Journal of Applied Statistics, 2013
Vol. 40, No. 5, 1120–1131, http://dx.doi.org/10.1080/02664763.2013.780160

Taylor & Francis
Taylor & Francis Group

An empirical goodness-of-fit test for multivariate distributions

Michael P. McAssey*

Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

(*Received 9 June 2012; final version received 23 February 2013*)

An empirical test is presented as a tool for assessing whether a specified multivariate probability model is suitable to describe the underlying distribution of a set of observations. This test is based on the premise that, given any probability distribution, the Mahalanobis distances corresponding to data generated from that distribution will likewise follow a distinct distribution that can be estimated well by means of a large sample. We demonstrate the effectiveness of the test for detecting departures from several multivariate distributions. We then apply the test to a real multivariate data set to confirm that it is consistent with a multivariate beta model.

Keywords: Mahalanobis distance; multivariate beta distribution; multivariate goodness-of-fit test; multivariate normal distribution

1. Introduction

In a recent paper [14], measurements of instantaneous coupling (IC) were computed between pairs of electroencephalogram signals in the gamma frequency band at selected tetrodes implanted in different regions of the brain of a rat. It was assumed that, upon selecting one tetrode as a reference, the distribution of its IC measurements with respect to any subset of the remaining tetrodes may be modeled with a multivariate beta distribution. While this assumption was intuitively valid based on inspection of univariate histograms, its validity was not verified. This was a consequence of the unavailability of a practical goodness-of-fit test for a multivariate beta distribution.

In fact, the literature on the topic of general multivariate goodness-of-fit tests is scarce. Tests for multivariate normality are abundant, beginning with the seminal work of Pearson on the chi-square goodness-of-fit test [1,15,19] and continuing with a multitude of additional approaches, *e.g.* application of the Rosenblatt transformation [22] to examine multivariate normality [21], tests using multivariate measures of skewness and kurtosis [11–13,25,26], a test based on the multivariate Shapiro–Wilk statistic [24], a radii and angles test [8], a test based on the multivariate Box–Cox transformation [28], and many other creative methods [2,5,16]. But these tests do not

*Email: m.p.mcassey@vu.nl

extend readily to the general case. Proposals for extending the Kolmogorov–Smirnov goodness-of-fit test to multiple dimensions have been published [4,6,9,18], but the required test statistic in each proposed method is extremely difficult to compute, even in the bivariate case, and a suggested simplification in [6] still requires a transformation whose derivation is analytically intractable for most multivariate distributions. Székely and Rizzo [27] proposed a test that is applicable to any multivariate distribution having finite second moments, but the test is applied only to the multivariate normal, as analytical derivation of the test statistic in other settings is likewise unmanageable. A multivariate goodness-of-fit test based on the empirical characteristic function has also been developed [3]. However, derivation of this test statistic is also not tractable for most multivariate distributions, and one is additionally required to creatively choose a weight matrix and a number of evaluation points.

What is needed is a goodness-of-fit test that is theoretically sound, is simple to implement in scientific applications, is adaptable to any multivariate distribution of any dimension, and has sufficient power. We propose below such an approach, with a focus on continuous multivariate distributions. We compare the performance of the proposed test with that of several established tests for multivariate normality to demonstrate its reliability in that setting, and then demonstrate its effectiveness for testing the suitability of the multivariate uniform and beta models. Finally, we apply this test to the IC measurement data referenced above to confirm the original multivariate beta assumption.

2. Method

Let $\mathbf{X} \in \Re^p$ be a sample from a population having known p-variate continuous distribution $F = F_{\boldsymbol{\theta}}$, with $\boldsymbol{\theta}$ in the parameter space Θ of F, and set $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathcal{E}_F(\mathbf{X})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathcal{V}_F(\mathbf{X})$. Recall [10] that the Mahalanobis distance $\Delta \in \Re$ between $\mathbf{X}$ and $\boldsymbol{\mu}$ is a continuous function from $\Re^p \to [0, \infty)$ computed as

$$\Delta = \Delta(\mathbf{X}|\ \boldsymbol{\theta}) = \|\boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})\| = \sqrt{(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})},$$

where $\|\cdot\|$ denotes the Euclidean norm. Given $t > 0$, the transformation Δ maps all points $\mathbf{x}$ lying on the p-dimensional ellipsoid

$$E_t = \{\mathbf{x} \in \Re^p \mid \Delta(\mathbf{x} \mid \boldsymbol{\theta}) = t\}$$

onto a p-dimensional sphere S_t of radius t centered at the origin. Define $G(t) = \mathcal{P}(\Delta \leq t) = \mathcal{P}(\Delta \in S_t)$. Hence, $\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{X} \in E_t) = \mathcal{P}_{\boldsymbol{\theta}}(\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})\| \leq t) = G(t)$.

Now let $\mathcal{T}$ denote the set of all invertible affine transformations from $\Re^p$ to $\Re^p$. Thus, for $T \in \mathcal{T}$, $T(\mathbf{X}) = \mathbb{A}\mathbf{X} + \mathbf{b}$ for some invertible $p \times p$ matrix $\mathbb{A}$ and some $\mathbf{b} \in \Re^p$. If $\mathbf{Y} = T(\mathbf{X})$, then $\mathbf{Y}$ has mean $\mathbb{A}\boldsymbol{\mu} + \mathbf{b}$ and variance $\mathbb{A}\boldsymbol{\Sigma}\mathbb{A}'$. However, although $X \sim F_{\boldsymbol{\theta}}$, the distribution of Y is in most settings not described by F for any choice of parameter. Nevertheless, given each $t > 0$, the transformation T results in a corresponding transformation of the ellipsoid E_t in $\Re^p$, producing the ellipsoid $T(E_t)$. Then, $\mathcal{P}(Y \in T(E_t)) = \mathcal{P}_{\boldsymbol{\theta}}(X \in E_t)$. It is straightforward to show that $\Delta(Y) = \Delta(X)$ when $T \in \mathcal{T}$:

$$\begin{aligned}\Delta(Y) &= \sqrt{(\mathbf{Y} - \mathbb{A}\boldsymbol{\mu} - \mathbf{b})'(\mathbb{A}\boldsymbol{\Sigma}\mathbb{A}')^{-1}(\mathbf{Y} - \mathbb{A}\boldsymbol{\mu} - \mathbf{b})},\\ &= \sqrt{(\mathbb{A}\mathbf{X} + \mathbf{b} - \mathbb{A}\boldsymbol{\mu} - \mathbf{b})'(\mathbb{A}\boldsymbol{\Sigma}\mathbb{A}')^{-1}(\mathbb{A}\mathbf{X} + \mathbf{b} - \mathbb{A}\boldsymbol{\mu} - \mathbf{b})},\\ &= \sqrt{(\mathbf{X} - \boldsymbol{\mu})'\mathbb{A}'(\mathbb{A}')^{-1}\boldsymbol{\Sigma}^{-1}\mathbb{A}^{-1}\mathbb{A}(\mathbf{X} - \boldsymbol{\mu})},\end{aligned}$$

$$= \sqrt{(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})},$$
$$= \Delta(X). \tag{1}$$

Hence, $\mathcal{P}(\Delta(Y) \leq t) = \mathcal{P}(\Delta(X) \leq t) = G(t)$ for all t. However, if $T \notin \mathcal{T}$, then the equality 1 does not hold, so that $\mathcal{P}(\Delta(Y) \leq t) \neq \mathcal{P}(\Delta(X) \leq t)$ for all t.

Therefore, given $F_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \Theta$, let Ξ denote the set of all distributions Φ such that, whenever $Y \sim \Phi$, there exists some $T \in \mathcal{T}$ such that $T^{-1}(Y) \in F_{\boldsymbol{\theta}}$. When $F_{\boldsymbol{\theta}}$ is a p-variate multivariate normal distribution, Ξ consists of all p-variate multivariate normal distributions. But, in general, the subset of elements of Ξ that belongs to the same distribution family as $F_{\boldsymbol{\theta}}$ has measure zero. Now suppose G is the distribution of Mahalanobis distances corresponding to Ξ, and G' is the distribution of Mahalanobis distances corresponding to some unknown distribution F'. On the basis of the foregoing discussion, we can conclude the following: If $G' \neq G$, then $F' \notin \Xi$, and if $F' \notin \Xi$, then $G' \neq G$. Consequently, a goodness-of-fit test for G is equivalent to a goodness-of-fit test for members of Ξ. In the multivariate normal and many other symmetric settings, a goodness-of-fit test for G will suffice for testing whether F' belongs to the same distribution family as F. In asymmetric settings, such as the multivariate beta, one must employ further analysis when the null hypothesis is not rejected to ensure that the hypothesized distribution is valid.

Consequently, we may generate a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from a known distribution F, compute $\Delta_1, \ldots, \Delta_n$, then estimate G via the corresponding empirical distribution function

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\Delta_i \leq t), \quad t > 0.$$

As $n \to \infty$, $G_n(t) \to G(t)$, by well-established theoretical results. If data $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ are generated from an unknown distribution F' belonging to Ξ, then the corresponding empirical distribution G'_m should not differ significantly from G_n. However, if $F' \notin \Xi$, then we should detect a significant difference between G_n and G'_m provided m and n are sufficiently large.

Figure 1 displays plots of $G_n(t)$ for samples randomly generated from four different bivariate distributions, with $n = 10,000$. The bivariate normal has a mean of $\boldsymbol{\mu} = (-1, 2)$ and a covariance of

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & -2 \\ -2 & 9 \end{bmatrix}.$$

The bivariate uniform consists of two independent samples, one from a uniform distribution of $(0, 1)$ and the other from a uniform distribution on $(0, 5)$. The bivariate Student's-t also consists of two independent samples, from t distributions with three and five degrees of freedom, respectively. The bivariate beta likewise consists of independent univariate beta samples, the $\mathrm{Be}(2, 3)$ and the $\mathrm{Be}(3, 2)$. For each bivariate distribution, we compute $\Delta_1, \ldots, \Delta_n$ based on the respective values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then obtain the corresponding functions $G_n(t)$. Since the sample size is quite large in each case, we have confidence that $G_n(t)$ is very near $G(t)$ for each example. As the figure demonstrates, the values of $G_n(t)$ among the four distributions are quite different, reflecting the intrinsic distinction in the scatter structure of the samples about their respective means. Note, however, that the distinction between the curves for the bivariate normal and the bivariate beta is not as sharp.

To test for a significant departure from G, we proceed as follows. Given any $p \in (0, 1)$, let q_p denote the p-quantile of G, i.e. $G(q_p) = \mathcal{P}(\Delta < q_p) = p$. Given any partition $0 = p_0 < p_1 < \cdots < p_T = 1$ of $(0, 1)$, we then obtain a corresponding partition $0 = q_{p_0} < q_{p_1} < \cdots < q_{p_T} = \infty$

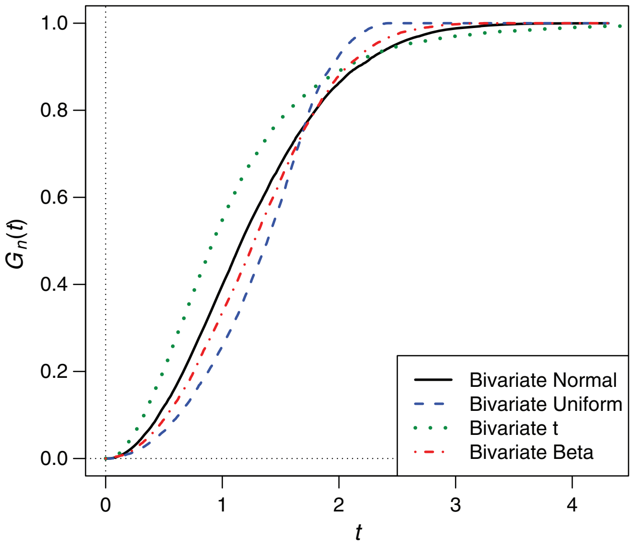

Figure 1. Plots of $G_n(t)$, with $n = 10,000$, for samples randomly generated from a bivariate normal distribution, a bivariate uniform distribution, a bivariate Student's-t distribution, and a bivariate beta distribution.

such that

$$P_j = G(q_{p_j}) - G(q_{p_{j-1}}) = p_j - p_{j-1}, \quad j = 1, \dots, T,$$

where $G(\infty) = 1$. Meanwhile, let $q_{p,n}$ denote the p-quantile of G_n, defined as

$$q_{p,n} = \min\{t \in \Re \mid G_n(t) \geq p\}.$$

As $n \to \infty$, $q_{p,n} \to q_p$, since $G_n(t) \to G(t)$. So, if we set

$$P_{j,n} = G_n(q_{p_j,n}) - G_n(q_{p_{j-1},n})$$

for $j = 1, \dots, T$, it is clear that $P_{j,n} \to P_j$ as $n \to \infty$. Thus, when n is very large, $P_{j,n}$ is a reliable estimate of P_j.

Consequently, given $G_n(t)$ with n large, and a new set of observations $y_1, \dots, y_m \in (0, \infty)$, one can test the null hypothesis that these data are realizations from the distribution G by

(1) selecting a partition $\{p_0, p_1, \dots, p_T\}$ as described above, with T large (but not so large that $mP_{j,n}$ becomes too small for any j),
(2) computing $E_j = mP_{j,n}$, i.e. the (approximate) expected number of observations in the interval $(q_{p_{j-1}}, q_{p_j}]$ under the null hypothesis, for $j = 1, \dots, T$,
(3) counting the number O_j of observations among $y_1, \dots, y_m$ in the interval $(q_{p_{j-1}}, q_{p_j}]$, for $j = 1, \dots, T$,
(4) calculating an appropriate test statistic to measure the global deviation of the counts of observed data from the expected counts among the intervals, e.g.

$$A_T = \sum_{j=1}^{T} \frac{|E_j - O_j|}{E_j},$$

(5) deciding whether A_T is too large to be plausible under the null hypothesis.

The basis for determining what constitutes 'too large' must be established empirically, since the distribution of A_T cannot be determined analytically.

However, our interest in this paper is not in deciding whether observed univariate data are realizations from G, but rather in deciding whether observed multivariate data are realizations from the distribution family corresponding to some $F_{\boldsymbol{\theta}}$. Moreover, while we shall assume that the mean and covariance of $F_{\boldsymbol{\theta}}$ exist, we shall not assume that $\boldsymbol{\theta}$, $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$ are known. We only require that the maximum-likelihood estimate (MLE) of $\boldsymbol{\theta}$ can be computed based on a sample drawn from $F_{\boldsymbol{\theta}}$. Consequently, let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathfrak{R}^p$ denote a sample from some unknown p-variate continuous distribution Φ. Let Ξ be the set of distributions generated by all invertible affine transformations of data generated from $F_{\boldsymbol{\theta}}$, as described above. We test

$$H_0 : \Phi \in \boldsymbol{\Xi}$$

against

$$H_1 : \Phi \notin \boldsymbol{\Xi}$$

at some specified significance level α. Of course, this assumes that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ lie within the support of some element of Ξ. Let $\bar{\mathbf{X}}$ and $\mathbb{S}$ denote the sample mean and sample covariance matrix, respectively. Then, the sample Mahalanobis distance D_i between $\mathbf{X}_i$ and $\bar{\mathbf{X}}$ is given by the familiar form

$$D_i = \sqrt{(\mathbf{X}_i - \bar{\mathbf{X}})'\mathbb{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})}, \quad i = 1, \ldots, n.$$

Since $\bar{\mathbf{X}} \xrightarrow{\text{P}} \boldsymbol{\mu}$ and $\mathbb{S} \xrightarrow{\text{P}} \boldsymbol{\Sigma}$ as $n \to \infty$ under the null hypothesis, the continuous mapping theorem guarantees that $D_i \xrightarrow{\text{P}} \Delta_i$ as $n \to \infty$. Now consider the empirical distribution function $\hat{G}_n(t)$ given by

$$\hat{G}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(D_i \leq t), \quad t > 0,$$

which is the sample version of $G_n(t)$. Under the null hypothesis, at each $t > 0$, $\mathbb{1}(D_i \leq t) \xrightarrow{\text{P}} \mathbb{1}(\Delta_i \leq t)$ as $n \to \infty$, so that the difference between $\hat{G}_n(t)$ and $G_n(t)$ will tend to zero as n increases. Since G_n converges to G as $n \to \infty$, we conclude that $\hat{G}_n$ likewise converges to G when H_0 holds.

Our test of H_0 is thus linked to a test of H_0': $D_1, \ldots, D_n \sim G$. If we reject H_0', we must also reject H_0. Otherwise, we do not reject H_0, and investigate further. But we cannot proceed exactly as when $F_{\boldsymbol{\theta}}$ is completely specified. Instead, we follow a longer, modified procedure. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathfrak{R}^p$ denote the observed multivariate data, and $d_1, \ldots, d_n$ the corresponding observed sample Mahalanobis distances. First check that the hypothesized distribution $F_{\boldsymbol{\theta}}$ is sensible for explaining the observed multivariate data. Then, execute the following steps:

(1) Assuming H_0 is true, estimate $\boldsymbol{\theta}$ with its MLE $\hat{\boldsymbol{\theta}}$ under $F_{\boldsymbol{\theta}}$ based on $\mathbf{x}_1, \ldots, \mathbf{x}_n$;
(2) Generate a very large sample $\mathbf{u}_1, \ldots, \mathbf{u}_N$ of size $N \gg n$ from $F_{\hat{\boldsymbol{\theta}}}$ (we use $N = 10,000$);
(3) Determine the sample mean $\hat{\boldsymbol{\mu}}$ and sample covariance $\hat{\boldsymbol{\Sigma}}$ for this very large sample;
(4) Compute the sample Mahalanobis distances $\hat{d}_1, \ldots, \hat{d}_N$ for this very large sample, where

$$\hat{d}_i = \sqrt{(\mathbf{u}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{u}_i - \hat{\boldsymbol{\mu}})};$$

(5) Compute

$$\hat{G}_N(t) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{d}_i \le t), \quad t > 0,$$

our estimate of $G(t)$, using very small increments of t from 0 to a sufficiently large value (we use $2\max_i \hat{d}_i$);

(6) Select a partition $\{p_0, p_1, \dots, p_T\}$ of [0, 1], with $p_0 = 0$ and $p_T = 1$ (where T is selected so that a sufficient number of observations are expected per bin, e.g., $T = 5$). Then, estimate the corresponding p-quantiles of $G(t)$ by setting

$$q_0 = 0, \quad q_j = \min\{t \in \Re \mid \hat{G}_N(t) \ge p_j\}, \quad j = 1, \dots, T-1, \quad \text{and} \quad q_T = \infty$$

(7) Compute $E_j = n(p_j - p_{j-1})$, i.e. the expected number of observations in the interval $(q_{j-1}, q_j]$ under H_0, for $j = 1, \dots, T$;

(8) Count the actual number O_j of observations among $d_1, \dots, d_n$ in the interval $(q_{j-1}, q_j]$, for $j = 1, \dots, T-1$, and in the interval (q_{T-1}, ∞) for $j = T$;

(9) Calculate an appropriate test statistic to measure the global deviation of the counts of observed data from the expected counts among the intervals, e.g.

$$A_T = \sum_{j=1}^{T}\frac{|E_j - O_j|}{E_j}.$$

Since A_T depends on T and N, A_T is not exact.

(10) Determine whether A_T is too large under H_0, based on an empirical p-value which can be obtained using a procedure to be described below.

Since the sample Mahalanobis distances computed in Step (4) depend on the sample $\mathbf{u}_1, \dots, \mathbf{u}_N$ generated in Step (2), which in turn affects all subsequent computations, we do not obtain consistent output at Step (9) for the statistic A_T unless we repeat Steps (2)–(5) R times, where R is not too small (we use $R = 100$, although smaller values are probably sufficient). Then, at Step (6), we use the pointwise average of $\hat{G}_N(t)$ over these R repetitions in order to obtain consistent values for the quantiles $q_0, \dots, q_T$, and ultimately for A_T.

Step (10) requires a simulation to obtain the empirical p-value. The idea is to obtain an empirical distribution (in lieu of an exact distribution) of values for A_T when H_0 holds, and determine whether the statistic A_T computed in Step (10) is plausible under this distribution. This is done in the setting where a sample of the same size n is generated from $F_{\hat{\boldsymbol{\theta}}}$, but after the sample is generated from $F_{\hat{\boldsymbol{\theta}}}$, its sample mean and sample covariance are used as was done with the original data. Then, for $b = 1, \dots, B$, where B is quite large, repeat the following procedure:

(a) Generate a sample $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$ from $F_{\hat{\boldsymbol{\theta}}}$;

(b) Obtain the sample mean and sample covariance, then compute the sample Mahalanobis distances $d_1^*, \dots, d_n^*$;

(c) Estimate $\hat{\boldsymbol{\theta}}$ with its MLE $\hat{\boldsymbol{\theta}}^*$ based on $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$;

(d) Follow Steps (2)–(9) (substituting $\hat{\boldsymbol{\theta}}^*$ for $\hat{\boldsymbol{\theta}}$ at Step (2), and substituting d_i^* for d_i at Step (8)) to obtain test statistic $A_{T,b}$. Again, repeat Steps (2)–(5) R times and use the average for $\hat{G}_N(t)$ at Step (6).

Having obtained $A_{T,1}, \dots, A_{T,B}$, the empirical p-value is then

$$p_e = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(A_{T,b} > A_T).$$

Therefore, Step (10) becomes: Reject H_0 if and only if $p_e < \alpha$. If H_0 is not rejected, additional work may be required to decide whether the observed data can be described adequately by $F_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta}$, depending on the symmetry of F.

We demonstrate the simplicity and effectiveness of this method in the next section.

3. Results

To illustrate the effectiveness of the method in multivariate goodness-of-fit testing, we first conduct a test of the null hypothesis

$$H_0 : \Phi = \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ for some } (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

against the alternative

$$H_1 : \Phi \neq \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ for any } (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

based on samples of size $n = 100$, where Φ denotes the true distribution family from which the data are sampled (Since the multivariate normal family is closed under affine transformations, Ξ is equivalent to the set of all bivariate normal distributions.). We conduct this test using the method presented above and, for comparison, four established methods: the multivariate Shapiro–Wilk test [23,24], the energy test of Székely and Rizzo for multivariate normality [27], and the tests of Kankainen *et al.* [7], based on skewness and on kurtosis. We generate samples from 10 distinct bivariate normal, bivariate uniform, and bivariate Student's-t distributions, with the parameters randomly selected for each of the 30 samples. We implement our method for each sample with $N = 10,000$, $R = 100$, $T = 20$ and $B = 100$. We then compute the mean empirical p-value for each method over the 10 samples of size 100 corresponding to each distribution family. We expect large empirical p-values for samples from the bivariate normal family, and small empirical p-values for the samples from the other two families. We also compute the false detection rate for each method, based on a significance level of $\alpha = 0.05$, for the bivariate normal samples, and the true detection rate for the bivariate uniform and Student's-t samples. Our results are as follows:

True distribution	Our method	Shapiro–Wilk	Energy	Skewness	Kurtosis
Mean empirical p-value					
Bivariate normal	0.504	0.320	0.487	0.417	0.486
Bivariate uniform	0.044	0.356	0.005	0.759	0.010
Bivariate Student's-t	0.148	0.025	0.015	0.126	181
Detection rates at $\alpha = 0.05$					
Bivariate normal	0.1	0.1	0.0	0.1	0.1
Bivariate uniform	0.7	0.0	1.0	0.0	1.0
Bivariate Student's-t	0.6	0.9	0.9	0.7	0.8

The energy test performs the best in detecting (and not falsely detecting) departures from normality for all three bivariate distribution families, followed closely by the kurtosis test, while the Shapiro–Wilk test outperforms our method only in the case of the Student's-t distribution.

But our method performs reasonably well in the case of the uniform distribution, while Shapiro–Wilk and the skewness test do very poorly. Overall, our method shows more promise than the multivariate Shapiro–Wilk test and the skewness test in this setting.

Next, we increase the dimension and test the null hypothesis

$$H_0 : \Phi = \mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ for some } (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

against the alternative

$$H_1 : \Phi \neq \mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ for any } (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

based on samples of size $n = 100$, where Φ denotes the true distribution family from which the data are sampled. We repeat the above simulation in this trivariate setting, and obtain the following results:

True distribution	Our method	Shapiro–Wilk	Energy	Skewness	Kurtosis
Mean empirical p-value					
Trivariate normal	0.431	0.225	0.447	0.583	0.584
Trivariate uniform	0.000	0.577	0.007	0.800	0.035
Trivariate Student's-*t*	0.061	0.004	0.011	0.068	0.003
Detection rates at $\alpha = 0.05$					
Trivariate normal	0.2	0.4	0.1	0.0	0.0
Trivariate uniform	1.0	0.0	1.0	0.0	0.9
Trivariate Student's-*t*	0.9	1.0	0.9	0.7	1.0

While the energy test again yields the best overall performance, followed by the kurtosis test, the performance of our method is almost identical. The Shapiro–Wilk test and the skewness test are once again unable to distinguish the uniform family from the normal family, while the other three methods succeed in almost all 10 samples of size 100. Our conclusion is that, with respect to goodness-of-fit tests for multivariate normality, the accuracy of our method is comparable to that of existing tests, and better in some situations. But our goodness-of-fit test is easily extended to other multivariate distribution families, while the other tests are not, in general, so versatile. We emphasize that the multivariate goodness-of-fit test proposed here is not intended as a competitor to the dozens of superb tests for multivariate normality available in the literature. It is intended as an accessible tool for evaluating goodness-of-fit to multivariate distributions which are not normal, concerning which alternative tractable methods cannot be found. We only compare the performance of our test to other tests in the multivariate normality setting in order to demonstrate its reliability, since we have no tractable tests against which to compare it in other settings.

As an additional example, we generate a sample of size $n = 100$ from a trivariate uniform distribution in the box $[0, 2] \times [-1, 1] \times [0, 1]$, and then use our approach to test the goodness-of-fit of the trivariate uniform distribution to these data. Here, the parameter $\boldsymbol{\theta}$ consists of the three upper and three lower bounds of the distribution, estimated by the maxima and minima in the three dimensions, respectively. Inspection of plots of the data confirm that such a distribution is plausible. Our test yields an empirical p-value of 0.91, so that the quality of fit is deemed good, as expected. Meanwhile, we generate a sample of size 1000 from a trivariate normal distribution with $\boldsymbol{\mu} = (0.5, 0.5, 0.5)$ and $\boldsymbol{\Sigma} = 0.01\mathbb{I}$, and randomly select $n = 100$ points from among those lying in the box $[0.3, 0.7]^3$. We test the goodness-of-fit of the trivariate uniform distribution to these data using our method, and obtain an empirical p-value of 0. Hence, we properly reject the trivariate uniform as an explanation for these data.

Being now confident that our method is sufficiently reliable, we investigate its performance in evaluating the goodness-of-fit of a sample from the p-variate beta (MVB$_p$) distribution. The

MVB$_p$ distribution is defined by a $(p+1)$-dimensional parameter $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)$ for random vectors $\mathbf{U} = (U_1, \ldots, U_p)$ in the p-dimensional cube $(0,1)^p$ according to the density

$$f(u_1, \ldots, u_n) = \frac{\Gamma\left(\sum_{j=0}^{p} \theta_j\right)}{\prod_{j=0}^{p} \Gamma(\theta_j)} \left(\prod_{j=1}^{p} \frac{u_j^{\theta_j - 1}}{(1-u_j)^{\theta_j+1}}\right) \left(1 + \sum_{j=1}^{p} \frac{u_j}{1-u_j}\right)^{-\sum_{j=0}^{p} \theta_j},$$

where $\Gamma(\cdot)$ denotes the gamma function [14,17]. Given $\boldsymbol{\theta}$, an observation $\mathbf{U} = (U_1, \ldots, U_p)$ from the MVB$_p$ distribution of dimension p can be randomly generated by first generating independent samples $X_0, X_1, \ldots, X_p$ from $p+1$ univariate gamma distributions with respective shape parameters $\theta_0, \theta_1, \ldots, \theta_p$. Then, for $j = 1, \ldots, p$, set

$$U_j = \frac{X_j}{X_0 + X_j}.$$

We therefore generate a sample of $n = 200$ observations from the MVB$_3$ distribution with $\boldsymbol{\theta}$ chosen arbitrarily as $(4.2, 5.8, 1.9, 3.6)$. The multivariate beta family is not closed under affine transformations. Consequently, let Ξ denote the set of all distributions resulting from invertible affine transformations of data generated by some trivariate beta distribution with parameter $\boldsymbol{\theta}$. We test the null hypothesis

$$H_0 : \Phi \in \Xi$$

against the alternative

$$H_1 : \Phi \notin \Xi,$$

where Φ denotes the hypothesized distribution underlying the sample. Since we already know that these data are generated from a trivariate beta distribution, failure to reject H_0 in this example will confirm that the method has been successful. We implement our procedure, with $N = 10{,}000$, $R = 100$, $B = 100$ and $T = 20$, to compute our test statistic A_T. This involves numerical optimization to obtain the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ based on the sample, and then generating R samples of size N as a basis for computing $\hat{G}_N(t)$ and its quantiles. We obtain $A_T = 14$. We then derive an empirical distribution for A_T by generating B samples from the MVB distribution with $\boldsymbol{\theta} = (2,2,2,2)$ and computing A_T for each sample using the same procedure. The subsequent empirical p-value is 0.42, so that H_0 is correctly affirmed.

We also generate a sample of the same size from the trivariate normal distribution with mean $\boldsymbol{\mu} = (0.5, 0.5, 0.5)$ and covariance $\boldsymbol{\Sigma} = 0.01\mathbb{I}$, where $\mathbb{I}$ is the identity matrix, and from a hybrid sample consisting of three independent samples from univariate χ_3^2, $\Gamma(5,2)$ and $F_{4,3}$ distributions. In both samples, we then apply a linear transformation to ensure that all observations fall within the unit cube, so that we can test for membership in Ξ. We implement our multivariate goodness-of-fit test for each sample, and obtain an empirical p-value of 0.07 for the first sample, and 0 for the second sample. Hence, there is strong grounds for concluding that the latter sample is not a trivariate beta sample, but we are not strongly convinced of the same conclusion regarding the former sample. It seems that the distinction between the distributions of Mahalanobis distances for the multivariate beta and the transformed multivariate normal is not very sharp, corresponding to our earlier observation concerning Figure 1. But there are still reasonable grounds to reject H_0 under our method, since the empirical p-value is below 0.10.

To confirm the validity of this test in the quadrivariate beta setting, we implement it on two samples also of size 200, one from a quadrivariate normal and the other a hybrid of independent samples from the χ_4^2, $\Gamma(2,3)$, $F_{3,2}$ and $\mathcal{U}(0,1)$ distributions. Again we transform the data before proceeding with the test. The respective empirical p-values are 0.07 and 0, just as we found using the three-dimensional versions of these samples. The power of our test is not diminished with the increase of dimension.

We are unable to identify any other tractable goodness-of-fit test with which to compare the performance of our method in this setting, but we are convinced that the proposed goodness-of-fit test procedure has sufficient power to detect departures in data from a multivariate beta model or some invertible affine transformation thereof. Hence, we return to our earlier investigation of IC measurements from the brain of a rat, as introduced in [14]. In particular, our null hypothesis is that the $n = 746$ four-dimensional observations may be modeled with some distribution belonging to the set Ξ corresponding to some quadrivariate beta distribution. All these data fall within the four-dimensional unit cube, so if we do not reject H_0, we will further argue that the quadrivariate beta model is valid. Under H_0, the MLE of $\boldsymbol{\theta}$ is approximately $\hat{\boldsymbol{\theta}} = (1.7, 3.6, 1.8, 1.8, 1.4)$. We implement our goodness-of-fit procedure to compute our test statistic A_T, with $N = 10,000$, $R = 100$ and $T = 20$, and obtain $A_T = 55.4$. We then generate $B = 100$ samples of size n from the quadrivariate beta distribution with parameter $\hat{\boldsymbol{\theta}}$, and compute the test statistic for each sample using the same settings. Given the empirical distribution of A_T over these samples, we arrive at an empirical p-value of 0.17, so that we do not reject our null hypothesis. We compare histograms for each of the four dimensions of real data with corresponding histograms for simulated data of the same size from a quadrivariate beta distribution with parameter $\hat{\boldsymbol{\theta}}$. As can be observed from Figure 2, the histograms are very similar. Consequently, it is quite reasonable to conclude that the data can be fit by a quadrivariate beta model without requiring an affine transformation. Based on our goodness-of-fit test and this additional evidence, there is no basis to dispute our hypothesis that the multivariate beta distribution is a suitable model for the IC data.

The computational burden for our goodness-of-fit procedure is quite reasonable. The choices for the settings R and B have the greatest impact on the time demand. With $R = 100$, the computation of our test statistic in each of the provided examples was under five minutes using the `R` statistical environment [20], for sample sizes up to 750 and dimensions up to four. Increasing the dimension significantly will force one to increase R to maintain an accurate estimate of $G(t)$, while increasing the sample size will allow one to decrease R. Setting $B = 100$ adds an additional 500 minutes to

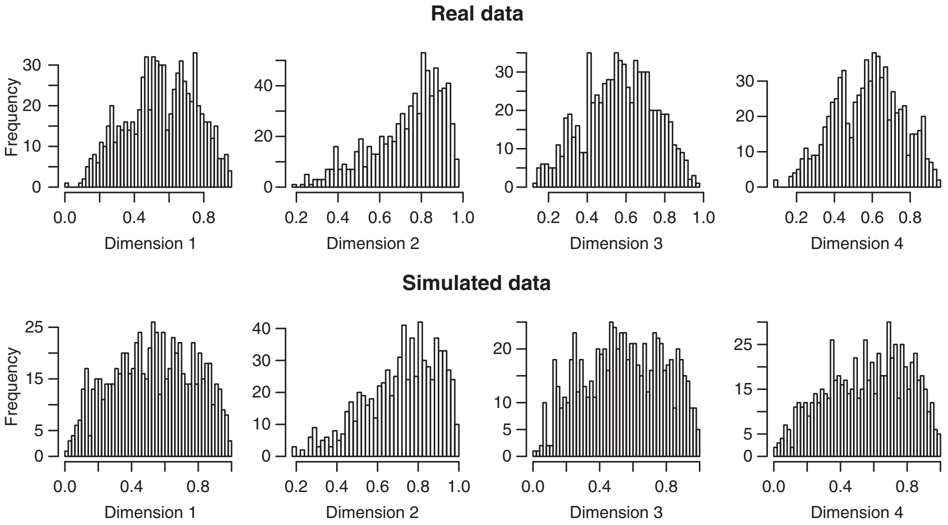

Figure 2. Histograms of the four dimensions of IC data ($n = 746$) and corresponding histograms of simulated data from a quadrivariate beta distribution having parameter $\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ based on the data under the quadrivariate beta distribution.

obtain a sufficient empirical distribution of the test statistic under the null hypothesis, but one can break up these B computations into smaller chunks that can run in parallel. Once an empirical distribution of the test statistic has been obtained, it can be retained for further use when testing the same hypotheses based on different samples of the same size and dimension. Moreover, the computational burden can certainly be further reduced in the hands of skilled programmers using a more powerful programming platform.

4. Discussion

We present a goodness-of-fit test for general multivariate distributions that is simple to implement, involves a reasonable computational burden, does not require analytically intractable derivations, and proves to have sufficient power to detect a poor fit in most situations. One must be able to compute the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ corresponding to a hypothesized distribution $F_{\boldsymbol{\theta}}$ from $\mathcal{F}$, and be able to generate samples from $F_{\hat{\boldsymbol{\theta}}}$. While we have focused here on continuous $F_{\boldsymbol{\theta}}$, the method should easily extend to the discrete case. We anticipate that this straightforward test procedure will prove widely useful in many statistical applications involving multivariate data analysis, particularly outside the multivariate normality setting.

References

[1] W.G. Cochran, *The χ^2 test of goodness of fit*, Ann. Math. Statist. 23(3) (1952), pp. 315–345.

[2] S. Csörgõ, *Testing for normality in arbitrary dimension*, Ann. Statist. 14 (1986), pp. 708–723.

[3] Y. Fan, *Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function*, J. Multivar. Anal. 62 (1997), pp. 36–63.

[4] G. Fasano and A. Franceschini, *A multidimensional version of the Kolmogorov-Smirnov test*, Mon. Not. Roy. Astron. Soc. Lett. 225 (1987), pp. 155–170.

[5] S. Ghosh and F.H. Ruymgaart, *Applications of empirical characteristic functions in some multivariate problems*, Can. J. Statist. 20 (1992), pp. 429–440.

[6] A. Justel, D. Peña, and R. Zamar, *A multivariate Kolmogorov–Smirnov test of goodness of fit*, Stat. Probab. Lett. 35 (1997), pp. 251–259.

[7] A. Kankainen, S. Taskinen, and H. Oja, *Tests of multinormality based on location vectors and scatter matrices*, Stat. Meth. Appl. 16 (2007), pp. 357–379.

[8] J.A. Koziol, *Assessing multivariate normality: A compendium*, Commun. Statist. Theoret. Meth. 15 (1986), pp. 2763–2783.

[9] R.H.C. Lopes, I. Reid, and P.R. Hobson, *The two-dimensional Kolmogorov-Smirnov test*, XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research (April 23–27), Amsterdam, the Netherlands, 2007.

[10] P.C. Mahalanobis, *On the generalised distance in statistics*, Proc. Natl. Inst. Sci. India 2(1) (1936), pp. 49–55.

[11] J.F. Malkovich and A.A. Afifi, *On tests for multivariate normality*, J. Am. Statist. Assoc. 68 (1973), pp. 176–179.

[12] K.V. Mardia, *Measures of multivariate skewness and kurtosis with applications*, Biometrika 57 (1970), pp. 519–530.

[13] K.V. Mardia, *Tests of univariate and multivariate normality*, in *Handbook of Statistics*, P.R. Krishnaiah, ed., North-Holland, Amsterdam, The Netherlands, 1980, pp. 279–320.

[14] M.P. McAssey, F. Hsieh, and A.C. Smith, *Coupling among electroencephalogram gamma signals on a short time scale*, Comput. Intell. Neurosci. 2010 (2010), pp. 1–12. Article ID 946089, doi: 10.1155/2010/946089.

[15] D.S. Moore and J.B. Stubblebine, *Chi-square tests for multivariate normality, with applications to common stock prices*, Commun. Statist. Theoret. Meth. A 10 (1981), pp. 713–733.

[16] G.S. Mudholkar, M. McDermott, and D.J. Srivastava, *A test of p-variate normality*, Biometrika 79(4) (1992), pp. 850–854.

[17] I. Olkin and R. Liu, *A bivariate beta distribution*, Stat. Probab. Lett. 62 (2003), pp. 407–412.

[18] J.A. Peacock, *Two-dimensional goodness-of-fit testing in astronomy*, Mon. Not. Roy. Astron. Soc. Lett. 202 (1983), pp. 615–627.

[19] K. Pearson, *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, Philos. Mag. Series 5. 50 (1900), pp. 157–172.

[20] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. Available at http://www.R-project.org/.

[21] S. Rincón-Gallardo, C.P. Quesenberry, and F.J. O'Reilly, *Conditional probability integral transformations and goodness-of-fit tests for multivariate normal distributions*, Ann. Stat. 7 (1979), pp. 1052–1057.

[22] M. Rosenblatt, *Remarks on a multivariate transformation*, Ann. Math. Statist. 23 (1952), pp. 470–472.

[23] J.P. Royston, *Algorithm AS 181: The W test for normality*, Appl. Statist. 31 (1982), pp. 176–180.

[24] J.P. Royston, *Some techniques for assessing multivariate normality based on the Shapiro–Wilk W*, Appl. Statist. 32 (1983), pp. 121–133.

[25] S.J. Schwager and B.H. Margolin, *Detection of multivariate normal outliers*, Ann. Statist. 10 (1982), pp. 943–954.

[26] N.J.H. Small, *Marginal skewness and kurtosis in testing multivariate normality*, Appl. Statist. 29 (1980), pp. 85–87.

[27] G.J. Székely and M.L. Rizzo, *A new test for multivariate normality*, J. Multivar. Anal. 93 (2005), pp. 58–80.

[28] S. Velilla, *Diagnostics and robust estimation in multivariate data transformation*, J. Amer. Statist. Assoc. 90 (1995), pp. 945–951.