogeneity for two multivariate populations. In: *2002 Proceedings of the American Statistical Association, Physical and Engineering Sciences Section*. Alexandria, VA: American Statistical Association, 2003.
12. Szekely GJ, Rizzo ML. Testing for equal distributions in high dimension. *InterStat* 2004, Nov.
13. Rizzo ML, Székely GJ. DISCO analysis: a nonparametric extension of analysis of variance. *Ann Appl Stat* 2010, 4:1034–1055.
14. Yang G. The energy goodness-of-fit test for univariate stable distributions. PhD Thesis, Bowling Green State University, 2012.
15. Rizzo ML. New goodness-of-fit tests for Pareto distributions. *ASTIN Bull* 2009, 39:691–715.
16. Li Y. Goodness-of-fit tests for Dirichlet distributions with applications. PhD Thesis, Bowling Green State University, 2015.
17. Székely GJ, Rizzo ML. Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. *J Classif* 2005, 22:151–183.
18. Li S. k-groups: a generalization of k-means by energy distance. PhD Thesis, Bowling Green State University, 2015.
19. Baringhaus L, Franz C. Rigid motion invariant two-sample tests. *Stat Sin* 2010, 20:1333–1361.
20. Lyons R. Distance covariance in metric spaces. *Ann Probab* 2013, 41:3284–3305.
21. Sejdinovic D, Gretton A, Sriperumbudur B, Fukumizu K. Hypothesis testing using pairwise distances and associated kernels. In: *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012. Available at: http://arxiv.org/abs/1205.0411v2.
22. Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann Stat* 2013, 41:2263–2291.
23. Klebanov LB. *N-distances and Their Applications*. Charles University, Prague: Karolinum Press; 2005.
24. Rizzo ML, Székely GJ. Energy: E-statistics (energy statistics). R package version 1.6.2, 2014. Available at: http://CRAN.R-project.org/package=energy.
25. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2015. Available at: http://www.R-project.org/.
26. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC; 1993.
27. Davison AC, Hinkley DV. *Bootstrap Methods and their Application*. Oxford: Cambridge University Press; 1997.
28. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing independence by correlation of distances. *Ann Stat* 2007, 35:2769–2794. doi:10.1214/009053607000000505.
29. Székely GJ, Rizzo ML. Brownian distance covariance. *Ann Appl Stat* 2009, 3:1236–1265. doi:10.1214/09-AOAS312.
30. Székely GJ, Rizzo ML. The distance correlation t-test of independence in high dimension. *J Multivar Anal* 2013, 117:193–213. doi:10.1016/j.jmva.2013.02.012.
31. Székely GJ, Rizzo ML. Partial distance correlation with methods for dissimilarities. *Ann Stat* 2014, 42:2382–2412.
32. Huo X, Székely G. Fast computing for distance covariance. *Technometrics* 2015. doi:10.1080/00401706.2015.1054435.
33. Rizzo ML, Székely GJ. pdcor: Partial distance correlation. R package version 1.0.0, 2014.
34. Henze N, Zirkler B. A class of invariant consistent tests for multivariate normality. *Commun Stat Theory Methods* 1990, 19:3595–3618.
35. Henze N, Wagner T. A New Approach to the BHEP tests for multivariate normality. *J Multivar Anal* 1997, 62:1–23.