ASIAN JOURNAL OF STATISTICS AND APPLICATIONS

VOL NO. 2, ISSUE 2, PAGE NO. 14-27, YEAR 2025

ISSN: 3048-7455

https://DOI:10.47509/AJSA.2025.V02I02.02

Goodness-of-fit tests for multivariate normality with missing components

John Lawrence

US Food and Drug Administration,
10903 New Hampshire Ave., Silver Spring, MD, USA 20993-002,
John.Lawrence@fda.hhs.gov

ABSTRACT
Suppose $\mathrm{x}_1, \ldots, \mathrm{x}_n$ is a random sample and it is desired to test whether the distribution is multivariate normal. Many likelihood-based methods such as mixed effects models for repeated measures are based on an assumption that the data have a multivariate normal distribution. There are many proposed goodness of fit tests for this problem. However, in many practical applications some of the components are missing for some of the vectors. There are few methods available for this very common problem. This paper describes three different methods using interpoint distances and compares their performance in several scenarios using simulation and one scenario with clinical trial data.

KEYWORDS
Missing data, energy statistic, interpoint distance, data depth

Declarations
Funding: none
Conflicts of interest/Competing interests: none
Availability of data and material: Summary data are provided.
Code availability: Software is provided.

1. Introduction

Suppose $\mathrm{x}_1, \ldots, \mathrm{x}_n$ is a random sample from some unknown multivariate distribution F and it is desired to test the hypothesis that F is a multivariate normal distribution. There are many goodness-of-fit tests for this scenario. The testing scenario can be either a specified distribution with a given mean and covariance matrix or the composite null hypothesis that F has a multivariate normal distribution with any mean and covariance. A theorem in Maa et al [8] demonstrates that the interpoint distance distribution can be used to test this hypothesis. Two multivariate distributions are equal if and only if the three interpoint distance distributions are equal (the distance between two vectors from the first distribution, two vectors from the second

CONTACT John Lawrence. Email: John.Lawrence@fda.hhs.gov

distribution, or one from each). The theorem applies to general distance functions, but this article only uses Euclidean distance. Three examples of goodness-of-fit tests based on interpoint distances are given in Bartoszyński et al [1], Székely and Rizzo [2], and Bonetti and Pagano [13].

The test in Bartoszyński et al [1] is motivated by the idea that if a random triangle is formed with two vertices from the distribution F and one vertex from a specified distribution G, then each leg of the triangle has the same chance of being the shortest leg in the triangle if F is equal to that specified distribution G. Moreover, each leg has the same chance of being the longest leg (or being neither the shortest nor longest) in the triangle if $F = G$. If the data can be used to estimate these probabilities and it is possible to find the mean and variance of these estimates under the null hypothesis $F = G$ to construct a test statistic. However, the calculation of the estimates and their variance under the null hypothesis is not as trivial as it first seems. An accurate and efficient method has recently been proposed by Lawrence [7, in submission]. For the composite null hypothesis, the sample is first standardized. The standardized data will no longer be independent, but now they will have a standard normal distribution under the null hypothesis. Hence, the statistic can be calculated as before but the distribution must be calculated for the scenario with dependent data.

The statistic described in Székely and Rizzo [13] is based on comparing the mean distances between two vectors. The sample can be used to estimate the mean distance if both vectors have distribution F and if one vector has distribution F and the other has distribution G. These two mean distances are the same as the mean distance between two vectors with distribution G if and only if $F = G$. A test statistic is defined and the single or composite null hypothesis can be tested. More detailed description of the test statistic is provided in Section 2.

The test described in Bonetti and Pagano [2] is based on directly comparing the empirical distribution function of the interpoint distances for the observed data to the interpoint distance distribution of the postulated multivariate normal distribution. This seems like a natural test using the interpoint distances and it is shown to have high power for many alternative distributions. However, it is not a consistent test in general because all three interpoint distance distributions need to be equal to ensure that the multivariate distributions are equal. An example is provided in Maa et al [8] that shows that merely having two interpoint distance distributions that are equal is not sufficient.

This article also introduces two new tests based on interpoint distances. In order to directly use the results in Maa et al [8], a test is needed that compares all three univariate interpoint distance distributions. David [3] describes several different extensions of the Kolmogorov-Smirnov test to the three sample scenario. The two new tests are described in more detail in Section 2.

Tests that are based on interpoint distances reduce multidimensional data into univariate distances. Despite that reduction in dimensionality, all of the previously published tests have demonstrated high power against many alternative distributions. It is for these reasons that interpoint distance approaches are chosen for the problem considered here with missing components.

According to National Research Council [10], missing data is prevalent in clinical trials and can occur for many different reasons. Types of missing data are broadly classified into missing completely at random, missing at random, or missing not at random. For example, suppose a trial is 5 years long with staggered entry of patients during the first 3 years. The protocol specifies that all patients have a measurement every 3 months from the time they enter until the end of the study. The first patient

in the study may have 20 measurements whereas the last patient entering the study may have only 8. It could be reasonable to assume that the last 12 components of the 20-dimensional vector of observations are missing completely at random for the last patient. A patient could be too sick to have the measurement taken during some scheduled visits, but have measurements taken at later visits. That type of missing data could be missing at random or missing not at random. In general, it is not possible to know the reason for missingness. Common approaches to analyzing missing data include deleting cases with any missing values, imputing values for missing components, or multiple imputation of missing components.

In this article, a different approach will be taken to analyzing the data with missing components. The missing components will be assumed to have the distribution specified by the null hypothesis. A value for the missing components will not be imputed. Instead, the probabilities involving lengths of legs in the triangle or the average interpoint distance will be calculated assuming the missing components are random variables. This idea will be explained in more detail in Section 2. The distribution calculation of the statistic under the null hypothesis will be conditioned on the observed pattern of missing data. If the data are missing completely at random, the resulting tests are unbiased. Unfortunately, when there is missing data no test can be consistent for all alternatives. It will always be possible to construct scenarios where $F \neq G$ but the missing data mask the difference. For example, suppose G is the standard bivariate normal distribution and the first component of F has the standard normal distribution but the second component is a mixture where 90% of the time it is standard normal but 10% of the time it has a t-distribution. Furthermore, suppose the second component is only observed in the 90% of cases where it has the standard normal distribution. This is missing not at random and in this scenario, the data would appear to show the null hypothesis is true. Conversely, some types of missingness could make data appear to be not normal when in fact it is normal.

Multidimensional goodness-of-fit tests are not only used for testing model assumptions. Pearl et al [11] describes a simulation-based technique for estimating the parameters of a high-dimensional stochastic model by optimizing a criterion based on the goodness-of-fit test. This, an extension to the missing data scenario can be useful for estimating parameters in a stochastic model when there are missing data.

2. Mathematical description of tests

First, consider the case of a completely specified multivariate normal distribution G. $\mathrm{x}_1, \ldots, \mathrm{x}_n$ is a random sample from some unknown multivariate distribution F with some components missing. The data can be standardized by subtracting the postulated mean and multiplying by the square root of the variance matrix so that we can assume G is the standard multivariate normal distribution. The null hypothesis is $F = G$. It will be assumed that all random variables have finite first absolute moment, $E\,|X|$. This condition is necessary for the proof of the theorems in Székely and Rizzo [13] and any corollaries or theorems derived from those.

If no components of x_i or x_j are missing, then the Euclidean distance between the two vectors is the square root of the sum of the squared differences in coordinates $\|\mathrm{x}_i - \mathrm{x}_i\| = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2}$.

Suppose some of the components of x_i or x_j could be missing. As an example,

suppose $d = 4$ and only the second and fourth components of x_i are observed while the first three components of x_j are observed. Use the notation where the missing coordinates are represented by boxes so that $\mathrm{x}_i = \begin{pmatrix} \blacksquare \\ x_{i,2} \\ \blacksquare \\ x_{i,4} \end{pmatrix}$ and $\mathrm{x}_j = \begin{pmatrix} x_{j,1} \\ x_{j,2} \\ x_{j,3} \\ \blacksquare \end{pmatrix}$. Define the random variables $(Z_{i,1}, Z_{i,3})'$ and $Z_{j,4}$ with a distribution equal to the conditional distribution given the observed coordinates. In other words, $(Z_{i,1}, Z_{i,3})'$ is bivariate normal with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $Z_{j,4}$ is normal with mean 0 and variance 1. Then, we can replace x_i and x_j by the random vectors $\begin{pmatrix} Z_{i,1} \\ x_{i,2} \\ Z_{i,3} \\ x_{i,4} \end{pmatrix}$ and $\begin{pmatrix} x_{j,1} \\ x_{j,2} \\ x_{j,3} \\ Z_{j,4} \end{pmatrix}$ and calculate the expected value of the distance between these random vectors.

With that motivating example, we now switch to the general case. The vectors have dimension d and are associated with each vector x_i are indicator variables with $\delta_{i,k} = \begin{cases} 1 if x_{i,k} is observed \\ 0 if x_{i,k} is missing \end{cases}$. The expected difference between two vectors with possibly missing components is

$$E \left\| \mathrm{x}_i - \mathrm{x}_j \right\|$$

$$= E \sqrt{\sum_{\substack{k:\delta_{i,k} \\ =\delta_{j,k}=1}} (x_{i,k} - x_{j,k})^2 + \sum_{\substack{k: \\ \delta_{i,k}=0 \\ \delta_{j,k}=1}} (Z_{i,k} - x_{j,k})^2 + \sum_{\substack{k: \\ \delta_{i,k}=1 \\ \delta_{j,k}=0}} (x_{i,k} - Z_{j,k})^2 + \sum_{\substack{k:\delta_{i,k} \\ =\delta_{j,k}=0}} (Z_{i,k} - Z_{j,k})^2}$$

$$= E \sqrt{\sum_{k:\delta_{i,k}=\delta_{j,k}=1} (x_{i,k} - x_{j,k})^2 + \chi^2_{\nu_{i,j,1}}(\lambda_{i,j}) + 2\chi^2_{\nu_{i,j,2}}}$$

the expected value of the square root of the sum of a constant, a noncentral chi-square random variable with

$$\nu_{i,j,1} = \sum_{k=1}^{d} \{\delta_{i,k}(1 - \delta_{j,k}) + \delta_{j,k}(1 - \delta_{i,k})\}$$

degrees of freedom and non-centrality parameter

$$\lambda_{i,j} = \sum_{k:\delta_{i,k}=0,\delta_{j,k}=1} x^2_{j,k} + \sum_{k:\delta_{i,k}=1,\delta_{j,k}=0} x^2_{i,k}$$

and twice an independent chi-square random variable with

$$\nu_{i,j,2} = \sum_{k=1}^{d} \{(1-\delta_{i,k})(1-\delta_{j,k})\}$$

degrees of freedom. There are some special cases where no numerical integration is needed. If $\nu_{i,j,2} = \nu_{i,j,1} = 0$, then there are no missing components in either vector and the mean is just the ordinary distance between the two vectors $\sqrt{\sum_k (x_{i,k} - x_{j,k})^2}$. If $\nu_{i,j,2} > \nu_{i,j,1} = 0$, then the mean is

$$E\sqrt{a + 2\chi^2_\nu} = \frac{2^{2-\nu}\sqrt{\pi}\Gamma(\nu) U\left(\frac{-1}{2}; \frac{1-\nu}{2}; \frac{a}{4}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}$$

with $a = \sum_{k:\delta_{i,k}=\delta_{j,k}=1} (x_{i,k} - x_{j,k})^2$ and $\nu = \nu_{i,j,2}$ where U denotes Kummer′s confluent hypergeometric function of the second kind.

If $\nu_{i,j,1} > \nu_{i,j,2} = 0$, then using the fact that a non-central chi-square random variable is the Poisson-weighted sum of central chi-square random variables, the mean can be expressed as

$$E\sqrt{a + \chi^2_\nu(\lambda)} = \sum_{i=0}^{\infty} \frac{2^{-i} e^{-\lambda/2} \lambda^i E\sqrt{a + \chi^2_{\nu+2i}}}{i!} = \sum_{i=0}^{\infty} \frac{2^{-i-1/2} e^{-\lambda/2} \lambda^i E\sqrt{2a + 2\chi^2_{\nu+2i}}}{i!}$$

with $a = \sum_{k:\delta_{i,k}=\delta_{j,k}=1} (x_{i,k} - x_{j,k})^2$ and $\nu = \nu_{i,j,1}$. The terms in the sum can be evaluated using the formula for $E\sqrt{a + 2\chi^2_\nu}$ derived earlier. If both $\nu_{i,j,1}$ and $\nu_{i,j,2}$ are positive, then the expected value can be evaluated by numerical evaluation of a one-dimensional integral since

$$E\sqrt{\sum_{k:\delta_{i,k}=\delta_{j,k}=1} (x_{i,k} - x_{j,k})^2 + \chi^2_{\nu_{i,j,1}}(\lambda_{i,j}) + 2\chi^2_{\nu_{i,j,2}}}$$

$$= E\left[E\left\{\sqrt{\sum_{k:\delta_{i,k}=\delta_{j,k}=1} (x_{i,k} - x_{j,k})^2 + \chi^2_{\nu_{i,j,1}}(\lambda_{i,j}) + 2\chi^2_{\nu_{i,j,2}}} \,\middle|\, \chi^2_{\nu_{i,j,1}}(\lambda_{i,j})\right\}\right]$$

The expression given earlier for $E\sqrt{a + 2\chi^2_\nu}$ can be used to find the conditional expectation and the unconditional expected value can be found by numerical integration.

Estimate the mean distance between two independent random vectors with distribution F by

$$\binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E \left\| \mathrm{x}_i - \mathrm{x}_j \right\|$$

where the expectation is conditional on all observed components and all missing components are filled in with random variables that have the distribution of those missing components conditional on the observed components. Note, this is an unbiased estimate of the true mean distance if $F = G$ and the missing components are missing completely at random. The reason is that the expected value of a conditional expectation is the unconditional expected value. Otherwise, it is an estimate of the distance between two random vectors that have a distribution that is a mixture of F and some other distributions where some components have the marginal distribution of components of F and others have conditional distribution of components of G. The following theorem demonstrates that $F = G$ if and only if this mixture distribution is equal to G.

Theorem 1. Let x be a random vector where with probability p, where $0 < p \leq 1$, the distribution is F but with probability $1 - p$ a fixed subset of the components are missing and replaced by random variables that have distribution of the conditional distribution of those components given the observed components from the distribution G. Furthermore, suppose the probability of having the components missing is independent of the observed components and the unobserved components. Then, x has the distribution G if and only if $F = G$.

Proof. The distribution of the observed components of x is the marginal distribution of those components from the distribution F. The distribution of the remaining components conditional on those components that are always observed is a mixture of the conditional distribution of those components from F and the conditional distribution of those components from G. Thus, if $F = G$, clearly x will have the distribution of G. However, if $F \neq G$, then either the marginal distribution of the observed components is different in F compared to that in G or the conditional distribution of the possibly unobserved components is different in F compared to that in G. In either case x cannot have the distribution in G.

Theorem 1 applies to a scenario where only one missing data structure is allowed. The following corollary extends that to the scenario where arbitrary missing data structures are allowed. Each possible structure has a probability p_i of occurring where i indexes the different missing component structures; there are 2^d possible structures including the case where no components are missing. The missing component structures need to be listed and ordered in such a way that $i = 1$ corresponds to the case where no components are missing. It is given that $\sum p_i = 1$.

Corollary 1. Let x be a random vector where with probability p_1, where $0 < p_1 \leq 1$, the distribution is F but with probability p_i; $i = 2, \ldots, 2^d$, the subset with index i of the components are missing and replaced by random variables that have the conditional distribution of those components given the observed components from the distribution G. Then, x has the distribution G if and only if $F = G$.

Proof. The proof follows by induction from the fact that a mixture of k distributions can be written as a mixture of two distributions where the first distribution is a mixture of k-1 and using Theorem 1.

Given a random sample $\mathrm{x}_1, \ldots, \mathrm{x}_n$ from a distribution F, estimate the mean distance between two random vectors where one has distribution F and the other, y, has distribution G by

$$n^{-1} \sum_{i=1}^{n} E \left\| \mathrm{x}_i - y \right\|$$

Note that

$$E \left\| \mathrm{x}_i - y \right\| = E \sqrt{\sum_{k:\delta_{i,k}=1} (x_{i,k} - Y_k)^2 + \sum_{k:\delta_{i,k}=0} (Z_{i,k} - Y_k)^2}$$

$$= E \sqrt{\chi^2_{\sum_{k=1}^{d} \delta_{i,k}} \left(\sum_{k:\delta_{i,k}=1} x_{i,k}^2 \right) + 2\chi^2_{\sum_{k=1}^{d} (1-\delta_{i,k})}}$$

When x_i has no missing components, then $E \left\| \mathrm{x}_i - y \right\|$ can be evaluated using Lawrence (6). Otherwise, it can be evaluated as a one-dimensional integral using the same iterated expectation technique described earlier. Finally, the mean distance between two independent vectors with distribution G, $\left\| \mathrm{y}_1 - \mathrm{y}_2 \right\|$, is $2\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}$. The energy test statistic generalized to the missing component scenario is then defined as

$$2n^{-1} \sum_{i=1}^{n} E \left\| \mathrm{x}_i - y \right\| - \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E \left\| \mathrm{x}_i - \mathrm{x}_j \right\| - E \left\| \mathrm{y}_1 - \mathrm{y}_2 \right\|$$

This is the same as the definition in Székely and Rizzo (13) when there are no missing components. But, when there are missing components, the distances between observed vectors are replaced by conditional expectations. The null hypothesis is rejected for large values of the statistic. The distribution of the statistic under the null hypothesis is estimated by repeatedly drawing samples from G with the same missing component structure as in the observed data.

For the composite null hypothesis testing scenario, there are two modifications. First, in the original sample, the data are standardized by an estimate of the mean and covariance. One way to estimate the mean and covariance is by using the maximum likelihood estimates. A simpler alternative is to use the sample mean and covariance from the complete cases. Second, when calculating the distribution under the null hypothesis, the simulated data need to be standardized using the maximum likelihood estimates of the mean and variance.

The energy test describe above compares the mean interpoint distances. Now, consider the test statistic that uses a three-sample generalization of the Kolmogorov-Smirnov test to compare the full distribution of the interpoint distances.

Let $H_{YY}(t) = P[\|\mathrm{y}_1 - \mathrm{y}_2\| \leq t] = P\left[\chi^2_d \leq \frac{t^2}{2}\right]$ where y_1 and y_2 are independent standard multivariate normal random variables with distribution G. Also, let H_{XX} denote the interpoint distance distributions for two independent vectors with distribution F and H_{XY} denote the interpoint distance distribution where one vector has distribution and let F and the other has distribution and G.

Estimate the distribution function H_{XX} as follows. First, define the support for the estimated distribution. The support could be taken as the set of all observed interpoint distances for pairs of complete observations. Alternatively, it could be taken as the union of two sets: the set of all observed interpoint distances for pairs of complete observations and the $\frac{i}{N+1}$ quantiles of the distribution $H_{YY}(t)$ for $i = 1, \ldots, N$. The integer N can be any number desired. This union could be used in situations where there are not many pairs of observed complete data vectors. Assume the points in the support are labeled such that $0 < t_1 < \ldots < t_K$. Now, each pair x_i and x_j will contribute a total mass of $\frac{1}{\binom{n}{2}}$ to the estimated distribution. Let $h^{i,j}_{XX}(t)$ denote the mass allocated by that pair to a specific value t in the support. To summarize, for all $1 \leq i < j \leq n$, $\sum_{t \in support} h^{i,j}_{XX} = \frac{1}{\binom{n}{2}}$ and $\sum_{1 \leq i < j \leq n} \sum_{t \in support} h^{i,j}_{XX} = 1$. If both x_i and x_j are complete, then the mass is assigned to the point $\|\mathrm{x}_i - \mathrm{x}_j\|$. Otherwise, the mass is allocated to all t_k in the support with

$$h^{i,j}_{XX}(t_1) = \frac{P[\|\mathrm{x}_i - \mathrm{x}_j\| \leq t_1]}{\binom{n}{2}} \quad \text{and} \quad h^{i,j}_{XX}(t_k) = \frac{P[\|\mathrm{x}_i - \mathrm{x}_j\| \leq t_k] - P[\|\mathrm{x}_i - \mathrm{x}_j\| < t_{k-1}]}{\binom{n}{2}} \quad \text{for } k > 1$$

where x_i and x_j are the random variables defined previously replacing the missing components by random variables with the appropriate conditional distribution. To calculate $h^{i,j}_{XX}(t)$ when there are missing components, first note that

$$P[\|\mathrm{x}_i - \mathrm{x}_j\| \leq t]$$

$$= P\left[\sqrt{\sum_{k:\delta_{i,k}=\delta_{j,k}=1} (x_{i,k} - x_{j,k})^2 + \chi^2_{\sum_{k=1}^d \{\delta_{i,k}(1-\delta_{j,k}) + \delta_{j,k}(1-\delta_{i,k})\}}(\lambda_{i,j}) + 2\chi^2_{\sum_{k=1}^d \{(1-\delta_{i,k})(1-\delta_{j,k})\}}} \leq t\right]$$

Therefore, we can calculate

$$P[\|\mathrm{x}_i - \mathrm{x}_j\| \leq t]$$

$$= P\left[\sum_{k:\delta_{i,k}=\delta_{j,k}=1} (x_{i,k} - x_{j,k})^2 + \chi^2_{\sum_{k=1}^d \{\delta_{i,k}(1-\delta_{j,k}) + \delta_{j,k}(1-\delta_{i,k})\}}(\lambda_{i,j}) + 2\chi^2_{\sum_{k=1}^d \{(1-\delta_{i,k})(1-\delta_{j,k})\}} \leq t^2\right]$$

$$= P\left[\chi^2_{\sum_{k=1}^{d}\{\delta_{i,k}(1-\delta_{j,k})+\delta_{j,k}(1-\delta_{i,k})\}}(\lambda_{i,j}) + 2\chi^2_{\sum_{k=1}^{d}\{(1-\delta_{i,k})(1-\delta_{j,k})\}} \leq t^2 - \sum_{k:\delta_{i,k}=\delta_{j,k}=1}(x_{i,k}-x_{j,k})^2\right]$$

Finally, define the estimated cumulative distribution function by

$$\widehat{H}_{XX}(t) = \sum_{1\leq i<j\leq n}\sum_{\substack{t' \in support \\ t'\leq t}} h^{i,j}_{XX}(t')$$

The third interpoint distance distribution, denoted H_{XY}, is estimated similarly. The estimate is

$$\widehat{H}_{XY}(t) = \sum_{i=1}^{n}\sum_{\substack{t' \in support \\ t'\leq t}} h^{i}_{XY}(t')$$

where $h^i_{XY}(t_1) = \frac{P[\|\mathrm{x}_i-\mathrm{y}\|\leq t_1]}{n}$ and $h^i_{XY}(t_k) = \frac{P[\|\mathrm{x}_i-\mathrm{y}\|\leq t_k]-P[\|\mathrm{x}_i-\mathrm{y}\|<t_{k-1}]}{n}$ for $k>1$.

The generalization of the Kolmogorov-Smirnov test used here will be:

$$L_{KS} = max\left\{\max_k\left|\widehat{H}_{XY}(t_k)-\widehat{H}_{XX}(t_k)\right|, \max_k\left|\widehat{H}_{XY}(t_k)-H_{YY}(t_k)\right|, \max_k\left|H_{YY}(t_k)-\widehat{H}_{XX}(t_k)\right|\right\}$$

As with the energy test, the distribution of the statistic is found under the null hypothesis by conditioning on the missing data structure, sampling from G, standardizing the sample by the estimated mean and variance, and then calculating the statistic based on the standardized sample. The testing procedure proceeds by finding the p-value by the estimated probability of exceeding the observed value of the test statistic under the null hypothesis.

The second new test is also based on comparing the three interpoint distance distribution functions. H_{XX} and H_{YY} are defined by interpoint distances within the same distribution (either F or G) while H_{XY} is defined by the interpoint distance between vectors with possibly different distributions. The test statistic defined by

$$L = \max_k\left\{\frac{\widehat{H}_{XX}(t_k)+H_{YY}(t_k)}{2} - \widehat{H}_{XY}(t_k)\right\}$$

compares the average within interpoint distance distribution to the between. At first at first seem strange to only define the statistic in one direction rather than the maximum absolute value of the difference, but we will show later that only one direction is needed. The tests using either L_{KS} or L are consistent when there are no missing components or if there are missing components that are missing completely at random.

Both L_{KS} or L are reasonable goodness-of-fit tests because they are consistent when there are no missing components. A mathematical proof follows.

Theorem 2. If there are no missing components,

$$L_{KS} = max\left\{ \max_k \left|\widehat{H}_{XY}(t_k) - \widehat{H}_{XX}(t_k)\right|, \max_k \left|\widehat{H}_{XY}(t_k) - H_{YY}(t_k)\right|, \max_k \left|H_{YY}(t_k) - \widehat{H}_{XX}(t_k)\right| \right\}$$

and $L = \max_k \left\{ \frac{\widehat{H}_{XX}(t_k)+H_{YY}(t_k)}{2} - \widehat{H}_{XY}(t_k) \right\}$ are consistent tests of the null hypothesis $F = G$.

Proof. If there are no missing components, then under the null hypothesis, both L_{KS} and L converge to 0 by the theorem in Maa et al [8]. The variance of both statistics converges to 0 [Silverman (12)]. Also, under any alternative, L_{KS} does not converge to 0. Therefore, L_{KS} is consistent.

The consistency of L hinges on showing that under any alternative, there is some t for which $\frac{H_{XX}(t)+H_{YY}(t)}{2} - H_{XY}(t) > 0$. If that were true, then L would converge to a positive number and the consistency is established. Suppose that there is no such t; in other words, $\frac{H_{XX}(t)+H_{YY}(t)}{2} - H_{XY}(t) \leq 0$ for all t. Then, using the Darth Vader rule [9], it would follow that

$$0 \geq \int_0^\infty \frac{H_{XX}(t) + H_{YY}(t)}{2} - H_{XY}(t)\, dt = -\int_0^\infty \frac{1 - H_{XX}(t) + 1 - H_{YY}(t)}{2} - (1 - H_{XY}(t))\, dt$$

$$= -\left\{ \frac{E\|\mathrm{x}_1 - \mathrm{x}_2\| + E\|\mathrm{y}_1 - \mathrm{y}_2\|}{2} - E\|\mathrm{x}_1 - \mathrm{y}_1\| \right\} = E\|\mathrm{x}_1 - \mathrm{y}_1\| - \frac{E\|\mathrm{x}_1 - \mathrm{x}_2\| + E\|\mathrm{y}_1 - \mathrm{y}_2\|}{2}$$

But, Corollary 1 of Székely and Rizzo (13) states that the last expression is strictly positive under any alternative.

3. Simulation comparison

Data are simulated from several types of multivariate distributions. Multivariate normal, MVN, and multivariate T, MVT_ν, vectors are simulated using the mvtnorm package [Genz et al (4), Genz and Bretz (5)]. The notation $\overrightarrow{a}$ represents a vector of length d with all coordinates equal to a, I is the dxd identity matrix, $CS(\rho)$ is a compound symmetric matrix where the diagonal elements are 1 and all other elements are ρ. If there is a subscript, then that indicates the length of the vector or size of the matrix. For example, $\overrightarrow{0}_5$, is the vector of length 5 with all elements equal to 0. If there is no subscript, the size is understood to be $d = 10$. Mixture distributions are included with the notation p MVN+ $(1-p)\, MVT_\nu$. A vector where the first components are multivariate normal and the last components are multivariate T is denoted $MVN \otimes MVT_\nu$. The missing data is artificially created so that 15% of vectors, chosen completely at random, have missing components. Within those vectors, data are monotonically missing for the final 4 coordinates $d-3, \ldots, d$. For each scenario,

the power is estimated using 5 thousand replications. Estimation of the mean and covariance of the population is done using the sample mean and covariance of the complete cases. The type 1 error rate is 0.05 and the null hypothesis is rejected if the observed value of the statistic is larger than the 95^{th} percentile of the distribution under the null hypothesis. By construction of the critical value used, the type 1 error rate under the null hypothesis will always be 0.05.

Table 1 shows the estimated power of each test under different alternative scenarios with d =10. In scenarios shown in rows 1, 2, and 7 of the table, all 3 tests have similar power. In the scenarios in rows 3 and 5, L has higher power. The energy test has higher power than the other tests in the scenario in row 4. For the remaining scenarios, the test statistic L has the highest power. In scenario 6, the test statistic L_{KS} has the lowest power and the other two tests have equal power. Although this is not a comprehensive comparison, it demonstrates that the three tests do not always have the same power for all alternatives.

Table 1. Estimated power (probability of rejecting the null hypothesis of multivariate normality) using modified energy test or Three Sample Kolmogorov-Smirnov Test of interpoint distance distributions (5,000 replications, $\alpha = 0.05, d = 10$).

Distribution F	n	Energy Test	Power of L_{KS}	Power of L
$MVT_{20}\left(\overrightarrow{0}, I\right)$	40	7	6	6
$0.5MVN\left(\overrightarrow{0}, I\right) + 0.5MVN\left(\overrightarrow{3}, I\right)$	40	5	4	3
$MVT_{20}\left(\overrightarrow{0}, I\right)$	100	14	14	27
$\frac{3}{4}MVN\left(\overrightarrow{0}, I\right) + \frac{1}{4}MVN\left(\overrightarrow{3}, I\right)$	100	9	4	4
$\frac{1}{2}MVN\left(\overrightarrow{0}, I\right) + \frac{1}{2}MVT_{10}\left(\overrightarrow{0}, I\right)$	100	9	8	17
$\frac{1}{2}MVN\left(\overrightarrow{0}, I\right) + \frac{1}{2}MVT_{10}\left(\overrightarrow{0}, I\right)$	100	58	35	58
$MVN\left(\overrightarrow{0}_5, I_5\right) \otimes MVT_{10}\left(\overrightarrow{0}_5, I_5\right)$	100	16	15	15

4. Kidney trial data analysis

The TEMPO 3:4 trial [14] randomized patients with autosomal dominant polycystic kidney disease to either the experimental drug tolvaptan or placebo. The patients were followed for 3 years and their estimated kidney function was measured every 4 months. Thus, patients who did not miss any measurements had 9 values measured including the baseline value at month 0. A mixed effects model was used that has fixed effects treatment arm, age at baseline, months from randomization, treatment by month interaction, and baseline total kidney volume. The model has random effects for intercept and months within patients.

Here, the focus is on the patients from the placebo arm who were in stage 3A at baseline (that is, estimated kidney function between 45 and 60 mL/min per 1.73 m^2 body surface area). This subgroup was chosen in order to attempt to make the patients more homogenous. The residuals from the fitted mixed effects model for this subgroup are used to illustrate the goodness of fit test for multivariate normality. There were

66 patients in this subgroup. 58 of those subjects had complete data (all 9 possible measurements observed); 5 had 8 measurements; 1 had 6 measurements; 2 had 4 measurements. The null hypothesis is that these residuals have a multivariate normal distribution and the alternative hypothesis is that they have some other distribution.

The estimated differences in the cumulative distribution functions for the interpoint distances, that is $\widehat{H}_{XX}(t_k) - \widehat{H}_{XY}(t_k)$ (in red) and $H_{YY}(t_k) - \widehat{H}_{XY}(t_k)$ (in black), are shown in Figure 1. The third estimated difference, $\widehat{H}_{XX}(t_k) - H_{YY}(t_k)$, is not shown directly but it can be found by subtracting y-values in the two curves. Furthermore, $\frac{\widehat{H}_{XX}(t_k)+H_{YY}(t_k)}{2} - \widehat{H}_{XY}(t_k)$ is the average of the two y-values of the curves shown. Using all 66 subjects, the p-values of these new tests including all 66 subjects are 0.14 (energy), 0.08 (L), and 0.59 (L_{KS}). In contrast, if the energy test is used with only the 58 subjects with complete data, the p-value is 0.01. The relatively large difference in the p-values illustrates that it can be important to use all of the data and not ignore the data with missing components. Moreover, it corroborates what was found in the simulation section where L_{KS} was less powerful than the other two statistics.

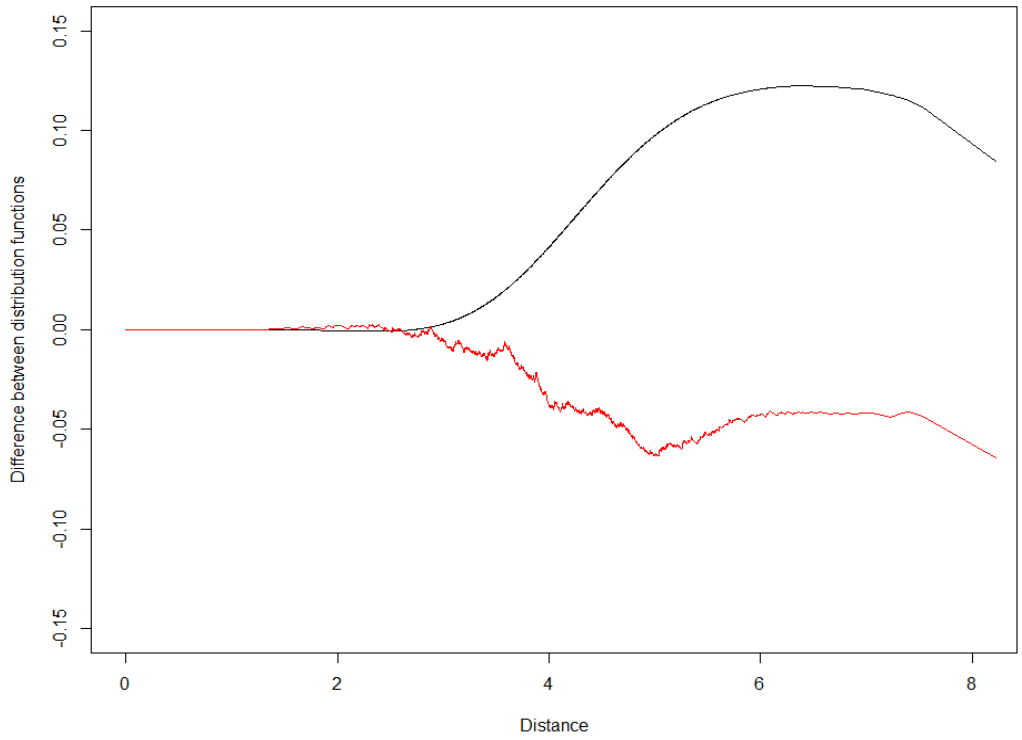

Figure 1. Estimated differences in interpoint distance distribution for kidney disease data. The red curve is $\widehat{H}_{XX}(t_k) - \widehat{H}_{XY}(t_k)$ and the black curve is $H_{YY}(t_k) - \widehat{H}_{XY}(t_k)$.

5. Discussion

Three multivariate goodness-of-fit tests from the interpoint distance were derived. All three tests developed here are applicable for the case when the data are complete or

when some vectors have missing components. The first is a modification of the energy test devised by [13]. The estimate of the mean distance between observed sample points was replaced by an unbiased estimate when the data was complete. The other two tests are completely new and are derived from a theorem that states that two multivariate distributions are equal if and only if the three interpoint distance distributions are equal [8]. Although the tests seem to be derived from the same basic principle, they exhibit different power for some alternative scenarios in the simulation.

A real dataset was analyzed using all three tests. The data consists of the residuals from a mixed effects model from a drug trial used to treat a type of chronic kidney disease. Larger subgroups and the entire study population were also analyzed (not shown here), but those cases were not very interesting because all the tests clearly demonstrated the data were not multivariate normal. For the subgroup, the complete cases were first tested for multivariate normality and it was found that there was fairly strong evidence against the null hypothesis ($p=0.01$). However, when the entire dataset was analyzed, the null hypothesis could not be rejected ($p>0.05$ for all three tests).

These tests can be used in situations where the null hypothesis is a specific multivariate normal distribution or in the situation where there is a composite null hypothesis of multivariate normality with arbitrary mean and covariance. None of the tests are distribution-free. Hence, the distribution of the test statistic under the null hypothesis must be estimated using simulation. For the composite null hypothesis scenario, the observed data and each simulated dataset are standardized using estimates of the mean and covariance. For multivariate data, it is not possible to view Q-Q plots or similar diagnostic plots. However, Figure 1 is an informative plot that can be viewed to assess whether the interpoint distance distributions are equal and if not, in what ways they differ.

References

[1] Bartoszyński, Robert, Dennis K Pearl, John Lawrence.: A multidimensional goodness-of-fit test based on interpoint distances. J Am Stat Assoc 1997; 92: 577-586.

[2] Bonetti, Marco, and Marcello Pagano.: The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. *Statistics in medicine* 24.5 (2005): 753-773.

[3] David, Herbert T. 1958.: A three-sample Kolmogorov-Smirnov test. Ann. Math. Statist. 29 842:851.

[4] Genz, Alan, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Torsten Hothorn (2020): mvtnorm: Multivariate Normal and t Distributions. R package version 1.1-0. URL `http://CRAN.R-project.org/package=mvtnorm`

[5] Genz, Alan and Frank Bretz. (2009): Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195. Springer-Verlag, Heidelberg.

[6] Lawrence, John.: Moments of the noncentral chi distribution Sankhya A 85.2 (2023): 1243-1259.

[7] Lawrence, John.: Distribution of distances between random vectors and two fixed points. In submission.

[8] Maa, Jen-Fue, Dennis K. Pearl, and Robert Bartoszyński.: Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *The annals of statistics* 24.3 (1996): 1069-1074.

[9] Muldowney, Pat, Krzysztof Ostaszewski, and Wojciech Wojdowski.: The darth vader rule. Tatra Mountains Mathematical Publications 52, no. 1 (2012): 53-63.
[10] National Research Council.: The prevention and treatment of missing data in clinical trials. National Academies Press, 2010.
[11] Pearl, Dennis K., et al.: High-dimensional simulation-based estimation. Mathematical and computer modelling 32.1-2 (2000): 27-51.
[12] Silverman, Bernard W.: Limit theorems for dissociated random variables.Advances in Applied Probability 8, no. 4 (1976): 806-819.
[13] Székely, Gábor J., and Maria L. Rizzo.: A new test for multivariate normality. Journal of Multivariate Analysis 93, no. 1 (2005): 58-80.
[14] Torres, Vicente E., et al.: Tolvaptan in patients with autosomal dominant polycystic kidney disease. New England Journal of Medicine 367, no. 25 (2012): 2407-2418.