

# Some Improved Tests for Multivariate One-Sided Hypotheses\*

Michael D. Perlman<sup>1</sup> and Lang Wu<sup>2</sup>

<sup>1</sup>*Department of Statistics, University of Washington, Seattle, WA 98195, USA*

*Email: michael@ms.washington.edu*

and

<sup>2</sup>*Department of Statistics, University of British Columbia, Vancouver, B.C., V6T 1Z2, Canada*

*Email: lang@stat.ubc.ca*

July, 2004

## Abstract

Multivariate one-sided hypothesis-testing problems are very common in clinical trials with multiple endpoints. The likelihood ratio test (LRT) and union-intersection test (UIT) are widely used for testing such problems. It is argued that, for many important multivariate one-sided testing problems, the LRT and UIT fail to adapt to the presence of subregions of varying dimensionalities on the boundary of the null parameter space and thus give undesirable results. Several improved tests are proposed that do adapt to the varying dimensionalities and hence reflect the evidence provided by the data more accurately than the LRT and UIT. Moreover, the proposed tests are often less biased and more powerful than the LRT and UIT.

KEY WORDS: One-sided hypothesis, multiple endpoints, likelihood ratio test, union-intersection test,  $p$ -value.

\*Research supported in part by U. S. National Science Foundation Grant No. DMS00-71818 and Canada Natural Sciences and Engineering Research Council grant No. 22R80742.

# 1 Introduction

One-sided tests for comparing multivariate treatment effects have received much attention in the literature (e.g., O'Brien (1984), Bloch *et al.* (2001), Tamhane and Logan (2004)). These tests are useful in clinical trials with multiple endpoints. For example, in clinical trials, treatment effects are often measured by both efficacy and toxicity, which may be measured by more than one response variable. In principle, a treatment is usually deemed better than its competitor if all components of its mean responses are larger (say). In some practical situations, it may be difficult to show that each component is better – instead the treatment will be preferred if at least one of its response components is greater than that of the competitor and if none of the remaining components are significantly worse (Bloch *et al.* 2001).

Suppose that there are two treatment groups, with  $n_i$  subjects in group  $i$ ,  $i = 1, 2$ . Let  $\bar{Y}_{ij}$  be the sample mean for the  $j$ th response to treatment  $i$ , and let  $\eta_{ij}$  be the population mean of the  $j$ th response to treatment  $i$ ,  $j = 1, \dots, p$ . Without loss of generality, we assume that larger values of  $\eta_{ij}$  correspond to better responses. Let  $X_j = (n_1^{-1} + n_2^{-1})^{-1/2}(\bar{Y}_{1j} - \bar{Y}_{2j})$ ,  $\mu_j = \eta_{1j} - \eta_{2j}$ ,  $X = (X_1, \dots, X_p)$ , and  $\mu = (\mu_1, \dots, \mu_p)$ . We assume that  $X$  follows a multivariate normal distribution with mean vector  $\mu$  and unknown covariance matrix  $\Sigma$ . We first focus on testing that at least one response for the first treatment is better than the corresponding response of the second, formally, testing

$$H_0 : \max\{\mu_j \mid 1 \leq j \leq p\} \leq 0, \quad \text{vs.} \quad H_1 : \max\{\mu_j \mid 1 \leq j \leq p\} > 0, \quad (1)$$

Testing problem (1) is perhaps more common in selection and ranking problem for finding the largest element of several normal means (Gupta 1965; Hsu 1996, Shimodaira and Hasegawa 1999; Shimodaira 2000). In such problems, we often want to construct the confidence set for the index of the largest mean by simultaneously testing several normal mean differences, which is closely related to multiple comparisons with the unknown best (Hsu 1996; Shimodaira 2000). In Section 5, we

present such an application in finding true phylogenies.

For multi-parameter order-restricted hypothesis testing problems such as (1), the standard Hotelling  $T^2$  test is undesirable since it fails to incorporate the restrictions on the null and alternative parameter spaces. Under the normality assumption, the likelihood ratio test (LRT) has been proposed for testing (1), cf. Perlman (1969), Robertson et al. (1988). Since the null space in (1) can be expressed as an intersection of halfspaces, a reasonable alternative test is the union-intersection test (UIT). Recently, however, Perlman and Wu [PW] (2002) note that both the LRT and UIT may exhibit anomalous behavior for testing problems such as (1), since they are unable to adapt to the differing dimensionalities of the boundary of  $H_0$  (the nonpositive orthant in  $\mathbf{R}^p$ ). For the case of known  $\Sigma$ , [PW] (2002) propose a new test for testing (1) which better adapts to these varying dimensionalities and which they deem preferable to the LRT and UIT.

In this paper the new test of [PW] (2002) is both improved and extended to more general and realistic multivariate one-sided testing problems. The anomalies of the LRT and UIT for (1) are reviewed in Section 2. In §3.1 improvements of the new test of [PW] (2002) are offered when  $\Sigma$  is assumed known, then extended to the general case of unknown  $\Sigma$  in §3.3. In §3.5 these ideas are extended further to a possibly more realistic and practical version of (1), first proposed by Bloch *al.* (2001) and studied in [PW] (2004). In Section 4 these ideas are then applied to the well-known and important problem of testing the simple-order restriction. Two real data examples are presented in Section 5 to illustrate the new tests.

## 2 Anomalies of the LRT and UIT for (1)

The testing problem (1) can be expressed as that of testing

$$H_0 : \mu \in \mathcal{N}^p \quad \text{vs.} \quad H_1 : \mu \in \mathbf{R}^p \setminus \mathcal{N}^p, \quad (2)$$

where  $\mathcal{N}^p \equiv \{(\mu_1, \dots, \mu_p) : \mu_1 \leq 0, \dots, \mu_p \leq 0\}$  is the nonpositive orthant in  $\mathbf{R}^p$ . Here we review the anomalies of the LRT and UIT for testing (1)  $\equiv$  (2) when  $\Sigma$  is known, say  $\Sigma = I$  for simplicity.

The size  $\alpha$  LRT for (2) (cf. Robertson et al. (1988, §2.2, 2.3)) *accepts*  $H_0$  iff

$$\|X - \mathcal{N}^p\|^2 \equiv (X_1^+)^2 + \dots + (X_p^+)^2 \leq a_{p,\alpha}^2, \quad (3)$$

where  $X_i^+ \equiv \max(0, X_i)$  and  $a_{p,\alpha}^2 > 0$  is the unique solution to the equation

$$\alpha = \sum_{i=1}^p 2^{-p} \binom{p}{i} P(\chi_i^2 > a_{p,\alpha}^2). \quad (4)$$

Next, because  $H_0$  can also be written as  $H_0 : \cap_{i=1}^p H_{0i}$ , where  $H_{0i} : \mu_i \leq 0$ , the size  $\alpha$  UIT *accepts*  $H_0$  iff

$$\max(X_1, \dots, X_p) \leq u_{p,\alpha}, \quad (5)$$

where  $u_{p,\alpha} = \Phi^{-1}(\sqrt[2]{1 - \alpha})$ . It is straightforward to verify that  $u_{p,\alpha} < a_{p,\alpha}$ . (See Figure 1.)

Although the LRT and UIT have been popular for testing (1)  $\equiv$  (2), [PW] (2002) argued that both tests may yield inappropriate inferences, especially for large  $p$ . For example, suppose that  $p = 2$  and  $\alpha = 0.05$ . It can be shown that the size  $\alpha$  LRT rejects  $H_0$  if  $[(X_1^+)^2 + (X_2^+)^2]^{1/2} > 2.05$ , and the size  $\alpha$  UIT rejects  $H_0$  if  $\max(X_1, X_2) > 1.95$ . Now, suppose that we observe  $X^* = (1.8, -10)$ . Then, neither the LRT nor the UIT reject  $H_0$  (see Figure 1). However, if we consider testing  $H_{01} : \mu_1 \leq 0$  vs  $H_{11} : \mu_1 > 0$ , we see that  $H_{01}$  is rejected (note that  $z_\alpha := \Phi^{-1}(.95) = 1.64$ ), and therefore  $H_0$  should also be rejected since  $H_0 \subset H_{01}$ .

This anomalous behavior becomes more emphatic as  $p$  increases. Suppose that  $X^*$  is a sample point such that (i)  $X_i^* \gg z_\alpha$  and (ii)  $\max_{j \neq i} X_j^* \rightarrow -\infty$ , where  $1 \leq i \leq p$  is fixed. Since  $u_{p,\alpha} \rightarrow \infty$  as  $p \rightarrow \infty$ , the size  $\alpha$  UIT fails to reject  $H_0$ . However, (ii) strongly indicates that  $\max_{j \neq i} \mu_j \ll 0$ , so the null hypothesis in (1) reduces to  $H_{0,i} : \mu_i \leq 0$ , while (i) indicates that  $H_{0,i}$  should be strongly rejected. Since  $X_1, \dots, X_p$  are independent with standard deviation 1, the observation  $X^*$  therefore provides strong evidence against  $H_{0i}$  thus also against  $H_0$  (since  $H_0 \subset H_{0i}$ ), so the UIT

exhibits contradictory behavior. Since  $a_{p,\alpha} > u_{p,\alpha} \rightarrow \infty$  as  $p \rightarrow \infty$ , the LRT exhibits similarly contradictory behavior. In fact, for a sequence of alternatives  $(\mu_1, \dots, \mu_p)$  with  $\mu_1$  arbitrarily large but  $\max_{i \neq 1} \mu_j \rightarrow -\infty$  as  $p \rightarrow \infty$ , the powers of the LRT and UIT approach zero, but for such alternatives, any appropriate test procedure should have reasonable power to reject  $H_0$ . This anti-evidence property of the LRT and UIT renders these two tests undesirable for problem (1)  $\equiv$  (2) with  $\Sigma$  known.

This contradictory behavior is explained as follows. The orthant  $\mathcal{N}^p$  is a convex polyhedral cone whose boundary consists of a union of faces of dimensions 0 (the origin),  $1, \dots, p-1$ . Thus the boundary of  $H_0$  can be thought of as a union of statistical models of varying dimensionalities. Because the least favorable distribution occurs at  $\mu = 0$ , the LRT and UIT thus determine their critical values with reference to the model (face) of lowest dimension and are therefore biased in favor of the models (faces) of highest dimension, sometimes failing to reject these models (and therefore failing to reject  $H_0$ ) despite strong evidence to the contrary (for example, as provided by  $X^*$ ). We conclude that the contradictory behavior of the LRT and UIT is due to their inability to adapt to the different dimensions of the faces of  $\mathcal{N}^p$ . Such contradictory behavior of the LRT and UIT also occur in other multi-parameter testing problems where the null parameter space or its boundary consists of subregions of varying dimensionalities (cf. [PW] (2002, 2004)).

### 3 The New Tests for (1) $\equiv$ (2)

#### 3.1 The case $\Sigma$ known ( $\Sigma = I$ )

Tests more appropriate for (1)  $\equiv$  (2) than the LRT and UIT should adapt to the varying dimensionalities in the boundary of  $H_0$  – in this case, to the dimension of the face of  $\mathcal{N}^p$  closest to the sample point  $X$ . Shimodaira (2000) proposed such a test based on bootstrap methods. However,

his method is computational intensive, especially when  $p$  is large, and is limited to the case where  $\Sigma$  is known. For the simple case  $\Sigma = I$ , [PW] (2002) proposed a new test (see (7) below) for (1)  $\equiv$  (2) that utilizes the  $p$ -values associated with the LRTs for testing the individual faces of  $\mathcal{N}^p$  against  $H_1$ . Since a  $p$ -value is “self-weighting” according to the dimensionality of its null hypothesis, this new test thus adapts to the varying dimensionalities of the faces.

Let  $\mathcal{S}^p = 2^{\{1, \dots, p\}} \setminus \emptyset$  denote the collection of all nonempty subsets of  $\{1, \dots, p\}$ . For each  $\sigma \in \mathcal{S}^p$ , define  $L_\sigma = \{\mu \in \mathbf{R}^p \mid \mu_i = 0, \forall i \in \sigma\}$  and  $F_\sigma = L_\sigma \cap \mathcal{N}^p$ . Then  $L_\sigma$  is a linear subspace of dimension  $p - |\sigma|$  and  $\{F_\sigma \mid \sigma \in \mathcal{S}^p\}$  is the family of *faces* of  $\mathcal{N}^p \equiv \cup_{\sigma \in \mathcal{S}^p} F_\sigma$ . For  $\sigma \in \mathcal{S}^p \setminus \emptyset$ , let  $T_\sigma \equiv T_\sigma(X)$  be a statistic appropriate for testing

$$H_{0,\sigma} : \mu \in F_\sigma \quad \text{vs.} \quad H_1 : \mu \in \mathbf{R}^p \setminus \mathcal{N}^p \quad (6)$$

and let  $\pi_\sigma(T_\sigma)$  be the associated  $p$ -value. Then the test that *accepts*  $H_0$  iff

$$\max_{\sigma \in \mathcal{S}^p \setminus \emptyset} \pi_\sigma(T_\sigma(X)) \geq \alpha \quad (7)$$

is expected to be an approximately size  $\alpha$  test for (1)  $\equiv$  (2). This test uses the individual  $p$ -values  $\pi_\sigma$  to adapt to the varying dimensionalities of the faces  $F_\sigma$ .

Because it may be difficult to choose appropriate statistics  $T_\sigma$  and calculate the corresponding  $p$ -values  $\pi_\sigma$ , [PW] (2002) instead proposed an explicit and computationally simpler test (here denoted by PW1) that approximates the test (7). Test PW1 *accepts*  $H_0$  iff

$$(1 - 1_{\mathcal{N}^p}(X)) \sum_{i \in \sigma} X_i^2 \leq \tilde{a}_{|\sigma|,\alpha}^2 \quad \text{and} \quad \max_{i \notin \sigma} X_i \leq 0 \quad (8)$$

for at least one  $\sigma \in \mathcal{S}^p \setminus \emptyset$  (see Figure 1), where

$$\tilde{a}_{k,\alpha}^2 := \bar{G}_k^{-1} \left( \frac{\alpha}{1 - 2^{-k}} \right), \quad k = 1, \dots, p, \quad (9)$$

and  $G_k \equiv 1 - \bar{G}_k$  is the cdf of the  $\chi_k^2$  distribution. Note that  $\sum_{i \in \sigma} X_i^2$  is the appropriate  $\chi_k^2$  statistic for testing  $\mu \in L_\sigma$  vs.  $\mu \notin L_\sigma$ . The divisor  $1 - 2^{-k}$  in (9) renders this statistic appropriate for testing  $\mu \in L_\sigma$  vs.  $\max\{\mu_i \mid i \in \sigma\} > 0$ .

The test PW1 in (8) better adapts to the varying dimensionalities of the faces  $F_\sigma$  of  $\mathcal{N}^p$  and so reduces the undesirable behavior of the LRT and UIT. Simulation results in [PW] (2002) show that PW1 is approximately size  $\alpha$  for (1)  $\equiv$  (2), is more nearly similar on the boundary of  $H_0$ , and is more nearly unbiased than the LRT and the UIT. Moreover, PW1 has better overall power performance than the LRT and the UIT. Note, however, that as argued by [PW] (1999), our preference for PW1 is based not mainly on consideration of power and unbiasedness but rather on the fact that it better reflects the evidence that the data provides regarding the competing hypotheses in (1).

We now propose two modifications of PW1 that reflect this evidence even more accurately and also improve its power and unbiasedness properties. The first modified test, denoted as PW2, has acceptance region obtained by replacing  $\tilde{a}_{|\sigma|,\alpha}$  in (8) by  $a_{|\sigma|,\alpha}$  (cf. (3)). This is motivated by the fact that it is  $a_{|\sigma|,\alpha}$ , not  $\tilde{a}_{|\sigma|,\alpha}$ , that is the critical value for the LRT for testing the individual face  $F_\sigma$ . Another alternative test, denoted by PW3, replaces  $\tilde{a}_{|\sigma|,\alpha}$  in (8) by the average  $l_{|\sigma|,\alpha} \equiv (\tilde{a}_{|\sigma|,\alpha} + a_{|\sigma|,\alpha})/2$ . Figure 1 shows the acceptance regions of the LRT, UIT, PW1, and PW3 tests. (To avoid overcrowding, PW2, which is qualitatively similar to PW1 and PW3, is not depicted.)

**Figure 1 here: rejection regions of the LRT, UIT, PW1, and PW3.**

It is seen from Figure 1 that, unlike the LRT and UIT, the new tests PW1, PW2, and PW3 do not have monotone acceptance regions. Since  $H_0$  is composite, however, monotonicity is not a natural requirement for the testing problem (1)  $\equiv$  (2). For example, it is easy to construct prior distributions for which the corresponding Bayes (therefore admissible) tests are not monotone, so the class of monotone tests does not form a complete class.

**3.2 Simulations for the case  $\Sigma$  known ( $\Sigma = I$ ).**

The newly proposed tests PW2 and PW3 are compared to the LRT, UIT, and PW1 via simulations for dimensions  $p = 2, 5$  and various representative choices of the mean vector  $\mu$ . In all simulations

throughout the article, we choose the nominal size  $\alpha = 0.05 (= 5\%)$ , sample sizes  $n_1 = n_2 = 40$ , and 5000 iterations in each case. Table 1 shows the type I error rates and powers for the five tests. The new test PW2 appears to be slightly too liberal in that its type I error rate sometimes exceeds the nominal level  $\alpha = 0.05$ , so is eliminated from contention. Of the remaining four tests, PW3 is more nearly similar (hence more nearly unbiased) on the boundary of  $H_0$  at its faces of higher dimension, while the UIT appears more nearly similar at the faces of lowest dimension but the differences here are not as pronounced. PW3 is more nearly similar than PW1 in all cases, both tests better adapt to the varying dimensionalities of the boundary of  $H_0$  than UIT/LRT (e.g., for boundary points far from the origin such as  $\mu = (-5, 0)$  or  $\mu = (-2^4, 0)$ , the type I error rates for PW3/PW1 are much closer to the nominal level  $\alpha = 5\%$  than for UIT/LRT). Furthermore, the power of PW3 is greater than that of the others, except in the central region of the positive orthant of the alternative parameter space  $H_1$  where the power of the LRT is somewhat greater. (This conforms to the behavior to be expected in light of the nature of the acceptance regions in Figure 1.) PW3 is more powerful than PW1 in all cases, and both PW3 and PW1 can be substantially more powerful than LRT/UIT in many cases (see, e.g., when  $\mu = (-5^4, 1)$ ). Therefore, we conclude that the new test PW3 gives better results than all other tests, so PW3 is our recommended choice for testing  $(1) \equiv (2)$  when  $\Sigma = I$ .

### 3.3 The general case: $\Sigma$ unknown

In this section we extend the results in the previous section to the general and more realistic case where the covariance matrix  $\Sigma$  is completely unknown. Let  $\hat{\Sigma}$  denote the pooled sample covariance matrix and  $n = n_1 + n_2 - 2$ , so  $S \equiv n\hat{\Sigma}$  has the Wishart distribution  $W_p(\Sigma, n)$  (assume that  $n_1 + n_2 \geq p + 2$ ). From Perlman (1969), the size  $\alpha$  LRT for  $(1) \equiv (2)$  with  $\Sigma$  unknown *accepts*  $H_0$

Table 1. Simulation results: sizes and powers of tests for (1)  $\equiv$  (2) with  $\Sigma$  known ( $\Sigma = I$ ).

Dim.	$\mu$	PW3	PW2	PW1	LRT	UIT	$\mu$	PW3	PW2	PW1	LRT	UIT
Type I Error Rates (in %)												
$p = 2$	$\mu = (0, 0)$	4.7	6.1	3.6	5.2	5.1	$\mu = (-1, 0)$	3.2	4.0	2.7	2.0	2.4
	$\mu = (-2, 0)$	4.2	4.5	3.9	2.0	2.7	$\mu = (-5, 0)$	4.8	4.8	4.8	1.9	2.3
$p = 5$	$\mu = (0, 0^4)$	4.1	7.8	2.0	5.1	4.9	$\mu = (-1, 0^4)$	2.2	3.8	1.0	0.9	1.9
	$\mu = (-2, 0^4)$	3.6	6.3	1.8	3.3	3.9	$\mu = (-5, 0^4)$	3.3	5.1	2.3	0.9	2.0
	$\mu = (-1^2, 0^3)$	4.0	6.7	2.1	3.2	3.9	$\mu = (-2^2, 0^3)$	4.4	5.8	3.5	0.7	1.9
	$\mu = (-5^2, 0^3)$	4.2	7.0	2.2	3.5	3.9	$\mu = (-1^3, 0^2)$	2.1	2.4	0.7	0.4	1.4
	$\mu = (-2^3, 0^2)$	2.8	5.2	1.2	2.2	3.2	$\mu = (-5^3, 0^2)$	3.1	3.8	2.5	0.3	0.9
	$\mu = (-1^4, 0)$	3.8	6.3	2.0	1.8	2.9	$\mu = (-2^4, 0)$	5.3	5.3	5.3	0.3	0.9
Power Comparison (in %)												
$p = 2$	$\mu = (1, 1)$	33	38	28	37	32	$\mu = (2, 2)$	80	84	76	84	77
	$\mu = (-2, 1)$	24	25	22	14	17	$\mu = (-5, 1)$	64	64	64	47	51
$p = 5$	$\mu = (1, 1^4)$	50	62	37	59	40	$\mu = (2, 2^4)$	97	99	95	99	91
	$\mu = (-2, 1^4)$	43	55	32	46	34	$\mu = (-5, 1^4)$	43	54	33	44	32
	$\mu = (-2^2, 1^3)$	36	45	28	30	27	$\mu = (-5^2, 1^3)$	37	46	29	29	25
	$\mu = (-2^3, 1^2)$	27	33	21	15	18	$\mu = (-5^3, 1^2)$	32	37	27	14	18
	$\mu = (-2^4, 1)$	18	21	15	4	9	$\mu = (-5^4, 1)$	26	26	26	4	10

Notes: (i) In all simulations, iterations = 5,000, nominal  $\alpha = 5\%$ . (ii)  $(-1^3, 0^2) \equiv (-1, -1, -1, 0, 0)$ , etc.

iff

$$\|X - \mathcal{N}^p\|_S^2 \equiv \|X - \pi_S(X; \mathcal{N}^p)\|_S^2 \leq a_{p,\alpha}^*, \quad (10)$$

where  $\|x\|_S^2 \equiv x^t S^{-1} x$  is the Euclidean norm determined by  $S$ ,  $\pi_S(X; \mathcal{N}^p)$  is the projection of  $X$  onto  $\mathcal{N}^p$  with respect to this norm, and  $a_{p,\alpha}^*$  is the critical value of the size  $\alpha$  LRT for  $H_0$  determined by the equation

$$\alpha = \frac{1}{2} \Pr \left[ \frac{\chi_{p-1}^2}{\chi_{n_1+n_2-p}^2} > a_{p,\alpha}^* \right] + \frac{1}{2} \Pr \left[ \frac{\chi_p^2}{\chi_{n_1+n_2-p-1}^2} > a_{p,\alpha}^* \right] \quad (11)$$

$$\equiv \sup_{\mu \in \mathcal{N}^p, \Sigma > 0} \Pr_{\mu, \Sigma} [\|X - \mathcal{N}^p\|_S^2 > a_{p,\alpha}^*]. \quad (12)$$

Perlman (1969) showed that  $\pi_S(X; \mathcal{N}^p)$  is the MLE of  $\mu$  under  $H_0 : \mu \in \mathcal{N}^p$  with  $\Sigma$  unknown. Tang (1994) has tabulated some values of  $a_{p,\alpha}^*$ . The values of  $a_{p,\alpha}^*$  can also be determined by numerical methods. The LR statistic can be computed by a program for minimizing linear inequality-constrained Mahalanobis distance (Wollan and Dykstra (1987)).

The UIT for (1)  $\equiv$  (2) with  $\Sigma$  unknown *accepts*  $H_0$  iff

$$\max(t_1, \dots, t_p) \leq v_{p,\alpha}, \quad (13)$$

where  $t_i = \frac{X_i}{\sqrt{s_{ii}}}$  and  $s_{11}, \dots, s_{pp}$  are the diagonal elements of  $S$ . Under  $H_0$  the distribution of  $\max(t_1, \dots, t_p)$  is stochastically largest when  $\mu = 0$ , so the critical value  $v_{p,\alpha}$  may be approximated from the Bonferroni inequality as follows:  $v_{p,\alpha} \approx t_{n,\alpha/p}$ , the upper  $(\alpha/p)$ -quantile of the Student  $t$ -distribution with  $n$  degrees of freedom (cf. Tamhane and Logan (2004)).

As in §3.1, the LRT and UIT do not adapt to the varying dimensionalities of the faces of  $\mathcal{N}^p$ . We thus propose the following new test, similar to the PW tests in §3.1 and designated here as PW4, which is intended to adapt to the varying dimensionalities: *accept*  $H_0$  iff

$$[1 - 1_{\mathcal{N}^p}(X)] \cdot \|X - L_\sigma\|_S^2 \leq a_{|\sigma|,\alpha}^* \quad \text{and} \quad \pi_S(X; L_\sigma) \in \mathcal{N}^p \quad (14)$$

for at least one  $\sigma \in \mathcal{S}^p$ , where the critical values  $a_{|\sigma|,\alpha}^*$  are given by (11). Test (14) is a generalization of test (8), and is motivated by the idea of combining the  $p$ -values associated with testing the faces

of  $\mathcal{N}^p$  individually, as in Perlman and Wu (2002). Thus, the new test should be preferable to the LRT and UIT, as shown next.

### 3.4 Simulations for the general case: $\Sigma$ unknown.

In this section we compare the new test PW4 (14) with the LRT and UIT via simulations. For dimensions  $p = 2$  and  $p = 5$ , the size and power of the tests are simulated for various representative choices of the mean vector  $\mu$  and four covariance matrices  $\Sigma_1, \dots, \Sigma_4$ . Each  $\Sigma_i$  is an intraclass correlation matrix with all diagonal elements = 1 and all off-diagonal elements =  $\rho_i$ , with  $\rho_1 = 0$ ,  $\rho_2 = 0.4$ ,  $\rho_3 = 0.8$ ,  $\rho_4 = -0.2$ . In particular,  $\Sigma_1 = I$ , the identity matrix. The sample sizes in all simulations are  $n_1 = n_2 = 40$  (results for sample sizes  $n_1 = n_2 = 20$  were similar) and the iteration number is 5000, as noted earlier.

Simulated values of Type I error rates and powers for the three tests are given in Table 2. The new test PW4 in (14) more nearly attains the nominal level  $\alpha = 5\%$  while the LRT and UIT can be very conservative in some cases (e.g., when  $\mu = (-1^4, 0)$ , etc). PW4 appears to better adapt to the dimensionality of the faces of  $H_0$  (e.g., for boundary points far from the origin  $\mathbf{0}$ , the type I errors for PW4 are much closer to the nominal level  $\alpha$  than for UIT/LRT), and is more nearly similar on the boundary (so less biased) than the LRT and UIT. Also, PW4 is more powerful than the LRT and UIT in most cases, except perhaps in the central region of the positive orthant of the alternative parameter space  $H_1$  when the correlation is negative. The power advantage of PW4 over LRT and UIT can be substantial (see, e.g., when  $\mu = (-1^4, 0.5)$ , etc). The performance of LRT and UIT seems to be mixed: neither test dominates the other. It appears that the UIT adapts the dimensionality slightly better than the LRT, but both are usually worse than the new test PW4. We conclude that PW4 is the test of preference.

Table 2. Simulation results: sizes and powers of tests for (1)  $\equiv$  (2) with  $\Sigma$  unknown.

Dimension	Mean $\mu$	LRT	UIT	PW4	LRT	UIT	PW4	LRT	UIT	PW4	LRT	UIT	PW4
Type I Error Rates (in %)													
		$\Sigma = \Sigma_1$			$\Sigma = \Sigma_2$			$\Sigma = \Sigma_3$			$\Sigma = \Sigma_4$		
$p = 2$	(0, 0)	3.0	4.8	3.8	2.8	4.8	3.8	1.6	3.5	2.6	3.1	5.0	4.1
	(-1, 0)	1.1	2.6	5.0	1.2	2.8	5.1	1.0	2.5	4.8	0.9	2.6	5.1
	(-5, 0)	1.2	2.4	4.9	1.0	2.2	4.5	1.1	2.4	4.8	1.4	2.8	5.4
$p = 5$	(0, 0 <sup>4</sup> )	1.7	5.3	2.9	0.7	4.6	1.8	0.2	2.6	0.9	3.2	4.8	4.1
	(-1, 0 <sup>4</sup> )	0.8	4.0	3.1	0.6	3.2	2.0	0.2	2.3	1.3	1.2	3.4	3.4
	(-1 <sup>2</sup> , 0 <sup>3</sup> )	0.3	3.0	3.6	0.2	2.8	2.8	0.1	2.2	1.8	0.5	3.2	3.9
	(-1 <sup>3</sup> , 0 <sup>2</sup> )	0.2	2.1	4.2	0.1	1.5	3.0	0.1	1.4	2.6	0.2	1.6	4.0
	(-1 <sup>4</sup> , 0)	0.0	1.0	5.0	0.1	1.0	4.7	0.1	0.9	4.8	0.1	0.9	4.4
Power Comparison (in %)													
		$\Sigma = \Sigma_1$			$\Sigma = \Sigma_2$			$\Sigma = \Sigma_3$			$\Sigma = \Sigma_4$		
$p = 2$	(0.2, 0.2)	22	25	23	17	24	19	13	21	15	28	26	28
	(-1, 0.3)	17	26	37	16	25	36	17	27	38	17	27	37
	(-5, 0.3)	17	25	36	17	26	38	17	27	37	17	26	36
$p = 5$	(0.1, 0.1 <sup>4</sup> )	9	14	11	2	11	4	1	7	2	39	14	39
	(-1, 0.5 <sup>4</sup> )	94	91	97	53	79	68	28	63	46	99	96	100
	(-1 <sup>2</sup> , 0.5 <sup>3</sup> )	79	84	93	44	73	71	24	60	52	96	88	99
	(-1 <sup>3</sup> , 0.5 <sup>2</sup> )	50	70	85	32	64	73	21	55	61	62	72	92
	(-1 <sup>4</sup> , 0.5)	14	45	72	13	43	71	14	44	71	14	46	72

Notes: (i) PW4 = the new test (14). (ii) In all simulations, iterations = 5,000, nominal  $\alpha = 5\%$ ,  $n_1 = n_2 = 40$ .

(iii)  $(-1^4, 0.5) = (-1, -1, -1, -1, 0.5)$ , etc.

### 3.5 A related testing problem

Testing problem (1)  $\equiv$  (2) is formulated to determine whether or not at least one endpoint in treatment 1 (say) is significantly superior than the corresponding endpoint in treatment 2. As noted in Section 1, however, Bloch *et al.* (2001) argue that sometimes it is more practical to assert that treatment 1 is preferred if it is superior for at least one of the endpoints and biologically “noninferior” for the remaining endpoints. This leads to the following reformulated testing problem: test

$$H'_0 : \mu \in \Theta_0 \equiv \left\{ \max_{1 \leq j \leq p} \mu_j \leq 0 \right\} \cup \left\{ \max_{1 \leq j \leq p} \mu_j > 0 \text{ and } \mu_j \leq -\epsilon_j \text{ for some } j \right\}, \quad (15)$$

versus  $H'_1 : \text{not } H'_0$ , where  $\epsilon_j$ 's are pre-specified positive numbers. Again assume that  $\Sigma$  is unknown.

Bloch *et al.* (2001) noted that  $H'_0$  is a union of  $H_0 : \mu \in \mathcal{N}^p$  and  $H_0^{(j)} : \mu_j \leq -\epsilon_j, j = 1, \dots, p$ , so an intersection-union test (IUT) is appropriate for (15). They combined the Hotelling  $T^2$ -test for  $H_0^* : \mu = 0$  with the usual  $t$ -tests for  $H_0^{(j)}, j = 1, \dots, p$ , thus obtaining the following approximate size  $\alpha$  test: *reject*  $H'_0$  iff

$$T^2 \equiv \hat{\mu}^t \hat{\Sigma}^{-1} \hat{\mu} > c_\alpha, \quad \text{and} \quad \hat{\mu}_j + \epsilon_j / s_{jj}^{1/2} > u_\alpha \quad \text{for all } j = 1, \dots, p, \quad (16)$$

where  $u_\alpha = t_{n,\alpha}$ , the upper  $\alpha$ -quantile of the  $t_n$ -distribution, and  $c_\alpha$  is the size  $\alpha$  critical value for the Hotelling  $T^2$  statistic.

Because the Hotelling  $T^2$ -test for  $H_0^*$  is not designed to test the one-sided hypothesis  $H_0$ , [PW] (2004) asserted that a more appropriate test, here designated as PW5, is obtained by replacing  $T^2$  by the LRT statistic for testing  $H_0$  in (1)  $\equiv$  (2) ( $\Sigma$  unknown), which better reflects its one-sided form. As noted above, however, this LRT does not adapt to the varying dimensionalities in the boundary of  $H_0$ . Therefore, here we propose an even more appropriate IUT for (15) with  $\Sigma$  unknown, obtained by replacing the  $T^2$ -test in (16) by the new test (14). The resulting size  $\alpha$  IUT test, here designated

Table 3. Simulation results: sizes and powers of tests for (15) with  $\Sigma$  unknown.

Mean $\mu$	PW5	PW6	PW5	PW6	PW5	PW6	PW5	PW6
Type I Error Rates (in %)								
	$\Sigma = \Sigma_1$		$\Sigma = \Sigma_2$		$\Sigma = \Sigma_3$		$\Sigma = \Sigma_4$	
(0, 0)	3.0	3.9	2.3	3.1	2.1	2.8	3.1	4.1
(-0.5, 0)	0.3	1.1	0.6	2.9	1.2	4.9	0.2	0.6
(-1, 0)	2.5	3.6	4.1	4.8	4.6	4.6	1.7	3.1
Power Comparison (in %)								
	$\Sigma = \Sigma_1$		$\Sigma = \Sigma_2$		$\Sigma = \Sigma_3$		$\Sigma = \Sigma_4$	
(0.1, 0.1)	10	11	8	9	6	8	11	12
(-0.5, 0.5)	34	49	40	56	44	62	32	44
(-0.2, 0.6)	64	72	64	80	65	85	64	69

Note: In all simulations, iterations = 5,000, nominal  $\alpha = 5\%$ ,  $p = 2$ ,  $\epsilon = 1$ ,  $n_1 = n_2 = 40$ .

as PW6, has an *acceptance* region given by

$$\begin{aligned} \mathcal{A} = & \{X : [1 - 1_{\mathcal{N}^p}(X)] \cdot \|X - L_\sigma\|_S^2 \leq a_{|\sigma|, \alpha}^* \text{ and } \pi_S(X; L_\sigma) \in \mathcal{N}^p \text{ for at least one } \sigma \in \mathcal{S}^p \setminus \{\emptyset\}\} \\ & \cup \{X : \hat{\mu}_j + \epsilon_j / \hat{\sigma}_{jj}^{1/2} \leq u_\alpha \text{ for at least one } j\}. \end{aligned} \tag{17}$$

Test PW6 not only respects the one-sided form of  $H_0$  but should also better adapt to the varying dimensionalities in its boundary.

To verify this, we have compared the PW6 test (17) to PW5 via simulations. The design of the simulation is similar to earlier ones and is briefly noted at the bottom of Table 3. The simulation results, given in Table 3, confirm that PW6 shows some improvement over PW5 in terms of both nominal size and power, i.e., PW6 is more similar and more powerful than PW5 in all cases considered in the simulations.

Recently, Tamhane and Logan (2004) consider the same testing problem (15) and propose a new test with sharper critical bound than that of Bloch *et al.* (2001). However, their simulation

results indicate that PW5 may still be better than their new test. Since [PW] (2004) already demonstrated that PW5 performs better than the test of Bloch *et al.* (2001) and Bloch *et al.* (2001) showed that their test is better than previous tests in the literature, we conclude that PW6 is the best test among all previously proposed tests for testing (15).

## 4 Testing the Simple-Order Restriction

### 4.1 The case $\Sigma$ known ( $\Sigma = I$ )

The anomalous behavior of the LRT and IUT occurs in any multivariate testing problem where the null parameter space or its boundary is the union of sets of varying dimensionalities, in particular a polyhedral cone, and, as for the nonpositive orthant cone, new tests that adapt to these dimensionalities may be obtained. Here we focus on the LRT for the well-known problem of testing the simple-order cone (cf. Robertson *et al.* 1988, Ch.2).

Let  $\mathcal{C}^p$  denote the *simple-order cone* in  $\mathbf{R}^p$  defined by

$$\mathcal{C}^p = \{\mu \equiv (\mu_1, \dots, \mu_p) \in \mathbf{R}^p \mid \mu_1 \leq \dots \leq \mu_p\}, \quad (18)$$

which is a non-pointed<sup>1</sup> acute polyhedral cone with “spine” given by the line  $\{\mu \mid \mu_1 = \dots = \mu_p\}$ .

Consider the problem of testing

$$\bar{H}_0 : \mu \in \mathcal{C}^p \quad \text{vs} \quad \bar{H}_1 : \mu \in \mathbf{R}^p \setminus \mathcal{C}^p \quad (19)$$

based on the data  $X \equiv (X_1, \dots, X_p) \sim N(\mu, \Sigma)$ . For the case  $\Sigma$  known with  $\Sigma = I$ , the LRT for (19) *accepts*  $\bar{H}_0$  iff

$$\|X - \mathcal{C}^p\|^2 \equiv \|X - \pi(X; \mathcal{C}^p)\|^2 \leq d_{p,\alpha}^2, \quad (20)$$

---

<sup>1</sup>A convex cone  $C$  is non-pointed if it contains a nontrivial linear subspace. Its spine  $L$  is the maximal such subspace, and  $C$  can be uniquely represented as the product  $L \times \tilde{C}$ , where  $\tilde{C}$  is the pointed cone obtained by projecting  $C$  onto the orthogonal complement  $L^\perp$ .

where  $d_{p,\alpha}^2$  satisfies

$$\alpha = \sum_{i=1}^{p-1} P(i, p) \Pr[\chi_{p-i}^2 > d_{p,\alpha}^2] \quad (21)$$

and  $\{P(i, p) \mid i = 1, \dots, p\}$  are the *level probabilities* associated with the LRT statistic for the case of equal weights – cf. Robertson *et al.* (1988, pp.69-70, 79-82).

We now propose a new test that combines the  $p$ -values associated with the faces of  $\mathcal{C}^p$ . Let  $\mathcal{F}_p$  denote the set of faces of  $\mathcal{C}^p$ . There is a 1-1 correspondence between  $\mathcal{F}_p$  and  $\mathcal{S}^{p-1} \equiv 2^{\{1, \dots, p-1\}} \setminus \emptyset$ , described as follows. Each face  $F \in \mathcal{F}_p$  has the form  $F_\tau \equiv L_\tau \cap \mathcal{C}^p$  for some unique  $\tau \in \mathcal{S}^{p-1}$ , where  $L_\tau \subset \mathbf{R}^p$  is the linear subspace given by  $L_\tau = \{\mu \mid \mu_i = \mu_{i+1}, \forall i \in \tau\}$ . Conversely, each  $\tau \in \mathcal{S}^{p-1}$  determines a unique face  $F_\tau$ . The dimension of both  $L_\tau$  and  $F_\tau$  is  $p - |\tau|$ . For example, if  $p = 5$  and  $F$  is the face of  $\mathcal{C}^5$  that spans the linear subspace given by the constraints  $\mu_1 = \mu_2$  and  $\mu_3 = \mu_4 = \mu_5$ , then  $F = F_\tau$  where  $\tau = \{1, 3, 4\}$  and  $\dim(F_\tau) = 2$ .

For each  $\tau \in \mathcal{S}^{p-1}$ , let  $T_\tau \equiv T_\tau(X)$  be a statistic appropriate for testing

$$H_{0,\tau} : \mu \in F_\tau \quad \text{vs.} \quad H : \mu \in \mathbf{R}^p \setminus \mathcal{C}^p \quad (22)$$

and let  $p_\tau \equiv p(T_\tau)$  be the associated  $p$ -value. Then the test (compare to (7)) that *accepts*  $\bar{H}_0$  iff

$$\max_{\tau \in \mathcal{S}^{p-1}} p_\tau \geq \alpha \quad (23)$$

will be an approximately size  $\alpha$  test for (19). This test uses the individual  $p$ -values  $p_\tau$  to adapt to the varying dimensionalities of the faces  $F_\tau$  of the polyhedral cone  $\mathcal{C}^p$ .

Because it may be difficult to choose appropriate statistics  $T_\tau$  and/or to calculate the corresponding  $p$ -values  $p_\tau$ , we propose instead the following test, denoted by PW7, which is qualitatively similar to (23) and somewhat easier to implement: *accept*  $\bar{H}_0$  iff

$$[1 - 1_{\mathcal{C}^p}(X)] \cdot \|X - L_\tau\|^2 \leq b_{|\tau|,\alpha}^2 \quad \text{and} \quad \pi_I(X; L_\tau) \in \mathcal{C}^p \quad (24)$$

Table 4. Simulation results: sizes and powers of tests for (19) with  $\Sigma$  known ( $\Sigma = I$ ).

$\mu$	LRT	PW7	PW8	PW9	$\mu$	LRT	PW7	PW8	PW9
Type I Error Rates					Power Comparison				
$\mu = (0, 0, 0)$	5.3	4.1	6.6	5.1	$\mu = (0.5, 0, 0)$	89	83	89	86
$\mu = (0, 0, 1)$	1.5	5.4	5.4	5.4	$\mu = (0.5, 0.5, 0)$	30	33	39	36
$\mu = (0, 1, 1)$	1.8	5.3	5.3	5.3	$\mu = (0.5, 0, 0.5)$	29	31	38	34
					$\mu = (0, 0, -0.5)$	84	89	91	90

Notes: In the simulations,  $p = 3$ , iterations = 5,000, nominal  $\alpha = 5\%$ .

for at least one  $\tau \in \mathcal{S}^{p-1}$ , where

$$b_{k,\alpha}^2 \equiv \bar{G}_k^{-1} \left( \frac{\alpha}{1 - \frac{1}{(k+1)!}} \right) \quad (25)$$

with  $G_k \equiv 1 - \bar{G}_k$  the cdf of the  $\chi_k^2$  distribution (compare to (8) and (9)).

**Figure 2 here: showing the acceptance region of the new test.**

As in Section 3.2, we may also consider two modifications of the new test PW7. The first modified test, denoted by PW8, is obtained by replacing  $b_{|\tau|,\alpha}^2$  in (24) with  $d_{|\tau|,\alpha}^2$  in (20), while the second modified test, denoted by PW9, is obtained by replacing  $b_{|\tau|,\alpha}^2$  in (24) with the average  $(b_{|\tau|,\alpha}^2 + d_{|\tau|,\alpha}^2)/2$ . Figure 2 shows the acceptance regions of the LRT and PW7 for (19), while Table 4 presents some size and power comparisons (the design of the simulation is similar to earlier ones and is briefly noted at the bottom of Table 4). We see that the new tests PW7, PW8, and especially PW9, are more nearly similar and more powerful than the LRT in most cases, and they also better adapt to the varying dimensionalities of the faces of  $\bar{H}_0$  (e.g., their type I error rates remain closer to the nominal level  $\alpha$  for boundary points far from the line  $L \equiv \{\mu \mid \mu_1 = \dots = \mu_p\}$  while those of the LRT do not).

## 4.2 The general case: $\Sigma$ unknown

Here we consider the problem (19) of testing the simple-order restriction under the more realistic assumption that  $\Sigma$  is completely unknown. The LRT (cf. Perlman (1969)) *accepts*  $\bar{H}_0$  iff

$$\|X - \mathcal{C}^p\|_S^2 \equiv \|X - \pi_S(X; \mathcal{C}^p)\|_S^2 \leq d_{p,\alpha}^{*2}, \quad (26)$$

where  $d_{p,\alpha}^{*2}$  is defined by

$$\alpha = \frac{1}{2} \Pr \left[ \frac{\chi_{p-1}^2}{\chi_{n_1+n_2-p}^2} > d_{p,\alpha}^{*2} \right] + \frac{1}{2} \Pr \left[ \frac{\chi_p^2}{\chi_{n_1+n_2-p-1}^2} > d_{p,\alpha}^{*2} \right] \quad (27)$$

$$\equiv \sup_{\mu \in \mathcal{C}^p, \Sigma > 0} \Pr_{\mu, \Sigma} [\|X - \mathcal{C}^p\|_S^2 > d_{p,\alpha}^{*2}]. \quad (28)$$

Instead, we propose a new test, denoted by PW10, which *accepts*  $\bar{H}_0$  iff

$$[1 - 1_{\mathcal{C}^p}(X)] \cdot \|X - L_\tau\|_S^2 \leq d_{|\tau|,\alpha}^{*2} \text{ and } \pi_S(X, L_\tau) \in \mathcal{C}^p \quad (29)$$

for at least one  $\tau \in \mathcal{S}^{p-1}$ .

We conducted a simulation study to compare the new test PW10 in (29) with the LRT in (26) for representative values of the mean vector  $\mu$  and for the four intraclass correlation matrices  $\Sigma_1, \dots, \Sigma_4$  already defined in conjunction with Table 2. We considered the cases of  $p = 3$  and  $p = 5$ . The sample sizes in all simulations are  $n_1 = n_2 = 40$  (results for sample sizes  $n_1 = n_2 = 20$  are similar), and the iteration number is 5000. Simulated values of Type I error rates and powers for the two tests are given in Table 5. The new test PW10 more nearly attains the nominal level  $\alpha = 5\%$  while the LRT is much more conservative, especially for  $p = 5$ . PW10 is more nearly similar on the boundary of  $\bar{H}_0$  and less biased than the LRT, and is often substantially more powerful than the LRT. Table 5 also indicates that PW10 adapts to the varying dimensionality of the faces of  $\bar{H}_0$ , while the LRT fails to do so (i.e., the type I error rates of PW10 remain close to the nominal level  $\alpha = 5\%$  for all points  $\mu$  on the boundary while those of LRT do not). Thus, PW10 is clearly preferable to the LRT. Note again, however, that our preference for PW10 is based mainly on its

Table 5. Simulation results: tests for the simple-order hypothesis (19) with  $\Sigma$  unknown.

Dimension $p$	Mean $\mu$	LRT	PW10	LRT	PW10	LRT	PW10	LRT	PW10
Type I Error Rates (in %)									
		$\Sigma = \Sigma_1$		$\Sigma = \Sigma_2$		$\Sigma = \Sigma_3$		$\Sigma = \Sigma_4$	
$p = 3$	(0, 0, 0)	1.3	4.8	1.3	4.0	1.3	4.4	1.4	4.8
	(0, 0, 1)	0.5	4.2	0.4	4.3	0.4	4.7	0.4	4.8
	(0, 1, 1)	0.5	5.2	0.3	4.8	0.3	4.4	1.6	4.6
$p = 5$	(0 <sup>4</sup> , 0)	1.6	3.9	1.5	4.0	1.6	4.4	1.6	4.8
	(0 <sup>4</sup> , 1)	0.7	4.4	0.6	4.1	0.6	4.1	0.4	3.6
	(0 <sup>3</sup> , 1 <sup>2</sup> )	0.7	4.1	0.4	3.3	0.5	4.0	0.4	4.0
	(0 <sup>2</sup> , 1 <sup>3</sup> )	0.6	3.7	0.6	3.5	0.6	3.8	0.4	4.0
	(0, 1 <sup>4</sup> )	0.6	4.9	0.9	4.3	0.7	4.3	1.4	4.4
Power Comparison (in %)									
		$\Sigma = \Sigma_1$		$\Sigma = \Sigma_2$		$\Sigma = \Sigma_3$		$\Sigma = \Sigma_4$	
$p = 3$	(0.5, 0, 0)	27	43	38	56	86	94	24	38
	(0.5, 0.5, 0)	27	42	22	38	20	35	28	45
	(0.5, 0, 0.5)	15	34	27	50	80	92	11	29
	(0, 0, -0.5)	26	41	23	38	20	34	28	42
$p = 5$	(0.4, 0 <sup>4</sup> )	14	24	26	38	76	84	11	19
	(0, 0.4, 0 <sup>3</sup> )	11	21	19	33	66	81	8	17
	(0, 0.4 <sup>2</sup> , 0 <sup>2</sup> )	14	28	26	45	82	93	11	22
	(0.3 <sup>4</sup> , -0.3)	36	48	59	71	99	100	29	42
	(0.4, 0, 0.4, 0, 0.4)	13	26	24	43	79	93	9	21

Notes: (i) PW10 = the new test (29). (ii) In all simulations, iterations = 5,000, nominal  $\alpha = 5\%$ ,  $n_1 = n_2 = 40$ . (iii)  $(-1^4, 0.5) = (-1, -1, -1, -1, 0.5)$ , etc.

better adaption to the dimensionalities of the boundary faces of  $\bar{H}_0$ , thus better representing the evidence provided by the data, as discussed in Section 2, rather than directly on consideration of sizes and powers.

## 5 Examples

### 5.1 Example 1

Here we consider a model selection example for finding true phylogenies. The data set, previously analyzed in Shimodaira and Hasegawa (1999) and Shimodaira (2000), contains mitochondrial protein sequences of 3414 amino acids for six mammal species (human, harbor seal, cow, rabbit, mouse, and opossum). There are 105 possible phylogenies for the six species. We are interested in finding the true phylogeny — the hypothetical tree of the evolution history. As discussed in Shimodaira and Hasegawa (1999), each phylogeny can be represented as a probabilistic model  $M_i$ ,  $i = 0, 1, \dots, 105$ , and the maximized log-likelihood of  $M_i$ , denoted by  $Y_i$ , is approximately normal. As illustration, here we consider  $p = 5$  phylogenies selected from 15 most probable phylogenies (see Shimodaira, 2000). Let  $E(Y_i) = \eta_i$  and  $\mu_{jk} = \eta_j - \eta_k = E(Y_j - Y_k)$ ,  $j, k = 0, 1, \dots, p$ . As in Shimodaira (2000), we consider a  $(1 - \alpha) \times 100\%$  confidence set for the true phylogenies. This can be achieved by testing each  $H_0^{(k)} : \max_{j \neq k} \mu_{jk} \leq 0$  versus  $H_1^{(k)} : \text{not } H_0^{(k)}$ ,  $k = 0, 1, \dots, p$ , and determining the indices  $k$  for which  $H_0^{(k)}$  is not rejected at level  $\alpha$ . Thus each test corresponds to testing problem (1)  $\equiv$  (2) with  $\Sigma$  unknown (see Section 3.3).

We index the  $p = 5$  phylogenies based on the order of  $\Delta Y_k = \max_{j \neq k} (Y_j - Y_k)$ ,  $k = 0, 1, 2, 3, 4$  (from smallest to largest), and obtain  $(\Delta Y_0, \Delta Y_1, \Delta Y_2, \Delta Y_3, \Delta Y_4) = (0.0, 19.5, 22.7, 29.1, 33.6)$ . At the 90% confidence level, the three tests in Section 3.3, LRT, UIT, and the new test PW4, produce the same confidence set which covers all 5 phylogenies. However, at 80% confidence level, the

confidence set based on LRT is  $\{1, 2, 3, 4, 5\}$ , the confidence set based on UIT is  $\{1, 2, 3\}$ , and the confidence set based on PW4 is  $\{1, 2, 3, 4\}$ . Therefore, the three tests may produce different results. Based on the simulations in Section 3.3, the new test PW4 should provide the most reliable results.

## 5.2 Example 2

We consider a longitudinal study on the change of mental distress of parents whose children died by accident. Data were collected on parents at 3 month, 6 month, and 18 month after their children's death (Murphy et al. 2003). Mental distress was measured in terms of depression, anxiety, hostility, somatization, and so on at each time point. We focus on the depression data, and only consider male parents in the treatment group whose children were over 20 year-old and died by accident. A total of  $n = 11$  parents have depression data available at all three time points. Let  $Y_1, Y_2$ , and  $Y_3$  denote depression measurements at month 3, 6, and 18 post-death. Let  $\mu_i = E(Y_i)$ ,  $i = 1, 2, 3$ . We are interested in testing whether depression decreases over time, that is, we would like to test  $H_0 : \mu_3 \leq \mu_2 \leq \mu_1$  versus  $H_1 : \text{not } H_0$ , which corresponds to testing problem (19) in Section 4.2.

It appears that a multivariate normality assumption may be reasonable based on exploratory analyses. The sample mean, sample covariance, and correlation matrices are

$$(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3) = (0.60, 0.97, 0.73), \quad \hat{\Sigma} = \begin{pmatrix} 0.32 & 0.63 & 0.40 \\ 0.63 & 1.51 & 0.92 \\ 0.40 & 0.92 & 0.57 \end{pmatrix}, \quad \hat{R} = \begin{pmatrix} 1 & 0.91 & 0.94 \\ 0.91 & 1 & 0.99 \\ 0.94 & 0.99 & 1 \end{pmatrix},$$

respectively. At the 5% level, the LRT in Section 4.2 fails to reject  $H_0$ , while the new test PW10 rejects  $H_0$ , suggesting that the depression of fathers does not decrease over time. Based on the simulation studies, the results from the new test PW10 should be more reliable. Thus, we conclude that depression of the fathers does not decrease over time. A more thorough analysis will be reported separately.

## 6 Discussion

The normality assumption in the article may be relaxed. When normality is not satisfied, we may use bootstrap methods, as in Shimodaira (2000) and Bloch et al. (2001). In such cases, we may let the test statistics remain the same, but use bootstrap methods to obtain the critical values. The implementation may be straightforward but it can be computationally intensive.

The testing problems we have considered are common in applications such as selection and ranking problems and clinical trials with multiple endpoints. The LRT/UIT are still widely used in these situations, but our results indicate that our proposed new tests do provide substantial improvement for testing multivariate one-side hypotheses.

**Acknowledgement.** The authors thank Professor Ajit Tamhane for helpful comments and suggestions.

## References

- Bloch, D.A., Lai, T.L., and Tubert-Bitter, P. (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics* **57**, 1039-1047.
- Hsu, J.C. (1996). *Multiple Comparisons - Theory and Methods*. Chapman and Hall, London/New York.
- Murphy, S.A., Johnson, L.C., Wu, L., Fan, J.J., Lohan, J. (2003). "Bereaved Parents' Outcomes 4 to 60 Months After Their Children's Deaths by Accident, Suicide, or Homicide: A Comparative Study Demonstrating Differences." *Death Studies*, **27**, 39-61.
- O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079-1087.

- Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *Annals of Statistics* **40**, 549-567.
- Perlman, M. D. and Wu, L. (1999). The Emperor's new tests (with discussion). *Statistical Science* **14** 355-381.
- Perlman, M. D. and Wu, L. (2002). On the Validity of the Likelihood Ratio and Maximum Likelihood Methods. *Journal of Statistical Planning and Inference* **117**, 59-81.
- Perlman, M. D. and Wu, L. (2004). A Note on One-Sided Tests with Multiple Endpoints. *Biometrics*, to appear.
- Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order-Restricted Statistical Inference*. New York: Wiley.
- Shimodaira, H. (2000). Approximately unbiased one-sided tests of the maximum of normal means using iterated bootstrap corrections. Tech. Report No. 2000-07, Dept. of Statistics, Stanford University, Stanford, CA.
- Shimodaira and Hasegawa (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114-1116.
- Tamhane, A.C. and Logan, B.R. (2004). A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. To appear in *Biometrika*.
- Tang, D.-I. (1994). Uniformly more powerful tests in a one-sided multivariate problem. *Journal of the American Statistical Association* **89**, 1006 - 1011.
- Wollan, P.C. and Dykstra, R.L. (1987). Algorithm AS 225, *Applied Statistics*, **36**, 234-240.



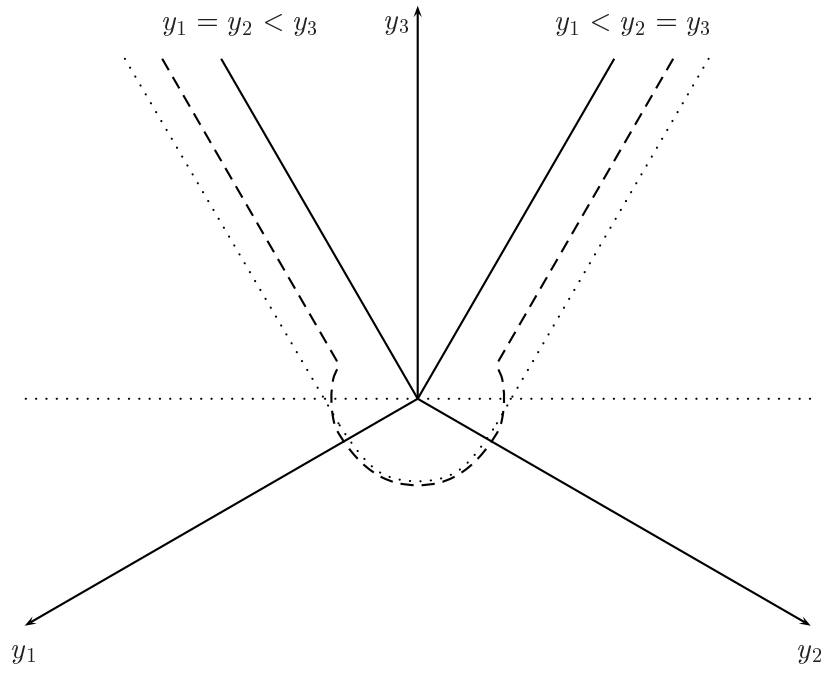


Figure 2. *Rejection/acceptance regions of the LRT and the new test PW9 for (19) with  $\Sigma$  known ( $\Sigma = I$ ). LRT: dotted line, PW9: dashed line*