Energy distance
Maria L. Rizzo¹* and Gábor J. Székely²,³

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.

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.
BACKGROUND AND APPLICATION OF ENERGY DISTANCE

The notion of ‘energy statistics’ 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 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:

\[ D^2(F, G) = 2 \int_{-\infty}^{\infty} (F(x) - G(x))^2 \, dx. \]

However, the equivalence of energy distance with Cramèr’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. An alternate proof related to a result of Morgenstern appears inRefs 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 normality8,9 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

\[ E_{n,m}(X, Y) = 2A - B - C, \]

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

\[ A = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} || x_i - y_j ||, \quad B = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} || x_i - x_j ||, \]

\[ C = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} || y_i - y_j ||. \]

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

\[ T = \frac{nm}{n+m} E_{n,m}(X, Y). \]

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 tend 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 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.
To compute the energy statistic only:

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

then \( 0 \leq H \leq 1 \) 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 \( H_0 : F_1 = \ldots = F_K \), \( K \geq 2 \), one can apply the between-sample test statistic, or alternatively 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 \( \| X - Y \|^{a} \), where \( 0 < a < 2 \). The characterization does not hold if \( a = 2 \) because \( 2E\|X-Y\|^2 = E\|X-X'\|^2 + E\|Y-Y'\|^2 - 2E\|X-Y'\|^2 \equiv 0 \) whenever \( EX = EY \). We denote the corresponding two-sample energy statistic by \( E_{n_1,n_2}^{(a)} \).

Let \( A = \{a_1, \ldots, a_n\} \), \( B = \{b_1, \ldots, b_m\} \) be two samples, and define

\[ g_a(A, B) := \frac{1}{n_1n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \| a_i - b_j \|^{a}, \]

for \( 0 < a \leq 2 \). 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

\[ W_a = W_a(A_1, \ldots, A_K) = \sum_{j=1}^{K} \frac{n_j}{2} g_a(A_j, A_i), \]

and the total dispersion of the observed response is

\[ T_a = T_a(A_1, \ldots, A_K) = \frac{N}{2} g_a(A, A), \]

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

\[ S_{n,a} = \sum_{1 \leq i < k \leq K} \left( \frac{n_i + n_k}{2N} \right) \frac{n_i n_k}{n_i + n_k} E_{n_i,n_k}^{(a)}(A_i, A_k) \]

\[ = \sum_{1 \leq i < k \leq K} \left\{ \frac{n_i n_k}{2N} (2g_a(A_i, A_k) - g_a(A_i, A_i) - g_a(A_k, A_k)) \right\}. \]

Note that \( S_{n,a} \) 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

\[ T_a = S_a + W_a, \]

where both \( S_a \) and \( W_a \) are nonnegative.
For every $0 < \alpha < 2$, the statistic Eq. (4) determines a consistent test of the multi-sample hypothesis of equal distributions. 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 $\alpha$ moments. If any of the underlying distributions may have non-finite first moment, a suitable choice of $\alpha$ 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.

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

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.

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:

\[ d(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) + \frac{n_i + n_j}{n_i + n_j + n_k} d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j). \]
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

\[
d_{(i)k} := d(C_i \cup C_j, C_k) = a_i d(C_i, C_k) + a_j d(C_j, C_k) + \beta d(C_i, C_j) = a_i d_{ik} + a_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|
\]

where

\[
a_i = \frac{n_i + n_k}{n_i + n_j + n_k}; \quad \beta = \frac{-n_k}{n_i + n_j + n_k}; \quad \gamma = 0.
\]

If we substitute squared Euclidean distances for Euclidean distances in this recursive formula, keeping the same parameters \( (\alpha, \alpha, \beta, \gamma) \), 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\(^\text{17}\) 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 \textsc{hclust} hierarchical clustering function provided in R\(^\text{25}\) implements exactly the above recursive formula Eq. (5), and therefore one can use \textsc{hclust} to apply either the \( \mathcal{E} \)-clustering solution or Ward's method by specifying method ‘\textsc{ward.D}’ (energy) or ‘\textsc{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 \textit{distance covariance} (dCov) and \textit{distance correlation} (dCor), defined below.

### Distance Covariance

The simplest formula for the \textit{distance covariance} statistic is the square root of

\[
\gamma^2_n(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} \hat{A}_{ij} \hat{B}_{ij},
\]

where \( \hat{A} \) and \( \hat{B} \) 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 \((x, y) = (x_i, y_j) : i = 1, \ldots, n\) 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

\[
\hat{A}_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}_{..}, \quad i, j = 1, \ldots, n,
\]

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 \( \hat{B} \) is

\[
\hat{B}_{ij} = b_{ij} - \bar{b}_i - \bar{b}_j + \bar{b}_{..}, \quad i, j = 1, \ldots, n.
\]

The sample \textit{distance variance} is

\[
\gamma^2_n(X, X) = \gamma^2_n(Y, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} \hat{A}_{ij}^2.
\]

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

### Distance Correlation

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

\[
\gamma^2_n(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} \hat{A}_{ij} \hat{B}_{ij},
\]

and the squared distance correlation is defined by
\[ \mathcal{R}_n^2(X, Y) = \begin{cases} \frac{\nu_n^2(X, Y)}{\sqrt{\nu_n^2(X)} \nu_n^2(Y)}, & \nu_n^2(X) \nu_n^2(Y) > 0; \\ 0, & \nu_n^2(X) \nu_n^2(Y) = 0. \end{cases} \]

Distance correlation satisfies

1. \(0 \leq \mathcal{R}_n(X, Y) \leq 1.\)
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 \(d\text{Cov}\) as \(\nu_n^2\) and \(d\text{Cor}\) 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 can be written in terms of expected distances:

\[
\nu^2(X, Y) = \mathbb{E}\|X - X'\|\|Y - Y'\| + \mathbb{E}\|X - X'\|\|Y - Y''\| - 2\mathbb{E}\|X - X'\|\|Y - Y''\|,
\]

where \((X, Y), (X', Y')\), and \((X'', Y'')\) are iid.\(^{29}\) Here \(\nu^2(X, Y)\) 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: \(\nu^2(X, Y) \geq 0\) 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:

\[
\mathcal{R}^2(X, Y) = \begin{cases} \frac{\nu^2(X, Y)}{\sqrt{\nu^2(X)} \nu^2(Y)}, & \nu^2(X) \nu^2(Y) > 0; \\ 0, & \nu^2(X) \nu^2(Y) = 0. \end{cases}
\]

The empirical coefficients \(\nu_n(X, Y)\) and \(\mathcal{R}_n(X, Y)\) converge almost surely to the population coefficients \(\nu(X, Y)\) and \(\mathcal{R}(X, Y)\), as \(n \to \infty\). The distribution of \(Q_n = n \nu_n^2(X, Y)\) converges to a quadratic form \(\sum_{i=1}^{\infty} \lambda_i Z_i^2\), where \(Z_i\) are iid standard normal and \(\lambda_i\) 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 \(Q_n = n \nu_n^2\), and it can be implemented as a nonparametric permutation test. The \(d\text{Cov}\) test applies to random vectors in arbitrary, not necessarily equal dimensions, for any sample size \(n \geq 4\). 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 \(\mathbb{R}^{p,q}\). Functions `dcor`, `dcov`, `dcov.test` and `dcor.test` 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.

```r
> 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.

\[ \text{dcov.test}(x, ss) \]

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:

\[ \text{dcor}(x, ss) \]

1] 0.2996931

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

\[ \text{unlist(DCOR}(x, ss)) \]

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 \( \lambda^2(X,Y) \) 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

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

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 \( B \). Then

\[
(\tilde{A} \cdot \tilde{B}) = \frac{1}{n(n-3)} \sum_{i \neq j} \bar{A}_{ii} \bar{B}_{ii},
\]

is an unbiased estimator of squared population distance covariance \( \lambda^2(X,Y) \). 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.

A bias corrected \( R_n^2 \) 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 \( R_n^2 \):

\[ \text{dcor}(x, ss)^2 \]

[1] 0.08981596

\[ \text{bcdcor}(x, ss) \]

[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 \( V_n^2 \) and \( R_n^2 \). 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\(^\text{32}\) 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 symmetric 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 $V^2$) 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*, 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 $E(X, Y)$ 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

$$
E_n = n \left(2 - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} ||x_i - X||^a - \mathbb{E} ||x_i - X'||^a - \frac{1}{n} \sum_{i,j=1}^{n} ||x_i - x_j||^a \right),
$$

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

For most applications the exponent $a = 1$ (Euclidean distance) can be applied, but smaller exponents have been applied for testing distributions with heavy tails including Pareto, Cauchy, and stable distributions. The important special case of testing multivariate normality is fully implemented in the *energy* 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. 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 $\mu$ and variance $\sigma^2$. Then it can be derived that

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

where $F$ and $f$ denote the cdf and density of the hypothesized Normal($\mu, \sigma^2$) distribution. The last sum in the statistic $E_n$ can be linearized in terms of the ordered sample, which allows computation in $O(n \log n)$ time.

Generally, the parameters $\mu$ and $\sigma$ 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 $E_n$ 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

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

where

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

The expression for \( \mathbb{E} \| a - Z \|_d \) follows from the fact that if \( Z \) is a \( d \)-variate standard normal random vector, then \( \| a - Z \|_d^2 \) has a noncentral chi-square distribution \( \chi^2(\nu; \lambda) \) with noncentrality parameter \( \lambda = |a|_d^2/2 \), and degrees of freedom \( \nu = d + 2\psi \), where \( \psi \) is a Poisson random variable with mean \( \lambda \). Typically the sum in \( \mathbb{E} \| a - Z \|_d \) converges to within a small tolerance after 40–60 terms, except when \( a \) is a true outlier of the standard multivariate normal distribution (when \( \| a \| \) is very large). However, \( \mathbb{E} \| a - Z \| \) converges to \( \| a \| \) as \( \| a \| \to \infty \), so we can evaluate \( \mathbb{E} \| a - Z \| \approx \| a \| \) in that case. See the source code in ‘energy.c’ of the energy package\(^{34}\) for an implementation.

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 \( \lambda_i \) 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 \( Q_d \).

> 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

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 $\lambda_i$, 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 $n\tilde{E}_{n,d}$ 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 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, 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). 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 $c^2$, 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 $c^\alpha$ 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 $\mu$ and $\nu$ 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). 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 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 $\alpha$. 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 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**


REFERENCES


