Bootstrap Tests for Multivariate Directional Alternatives

Abu T. M. Minhajuddin [a,*], William H. Frawley [a], William R. Schucany [b], Wayne A. Woodward [b].

[a] *Division of Biostatistics, Department of Clinical Science, UT Southwestern Medical Center, Dallas, Texas, USA.*

[b] *Department of Statistical Science, Southern Methodist University, Dallas, Texas, USA.*

Abstract

Tests on multivariate means that are hypothesized to be in a specified direction have received attention from both theoretical and applied points of view. One of the most common procedures used to test this cone alternative is the likelihood ratio test (LRT) assuming a multivariate normal model for the data. However, the resulting test for an ordered alternative is biased in that the only usable critical values are bounds on the null distribution. The present paper provides empirical evidence that bootstrapping the null distribution of the likelihood ratio statistic results in a bootstrap test (BT) with comparable power properties without the additional burden of assuming multivariate normality. Additionally, the tests based on the LRT statistic can reject the null hypothesis in favor of the alternative even though the true means are far from the alternative region. The bootstrap test also has similar properties for normal and non-normal data. This anomalous behavior is due to the formulation of the null hypothesis and a possible remedy is to reformulate the null to be the complement of the alternative hypothesis. We discuss properties of a bootstrap test for the modified set of hypotheses (MBT) based on a simulation study. The resulting test is conservative in general and in some specific cases has power estimates comparable to those for existing methods. The BT has higher sensitivity but relatively lower specificity whereas the MBT has higher specificity but relatively lower sensitivity.

Key words: cone, correlated data, likelihood ratio, mean, nonparametric, one-sided, ordered, positive orthant.

* Correspondence to: Abu Minhajuddin, Division of Biostatistics, Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, Texas, USA.

Email address: abu.minhajuddin@utsouthwestern.edu (Abu T. M. Minhajuddin).

1 Introduction

In many clinical trials, the treatment is expected to have a positive effect on multiple endpoints as compared to the control. Thus, practitioners are often interested in testing a one-sided hypothesis. One of the ways of doing this is to perform a series of tests comparing each endpoint to its control using univariate one-sided test procedures. However, use of multiple significance tests increases the chance of false positive findings. Also, by resorting to running univariate analysis, one is ignoring the information available in the correlation structure. To utilize this information one needs a multivariate technique. Classical multivariate tests for means, e. g. Hotelling's T^2, are nondirectional. However, several authors have considered the topic of directional multivariate tests (see Perlman 1969, Follmann 1996, Tang 1994, Wang and McDermott 1998, Perlman and Wu 2002a, Larocque and Labarre 2004, etc.). These multivariate procedures have both advantages and disadvantages.

To formalize the idea, let $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ be a random sample of p-dimensional observations from a population with unknown mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ and unknown $p \times p$ covariance matrix Σ. Consider testing the hypotheses

$$H_0 : \boldsymbol{\mu} = 0 \text{ vs. } H_1 : \boldsymbol{\mu} \geq \mathbf{0}, \tag{1}$$

where at least one of the μ_i's is positive in the alternative. We refer to $(\mathbf{y} : y_i \geq 0, i = 1, 2, \ldots, p)$ as the positive orthant. Assuming a multivariate normal distribution for the X_is with an unknown covariance matrix, Perlman (1969) developed the likelihood ratio test (LRT) statistic

$$U(\mathbf{x}, \mathbf{m}, \mathbf{A}, \mathcal{O}^+) = \|\mathbf{m}\|^2_{\mathbf{A}} \left(1 + \|\mathbf{m} - \mathbf{x}\|^2_{\mathbf{A}}\right)^{-1} \tag{2}$$

for a positive orthant alternative, where $\mathbf{x} = \sqrt{n}\bar{\mathbf{x}}$, $\mathbf{A} = (n-1)S$, S is the sample covariance matrix, and $\mathbf{m}$ is the point in the closed positive orthant $\mathcal{O}^+$ that is closest to $\mathbf{x}$ in terms of Mahalanobis distance

$$\|\mathbf{m} - \mathbf{x}\|^2_{\mathbf{A}} = (\mathbf{x} - \mathbf{m})' \mathbf{A}^{-1} (\mathbf{x} - \mathbf{m}). \tag{3}$$

However, since the null distribution of the statistic (2) depends on the unknown covariance matrix Σ, the only usable critical points are upper and lower bounds on the null distribution. Consequently, the resulting test is biased. The issue has been discussed widely in the literature. Perlman and Wu (1999) give a nice review of the various approaches for testing hypotheses (1), while arguing in favor of the LRT and the likelihood principle of testing hypotheses in general. However, the issue is far from resolved.

Wang and McDermott (1998) develop an unbiased and uniformly more powerful test for the positive orthant alternative by conditioning on the complete

sufficient statistic, $\mathbf{X}\mathbf{X}'$, the sample sum of squares and cross-products matrix. However, the conditional test requires numerical integration of the conditional likelihood and hence is computationally expensive and not suitable for large values of p. Perlman and Wu (2002a) propose a different conditional test by using the conditional distribution of the LRT statistic $U(\cdot)$ given K, the number of strictly positive components of $\mathbf{m}$. Simulation results reported by Perlman and Wu (2002a) indicate that the conditional test performs better than the corresponding unconditional tests when the probability content of the non-negative orthant $\mathcal{O}^+$ is relatively small, and the power advantage is substantial as p increases.

The common feature of the tests discussed in the literature is that all heavily depend on the multivariate normal model. Therefore, all of them can suffer loss of power when the normality assumption is not satisfied. In practice, normality can only be an assumption and sometimes data provide sufficient evidence against it. The omnibus test based on Hotelling's T^2 statistic is reasonably robust. However, it has relatively low power when used to test (1), because it detects any directional departure from the null mean value of zero. In recent years, various authors have proposed robust and nonparametric test procedures for the orthant-restricted mean vector as a specific directional alternative.

The null distribution of the rank test proposed in Park, Na, and Desu (2001) relies on the permutation principle and hence is only suitable for small n and p. For relatively large values of n and/or p they derive a multivariate normal approximation, that uses numerical integration to compute the tail probabilities of the multivariate normal. Thus the test is computationally expensive. The robust test proposed by Mudholkar, Kost, and Subbaiah (2001) is based on trimmed estimates of the mean and covariance matrix. The test loses power as the percent trimmed increases. The sign test developed by Larocque and Labarre (2004) is distribution-free conditional on the number of observations in the second and fourth quadrants for bivariate data. For multivariate data, the test is conditionally distribution-free given the number of observations in the positive and negative orthant given a specific data configuration defined with perpendicular hyperplanes formed by the data points. The authors report simulation results indicating that the conditional sign test is competitive for a range of alternatives for nonnormal skewed data.

The need for a nonparametric test for the orthant-restricted hypothesis arises for two reasons. First, the main difficulty of Perlman's LRT is the lack of the exact null critical values for unknown Σ. Second, the assumption of multivariate normality imposes an undesirable constraint on the scope of inference. A distribution-free test based on the likelihood-ratio principle, if available, could solve both of these problems. Schucany, Frawley, Gray, and Wang (1999) addressed both of these issues satisfactorily by bootstrapping the null distribu-

tion of the statistic (2), which is the likelihood ratio statistic when the sampling is from a multivariate normal distribution. In Section 2 we present empirical evidence that the nonparametric bootstrap provides a superior estimate of the sampling distribution of the LRT statistic under the null hypothesis and hence yields a test with competitive power properties. Also, the bootstrap test does not require any parametric family assumption and thus these inferences have fewer constraints. We refer to the test as a *bootstrap test* rather than the bootstrap LRT to emphasize the fact that the proposed test remains valid for normal as well as nonnormal models. The test statistic in question is the LRT statistic for the normal model, even though it is not the likelihood for a nonnormal model.

Section 2 presents a bootstrap test (BT) first reported in Schucany et. al. (1999). We present simulation results involving BT for multivariate normal as well as multivariate gamma data in Section 3. In Section 4 we provide empirical evidence of an anomaly in tests based on the statistic (2) for testing hypotheses (1). We then propose a new nonparametric bootstrap test for testing a modified set of hypotheses in Section 5 and summarize some power results involving this new bootstrap test in Section 6. In Section 7 we report a power comparison of the bootstrap tests with the Wang and McDermott (1998) test. An example is discussed in Section 8 along with some concluding remarks in Section 9.

2 A Bootstrap Test

It is not possible to analytically calculate the critical points for the test statistic in equation (2) for the positive orthant $\mathcal{O}^+$. Schucany et. al. (1999) proposed the following bootstrap algorithm for estimating the critical points of the statistic under the null hypothesis that the mean is zero:

Step 1. Compute $u = U(\mathbf{x}, \mathrm{m}, \mathbf{A}, \mathcal{O}^+)$ using the original set of observations.
Step 2. Compute the residuals $\{\mathbf{e}_i = \mathbf{x}_i - \bar{\mathbf{x}},\ i = 1, 2, \ldots, n\}$.
Step 3. Resample with replacement $\{\mathbf{e}_i^*, i = 1, 2, \ldots, n\}$ from the residuals.
Step 4. Compute the statistic $u^* = U(\mathbf{e}^*, \mathbf{m}^*, \mathbf{A}^*, \mathcal{O}^+)$, where the * denotes the analogous quantities for bootstrap resamples.
Step 5. Repeat steps 3 and 4 for a total of B times and count J = number of values for which $u^* > u$.
Step 6. Estimate the p-value $= (J + 1)/(B + 1)$.
Step 7. If the p-value is less than α, then reject H_0 in favor of the alternative that $\mu \in \mathcal{O}^+$.

Here the resampling is done under the null hypothesis of zero mean by using the residuals e_i, $i = 1, 2, \ldots, n$. See Davison and Hinkley (1997) for further

discussion and rationale for using the residuals. Also, $\mathbf{m}$ is the point in $\mathcal{O}^+$ that is closest to $\mathbf{x}$ in terms of Mahalanobis distance in equation (3). Clearly, if all elements of $\mathbf{x}$ are non-negative, then $\mathbf{m} = \mathbf{x}$. However, if some or all of the components of $\mathbf{x}$ are negative, then we need to find the point $\mathbf{m} \in \mathcal{O}^+$ that is closest to $\mathbf{x}$ in terms of Mahalanobis distance. This can be found by a standard quadratic programming algorithm.

3 Simulation Comparison of the Tests

To investigate the power properties of the bootstrap test, we ran a Monte Carlo experiment using a Sun Ultra workstation. The program was codeed in C with IMSL$^{\text{TM}}$ routines for random number generation. The software is available for download at http://www.smu.edu/statistics/TechReports/tech-rpts.asp or by email from the corresponding author. Two distinctly different multivariate families are reported here, namely 1) multivariate normal and 2) multivariate gamma with shape parameter $r = 1$ and scale parameter $s = 3$. The multivariate gamma (1, 3) is used because it is highly right-skewed. The exchangeable multivariate gamma random numbers are generated using the algorithm by Minhajuddin, Harris, and Schucany (2004) and transformed to have the correct hypothesized marginal expectations and marginal variances of unity. Following Perlman and Wu (2002a), three types of covariance matrices Σ are used, denoted by $I, Q,$ and R. Here I is the identity matrix; $Q = (q_{ij})$ with $q_{ii} = 1$ for $i = 1, 2, \ldots, p$ and $q_{ij} = 0.9$ for $i \neq j$ (positive); and $R = (r_{ij})$ with $r_{ii} = 1$ for each i, $r_{12} = r_{34} = \ldots = -0.9$, and $r_{ij} = 0$ for all other $i \neq j$ (negative). The nominal size of the test $\alpha = 0.05$, the number of bootstrap resamples $B = 299$, and the number of Monte Carlo iterations $N = 2000$ are fixed throughout.

3.1 Size

Table 1 shows the observed significance levels for the three tests for multivariate normal and multivariate gamma (1, 3). The effect of sample size n and dimension p on the size of the tests for $n = 22, 40, 62,$ and 99 and $p = 2$ and 6 are reported. The full simulation experiment independently included all combinations of $p = 2, 4,$ and 6 for these and $n = 32$ and 75, as well. See Minhajuddin (2003). The tabled values are the percent of times the observed test statistic resulted in the rejection of the null hypothesis. The nominal expected value (EV) is 5% with standard error (SE) less than or equal to $\sqrt{(.05)(.95)/2000} = 0.5\%$. For $p = 2$, the estimated significance levels

Table 1
Comparison of estimated size of the three tests for selected values of sample size n and dimension p for multivariate normal and multivariate gamma (1, 3) data with covariance matrices R, I, and Q, EV = 5, and SE = 0.5.

Multivariate Normal													
		Covariance Matrix											
		R				I				Q			
p	Test \ n	22	40	62	99	22	40	62	99	22	40	62	99
2	BT	4.3	4.6	4.3	4.3	4.2	5.1	3.5	4.8	4.3	5.3	4.4	4.5
	PW	*2.6*	*3.0*	*2.8*	*2.4*	3.3	3.6	*2.6*	4.8	5.0	5.2	4.3	4.8
	T^2	4.6	5.4	4.9	5.1	4.8	5.5	4.0	5.2	5.1	5.4	4.9	5.1
6	BT	*1.5*	4.9	5.3	5.2	*1.8*	3.3	3.8	4.8	*1.3*	3.7	3.9	5.1
	PW	3.3	4.8	4.8	4.4	4.6	4.3	4.9	5.3	4.1	4.9	4.4	4.8
	T^2	4.9	5.1	5.1	5.0	4.0	5.0	4.8	5.9	4.5	5.3	5.8	5.1
Multivariate Gamma (1, 3)													
2	BT	6.4	5.1	5.1	4.0	*2.9*	3.9	3.8	4.0	4.0	4.4	4.6	3.7
	PW	5.6	4.3	4.0	*2.9*	*0.8*	*1.3*	*1.4*	4.0	4.9	4.5	4.3	3.4
	T^2	**11.1**	**8.9**	**7.4**	5.9	**10.0**	**7.6**	**7.7**	6.2	**9.5**	**7.9**	**7.6**	6.1
6	BT	*2.4*	3.8	4.4	4.0	*0.9*	*2.5*	*3.1*	3.5	*1.3*	*3.0*	4.4	4.4
	PW	**7.9**	6.6	6.7	5.5	*2.8*	*2.1*	*1.7*	*1.8*	**10.2**	**8.7**	**7.1**	6.4
	T^2	**13.9**	**10.4**	**8.9**	6.7	**13.9**	**11.5**	**8.5**	**8.4**	**15.7**	**13.5**	**11.3**	**8.8**

Note: BT: Bootstrap test, PW: Perlman and Wu's Conditional test, and T^2: Hotelling's T^2 test. Estimated sizes that are significantly higher than the nominal 5% are listed in **boldface** and those that are significantly lower are listed in *italic* fonts (controlled for 36 multiple comparisons).

for the bootstrap test are not significantly different from the nominal level of 5% even for a small sample of size 22. However, for nonnormal data, the bootstrap test requires a larger sample size to maintain the nominal level. See Polansky (1999) for a proof of the sample size limitation of bootstrap confidence intervals, which is applicable in the context of bootstrap testing as well. As p increases, the sample size required to maintain the nominal level also increases. For example, for $p = 6$, the estimated levels for the bootstrap test are significantly lower than the nominal 5% for $n = 22$, and 40, when the data are nonnormal (skewed to the right). The omnibus test based on Hotelling's T^2 statistic performs poorly for nonnormal data. The estimated levels are significantly higher than 5% for the T^2 test. Perlman and Wu's conditional test maintains the nominal size for multivariate normal data. However, it suffers severely from this departure from the multivariate normal model. The estimated levels are significantly higher than 5% when $\Sigma = Q$ (positive) and R (negative) at $p = 4$. These are more pronounced at $p = 6$.

3.2 Power

For $p = 2$ three different alternatives directions are considered: $\boldsymbol{\mu}_i = \lambda e_i,\ i = 1, 2, 3$, where $e_1 = (1, 0)^{'}$, $e_2 = (1, 2)^{'}$, $e_3 = (1, 1)^{'}$ with λ ranging from -0.5 to 0.5. The estimated power is the proportion of 2000 iterations in which the test statistic is in the critical region (SE $\leq \sqrt{(0.5)(0.5)/2000} = 0.011$). The sample size for these power curves is $n = 32$. McNemar's test (Lehmann, 1998) is used to assess the difference in powers for the bootstrap test and Perlman and Wu's conditional test. These paired comparisons are made for each value of the alternative means. Figure 1 displays estimated power curves of the three tests, BT, PW, and T^2 for bivariate normals. The first row of graphs are for covariance matrix Q, the second row for I, and the last row for R. The three columns are for e_1, e_2, and e_3, respectively. The plus signs indicate significant differences ($p < 0.05$) between the power of BT and that of PW for the specific alternative mean.

For normal data the bootstrap test is significantly more powerful than the conditional test PW at many of the points in the alternative region ($\lambda > 0$), especially when the data are independent. Moreover, both of the directional tests are clearly better than the omnibus T^2. Similar graphs for the bivariate gammas are in Figure 2. For these nonnormal data, significant improvements in power have been achieved by the BT test over the conditional test PW. The improvement in power is higher when the data are independent. For positively correlated data (Q), the BT test performs better if the alternative mean is along the diagonal. Both of the normal theory tests, PW and T^2, suffer in terms of power for these skewed nonnormal data.

In four dimensions the alternatives in simulations are $\boldsymbol{\mu}_i = \lambda e_i,\ i = 1, 2, 3$, where $e_1 = (1, 0, 0, 0),\ e_2 = (1, 1, 0, 0)$, and $e_3 = (1, 1, 1, 0)$ using the same three covariance structures $Q, I,$ and R. The results for the multivariate normal are given in Figure 3. For the independent multivariate normal, the bootstrap test has significantly higher power compared to the conditional test as depicted by the plus signs. For positive correlations (Q), the bootstrap test and the conditional test have comparable power. However, the conditional test performs better than the bootstrap test if the components are negatively correlated (R). Similar power curves for the nonnormal, skewed data are given in Figure 4 with the same covariance structures organized as before. Significant improvement of power is achieved for BT over PW when the data are independent (I). The conditional test has better power than the bootstrap test for positively correlated (Q) data and comparable power for covariance matrix (R). However, in these cases PW has inflated Type I error rates for $p = 4$ (see Table 1).

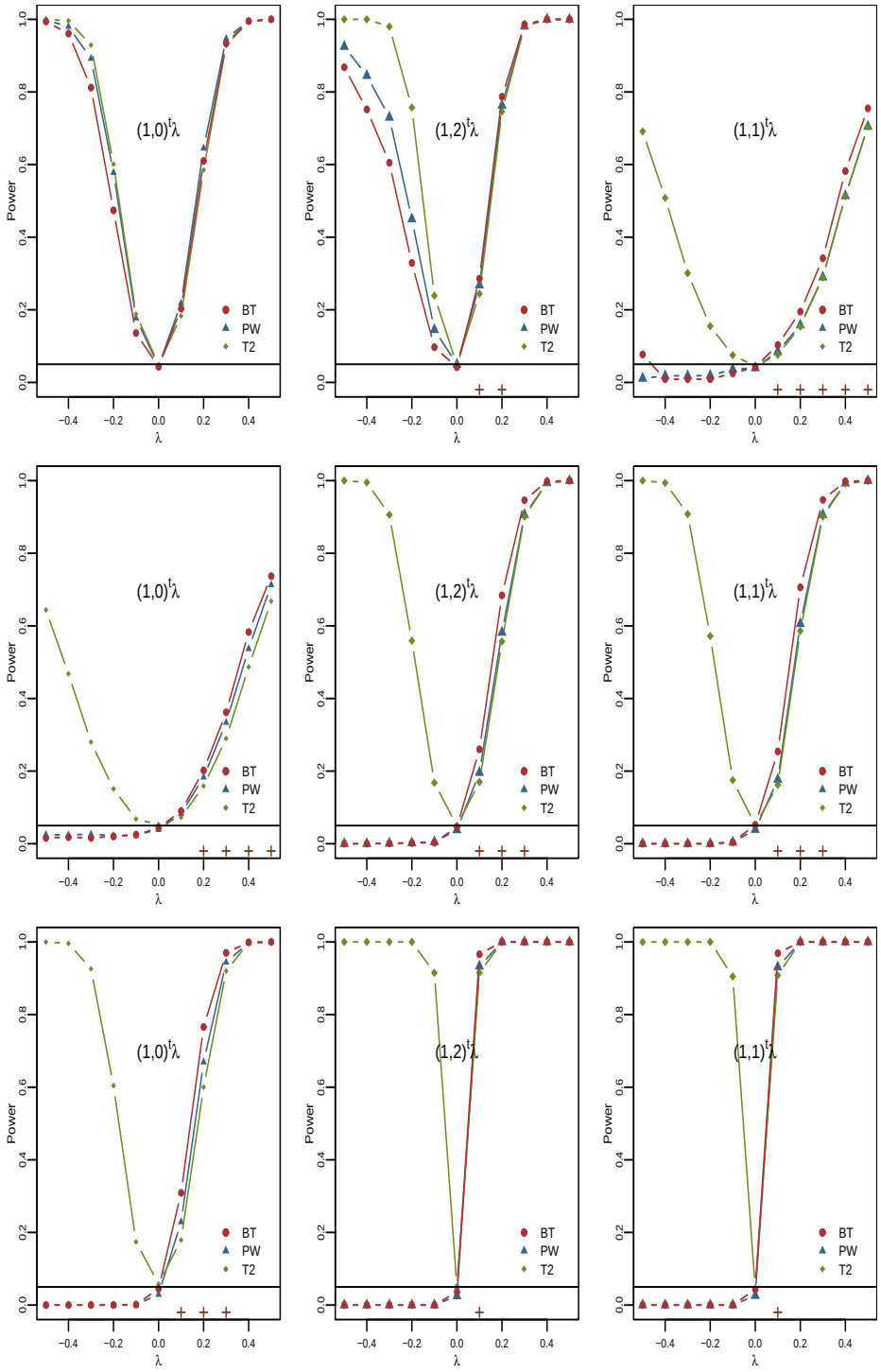

Fig. 1. Estimated power curves of three tests with bivariate normal data: The first row is for covariance matrix Q, the middle row for covariance matrix I, and the last row for covariance matrix R. The + signs indicate significant differences at the 5% level between the powers of the BT test and PW test.

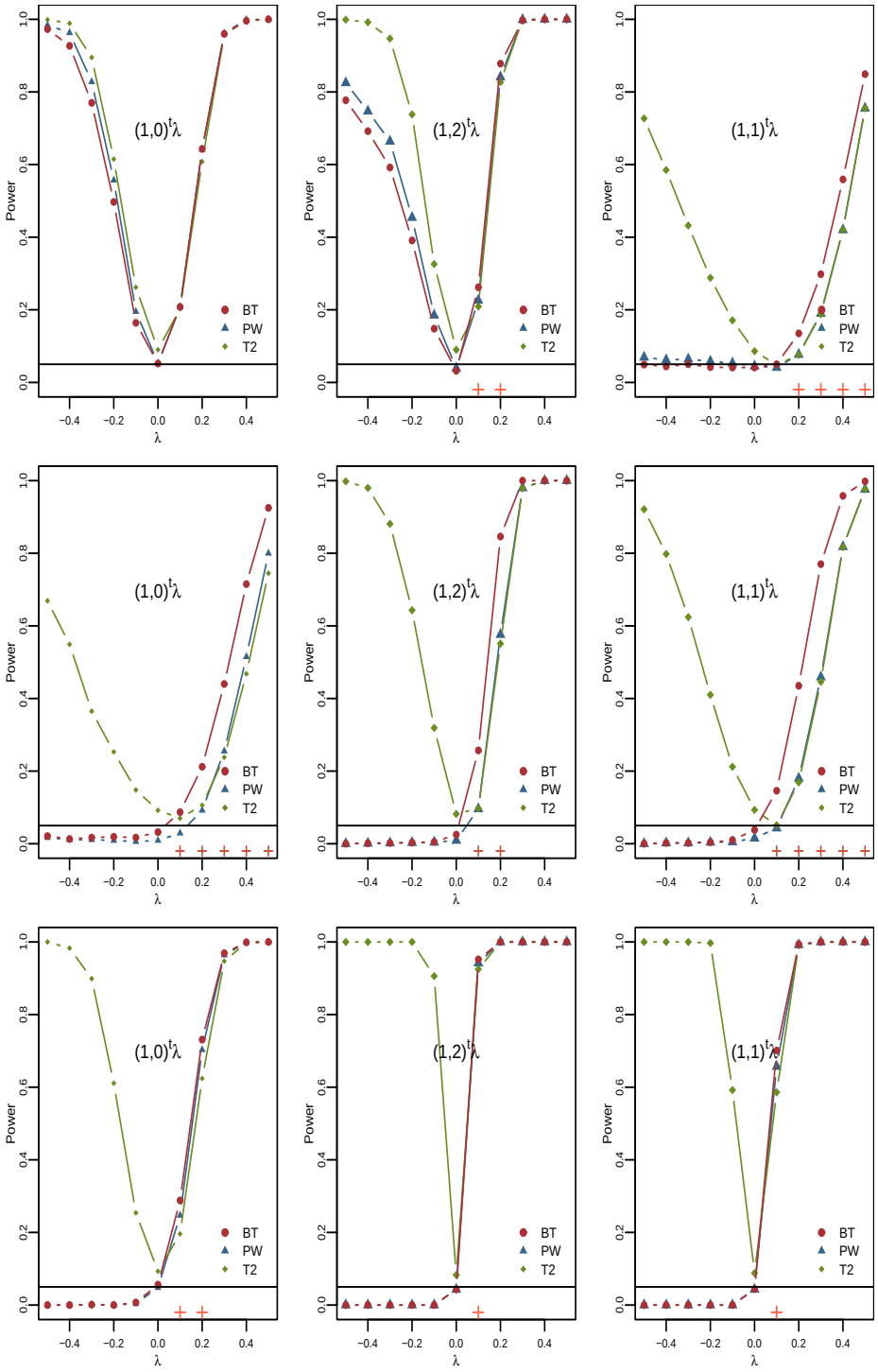

Fig. 2. Estimated power curves of three tests with bivariate gamma (1, 3) data: The first row is for covariance matrix Q, the middle row for covariance matrix I, and the last row for covariance matrix R. The + signs indicate significant differences at the 5% level between the powers of the BT test and PW test.

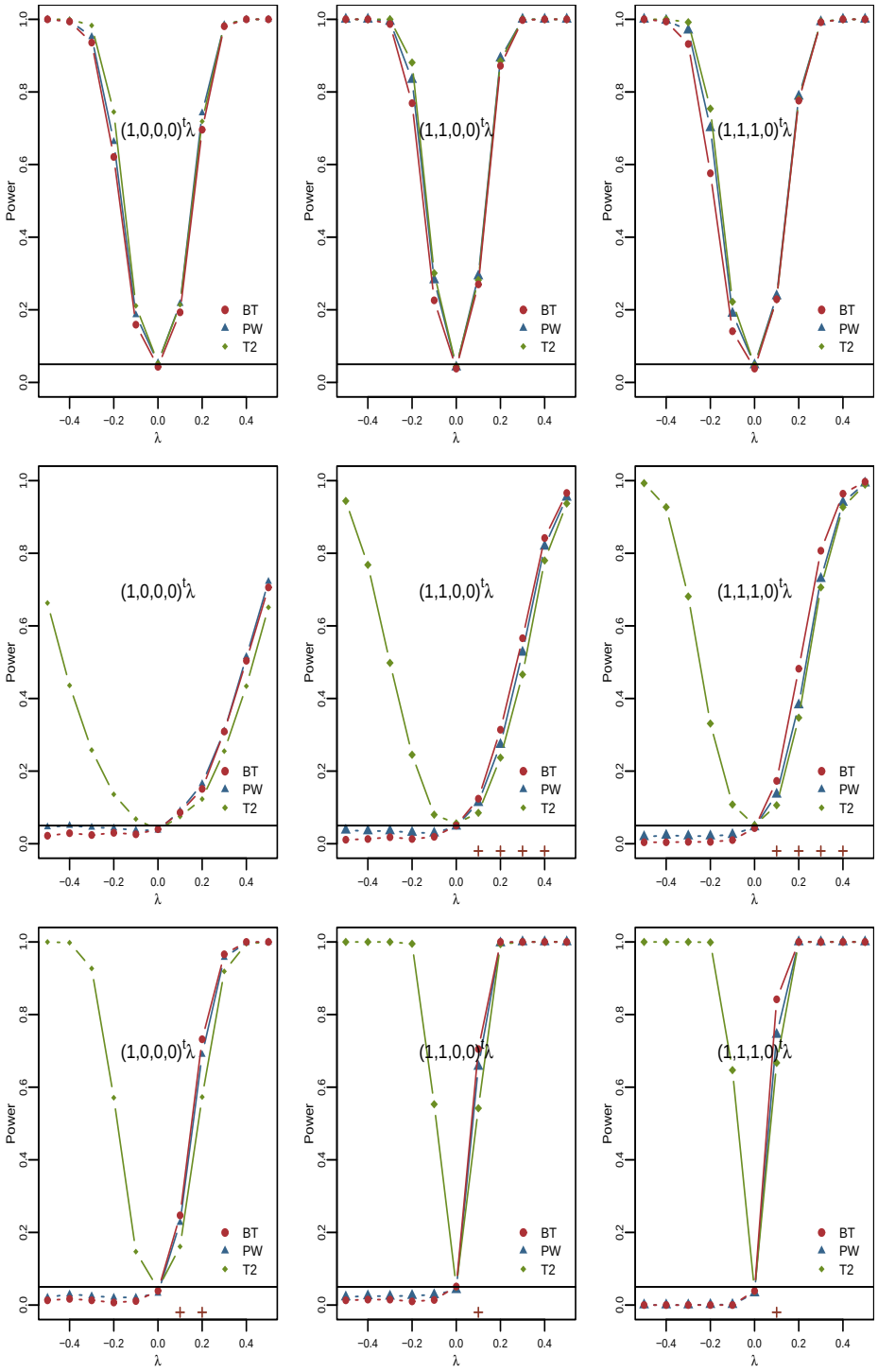

Fig. 3. Estimated power curves of three tests with 4-variate normal data: The first row is for covariance matrix Q, the middle row for covariance matrix I, and the last row for covariance matrix R. The + signs indicate significant differences at the 5% level between the powers of the BT test and PW test.

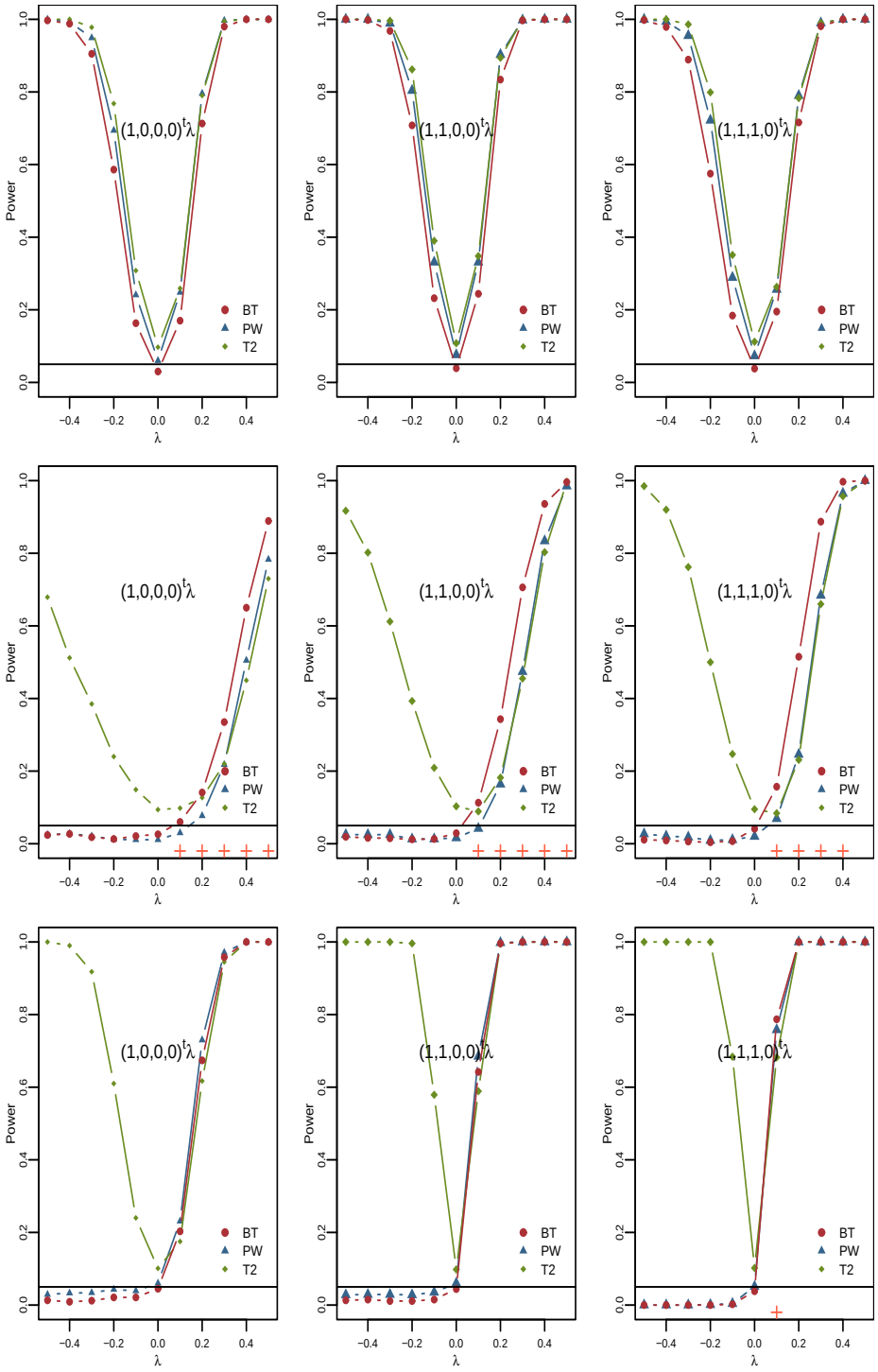

Fig. 4. Estimated power curves of three tests with 4-variate gamma (1, 3) data: The first row is for covariance matrix Q, the middle row for covariance matrix I, and the last row for covariance matrix R. The + signs indicate significant differences at the 5% level between the powers of the BT test and PW test.

4 An Anomaly of the Test Based on the Statistic $U(.)$

In Figures 1 to 4, an anomalous behavior of the directional test based on the statistic $U(\cdot)$ is evident for $(\lambda < 0)$. For positively correlated (Q) bivariate data, BT and PW tend to reject the null hypothesis of zero mean in favor of a positive orthant alternative even if the true mean lies "deep" inside the null region where both components of the mean are negative. This is true regardless of whether or not the data are multivariate normal. Silvapulle (1997) discussed this apparent "anomalous" behavior using a simple example in two dimensions. Figure 1 of Silvapulle (1997) clearly illustrates that the rejection region of the LRT contains areas of the sample space where the LRT rejects the null hypothesis in favor of the positive quadrant, even though one's intuition does not agree with the conclusion of the test.

At this point, we want to be clear that we are not concluding that the likelihood criterion is an invalid significance test procedure. Rather, we agree with Perlman and Wu (1999) that "the LR criterion, ... is a readily understood and generally useful tool for statistical inference". However, there are instances when a test criterion violates common sense. In such a situation one should not blindly reject the criterion. Rather, a careful review of the test components is required. We argue that this is one such case. Here the simple null hypothesis of zero mean vector and the positive orthant alternative do not comprise the entire parameter space. This is shown in the left panel of Figure 5 for $p = 2$ with positively correlated data. The point $\mathbf{x} = (-0.4, 0)'$ is closer to the alternative region than it is to the null mean value of $(0, 0)'$ in terms of Mahalanobis distance. Since the LRT statistic considers plausibility of the null hypothesis in terms of Mahalanobis distance, the null hypothesis would always be rejected for this observed $\mathbf{x}$, even though it is "far" from the alternative region. In fact, this type of outcome is possible for "virtually any test that uses the general ideas of likelihood ratio tests" (Silvapulle, 1997). Perlman and Wu (2002b) recognize this anomalous behavior of the LRT and suggest a remedy by reformulating the null hypothesis to include the complement of the alternative region as the null region. Hypothesis (1) then becomes

$$H_0 : \boldsymbol{\mu} \notin \mathcal{O}^+ \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \in \mathcal{O}^+. \tag{4}$$

For the multivariate normal distribution the likelihood ratio statistic for testing hypothesis (4) is then

$$D(\mathbf{x}, \boldsymbol{\pi}_\mathbf{x}, \mathbf{A}, \mathcal{O}^+) = \|\boldsymbol{\pi}_\mathbf{x} - \mathbf{x}\|^2_\mathbf{A}, \tag{5}$$

where $\boldsymbol{\pi}_\mathbf{x} = (\pi_1, \pi_2, \ldots, \pi_p)'$ is the point in the complement of $\mathcal{O}^+$, denoted by $\mathcal{O}^\circ$, that is closest to $\mathbf{x}$. This is illustrated in the right panel of Figure 5 for the positive orthant alternative. Clearly, if $\mathbf{x} \notin \mathcal{O}^+$ then $\boldsymbol{\pi}_\mathbf{x} = \mathbf{x}$ and $D(\cdot) = 0$, while $D(\cdot) > 0$ for $\mathbf{x} > \mathbf{0}$. For $\mathbf{x} \in \mathcal{O}^+$, the statistic $D(\mathbf{x}, \boldsymbol{\pi}_\mathbf{x}, \mathbf{A}, \mathcal{O}^+)$ is the

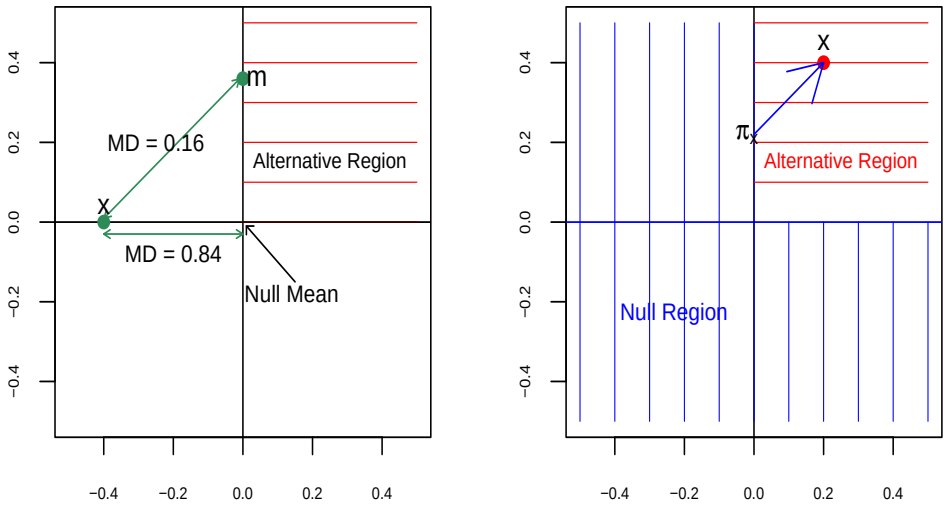

Fig. 5. The null and the alternative regions with projections $\mathbf{m}$ and $\boldsymbol{\pi}_{\mathbf{x}}$ of a point $\mathbf{x}$, (MD = Mahalanobis Distance).

Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\pi}_{\mathbf{x}}$.

5 A New Bootstrap Test

For the multivariate normal model, the critical values for the statistic $D(\cdot)$ can be found using χ^2_1 critical values (Perlman and Wu, 2002b). However, our interest is in a nonparametric bootstrap test that does not require the assumption of multivariate normality. Additionally, we want to develop a test procedure that is free of the anomaly discussed in the previous section. We claim that a test that uses the distance $D(\mathbf{x}, \pi)$ as the test statistic would provide such a test. Since the distribution of the test statistic $D(\cdot)$ is not readily known and also since we want to avoid making distributional assumptions, we propose the following modified bootstrap (MBT) algorithm for a size-α test for (4).

Step 1. Compute $d = D(\mathbf{x}, \boldsymbol{\pi}_{\mathbf{x}}, \mathbf{A}, \mathcal{O}^+)$ using the original set of observations.

Step 2. If $\bar{\mathbf{x}} \in \mathcal{O}^+$, shift the location of the sample by setting $\mathbf{y}_i = \mathbf{x}_i - \bar{\mathbf{x}} + \boldsymbol{\pi}_{\bar{\mathbf{x}}}$, $i = 1, 2, \ldots, n$. Otherwise, set $\mathbf{y}_i = \mathbf{x}_i$ for all i.

Step 3. Resample with replacement $\{\mathbf{y}^*_i, i = 1, 2, \ldots, n\}$ from $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$.

Step 4. Compute the statistic $d^* = D(\mathbf{y}^*, \boldsymbol{\pi}^*, \mathbf{A}^*, \mathcal{O}^+)$, where the * denotes the analogous quantities for bootstrap resamples.

Step 5. Repeat steps 3 and 4 a total of B times and count J = number of values for which $d^* > d$.

Step 6. Estimate the MBT p-value $= (J + 1)/(B + 1)$.

Step 7. If the p-value is less than α, then reject H_0 in favor of the alternative that $\mu \in \mathcal{O}^+$.

Notice that in the above algorithm, resampling is done under the null hypothesis. This is achieved by shifting the location of the sample whenever $\mathbf{x} > \mathbf{0}$ so that the sample mean is on the boundary of the null space. This results in a zero value of the statistic $D(\cdot)$. Also, since the alternative region consists of

$\boldsymbol{\mu} > \mathbf{0}$, the axes of the parameter space where at least one of the components is zero is a part of the null region. The test is expected to maintain its size when the mean vector lies on or inside the boundary of the null space.

6 Some Additional Simulation Results: Effect of Changing Hypotheses

A Monte Carlo experiment using the same design parameters as in Section 3 was conducted to estimate and compare the size and power of the modified bootstrap test (MBT) with those of BT. This assesses the effect of modifying the set of hypotheses. The worst case scenario of the mean vector on the boundary between the parameter spaces was used to examine the size. The estimated sizes of MBT for the multivariate normal data ($p = 2$ and 4) are graphed in Figure 6. Clearly, the estimated sizes of MBT are much lower than the nominal value of 0.05 in these small samples. Specifically, for covariance matrices I and R, the estimated size of MBT is substantially lower than 5%. Therefore, even though the test is unbiased, some loss of power will occur because of the conservative nature of the test. The conservatism arises from the dramatic increase in the size of the null parameter space. To investigate the power of the MBT with $D(\cdot)$ as the test statistic, multivariate normal and multivariate gamma (1, 3) data with covariance matrices R, I and Q were generated as in Section 3. For $p = 2$, alternatives of the form $\lambda \mathbf{e}_i,\ i = 1, 2$, where $\mathbf{e}_1 = (1, 2)'$ and $\mathbf{e}_2 = (1, 1)'$ are considered. The Monte Carlo power estimates are obtained for $n = 32$. As before, $B = 299$ and $N = 2000$ Monte Carlo replications, which implies that $SE \leq 0.011$.

Figure 7 plots these curves for the multivariate gamma (1, 3). It is evident from Figure 7 that the MBT is somewhat more powerful than BT, for positively correlated (Q) data when the alternative mean falls along the principal diagonal of the positive orthant. This is true for the bivariate normal data as well. However, if the alternative mean is not along the diagonal of the positive orthant, MBT often has substantially lower power. As expected, MBT is free of the anomalous behavior exhibited by the likelihood ratio based tests. It should be noted that the null hypothesis being tested by the MBT is different from that in case of BT test. For MBT $H_0 :\ \boldsymbol{\mu} \leq \mathbf{0}$, while for the BT test, $H_0 :\ \boldsymbol{\mu} = \mathbf{0}$. Thus, these simulation results show the effect of modifying the set of null hypotheses being tested. Changing the hypotheses yielded a rather conservative test. Similar results are observed for $p = 4$. The behavior of the bootstrap tests at $p = 6$ and for other sample sizes are reported in Minhajuddin (2003).

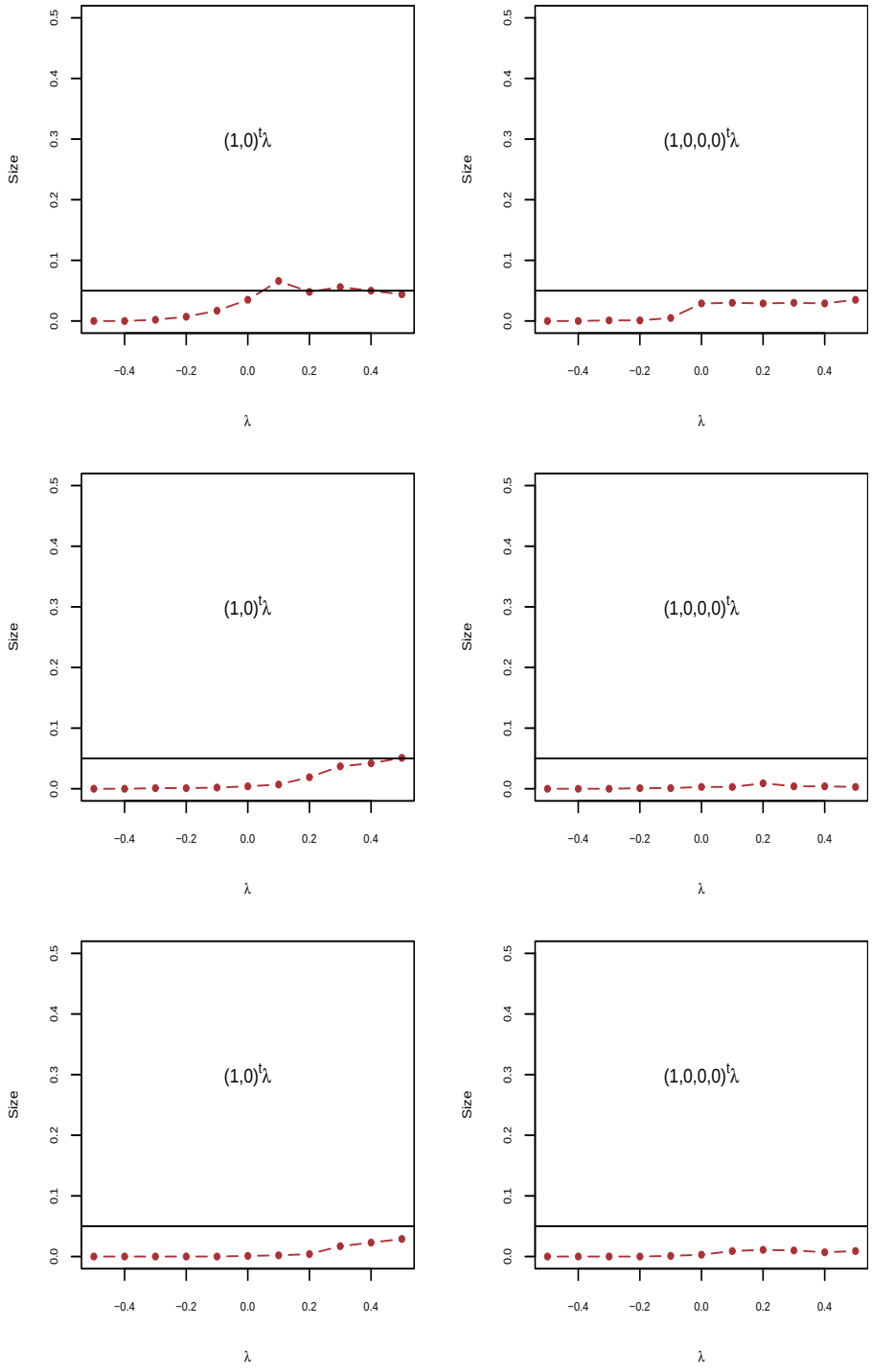

Fig. 6. Estimated size of MBT on two boundaries: Those in the top row are for covariance matrix Q, middle row for I, and the bottom row for R with $n = 22$ for $p = 2$ on the left and $n = 40$ for $p = 4$ on the right. In all cases, $B = 299$, $N = 2000$ so that SE ≤ 0.005.

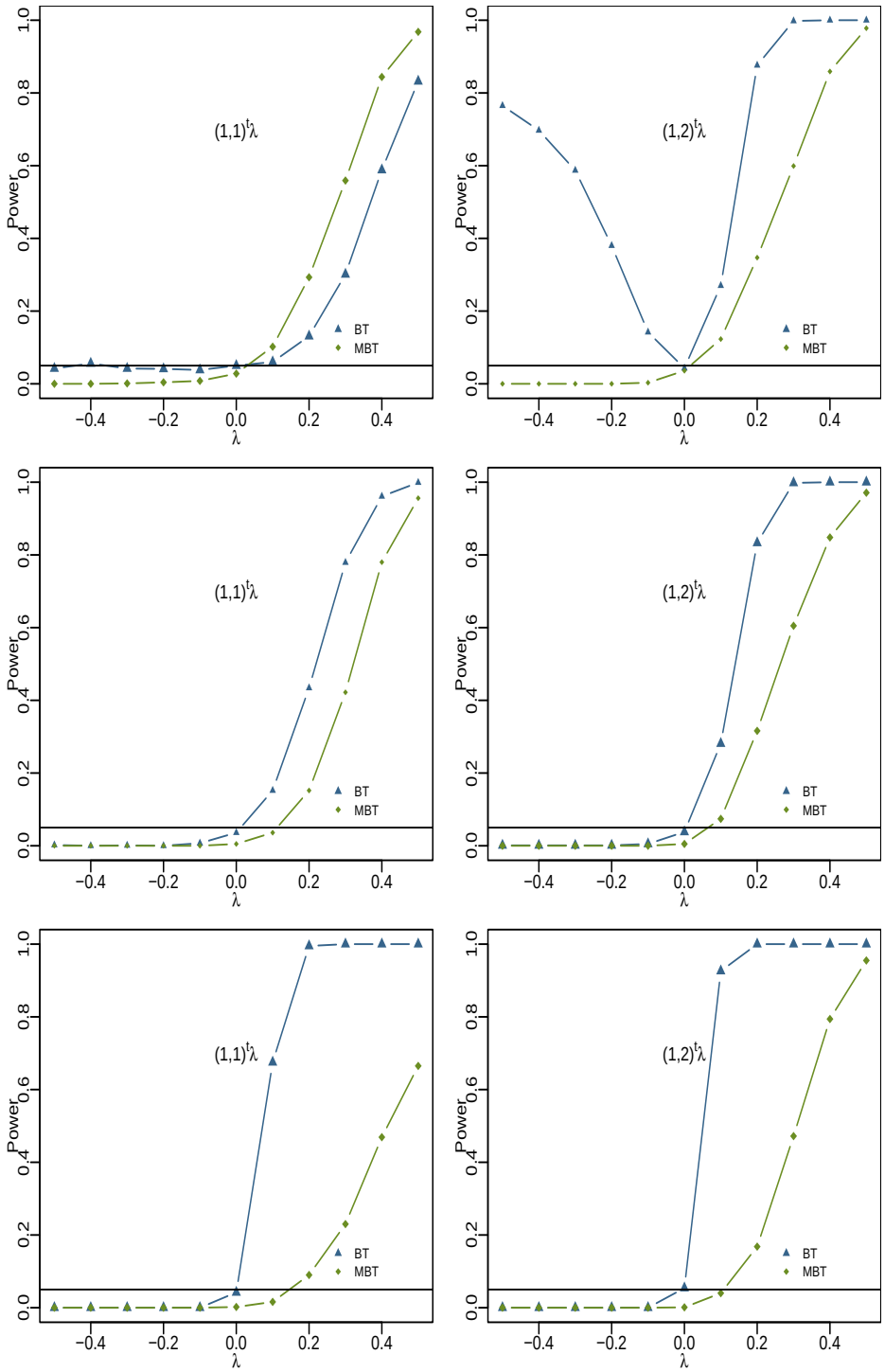

Fig. 7. Estimated power curves of BT and MBT tests with bivariate gamma (1, 3) data: The first row is for covariance matrix Q, the middle row for I, and the last row for R.

7 Power Comparison with Wang and McDermott's Test

Wang and McDermott (1998) developed a conditional test (WM) for the positive orthant alternative with good properties. Thus we compare the power of the bootstrap test (BT) and the modified bootstrap test (MBT) with that of the WM test. One of the difficulties of the WM test is its computationally expensive nature, requiring numerical integration of the conditional distribution. For comparison purposes, we repeat the power estimates for the WM test reported in Wang and McDermott (1998). They consider $n = 17$ for bivariate normal data with different values of the correlation coefficient. These estimates are obtained from $N = 10000$ Monte Carlo samples with the alternative means (0.5, 0.0) and (0.5, 0.5). An identical design is used here to obtain the power estimates for the four other tests. The results are summarized in Table 2. For these bivariate normal models the WM test is the most powerful among the tests under study for the alternative means considered. Recall that for the MBT, the point (0.5, 0.0) is on the boundary of the null region and MBT should maintain the nominal size of 5%. Indeed MBT maintains its size for the independent, as well as the positively correlated data. The bootstrap test BT has power estimates comparable to that of the WM test for all of the different correlation structures considered. The modified bootstrap test MBT has comparable power only when the alternative mean is on the main diagonal and there is positive correlation.

Table 2
Estimated powers (%) of the tests for bivariate normal data with different covariance matrices, $n = 17$, $B = 299$, $N = 10000$, SE $\leq 0.5\%$.

Mean $\boldsymbol{\mu}$	(0.5, 0.0)			(0.5, 0.5)		
Correlation	-0.75	0	0.75	-0.75	0	0.75
BT	82	44	72	100	74	47
PW	76	42	76	100	65	42
T^2	71	37	71	100	65	41
MBT	17	6	5	64	46	49
WM*	86	49	78	100	79	53

* from Wang and McDermott (1998)

8 An Example

Reaven and Miller (1979) present data on chemical and overt nonketotic diabetes in 145 non-obese adult subjects. Each subject is classified using existing medical criteria into one of the three groups namely, overt diabetic, chemical diabetic, and normal. There are 76 normal subjects in the sample along with

33 and 36 subjects from the overt and chemical diabetic groups, respectively. The chemical and overt diabetic subjects are expected to have higher levels of glucose intolerance and steady state plasma glucose (SSPG). On the other hand, the normal subjects are expected to have a higher level of insulin area representing better response to oral glucose. It is of interest to test these expectations using the data for the $n = 33$ overt diabetic subjects. In the absence of any gold standard mean levels for these $p = 3$ variables for the population of normal subjects, the respective sample means of "normal" subjects are used as the standard.

The hypothesized normal mean levels of glucose intolerance, SSPG, and insulin area are 350, 114, and 173, respectively. The corresponding sample means for the 33 overt diabetic subjects are 694, 205, and 67, respectively. The estimated p-value of the bootstrap test (B = 299) for testing the directional trivariate hypotheses is 0.003 and that of the modified bootstrap test is also $p = 0.003$. The p-values for the conditional test by Perlman and Wu (2002a) and the Hotelling's T^2 tests are both < 0.001. Therefore, these data provide enough evidence to conclude that indeed the mean levels for the glucose intolerance and SSPG for the diabetic subjects are higher than those for the nondiabetic subjects whereas the mean level for the insulin area for the diabetic subjects is lower than that for the nondiabetic subjects.

9 Concluding Remarks

In the present article two nonparametric bootstrap procedures for testing against a one-sided alternative in the multivariate setting have been examined. This issue of testing means in certain specified directions has received attention from both theoretical and applied points of views. However, most of the research is based on the multivariate normal distribution. The beauty of the bootstrap tests is that they are free of this additional assumption of multivariate normality and require a more relaxed assumption of existence of the first two moments. However, more robust measures of the location and scale could be incorporated in the bootstrap test procedures along with a more general distance metric instead of sample mean and sample covariance matrix and the traditional Mahalanobis Distance. Our simulation results show that for most cases considered, BT has competitive power compared to the normal theory tests. MBT also has competitive power for some alternative means considered. In some cases the bootstrap tests have significantly more power then their normal theory counterpart, especially for nonnormal data.

For positively correlated data (Q), the test based on the statistic $U(\cdot)$ has a tendency to reject the null hypothesis in favor of a positive orthant alternative even when the alternative mean is far from the rejection region. The problem

is caused by the hypotheses tested, not the likelihood principle. The MBT is free of this anomalous behavior which may lead to spurious rejection of the null hypothesis. However, the MBT is found to be quite conservative. It lacks power if the alternative mean is much removed from the diagonal of the alternative region and also for certain covariance structures. It should be noted here that, to our knowledge, MBT is the first workable test for the modified set of hypotheses that has been implemented in simulations. The choice between MBT and BT depends on the sensitivity and specificity required in a particular testing situation. If for a particular problem, it is important to protect against false rejections of the null hypothesis in favor of positive orthant alternative, then MBT should be used. However, as it was shown via simulation results, this may cost the investigator in terms of power. If the investigator is willing to accept a certain proportion of false rejections of the null hypothesis, then he/she should use the BT test, which is more powerful. In short, the BT has higher sensitivity but relatively lower specificity whereas the MBT has higher specificity but relatively lower sensitivity.

Based on these comparisons, it can be concluded that none of the tests considered is a uniformly most powerful test for the positive orthant hypothesis. Rather, the test criterion depends on the unknown covariance matrix and the hypotheses tested. It is interesting to note that the most powerful normal theory tests are conditional tests, even though the given condition is different. The WM test proposed by Wang and McDermott (1998) conditions on the sample matrix of sums and crossproducts, while the conditional test by Perlman and Wu (2002a) conditions on the number of strictly positive components of the MLE of the mean vector under the null hypothesis. The bootstrap tests perform reasonably well compared to their normal theory competitors. The normal theory tests break down by not having the proper size and can have lower power in some cases if the data are from a skewed distribution. On the other hand, the bootstrap test has the desired level for both normal and nonnormal joint distributions.

References

[1] Davison, AC. and Hinkley, DV. *Bootstrap Methods and Their Applications*, Cambridge University Press: Cambridge, 1997.

[2] Follmann, D. A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association* 1996; **91**: 854–861.

[3] Larocque, D, and Labarre, M. A conditionally distribution-free multivariate sign test for one-sided alternatives. *Journal of American Statistical Association* 2004; **99**: 499–509.

[4] Lehmann, EL. *Nonparametrics: Statistical Methods Based on Ranks.* Prentice Hall: New Jersey, 1998.

[5] Minhajuddin, ATM. Bootstrap Test for Multivariate Directional Alternatives.; Ph. D. dissertation 2003; Department of Statistical Sciences, Southern Methodist University, Dallas, Texas.

[6] Minhajuddin, ATM, Harris, IR, and Schucany WR. Simulating multivariate distributions with specific correlations. *Journal of Statistical Computing and Simulation* 2004; **74**: 599–607.

[7] Mudholkar, GS, Kost, J, and Subbaiah, P. Robust tests for the significance of orthant restricted mean vector. *Communications in Statistics: Theory and Methods* 2001; **30**: 1789–1810.

[8] Park, H, Na, JH, and Desu, MM. Nonparametric one-sided tests for multivariate data. *Sankhya* 2001; **63**: 286–297.

[9] Perlman, MD. One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics* 1969; **40**: 549–567.

[10] Perlman, MD, and Wu, L. The emperor's new tests. *Statistical Science* 1999; **14**: 355–381.

[11] Perlman, MD, and Wu, L. A class of conditional tests for a multivariate one-sided alternative. *Journal of Statistical Planning and Inference* 2002a; **107**: 155–171.

[12] Perlman, MD, and Wu, L. A defense of the likelihood ratio test for one-sided and order restricted alternatives. *Journal of Statistical Planning and Inference* 2002b; **107**: 173–186.

[13] Polansky, A. Upper bounds on the true coverage of bootstrap percentile type confidence intervals. *The American Statistician* 1999; **53**: 362–369.

[14] Reaven, GM, and Miller, RG. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia* 1979; **16**: 17–24.

[15] Schucany, WR, Frawley, WH, Gray, HL, and Wang, S. Bootstrap testing for ordered multivariate means. *Technical Report* 1999; Southern Methodist University, Department of Statistical Science.

[16] Silvapulle, MJ. A curious example involving the likelihood ratio test against one-sided hypotheses. *The American Statistician* 1997; **51**: 178–180.

[17] Tang, D. Uniformly more powerful tests in a one-sided multivariate problem. *Journal of the American Statistical Association* 1994; **89**: 1006–1011.

[18] Wang, Y, and McDermott, MP. Conditional likelihood ratio test for a nonnegative normal mean vector. *Journal of the American Statistical Association* 1998; **93**: 380–386.