On the Validity of the Likelihood Ratio and Maximum Likelihood Methods*

Michael D. Perlman[1] and Lang Wu[2]

[1] *Department of Statistics, University of Washington, Seattle, WA 98195, USA*

[2] *Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada*

May 5, 2002

Abstract

When the null or alternative hypothesis of a statistical testing problem is a union of finitely many regions of varying dimensionality, the likelihood ratio test is statistically inappropriate. Its inappropriateness is revealed not by its performance under the Neyman-Pearson criterion but by the fact that it yields incorrect inferences in certain regions of the sample space due to its inability to adapt to the differing dimensions in the composite hypothesis. Maximum likelihood estimators and associated model selection procedures also are inappropriate for such composite models. Tests and estimators based on the p-values associated with each of the regions that constitute the composite model are more appropriate for this geometry. Similar issues arise when the boundary of the null hypothesis is a union of finitely many regions of varying dimensionality.

*Corresponding author: Michael D. Perlman, Department of Statistics, University of Washington, Seattle, WA 98195. *email: michael@ms.washington.edu.*

KEY WORDS: Non-nested hypotheses, varying dimensionality, likelihood ratio test, intersection-union test, union-intersection test, model selection, maximum likelihood estimator, preliminary-test estimator, p-value, multivariate one-sided hypothesis, multivariate dependence models.

1 Introduction

Numerous examples of allegedly inferior likelihood ratio tests (LRT) have appeared in the past two decades, including those in Berger (1989), Brown *et al* (1997), Cohen *et al* (2000), Cohen and Sackrowitz (1998), Liu (2000), Liu and Berger (1995), McDermott and Wang (2002), Mendendez *et al* (1991, 1992), Mukerjee and Tu (1995), Munk (1999), Tsuda (2000), Wang (1997), Wang *et al* (1999), Warrack and Robertson (1984). These examples involve relatively simple testing problems, most based on univariate or multivariate normal observations. Perlman and Wu (1999, 2002) demonstrated that in these examples the allegations against the LRT are incorrect; the fault lies either with the Neyman-Pearson criterion itself (most powerful size α test), or else with an incorrect interpretation of order-restricted testing problems.

In apparently similar and equally simple examples, such as that presented by Berger and Sinclair (1984), however, the LRTs are indeed statistically inappropriate. In Section 2 we note that their inappropriateness stems not from inferior performance under the NP criterion, as asserted by Berger and Sinclair, but rather from the geometry of the statistical models in these examples, namely, a null or alternative hypothesis that is a union of finitely many pairwise non-nested regions (here, linear subspaces) of varying dimensionality. Tests, estimates, and model selection procedures based on maximum likelihood (Section 3) are unable to adapt to the varying dimensions in the composite hypothesis. However, the p-values associated with the various regions that constitute the composite hypothesis can be combined, often via the intersection-union or union-intersection methods, to construct statistical procedures that are more appropriate for this geometry.

Similar issues arise for model selection with partially non-nested models (Section 4) or when testing a multivariate one-sided null hypothesis whose boundary is a union of regions of varying dimensionality (Section 6). This is illustrated by an example of Shimodaira (2000a) in which the null hypothesis is the nonpositive orthant in $\mathbf{R}^p$. Again the LRT is unable to adapt to the differing dimensions in the models, but the p-values associated with the various regions that constitute the composite model can be combined to yield reasonable tests, estimates, and model selection procedures.

Our examples involve elementary normal-theory models with fixed (non-asymptotic) sample size $n = 1$ and fixed dimension p sometimes assumed to be large. Our viewpoint is neither frequentist nor Bayesian, but instead is based simply on model significance as measured by p-values.

2 Testing a Hypothesis with Varying Dimensionalities

2.1 A composite null hypothesis; the LRT vs. the IUT

The inappropriateness of the LRT for a null hypothesis composed of finitely many regions of varying dimensionality is readily illustrated by the following elementary normal-theory example of Berger and Sinclair (1984). For simplicity, we assume that the variances are known and equal. Based on the observation

$$X \equiv (X_1, X_2, \ldots, X_p) \sim N_p(\mu \equiv (\mu_1, \mu_2, \ldots, \mu_p), I_p) \qquad (p \geq 3), \tag{1}$$

test

$$H_0 : \mu \in L_1 \cup L_2 \qquad \text{vs.} \qquad H : \mu \in \mathbf{R}^p \setminus (L_1 \cup L_2), \tag{2}$$

where

$$L_1 := \{\mu \mid \mu_1 = 0\}, \tag{3}$$
$$L_2 := \{\mu \mid \mu_2 = 0, \ldots, \mu_p = 0\}, \tag{4}$$

are orthogonal linear subspaces of dimensions $p-1$ and 1, respectively (see Figure 1). Equivalently, test

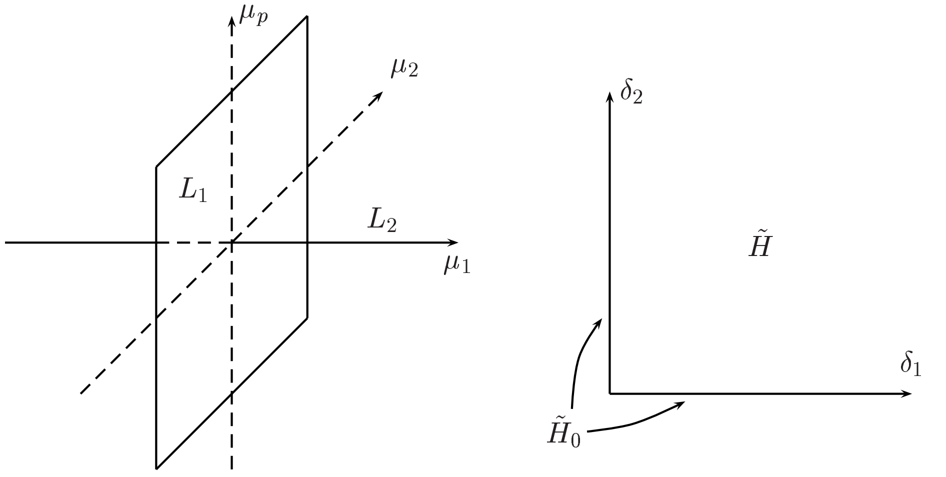
$$\tilde{H}_0 : \delta_1 = 0 \text{ or } \delta_2 = 0 \qquad \text{vs.} \qquad \tilde{H} : \delta_1 > 0 \text{ and } \delta_2 > 0, \tag{5}$$

where

$$\begin{aligned} \delta_1 &:= \|\mu - L_1\|^2 \equiv \min_{\nu \in L_1} \|\mu - \nu\|^2 = \mu_1^2, \\ \delta_2 &:= \|\mu - L_2\|^2 \equiv \min_{\nu \in L_2} \|\mu - \nu\|^2 = \mu_2^2 + \cdots + \mu_p^2. \end{aligned}$$

Define

$$Y_1 := \|X - L_1\|^2 = X_1^2 \sim \chi_1^2(\delta_1), \tag{6}$$
$$Y_2 := \|X - L_2\|^2 = X_2^2 + \cdots + X_p^2 \sim \chi_{p-1}^2(\delta_2), \tag{7}$$



(a) $H_0 : \mu \in L_1 \cup L_2$ (b) $\tilde{H}_0 : \delta_1 = 0$ or $\delta_2 = 0$

Figure 1. *Two representations of the null hypothesis $H_0 \equiv \tilde{H}_0$ in (2) $\equiv$ (5).*

where $\chi^2_k(\delta)$ denotes a noncentral chi-square random variable with noncentrality parameter $\delta \geq 0$ and k degrees of freedom; we often write χ^2_k for $\chi^2_k(0)$. Note that Y_1 and Y_2 are independent. The size α LRT for (2) $\equiv$ (5) rejects H_0 iff

$$\|X - (L_1 \cup L_2)\|^2 \equiv \min(Y_1, Y_2) > \chi^2_{p-1,\alpha}. \tag{8}$$

Berger and Sinclair (1984) noted that the LRT is not similar on the boundary of H_0 and therefore biased; the least favorable distribution is not attained in H_0 but is only approached by $N_p((\mu_1, 0, \ldots, 0), I_p)$ as $\mu_1 \to \infty$. (In this context, the least favorable distribution is that distribution in H_0 that is most difficult to distinguish from H.) Instead, they proposed the size α intersection-union test (IUT), which here rejects H_0 iff

$$\min\left(\frac{Y_1}{\chi^2_{1,\alpha}}, \frac{Y_2}{\chi^2_{p-1,\alpha}}\right) > 1. \tag{9}$$

They noted that this test is also size α and that its rejection region properly contains that of the LRT (see Figure 2), hence the IUT is less biased and everywhere more powerful than the LRT. Berger and Sinclair declared the LRT (8) to be inferior on this basis.

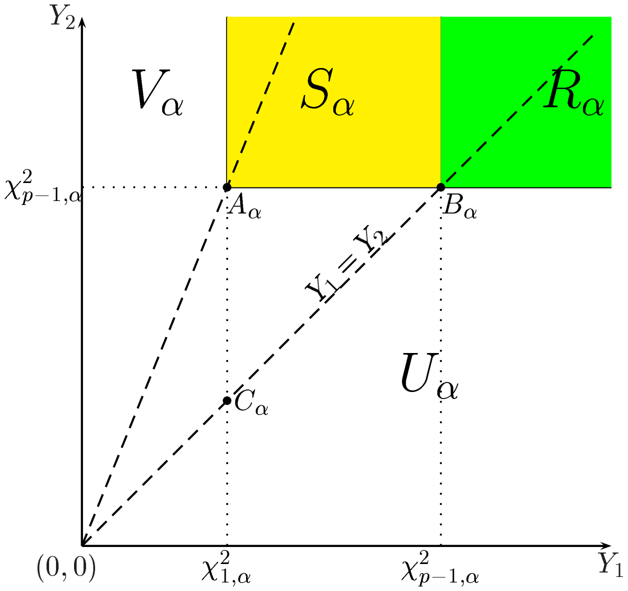


Figure 2. *The LRT (resp., IUT) for (2) ≡ (5) has rejection region* R_α $(R_\alpha \cup S_\alpha)$
$(p = 5,\ \alpha = 0.05,\ \chi^2_{1,\alpha} = 3.84,\ \chi^2_{p-1,\alpha} = 9.49)$.

Perlman and Wu (1999, Rejoinder) agreed with Berger and Sinclair that for this problem the LRT is inappropriate for inference, but *not* on the basis of its performance under the NP criterion. Perlman and Wu had already demonstrated the unreliability of the NP criterion for testing problems with composite null hypotheses by exhibiting tests that are everywhere more powerful than the LRTs of the same size and less biased or unbiased, but which can lead to absurd inferences. Instead, the inappropriateness of the LRT for (2) ≡ (5) has a different basis.

For any sample point in S_α (the shaded region in Figure 2), the LRT fails to reject $H_0 : \mu \in L_1 \cup L_2$ even though both component hypotheses $\mu \in L_1$ and $\mu \in L_2$ are rejected individually. In fact, S_α contains points such that $Y_1 \approx \chi^2_{p-1,\alpha} = \mathrm{O}(p)$ and $Y_2 \gg \chi^2_{p-1,\alpha}$, so *both component hypotheses are resoundingly rejected, yet the LRT fails to reject* H_0. The IUT avoids this strongly contradictory behavior.

This demonstrates that the LRT is much too conservative for this problem, especially when p is large. The LRT fails to account for the fact that the null hypothesis H_0 is the union of subspaces of different dimensions. Instead, it acts as if each subspace has the *minimum* dimension, and therefore is too conservative.

2.2 A composite alternative hypothesis; the LRT vs. the UIT

The inappropriateness of the LRT for an alternative hypothesis composed of finitely many regions of varying dimensionality is illustrated by an equally elementary normal-theory example. (We again assume that the variances are known and equal.) Based on the observation $X \sim N_p(\mu, I_p)$ as in (1) with $p \geq 3$, test

$$H_0' : \mu = 0 \qquad \text{vs.} \qquad H' : \mu \in (L_1 \cup L_2) \setminus \{0\}, \tag{10}$$

where L_1 and L_2 are as in (3) and (4). Equivalently, test

$$\tilde{H}_0' : \eta_1 = \eta_2 = 0 \qquad \text{vs.} \qquad \tilde{H}' : \eta_1 > 0, \eta_2 = 0 \text{ or } \eta_1 = 0, \eta_2 > 0, \tag{11}$$

where

$$\begin{aligned} \eta_1 &:= \|\mu_{L_1}\|^2 = \mu_2^2 + \cdots + \mu_p^2, \\ \eta_2 &:= \|\mu_{L_2}\|^2 = \mu_1^2, \end{aligned}$$

and $\mu_{L_1} \equiv (0, \mu_2, \ldots, \mu_p)$ and $\mu_{L_2} \equiv (\mu_1, 0, \ldots, 0)$ are the projections of μ onto L_1 and L_2 respectively.

Define

$$W_1 := \|X_{L_1}\|^2 = X_2^2 + \cdots + X_p^2 \sim \chi^2_{p-1}(\eta_1), \tag{12}$$

$$W_2 := \|X_{L_2}\|^2 = X_1^2 \sim \chi^2_1(\eta_2), \tag{13}$$

where $X_{L_1} \equiv (0, X_2, \ldots, X_p)$ and $X_{L_2} \equiv (X_1, 0, \ldots, 0)$ are the projections of X onto L_1 and L_2 respectively. The size α LRT for (10) $\equiv$ (11) rejects H_0' iff

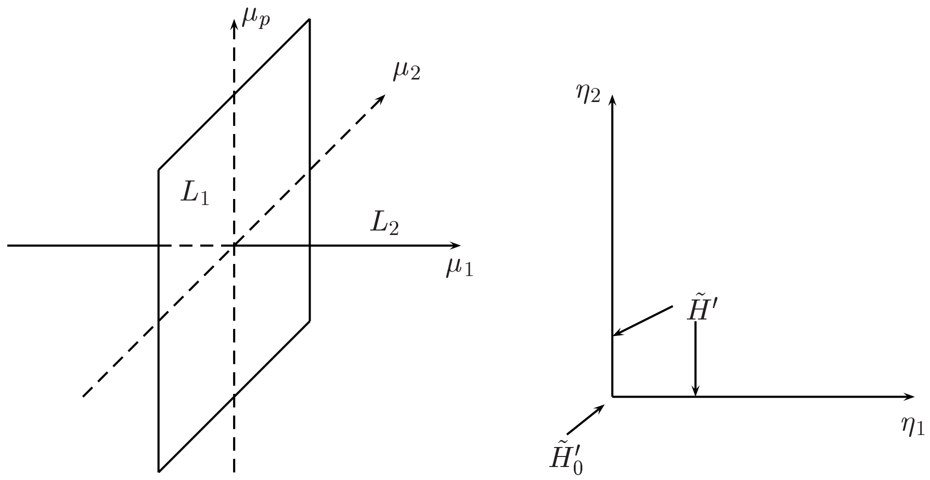
$$\max(W_1, W_2) > c_{p,\alpha}, \tag{14}$$

where, because W_1 and W_2 are independent, $c_{p,\alpha} > 0$ satisfies

$$1 - \alpha = P[\chi^2_1 \leq c_{p,\alpha}]\, P[\chi^2_{p-1} \leq c_{p,\alpha}]. \tag{15}$$

The size α union-intersection test (UIT) rejects H_0' iff

$$\max\Bigg(\frac{W_1}{\chi^2_{1,1-\sqrt{1-\alpha}}}, \frac{W_2}{\chi^2_{p-1,1-\sqrt{1-\alpha}}}\ biggr) > 1. \tag{16}$$



(a) $H' : \mu \in (L_1 \cup L_2) \setminus \{0\}$ (b) $\tilde{H}' : \eta_1 > 0, \eta_2 = 0$ or $\eta_1 = 0, \eta_2 > 0$

Figure 3. *Two representations of the alternative hypothesis $H' \equiv \tilde{H}'$ in (10) $\equiv$ (11).*

Because the acceptance regions of the LRT and the UIT are convex in X, the theorem of Stein (1956) on admissibility of such tests in exponential families can be applied to show that both the LRT and the UIT are admissible - neither dominates the other in terms of power. Nonetheless, in the absence of prior weights for L_1 and L_2 under H', we now argue that the UIT is preferable to the LRT for this problem, especially for large p.

It follows from (15) (see Appendix) that for large p,

$$\chi^2_{1,\alpha} < \chi^2_{1,1-\sqrt{1-\alpha}} \ll \chi^2_{p-1,\alpha} < c_{p,\alpha} < \chi^2_{p-1,1-\sqrt{1-\alpha}}. \tag{17}$$

For any sample point $(W_1, W_2) \in Q_\alpha$ (the shaded region in Figure 4), the size α LRT fails to reject $H'_0 : \mu = 0$ in favor of $H' : \mu \in (L_1 \cup L_2) \setminus \{0\}$ even though $\mu = 0$ is rejected in favor of *both* $\mu \in L_1$ and $\mu \in L_2$ individually. In fact, Q_α contains points such that $W_1 \approx c_{p,\alpha} \gg \chi^2_{1,\alpha}$, which provides overwhelming evidence against $\mu = 0$ in favor of $\mu \in L_1$ if p is large, yet the LRT fails to reject $\mu = 0$. Here the LRT fails to account for the fact that the alternative H' is the union of subspaces of different dimensions. Instead, it acts as if each subspace in H' has the *maximum* dimension, and therefore is too conservative. By contrast, the UIT avoids this contradictory behavior.

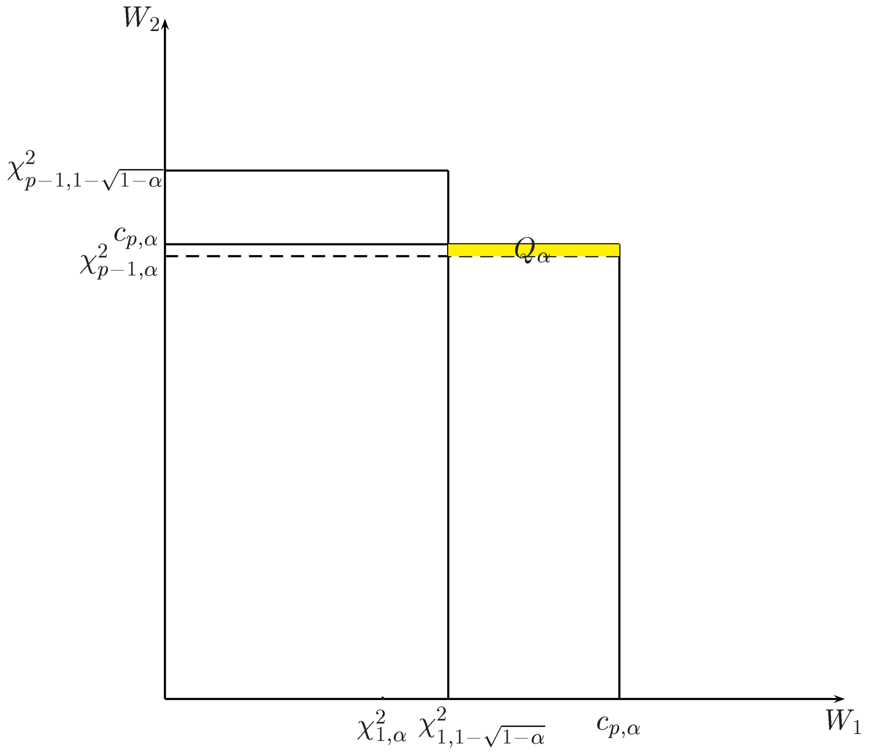


Figure 4. *The acceptance regions for the LRT (14) and UIT (16) for (10) ≡ (11)* ($p = 4$, $\alpha = 0.05$, $\chi^2_{p-1,\alpha} = 7.81$, $\chi^2_{p-1,1-\sqrt{1-\alpha}} = 9.32$, $\chi^2_{1,\alpha} = 3.84$, $\chi^2_{1,1-\sqrt{1-\alpha}} = 5.00$, $c_{p,\alpha} = 8.02$).

3 Estimation and model selection for $\{L_1, L_2\}$

The anomalous behaviors of the LRTs for (2) ≡ (5) and (10) ≡ (11) cast doubt on the behavior of the MLE $\hat{\mu}$ under the model $\mu \in L_1 \cup L_2$. Estimation of μ under $L_1 \cup L_2$ can be thought of as model selection (L_1 or L_2) followed by estimation within the chosen submodel. It can be seen from (14) that $\hat{\mu}(X)$ is just the projection of X onto the nearer of L_1 or L_2 as determined by the minimum of the squared residual distances $\|X - L_i\|^2$ from X to these two subspaces; that is,

$$\hat{\mu}(X) = \begin{cases} X_{L_1} := (0, X_2, \ldots, X_p), & \text{if } Y_1 < Y_2; \\ X_{L_2} := (X_1, 0, \ldots, 0), & \text{if } Y_1 > Y_2. \end{cases} \tag{18}$$

(Either projection X_{L_1} or X_{L_2} may be selected if $Y_1 = Y_2$, an event of probability 0.) The associated model selection procedure is, therefore,

$$\hat{L}(X) = \begin{cases} L_1, & \text{if } Y_1 < Y_2; \\ L_2, & \text{if } Y_1 > Y_2. \end{cases} \tag{19}$$

We call (19) the *maximum likelihood model selection procedure (MLMSP)*.

Define the regions (see Figure 2)

$$U_\alpha := \{(Y_1, Y_2) \mid Y_1 > \chi^2_{1,\alpha}, Y_2 < \chi^2_{p-1,\alpha},\} \tag{20}$$

$$V_\alpha := \{(Y_1, Y_2) \mid Y_1 < \chi^2_{1,\alpha}, Y_2 > \chi^2_{p-1,\alpha}\}. \tag{21}$$

When $(Y_1, Y_2) \in U_\alpha$, $\mu \in L_1$ is rejected and $\mu \in L_2$ accepted if tested individually at level α against $\mu \in \mathbf{R}^p$, so it can be argued that a model selection procedure should select L_2 in this case, and similarly should select L_1 when $(Y_1, Y_2) \in V_\alpha$. However, the MLMSP and MLE do not behave in this way: for X such that (Y_1, Y_2) lies in the triangle $A_\alpha B_\alpha C_\alpha$ in Figure 2, $\hat{L}(X) = L_1$ even though $(Y_1, Y_2) \in U_\alpha$. This behavior reflects the fact that the line $Y_1 = Y_2$, which delineates the two cases in (18) and (19), passes through the point B_α in Figure 2 and therefore fails to separate the two regions U_α and V_α.

This argument may not be entirely convincing, however, because the alternative $\mu \in \mathbf{R}^p$ includes regions not contained in the model $L_1 \cup L_2$. Nonetheless, our preference for the IUT in (9) suggests that, under $L_1 \cup L_2$, more satisfactory model selection procedures $\tilde{L}_\alpha$ and associated estimators $\tilde{\mu}_\alpha$ may be obtained by weighting the squared residual distances $\|X - L_i\|^2$ approximately in inverse proportion to their degrees of freedom, leading to the following model selection procedure and associated estimator:

$$\tilde{L}_\alpha(X) = \begin{cases} L_1, & \text{if } Y_1/\chi^2_{1,\alpha} < Y_2/\chi^2_{p-1,\alpha}; \\ L_2, & \text{if } Y_1/\chi^2_{1,\alpha} > Y_2/\chi^2_{p-1,\alpha}; \end{cases} \tag{22}$$

$$\tilde{\mu}_\alpha(X) = \begin{cases} X_{L_1}, & \text{if } Y_1/\chi^2_{1,\alpha} < Y_2/\chi^2_{p-1,\alpha}; \\ X_{L_2}, & \text{if } Y_1/\chi^2_{1,\alpha} > Y_2/\chi^2_{p-1,\alpha}; \end{cases} \tag{23}$$

The inequality $\frac{Y_1}{\chi^2_{1,\alpha}} < \frac{Y_2}{\chi^2_{p-1,\alpha}}$ suggests stronger evidence against $\mu \in L_2$ than against $\mu \in L_1$. Note that the line $\frac{Y_1}{\chi^2_{1,\alpha}} = \frac{Y_2}{\chi^2_{p-1,\alpha}}$, which delineates the two cases in (22) and (23), passes through the point A_α in Figure 2, so does separate U_α and V_α.

Because α plays no direct role in the estimation problem, however, we actually recommend a somewhat different model selection procedure $\bar{L}$ and estimator $\bar{\mu}$. These rely on a comparison of the p-values ($\equiv$ attained significance levels) $\pi_i \equiv \pi_i(Y_i)$ for testing $\mu_i \in L_i$ vs. $\mu \in \mathbf{R}^p \setminus L_i$, equivalently, for testing $\delta_i = 0$ vs. $\delta_i > 0$, based on Y_i, $i = 1, 2$. That is, for $i = 1, 2$ and $y \geq 0$,

$$\pi_i(y) = P_{\delta_i=0}[Y_i > y]. \tag{24}$$

Based on these p-values, we propose the following MSP $\bar{L}$ and estimator $\bar{\mu}$ under $L_1 \cup L_2$:

$$\bar{L}(X) = \begin{cases} L_1, & \text{if } \pi_1(Y_1) > \pi_2(Y_2); \\ L_2, & \text{if } \pi_1(Y_1) < \pi_2(Y_2); \end{cases} \tag{25}$$

$$\bar{\mu}(X) = \begin{cases} X_{L_1}, & \text{if } \pi_1(Y_1) > \pi_2(Y_2); \\ X_{L_2}, & \text{if } \pi_1(Y_1) < \pi_2(Y_2). \end{cases} \tag{26}$$

Because $\pi_i(y)$ is decreasing in y, $\bar{L}$ and $\bar{\mu}$ coincide with $\hat{L}$ and $\hat{\mu}$, respectively, if $\dim(L_1) = \dim(L_2)$ but not otherwise.

The estimator $\bar{\mu}$ is a significance-based *preliminary test ($\equiv$ pre-test) estimator* and $\bar{L}$ is called a significance-based *pre-test model selection procedure (PTMSP)*. Such estimators have occurred before in the literature, but usually in the context of Stein shrinkage estimation, where the models being compared are nested. A notable exception is Efron (1984).

Does the delineating boundary for $\bar{L}$ and $\bar{\mu}$, namely the curve determined by the equation $\pi_1(Y_1) = \pi_2(Y_2)$, separate the regions U_α and V_α? Recall that we criticized the MLMSP and MLE for their failure to satisfy this separation property. Because this curve is the graph of an increasing function, it would suffice to show that it passes through the point $A_\alpha \equiv (\chi^2_{1,\alpha}, \chi^2_{p-1,\alpha})$. It may appear unlikely that $\bar{\mu}$ satisfies this property for *all* $0 < \alpha < 1$, but an easy calculation (see Appendix) shows that the delineating boundary $\{\pi_1(Y_1) = \pi_2(Y_2)\}$ is exactly the locus $\{A_\alpha \mid 0 < \alpha < 1\}$, and that

$$\{(Y_1, Y_2) \mid \pi_1(Y_1) > \pi_2(Y_2)\} = \bigcup_{0<\alpha<1} U_\alpha, \tag{27}$$

$$\{(Y_1, Y_2) \mid \pi_1(Y_1) < \pi_2(Y_2)\} = \bigcup_{0<\alpha<1} V_\alpha, \tag{28}$$

so this boundary in fact separates U_α and V_α for every α! This is illustrated in Figure 5. Thus, $\bar{L}(X) = L_1$ if and only if $\mu \in L_1$ is accepted and $\mu \in L_2$ rejected for least one level $\alpha \in (0, 1)$, or equivalently, if there exists no $\alpha \in (0, 1)$ for which $\mu \in L_1$ is rejected and $\mu \in L_2$ accepted at level α (and similarly for $\bar{L}(X) = L_2$).

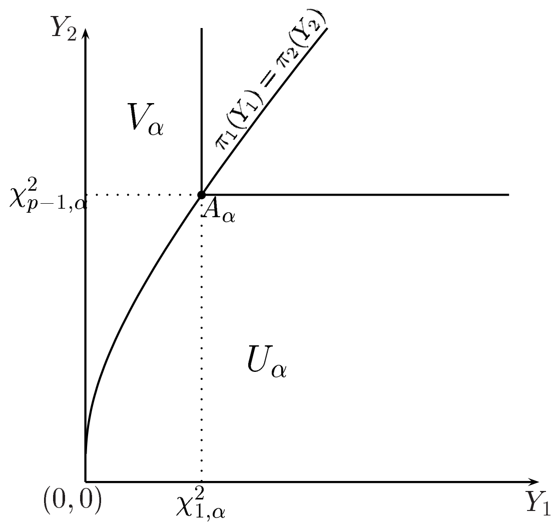


Figure 5. *The delineating boundary* $\{(Y_1, Y_2) \mid \pi_1(Y_1) = \pi_2(Y_2)\} \equiv \{A_\alpha \mid 0 < \alpha < 1\}$ *(*$p = 5,\ \alpha = 0.05,\ \chi^2_{1,\alpha} = 3.84,\ \chi^2_{p-1,\alpha} = 9.49$*).*

This seems to be a compelling argument in favor of $\bar{\mu}$ and $\bar{L}$. As already noted, however, because the parameter space is restricted here to $L_1 \cup L_2$, it may be inappropriate to base an estimator and model selection procedure on the p-values relevant for testing L_1 and L_2 individually against the full alternative $\mathbf{R}^p$. Instead, one might prefer a model selection procedure and pre-test estimator based on a p-value for testing L_1 against L_2 directly. The standard test statistic for this problem is the generalized LR statistic

$$\frac{\sup_{\mu \in L_2} f_\mu(X)}{\sup_{\mu \in L_1} f_\mu(X)} = \exp(Y_1 - Y_2), \tag{29}$$

where f_μ denotes the $N_p(\mu, I_p)$ density. This leads to the following model selection procedure and associated pre-test estimator under $L_1 \cup L_2$:

$$\check{L}_c(X) = \begin{cases} L_1, & \text{if } Y_1 - Y_2 < c; \\ L_2, & \text{if } Y_1 - Y_2 > c, \end{cases} \tag{30}$$

$$\check{\mu}_c(X) = \begin{cases} X_{L_1}, & \text{if } Y_1 - Y_2 < c; \\ X_{L_2}, & \text{if } Y_1 - Y_2 > c; \end{cases} \tag{31}$$

where $-\infty < c < \infty$.

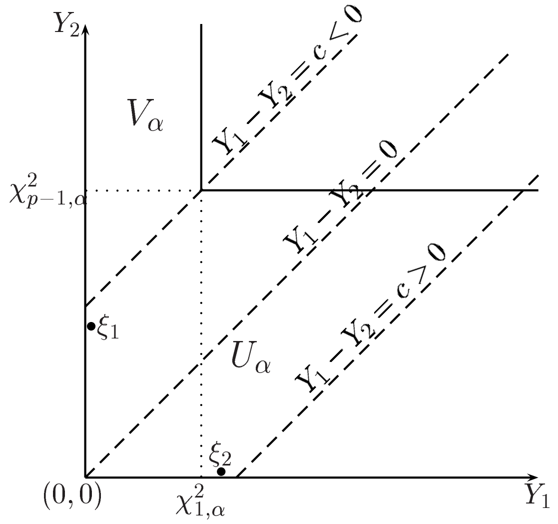


Figure 6. *The delineating boundaries of $\check{L}_c$ and $\check{\mu}_c$ for $c < 0$, $c = 0$, $c > 0$ ($p = 5,\ \alpha = 0.05,\ \chi^2_{1,\alpha} = 3.84,\ \chi^2_{p-1,\alpha} = 9.49$).*

Both $\check{L}_c$ and $\check{\mu}$ exhibit the same questionable behavior as $\hat{L}$ and $\hat{\mu}$: if $c \geq 0$ then the delineating boundary $Y_1 - Y_2 = c$ fails to separate U_α and V_α for any $\alpha \in (0,1)$, while if $c < 0$ this boundary separates U_α and V_α for only one α (see Figure 6). We have already conceded, however, that the regions U_α and V_α are directly relevant only for testing against the full alternative $\mathbf{R}^p$, so an estimator's failure to separate these regions may not be definitive under the model $L_1 \cup L_2$. More compelling is the fact that for $c > 0$ (resp., $c < 0$) the delineating boundary $Y_1 - Y_2 = c$ does not pass through the origin, so it is possible that $\frac{\|X - L_2\|}{\|X - L_1\|} \approx \infty$ (≈ 0) but $\check{L}_c(X) = L_2$ ($= L_1$). This occurs, for example, if $(Y_1, Y_2) \approx \xi_1$ ($\approx \xi_2$) in Figure 6.

Because this behavior does not occur when $c = 0$, in which case $\check{L}_0$ and $\check{\mu}_0$ reduce to $\hat{L}$ and $\hat{\mu}$ respectively, even this argument does not definitively condemn the MLMSP and the MLE. It is easy, however, to compare directly the probabilities of correct model selection for $\hat{L}$ and $\bar{L}$. For the former,

$$P_{L_1}[\hat{L}(X) = L_1] = P_{\delta_1=0}[Y_1 < Y_2] = P[\chi^2_1(0) < \chi^2_{p-1}(\delta_2)] \quad > \frac{1}{2}, \tag{32}$$

$$P_{L_2}[\hat{L}(X) = L_2] = P_{\delta_2=0}[Y_1 > Y_2] = P[\chi^2_1(\delta_1) > \chi^2_{p-1}(0)], \tag{33}$$

while for the latter (see Appendix),

$$P_{L_1}[\bar{L}(X) = L_1] = P_{\delta_1=0}[\pi_1(Y_1) > \pi_2(Y_2)] = P[\chi^2_{p-1}(\delta_2) > \chi^2_{p-1}(0)] > \frac{1}{2}, \tag{34}$$

$$P_{L_2}[\bar{L}(X) = L_2] = P_{\delta_2=0}[\pi_1(Y_1) < \pi_2(Y_2)] = P[\chi^2_1(\delta_1) > \chi^2_1(0)] > \frac{1}{2}. \tag{35}$$

(The chi-squares variates in (32)-(35) are independent.) Thus the probability that $\bar{L}$ selects the correct model exceeds 1/2 under both L_1 and L_2, whereas this holds for $\hat{L}$ only under L_1; if L_2 is true then the probability that $\hat{L}$ correctly selects L_2 will be small unless $\delta_1 \gg p-2$. (Note that $\mathrm{E}(\chi^2_1(\delta_1)) = 1 + \delta_1$.) In fact, for fixed δ_1 the probability that $\hat{L}$ correctly selects L_2 approaches 0 as $p \to \infty$.

This shows that the maximum likelihood-based model selection procedure $\hat{L}$ is intrinsically biased in favor of the model of higher dimensionality. By contrast, the procedure $\bar{L}$ is based on the p-values π_1 and π_2 which are known to be "self-weighting" (cf. Berk and Jones (1978), Koziol and Perlman (1978)), i.e., they tend to automatically adapt to the dimension of the model being tested – compare (34)-(35) to (32)-(33). For the same reason we conclude that the significance-based pre-test estimator $\bar{\mu}$ is preferable to the MLE $\hat{\mu}$ under $L_1 \cup L_2$.

4 Model selection for $\{L_0, L_1, L_2\}$

We extend the discussion in Section 3 to the problem of model selection when the nested model $\mu \in L_0 := L_1 \cap L_2 \equiv \{0\}$ is also included. maximum likelihood-based procedures again are unable to adapt to the varying dimensionalities of the partially non-nested subspaces L_0, L_1, L_2. This is already expected if, as in Section 3, L_1 and L_2 have different dimensions. Therefore we shall assume that $p = 2k$, $\dim(L_1) = \dim(L_2) = k > 0 = \dim(L_0)$, and

$$L_1 := \{\mu \mid \mu_1 = \cdots = \mu_k = 0\}, \tag{36}$$

$$L_2 := \{\mu \mid \mu_{k+1} = \cdots = \mu_{2k} = 0\}, \tag{37}$$

Again, ML-based procedures will fail to adapt to the lower dimension of the subspace L_0.

The MLMSP itself is clearly unsatisfactory for this model selection problem since with probability one the maximum of the likelihood over $\{L_0, L_1, L_2\}$ cannot be attained at L_0, hence L_0 can never be selected. It is customary, therefore, to penalize models of higher dimensionality and base model selection on the penalized maximum likelihoods of the competing models.

Standard ML-based procedures such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) attach *additive* penalties to the maximum log likelihoods, hence lead here to model selection procedures $\check{L}_c$ of the following general form: $\check{L}_c(X) = L_0,\ L_1,\ L_2$ according to the minimum of $Y_0,\ Y_1 + c,\ Y_2 + c$, where now

$$\begin{aligned}
Y_1 &:= \|X - L_1\|^2 = X_1^2 + \cdots + X_k^2 \sim \chi_k^2(\delta_1), &(38)\\
Y_2 &:= \|X - L_2\|^2 = X_{k+1}^2 + \cdots + X_{2k}^2 \sim \chi_k^2(\delta_2), &(39)\\
\delta_1 &:= \|\mu - L_1\|^2 = \mu_1^2 + \cdots + \mu_k^2, &\\
\delta_2 &:= \|\mu - L_2\|^2 = \mu_{k+1}^2 + \cdots + \mu_{2k}^2, &\\
Y_0 &:= \|X - L_0\|^2 = Y_1 + Y_2 \sim \chi_{2k}^2(\delta_1 + \delta_2 := \delta_0), &(40)
\end{aligned}$$

and c is a positive constant. Equivalently,

$$\check{L}_c(X) = \begin{cases} L_0, & \text{if } \max(Y_1, Y_2) < c; \\ L_1, & \text{if } \max(Y_1, c) < Y_2; \\ L_2, & \text{if } \max(Y_2, c) < Y_1. \end{cases} \tag{41}$$

Examination of Figure 7 reveals, however, that the additive penalty overcompensates for the lower dimensionality of L_0, in the following sense: for $i = 1, 2$, it is possible that $\frac{\|X - L_i\|}{\|X - L_0\|} \approx 0$ but $\check{L}_c(X) = L_0$ rather than L_i. This occurs in Figure 7, for example, if $(Y_1, Y_2) \approx \xi_i$, $i = 1, 2$.

By contrast, the significance-based PTMSP $\bar{L}$ does not behave in this way. This PTMSP is given by

$$\bar{L}(X) = \begin{cases} L_0, & \text{if } \pi_0(Y_0) > \max(\pi_1(Y_1), \pi_2(Y_2)); \\ L_1, & \text{if } \pi_1(Y_1) > \max(\pi_0(Y_0), \pi_2(Y_2)); \\ L_2, & \text{if } \pi_2(Y_2) > \max(\pi_0(Y_0), \pi_1(Y_1)); \end{cases} \tag{42}$$

where the p-values $\pi_i(Y_i)$ are defined as in (24) for $i = 0, 1, 2$. For both $\bar{L}$ and $\check{L}_c$, $P_{L_i}[L_j \text{ is selected}] \to 1$ as $\delta_j \to \infty$ if $i = j$, and $\to 0$ if $i \neq j$ $(i, j \in \{1, 2\}, i \neq j)$. Unlike $\check{L}_c$, however, the PTMSP $\bar{L}$ selects L_i if $\frac{\|X - L_i\|}{\|X - L_0\|} \approx 0$, $i = 1, 2$ (see Figure 8). Furthermore, in the absence of prior information about the relative likelihoods of L_0, L_1, and L_2, $\check{L}_c$ requires an arbitrary choice of the constant c.

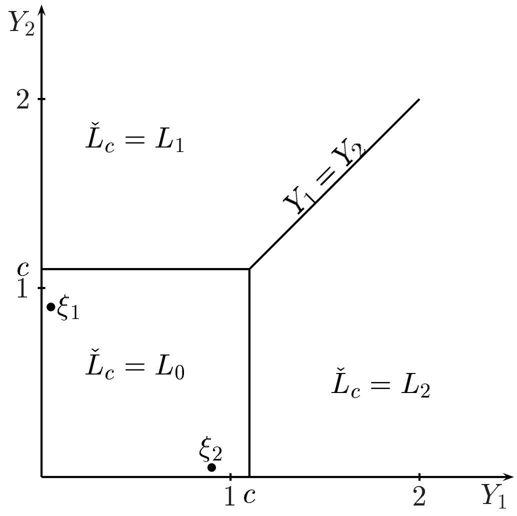


Figure 7. *The maximum-likelihood based model selection procedure* $\check{L}$ *given by (41).*

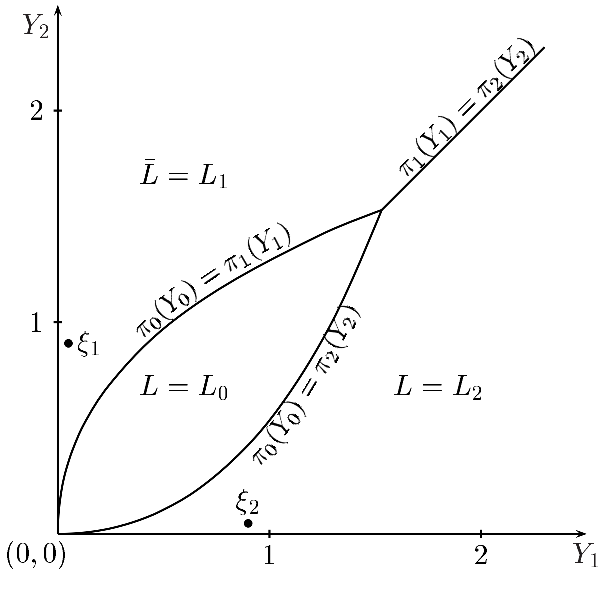


Figure 8. *The preliminary-test model selection procedure* $\bar{L}$ *given by (42), with* $dim(L_1) = \dim(L_2) = 1$. *The figure for* $dim(L_1) = \dim(L_2) = 5$ *is similar in appearance.*

Table 1. Error Rates for PTMSP and AIC

$\dim(L_1) = \dim(L_2) = 1$					$\dim(L_1) = \dim(L_2) = 5$				
True parameter	method	selected space			True parameter	method	selected space		
$\mu = (\mu_1, \mu_2)$		L_0	L_1	L_2	$\mu = (\mu_1, \mu_2)$		L_0	L_1	L_2
(3, 0) in L_1	PTMSP	0.002	0.97	0.02	$((1.5)^5, 0^5)$ in L_1	PTMSP	0.01	0.94	0.05
	AIC	0.03	0.95	0.02		AIC	0.17	0.80	0.02
(2, 0) in L_1	PTMSP	0.06	0.81	0.12	$((1.0)^5, 0^5)$ in L_1	PTMSP	0.05	0.78	0.17
	AIC	0.24	0.68	0.06		AIC	0.52	0.43	0.04
(1, 0) in L_1	PTMSP	0.18	0.54	0.26	$((0.5)^5, 0^5)$ in L_1	PTMSP	0.13	0.51	0.34
	AIC	0.55	0.33	0.11		AIC	0.74	0.17	0.08
(0.1, 0) in L_1	PTMSP	0.26	0.37	0.36	$((0.1)^5, 0^5)$ in L_1	PTMSP	0.18	0.39	0.42
	AIC	0.69	0.16	0.14		AIC	0.84	0.07	0.08
(0, 0) in L_0	PTMSP	0.28	0.36	0.35	$(0^5, 0^5)$ in L_0	PTMSP	0.22	0.40	0.37
	AIC	0.71	0.14	0.14		AIC	0.86	0.07	0.06

Note: $(1.5)^5$ denotes $(1.5, 1.5, 1.5, 1.5, 1.5)$, 0^5 denotes $(0, 0, 0, 0, 0)$, and $(1.0)^5, (0.5)^5, (0.1)^5$ denote similar vectors. The figures in the table are the estimated probabilities of selecting the corresponding parameter spaces. Simulations are based on 1000 iterations.

Table 1 shows the simulation results comparing the performances of the PTMSP $\bar{L}$ (42) and the AIC procedure, which is the particular case of the ML-based MSP $\check{L}_c$ (41) with $c = k \equiv \dim(L_i)$, $i = 1, 2$. As predicted by our comparisons of Figures 7 and 8, the ML-based procedure is too heavily weighted in favor of L_0, the model of lower dimension. The PTMSP selects the true parameter space more often than AIC, except for the case of L_0. The difference between the two methods increases as the dimensions of the parameter spaces L_1 and L_2 increase.

Remark 4.1 If the subspaces L_1 and L_2 are non-orthogonal, then the the chi-square variates Y_1 and Y_2 in (6) and (7) are no longer independent, nor are W_1 and W_2 in (12) and (13). Nonetheless, our assertions in Sections 2.1, 2.2, 3, and 4 about the inappropriateness of the LRT, MLE, and MLMSP, and the preferability of procedures based on combining the individual p-values, remain qualitatively valid for composite models of varying dimensionalities. Our assertions also remain valid if the common variance is unknown but estimated independently.

Remark 4.2 The inappropriateness of the LRTs (8) and (14), the MLE $\hat{\mu}$ (18), the MLMSP $\hat{L}$ (19), the maximum likelihood-based procedures $\check{\mu}_c$ (30) and $\check{L}_c$ (31), (41), all persist if the line L_1 is expanded to a cylinder of small radius with L_1 as its axis; here we say that the cylinder is *approximable* by a one-dimensional subspace. More generally, the inappropriateness of these maximum likelihood-based tests, estimators, and model selection procedures persist whenever the composite model is approximable by a union of two (or more - see Section 5) pairwise or partially non-nested regions (not necessarily entire subspaces or orthogonal) of varying dimensionality. Furthermore, these features persist for non-normal but regular parametric families, at least for moderate or large sample sizes.

5 Three or more non-nested linear subspaces

The discussions in Sections 2.1, 3, and 4 extend readily to hypothesis testing, estimation, and model selection problems involving a union of three or more non-nested linear subspaces.

First we extend the discussion in §2.1. Consider the problem of testing

$$H_0 : \mu \in L_1 \cup \cdots \cup L_r \qquad \text{vs.} \qquad H : \mu \in \mathbf{R}^p \setminus (L_1 \cup \cdots \cup L_r) \tag{43}$$

based on $X \sim N_p(\mu, I_p)$, where $L_1, \ldots, L_r$ are $r \geq 3$ pairwise non-nested subspaces of $\mathbf{R}^p$. It need not be assumed that these subspaces are pairwise orthogonal or that $L_i \cap L_j = \{0\}$, but for simplicity we continue to assume that $L_1, \ldots, L_r$ span $\mathbf{R}^p$. For each $i = 1, \ldots, r$,

$$Y_i := \|X - L_i\|^2 \sim \chi^2_{p-d_i}(\delta_i), \tag{44}$$

where $\delta_i = \|\mu - L_i\|^2$ and $d_i = \dim(L_i)$. (These chi-square variates are not necessarily independent.) The size α LRT for (43) rejects H_0 iff

$$\|X - (L_1 \cup \cdots \cup L_r)\|^2 \equiv \min(Y_1, \ldots, Y_r) > \chi^2_{p-m,\alpha}, \tag{45}$$

where $m := \min(d_1, \ldots, d_r)$, while the size α IUT rejects H_0 iff

$$\min\Big(\frac{Y_1}{\chi^2_{p-d_1,\alpha}}, \ldots, \frac{Y_r}{\chi^2_{p-d_r,\alpha}}\Big) > 1. \tag{46}$$

(cf. Berger and Sinclair (1984, Theorems 1 and 2)). As before, if $d_1, \ldots, d_r$ are not identical then the IUT is preferable to the LRT, not because it dominates the LRT on the basis of the NP criterion but because it is consonant with the individual tests for the hypotheses $\mu \in L_i$ vs. $\mu \in \mathbf{R}^p \setminus \{L_i\}$, $i = 1, \ldots, r$, whereas the LRT is not consonant – the LRT remains overly conservative.

Similarly, the inappropriate behavior (cf. Section 3) of the MLE $\hat{\mu}$ and MLMSP $\hat{L}$ under the model H_0 in (2) is shared by the MLE $\hat{\mu}$ and MLMSP $\hat{L}$ under H_0 in (43). Here $\hat{\mu}(X)$ is simply the projection of X onto the nearest of the subspaces $L_1, \ldots, L_r$, while the associated MLMSP $\hat{L}$ selects this nearest subspace. Specifically, $\hat{L}(X) = L_{\hat{i}(X)}$ and $\hat{\mu}(X) = X_{L_{\hat{i}(X)}}$, where

$$\hat{i}(X) := \mathrm{argmin}_{i=1,\ldots,r}\{\, Y_i\}. \tag{47}$$

Again it can be shown that the behaviors of $\hat{\mu}$ and $\hat{L}$ are not consonant with the individual tests for the hypotheses $\mu \in L_i$ vs. $\mu \in \mathbf{R}^p \setminus \{L_i\}$, $i = 1, \ldots, r$, because $\hat{\mu}$ and $\hat{L}$ favor the subspaces of largest dimension.

Instead, we propose the significance-based pre-test estimator $\bar{\mu}(X) := X_{L_{\bar{i}(X)}}$ and pre-test model selection procedure $\bar{L}(X) := L_{\bar{i}(X)}$, where

$$\bar{i}(X) := \mathrm{argmax}_{i=1,\ldots,r}\{\, \pi_i(Y_i)\}; \tag{48}$$

here

$$\pi_i(y) := P_{\mu \in L_i}[Y_i > y], \quad i = 1, \ldots, r, \tag{49}$$

are the p-values for testing $\mu \in L_i$ vs. $\mu \in \mathbf{R}^p \setminus \{L_i\}$ individually. Then it can be shown (see Appendix) that for each $i = 1, \ldots, r$,

$$\{X \mid \bar{L}(X) = L_i\} = \bigcup_{0<\alpha<1} U_{i,\alpha}, \tag{50}$$

where

$$U_{i,\alpha} := \{X \mid Y_i < \chi^2_{p-d_i,\alpha},\, Y_j > \chi^2_{p-d_j,\alpha} \text{ for } j \neq i\}. \tag{51}$$

Thus, $\bar{L}(X) = L_i$ iff for at least one $\alpha \in (0,1)$, $\mu \in L_i$ is accepted at level α and $\mu \in L_j$ rejected at level α for each $j \neq i$. Equivalently, $\bar{L}(X) = L_i$ iff for each $j \neq i$, there exists no $\alpha \in (0,1)$ for which $\mu \in L_j$ is accepted at level α and $\mu \in L_{j'}$ rejected for each $j' \neq j$ at level α.

The probabilities of correct selection for $\hat{L}$ and $\bar{L}$ are given by

$$P_{L_i}[\hat{L}(X) = L_i] = P_{\delta_i=0}[Y_i < Y_j,\, j \neq i] = P[\chi^2_{p-d_i}(0) < \chi^2_{p-d_j}(\delta_j),\, j \neq i], \tag{52}$$

$$P_{L_i}[\bar{L}(X) = L_i] = P_{\delta_i=0}[\pi_i(Y_i) > \pi_j(Y_j),\, j \neq i] = P[\chi^2_{p-d_j}(0) < \chi^2_{p-d_j}(\delta_j),\, j \neq i], \quad (53)$$

respectively (see Appendix), where the chi-square variates are possibly dependent. Nonetheless, a comparison of (52) and (53) indicates that the PTMSP $\bar{L}$, unlike the MLMSP $\hat{L}$, again compensates for the differing dimensionalities of the models under consideration. The pre-test estimator $\bar{\mu}$ is preferable to the MLE $\hat{\mu}$ under H_0 for the same reason.

If, instead, $L_1, \ldots, L_r$ are only partially non-nested (as in Section 4), then the PTMSP $\bar{L} \equiv L_{\bar{i}}$ with $\bar{i}$ given by (48) remains a reasonable model selection procedure.

6 Testing a Multivariate One-sided Hypothesis

In §2.1 it was demonstrated that the LRT is unable to adapt when the null hypothesis is a union of regions of varying dimensionalities. This phenomenon also occurs when the boundary of the null hypothesis is a union of such regions. This is demonstrated by the following simple example, due to Shimodaira (2000a).

Based on the observation $X \sim N_p(\mu, I_p)$ as in (1) with $p \geq 2$, consider the problem of testing

$$H_0 : \max(\mu_1, \ldots, \mu_p) \leq 0 \qquad \text{vs.} \qquad H : \max(\mu_1, \ldots, \mu_p) > 0, \quad (54)$$

Equivalently, (54) can be expressed as

$$H_0 : \mu \in \mathcal{N}^p \qquad \text{vs.} \qquad H : \mu \in \mathbf{R}^p \backslash \mathcal{N}^p, \quad (55)$$

where $\mathcal{N}^p$ is the nonpositive orthant in $\mathbf{R}^p$. For $0 < \alpha < 1 - 2^{-p}$, the size α LRT for (54) $\equiv$ (55) rejects H_0 iff

$$\|X - \mathcal{N}^p\| \equiv \left((X_1^+)^2 + \cdots + (X_p^+)^2\right)^{1/2} > l_{p,\alpha}, \quad (56)$$

where $l_{p,\alpha} > 0$ is the unique solution to the equation

$$\alpha = \sum_{i=1}^{p} 2^{-p} \binom{p}{i} P[\chi_i^2 > l_{p,\alpha}^2] \quad (57)$$

(cf. Robertson *et al* (1988, §2.2, 2.3)). A second apparently reasonable test for this problem is the size α union-intersection test (UIT) (Roy (1958)), which for $0 < \alpha < 1 - 2^{-p}$ rejects H_0 iff

$$\max(X_1, \ldots, X_p) > u_{p,\alpha}, \quad (58)$$

where $u_{p,\alpha} = \Phi^{-1}(\sqrt[p]{1-\alpha})$. It is straightforward to verify that $u_{p,\alpha} < l_{p,\alpha}$ (see Appendix). The acceptance regions of these two tests are shown in Figure 9 for the case $p = 2$.

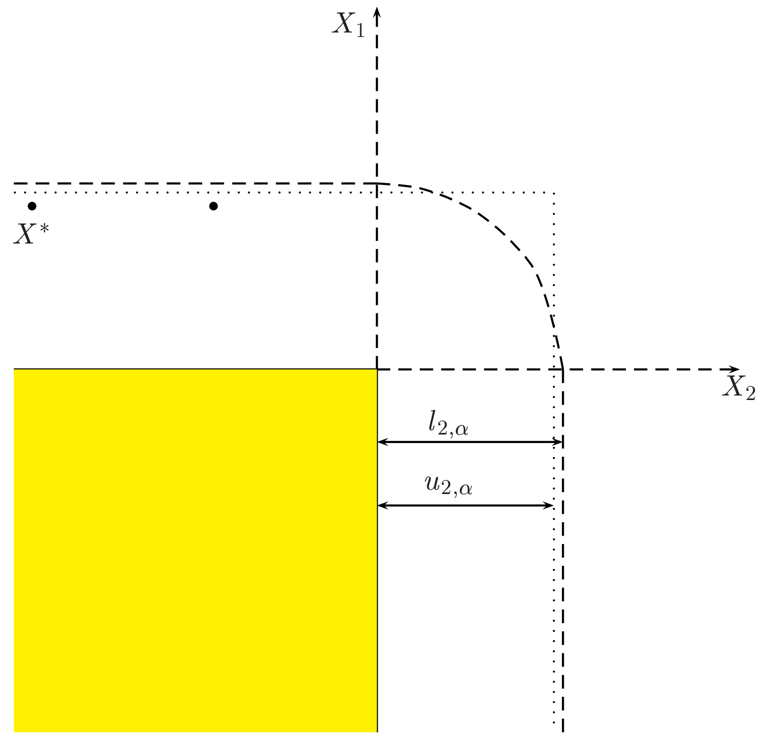


Figure 9. *The size α LRT and UIT for (54) $\equiv$ (55) ($p = 2$, $\alpha = 0.05$, $u_{2,\alpha} = 1.95$, $l_{2,\alpha} = 2.05$).*

The least favorable distribution in H_0 for (54) $\equiv$ (55) occurs at $\mu = 0$ – see Robertson *et al* (1988, §2.3). For both the LRT and UIT, the probability of rejecting H_0 attains its maximum value α over $\mu \in H_0$ at $\mu = 0$ and decreases as each μ_i decreases. (This is true for any test with monotone rejection region – cf. Lehmann (1952).) Thus, neither test is similar on the boundary of H_0, and both are biased.

It is well known, however, that similarity and unbiasedness *per se* are not necessarily desirable for testing problems with composite null hypotheses (cf. Perlman and Wu (1999, Sections 4, 5, and 7)). We now argue that *both* the LRT and UIT can yield inappropriate inferences, especially for large p.

Because $u_{p,\alpha} \to \infty$ as $p \to \infty$, if p is large then for each $i = 1, \ldots, p$ the acceptance region of the UIT contains sample points X^* such that $X_i^* \gg z_\alpha := \Phi^{-1}(1-\alpha)$ but $\max_{j \neq i} X_j^* \to -\infty$ (see Figure 9). This second property of X^* indicates that $\max_{j \neq i} \mu_j \ll 0$, so the null hypothesis in (54) reduces to $H_{0,i} : \mu_i \leq 0$, while the first property of X^* tells us to strongly reject $H_{0,i}$. Since $X_1, \ldots, X_p$ are independent with standard deviation 1, such observations X^* therefore provide strong evidence against H_0 ($\subset H_{0,i}$), yet the UIT fails to reject H_0. Because $l_{p,\alpha} \to \infty$ as $p \to \infty$, the LRT exhibits similarly contradictory behavior.

[Admittedly, the convergence of $u_{p,\alpha}$ and l_α to ∞ is slow, occurring at the rate $\mathrm{O}((\log p)^{1/2})$. For $u_{p,\alpha}$, this follows from standard extreme value theory, e.g., Resnick (1987); for $l_{p,\alpha}$, this is suggested by the approximation given by Barlow *et al* (1972, pp.150-153.]

We offer the following explanation for this behavior. The orthant $\mathcal{N}^p$ is a convex polyhedral cone whose boundary consists of a union of faces of dimensions 0 (the origin), $1, \ldots, p-1$. Thus the boundary of H_0 can be thought of as a union of statistical models with varying dimensions. Because the least favorable distribution occurs at $\mu = 0$, the LRT and UIT determine their critical values with reference to the model of lowest dimension and are therefore biased in favor of the models 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^*). This behavior is qualitatively similar to that described in §2.1. e acceptance region of the UIT contains sample As in §2.1, more appropriate tests for (54) $\equiv$ (55) should adapt to the varying dimensionalities in H_0 – in this case, to the dimension of the face of $\mathcal{N}^p$ closest to the sample point X. Shimodaira (2000a) proposes such a test based on bootstrap resampling – also see Efron and Tibshirani (1998). We shall propose an explicit test qualitatively similar to Shimodaira's procedure.

For each $\sigma \in S_p := 2^{\{1,\ldots,p\}}$, define

$$L_\sigma := \{\mu \in \mathbf{R}^p \mid \mu_i = 0 \ \forall \ i \in \sigma\}, \tag{59}$$

$$F_\sigma := L_\sigma \cap \mathcal{N}^p. \tag{60}$$

($L_\emptyset = \mathbf{R}^p$ and $F_\emptyset = \mathcal{N}^p$.) Thus L_σ is a linear subspace of dimension $p - |\sigma|$ and

$$\mathcal{N}^p = \cup_{\sigma \in S_p} F_\sigma, \tag{61}$$

where $(F_\sigma \mid \sigma \in S_p \backslash \emptyset)$ is the family of faces of $\mathcal{N}^p$ (see Figure 10). For $\sigma \in S_p \backslash \emptyset$, let

$T_\sigma \equiv T_\sigma(X)$ be a statistic appropriate for testing

$$H_{0,\sigma} : \mu \in F_\sigma \qquad \text{vs.} \qquad H : \mu \in \mathbf{R}^p \backslash \mathcal{N}^p \tag{62}$$

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

$$\max_{\sigma \in S_p \backslash \emptyset} \pi(T_\sigma(X)) < \alpha \tag{63}$$

is expected to be an approximately size α test for (54) $\equiv$ (55). Like the IUT in §2.1, this test uses the individual p-values π_σ to adapt to the varying dimensionalities of the faces F_σ.

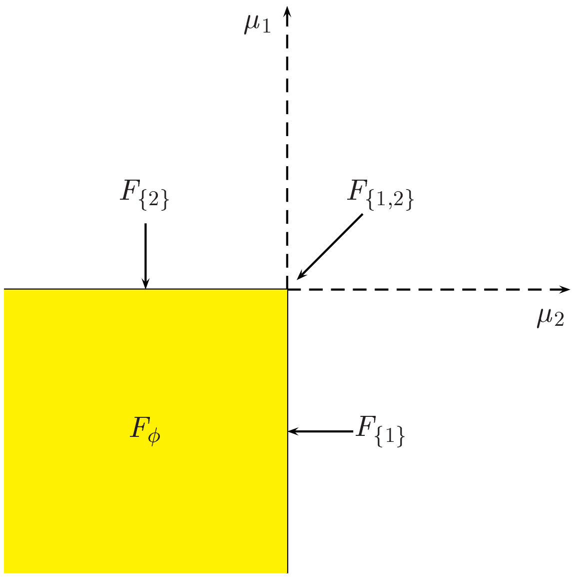


Figure 10. *The faces of $\mathcal{N}^p$ for $p = 2$.*

Because it may be difficult to choose appropriate statistics T_σ and calculate the corresponding p-values π_σ, we propose instead the following test which is qualitatively similar to (63) and somewhat easier to implement: *accept* H_0 iff

$$(1 - 1_{\mathcal{N}^p}(X)) \sum_{i \in \sigma} X_i^2 \leq a^2_{|\sigma|,\alpha} \quad \textit{and} \quad \max_{i \notin \sigma} X_i \leq 0 \tag{64}$$

for at least one $\sigma \in S_p \backslash \emptyset$, where for $k = 1, \ldots, p$,

$$a^2_{k,\alpha} := \bar{G}_k^{-1} \left(\frac{\alpha}{1 - 2^{-k}} \right) \tag{65}$$

and $G_k \equiv 1 - \bar{G}_k$ is the cdf of the χ_k^2 distribution. For $0 < \alpha < 1/2$, simulation results (cf. Tables 2 and 3) suggest that this test is approximately size α for (54) $\equiv$ (55).

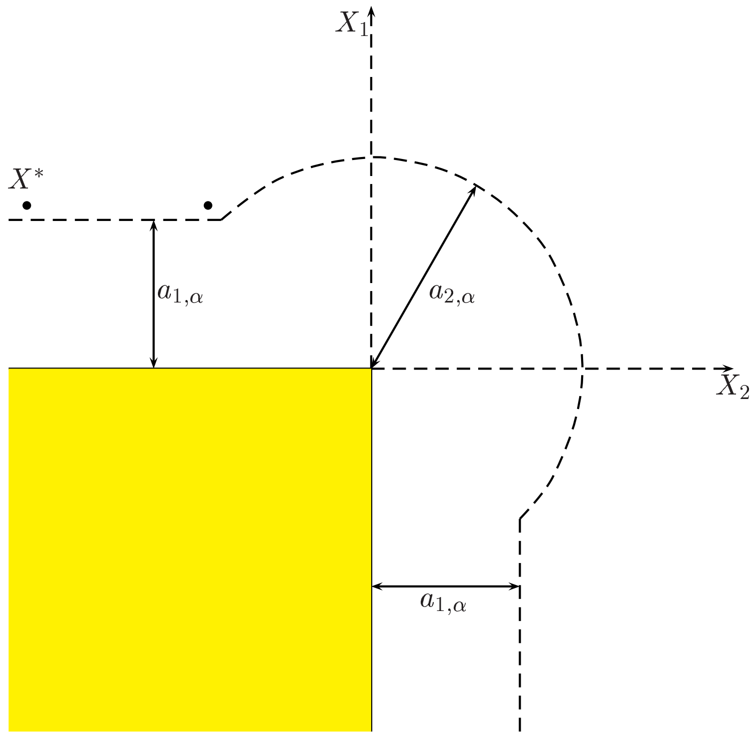


Figure 11. *The size α adaptive test (64) for (54) $\equiv$ (55) ($p = 2,\ \alpha = 0.05,\ a_{1,\alpha} = 1.64,\ a_{2,\alpha} = 2.33$).*

The form of the test (64), qualitatively similar to that of the bootstrap-based test of Shimodaira (2000a, Figure 1), is indicated in Figure 11 for the case $p = 2$. By comparing this to the forms of the LRT (56) and UIT (58) in Figure 9, it is seen that the proposed test adapts to the varying dimensionalities of the faces F_σ of $\mathcal{N}^p$, unlike the LRT and IUT, and does not share their anomalous behavior as described above. In fact, because $G_k(\cdot)$ is decreasing in k, necessarily $G_k^{-1}(\cdot)$ is increasing in k, hence so is $a_{k,\alpha}$. Note in particular that for large p,

$$a_{1,\alpha} = z_\alpha \equiv \Phi^{-1}(1-\alpha) \quad \ll \quad \Phi^{-1}(\sqrt[p]{1-\alpha}) \equiv u_{p,\alpha} < l_{p,\alpha}, \tag{66}$$

$$a_{p,\alpha} \approx \sqrt{\chi^2_{p,\alpha}} = \mathrm{O}(\sqrt{p}) \quad \gg \quad l_{p,\alpha} > u_{p,\alpha}. \tag{67}$$

Table 2. Type I Error Rates for the Three Tests

Dimension	μ	PRE	LRT	UIT	μ	PRE	LRT	UIT
$p = 2$	(0,0)	0.034	0.051	0.051	(-2,0)	0.041	0.020	0.026
	(-1,0)	0.027	0.022	0.027	(-5,0)	0.052	0.021	0.027
$p = 3$	(0,0,0)	0.027	0.047	0.047	(-1,-1,0)	0.015	0.012	0.018
	(-1,0,0)	0.022	0.028	0.033	(-2,-2,0)	0.031	0.009	0.016
	(-2,0,0)	0.030	0.025	0.032	(-5,-5,0)	0.050	0.011	0.018
	(-5,0,0)	0.033	0.023	0.030				
$p = 5$	(0,0,0,0,0)	0.020	0.052	0.049	(-1,-1,-1,0,0)	0.012	0.012	0.021
	(-1,0,0,0,0)	0.015	0.031	0.035	(-2,-2,-2,0,0)	0.023	0.008	0.020
	(-2,0,0,0,0)	0.020	0.032	0.041	(-5,-5,-5,0,0)	0.035	0.009	0.020
	(-5,0,0,0,0)	0.023	0.032	0.040	(-1,-1,-1,-1,0)	0.006	0.004	0.012
	(-1,-1,0,0,0)	0.015	0.022	0.030	(-2,-2,-2,-2,0)	0.023	0.003	0.011
	(-2,-2,0,0,0)	0.022	0.018	0.029	(-5,-5,-5,-5,0)	0.051	0.003	0.011
	(-5,-5,0,0,0)	0.029	0.019	0.030				

Note: 20,000 iterations, α=0.05, standard errors are not larger than 0.003.

Tables 2 and 3 show that our proposed explicit test (64), like Shimodaira's bootstrap test, is more nearly similar on the boundary of H_0 and more nearly unbiased than the LRT and the UIT. As already noted, however, similarity and unbiasedness are not intrinsically desirable properties for testing problems with composite null hypotheses. We reiterate that our preference for adaptive tests such as Shimodaira's test and (64) is based not on consideration of power and unbiasedness but rather on the fact that these tests better reflect the evidence that the data provides regarding the competing hypotheses in (55).

Remark 6.1. The acceptance and rejection regions of our proposed test (64) and Shimodaira's test are not monotone in the components of X. However, Perlman and Wu (1999, Rejoinder, penultimate paragraph) show that monotonicity is not necessarily relevant for testing problems with multivariate one-sided hypotheses.

Table 3. Power Comparison for the Three Tests

Dimension	μ	PRE	LRT	UIT	μ	PRE	LRT	UIT
$p = 2$	(1,1)	0.26	0.36	0.31	(2,2)	0.76	0.84	0.77
	(-2,1)	0.22	0.14	0.17	(-2,3)	0.88	0.82	0.85
$p = 3$	(1,1,1)	0.29	0.45	0.34	(2,2,2)	0.85	0.93	0.83
	(-2,1,1)	0.24	0.26	0.24	(-2,2,2)	0.73	0.75	0.69
	(-2,-2,1)	0.19	0.09	0.13	(-2,-2,2)	0.54	0.37	0.45
$p = 5$	(1,1,1,1,1)	0.37	0.59	0.39	(2,2,2,2,2)	0.96	0.99	0.90
	(-2,1,1,1,1)	0.31	0.44	0.32	(-2,2,2,2,2)	0.91	0.95	0.84
	(-2,-2,1,1,1)	0.26	0.28	0.24	(-2,-2,2,2,2)	0.83	0.85	0.75
	(-2,-2,-2,1,1)	0.21	0.14	0.17	(-2,-2,-2,2,2)	0.68	0.60	0.60
	(-2,-2,-2,-2,1)	0.15	0.04	0.09	(-2,-2,-2,-2,2)	0.46	0.23	0.37

Note: 20,000 iterations, α=0.05, standard errors are not larger than 0.003.

7 Concluding Remarks

We have presented several examples of statistical testing and estimation problems where the LRTs, MLEs, and associated model selection procedures lead to inappropriate inferences. These statistical models are simple and well-behaved, based on families of univariate normal distributions with known variances, so the undesirable behavior of the LRT and MLE are not artifacts of irregularity, non-measurability, contamination, unboundedness, or asymptotics. (See Le Cam (1990) for an excellent survey of such anomalies of MLEs.) Rather, this behavior is rooted in the geometric nature of the statistical models or their boundaries, which comprise regions of varying dimensionalities.

This underscores the need to examine carefully the geometric structure of the statistical model before applying maximum-likelihood-based methods for its analysis. Recent studies suggest that latent-variable models, hidden Markov models, LISREL models including linear structural equation models, and other complex multivariate dependence models may have non-standard geometric forms – cf. Settimi and Smith (2000), Geiger *et al* (2001), Bergsma and Rudas (2002, Example 7). Maximum likelihood estimation and model selection for phylogenetic labeled trees also may be problematic due to the geometry of the models – cf. Felsenstein (1981, 1983)), Holmes (1997), Shimodaira and Hasegawa (1999), Shimodaira (2000b), Billera, Holmes, and Vogtmann (2000). In all cases, caution should be exercised before the use of computer packages that automatically apply maximum likelihood-based methods, including the LRT, MLE, AIC, and BIC.

Acknowledgement. We warmly thank Steen Andersson, David Cox, Persi Diaconis, Brad Efron, Joe Felsenstein, Leon Gleser, Susan Holmes, Michael Lavine, Erich Lehmann, Adrian Raftery, Hidetoshi Shimodaira, Elizabeth Thompson, Paul Vos, and Jon Wellner for many helpful discussions. The research was supported in part by grants from the U.S. National Science Foundation and the Natural Sciences and Engineering Research Council of Canada.

Appendix

Verification of (17)*:*

First note that $\chi^2_{k,\alpha} = \bar{G}_k^{-1}(\alpha)$, where $G_k \equiv 1 - \bar{G}_k$ is the cdf of the $\chi^2_k(0)$ distribution; G_k (G_k^{-1}) is an increasing (decreasing) function. From (15), $1-\alpha > G^2_{p-1}(c_{p,\alpha})$, so $1-\sqrt{1-\alpha} < \bar{G}_{p-1}(c_{p,\alpha})$, hence $\bar{G}^{-1}_{p-1}(1-\sqrt{1-\alpha}) > c_{p,\alpha}$, which verifies the fourth inequality in (17). Again from (15), $1-\alpha < G_{p-1}(c_{p,\alpha})$, so $\bar{G}_{p-1}(c_{p,\alpha}) < \alpha$, hence $c_{p,\alpha} > \bar{G}^{-1}_{p-1}(\alpha)$, which yields the third inequality in (15). The second and first inequalities are straightforward.

Verification of (27), (28)*, and* (50)*:*

By (24), $\pi_1 = \bar{G}_1$ and $\pi_2 = \bar{G}_{p-1}$, hence $A_\alpha \equiv (\chi^2_{1,\alpha}, \chi^2_{p-1,\alpha})$ lies on the delineating boundary $\{(Y_1, Y_2) \mid \pi_1(Y_1) = \pi_2(Y_2)\}$ and in fact this boundary is the locus A_α, $0 < \alpha < 1$. Because π_1 and π_2 are strictly decreasing functions, this boundary corresponds to a strictly increasing function of Y_1 (see Figure 3), hence (27) and (28) follow immediately. The proof of (50) is similar.

Derivation of (34)*,* (35)*, and* (53)*:*

By (24),

$$P_{\delta_1=0}[\pi_1(Y_1) > \pi_2(Y_2)] \quad = \quad P_{\delta_1=0}[G_{p-1}(Y_2) > G_1(Y_1) \equiv U]$$

$$
\begin{aligned}
&= P_{\delta_1=0}[Y_2 > G_{p-1}^{-1}(U)] \\
&= P[\chi^2_{p-1}(\delta_2) > \chi^2_{p-1}(0)],
\end{aligned}
$$

where $U \sim \text{Uniform}[0,1]$. The proofs of (35) and (53) are similar.

Verification that $u_{p,\alpha} < l_{p,\alpha}$ (cf. (57) and (58)):

From its definition, $u_{p,\alpha} \equiv \Phi^{-1}(\sqrt[p]{1-\alpha})$ satisfies

$$\sqrt[p]{1-\alpha} = \frac{1}{2}\left(1 + P[\chi^2_1 < u^2_{p,\alpha}]\right), \tag{68}$$

hence

$$1-\alpha = \frac{1}{2^p}\sum_{i=0}^{p}\binom{p}{i}(P[\chi^2_1 < u^2_{p,\alpha}])^i. \tag{69}$$

By (57), however,

$$1-\alpha = \frac{1}{2^p}\sum_{i=0}^{p}\binom{p}{i}P[\chi^2_i < l^2_{p,\alpha}] =: K(l_{p,\alpha}), \tag{70}$$

where $\chi^2_0 := 0$. Because $(P[\chi^2_1 < v])^i > P[\chi^2_i < v]$ for $v > 0$ and $i = 1, \ldots, p$, therefore $K(l_{p,\alpha}) > K(u_{p,\alpha})$. Thus, since $K(\cdot)$ is strictly increasing, necessarily $l_{p,\alpha} > u_{p,\alpha}$.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions.* New York: Wiley.

Berger, R. L. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *J. Amer. Statist. Assoc.* **84** 192-199.

Berger, R. and Sinclair, D. (1984). Testing hypotheses concerning unions of linear subspaces. *J. Amer. Statist. Assoc.* **79**, 158-163.

Bergsma, W. P. and Rudas, T. (2002). Marginal models for categorical data. *Ann. Statist.* **30** 140-159.

Berk, R. and Jones, D. (1978). Relatively optimal combinations of test statistics. *Scand. J. Statist.* **5**, 158-162.

Billera, L., Holmes, S. P., Vogtmann, K. (2000). Geometry of the space of phylogenetic trees. Tech. Report, Stanford University, Stanford, CA. http://www-stat.stanford.edu/ susan/papers/june1.ps

Brown, L. D., Hwang, J. T. G., and Munk, A. (1997). An unbiased test for the bioequivalence problem. *Ann. Statist.* **25** 2345-2367.

Cohen, A., Kemperman, J. H. B., and Sackrowitz, H. B. (2000). Properties of likelihood inference for order restricted models. *J. Multivariate Analysis* **72** 50-77.

Cohen, A. and Sackrowitz, H. B. (1998). Directional tests for one-sided alternatives in multivariate models. *Ann. Statist.* **26** 2321-2338.

Eaton, M. L. (1970). A complete class theorem for multidimensional one-sided alternatives. *Ann. Math. Statist.* **41** 1884-1888.

Efron, B. (1984). Comparing non-nested linear models. *J. Amer. Statist. Assoc.* **79** 791-803.

Efron, B. and Tibshirani, R. (1998). The problem of regions. *Ann. Statist.* **26** 1687-1718.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17** 246-272.

Felsenstein, J. (1983). Statistical inference of phylogenies (with discussion). *J. Royal Statist. Soc. A* **146** 368-376.

Geiger, D., Heckerman, D., King, H., Meek, C. (2001). Stratified exponential families: graphical models and model selection. *Ann. Statist.* **29 (2)**, 505-529.

Holmes, S. P. (1997). Phylogenetic trees: an overview. In *Statistics and Genetics*, Institute of Mathematics and its Applications, Minneapolis, MN, 81-118. (Also published by Springer, NY, 1999.) http://www-stat.stanford.edu/ susan/papers/ima.ps

Koziol, J. A. and Perlman, M. D. (1978). Combining independent non-central chi-square tests. *J. Amer. Statist. Assoc.* **73** 753-763.

Le Cam (1990). Maximum likelihood: an introduction. *International Statist. Review* **58** 153-171.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Ann. Math. Statist.* **23** 541-562.

Liu, H. (2000). Uniformly more powerful, two-sided tests for hypotheses about linear inequalities. *Ann. Inst. Statist. Math.* **52** 15-27.

Liu, H. and Berger, R. L., (1995). Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. *Ann. Statist.* **23** 55-72.

McDermott, M. P. and Wang, Y. (2002). Construction of uniformly more powerful tests for hypotheses about linear inequalities. *J. Statist. Planning Inference.* To appear.

Menendez, J. A. and Salvador, B. (1991). Anomalies of the likelihood ratio test for testing restricted hypotheses. *Ann. Statist.* **19** 889-898.

Menendez, J. A., Rueda, C. and Salvador, B. (1992). Dominance of likelihood ratio tests under cone constraints. *Ann. Statist.* **20** 2087-2099.

Mukerjee, H. and Tu, R. (1995). Order-restricted inferences in linear regression. *J. Amer. Statist. Assoc.* **90** 717-728.

Munk, A. (1999). A note on unbiased testing for the equivalence problem. *Statist. Prob. Letters* **41** 401-406.

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). A defense of the likelihood ratio criterion for testing one-sided and order-restricted alternatives. *J. Statist. Planning Inference*, in press.

Resnick, S. I. (1987), *Extreme Values, Regular Variation, and Point Processes*, New York: Springer.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order-Restricted Statistical Inference.* New York: Wiley.

Roy, S. N. (1958). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.* **24** 220-238.

Settimi, R. and Smith, J. Q. (2000). Geometry, moments, and conditional independence trees with hidden variables. *Ann. Statist.* **28**, 1179-1205.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16** 1114-1116.

Shimodaira, H. (2000a). 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, H. (2000b). Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. Tech. Report, Institute of Statistical Mathematics, Tokyo, Japan.

Stein, C. (1956). The admissibility of Hotellings's T^2-test. *Ann. Math. Statist.* **27** 616-623.

Tsuda, Y. (2000). On the bioequivalence problem and a testing hypothesis problem for the bivariate normal distribution. *J. Japan. Statist. Soc.* **30** 213-236.

Wang, W. (1997). Optimal unbiased tests for equivalence intrasubject variability. *J. Amer. Statist. Assoc.* **92** 1163-1170.

Wang, W., Hwang, J. T. G., and Dasgupta, A. (1999). Statistical tests for multivariate bioequivalence. *Biometrika* **86** 395-402.

Warrack, G. and Robertson, T. (1984). A likelihood ratio test regarding two nested but oblique order-restricted hypotheses. *J. Amer. Statist. Assoc.* **79** 881-886.