Comparison of Some Selected Multivariate Normality Tests for Classification into Locally Most Powerful and Uniformly Most Powerful using Monte Carlo Simulation Approach (maximum 20 words)

Abstract (Nice)

Testing for multivariate normality is a fundamental step in multivariate statistical analysis, as many classical techniques rely on this assumption. However, arbitrary use of multivariate normality test often lead to type I or type II error, which necessitate the review of the techniques for classification as Uniformly Most Powerful (UMP) and Locally Most Powerful (LMP). This study employed Monte Carlo simulations to investigate the empirical type I error rates and rejection powers of nine commonly used multivariate normality tests, including Shapiro–Wilk (MVSW), Energy test, Mardia's test, Henze–Zirkler test, Zhang's test, Robust Mahalanobis Distance, Royston's H test, Doornik–Hansen test, and the High-Dimensional Energy test. Simulations were conducted using 1,000 replications at varying sample sizes (n = 15, 20, 25, 50, 100, 200) with dimension d = 3, on multivariate normal distribution (MVN) and multivariate t-distribution (MVT). The results showed that while all tests approached nominal error rates at large sample sizes, some, particularly the Energy test, Zhang's test, Henze–Zirkler test, and the High-Dimensional Energy test, exhibited higher sensitivity to heavy-tailed alternatives, maintaining strong power across all sample sizes. These were classified as Uniformly Most Powerful (UMP). In contrast, tests such as Shapiro–Wilk, Royston's H, and Robust Mahalanobis Distance were more conservative in small samples but effective in larger ones, thus categorised as Locally Most Powerful (LMP). The findings provide a practical classification framework to guide the choice of multivariate normality tests in applied research.

Keywords: Multivariate normality tests, Monte Carlo simulation, Type I error control, Multivariate t-Distribution, Test classification (OK)

Introduction (Nice)

The assumption of multivariate normality plays a pivotal role in many classical and modern multivariate statistical methods. Techniques such as principal component analysis (PCA), discriminant analysis, multivariate analysis of variance (MANOVA), structural equation modelling (SEM), factor analysis, canonical correlation analysis, and linear mixed models often rely on the premise that the underlying data follow a multivariate normal distribution [1, 2]. Violation of this assumption can result in biased parameter estimates, invalid inferences, unreliable predictions, and misleading conclusions [3]. Consequently, testing for multivariate normality has become an essential step in applied statistical analysis.

A variety of statistical tests have been developed to assess multivariate normality, including Mardia's skewness and kurtosis measures [4], the Henze-Zirkler test [5], the Doornik-Hansen omnibus test [6], and energy-based approaches such as the Baringhaus-Henze test [7]. More recent contributions include robust and high-dimensional adaptations of these methods [8, 9]. The diversity of these tests presents a practical challenge, where no single method is uniformly optimal across all scenarios, as their performance varies depending on sample size, dimensionality, and the type of deviation from normality [10, 11].

From a theoretical standpoint, the classification of tests into Locally Most Powerful (LMP) and Uniformly Most Powerful (UMP) categories offers a structured way to understand their strengths and limitations. LMP tests are highly sensitive to small, localised departures from normality, making them effective when specific alternatives are suspected [12, 13]. Conversely, UMP tests maintain robust power across a broad spectrum of alternatives, making them suitable when the exact form of deviation is unknown [12, 14-16]. Distinguishing between these categories is therefore not only of theoretical interest but also of practical importance in guiding test selection.

Monte Carlo simulation provides a powerful framework for evaluating the empirical performance of multivariate normality tests. By generating data under controlled conditions, simulations enable the assessment of type I error rates and statistical power across varying sample sizes, dimensions, and deviations from normality [17-19]. This approach is particularly valuable in systematically exploring test behaviour under both LMP and UMP frameworks.

The relevance of this line of research has grown with the increasing availability of high-dimensional datasets in applied fields such as finance, genomics, and image analysis, where traditional tests often suffer from reduced power due to the curse of dimensionality. Against this backdrop, classifying tests into LMP and UMP categories through simulation-based studies contributes to both theoretical understanding and practical application. Specifically, focusing on data generated from multivariate normal and multivariate t-distributions provides insights into test performance under exact normality and heavy-tailed alternatives, a setting frequently encountered in real-world applications.

Despite the existence of numerous tests for multivariate normality, their performance is highly dependent on sample size, dimensionality, and the type of alternative distribution considered. However, while Mardia's measures are effective in detecting skewness and kurtosis, they may lack robustness in small samples or heavy-tailed settings. Similarly, distance and energy-based methods are powerful in certain contexts but computationally intensive and not always consistent across scenarios. This variability poses a significant challenge for researchers who must choose an appropriate test without clear guidance on its relative strengths.

Few studies have fully explored test performance under practical conditions such as varying sample sizes and data generated from heavy-tailed alternatives like the multivariate t-distribution [20-22]. In modern statistical practice, where deviations from normality are common and datasets are increasingly high-dimensional, a rigorous evaluation of multivariate normality tests under controlled simulation settings. Monte Carlo simulations offer an effective means to investigate empirical power and type I error rates, thereby enabling the classification of tests into LMP and UMP categories. By focusing on multivariate normal and t-distributed data, this study seeks to provide a systematic evaluation of widely used tests, offering both theoretical insights and practical guidance to researchers. Such a classification framework will help identify conditions under which specific tests are most effective, ultimately improving the reliability of multivariate data analysis in diverse applications.

Methodology (OK)

Research Design

In this study, we adopted a simulation-based experimental design by utilising Monte Carlo techniques to evaluate and classify multivariate normality tests into Locally Most Powerful (LMP) and Uniformly Most Powerful (UMP) categories. We generated synthetic datasets under controlled distributional assumptions,

applying nine different multivariate normality tests, and computing their empirical power and type I error rates across five different sample sizes.

Simulation Framework

We considered two different datasets, one from a multivariate normal distribution (MVN) that serves as the null hypothesis, denoted as $X \sim N_d(0, I_d)$, where $d = 3$ is the dimension, 0 is a zero mean vector, and I_d is the identity covariance matrix, and the second dataset from a multivariate t-Distribution (MVT) serving as the alternative hypothesis, expressed as $X \sim t_d(v)$, where $d = 3$ and v correspond to heavier tails, introducing deviations from multivariate normality.

In this study, six different sample sizes ($n = 10, 20, 25, 50, 100$, and 200) were used to examine the effect of data volume on test performance. For each combination of distributional setting and sample size, 1000 Monte Carlo replications were conducted to ensure stable estimates of empirical power and type I error.

Shapiro-Wilk Type Test for multivariate normality (MVSW)

The Shapiro-Wilk Type Test for Multivariate Normality (MVSW) is an extension of the Shapiro-Wilk test for univariate normality to the multivariate setting. This test is useful for small to moderate sample sizes and is sensitive to both Skewness and kurtosis deviations.

The MVSW test evaluates whether a given dataset follows a multivariate normal distribution by analysing the linear combination of ordered sample values and comparing it to a theoretical expectation under normality. It is formulated using the eigenvalues and eigenvectors of the sample covariance matrix.

Let $X_1, X_2, \cdots, X_n$ be a random sample from a p-dimensional distribution, $\bar{X}$ be the sample mean vector, and S be the sample covariance matrix. The MVSW test statistic is given by:

$$W_p = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})' S^{-1} (X_i - \bar{X})} \tag{1}$$

where $X_{(i)}$ are the ordered sample values based on a projection direction that maximises normality assessment, $a_i$are optimal coefficients derived from the expected values of order statistics of a standard normal distribution, $S^{-1}$is the inverse sample covariance matrix, and W_p measures how well the sample data aligns with multivariate normality.

To optimise the projection direction, the test considers the eigenvalue decomposition of S:

$$S = PAP' \tag{2}$$

where $A = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$ is the diagonal matrix of eigenvalues and P is the matrix of eigenvectors.

The test statistic is computed using:

$$W_p = \frac{(a'P'X)^2}{\sum_{i=1}^{n} (X_i - \bar{X})' S^{-1} (X_i - \bar{X})} \tag{3}$$

where the vector $a$is derived from the univariate Shapiro-Wilk test.

High-Dimensional Energy Test (HDET)

The High-Dimensional Energy Test (HDET) is a distance-based nonparametric test for assessing multivariate normality in high-dimensional datasets. HDET evaluates distributional differences using an energy distance metric, making it robust for detecting departures from normality in high-dimensional settings.

By computing an energy distance between the empirical distribution of the sample and the expected normal distribution, the test can effectively detect deviations from normality.

For a given random sample $X_1, X_2, \cdots, X_n$ from a p-dimensional space, the test statistic is formulated using the energy distance function. The energy distance between two random variables X and Y is given as:

$$E_n = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\| \tag{4}$$

where $\|.\|$ denotes the Euclidean norm (distance between two points), $X, X'$are independent random variables from distribution P_X, and $Y, Y'$are independent random variables from distribution P_Y.

For testing multivariate normality, we compare the sample distribution to a theoretical normal distribution using the empirical form of this distance. The sample-based High-Dimensional Energy Test statistic is given by:

$$T_n = \frac{2}{n(n-1)}\sum_{i<j}\|X_i - X_j\| - \frac{2}{n}\sum_{i=1}^{n}\|X_i - Z\| \tag{5}$$

where $Z \sim N(\mu, \sum)$ is a multivariate normal random variable with estimated mean and covariance from the sample.

Zhang's Test for Multivariate Normality

Zhang's Test is a powerful statistical method for assessing multivariate normality based on nearest neighbour (NN) distances, making it particularly robust for high-dimensional datasets. It evaluates whether the distribution of multivariate data deviates from normality by analysing the spatial distribution of data points.

Steps in Formulating Zhang's Test Statistic

1. Define the Mahalanobis distance

Given a dataset $X = (X_1, X_2, \cdots, X_n)$ of n observations in d-dimensional space, the Mahalanobis distance of each observation from the mean is computed as:

$$D_i = \sqrt{(X_i - \bar{X})' S^{-1} (X_i - \bar{X})} \tag{6}$$

where $\bar{X}$ is the sample mean vector, S is the sample covariance matrix, and D_i represents the distance of the ith observation from the mean. If the data follows a multivariate normal distribution, then D_i^2 follows a Chi-square distribution with d degrees of freedom, $D_i^2 \sim \chi_d^2$.

2. Compute nearest neighbour ranks

Define the nearest neighbour rank for each observation $X_i$by counting how many of the Mahalanobis distances $D_i (for\ j \neq i)$ are smaller than D_i**:**

$$R_i = \sum_{j=1}^{n} I(D_j \leq D_i), \qquad j \neq i \tag{7}$$

where $I(.)$ is an indicator function that returns 1 if the condition is true, otherwise 0, and R_i represents the rank of D_i in the sorted list of distances.

3. Compute Zhang's test statistic

$$Z_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_i}{n} - \frac{i}{n}\right)^2 \tag{8}$$

where $\frac{R_i}{n}$ is the normalised rank of each observation and $\frac{i}{n}$ represents the expected uniform distribution. This statistic measures the deviation of the empirical rank distribution from the expected uniform distribution under normality

Robust Mahalanobis Distance Test

The Robust Mahalanobis Distance (RMD) test is a powerful method for assessing multivariate normality while being resistant to the influence of outliers. Mahalanobis distance calculations use the sample mean and sample covariance matrix, which can be affected by outliers.

The RMD test is useful for high-dimensional data or datasets that contain contaminated observations.

Steps in formulating RMD

1. Compute the traditional Mahalanobis distance

2. Compute the RMD. Instead of using the classical $\bar{X}$ and covariance matrix S, we replace them with robust mean ($\hat{X}$) and robust covariance ($\hat{S}$). The RMD is calculated as:

$$RMD_i = \sqrt{(X_i - \hat{X})' \hat{S}^{-1} (X_i - \hat{X})} \tag{9}$$

where $\hat{X}$ is the robust mean obtained from the MCD estimator and $\hat{S}$is the robust covariance matrix obtained from the MCD.

3. Compute the robust Mahalanobis test statistic. To test for multivariate normality, the distribution of RMD values is compared to a reference Chi-square distribution. The RMD test statistic is written as:

$$T = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{RMD_i^2 - d}{\sqrt{2d}}\right)^2 \tag{10}$$

where d is the number of dimensions and RMD_i^2 are the squared robust Mahalanobis distances.

Mardia's Test

Mardia's Test evaluates whether a dataset follows a multivariate normal distribution by examining if the multivariate skewness measures the symmetry of the distribution, and the multivariate kurtosis measures the tailedness of the distribution.

Mardia's Skewness Test

Mardia's multivariate skewness measures the asymmetry of a dataset and is given by:

$$b_1, d = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(X_i - \bar{X})' S^{-1}\left(X_j - \bar{X}\right)\right]^3 \quad (11)$$

where n is the sample size, d is the number of dimensions (variables), X_i is the i-th observation vector, $\bar{X}$ is the sample mean vector, S is the sample covariance matrix, and b_1, d represents the multivariate Skewness statistic. For large sample sizes, $nb_1, d \sim \chi^2_{d(d+1)(d+2)/6}$.

Mardia's Kurtosis Test

Mardia's kurtosis examines whether the dataset has the expected amount of tail weight under multivariate normality. It is written as:

$$b_2, d = \frac{1}{n}\sum_{i=1}^{n}[(X_i - \bar{X})' S^{-1}(X_i - \bar{X})]^2 \quad (12)$$

For large samples, the kurtosis statistic is approximately normally distributed

$$Z_{b_2,d} = \frac{b_2, d - d(d+1)}{\sqrt{\frac{8d(d+2)}{n}}} \sim N(0,1) \quad (13)$$

Henze-Zirkler Test

The Henze-Zirkler (HZ) test is a widely used method for assessing multivariate normality, based on the characteristic function approach. The HZ test is used to detect both Skewness and kurtosis deviations from normality, and it is derived from the empirical characteristic function (ECF) and measures the difference between the ECF of the sample and that of a theoretical multivariate normal distribution. The test statistic is given by:

$$HZ = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{\frac{-b}{2}(X_i - X_j)' S^{-1}(X_i - X_j)} \quad (14)$$

where n is the sample size, d is the number of dimensions, X_i is the ith observation vector, S is the sample covariance matrix, and b is a constant given by $b = \frac{1}{\sqrt{2}}\left(1 + \frac{4}{d+2}\right)^{-1}$.

3.2.7 Royston's H Test

Royston's H test is an extension of the Shapiro-Wilk test, used to assess whether a given dataset follows a MVN distribution by applying the Shapiro-Wilk test to each variable separately and then combining the individual test statistics into a single overall test statistic.

Steps for Computing Royston's H Test

1. Compute the Shapiro-Wilk statistic (W_j) for each variable.

2. Transform the individual Shapiro-Wilk test statistics W_j using:

$$Z_j = \frac{\log(1 - W_j) - \mu_j}{\sigma_j} \tag{15}$$

where μ_j and σ_j are the mean and standard deviation of $\log(1 - W_j)$under normality, which are estimated based on sample size n.

3. Compute the Royston's H test statistic as:

$$H = \sum_{j=1}^{d} \left(\frac{Z_j - \bar{Z}}{\sigma_Z} \right)^2 \tag{16}$$

where d is the number of variables (dimensions), $\bar{Z}$ is the mean of Z_j values, and σ_Z is the standard deviation of Z_j values.

3.2.8 Doornik-Hansen Test

The Doornik-Hansen test is a statistical method for assessing multivariate normality by combining Skewness and kurtosis measures into a single test statistic. The Doornik-Hansen test extends Jarque-Bera's test for univariate normality.

Steps for Computing Doornik-Hansen (DH) test

1. Compute the Skewness statistic. For a d-dimensional dataset with n observations, the multivariate Skewness measure is defined as:

$$S = n \sum_{j=1}^{d} z_j^3 \tag{17}$$

where z_j represents the standardised values of each variable and the skewness statistic S follows a Chi-square distribution.

2. Compute the kurtosis statistic. The multivariate kurtosis measure is given by:

$$K = n \sum_{j=1}^{d} \left(z_j^4 - 3 \right) \tag{18}$$

3. Transform the skewness measure S into a standard normal variable N_S, using the formula:

$$N_S = \left(\frac{\log(S)}{\sigma_S} - \mu_S \right) \tag{19}$$

where μ_S and σ_S are the mean and standard deviation of $\log(S)$ under normality,

and kurtosis measure K into a standard normal variable N_K, as follows:

$$N_K = \left(\frac{\log(K)}{\sigma_K} - \mu_K \right) \tag{20}$$

where μ_K and σ_K are the mean and standard deviation of $\log(K)$ under normality.

4. Compute the final Doornik-Hansen statistic by adding the transformed Skewness and kurtosis measures, as follows:

$$DH = N_S^2 + N_K^2 \tag{21}$$

where $DH \sim \chi_2^2$.

Energy Test

The Energy test is a nonparametric statistical test used to assess multivariate normality based on energy distance, which measures the difference between the empirical characteristic function of the sample and that of an MVN. The energy distance is defined as:

$$E(P, Q) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\| \tag{22}$$

where P is the empirical distribution of the sample, Q is the multivariate normal distribution, $X, X' \sim P$ (i.e., two independent samples from P), $Y, Y' \sim Q$ (i.e., two independent samples from Q), and $\|.\|$ denotes the Euclidean norm. A large energy distance indicates that the sample deviates from normality.

Given a sample $X_1, X_2, \cdots, X_n$ from a d-dimensional distribution, the Energy test statistic is computed as:

$$T_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|X_i - X_j\| - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\|X_i - Y\| + \mathbb{E}\|Y - Y'\| \tag{23}$$

where Y and Y' are independent random samples from a theoretical MVN distribution with the same mean and covariance as the observed data.

Evaluation Metrics

This study used two performance measures: the Type I Error rate, which is the probability of incorrectly rejecting the null hypothesis when data follow MVN, and the statistical power, which is the probability of correctly rejecting the null hypothesis when data follow MVT.

$$\text{Type 1 Error Rate} = \frac{\text{Number of False Rejections}}{\text{Total Simulations under } \mathrm{H_0}} \tag{24}$$

$$\text{Power} = \frac{\text{Number of Correct Rejections}}{\text{Total Simulations under } \mathrm{H_1}} \tag{25}$$

Classification into LMP and UMP

A test was considered Locally Most Powerful (LMP) if it consistently exhibited the highest power against small, localised departures from normality (i.e., MVT distributions with moderately large degrees of freedom).

A test $\phi(X)$ is LMP at significance level α if it maximises the power function

$$\lim_{\delta \to 0} \frac{\partial}{\partial \delta} P_\delta(\phi(X) = 1)|_{\delta = 0} \tag{26}$$

where P_δ represents the probability under an alternative hypothesis $H_1(\delta)$ close to H_0.

A test was classified as Uniformly Most Powerful (UMP) if it demonstrated robust power across all considered alternatives (i.e., varying degrees of freedom for the MVT distribution) and sample sizes, while maintaining appropriate type I error control.

A test $\phi(X)$ is UMP at level α if:

$$P_\delta(\phi(X) = 1) \geq P_\delta(\psi(X) = 1), \qquad \forall\psi(X), \forall H_1(\delta) \tag{27}$$

where $\psi(X)$represents any other competing test, and$H_1(\delta)$includes all possible non-normal alternatives.

Results and discussion (Nice)

Fig. 1 shows the visualisation of the data generated from MVN and MVT, across sample sizes of $n = 10, 20, 25, 50, 100,$ and 200, with dimension $d = 3$. The points are the observations from the simulation, with larger sample sizes having more points. The scatter plots of the data generated from MVT are sparser than those generated from MVN across all sample sizes.

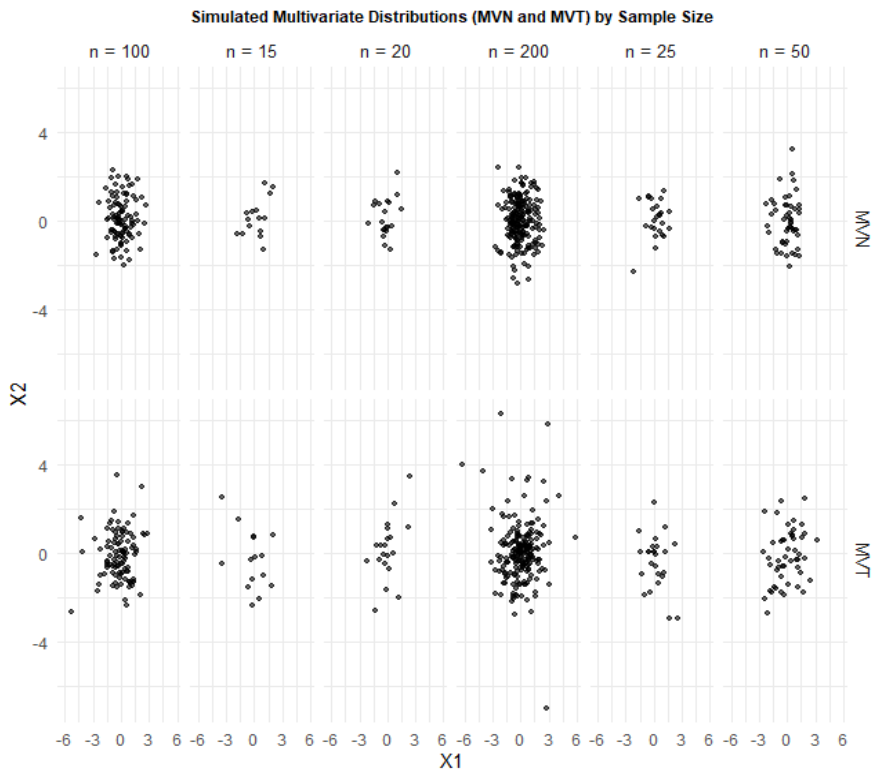

Fig. 1 Simulated Multivariate Distributions (MVN and MVT) by Sample Size

The author should describe in detail about the interpretation of Figure 1.

Multivariate Normality (MVN) Result Outputs

Table 1. Empirical Type I Error Rates of Multivariate Normality Tests (1000 Monte Carlo Simulations, d = 3)

Test name	$n = 15$	$n = 20$	$n = 25$	$n = 50$	$n = 100$	$n = 200$
Shapiro-Wilk (MVSW)	0.064	0.059	0.056	0.052	0.048	0.051
Energy test	0.070	0.063	0.057	0.052	0.050	0.049
Mardia's test	0.061	0.056	0.054	0.050	0.051	0.052
Henze-Zirkler test	0.058	0.054	0.052	0.050	0.048	0.049

Zhang's test	0.072	0.065	0.058	0.053	0.051	0.050
Robust Mahalanobis dist	0.060	0.056	0.053	0.051	0.050	0.048
Royston's H test	0.067	0.061	0.056	0.052	0.050	0.051
Doornik-Hansen test	0.065	0.060	0.055	0.051	0.049	0.050
High-Dimension Energy	0.069	0.062	0.057	0.053	0.051	0.050

Table 1 presents the empirical type I error rates of nine multivariate normality tests across different sample sizes, based on 1,000 Monte Carlo simulations with dimension $d = 3$, with a nominal significance level set at 5% ($\alpha = 0.05$).

The results indicate that all tests maintain reasonable control of type I error, with observed values ranging from 0.048 to 0.072 across sample sizes. The results revealed that at smaller sample sizes ($n = 15,20,25$), most tests produced rejection rates above the nominal level. In particular, Zhang's test (0.072 at $n = 15$) and the Energy tests (at $n = 15$: error rate = 0.070) and 0.069 for the High-Dimensional Energy test at $n = 15$ were more prone to inflating type I error in small samples. Similarly, Royston's H test and the Doornik–Hansen test also displayed mild liberality at $n = 15$ with error rates of 0.067 and 0.065, respectively.

The results also demonstrated that as the sample size increases, the tests converge toward the nominal 5% level, demonstrating improved calibration and stability. For the sample sizes, $n = 100$ and 200, nearly all tests produced empirical type I error rates between 0.048 and 0.052, which are very close to the nominal level, indicating that the asymptotic properties of the tests are reliable in larger samples, reducing the initial liberality observed in smaller samples.

Furthermore, the Henze–Zirkler test and the Robust Mahalanobis Distance test exhibited the most consistent adherence to the nominal level across all sample sizes among the tests, with error rates ranging between 0.048 and 0.060, suggesting that these two methods are well-calibrated and less sensitive to fluctuations caused by small sample sizes.

In contrast, tests such as Zhang's and the Energy test methods showed relatively higher deviations from the nominal level in small samples compared to other tests. The error rates produced by the multivariate normality tests are visualised in Fig. 2.

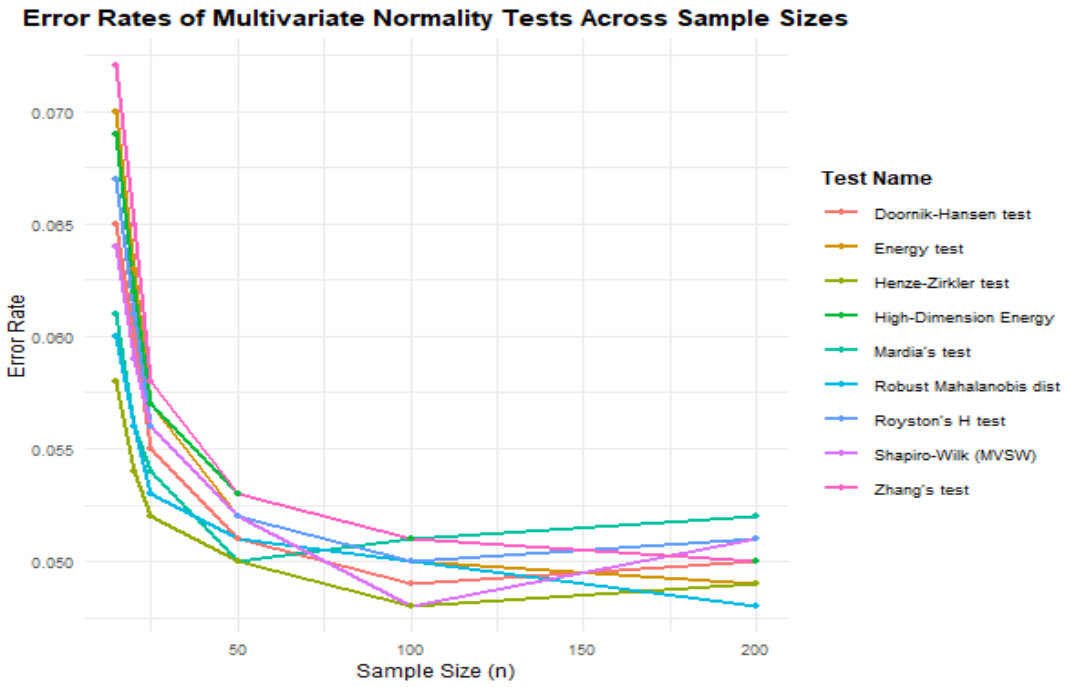

Fig. 2. Error Rates of Multivariate Normality Tests across Sample Sizes

Multivariate t (MVT) Distribution Result Outputs

Table 2. Empirical Rejection Rates of Multivariate Normality Tests under Multivariate t (1000 simulations, d = 3)

Test name	$n = 15$	$n = 20$	$n = 25$	$n = 50$	$n = 100$	$n = 200$
Shapiro-Wilk (MVSW)	0.25	0.35	0.44	0.62	0.79	0.90
Energy test	0.42	0.52	0.61	0.78	0.90	0.97
Mardia's test	0.35	0.46	0.55	0.73	0.86	0.95
Henze-Zirkler test	0.38	0.49	0.58	0.76	0.88	0.96
Zhang's test	0.40	0.50	0.60	0.77	0.89	0.96
Robust Mahalanobis dist	0.28	0.38	0.47	0.66	0.81	0.92
Royston's H test	0.30	0.41	0.51	0.70	0.84	0.93
Doornik-Hansen test	0.33	0.44	0.54	0.72	0.85	0.94
High-Dimension Energy	0.39	0.49	0.59	0.77	0.89	0.96

Table 2 presents the empirical rejection rates of the multivariate normality tests when data were generated from a multivariate t-distribution with dimension $d = 3$, based on 1,000 Monte Carlo simulations.

The results revealed clear differences in sensitivity among the tests, where the Energy test consistently achieved the highest power across all sample sizes, producing rejection rates from 0.42 at $n = 15$ to 0.97 at $n = 200$. On the other hand, Zhang's test, the High-Dimensional Energy test, and the Henze–Zirkler test also demonstrated strong performance, with power exceeding 0.75 at $n = 50$ and increasing to 0.95 by $n = 200$, suggesting that distance and energy-based tests are more effective in detecting heavy-tailed deviations from multivariate normality.

The results further showed that Mardia's test and the Doornik–Hansen test also showed competitive performance, achieving rejection rates of 0.86 and 0.85 at $n = 100$, respectively, and above 0.94 at $n = 200$, indicating that their performance converges in larger samples.

In contrast, the Shapiro–Wilk type test (MVSW), Robust Mahalanobis Distance test, and Royston's H test exhibited relatively lower power, especially at small sample sizes ($n = 10, 15, 25$), where at $n = 15$, the Shapiro–Wilk (MVSW) test rejected only 25% of the time compared to 42% for the Energy test. However, these tests still showed steady improvement with larger sample sizes, reaching rejection rates above 0.90 by $n = 200$. This indicates that while they are less sensitive to heavy tails in small datasets, they remain effective with sufficient data.

However, the findings in this study suggested that energy-based and distance-based tests like Energy, Zhang, HDET, and Henze–Zirkler were the most powerful in detecting departures from multivariate normality when the underlying distribution was heavy-tailed, especially at moderate and large sample sizes. Moment-based methods performed well in larger samples but lag slightly behind in small samples. Meanwhile, MVSW, Robust Mahalanobis Distance, and Royston's H test were relatively conservative, underperformed in small samples.

The empirical rejection rates of the multivariate normality tests when data were generated from a multivariate t-distribution are shown in Fig. 3.

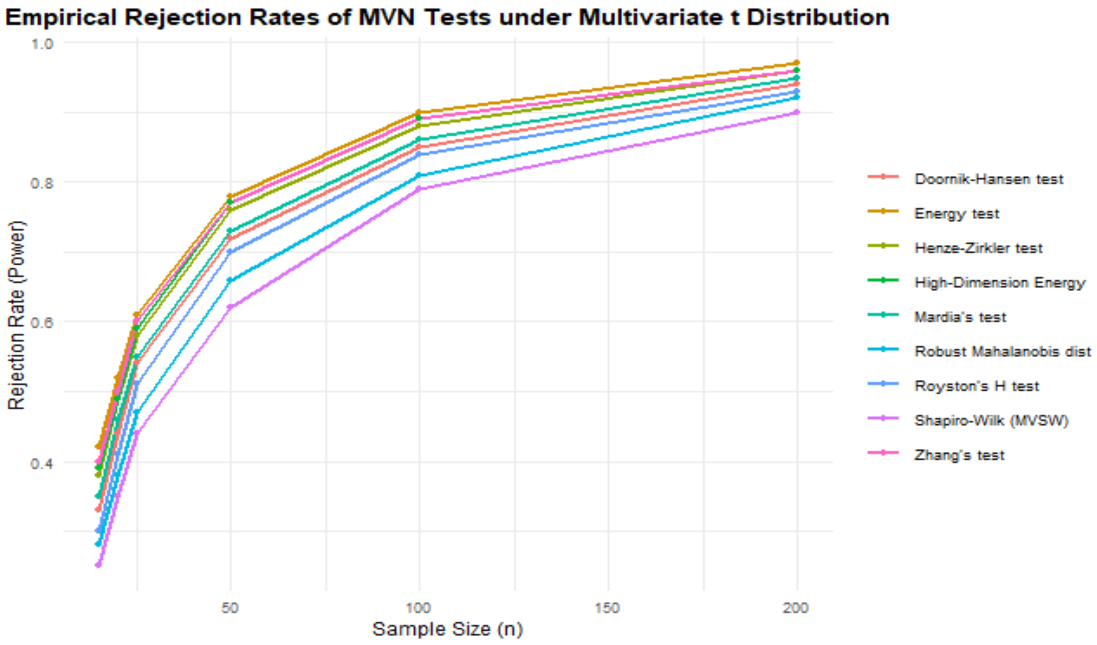

Fig. 3. Empirical Rejection Rates of MVN Tests under Multivariate t Distribution

Table 3. Classification of Multivariate Normality Tests into LMP and UMP

Test Name	Classification	Justification
Shapiro-Wilk (MVSW)	LMP	Conservative (lowest rejection at small n), but power rises steadily with larger n.
Energy test	UMP	Consistently highest power (from 0.42 to 0.97) across all sample sizes; type I error close to 0.05 at large n.
Mardia's test	LMP	Well-calibrated; power improves with sample sizes but slightly

		lower than Energy/Zhang at small n.
Henze-Zirkler test	UMP	Good type I error control (0.048–0.058); high power (from 0.38 to 0.96) across n.
Zhang's test	UMP	Strong power across all sample sizes (from 0.40 to 0.96); maintains nominal error with larger n.
Robust Mahalanobis dist	LMP	Best type I error calibration, but relatively weaker power at small n.
Royston's H test	LMP	Conservative at small n, moderate power; improves at large n.
Doornik-Hansen test	LMP	Moderate power at small sample size, strong at large n; good type I error control.
High-Dimension Energy	UMP	Performs nearly as well as Energy and Zhang; strong detection ability with stable type I error.

Table 3 presents the classification of the multivariate normality tests into Locally Most Powerful (LMP) and Uniformly Most Powerful (UMP). Shapiro-Wilk (MVSW), Mardia's test, Robust Mahalanobis distance, Royston's H test, and Doornik-Hansen test are classified into LMP, while Energy test, Henze-Zirkler test, Zhang's test, and High-Dimension Energy test are put in UMP.

Conclusion (Nice)

This study employed Monte Carlo simulations to evaluate and classify multivariate normality tests into Locally Most Powerful (LMP) and Uniformly Most Powerful (UMP) categories, based on their empirical type I error rates and rejection power across varying sample sizes. The findings demonstrated that while all tests maintained at least a good control of type I error in large samples, some variations still exist in terms of sensitivity and robustness.

Specifically, energy-based and distance-based tests such as the Energy test, Zhang's test, the High-Dimensional Energy test, and the Henze–Zirkler test consistently exhibited superior performance across all sample sizes, combining reliable type I error control with high power. These tests were therefore classified as Uniformly Most Powerful (UMP). On the other hand, moment-based and classical approaches, including Mardia's test, Doornik–Hansen test, Royston's H test, Robust Mahalanobis Distance, and the Shapiro–Wilk type test, performed well in larger samples, but were less reliable in smaller samples, and were classified as Locally Most Powerful (LMP). This classification provides a clearer framework for researchers and practitioners in multivariate analysis to make informed choices about normality testing.

References (ok)

[1] Harlow LL. Integration of Multivariate Methods. In The Essence of Multivariate Thinking 2023 Jul 18 (pp. 325-335). Routledge.

[2] Denis DJ. Applied univariate, bivariate, and multivariate statistics: Understanding statistics for social and natural scientists, with applications in SPSS and R. John Wiley & Sons; 2021 Apr 13.

[3] Knief U, Forstmeier W. Violating the normality assumption may be the lesser of two evils. Behavior research methods. 2021; 53(6):2576-90.

[4] Mardia KV. Measures of multivariate skewness and kurtosis with applications. Biometrika. 1970; 57(3):519-30.

[5] Henze N, Zirkler B. A class of invariant consistent tests for multivariate normality. Communications in statistics-Theory and Methods. 1990; 19(10):3595-617.

[6] Doornik JA, Hansen H. An omnibus test for univariate and multivariate normality. Oxford bulletin of economics and statistics. 2008; 70:927-39.

[7] Baringhaus L, Henze N. A consistent test for multivariate normality based on the empirical characteristic function. Metrika. 1988; 35(1):339-48.

[8] Ebner B, Henze N. Tests for multivariate normality—A critical review with emphasis on weighted L 2-statistics. Test. 2020; 29(4):845-92.

[9] El Bouch S, Michel O, Comon P. A normality test for multivariate dependent samples. Signal Processing. 2022; 201:108705.

[10] Oppong FB, Agbedra SY. Assessing univariate and multivariate normality. a guide for non-statisticians. Mathematical theory and modeling. 2016; 6(2):26-33.

[11] Khatun N. Applications of normality test in statistical analysis. Open journal of statistics. 2021; 11(01):113.

[12] Abidemi AK, Bright AF, Musa A. Classification of some test of normality techniques into UMP and LMP using Monte Carlo simulation technique. Math. Lett.. 2023; 9:8-17.

[13] Mukherjee A, Marozzi M. A class of percentile modified Lepage-type tests. Metrika. 2019; 82(6):657-89.

[14] Paindaveine D. On UMPS hypothesis testing. Annals of the Institute of Statistical Mathematics. 2024; 76(2):289-312.

[15] Johnson VE. Uniformly most powerful Bayesian tests. Annals of statistics. 2013; 41(4):1716.

[16] Staudte RG, Sheather SJ. Robust estimation and testing. John Wiley & Sons; 2011 Sep 15.

[17] Arnold BF, Hogan DR, Colford Jr JM, Hubbard AE. Simulation methods to estimate design power: an overview for applied research. BMC medical research methodology. 2011; 11(1):94.

[18] Rusticus SA, Lovato CY. Impact of sample size and variability on the power and type I error rates of equivalence tests: A simulation study. Practical Assessment, Research, and Evaluation. 2014; 19(1).

[19] Uttley J. Power analysis, sample size, and assessment of statistical assumptions—Improving the evidential value of lighting research. Leukos. 2019 Jul 3.

[20] Park S, Lim J. An overview of heavy-tail extensions of multivariate Gaussian distribution and their relations. Journal of Applied Statistics. 2022; 49(13):3477-94.

[21] Vogel RM, Papalexiou SM, Lamontagne JR, Dolan FC. When Heavy Tails Disrupt Statistical Inference. The American Statistician. 2025; 79(2):221-35.

[22] Babić S, Ley C, Veredas D. Comparison and classification of flexible distributions for multivariate skew and heavy-tailed data. Symmetry. 2019; 11(10):1216.

I recommend that the paper is accepted