Journal of Statistical Computation & Simulation
Vol. 00, No. 00, January 2006, 1–14

On tests for multivariate normality and associated simulation studies

Patrick J. Farrell[†] Matias Salibian-Barrera[*] Katarzyna Naczk[‡]

(Received 00 Month 200x; In final form 00 Month 200x)

We study the empirical size and power of some recently proposed tests for multivariate normality and compare them with existing proposals that performed best in previously published studies. We show that the Royston (1983b) extension to the Shapiro and Wilk (1965) test is unable to achieve the nominal significance level, and consider a subsequent extension proposed by Royston (1992) to correct this problem, which earlier studies appear to have ignored. A consistent and invariant test proposed by Henze and Zirkler (1990) is found to have good power properties, particularly for sample sizes of seventy-five or more, while an approach suggested by Royston (1992) performs effectively at detecting departures from multivariate normality for smaller sample sizes. We also compare our results to those of previous simulation studies, and discuss the challenges associated with generating multivariate data for such investigations.

[†] [‡] School of Mathematics and Statistics, Carleton University, 1125 Colonel By Drive, Room 4302 Herzberg Building, Ottawa, ON, K1S 5B6 - Canada.

[*] Department of Statistics, 333 - 6356 Agricultural Road, The University of British Columbia, Vancouver, BC, V6T 1Z2 - Canada

1 Introduction

It is often the case that studies such as clinical trials, business marketing investigations, and sociology and psychology experiments involve multivariate response data. Many of the procedures required to analyze such data, including MANOVA, discriminant analysis and multivariate regression, assume multivariate normality (MVN). Simulation studies conducted by Hopkins and Clay (1963), Mardia (1975), and Conover and Iman (1980) emphasize the importance of the MVN assumption for many of these procedures, illustrating that many of them lack robustness when they are applied to non-multivariate normal data.

Despite the sensitivity of these multivariate inferential techniques to the MVN assumption, and the vast number of tests that have been proposed for detecting departures from MVN, the assumption frequently goes untested. Looney (1995) lists a number of reasons for the reluctance to test for MVN, including the lack of awareness of the existence of the tests, the limited availability of software, and the lack of information regarding size and power. This article focuses on the latter issue of size and power. We examine via a Monte Carlo simulation the performance of some of the more promising tests; some of which appear to have received little or no attention in the literature.

There exists a vast number of proposed methods for testing MVN. A recent review by Mecklin and Mundfrom (2005) lists over fifty different procedures. However, despite the large number of approaches, these authors also found that extremely little effort has been directed towards assessing the size and power of these tests. In addition to their extensive simulation study involving a power comparison of thirteen different approaches for a wide variety of alternative distributions, Mecklin and Mundfrom (2005) also cite Ward (1988), Horswell (1990), Horswell and Looney (1992), Romeu and Ozturk (1993), and Bogdan (1999) as being among the few studies that have been concerned with the size and power of tests for MVN. The study of Mecklin and Mundfrom (2005) is based on the results of Mecklin (2000).

In their review of tests for MVN, Mecklin and Mundfrom (2005) indicate that these tests can be categorized into one of four groups: goodness of fit techniques, procedures based on skewness and kurtosis, consistent and invariant tests, and graphical and correlational approaches. They also observed that none of the studies listed above was exhaustive, and that most were designed to focus on tests that fell into only one of the categories above. For example, Romeu and Ozturk (1993) investigated the power of tests based on goodness of fit techniques, while studies conducted by Horswell (1990), Horswell and Looney (1992), and Doornik and Hansen (1994) focused on tests based on skewness and kurtosis. Henze and Zirkler (1990) compared the power of consistent and invariant tests.

Similar to Mecklin and Mundfrom (2005), we shall investigate the size and power of different tests that cut across the major categories. However, our study differs from theirs in that we do not examine a large number of tests. Rather, we make use of the results of previous studies conducted by Romeu and Ozturk (1993), Doornik and Hansen (1994), and Henze and Zirkler (1990) and do not consider tests that have been shown to possess relatively low power. Since each of these studies focus predominantly on tests that fall into one of the first three major categories listed above (goodness of fit, skewness and kurtosis, consistent and invariant tests), we choose to compare one of the more powerful tests from each category. Therefore, despite the fact that only three tests are evaluated here, our findings, when combined with those of previous investigations, are quite comprehensive.

Specifically, among the tests we consider is one based on a revision given in Royston (1992) of Royston's (1983b) extension of the Shapiro and Wilk (1965) goodness of fit test for univariate normality. Royston (1992) warned that this revision to the Royston (1983b) extension was necessary, pointing to a problem that leads to an incorrect specification of the null distribution. To our knowledge, the present study marks the first effort to assess the size and power of the 1992 revised version. In fact, we illustrate via simulation that the Royston (1983b) test statistic does not achieve the nominal significance level.

We choose the Royston (1992) test for investigation here as the Shapiro and Wilk (1965) test has been found to be among the more powerful tests for detecting departures from univariate normality, yielding comparable results for small samples to those of the Spiegelhalter (1977, 1980) tests for many different alternative distributions. In fact, Srivastava and Hui (1987) state that the Shapiro and Wilk (1965) test "... has been found to be the best omnibus test for detecting departures from univariate normality". We shall also consider a relatively new test of MVN proposed by Doornik and Hansen (1994) in a working paper that is based on multivariate measures of skewness and kurtosis. These authors conducted a small simulation study that suggests that their statistic has better power properties than other tests based on skewness and kurtosis. Thus, the promising power results associated with this test relative to others within the same category that appear in refereed journals prompted us to explore it further. Finally, we also study a consistent and invariant test proposed by Henze and Zirkler (1990) that is based on an extension of the Epps and Pulley (1983) test to the multivariate case. This test was found by Henze and Zirkler (1990) to be relatively powerful for detecting departures from MVN.

In Section 2, we discuss the reviews of tests for MVN and the associated simulations that have been conducted. In Section 3, we describe and report the results of a simulation designed to estimate the size and power of the three tests considered here. We note that our results differ noticeably from those obtained

in previous studies. We also illustrate that the Royston (1983b) extension to the Shapiro and Wilk (1965) test is unable to achieve the nominal significance level. Finally, conclusions and discussion are provided in Section 4.

2 Tests for Multivariate Normality

In this section we briefly review consistent and invariant tests for assessing MVN, along with procedures based on goodness of fit and measures derived from skewness and kurtosis. We use this review to motivate our choice of tests investigated in this study. Note that we do not consider MVN tests arising from correlational procedures due to their relatively poor size and power properties; see, for example, the simulation studies of Young, Seaman, and Seaman (1995), Mecklin (2000), and Mecklin and Mundfrom (2005).

Goodness of fit tests. Romeu and Ozturk (1993) provide a review and comparative study of many goodness of fit tests for MVN. In addition, Mudholkar, McDermott, and Srivastava (1992) proposed a test that is a simple adaptation of the Lin and Mudholkar (1980) test for univariate normality. A simulation study demonstrated that this test is able to achieve the nominal significance level for p variates in the range $2 \leq p \leq 6$ and for sample size $n \geq 10$, and that it is reasonably powerful against long-tailed alternatives. Justel, Pena, and Zamar (1997) introduced a distribution-free multivariate Kolmogorov-Smirnov test and developed an algorithm to compute the associated statistic in the bivariate case. A preliminary investigation suggested that its power does not compare favourably to that of the other tests selected for our study. More recently, goodness of fit tests based on generalized chi-square quantiles as discussed by Einmahl and Mason (1992) have also been proposed by Beirlant, Mason, and Vynckier (1999).

Most reviews and comparative studies of tests for MVN refer to the Royston (1983b) extension of the powerful Shapiro and Wilk (1965) goodness of fit test for univariate normality (for example, see Romeu and Ozturk 1993, Doornik and Hansen 1994, Looney 1995, Mecklin 2000, and Mecklin and Mundfrom 2005). The Shapiro and Wilk (1965) test was originally proposed for sample sizes n between 3 and 50. Royston (1982b) extended this test to the cases $3 \leq n \leq 2000$ and provided a suitable normalizing transformation. An algorithm for computing this extension is given in Royston (1982a) and Royston (1983a). Specifically, if $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ represents an ordered univariate sample, the Shapiro and Wilk (1965) test statistic is given by

$$W = (\sum_{i=1}^{n} a_i X_{(i)})^2 / \sum_{i=1}^{n} (X_{(i)} - \bar{X})^2, \tag{1}$$

where the vector of weights $a = (a_1, \ldots, a_n)'$ are normalized "best linear unbiased" coefficients. Values for a_i are tabulated for $n \leq 20$ by Sarhan and Greenberg (1956). For $n > 20$, Royston (1982b) points to an approximation for a_i proposed by Shapiro and Wilk (1965).

Provided that the sample $X_1, \ldots, X_n$ comes from a normal distribution, Royston (1982b) suggested the normalizing transformation $Z = [(1 - W)^\lambda - \mu]/\sigma$, where λ, μ, and σ are calculated according to Royston (1982b). Royston (1983b) extended this idea to the multivariate case. For the jth variate, $j = 1, \ldots, p$, the data vector $X_j = (X_{1j}, \ldots, X_{nj})'$ is used to compute W_j, the corresponding univariate Shapiro and Wilk (1965) test statistic. The normalizing transformation is then applied to each W_j to determine $Z_j = [(1 - W_j)^\lambda - \mu]/\sigma$, where λ, μ, and σ are obtained as before. Next, one computes $K_j = \{\phi^{-1}[\phi(-Z_j)/2]\}^2$, where $\phi(\cdot)$ denotes the standard normal cumulative distribution function. Then, if $(X_1, \ldots, X_p)'$ is jointly multivariate normal and its components mutually independent, then $G = \sum_{j=1}^{p} K_j$ is approximately χ^2_p. If the X_j's are not independent, then $H = e\,G$ is approximately χ^2_e, where e is referred to as the *equivalent degrees of freedom*. Royston (1983b) suggests an estimate for e based on the method of moments, $\hat{e} = p/[1 + (p - 1)\bar{c}]$, where $\bar{c}$ is an estimate for the average correlation among the K_j.

Unfortunately, the results of the simulation in Naczk (2004) show that the statistic based on Royston (1983b) is unable to achieve the nominal significance level. Indeed, Royston (1992) revised the procedure to approximate a. Referring to this article, Royston (1993) stated that "Investigation revealed that Shapiro and Wilk's (1965) approximation to the weights (a), used by Royston (1982a,b) was inadequate, therefore the earlier 'W test' differed seriously from the true test". Royston (1995) provides an algorithm that uses the revised procedure for any $3 \leq n \leq 5000$. Despite the findings of Royston (1992), the recent reviews and simulation studies of MVN tests listed above only refer to the Royston (1983b) approach that employs the weights in Royston (1982a,b).

We have consequently chosen to investigate the Royston (1992) revised procedure as a promising representative from the goodness of fit test category as a result of the power of the univariate Shapiro and Wilk (1965) test upon which it is based. Moreover, to our knowledge, no study published to date has included Royston's (1992) approach. We also consider values of p larger than those studied in Royston (1992), namely $p = 4$, 5, and 10.

Skewness and kurtosis tests. Horswell (1990) reports results from a simulation study on the performance of tests in this class. Among others, Horswell considered the tests proposed in the landmark paper of Mardia (1970). Mardia (1970) introduced measures of skewness and kurtosis, demonstrated that functions of these variables were asymptotically distributed as chi-square and

standard normal, respectively and derived two tests of multivariate normality. Despite the widespread use of Mardia's (1970) statistics, Horswell (1990) demonstrated that, generally speaking, MVN tests based on measures of skewness and kurtosis did not distinguish well between 'skewed' and 'non-skewed' distributions. Subsequent simulation studies by Horswell and Looney (1992), Mecklin (2000), and Mecklin and Mundfrom (2005) confirm the relatively low power of these tests. To improve upon power, some authors have attempted to combine measures of skewness and kurtosis into a single 'omnibus' test statistic. Mardia and Foster (1983) derived six statistics, including one that uses the Wilson-Hilferty approximation (Wilson and Hilferty, 1931). However, Horswell and Looney (1992) found that this statistic lacked power.

More recently, Doornik and Hansen (1994) have proposed a simple omnibus MVN test based on measures of skewness and kurtosis that is an extension of the univariate test proposed by Shenton and Bowman (1977). For each of the p variates, the measure of skewness, $\sqrt{b_{1j}}$ for $j = 1, \ldots, p$, is transformed to a standard normal Z_{1j} as in D'Agostino (1970), while the measure of kurtosis, b_{2j}, is transformed from a gamma distribution to a chi-square and then to a standard normal Z_{2j} using the Wilson-Hilferty cubed root transformation. Doornik and Hansen (1994) propose the statistic $Z_1'Z_1 + Z_2'Z_2$, where $Z_1 = (Z_{11}, \ldots, Z_{1j}, \ldots, Z_{1p})'$ and $Z_2 = (Z_{21}, \ldots, Z_{2j}, \ldots, Z_{2p})'$. This statistic has a χ^2_{2p} asymptotic distribution when the data are MVN.

Using a Monte Carlo simulation study, Doornik and Hansen (1994) compared their proposed method against four other statistics for testing MVN, including those of Mardia (1970), and the Royston (1983b) extension of the Shapiro and Wilk (1965) test. They demonstrated that their test was able to achieve the nominal significance level, and that it possessed good power properties, bettering the other tests in the comparison in this regard.

We have selected the Doornik and Hansen (1994) statistic from the group of tests based on multivariate measures of skewness and kurtosis for evaluation in this study. The initial work of these authors suggest that the test is promising.

Consistent and invariant tests. Many tests in the three classes mentioned above have been criticized because they may not be consistent and/or invariant under linear transformations of the data. One test that is simultaneously consistent for any non-MVN distribution and invariant has been proposed by Epps and Pulley (1983). It is based on

$$T = \int_{-\infty}^{\infty} |\phi_n(t) - \hat{\phi}_0(t)|^2 dG(t)\,, \tag{2}$$

where $\phi_n(t)$ is the empirical characteristic function, $\hat{\phi}_0(t)$ is an estimate of the characteristic function of the normal distribution, and $G(t)$ is a weight

function.

Csörgő (1989) proved that the test statistic in (2) is consistent, and also showed the consistency of an extension of this test to the multivariate case proposed by Baringhaus and Henze (1988). Henze and Zirkler (1990) proposed a multivariate extension of (2), namely

$$D_{n\,\beta} = \int_{\mathbb{R}^d} |\Phi_n(t) - \hat{\Phi}_0(t)|^2 \varphi_\beta(t) dt\,, \tag{3}$$

where $\Phi_n(t)$ is the empirical characteristic function of the standardized observations, $\hat{\Phi}_0(t)$ is the characteristic function of a multivariate standard normal distribution, and $\varphi_\beta(t)$ is a kernel function. Henze and Zirkler (1990) use the density function of a $N_p(0, \beta^2\, I_p)$ random vector ($\beta \in \mathbb{R}$) as the kernel in (3), they show that the test statistic has a lognormal asymptotic distribution and derive a closed form expression for $D_{n\,\beta}$. They also provide a formula for determining an optimal choice of β for each n and p. The consistency of this test follows directly from Csörgő (1989).

Using various values of β, Henze and Zirkler (1990) conducted a simulation study to compare their statistic with others, including Mardia's (1970) multivariate measures of skewness and kurtosis. A number of alternative distributions were considered, including those with independent marginals, mixtures of normal distributions, and spherically symmetric distributions. Henze and Zirkler (1990) demonstrated that their test had good power, and also found that the choice $\beta = 0.5$ produced a powerful test against alternative distributions with heavy tails. The Henze and Zirkler (1990) statistic was also one of the thirteen considered in the simulation study conducted by Mecklin and Mundfrom (2005) that evaluated the size of these tests, and also considered their power over a wide variety of alternative distributions. Based on the results of their simulation study, Mecklin and Mundfrom (2005) recommend the Henze and Zirkler (1990) test for assessing MVN.

We have therefore selected the Henze and Zirkler (1990) test from the class of invariant and consistent tests. Moreover, note that this test has not yet been compared against the corrected Royston test (Royston 1992).

3 Simulation Study

A simulation was conducted to compare the size and power of the Royston (1992), Doornik and Hansen (1994), and Henze and Zirkler (1990) tests (R92, DH, and HZ). We also include the Royston (1983b) statistic, R83, in the size comparison to illustrate that it is not possible to achieve the nominal significance level with this test. To assess the size of the four tests we used a

significance level of 0.05. We generated 10,000 multivariate normal samples for specified values of n and p using the functions `rnorm` and `rmvnorm` in S-PLUS 6.2. For each data set, the R83, R92, DH, and HZ statistics were calculated, along with associated P-values. The size of each test was estimated by the proportion of the 10,000 samples for which the P-value was less than 0.05. We considered all combinations of n = 25, 50, 75, 100, and 250 with p = 2, 3, 4, 5, and 10. The S-PLUS and functions used to compute the test statistics are available at `http://hajek.stat.ubc.ca/~matias/MVN`.

Table 1 presents the empirical Type I error rates for the four tests. The results clearly show that R83 does not achieve the nominal significance level; the maximum empirical Type I error rate for this test was 1.01% with n = 75 and p = 2. In addition, all estimates were zero for R83 when n = 25 or 250. These findings are in rather sharp contrast to those obtained by Mecklin (2000) and Mecklin and Mundfrom (2005) who found empirical levels ranging between 4.7% and 5.3%. Our results are in agreement with the comments of Royston (1992) regarding the inadequacy of the choice of weights suggested in Royston (1983b). Based on these observations, we did not to include the R83 test in our power comparisons.

Royston's (1992) test produced the best results regarding empirical Type I error rates, which ranged between 4.54% and 5.26% over all combinations of n and p. The estimates for the DH test were also extremely good in all cases. The empirical Type I error rates for the HZ test were conservative for small n, in particular for n = 25, where rates ranging from 3.23% to 4.09% were obtained. The test was still somewhat conservative for n = 50, but yielded estimates that approached the 0.05 nominal rate for values of n = 75 or more.

Regarding a comparison of the power of the tests, according to Mecklin and Mundfrom (2005), "In a Monte Carlo study, it is important to carefully choose the distributions to be simulated." The set of alternative distributions employed here was chosen for comparative purposes to emulate those used in the Mecklin and Mundfrom (2005) study that summarizes the results in Mecklin (2000). We chose to work with these distributions since, according to Mecklin (2000), the comparison of results from simulation studies for MVN tests "... is difficult, since there has been no complete uniformity in the test procedures studied or the alternatives to multivariate normality that were considered". Our choice of distributions addresses the latter concern. Following Mecklin (2000), to assess the power of the R92, DH and HZ tests, we used a variety of alternative distributions that included fifteen different mixtures of two multivariate normals, that were distinguished by three levels of contamination and five combinations of means and covariance matrices. In this regard, Mecklin (2000) wisely selected degrees of mixing that covered a wide range of contamination levels ranging from mild (skewed and leptokurtic) to moderate (skewed and mesokurtic) to severe (symmetric and platykurtic). The second contami-

nation level was of particular interest, as it reflects a non-normal distribution with normal kurtosis. In addition, to assess the effect on power of differences in certain parameters in the two multivariate normal distributions comprising the mixture, five different combinations of equal versus unequal means and variances were considered. We also investigated ten symmetric distributions from the elliptically contoured family that included the multivariate Cauchy, the multivariate t_{10}, and eight members of the Pearson Type II family, one of which was the multivariate uniform. Non-normal distributions with some characteristics identical to those of the multivariate normal were also studied. These distributions include the Khintchine and generalized exponential power distributions. While neither of these distributions is jointly multivariate normal, the former has normal marginals, while the latter possesses the same skewness and kurtosis as the multivariate normal. The power of tests against two heavily skewed distributions, the multivariate χ^2_1, and the multivariate lognormal, were also evaluated. Further details regarding the alternative distributions considered here can be found in Naczk (2004).

To estimate the power of the R92, DH, and HZ tests against each alternative distribution, we used S-PLUS 6.2 to generate 10,000 samples from a particular distribution for specified n and p and then computed the statistics for each test, along with the associated P-values. We used the algorithms described in Johnson (1987) to generate the multivariate samples. The S-PLUS functions used to generate these distributions are available at `http://hajek.stat.ubc.ca/~matias/MVN`. For each combination of n and p we estimated the power for each statistic at a significance level of 0.05 as the proportion of samples where the P-value was less than 0.05. The same values of n and p used for assessing size were used for the power comparison.

None of the tests performed well in detecting the multivariate normal mixtures. Most estimates of power were under 10%, even for large n and p. There were a few mixtures where the powers of all three tests exceeded 30%; these cases occurred when n and p were both large. For such situations, R92 yielded higher estimates than DH and HZ. In fact, R92 almost always had the highest power for all values of n and p across all mixtures, although in most cases the estimates for all tests were low (under 10%) and extremely similar.

By contrast, all three tests exhibited high power for skewed distributions. For both the multivariate χ^2_1 and the lognormal, for each test all estimates of power were 100% at all combinations of n and p with the exception of a few cases when $n = 25$ for the DH and HZ tests where the estimated power was almost 100%. The lowest power achieved was 95.28%; this occurred with HZ for $n = 25$ and $p = 10$. The R92 test always attained 100% power.

Similar findings were observed for the multivariate Cauchy (t_1). For $n = 50$ and larger, all tests achieved 100% power regardless of p. For $n = 25$, the powers for the tests ranged from 97.50% to 99.90% across all values of p, with

HZ having slightly larger power than the other tests at each p. By contrast, with n ranging from 25 to 100 and $2 \leq p \leq 5$, Mecklin (2000) obtains powers for HZ that range from 2.6% to 4.9%. Mecklin (2000) and Mecklin and Mundfrom (2005) do not consider R92 and DH. We attempt to explain our discrepancy with Mecklin (2000) below; however, our finding is consistent with the high power of univariate tests when applied to data from a Cauchy distribution, in particular the Shapiro and Wilk (1965) test, upon which R92 is based.

Our findings for HZ also differ from Mecklin (2000) for the multivariate t_{10} distribution. Figure 1 presents our power estimates for R92, DH, and HZ at various values of n and p. Across all tests and the values of p, the powers range from 15% to 46% when $n = 50$, and increase steadily with increasing n, ranging from 41% to 100% when $n = 250$. With the exception of some cases where n is small (25 or 50) or p is large, DH has the highest power. This test is definitely the one of choice when $n \geq 50$, and $p \leq 5$. When n is small and p is large, R92 possesses the highest power, while HZ is best when n and p are both large.

For $n = 100$ and $2 \leq p \leq 5$, Mecklin (2000) obtains powers for the multivariate t_{10} based on HZ that decrease from 5.2% to 3.6% as p increases. Over the same values of p, our counterpart estimates increase from 21.87% to 47.30%. However, note that we have not used the formula cited in Mecklin (2000) for generating multivariate t_ν data (also given in Johnson 1987, page 118). For such data, Mecklin (2000, page 90) gives the following formula:

$$X = (\frac{\sqrt{S}}{\nu})^{-1} Z + \mu \,, \tag{4}$$

where μ is the mean vector of X, Z is generated from a p-variate normal with mean vector zero and known covariance matrix, and S is generated independently from Z as a χ^2_ν variable. Unfortunately (4) does not reduce to a univariate t_ν distribution when $p = 1$ since the degrees of freedom parameter is not in the square root. Instead, Johnson and Kotz (1972, page 133) give

$$X = (\sqrt{\frac{S}{\nu}})^{-1} Z + \mu \,, \tag{5}$$

which is the formula we have used to generate multivariate Cauchy (t_1) and t_{10} data.

In Naczk (2004) we show that the empirical powers of R92, DH, and HZ when the data come from a multivariate Cauchy are very high. By contrast, the powers observed by Mecklin (2000) notably resemble the nominal level of the tests. This difference might be explained by the fact that if the same value for the χ^2_1 variable S is used for all n observation vectors generated with (4)

or (5) then the generated X's are simply translated multiples of Z, and thus have a MVN distribution. For such a situation, one would expect the observed "power" to be close to the nominal level, since the data distribution is in fact normal.

Figures 2 and 3 present the power estimates for the three tests for two of the Pearson Type II distributions. This family of distributions is indexed by a shape parameter $m > -1$. When $m = 0$ the distribution is also called multivariate uniform. The larger the value of m, the more this distribution resembles a multivariate normal (Johnson 1987, page 114). The results in Figure 2 are for the multivariate uniform distribution ($m = 0$), while those in Figure 3 are associated with a distribution having shape parameter $m = 10$. For $n = 50$ or larger, Figure 2 demonstrates that the HZ test is clearly the most powerful for the multivariate uniform distribution, especially for large p. This is true in general for the Pearson Type II distributions; however the power of all three tests decreases as the shape parameter m increases. For example, for $n = 100$ and $p = 3$, the power of the HZ test is 98.39% when $m = 0$, and only 6.22% when $m = 10$. Our results for the HZ test again differ noticeably from those in Mecklin (2000). Specifically, for values of $n = 25$, 50, and 100, and $p = 2$, 3, 4, and 5, Mecklin (2000) obtains nearly identical power estimates for the multivariate uniform ($m = 0$) distribution and the Pearson Type II distribution with $m = 10$, all of them very close to 100% for almost all combinations of n and p. However, when $m = 10$, the Pearson Type II distribution is already very close to a multivariate normal and, thus, one would expect a significant decrease in the empirical powers, as the one we observed in our study.

An interesting insight into the behaviour of the R92, DH, and HZ tests can be obtained from Figure 5. In this plot we considered the case $p = 3$, $n = 25$, 50, 75 and 100, and data following a multivariate uniform distribution (MUD). For each test and each value of n we display two boxplots comparing the empirical distributions of the test statistic when the data follow MVN and MUD distributions, respectively. For example, panel (c) ($n = 75$) consists of three pairs of boxplots, one for each test. The left boxplot in each pair contains the test statistics obtained with 10,000 simulated samples following a MVN distribution, while the right boxplot corresponds to test statistics based on 10,000 samples from a MUD distribution. The solid horizontal line for each pair of boxplots corresponds to the empirical 95% quantile of the null distribution (MVN) of the test (the 5%-level critical value). Similar to Figure 2, the plots in Figure 5 illustrate the improvement in power that is achieved by the HZ test as opposed to R92 and DH. However, we believe that the plot provides further information about how the underlying distribution of the data affects the behaviour of these test statistics, in particular how their shape, scale and location change, and how these changes subsequently affect

the power of the tests. For example, for a given n, the HZ test appears to obtain its power solely through a shift in the location of the distribution of the test statistic under MUD data relative to MVN. Moreover, the larger the sample size, n, the greater the shift, and hence the power.

We next consider the two distributions that possess some properties of the MVN, namely the Khintchine (KHN) and generalized exponential power (GEP) distributions. Figure 4 presents the power estimates based on KHN; the results indicate that HZ is the clear winner here. The power estimates for R92 are all around 5%, and those for DH are even worse. The power for HZ increases steadily with increasing n and p, ranging from 51.81% when $n = 75$ and $p = 3$, to 100% when n and p are both large. For the GEP distribution, R92 had the best power estimates for small $n = 25$, ranging from 90.12% to 100% as p increased from 2 to 5, followed by the DH test with estimates from 71.67% to 94.07%, and finally the HZ test that for $n = 25$ had maximum power of 61.75% with $p = 2$. For $n = 75$ of larger, all three tests achieved the maximum possible power of 100% for all p. While the results of Mecklin (2000) were somewhat similar to ours for HZ under KHN, those for GEP were quite different. For example, our power estimates for HZ under GEP were all 100% for $n = 100$ and $2 \leq p \leq 5$, while those of Mecklin (2000) ranged from 5.9% to 7.8%. We are unable to explain this difference; however throughout our investigation we verified to the best of our ability that our routines for multivariate data generation were producing samples from the appropriate distribution by constructing quantile-quantile plots and plotting histograms of the marginals.

4 Conclusion and Discussion

Many of the inferential procedures for analyzing multivariate data assume MVN, and it has been shown that their performance can be affected by violations of this assumption. We have reviewed many of the tests for assessing MVN and conducted a simulation to compare some of the more recent and promising ones, including those of Royston (1992), Doornik and Hansen (1994), and Henze and Zirkler (1990). We showed in the simulation that while these tests are able to maintain the nominal level, this was not the case for Royston (1983b).

In comparing our simulation results with those of other studies, we note some important differences that we believe may be due to the use of different data generation routines. The generation of multivariate data is no easy task, and must be performed with care. Throughout our investigation, we ensured to the best of our ability that the routines that we programmed for data generation were performing as expected by constructing quantile-quantile plots and exploring the corresponding marginal distributions.

The results of our simulation suggest that, relative to the other two tests considered, the Henze and Zirkler (1990) test generally possesses good power across the alternative distributions investigated, in particular for $n \geq 75$. The fact that the test is slightly conservative for small n may explain why it does not perform as well in these situations. In addition, the Henze and Zirkler (1990) test is not useful in detecting the reason(s) for departure from MVN. For small sample sizes around $n = 25$, the Royston (1992) test offers good power, relatively speaking, across all alternative distributions, while the power of the Doornik and Hansen (1994) test for the multivariate t_{10} is worthy of note. Regardless of the test used for assessing MVN, we also recommend the simultaneous use of graphical methods and numerical summaries as aids to diagnose the specific departure(s) from MVN that may exist.

Acknowledgments. This research was supported through funds from the Natural Sciences and Engineering Research Council of Canada. The authors are grateful to the Editor, an Associate Editor, and a referee for useful comments.

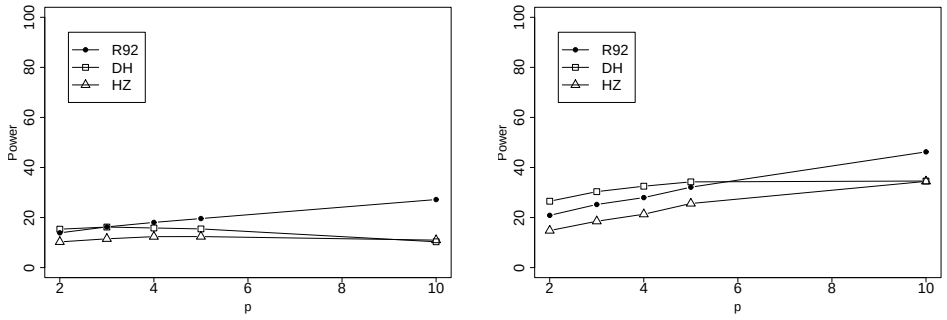
References

[1] Baringhaus, L., Henze, N., 1988. A consistent test for multivariate normality based on the empirical characteristic function. Metrika, 35, 339-348.
[2] Beirlant, J., Mason, D.M., Vynckier, C., 1999. Goodness of fit analysis for multivariate normality based on generalized quantiles. Computational Statistics and Data Analysis, 30, 119-142.
[3] Bogdan, M., 1999. Data driven smooth tests for bivariate normality. Journal of Multivariate Analysis, 68, 26-53.
[4] Conover, W.J., Iman, R.L., 1980. The rank transformation as a method of discrimination with some examples. Communications in Statistics - Theory and Methods, 9, 465-487.
[5] Csörgő, S., 1989. Consistency of some tests for multivariate normality. Metrika, 36, 107-116.
[6] D'Agostino, R.B., 1970. Transformation to normality of the null distribution of g_1. Biometrika, 57, 679-681.
[7] Doornik, J.A. and Hansen, D., 1994. An omnibus test for univariate and multivariate normality. Working Paper, Nuffield College, Oxford.
[8] Einmahl, J.H.J., Mason, D.M., 1992. Generalized quantile processes. Annals of Statistics, 20, 1062-1078.
[9] Epps, T.W., Pulley, L.B., 1983. A test for normality based on the empirical characteristic function. Biometrika, 70, 723-726.
[10] Henze, N., Zirkler, B., 1990. A class of invariant consistent tests for multivariate normality. Communications in Statistics - Theory and Methods, 19, 3595-3617.
[11] Hopkins, J.W., Clay, P.P.F., 1963. Some empirical distributions of bivariate T^2 and homoscredasticity criterion M under unequal variance and leptokurtosis. Journal of the American Statistical Association, 58, 1048-1053.
[12] Horswell, R.L., 1990. A Monte Carlo comparison of tests of multivariate normality based on multivariate skewness and kurtosis. Doctoral Dissertation, Louisiana State University.
[13] Horswell, R.L., Looney, S.W., 1992. A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis. Journal of Statistical Computation and Simulation, 42, 21-38.
[14] Johnson, M.E., 1987. Multivariate statistical simulation. Wiley, New York.
[15] Johnson, N.L., Kotz, S., 1972. Distributions in statistics: continuous multivariate distributions. Wiley, New York.
[16] Justel, A., Pena, D., Zamar, R., 1997. A multivariate Kolmogorov-Smirnov test of goodness of fit. Statistics and Probability Letters, 35, 251-259.

[17] Lin, C.C., Mudholkar, G.S., 1980. A simple test for normality against asymmetric alternatives. Biometrika, 67, 455-461.
[18] Looney, S.W., 1995. How to use tests for univariate normality to assess multivariate normality. American Statistician, 49, 64-70.
[19] Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika, 57, 519-530.
[20] Mardia, K.V., 1975. Assessment of multinormality and the robustness of Hotelling's T^2 test. Applied Statistics, 24, 163-171.
[21] Mardia, K.V., Foster, K., 1983. Omnibus tests of multinormality based on skewness and kurtosis. Communications in Statistics - Theory and Methods, 12, 207-221.
[22] Mecklin, C.J., 2000. A comparison of the power of classical and newer tests of multivariate normality. Doctoral Dissertation, University of Northern Colorado.
[23] Mecklin, C.J., Mundfrom, D.J. 2005. A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. Journal of Statistical Computation and Simulation, 75, 93 - 107.
[24] Mudholkar, G.S., McDermott, D., Srivastava, D.K. (1992). A test of p-variate normality. Biometrika, 79, 850-854.
[25] Naczk, K., 2004. Assessing tests for multivariate normality. MSc Thesis, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario.
[26] Romeu, J.L., Ozturk, A., 1993. A comparative study of goodness of fit tests for multivariate normality. Journal of Multivariate Analysis, 46, 309-334.
[27] Royston, J.P., 1982a. Algorithm AS 181: The W test for normality. Applied Statistics, 31, 176-180.
[28] Royston, J.P., 1982b. An extension of the Shapiro and Wilk's W test for normality to large samples. Applied Statistics, 31, 115-124.
[29] Royston, J.P., 1983a. Correction: Algorithm AS 181: The W test for normality. Applied Statistics, 32, 224.
[30] Royston, J.P., 1983b. Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. Applied Statistics, 32, 121-133.
[31] Royston, J.P., 1992. Approximating the Shapiro-Wilk W-Test for non-normality. Statistics and Computing, 2, 117-119.
[32] Royston, J.P., 1993. A Toolkit for testing for non-normality in complete and censored samples. The Statistician, 42, 37-43.
[33] Royston, J.P., 1995. Remark AS R94: A remark on Algorithm AS 181: The W test for normality. Applied Statistics, 44, 547-551.
[34] Sarhan, A.E., Greenberg, B.G., 1956. Estimation of location and scale parameters by order statistics from singly and double censored samples. Annals of Mathematical Statistics, 27, 427-451.
[35] Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika, 52, 591-611.
[36] Shenton, L.R., Bowman, K.O., 1977. A bivariate model for the distribution of $\sqrt{b_1}$ and b_2. Journal of the American Statistical Association, 72, 206-211.
[37] Spiegelhalter, D.J., 1977. A test for normality against symmetric alternatives. Biometrika, 64, 415-418.
[38] Spiegelhalter, D.J., 1980. An omnibus test for normality for small samples. Biometrika, 67, 493-496.
[39] Srivastava, M.S., Hui, T.K., 1987. On assessing multivariate normality based on the Shapiro Wilk W statistic. Statistics and Probability Letters, 5, 15-18.
[40] Ward, P.J., 1988. Goodness of fit tests for multivariate normality. Doctoral Dissertation, University of Alabama.
[41] Wilson, E., Hilferty, M., 1931. The distribution of chi-square. Proc. Nat. Acad. Sci, 17, 684-688.
[42] Young, D.M., Seaman, S.L., Seaman, J.W., 1995. A comparison of six test statistics for detecting multivariate nonnormality which utilize the multivariate squared-radii statistic", Texas Journal of Science, 47, 21-38.

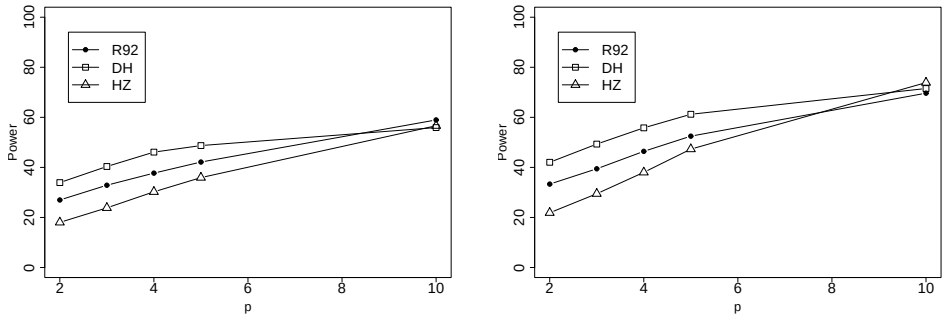
Table 1. Empirical Type I error rates (in percent) based on 10,000 samples and a 5% significance level for the Royston (1983b), Royston (1992), Doornik and Hansen (1994), and Henze and Zirkler (1990) test statistics for various values of n and p.

n	Test	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 10$
25	R83	0.00	0.00	0.00	0.00	0.00
	R92	4.54	5.17	5.07	5.11	4.93
	DH	4.64	5.21	5.39	5.06	4.25
	HZ	4.09	3.39	3.79	3.23	3.44
50	R83	0.62	0.28	0.32	0.16	0.03
	R92	5.14	4.75	4.87	4.87	5.04
	DH	5.52	4.64	5.09	4.85	4.48
	HZ	4.80	4.53	4.44	4.65	4.17
75	R83	1.01	0.62	0.56	0.77	0.22
	R92	4.62	4.61	4.77	5.11	5.23
	DH	5.03	4.66	4.78	5.54	4.89
	HZ	4.57	4.35	4.82	4.82	4.88
100	R83	0.60	0.50	0.32	0.34	0.09
	R92	4.94	4.93	5.04	4.75	5.05
	DH	4.98	5.17	5.17	5.06	5.15
	HZ	4.95	5.13	5.09	4.99	4.55
250	R83	0.01	0.00	0.00	0.00	0.00
	R92	5.06	4.97	5.26	5.16	5.12
	DH	5.26	5.41	5.60	5.46	5.57
	HZ	4.72	4.84	5.06	5.10	5.85



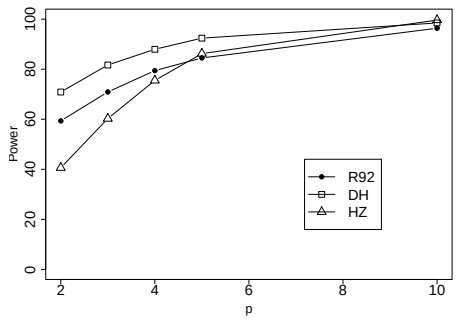
(a) $n = 25$

(b) $n = 50$



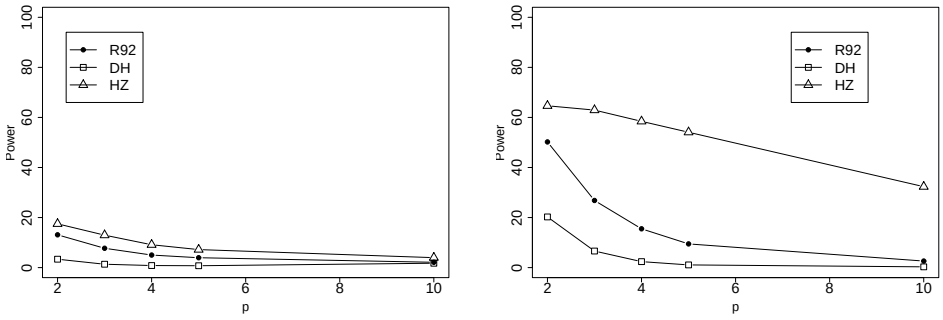
(c) $n = 75$

(d) $n = 100$



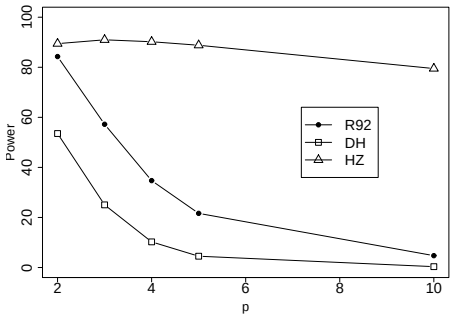
(e) $n = 250$

Figure 1. Empirical powers for the Royston (1992) [R92], Doornik and Hansen (1994) [DH], and Henze and Zirkler (1990) [HZ] test statistics for the Multivariate T distribution with 10 degrees of freedom. Based on 10,000 samples of sizes $n = 25$, 50, 75, 100 and 250, and $p = 2$, 3, 4, 5 and 10.

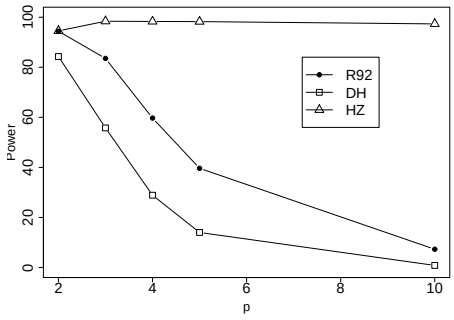


(a) $n = 25$

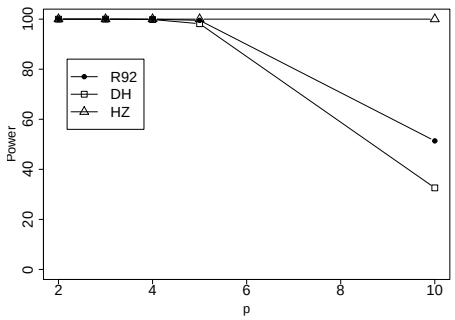
(b) $n = 50$



(c) $n = 75$

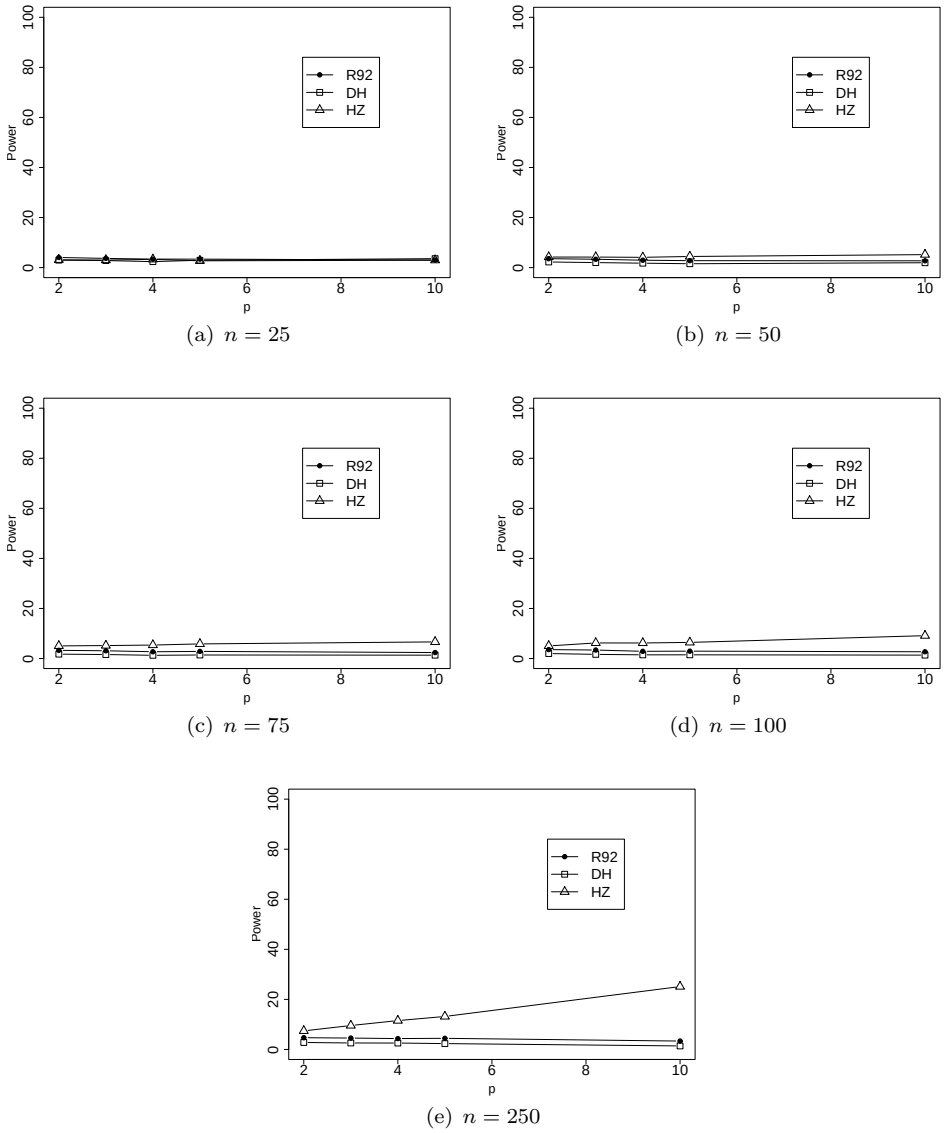


(d) $n = 100$



(e) $n = 250$

Figure 2. Empirical powers for the Royston (1992) [R92], Doornik and Hansen (1994) [DH], and Henze and Zirkler (1990) [HZ] test statistics for the Multivariate Uniform distribution. Based on 10,000 samples of sizes $n = 25$, 50, 75, 100 and 250, and $p = 2$, 3, 4, 5 and 10.

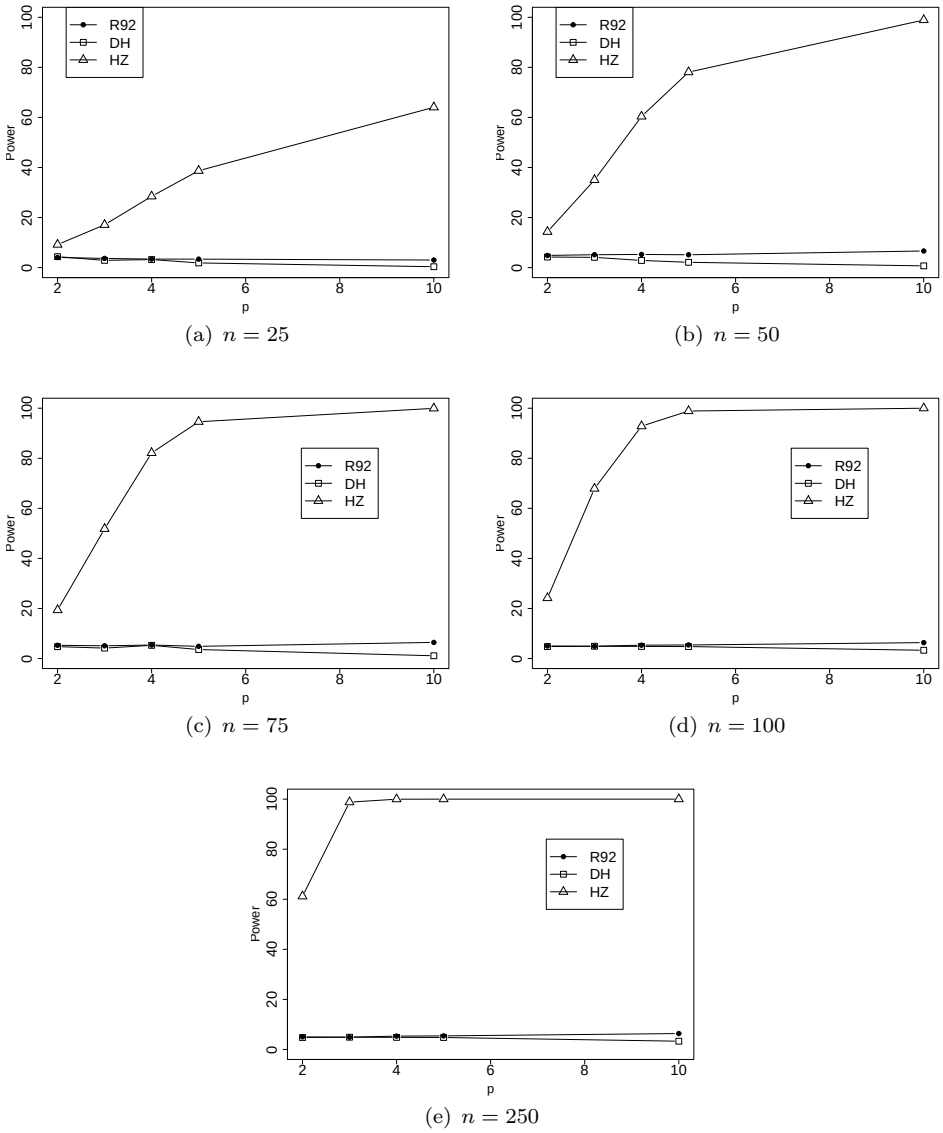


(a) $n = 25$ (b) $n = 50$

(c) $n = 75$ (d) $n = 100$

(e) $n = 250$

Figure 3. Empirical powers for the Royston (1992) [R92], Doornik and Hansen (1994) [DH], and Henze and Zirkler (1990) [HZ] test statistics for the Pearson Type II distribution with $m = 10$. Based on 10,000 samples of sizes $n = 25$, 50, 75, 100 and 250, and $p = 2$, 3, 4, 5 and 10.

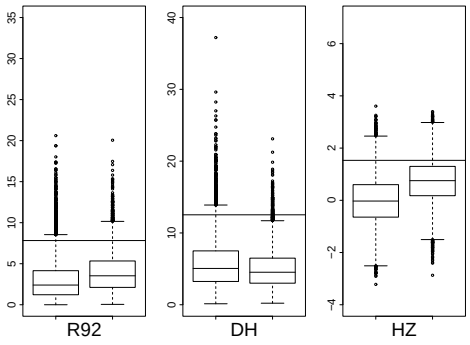


(a) $n = 25$ (b) $n = 50$

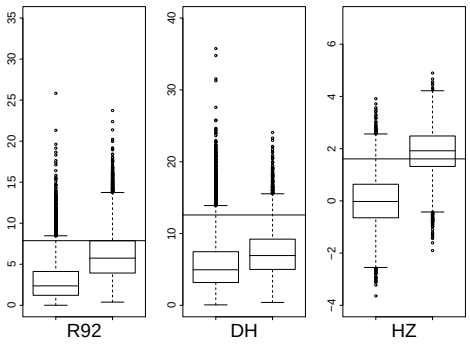
(c) $n = 75$ (d) $n = 100$

(e) $n = 250$

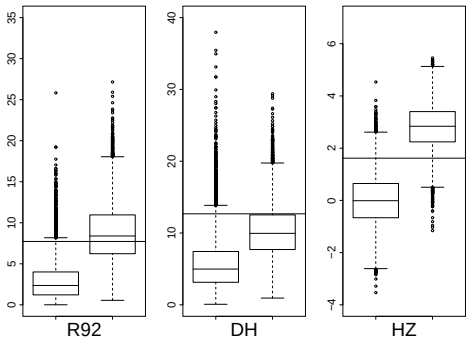
Figure 4. Empirical powers for the Royston (1992) [R92], Doornik and Hansen (1994) [DH], and Henze and Zirkler (1990) [HZ] test statistics for the Khintchine distribution. Based on 10,000 samples of sizes $n = 25$, 50, 75, 100 and 250, and $p = 2$, 3, 4, 5 and 10.



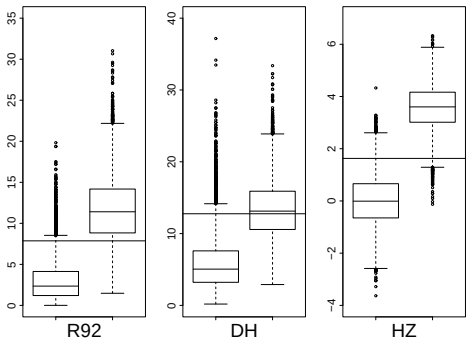
(a) $n = 25$



(b) $n = 50$



(c) $n = 75$



(d) $n = 100$

Figure 5. Empirical distributions of the Royston (1992) [R92], Doornik and Hansen (1994) [DH], and Henze and Zirkler (1990) [HZ] test statistics for the multivariate normal and multivariate uniform distributions (left and right boxplots respectively), based on 10,000 samples of sizes $n = 25$, 50, 75 and 100 and $p = 3$. The solid horizontal lines correspond to the 95% quantile of the distribution under multivariate normal data.