

An empirical goodness-of-fit test for multivariate distributions

Michael P. McAssey*

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

Abstract

An empirical test is presented by which one may determine 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 the multivariate normal and from the multivariate beta distributions. In the latter case, we apply the test to real multivariate data 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 [?], measurements of Instantaneous Coupling (IC) were computed between pairs of electroencephalogram (EEG) 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 [?, ?, ?] and continuing with a multitude of additional approaches, e.g., application of the Rosenblatt transformation [?] to examine multivariate normality [?], tests using multivariate measures of skewness and kurtosis [?, ?, ?, ?, ?], a test based on the multivariate Shapiro–Wilk statistic [?], a radii and angles test [?], a test based on the multivariate Box–Cox transformation [?], and many other creative methods [?, ?, ?, ?]. But these tests do not extend readily to the general case. Proposals for extending the Kolmogorov–Smirnov (KS) goodness-of-fit test to multiple dimensions have been published [?, ?, ?, ?], but the required test statistic in each proposed method is extremely difficult to compute, even in the bivariate case, and a suggested simplification in [?] still

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

requires a transformation whose derivation is analytically intractable for most multivariate distributions. A multivariate goodness-of-fit test based on the empirical characteristic function has also been developed [?]. 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 two established tests for multivariate normality to demonstrate its reliability in that setting, and then demonstrate its effectiveness for testing suitability of the multivariate beta model. Finally, we apply this test to the IC measurement data referenced above to confirm the original multivariate beta assumption.

2 Method

Let $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathcal{R}^p$ be a sample from a population having known p -variate continuous distribution $F = F_{\boldsymbol{\theta}}$, with $\boldsymbol{\theta} \in \mathcal{R}^k$ for some $k \geq 1$, 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 [?] that the Mahalanobis distance $\Delta \in \mathcal{R}$ between \mathbf{X} and $\boldsymbol{\mu}$ is a continuous function from $\mathcal{R}^p \rightarrow [0, \infty)$ computed as

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

Essentially this function assigns the same distance to all points in a p -dimensional ellipsoid centered at $\boldsymbol{\mu}$ whose shape is uniquely determined by $\boldsymbol{\Sigma}$. Hence given $\mathbf{X}_1, \dots, \mathbf{X}_n \sim F$ we obtain $\Delta_1, \dots, \Delta_n$, where $\Delta_i = \Delta(\mathbf{X}_i)$. Let G denote the probability distribution underlying $\Delta_1, \dots, \Delta_n$, i.e., $G(t) = \mathcal{P}(\Delta_i < t)$ for $t \in \mathcal{R}$.

Let ϕ denote the characteristic function of G , that is,

$$\begin{aligned} \phi(t) &= \mathcal{E}(e^{it\Delta}) = \mathcal{E} \left[e^{it\sqrt{(\mathbf{X}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu})}} \right] \\ &= \int_{\mathcal{R}^p} e^{it\sqrt{(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}} dF(\mathbf{x}). \end{aligned}$$

This relationship shows that the distribution G of Mahalanobis distances is specified by the distribution F of \mathbf{X} , although not necessarily uniquely. That is, if $F_1 = F_2$, then $G_1 = G_2$, and if $G_1 \neq G_2$, then $F_1 \neq F_2$.

Therefore, if we generate a sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ from a known distribution F , we may compute $\Delta_1, \dots, \Delta_n$, 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 $n \rightarrow \infty$, $G_n(t) \rightarrow G(t)$, by well-established theoretical results.

Figure ?? displays plots of $G_n(t)$ for samples randomly generated from four different bivariate distributions, with $n = 10000$. 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 $\text{Be}(2, 3)$ and the $\text{Be}(3, 2)$. For each bivariate distribution, we compute $\Delta_1, \dots, \Delta_n$ 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.

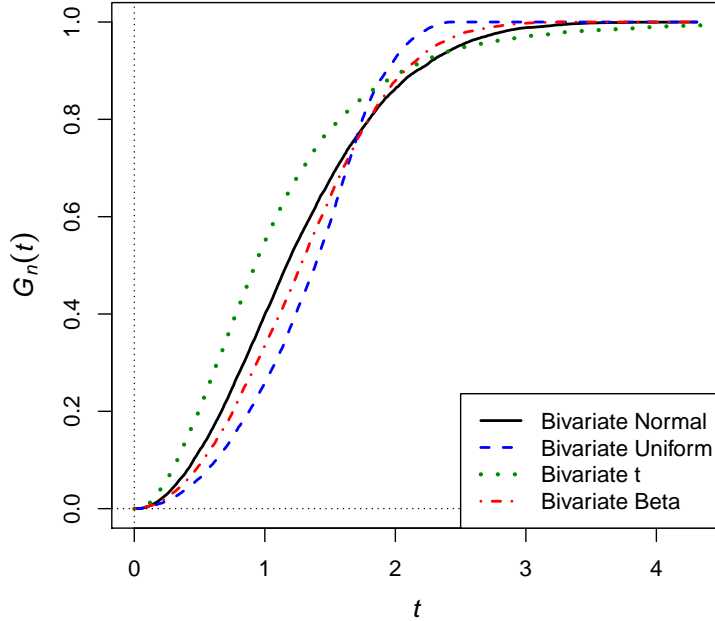


Figure 1: Plots of $G_n(t)$, with $n = 10000$, for samples randomly generated from a bivariate normal distribution, a bivariate uniform distribution, a bivariate Student's- t distribution, and a bivariate beta distribution.

Now, 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 $0 = p_0 < p_1 < \dots < p_T = 1$ of $(0, 1)$, we then obtain a corresponding partition $0 = q_{p_0} < q_{p_1} < \dots < q_{p_T} = \infty$ 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

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

As $n \rightarrow \infty$, $q_{p,n} \rightarrow q_p$, since $G_n(t) \rightarrow G(t)$. 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 $P_{j,n} \rightarrow P_j$ as $n \rightarrow \infty$. 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 F_{θ} . Moreover, while we shall assume that the mean and covariance of F_{θ} exist, we shall not assume that θ , μ or Σ are known. We only require that the maximum-likelihood estimate (MLE) of θ can be computed based on a sample drawn from F_{θ} . Consequently, let $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathcal{R}^p$ denote a sample from some unknown p -variate continuous distribution Φ . We wish to test

$$H_0 : \Phi = F_{\theta} \text{ for some } \theta$$

against

$$H_1 : \Phi \neq F_{\theta} \text{ for any } \theta$$

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

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

Since $\bar{\mathbf{X}} \xrightarrow{P} \mu$ and $\mathbb{S} \xrightarrow{P} \Sigma$ as $n \rightarrow \infty$ under the null hypothesis, the continuous mapping theorem guarantees that $D_i \xrightarrow{P} \Delta_i$ as $n \rightarrow \infty$. 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 $t > 0$, $\mathbb{1}(D_i \leq t) \xrightarrow{P} \mathbb{1}(\Delta_i \leq t)$ as $n \rightarrow \infty$, 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 $n \rightarrow \infty$, 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' : $D_1, \dots, D_n \sim G$. If we reject H_0' we must also reject H_0 . But we cannot proceed exactly as before when F_{θ} was completely specified. Instead, we follow a longer, modified procedure. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ denote the observed multivariate data, and d_1, \dots, d_n the corresponding observed sample Mahalanobis distances. Then execute the following steps:

1. Assuming H_0 is true, estimate $\boldsymbol{\theta}$ with its MLE $\hat{\boldsymbol{\theta}}$ based on $\mathbf{x}_1, \dots, \mathbf{x}_n$;
2. Generate a very large sample $\mathbf{u}_1, \dots, \mathbf{u}_N$ of size $N \gg n$ from $F_{\hat{\boldsymbol{\theta}}}$ (we use $N = 10000$);
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 $\hat{d}_1, \dots, \hat{d}_N$ 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

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

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

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

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.,

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

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 through 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 of values for A_T when H_0 holds with $\boldsymbol{\theta}$ unspecified, and determine whether our statistic A_T computed in step 10 is plausible under this distribution. This must be done in the setting where a sample of the same size n indeed comes from $F_{\boldsymbol{\theta}}$ for some arbitrarily selected $\boldsymbol{\theta}$, but after the sample is generated from $F_{\boldsymbol{\theta}}$ we proceed as if we do not know $\boldsymbol{\theta}$, i.e., we use the sample mean and sample covariance as we did with the original data. Hence we first choose some sensible $\boldsymbol{\theta}^* \in \mathcal{R}^k$ that lies within the interior of the parameter space of F . For example, if F is multivariate normal, it would be sensible to choose the zero vector for the location parameter and the identity matrix for the scale parameter. 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_{θ^*} ;
- B. Obtain the sample mean and sample covariance, then compute the sample Mahalanobis distances d_1^*, \dots, d_n^* ;
- C. Estimate θ^* with its MLE $\hat{\theta}^*$ based on $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$;
- D. Follow steps 2 through 9 (substituting $\hat{\theta}^*$ for $\hat{\theta}$ at step 2, and substituting d_i^* for d_i at step 8) to obtain test statistic $A_{T,b}$. Again, we repeat Steps 2 through 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 $p_e < \alpha$.

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. We conduct this test using the method presented above and, for comparison, two established methods: the multivariate Shapiro–Wilk test [?, ?] and the energy test for multivariate normality [?]. We generate samples from ten distinct bivariate normal, bivariate uniform, and bivariate Student’s- t distributions, with the parameters randomly selected for each of the thirty samples. We implement our method for each sample with $N = 10000$, $R = 100$, $T = 20$ and $B = 100$. We then compute the mean p -value for each method over the ten samples corresponding to each distribution family. We expect large p -values for samples from the bivariate normal family, and small 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 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	Mean p -value		
	Our Method	Shapiro–Wilk	Energy Test
Bivariate Normal	0.504	0.320	0.487
Bivariate Uniform	0.044	0.356	0.005
Bivariate Student’s- t	0.148	0.025	0.015
True distribution	Detection Rates		
	Our Method	Shapiro–Wilk	Energy Test
Bivariate Normal	0.1	0.1	0.0
Bivariate Uniform	0.7	0.0	1.0
Bivariate Student’s- t	0.6	0.9	0.9

The energy test performs the best in detecting (and not falsely detecting) departures from normality for all three bivariate distribution families, while the Shapiro–Wilk test outperforms our method 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 does very poorly. Overall, our method shows more promise than the multivariate Shapiro–Wilk 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	Mean p -value		
	Our Method	Shapiro–Wilk	Energy Test
Trivariate Normal	0.431	0.225	0.447
Trivariate Uniform	0.000	0.577	0.007
Trivariate Student’s- t	0.061	0.004	0.011
True distribution	Detection Rates		
	Our Method	Shapiro–Wilk	Energy Test
Trivariate Normal	0.2	0.4	0.1
Trivariate Uniform	1.0	0.0	1.0
Trivariate Student’s- t	0.9	1.0	0.9

While the energy test again yields the best overall performance, the performance of our method is almost identical. Shapiro–Wilk is once again unable to distinguish the uniform family from the normal family, while the other two methods succeed in all 10 samples. 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.

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 $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_p)$ for random vectors $\mathbf{U} = (U_1, \dots, U_p)$ in the p -dimensional cube $(0, 1)^p$ according to the density

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

where $\Gamma(\cdot)$ denotes the gamma function [?, ?]. Given $\boldsymbol{\theta}$, an observation $\mathbf{U} = (U_1, \dots, U_p)$ from the MVB_p distribution of dimension p can be randomly generated by first generating independent samples X_0, X_1, \dots, X_p from $p + 1$ univariate gamma distributions with respective shape parameters $\theta_0, \theta_1, \dots, \theta_p$. Then, for $j = 1, \dots, p$, 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 $(4.2, 5.8, 1.9, 3.6)$. We wish to test the null hypothesis

$$H_0 : \Phi = \text{MVB}_3(\boldsymbol{\theta}) \text{ for some } \boldsymbol{\theta}$$

against the alternative

$$H_1 : \Phi \neq \text{MVB}_3(\boldsymbol{\theta}) \text{ for any } \boldsymbol{\theta},$$

where Φ denotes the proposed distribution underlying the sample. We implement our procedure, with $N = 10000$, $R = 100$, $B = 100$ 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 samples 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 combination 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. 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 ???. But there are still reasonable grounds to reject H_0 under our method, since the empirical p -value is below 0.10.

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. Hence we return to our earlier investigation of Instantaneous Coupling (IC) measurements from the brain of a rat, as introduced in [?]. In particular, our null hypothesis is that the $n = 746$ four-dimensional observations may be modeled with a quadrivariate beta distribution. Under H_0 , the MLE of $\boldsymbol{\theta}$ is approximately $(1.7, 3.6, 1.8, 1.8, 1.4)$. We implement our goodness-of-fit procedure to compute our test statistic A_T , with $N = 10000$, $R = 100$ and $T = 20$, and obtain $A_T = 55.4$. We then generate $B = 100$ samples of size n from the quadrivariate beta distribution with arbitrarily-chosen parameter $\boldsymbol{\theta} = (2, 2, 2, 2, 2)$, and compute the test statistic for each sample using the same settings. Given the distribution of A_T over these samples, we arrive at an empirical p -value of 0.13, so that we do not reject our null hypothesis. Based on our test there is no basis to dispute our assumption that the multivariate beta distribution is a suitable model for the IC data.

To confirm the validity of this test we implement it on two samples also of size n , one from a quadrivariate normal and the other from a combination of independent samples from the χ_4^2 , $\Gamma(2, 3)$, $F_{3,2}$ and $\mathcal{U}(0, 1)$ distributions. The respective 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 thus confident in the trustworthiness of our conclusion regarding the real 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 [?], for samples 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 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 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 $\hat{\theta}$ of θ corresponding to the hypothesized distribution F_{θ} , and be able to generate samples from $F_{\hat{\theta}}$. While we have focused here on continuous F_{θ} , 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.

References

- [1] Cochran, W.G. (1952). The χ^2 Test of Goodness of Fit. *Ann. Math. Statist.* **23**(3), pp. 315–345.
- [2] Csörgő, S. (1986). Testing for normality in arbitrary dimension. *Ann. Statist.* **14**, pp. 708–723.
- [3] Fan, Y. (1997). Goodness-of-Fit Tests for a Multivariate Distribution by the Empirical Characteristic Function. *Journal of Multivariate Analysis* **62**, pp. 36–63.
- [4] Fasano, G. and Franceschini, A. (1987). A multidimensional version of the Kolmogorov-Smirnov test, *Monthly Notices of the Royal Astronomy Society* **225**, pp. 155-170.
- [5] Ghosh, S. and Ruymgaart, F.H. (1992). Applications of empirical characteristic functions in some multivariate problems. *Can. J. Statist.* **20**, pp. 429–440.
- [6] Justel, A., Peña, D. and Zamar, R. (1997). A multivariate Kolmogorov–Smirnov test of goodness of fit. *Statistics and Probability Letters* **35**, pp. 251–259.
- [7] Koziol, J.A. (1986). Assessing multivariate normality: a compendium. *Commun. Statist. Theoret. Meth.* **15**, pp. 2763–2783.
- [8] Lopes, R.H.C., Reid, I. and Hobson, P.R. (2007). The two-dimensional Kolmogorov-Smirnov test. *XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research* (April 2327, 2007). Amsterdam, the Netherlands.
- [9] Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India* **2**(1), pp. 49-55.
- [10] Malkovich, J.F. and Afifi, A.A. (1973). On tests for multivariate normality. *J. Am. Statist. Assoc.* **68**, pp. 176–179.

- [11] Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, pp. 519–530.
- [12] Mardia, K.V. (1980). Tests of univariate and multivariate normality, in *Handbook of Statistics*, P.R. Krishnaiah (ed.), Amsterdam, The Netherlands: North-Holland, pp. 279–320.
- [13] McAssey, M.P., Hsieh, F. and Smith, A.C. (2010). Coupling among Electroencephalogram Gamma Signals on a Short Time Scale. *Computational Intelligence and Neuroscience*, Volume 2010, Article ID 946089, doi: 10.1155/2010/946089.
- [14] Moore, D.S. and Stubblebine, J.B. (1981). Chi-square tests for multivariate normality, with applications to common stock prices. *Commun. Statist. Theoret. Meth. A* **10**, pp. 713–733.
- [15] Mudholkar, G.S., McDermott, M. and Srivastava, D.J. (1992). A test of p -variate normality. *Biometrika* **79**(4), pp. 850–854.
- [16] Olkin, I. and Liu, R. (2003). A bivariate beta distribution. *Statistics and Probability Letters* **62**, pp. 407–412.
- [17] Peacock, J.A. (1983). Two-dimensional goodness-of-fit testing in astronomy, *Monthly Notices Royal Astronomy Society* **202**, pp. 615–627.
- [18] Pearson, K. (1900). 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**, pp. 157–172.
- [19] R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- [20] Rincón-Gallardo, S., Quesenberry, C.P. and O’Reilly, F.J. (1979). Conditional probability integral transformations and goodness-of-fit tests for multivariate normal distributions. *Ann. Stat.* **7**, pp. 1052–1057.
- [21] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23**, pp. 470–472.
- [22] Royston, J.P. (1982) Algorithm AS 181: The W Test for Normality. *Applied Statistics*, **31**, pp. 176–180.
- [23] Royston, J.P. (1983). Some techniques for assessing multivariate normality based on the Shapiro–Wilk W . *Appl. Statist.* **32**, pp. 121–133.
- [24] Schwager, S.J. and Margolin, B.H. (1982). Detection of multivariate normal outliers. *Ann. Statist.* **10**, pp. 943–954.
- [25] Small, N.J.H. (1980). Marginal Skewness and kurtosis in testing multivariate normality. *Appl. Statist.* **29**, pp. 85–87.
- [26] Székely, G.J. and Rizzo, M.L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis* **93**, pp. 58–80.
- [27] Velilla, S. (1995). Diagnostics and robust estimation in multivariate data transformation. *J. Amer. Statist. Assoc.* **90**, pp. 945–951.

Figure captions

Figure 1: Plots of $G_n(t)$, with $n = 10000$, for samples randomly generated from a bivariate normal distribution, a bivariate uniform distribution, a bivariate Student's- t distribution, and a bivariate beta distribution.