Some Improved Tests for Multivariate One-Sided Hypotheses

August, 2005

Lange Wu

Department of Statistics

University of British Columbia

Vancouver, BC, Canada

* This is a joint work with Professor Michael Perlman at the University of Washington.

One-sided tests for comparing multivariate responses

Examples:

- **Clinical trials with multiple endpoints.** Treatment effects may be measured by both efficacy and toxicity. Treatment A is better than Treatment B if all components of its mean responses are larger (say).
- **Selection and ranking problems.** Find the largest element of several normal means (Gupta 1965; Hsu 1996). E.g., construct a confidence set for the index of the largest mean $\equiv$ simultaneously test several normal mean differences (closely related to multiple comparisons with the unknown best).

Example I: Finding True Phylogenies

This is a selection and ranking application. In this dataset

- There are 6 mammal species (human, harbor seal, cow, rabbit, mouse, and opossum).
- We consider $p = 5$ most probable phylogenies, and want to find the *true phylogeny* — the hypothetical tree of the evolution history.
- Each phylogeny can be represented as a probabilistic model M_i.

Example I: Finding True Phylogenies

- We assume $Y_i \equiv$ maximized loglikelihood(M_i) to be approximately normal.
- Let $E(Y_i) = \mu_i$, and $\mu_{jk} = \mu_j - \mu_k = E(Y_j - Y_k)$, $j, k = 0, 1, \ldots, p$.
- We want to construct a $(1 - \alpha) \times 100\%$ confidence set for *the true phylogeny* – the one with the largest likelihood.

Example I: Finding True Phylogenies

The problem is equivalent to testing

$$H_0^{(k)} : \max_{j \neq k} \mu_{jk} \equiv \max_{j \neq k} (\mu_j - \mu_k) \leq 0$$

versus $H_1^{(k)}$: not $H_0^{(k)}$, for each k, $\quad k = 0, 1, \ldots, p$.

We then determine the indices k for which $H_0^{(k)}$ is not rejected at level α, and obtain a $(1 - \alpha) \times 100\%$ confidence set for *the true phylogeny*.

Example II: A Longitudinal Study

This is an example on testing *simple order* hypothesis. We consider

- a longitudinal study on parents whose children died by accident.
- *Research question*: does parents' depression change over time?
- Data were collected on 11 parents at 3 month, 6 month, and 18 month post-death.

Example II: A Longitudinal Study

- 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 want to test whether parents' depression decreases over time, i.e., test

$$H_0 : \mu_3 \leq \mu_2 \leq \mu_1 \qquad \text{versus} \qquad H_1 :\ \text{not } H_0$$

H_0 is a simple order hypothesis.

General Case

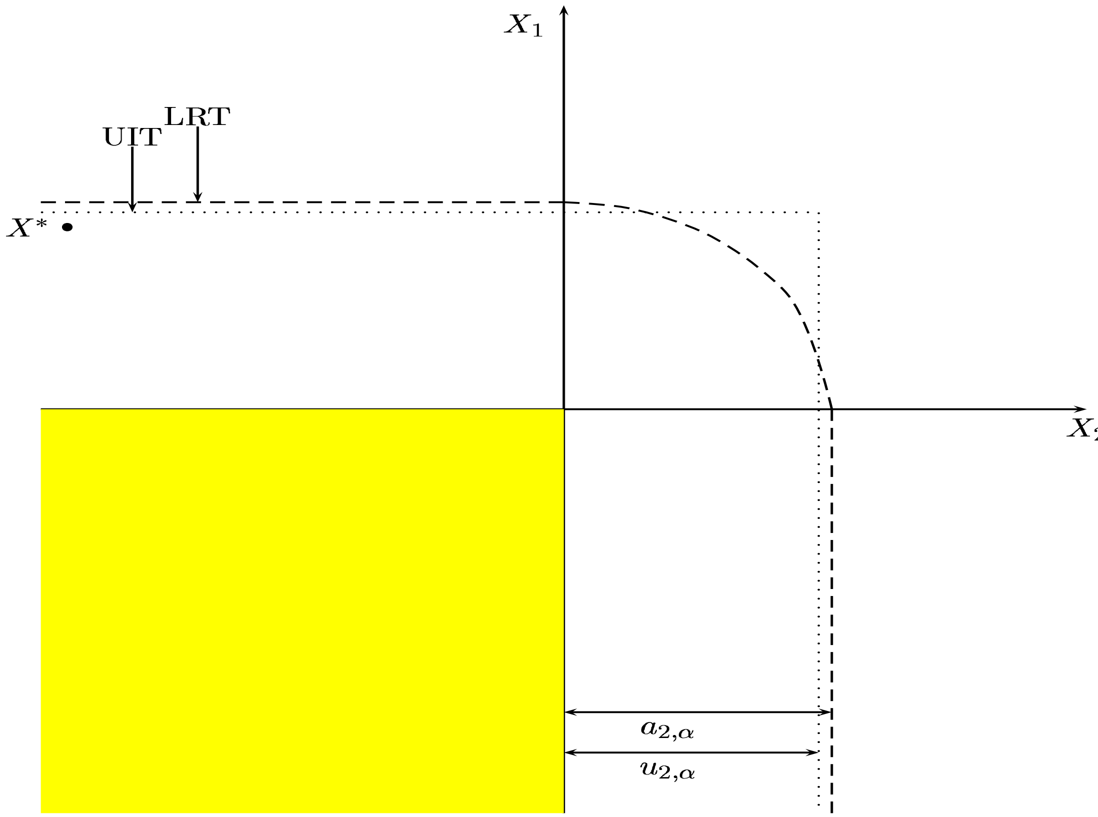
- Let $X \sim N(\mu, \Sigma)$ (μ and Σ unknown). Consider testing

  $$H_0 : \text{max}\{\mu_1, \cdots, \mu_p\} \leq 0, \quad \text{vs.} \quad H_1 : \text{max}\{\mu_1, \cdots, \mu_p\} > 0.$$

- Hotelling T^2 test may be undesirable since it fails to incorporate the constraints on the parameter spaces.

- *Commonly used tests*: likelihood ratio test (LRT), union-intersection test (UIT).

- *Problem with LRT and UIT:* they may exhibit anomalous behavior since they are unable to adapt to the *varying dimensionalities* of the boundary of H_0.



X_1
LRT
UIT
X^*
X_2
$a_{2,\alpha}$
$u_{2,\alpha}$

Anomalies of the LRT and UIT

Assume $\Sigma = I$ for simplicity. The size α LRT *accepts* H_0 iff

$$\|X - \mathcal{N}^p\|^2 \equiv (X_1^+)^2 + \cdots + (X_p^+)^2 \leq a_{p,\alpha}^2, \qquad (1)$$

where $X_i^+ \equiv \max(0, X_i)$ and $a_{p,\alpha}^2$ is a critical value.

The size α UIT *accepts* H_0 iff

$$\max(X_1, \ldots, X_p) \leq u_{p,\alpha}, \qquad (2)$$

where $u_{p,\alpha} = \Phi^{-1}(\sqrt[p]{1-\alpha}\,)$.

Anomalies of LRT and UIT: an example

Suppose $p = 2$ and $\alpha = 0.05$.

- The LRT rejects H_0 if $[(X_1^+)^2 + (X_2^+)^2]^{1/2} > 2.05$.
- The UIT rejects H_0 if max$(X_1, X_2) > 1.95$.

Now, if we observe $X^* = (1.8, -10)$. Then, neither LRT nor UIT reject H_0.

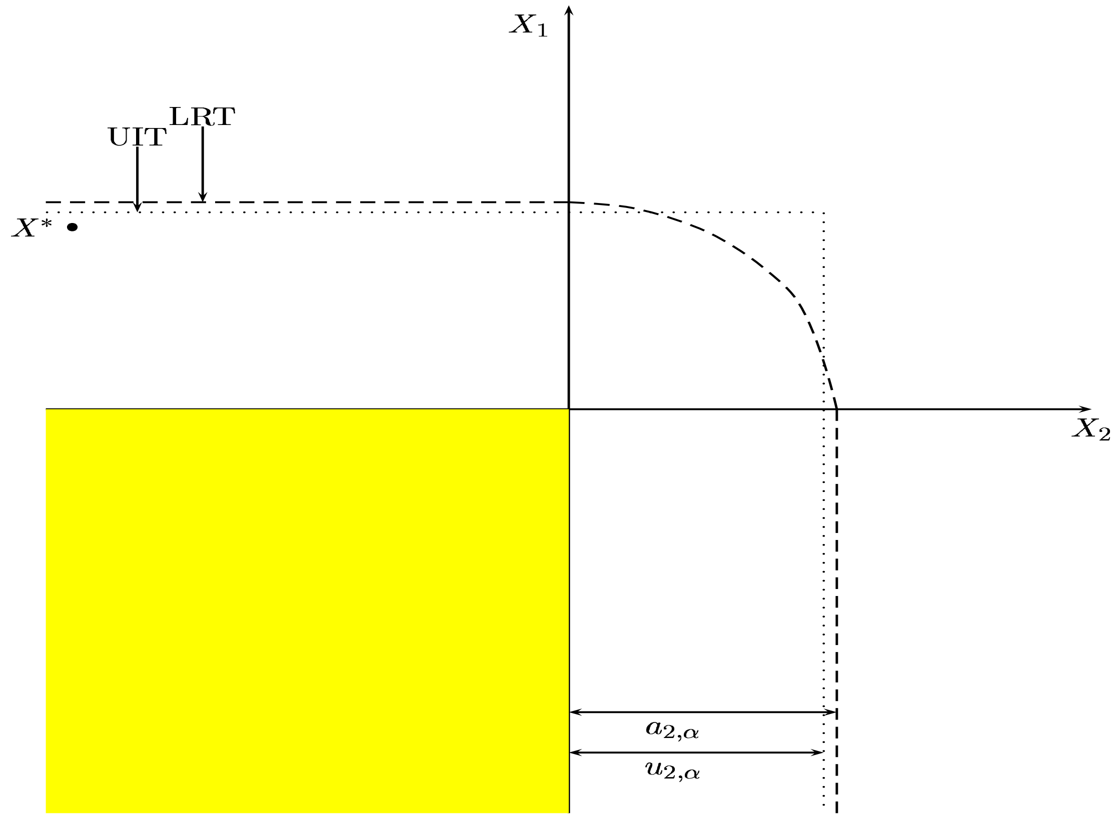
However, consider testing $H_{01} : \mu_1 \leq 0$ vs $H_{11} : \mu_1 > 0$ *individually*. H_{01} is clearly rejected. So H_0 should also be rejected since $H_0 \subset H_{01}$. **Contradiction!**

Anomalies of LRT and UIT

The anomalies of LRT and UIT become more emphatic as p increases.

In fact, for a sequence of alternatives $(\mu_1, \ldots, \mu_p)$ with μ_1 arbitrarily large but max$\{\mu_2, \cdots, \mu_p\} \to -\infty$ as $p \to \infty$, the powers of the LRT and UIT approach 0.

However, for such alternatives, any appropriate test procedure should have reasonable power to reject H_0.



X_1
LRT
UIT
X^*
X_2
$a_{2,\alpha}$
$u_{2,\alpha}$

Anomalies of LRT and UIT: Explanation

- The boundary of H_0 consists of a union of faces of varying dimensions (i.e., dimensions 0, $1, \ldots, p-1$).
- The LRT and UIT determine their critical values with reference to the face of *lowest* dimension. So they fail to adapt to the *varying dimensionalities* of the faces of H_0.
- Such contradictory behavior of the LRT and UIT also occur in other constrained multi-parameter testing problems.

A New Test

- We propose a new test which *adapts* to the varying dimensionalities in the boundary of H_0.
- The idea is to combine the p-values for testing the *individual faces* of H_0.
- Since a p-value is “self-weighting” according to the dimensionality of H_0, the new test adapts to the varying dimensionalities of the faces of H_0.
- The new test avoids the contradictory behavior of the LRT/UIT, so may better reflect the evidence provided by the data.

A New Test (Case I: $\Sigma = I$)

We *accepts* H_0 iff

$$(1 - 1_{\mathcal{N}^p}(X)) \sum_{i \in \sigma} X_i^2 \leq \tilde{a}^2_{|\sigma|,\alpha} \;\; and \;\; \max_{i \notin \sigma} X_i \leq 0 \qquad (3)$$

for at least one $\sigma \in \mathcal{S}^p$, where $\mathcal{S}^p = 2^{\{1,\ldots,p\}} \setminus \emptyset$, $\mathcal{N}^p \equiv \{(\mu_1, \ldots, \mu_p) : \mu_1 \leq 0, \ldots, \mu_p \leq 0\}$ is the nonpositive orthant in $\mathbf{R}^p$, and $\tilde{a}^2_{k,\alpha}$ is a critical value.

The above test is motivated by combining the p-values for testing the *individual faces* of H_0.

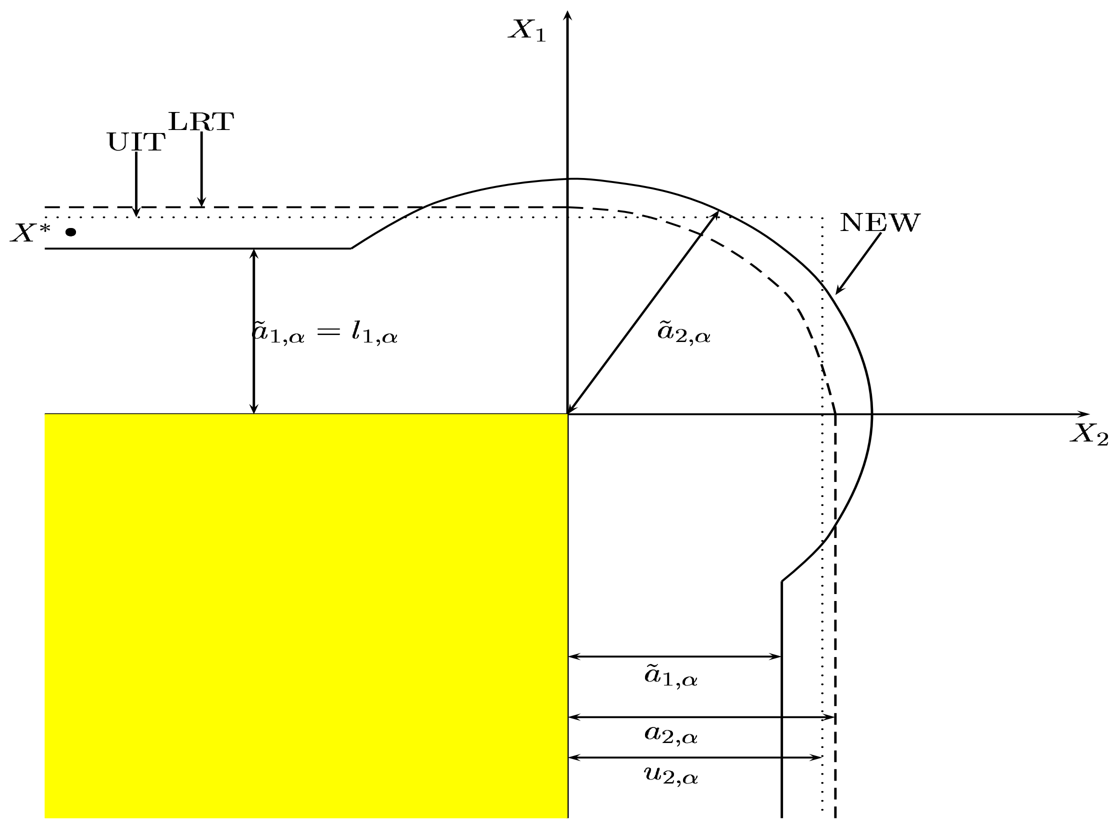


Figure 1. *Rejection/acceptance regions of various tests for (1) with Σ known ($\Sigma = I$).*
$p = 2,\ \alpha = 0.05,\ u_{2,\alpha} = 1.95,\ a_{2,\alpha} = 2.05,\ \tilde{a}_{1,\alpha} = 1.64,\ \tilde{a}_{2,\alpha} = 2.33.$
$l_{1,\alpha} = 1.64,\ l_{2,\alpha} = 2.19.$

A New Test (Case II: Σ unknown)

We *accept* H_0 iff

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

for at least one $\sigma \in \mathcal{S}^p$, where the critical values $a^*_{|\sigma|,\alpha}$ are given by

$$\begin{aligned} \alpha &= \frac{1}{2}\Pr\left[\frac{\chi^2_{p-1}}{\chi^2_{n_1+n_2-p}} > a^*_{p,\alpha}\right] + \frac{1}{2}\Pr\left[\frac{\chi^2_{p}}{\chi^2_{n_1+n_2-p-1}} > a^*_{p,\alpha}\right] \\ &\equiv \sup_{\mu \in \mathcal{N}^p, \Sigma > 0} \Pr_{\mu,\Sigma}[\, \|X - \mathcal{N}^p\|_S^2 > a^*_{p,\alpha}]. \end{aligned}$$

The New Tests

- The new tests are motivated by *combining the individual p-values* for testing the faces of H_0.
- The new tests better adapt to the varying dimensionalities of the boundaries of null parameter space.
- The new tests may be also more powerful than the LRT and UIT in many cases. The power advantage can be substantial (see simulation results).

Simulation

- We compare the new test (NEW) with the LRT and UIT via simulation.
- In all simulations, we have 5,000 iterations. We set nominal level $\alpha = 5\%,$ and sample sizes $n_1 = n_2 = 40$. We denote $(-1^4, 0.5) = (-1, -1, -1, -1, 0.5)$, etc.
- We consider several mean vectors μ and covariance matrices $\Sigma_1, \Sigma_2, \Sigma_3$. Each Σ_i is an intraclass correlation matrix with all diagonal elements $= 1$ and all off-diagonal elements $= \rho_i$, with $\rho_1 = 0$, $\rho_2 = 0.4$, $\rho_3 = 0.8$ respectively.

Table 1. Simulation results: sizes (type I error rates). Nominal level $\alpha = 5\%$*.*

Dimension	Mean μ	LRT	UIT	NEW	LRT	UIT	NEW	LRT	UIT	NEW
		$\Sigma = \Sigma_1$			$\Sigma = \Sigma_2$			$\Sigma = \Sigma_3$		
$p = 2$	$(0, 0)$	3.0	4.8	3.8	2.8	4.8	3.8	1.6	3.5	2.6
	$(-1, 0)$	1.1	2.6	5.0	1.2	2.8	5.1	1.0	2.5	4.8
	$(-5, 0)$	1.2	2.4	4.9	1.0	2.2	4.5	1.1	2.4	4.8
$p = 5$	$(0, 0^4)$	1.7	5.3	2.9	0.7	4.6	1.8	0.2	2.6	0.9
	$(-1, 0^4)$	0.8	4.0	3.1	0.6	3.2	2.0	0.2	2.3	1.3
	$(-1^2, 0^3)$	0.3	3.0	3.6	0.2	2.8	2.8	0.1	2.2	1.8
	$(-1^3, 0^2)$	0.2	2.1	4.2	0.1	1.5	3.0	0.1	1.4	2.6
	$(-1^4, 0)$	0.0	1.0	5.0	0.1	1.0	4.7	0.1	0.9	4.8

Table 2. Simulation results: powers comparison (in %).

Dimension	Mean μ	LRT	UIT	NEW	LRT	UIT	NEW	LRT	UIT	NEW
		$\Sigma = \Sigma_1$			$\Sigma = \Sigma_2$			$\Sigma = \Sigma_3$		
$p = 2$	$(0.2, 0.2)$	22	25	23	17	24	19	13	21	15
	$(-1, 0.3)$	17	26	37	16	25	36	17	27	38
	$(-5, 0.3)$	17	25	36	17	26	38	17	27	37
$p = 5$	$(0.1, 0.1^4)$	9	14	11	2	11	4	1	7	2
	$(-1, 0.5^4)$	94	91	97	53	79	68	28	63	46
	$(-1^2, 0.5^3)$	79	84	93	44	73	71	24	60	52
	$(-1^3, 0.5^2)$	50	70	85	32	64	73	21	55	61
	$(-1^4, 0.5)$	14	45	72	13	43	71	14	44	71

Simulation Results: Conclusions

- The new test better adapts to the *varying dimensionalities* of the faces of H_0, so reduces the undesirable behavior of the LRT and UIT.
- The new test is approximately size α, is more nearly similar on the boundary of H_0, and is more nearly unbiased than the LRT and the UIT.
- Our preference for the new test is based *not* mainly on consideration of power and unbiasedness but rather on the fact that it better reflects *the evidence the data provides* regarding the competing hypotheses.

A Related Test

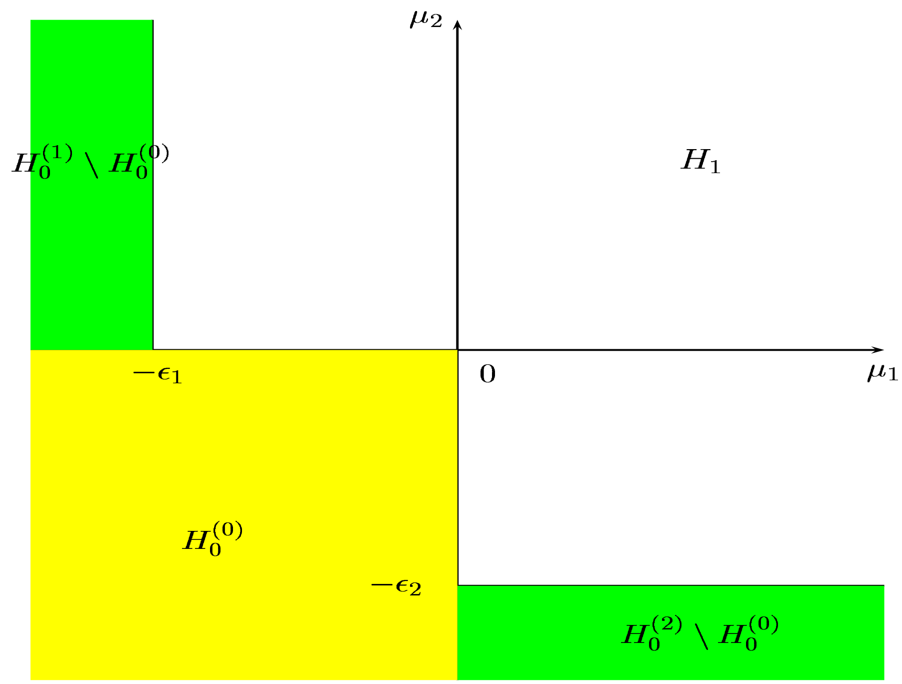
Sometimes it is more practical to assert that treatment 1 is preferred if it is superior for at least one of the endpoints and biologically "noninferior" for the remaining endpoints.

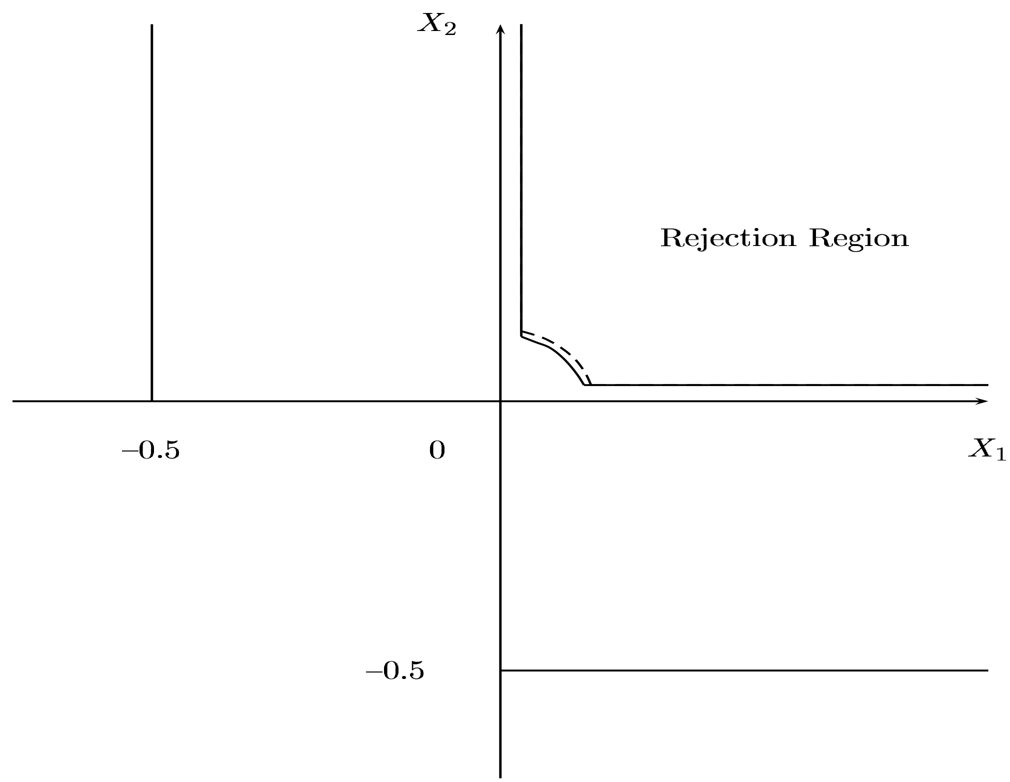
In other words, we want to test

$$H_0': \ \mu \in \Theta_0 \equiv \left\{ \max_{1 \le j \le p} \mu_j \le 0 \right\} \cup \left\{ \max_{1 \le j \le p} \mu_j > 0 \text{ and } \mu_j \le -\epsilon_j \text{ for some } j \right\}, \qquad (4)$$

versus H_1' : not H_0',

where ϵ_j's are pre-specified positive numbers. Again assume that Σ is unknown.


μ_2
$H_0^{(1)} \setminus H_0^{(0)}$
H_1
$-\epsilon_1$
0
μ_1
$H_0^{(0)}$
$-\epsilon_2$
$H_0^{(2)} \setminus H_0^{(0)}$


X_2
Rejection Region
-0.5
0
X_1
-0.5

A New Test for the Related Test

Noted that H_0' is a *union* of

$$H_0 : \mu \in \mathcal{N}^p \quad \text{and} \quad H_0^{(j)} : \mu_j \leq -\epsilon_j,\ j = 1, \ldots, p,$$

so an intersection-union test (IUT) is appropriate.

We can combine the new test for H_0 with the standard t-test for each $H_0^{(j)},\ j = 1, \ldots, p$, using the IUT idea, to obtain an overall NEW test.

A New Test for the Related Test

Since the new test for H_0 and each t-test for $H_0^{(j)}$ adapt to the varying dimensionality, the overall NEW test also adapts to the varying dimensionality of the boundary of H_0.

Simulation results show that the NEW test performs better than existing tests for this testing problem.

Testing the Simple-Order Restriction

Let $X \equiv (X_1, \ldots, X_p) \sim N(\mu, \Sigma)$. Consider testing the *simple-order*:

$$\bar{H}_0 : \mu_1 \leq \mu_2 \cdots \leq \mu_p \qquad \text{vs} \qquad \bar{H}_1 : \text{ not } \bar{H}_0. \tag{5}$$

This test is very common in practice. Denote $\mathcal{C}^p = \{\mu \equiv (\mu_1, \ldots, \mu_p) \mid \mu_1 \leq \cdots \leq \mu_p\}$.

The boundary of $\bar{H}_0$ is again a union of faces of *varying dimensionalities.* So the commonly used LRT may be undesirable.

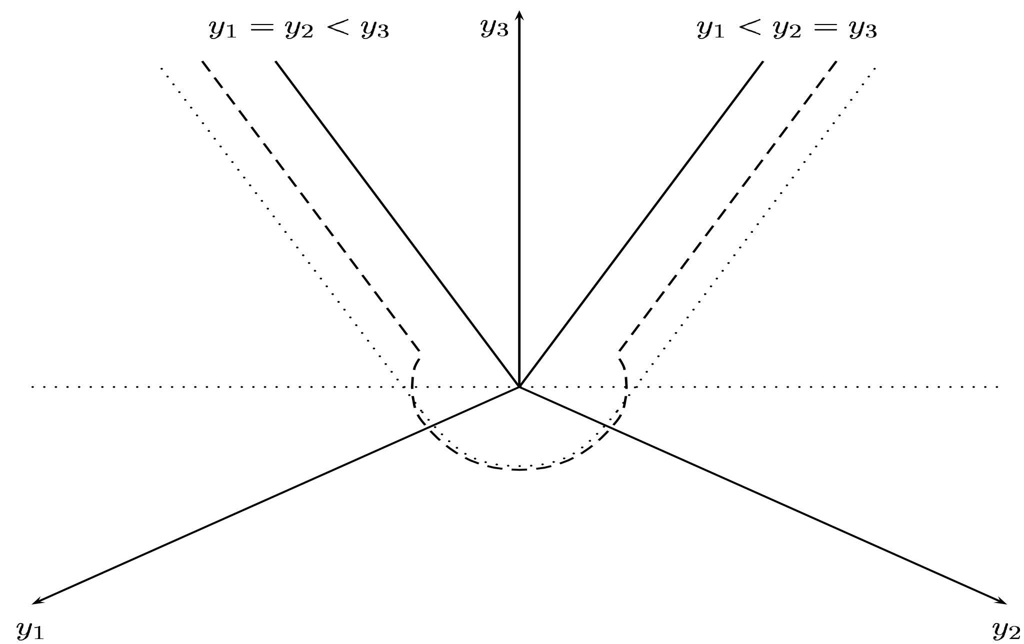


Figure 2. *Rejection/acceptance regions of the LRT and the new test PW9 for (19) with Σ known ($\Sigma = I$). LRT: dotted line, PW9: dashed line*

Testing the Simple-Order Restriction

The LRT *accepts* $\bar{H}_0$ iff

$$\|X - \mathcal{C}^p\|^2_{\hat{\Sigma}} \leq d^{*2}_{p,\alpha}. \tag{6}$$

Again, the LRT fails to adapt the *varying dimensionalities* of the faces of H_0.

A new test: *accepts* $\bar{H}_0$ iff

$$[1 - 1_{\mathcal{C}^p}(X)] \cdot \|X - L_\tau\|^2_{\hat{\Sigma}} \leq d^{*2}_{|\tau|,\alpha} \;\; and \;\; \pi_{\hat{\Sigma}}(X, L_\tau) \in \mathcal{C}^p$$

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

Testing the Simple-Order Restriction

The new test is obtained by combining *individual p-values* associated with testing each face of $\bar{H}_0$ (each individual test is a LRT).

Thus, unlike the LRT, the new test should adapt to the varying dimensionalities since a p-value is "self-weighting" according to the dimensionality of H_0, so the new test should better reflect the evidence provided by the data.

A Simulation Study

- We consider the cases of $p = 3$ and $p = 5$.
- Four covariance matrices $\Sigma_i, i = 1, 2, 3, 4$. Each covariance matrix has diagonal elements being all 1 and off-diagonal elements being 0.4, 0.8, –0.4, and –0.8 respectively.
- Sample sizes $n_1 = n_2 = 40$.
- 5,000 iterations.
- Nominal level $\alpha = 0.05$.

Table 5. Simulation results: Type I Error Rates (in %). Nominal level $\alpha = 5\%$.

Dimension p	Mean μ	LRT	NEW	LRT	NEW	LRT	NEW	LRT	NEW
		$\Sigma = \Sigma_1$		$\Sigma = \Sigma_2$		$\Sigma = \Sigma_3$		$\Sigma = \Sigma_4$	
$p = 3$	$(0, 0, 0)$	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
$p = 5$	$(0^4, 0)$	1.6	3.9	1.5	4.0	1.6	4.4	1.6	4.8
	$(0^4, 1)$	0.7	4.4	0.6	4.1	0.6	4.1	0.4	3.6
	$(0^3, 1^2)$	0.7	4.1	0.4	3.3	0.5	4.0	0.4	4.0
	$(0^2, 1^3)$	0.6	3.7	0.6	3.5	0.6	3.8	0.4	4.0
	$(0, 1^4)$	0.6	4.9	0.9	4.3	0.7	4.3	1.4	4.4

Table 6. Simulation Results: Power Comparison (in %)

Dimension p	Mean μ	LRT	NEW	LRT	NEW	LRT	NEW	LRT	NEW
		$\Sigma = \Sigma_1$		$\Sigma = \Sigma_2$		$\Sigma = \Sigma_3$		$\Sigma = \Sigma_4$	
$p = 3$	$(0.5, 0, 0)$	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
$p = 5$	$(0.4, 0^4)$	14	24	26	38	76	84	11	19
	$(0, 0.4, 0^3)$	11	21	19	33	66	81	8	17
	$(0, 0.4^2, 0^2)$	14	28	26	45	82	93	11	22
	$(0.3^4, -0.3)$	36	48	59	71	99	100	29	42

Simulation Results

- The new test adapts to the *varying dimensionality* of the faces of $\bar{H}_0$, while the LRT does not.
- The new test is more nearly similar and less biased than the LRT, and is often substantially more powerful than the LRT.
- **Our preference for the new test is based mainly on its better representing the evidence provided by the data (i.e., better adaption to the dimensionalities), rather than size/power.**

Example I (cont.): Finding True Phylogenies

- We consider again the 5 most probable (true) phylogenies.
- We test each $H_0^{(k)}$: $\max_{j \neq k}(\mu_j - \mu_k) \leq 0$ versus $H_1^{(k)}$: not $H_0^{(k)}$, where $\mu_j = E(Y_j)$.
- Let $\Delta Y_k = \max_{j \neq k}(Y_j - Y_k),\ k = 0, \cdots, 4$, where Y_k is the maximized likelihood for the k-th phylogeny. The data give

$$(\Delta Y_0, \cdots, \Delta Y_4) = (0.0, 19.5, 22.7, 29.1, 33.6).$$

- At 80% confidence level, the confidence sets are: LRT leads to $\{1, 2, 3, 4, 5\}$, the UIT leads to $\{1, 2, 3\}$, and the NEW test leads to $\{1, 2, 3, 4\}$.

Example II (cont.): A Longitudinal Study

- Let Y_i be the depression at time t_i. Let $\mu_i = E(Y_i)$. We want to test $H_0 : \mu_3 \leq \mu_2 \leq \mu_1$ versus H_1 : not H_0.
- The sample mean and sample covariance are

$$(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3) = (0.60, 0.97, 0.73), \qquad \hat{\Sigma} = \begin{pmatrix} 0.32 & 0.63 & 0.40 \\ 0.63 & 1.51 & 0.92 \\ 0.40 & 0.92 & 0.57 \end{pmatrix}.$$

- At the 5% level, the LRT fails to reject H_0, while the NEW test rejects H_0. The new test should be more reliable, suggesting that depression does not decrease over time.

Conclusions

- For testing problems where the parameter spaces have varying dimensionalities, the LRT and UIT fail to adapt to this varying dimensionality and thus may produce misleading results.
- The proposed new tests adjust the varying dimensionality, so better reflect the evidence provided by the data.
- The new tests are obtained by combining *individual p-values* associated with testing individual faces of the null space.
- Simulations show that the new tests are better than the LRT and UIT.

Acknowledgment

We thank Professor Ajit Tamhane for very helpful comments and suggestions.