Reconsidering the asymptotic null distribution of likelihood ratio tests for genetic linkage in multivariate variance components models under complete pleiotropy

SUMMER S. HAN*

Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20892, USA
summer.han@nih.gov

JOSEPH T. CHANG

Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06520, USA

SUMMARY

Accurate knowledge of the null distribution of hypothesis tests is important for valid application of the tests. In previous papers and software, the asymptotic null distribution of likelihood ratio tests for detecting genetic linkage in multivariate variance components models has been stated to be a mixture of chi-square distributions with binomial mixing probabilities. For variance components models under the complete pleiotropy assumption, we show by simulation and by theoretical arguments based on the geometry of the parameter space that all aspects of the previously stated asymptotic null distribution are incorrect—both the binomial mixing probabilities and the chi-square components. Correcting the null distribution gives more conservative critical values than previously stated, yielding $P$ values that can easily be 10 times larger. The true mixing probabilities give the highest probability to the case where all variance parameters are estimated positive, and the mixing components show severe departures from chi-square distributions. Thus, the asymptotic null distribution has complex features that raise challenges for the assessment of significance of multivariate linkage findings. We propose a method to generate an asymptotic null distribution that is much faster than other empirical methods such as permutation, enabling us to obtain $P$ values with higher precision more efficiently.

Keywords: Asymptotic null distribution; Likelihood ratio test; Mixing probabilities; Multivariate linkage; Nonstandard boundary condition; Single-factor model; Variance components.

1. INTRODUCTION AND MAIN RESULTS

Variance components methods are widely used for mapping quantitative trait loci (Almasy and Blangero, 1998; Amos, 1994; Goldgar, 1990; Schork, 1993; SOLAR, 2008). Multivariate extensions of variance components methods, which simultaneously test genetic effects on multiple traits by exploiting extra information related to correlations among the traits, have attracted much attention and have been

*To whom correspondence should be addressed.
found to be more powerful than univariate methods by several studies (Almasy and others, 1997; Amos and others, 2001; Boomsma and Dolan, 1998; Schmitz and others, 1998; Williams, Begleiter, and others, 1999; Williams, Van Eerdewegh, and others, 1999).

The asymptotic null distribution of a univariate variance components test, which is a likelihood ratio test for comparing the full model to the null model where the genetic effect variance parameter is constrained to zero, is well known to be a 50:50 mixture of a point mass at zero and \( \chi^2_1 \) (the chi-squared distribution with 1 degree of freedom). This nonstandard limiting distribution arises because the parameter of interest lies on the boundary of the parameter space under the null hypothesis (Self and Liang, 1987).

An analogous mixture distribution was applied to multivariate variance components tests by a number of papers (Almasy and others, 1997; Amos and others, 2001; Amos and de Andrade, 2001; Kraft and others, 2004; Williams, Begleiter, and others, 1999; Williams, Van Eerdewegh, and others, 1999) in which the asymptotic null distribution of the tests was stated to be a mixture of a point mass at zero and several chi-squared distributions, with binomial mixing probabilities. This multivariate method has become widely available; for example, it is implemented in SOLAR (Almasy and Blangero, 1998; SOLAR, 2008), one of the most commonly used software packages for mapping quantitative loci.

However, the evaluation and verification of the asymptotic null distribution of this test have not yet been done systematically. As part of a study comparing the power of univariate to bivariate variance components methods, Amos and others (2001) conducted a simulation of the null distribution. Their bivariate simulation result was not consistent with the previously stated asymptotic null distribution that they assumed, but the source and nature of the inconsistency have remained unexplained (Amos and de Andrade, 2001).

Here we study the null distribution of multivariate variance components tests under a complete pleiotropy assumption. We demonstrate that all aspects of the previously stated asymptotic null distribution are incorrect—both the binomial mixing probabilities and the chi-squared components. We propose a method to assess pointwise significance by generating the asymptotic null distribution making use of the Fisher information estimated from given data. In an example application to an illustrative data set, we compare our method to other well-known empirical methods such as gene-dropping, permutation, and bootstrap methods.

2. Background

2.1 Variance components linkage methods

Variance components linkage methods partition the variance of the traits into several random-effect components, one of which captures the influence of a hypothesized “major gene” that affects the traits. In univariate models, the response variable includes a single measured trait for each individual, modeled as a sum of 3 independent normally distributed random effects: an additive major gene effect, which is to be tested on each genetic marker, a polygenic effect, and an environmental effect with corresponding variance parameters \( a^2 \), \( g^2 \), and \( e^2 \), respectively. For simplicity, here let us assume one sib-pair in each family. The trait vector for the \( i \)th family follows a multivariate normal distribution: \( y_i = (y_{i1}, y_{i2}) \sim N(\mu, \Sigma_i) \) with

\[
\Sigma_i = \begin{bmatrix}
a^2 + g^2 + e^2 & \pi_{i,12}a^2 + 2\phi_{i,12}g^2 \\
\pi_{i,12}a^2 + 2\phi_{i,12}g^2 & a^2 + g^2 + e^2
\end{bmatrix}
\]

Here the identity by descent (IBD) sharing proportion \( \pi_{i,12} \) quantifies the allele sharing between individuals 1 and 2 at the locus being tested. The kinship coefficient \( \phi_{i,12} \) between individuals 1 and 2 is obtained from the pedigree without using any genetic marker information; for example, for a sib-pair,
\( \phi_{i,12} = \frac{1}{3} \). Linkage is tested by comparing the null hypothesis \( H_0: a^2 = 0 \) to the alternative \( H_1: a^2 > 0 \). The likelihood ratio test (LRT) statistic is twice the difference between the log-likelihood of the full model and the model restricted according to the null hypothesis. The parameter \( a^2 \), which must be nonnegative, lies on the boundary of the parameter space under the null hypothesis \( a^2 = 0 \), so that a nonstandard boundary condition applies (Self and Liang, 1987), and the asymptotic null distribution of LRT is \( \frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2 \), that is, the mixture distribution of 50% point mass at zero (equivalent to \( \chi_0^2 \)) and 50% \( \chi_1^2 \).

Multivariate variance components models are a natural extension of the above single-trait model, where the response variables include \( k \) measured traits for each individual. Continuing to illustrate with the case of one sib-pair per family, \( y_{ij} = (y_{i11}, y_{i12}, \ldots, y_{i1k}; y_{i21}, y_{i22}, \ldots, y_{i2k}) \sim N(\mu, \Sigma_i) \), where \( y_{it} \) is the value of trait \( t \) measured on individual \( j \) in family \( i \) and

\[
\Sigma_i = \begin{bmatrix}
A + G + E & \pi_{i,12}A + 2\phi_{i,12}G \\
\pi_{i,12}A + 2\phi_{i,12}G & A + G + E
\end{bmatrix}.
\] (2.1)

Here \( G = (g_{ij}) \) and \( E = (e_{ij}) \) are general \( k \times k \) covariance matrices for the polygenic effect and environmental effect, respectively. The covariance matrix for the additive major gene effect, \( A(a_{ij}) \), is assumed to be an outer product of the form \( A(a_1a_2\cdots a_k)^T(a_1a_2\cdots a_k) \). Here the variances, which are the diagonal entries \( a_1^2, \ldots, a_k^2 \), are of course nonnegative, but the entries \( a_1 \) can take both positive and negative values, so that the covariances \( a_{ij} = a_ia_j \) can be positive or negative. The null hypothesis to be tested, \( H_0: a_1^2 = \cdots = a_k^2 = 0 \), is that all the major gene effect variances are 0. The parameters in the matrices \( G \) and \( E \) are nuisance parameters.

This model, also known as the “single-factor model” (Evans and others, 2004; Vogler and others, 1997) or the “complete pleiotropic model” (Almasy and others, 1997; Amos and others, 2001; Amos and de Andrade, 2001; Kraft and others, 2004; Williams, Begleiter, and others, 1999; Williams, Van Eerdewegh, and others, 1999), arises when the dominance components of variance for the effects of a single major gene on each trait are assumed to be 0. The idea of the assumed form for \( A \) is that the effect on a given trait of a given major gene genotype is a proportionality constant that depends on the trait times an effect size that depends on the genotype. That is, the major gene genotype is modeled as a latent factor that affects the traits in a linear pleiotropic manner. The single-factor model has been extensively used (Amos and others, 2001; Amos and de Andrade, 2001; Evans, 2002; Evans and Duffy, 2004; Evans and others, 2004; Marlow and others, 2003; Monaco, 2007; Vogler and others, 1997). Later, we will use the term “general model” for models without the complete pleiotropic constraint.

### 2.2 Asymptotic null distributions and binomial mixing probabilities

The null hypothesis of the LRT in the models discussed above violates the regularity conditions that imply the typical asymptotic chi-square distribution, the most obvious violation being that constraining a variance to be 0 forces the variance to lie in the boundary of the parameter space rather than its interior. In the previous work (Amos and others, 2001; SOLAR, 2008), the asymptotic null distribution of the multivariate LRT statistic for testing \( k \) traits has been stated to be a mixture distribution of point mass at zero and chi-square distributions with degrees of freedom from 1 to \( k \), with mixing probabilities coming from the Binomial distribution \( B(k, \frac{1}{2}) \), that is,

\[
\frac{1}{2^k} \binom{k}{0} \chi_0^2 + \frac{1}{2^k} \binom{k}{1} \chi_1^2 + \cdots + \frac{1}{2^k} \binom{k}{k-1} \chi_{k-1}^2 + \frac{1}{2^k} \binom{k}{k} \chi_k^2.
\] (2.2)

Table 1 shows mixing probabilities and critical values according to formula (2.2) for significance levels \( \alpha = 0.01 \) and 0.05 for \( k \)-trait asymptotic null distributions for \( k = 2, 3, 4, \) and 5.
Table 1. Mixing probabilities and critical values of the previously stated asymptotic null distributions for k-trait multivariate tests according to the binomial mixture formula (2.2)

<table>
<thead>
<tr>
<th>k</th>
<th>$\chi^2_0$</th>
<th>$\chi^2_1$</th>
<th>$\chi^2_2$</th>
<th>$\chi^2_3$</th>
<th>$\chi^2_4$</th>
<th>$\chi^2_5$</th>
<th>Critical value ($\alpha = 0.01$)</th>
<th>Critical value ($\alpha = 0.05$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.2500</td>
<td>0.5000</td>
<td>0.2500</td>
<td></td>
<td></td>
<td></td>
<td>7.289</td>
<td>4.231</td>
</tr>
<tr>
<td>3</td>
<td>0.1250</td>
<td>0.3750</td>
<td>0.3750</td>
<td>0.1250</td>
<td></td>
<td></td>
<td>8.746</td>
<td>5.435</td>
</tr>
<tr>
<td>4</td>
<td>0.0625</td>
<td>0.2500</td>
<td>0.3750</td>
<td>0.2500</td>
<td>0.0625</td>
<td></td>
<td>10.019</td>
<td>6.498</td>
</tr>
<tr>
<td>5</td>
<td>0.0313</td>
<td>0.1563</td>
<td>0.3125</td>
<td>0.3125</td>
<td>0.1563</td>
<td>0.0313</td>
<td>11.183</td>
<td>7.480</td>
</tr>
</tbody>
</table>

3. SIMULATIONS

3.1 Methods

We simulated the null distributions of k-trait multivariate LRT statistics for $k = 2, 3, 4,$ and 5. For each test, we generated 1000 data sets each including 2000 independent sib-pairs with traits simulated under the truth of the null hypothesis $H_0$: $a_1^2 = \cdots = a_k^2 = 0$ using the multivariate normal distributions specified at (2.1). Each trait had total variance 1 with $g_{11} = \cdots = g_{kk} = 0.4$ (i.e. the polygenic effect explained 40% of the total variance of each trait) and $e_{11} = \cdots = e_{kk} = 0.6$ (i.e. the rest of the variance was due to environmental effects). Polygenic and environmental correlation coefficients among traits were assigned information and random sampling.

Optimization was done using the Mx software (Neale and others, 1999). We recorded the LRT statistic and estimates $\hat{a}_1^2, \ldots, \hat{a}_k^2$ for each replicate data set. In order to separate the mixture distribution of LRT into its components, we grouped the replicate results by the number of major gene variance parameters that were estimated positive, which we will denote by $v$. There were $k + 1$ groups, starting from 0 (where none of the $k$ parameters was estimated positive) up to $k$ (where all the $k$ parameters were estimated positive), so that $v$ took its values in $\{0, 1, 2, \ldots, k\}$. The empirical mixing probabilities were calculated simply as the fraction of each group within the total number of replicates.

3.2 Results

Table 2, which shows the results of the null distribution simulations, reveals that the mixing probabilities do not agree with the probabilities anticipated in Table 1, that is, the binomial distribution previously used in the literature and software. For example for $k = 2$, the probabilities for $v$ taking the values 0, 1, and 2 are 0.142, 0, and 0.858, respectively, whereas the corresponding binomial probabilities are 0.25, 0.5, and 0.25. The substantial discrepancies already seen in the case $k = 2$ become larger for higher $k$. Instead of a binomial distribution having probabilities $\binom{k}{v}2^{-k}$ for $v = 0, 1, 2, \ldots, k$, the true mixing distribution puts positive probability only on $v = 0$ and $v = k$. Furthermore, rather than having equal probabilities for the cases $v = 0$ and $v = k$, the true distribution puts larger probability on $v = k$, and this effect becomes more extreme as $k$ increases.

Figure 1 shows empirical distributions of the LRT statistics in the cases where $v = k$. We focused on these cases since, as shown in Table 2, only the $v = k$ component had large enough counts to be drawn as histogram. Figure 1 reveals some departures from the anticipated $\chi^2$ distributions. In $k = 2$ and 3, there are more points near zero than anticipated, in $k = 5$, the distribution is shifted to the right compared to the $\chi^2_5$ curve, and $k = 4$ is somewhat in between these 2 different patterns. To quantify the departure
Table 2. Mixing probabilities, K-S tests, and critical values for $k$-trait multivariate test null distributions obtained by simulations. The first 6 columns are probabilities for various values for $v$, calculated as the fraction of replicates that yielded that value for $v$. The Kolmogorov tests were conducted for the $v = k$ cases.

<table>
<thead>
<tr>
<th>Number of major gene effect variance parameters estimated positive ($v$)</th>
<th>K-S test ($P$ value)</th>
<th>Critical value $\alpha = 0.01$</th>
<th>Critical value $\alpha = 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha = 0.01$</td>
<td>$\alpha = 0.05$</td>
</tr>
<tr>
<td>$k = 2$</td>
<td>0.142</td>
<td>—</td>
<td>0.858</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>0.020</td>
<td>—</td>
<td>0.980</td>
</tr>
<tr>
<td>$k = 4$</td>
<td>0.001</td>
<td>—</td>
<td>0.999</td>
</tr>
<tr>
<td>$k = 5$</td>
<td>—</td>
<td>—</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Fig. 1. Empirical null distributions of LRT statistics for $k$-trait multivariate tests for the cases where $v = k$. The blue curve is the $\chi^2_k$ density.
from the anticipated $\chi^2_k$ distributions, we conducted Kolmogorov–Smirnov (K-S) tests, and the results are shown in Table 2. From the small $P$ values we conclude that the empirical distributions do not follow the anticipated $\chi^2_k$ distributions.

The last 2 columns in Table 2 are critical values for the significance levels $\alpha = 0.01$ and 0.05. Compared to the critical values anticipated in Table 1, these critical values are larger, that is, more conservative. This implies that the previously stated null distribution can lead to false-positive findings. For example, suppose we observe an LRT statistic value 14 in a multivariate test with $k = 5$ traits, which the previously stated null distribution assigns a $P$ value of 0.002849. According to the simulated null distribution, however, the true $P$ value is 0.025, which is about 10 times less significant.

4. GEOMETRIC EXPLANATIONS AND APPLICATIONS

Our simulation results contain a number of interesting features: (i) zero mixing proportion for the case where only a subset of boundary parameters are estimated positive and highest proportion on the case where all variance parameters are estimated positive and (ii) departure from chi-square distributions within the mixing components. Here we explain these results using theoretical arguments based on the geometry of the parameter space.

The LRT of the null hypothesis $\theta \in \Omega_0$ versus the alternative hypothesis $\theta \in \Omega_1$ uses the LRT statistic $2 \log \left( \sup_{\theta \in \Omega_1} L(\theta) \right) - 2 \log \left( \sup_{\theta \in \Omega_0} L(\theta) \right)$, where $L$ is the likelihood function. According to Theorem 3 in Self and Liang (1987), when $\theta = \theta_0$, under certain regularity conditions, the asymptotic distribution of LRT is the same as the distribution of the LRT for testing $\theta \in C_{\Omega_0}$ versus $\theta \in C_{\Omega_1}$ based on a single observation $Y$, where $Y \sim N(\theta, I^{-1}(\theta_0))$ and $C_{\Omega_0}$ and $C_{\Omega_1}$ are cones approximating $\Omega_0$ and $\Omega_1$ at $\theta_0$. This latter LRT takes the form $\inf_{\theta \in C_{\Omega_0}} (Y - \theta)^T I(\theta_0) (Y - \theta) - \inf_{\theta \in C_{\Omega_1}} (Y - \theta)^T I(\theta_0) (Y - \theta)$, so that, defining $Z = Y - \theta_0$, the LRT is equivalent to

$$LRT = \inf_{\theta \in C_{\Omega_0} - \theta_0} (Z - \theta)^T I(\theta_0) (Z - \theta) - \inf_{\theta \in C_{\Omega_1} - \theta_0} (Z - \theta)^T I(\theta_0) (Z - \theta),$$

where $C_{\Omega_0} - \theta_0 = \{ \theta - \theta_0 : \theta \in C_{\Omega_0} \}$ and $Z \sim N(0, I^{-1}(\theta_0))$ under the null hypothesis. For the special case when $I(\theta_0)$ is the identity matrix $I$,

$$LRT = \inf_{\theta \in C_{\Omega_0} - \theta_0} \| Z - \theta \|^2 - \inf_{\theta \in C_{\Omega_1} - \theta_0} \| Z - \theta \|^2. \tag{4.2}$$

So in the case $I(\theta_0) = I$, LRT is the difference between the squared (Euclidean) distance from $Z$ to $C_{\Omega_0} - \theta_0$ and squared distance from $Z$ to $C_{\Omega_1} - \theta_0$.

Below, we will mainly write expressions explicitly for the case of $k = 2$ traits with the information matrix $I(\theta_0)$ being the identity matrix, which reveals the essential ideas with minimal notational complexity.

4.1 Mixing probabilities

A way of thinking about the problem that is tempting but leads to incorrect conclusions, including the asymptotic distribution incorrectly stated in previous papers, is as follows. Denote the parameters of interest by $\theta = (\theta_1, \theta_2) = (a_{11}, a_{22}) = (a_1^2, a_2^2)$. The null hypothesis $\Omega_0$ is $\theta = (0, 0)$, and the alternative $\Omega_1$ is the first quadrant $\theta \in \{(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 > 0\}$, and the approximating cones $C_{\Omega_0}$ and $C_{\Omega_1}$ are the same as $\Omega_0$ and $\Omega_1$. Let $Z \sim N(0, I)$ and $LRT = \inf_{\theta \in C_{\Omega_0}} \| Z - \theta \|^2 - \inf_{\theta \in C_{\Omega_1}} \| Z - \theta \|^2 = \| Z \|^2 - \inf_{\theta \in \Omega_1} \| Z - \theta \|^2 = \| Z \|^2 - \| Z - \hat{\theta} \|^2$, where $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ is the vector $\theta$ that achieves the infimum in $\inf_{\theta \in \Omega_1} \| Z - \theta \|^2$, that is, the projection of $Z$ onto the first quadrant $\Omega_1$. The random point
$Z = (Z_1, Z_2)$ falls in each of the 4 quadrants (regions 1-4 in Figure 2) with equal probability 1/4. When $Z$ falls in the first quadrant (region 1, probability 1/4), LRT $= \|Z\|^2$ (because $\hat{\theta}$ is simply $Z$ itself, so that $\|Z - \hat{\theta}\|^2 = 0$), which is distributed as $\chi^2_2$, and both $\hat{\theta}_1$ and $\hat{\theta}_2$ are positive. When $Z$ lies in the second quadrant, $(\hat{\theta}_1, \hat{\theta}_2) = (0, Z_2)$, so that LRT $= \|Z\|^2 - Z_2^2 = Z_2^2 \sim \chi^2_1$, and only $\theta_2$ is estimated positive. Symmetric logic applies to the fourth quadrant, so that LRT $= Z_1^2 \sim \chi^2_1$, and only $\theta_1$ is estimated positive. When $Z$ lies in the third quadrant, the projection $(\hat{\theta}_1, \hat{\theta}_2) = (0, 0)$, so that LRT $= \|Z\|^2 - \|Z\|^2 = 0$ and both parameters are estimated zero. In summary, LRT is distributed as the mixture $\frac{1}{4}X_0^2 + \frac{1}{2}X_1^2 + \frac{1}{4}X_2^2$, with the different mixture components determined by the number of parameters estimated positive.

The above reasoning, however, ignores the covariance parameter $a_{12}$. For simplicity, let us focus our attention on the parameters of interest $a_{11}, a_{22}$, and $a_{12}$ and ignore the nuisance parameters in the model that describe the polygenic and environmental effects. The single-factor model is constrained according to $a_{12} = a_1a_2$. That is, letting $\theta = (\theta_1, \theta_2, \theta_3) = (a_{11}, a_{22}, a_{12}) = (a_1^2, a_2^2, a_1a_2)$, the constraint in the single-factor model is expressed by $\theta_3 = \pm \sqrt{\theta_1\theta_2}$. So for this problem, the null hypothesis is $\Omega_0 = \{\theta \in \mathbb{R}^3 : (\theta_1, \theta_2, \theta_3) = (0, 0, 0)\}$ and the alternative hypothesis is $\Omega_1 = \{\theta \in \mathbb{R}^3 : \theta_1 \geq 0, \theta_2 \geq 0, \theta_3 = \pm \sqrt{\theta_1\theta_2}\}$. Here we are thinking of the parameter space as a curved surface in 3-dimensional space, not just a single quadrant in the 2-dimensional space $\{(a_{11}, a_{22})\}$. Figures 3(a) and (b) show this surface from different angles. To apply Theorem 3 of Self and Liang (1987) to this problem, we note that the hypotheses $\Omega_0 = \{\theta = (0, 0, 0)\}$ and $\Omega_1$ are already cones, so that the approximating cones are $C_{\Omega_0} = \Omega_0$ and $C_{\Omega_1} = \Omega_1$. Suppose for illustration that the information matrix under the null hypothesis, $I(0, 0, 0)$, is the identity matrix $I = I_3$. A general case will be discussed in Section 4.3.

The above problem arises from the original genetic problem by using its asymptotic equivalence to a problem with a single observation from a Gaussian distribution, ignoring nuisance parameters, and considering the special case where the information matrix is the identity. Assume the null hypothesis holds. As in (4.2), the asymptotic distribution of the LRT is the same as the distribution of LRT $= \inf_{\theta \in \Omega_0} \|Z - \theta\|^2 - \inf_{\theta \in \Omega_1} \|Z - \hat{\theta}\|^2 = \|Z\|^2 - \inf_{\theta \in \Omega_1} \|Z - \theta\|^2 = \|Z\|^2 - \|Z - \hat{\theta}\|^2$, where $Z \sim N(0, I)$ and $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ is the vector $\theta$ that achieves the infimum in $\inf_{\theta \in \Omega_1} \|Z - \theta\|^2$, that is, the projection of $Z$ onto the alternative hypothesis surface $\Omega_1$.

The distribution of LRT is a mixture of components that are distinguished by the number of variance parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ that are estimated positive (which, by the constraint that defines $\Omega_1$, also determines whether or not $\hat{\theta}_3$ is estimated nonzero). As before, let $n$ be the number of variance parameters $\hat{\theta}_1$ and

Fig. 2. The parameter space $\theta \in \{(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 > 0\}$.
Asymptotic null distribution of LRTs in multivariate variance components models

Fig. 3. Plots (a) and (b) are the parameter space of points \( \theta = (\theta_1, \theta_2, \theta_3) \) such that \( \theta_3 = \pm \sqrt{\theta_1 \theta_2} \). Plot (c) shows the boundary of the polar cone of the parameter space in blue. Plot (d) shows the points \( Z \) that give small LRT values (\( \leq 0.02 \)) in green. Plot (e) shows the points \( Z \) giving \( \nu = 0 \) when \( Z \sim N(0, I) \). Plot (f) shows the points \( Z \) with \( \nu = 0 \) when the covariance of \( Z \) is not the identity matrix, but a general covariance matrix obtained from the Fisher information matrix of a genetic model.

\( \hat{\theta}_2 \) that are estimated positive, that is, the number of positive components among \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \); the possible values for \( \nu \) are 0, 1, and 2.

The determination of \( \nu \) can be explained using the polar cone of the parameter space \( \Omega_1 \). The polar cone of \( \Omega_1 \) is defined to be the set \( P_{\Omega_1} = \{ Z \in \mathbb{R}^3 : Z \cdot \theta \leq 0 \forall \theta \in \Omega_1 \} \). Figure 3(c) depicts the boundary of the polar cone of the parameter space.

When \( Z \) is in the polar cone, we claim \( \nu = 0 \). To see this, assume \( Z \in P_{\Omega_1} \) and note that if \( \theta \in \Omega_1 \), then \( Z \cdot \theta \leq 0 \) by definition of the polar cone, which implies \( \| Z - 0 \|^2 \leq \| Z - \theta \|^2 \). Thus, the origin is the point in \( \Omega_1 \) that is closest to \( Z \), which gives \( \nu = 0 \).

When \( Z \) lies outside the polar cone, we claim \( \nu = 2 \) with probability 1. First, by definition, for any point \( Z \) outside the polar cone \( P_{\Omega_1} \), there exists a vector in \( \Omega_1 \) that has positive inner product with \( Z \). This implies that the projection of \( Z \) onto that vector is closer to \( Z \) than the origin is, proving \( \nu \neq 0 \). Second, we want to show \( \nu \neq 1 \) with probability 1. Note that \( \nu = 1 \) can occur when \( Z \) lies on the \( Z_3 = 0 \) plane in the first, second, and fourth quadrants, which has probability 0. We claim that whenever \( Z_3 \) is nonzero (which occurs with probability 1), \( \nu \) cannot be 1. This is shown in each quadrant in the following paragraphs.

Suppose \( (Z_1, Z_2) \) falls in the first quadrant, that is, \( Z_1 > 0 \) and \( Z_2 > 0 \). Without loss of generality, suppose \( Z_3 > 0 \) (an analogous argument will handle the case \( Z_3 < 0 \) and \( Z_1 > Z_2 \). Then the point in
\[ \Omega_1 \] having \( v = 1 \) that is closest to \( Z \) is \((Z_1, 0, 0)\). However, it is obvious that for sufficiently small \( \varepsilon > 0 \), the point \((Z_1, \varepsilon, \sqrt{Z_1^2 - \varepsilon^2})\) \( \in \Omega_1 \), which has \( v = 2 \), is closer to \( Z \) than \((Z_1, 0, 0)\) is, proving that \( v \) cannot be 1. Thus, we have shown that when \((Z_1, Z_2)\) falls in the first quadrant, we have \( v = 2 \) with probability 1.

Suppose \((Z_1, Z_2)\) falls in the second quadrant, that is, \(Z_1 < 0 \) and \(Z_2 > 0\). Again suppose \( Z_3 > 0 \), and observe that the point in \( \Omega_1 \) having \( v < 1 \) that is closest to \( Z \) is \((0, Z_2, 0)\). Define \( \theta(\varepsilon) = (\varepsilon, Z_2, \sqrt{\varepsilon Z_2}) \), and define \( D(\varepsilon) \) to be the squared distance \( \|Z - \theta(\varepsilon)\|^2 \). The derivative \( D'(\varepsilon) = 2(\varepsilon - Z_1) + Z_2 - Z_3 \sqrt{Z_2/\varepsilon} \) approaches \(-\infty\) as \( \varepsilon \downarrow 0 \). Therefore, for sufficiently small positive \( \varepsilon \), we must have \( D(\varepsilon) < D(0) \), so that the point \( \theta(\varepsilon) \) is closer to \( Z \) than \( \theta(0) = (0, Z_2, 0) \) is, which implies that \( v < 1 \) cannot hold. The case where \((Z_1, Z_2)\) lies in the fourth quadrant is analogous to the case of the second quadrant. Thus, for both the cases of the second and fourth quadrants, we have \( v = 2 \) with probability 1.

In summary, when \( Z \) lies inside the polar cone, \( v = 0 \), and when \( Z \) lies outside the polar cone, \( v = 2 \) with probability 1. To illustrate this, we generated 10,000 vectors \( Z \) from the 3-dimensional Gaussian distribution with mean zero and identity covariance matrix. For each \( Z \), we calculated \( v \) by estimating \( \hat{\theta} \) with the optimization being performed under the constraints \( \theta_1 \geq 0, \theta_2 \geq 0, \theta_3 = \pm \sqrt{\theta_1 \theta_2} \), using the program Mx, which applies here because the projection \( \hat{\theta} \) is also the maximum likelihood estimate in this problem. The red points in Figure 3(e) are the points \( Z \) having \( v = 0 \), which form the polar cone of the parameter space.

### 4.2 Departures from chi-squared distributions

In Section 3.2, one of the departures from the anticipated \( \chi^2 \) distributions we observed was an excessive number of LRT values near 0. These small LRT values result from points \( Z \) that are located close to the boundary of the polar cone, where the difference between the squared distance from \( Z \) to the origin and the squared distance from \( Z \) to the parameter space is small. To visualize this, using the same method of generating \( Z \) as described in Section 4.1, we simulated 10,000 vectors \( Z \) from \( N(0, I_3) \), and for each \( Z \) we performed the LRT for testing \( H_0: \theta = 0 \) versus \( H_1: \theta \neq 0 \) with the optimization being performed under the constraints \( \theta_1 \geq 0, \theta_2 \geq 0, \theta_3 = \pm \sqrt{\theta_1 \theta_2} \). Then we identified the points \( Z \) showing small LRT values (\( \leq 0.02 \)) and plotted them as green in Figure 3(d). We see that they are located very close to the outer surface of the polar cone.

Another type of departure we observed from the simulation results, particularly in models with larger numbers of traits, was a pattern that the distribution was shifted to the right compared to the anticipated chi-square distribution. This can be explained by the fact that the parameter space consists of 2 surfaces joined at an angle of less than 180°. The angle between the 2 surfaces (one blue and the other red; see Figures 3(a) and (b)) in our parameter space \( \theta_3 = \pm \sqrt{\theta_1 \theta_2} \) is 70.54° in the direction \( \theta_1 = \theta_2 \). When \( Z \) lies between these 2 surfaces, the projection of \( Z \) onto the parameter space involves a choice of whether the point is closer to one surface or the other, and the availability of this choice affects the distribution of LRT. An analogous example for illustration in 2 dimensions can help explain this problem simply. Figure 4 shows 2 different cases. In the first, the parameter space is a 1-dimensional straight line, and in the second, the parameter space is 2 lines joined at the origin with an angle less than 180°, say, 90°. For both hypothesis-testing problems, the null hypothesis is the origin. For the first case, the LRT is exactly \( Z_1^2 \) which is distributed as \( \chi^2 \). On the other hand, in the second plot, consider the points \( A \) and \( B \) lying between the 2 lines forming the parameter space. The point \( A \) has LRT = \( Z_2^2 \), while the point \( B \) has LRT = \( Z_1^2 \); the choice of whether LRT is \( Z_1^2 \) or \( Z_2^2 \) depends on whether the point is closer to the first line or the second. Thus, in the first quadrant, LRT is \( \max(Z_1^2, Z_2^2) \), which in general is larger than, say, \( Z_1^2 \), which is distributed as \( \chi^2 \). Thus, in the first quadrant, LRT does not have a \( \chi^2 \) distribution, but rather a distribution that is stochastically larger than \( \chi^2 \). Note that if we change the parameter space
by decreasing the angle between the lines, the LRT in the region between the lines becomes closer to $Z_1^2 + Z_2^2$, which is distributed as $\chi^2_2$. Returning to the 3-dimensional problem, as the angle between the surfaces becomes closer to 0°, the distribution of LRT in the region between the surfaces would be closer to $\chi^2_3$ because the difference between the squared distance from $Z$ to the origin, which is $Z_1^2 + Z_2^2 + Z_3^2$, and the squared distance from $Z$ to the alternative space, which is close to zero, would be close to $Z_1^2 + Z_2^2 + Z_3^2$, which is distributed as $\chi^2_3$.

### 4.3 General case of $I(\theta_0) \neq I$ and genetic models

**General case of $I(\theta_0) \neq I$.** So far, for simplicity, we have focused on the special case where the Fisher information $I(\theta_0)$ was the identity matrix $I$. For the general case when $I(\theta_0) \neq I$, the results of the Self and Liang theorem stated in (4.1) show that the asymptotic distribution of the LRT can be analyzed as above (involving a difference between the squared distance from $Z$ to the origin and the squared distance from $Z$ to the parameter space), with the Euclidean distance replaced by the more general Mahalanobis distance $\|Z - \theta\|_M^{-1} = (Z - \theta)^T I(\theta_0)(Z - \theta)$. This distance takes into account the covariance of $Z$, where $Z \sim N(0, I^{-1}(\theta_0))$. For the multivariate normal model, $I(\theta_0)$ can be obtained from the formula (Searle *et al.*, 1992)

$$I_{ij}(\theta) = -E \left( \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right) = \frac{1}{2} \text{Tr} \left( \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right).$$

### Generating the asymptotic null distribution in a genetic model

The results in (4.1) and (4.3) above allow us to simulate from the asymptotic null distribution in genetic models that have general Fisher information matrices. For example, we did this for the $k = 2$ trait genetic model specified in Section 3, that is, the model with parameter values $g_{11} = g_{22} = 0.4, e_{11} = e_{22} = 0.6, e_{12} = 0.04$ and $e_{12} = 0.06$. After obtaining $\Sigma$ by plugging in the specified values in $G$ and $E$, using (4.3) we calculated the $9 \times 9$ Fisher information matrix for the full set of parameters $(a_{11}, a_{22}, a_{12}; g_{11}, g_{22}, g_{12}; e_{11}, g_{22}, g_{12})$. After extracting the $3 \times 3$ submatrix $V$ of the inverted Fisher information matrix corresponding to the parameters of interest $\theta = (a_{11}, a_{22}, a_{12})$, we generated 10,000 vectors $Z$ from $N(0, V)$. For each $Z$, we performed the LRT for testing $H_0: \theta = 0$ versus $H_1: \theta \neq 0$ under the constraints $\theta_1 \geq 0, \theta_2 \geq 0$, and $\theta_3 = \pm \sqrt{\theta_1 \theta_2}$ using Mx. In this application of Mx, we assigned fixed values for the known covariances in $V$, and Mx optimized the likelihood only over mean parameters. We also determined $\nu$ using the same technique of comparing a full model to a partially constrained model as described in Section 3.

According to the results, the mixing proportions for $\nu = 0, \nu = 1$, and $\nu = 2$ were 0.1477, 0, and 0.8523, and critical values for $\alpha = 0.01$ and $\alpha = 0.05$ were 8.713 and 5.405, which were quite similar to
Fig. 5. Asymptotic null distributions of tests under 2 models with $k = 4$ traits.

the results in Table 2. To confirm that the results in this section were consistent with the genetic simulation results for the $k = 2$ trait null distribution in Section 3, a K-S test gave a $P$ value of 0.2063.

Plot (f) in Figure 3 shows the points $Z$ with $v = 0$ colored as red, filling the polar cone. The contrast between the polar cones in plots (e) and (f) illustrates the effect of having a Fisher information matrix $I(\theta_0) \neq I$.

Different sets of nuisance parameter values. To investigate whether and how the asymptotic null distribution is influenced by different choices of the nuisance parameters, we selected 10 sets of nuisance parameters in $G$ and $E$ for each $k$-trait model, using inverse Wishart distributions to choose matrices that were well spread out over the range of possible covariance matrices. We used the same method described in Section 4.3 to generate the asymptotic null distribution for each set of nuisance parameters. In order to assess whether the asymptotic null distributions generated using different sets of nuisance parameters were different, we conducted pairwise K-S tests. Most of the results did not show significant differences, but the asymptotic distributions for a few of the models were found to have highly significant departures from the rest, showing that the asymptotic distribution does depend on nuisance parameters. For example, Figure 5 shows a case having K-S $P$ value $2.20 \times 10^{-16}$.

However, although differences among asymptotic distributions of LRT statistics are real, their practical effect on $P$ values for LRT tests seems to be quite minor. For example, even the pair of asymptotic distributions in Figure 5, which gave the smallest K-S $P$ value we observed, are quite similar.

5. Proposed method for generating asymptotic null distributions and calculating $P$ values

As we have shown, asymptotic null distributions of multivariate linkage tests have several complex features: the mixing probabilities are not binomial distributions and the mixture components show severe departures from chi-square distributions. The asymptotic null distributions also varied depending on nuisance parameter values to some extent. These complexities in the asymptotic null distribution raise challenges for the assessment of significance of multivariate linkage findings. Empirical methods for obtaining
of the parameters of interest \( \theta \) consists of the parameters in parameter values. A method that is applicable to the more realistic situation where we are given data and do not know nuisance in Section 4.3, although there we assumed knowledge of the nuisance parameter values. Here we specify a method that is applicable to the more realistic situation where we are given data and do not know nuisance parameter values.

The description of the method uses the following notation. In a \( k \)-trait model, the total parameter set consists of the parameters in \( A = (a_{ij}), G = (g_{ij}), \) and \( E = (e_{ij}) \), where each matrix has \( m = \frac{k(k+1)}{2} \) parameters. The full parameter set contains all \( 3m \) parameters in \( A, G, \) and \( E \). The parameters in \( A \) consist of the parameters of interest \( \theta = (\theta_1, \theta_2, \ldots, \theta_m) \), with constraints of the form \( \theta_i \geq 0 \) for \( i \leq k \) and \( \theta_i = \pm \sqrt{\theta_p(i) \theta_q(i)} \) for \( i > k \), where we write \( \theta_i = a_{p(i),q(i)}(i) \) for \( i > k \). For example, for the models with \( k = 3 \) traits, the parameters of interest are \( \theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) \) = \( (a_{11}, a_{22}, a_{33}, a_{12}, a_{13}, a_{23}) = (a_1^2, a_2^2, a_3^2, a_1 a_2, a_1 a_3, a_2 a_3) \), and the constraints are \( \theta_1 \geq 0, \theta_2 \geq 0, \theta_3 \geq 0, \theta_4 = \pm \sqrt{\theta_1 \theta_2}, \theta_5 = \pm \sqrt{\theta_1 \theta_3} \), and \( \theta_6 = \pm \sqrt{\theta_2 \theta_3} \).

### 5.1 Method

Suppose we have a given data set and have calculated an LRT statistic for a multivariate linkage test for that data set, and we wish to calculate a \( P \) value.

**STEP 1:** Estimate the Fisher information from the given data.

**STEP 1(a):** Under the null hypothesis, estimate all parameters in the model—\( \hat{G} \) and \( \hat{E} \). Use these to form the estimate \( \hat{\Sigma} = \begin{bmatrix} \hat{G} + \hat{E} & \frac{1}{2} \hat{G} \\ \frac{1}{2} \hat{G} & \hat{G} + \hat{E} \end{bmatrix} \).

**STEP 1(b):** Using \( \hat{\Sigma} \) and the formula (4.3), calculate the \( (3m) \times (3m) \) Fisher information matrix for the full parameter set and invert it.

**STEP 1(c):** Extract the \( m \times m \) submatrix of the inverted Fisher information matrix corresponding to, the parameters of interest, and denote it as \( V \).

**STEP 2:** Generate \( N \) random vectors \( Z \) from the multivariate Normal distribution \( N(0, V) \). An appropriate choice of \( N \) depends on the desired precision level.

**STEP 3:** For each \( Z \), perform the LRT for the null hypothesis \( H_0: \theta = 0 \) in the model \( Z \sim N(\theta, V) \), under the constraints \( \theta_i \geq 0 \) for \( 1 \leq i \leq k \) and \( \theta_i = \pm \sqrt{\theta_p(i) \theta_q(i)} \) for \( i > k \). Here the covariance matrix \( V \) is fixed at the value calculated in STEP 1(c), and the test is performed by optimizing the likelihood only over the mean parameters in the vector \( \theta \).

**STEP 4:** The \( P \) value is calculated as the fraction of LRT values that are larger than or equal to the LRT value calculated from the given data.

An example Mx script that performs the LRT in STEP 3 above is available in the web supplement at http://www.stat.yale.edu/~sh437/mxExample.

### 5.2 Applied example in an illustrative data set

In this section, we apply our proposed method and other well-known empirical methods for assessing \( P \) values to an illustrative data set and compare the times for completing the tasks. To create the illustrative data set, we simulated 2000 independent sib-pairs with 2 traits generated from the bivariate model with the same parameter values used in Section 3. For each sibling, we simulated a segment of chromosome with length 80 cM, having 16 markers located 5 cM apart. Each marker had 8 equally frequent alleles, and IBD
sharing proportions $\pi_{1,12}$ were estimated by Merlin (Abecasis and others, 2001) with grid size 2.5 cM. We prespecified a single test point at 22.5 cM on the chromosome where the LRT would be conducted.

We applied 3 well-known methods for obtaining empirical critical values and $P$ values: gene dropping (Terwilliger and others, 1993), permutation (Wan and others, 1997), and the bootstrap (Marlow and others, 2003). Keeping the phenotype data fixed, the gene-dropping method simulates genotypes on founders’ chromosomes using estimated marker allele frequencies and then segregates the chromosomes to offspring using marker recombination fraction information. For the permutation method, we fixed the trait values and permuted IBD estimates among families. The (parametric) bootstrap, which is not as commonly applied in linkage analysis as the previous 2 methods, estimates parameters from a given data set, treats the estimates as if they were true parameters, and generates replicate data sets, keeping the genotypes the same as in the original data. We applied the above 3 methods to our illustrative data set, generating 10 000 replicates for each method using R and running LRTs at the prespecified test point for each replicated data set using Mx. For STEP 2 in our proposed method, we used $N = 10 000$.

In this experiment, our method took 56 min to complete the task, while the gene-dropping, bootstrap, and permutation methods took 5893, 3182, and 1611 min, respectively (see Table 3). In order to measure the accuracy of the null distribution estimations, we estimated the type 1 error rates for $\alpha = 0.01$ and 0.05 based on the critical values shown in Table 2. The results in Table 3 show reasonably accurate performances of all methods in controlling type 1 errors. The K-S tests for comparing these 4 distributions to the reference distribution obtained in Section 3.2 were conducted, and $P$ values were 0.2063, 0.654, 0.2742, and 0.4982 for the proposed method, gene-dropping, bootstrap, and permutation, respectively, none of them showing severe departures.

The empirical methods took longer than our method since they require generating many replicate data sets with each data set including as many observations as in the given data set (2000 in our experiment), whereas our method requires generating only one Gaussian observation in each data set. Also, each Mx optimization took longer for the empirical methods than for our proposed method, again because our proposed method applies Mx to data sets of just one Gaussian observation.

### 6. Discussion

Evaluating the null distribution of a test statistic is of course a prerequisite for valid $P$ values. The previously stated asymptotic null distribution, which gave strongly anticonservative $P$ values, has been used in several studies. In fact, some researchers (Amos and de Andrade, 2001; Evans and others, 2004; Marlow and others, 2003; Monaco, 2007) have mentioned the possibility that the previously stated asymptotic null distribution might be incorrect. However, clear statements confirming and quantifying these discrepancies and explaining their nature and origin have not appeared previously, and some researchers (Evans and others, 2004; Monaco, 2007) have expressed the desirability of clarifying the asymptotic null distribution of this test. A part of the purpose of this paper is to provide such clarification. Our simulations establish that there is indeed a real problem underlying discrepancies that had been mentioned as a puzzle in these

---

**Table 3. Comparing the performances of proposed method and 3 empirical methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Computation time (min)</th>
<th>Type 1 error rates $\alpha = 0.01$</th>
<th>Type 1 error rates $\alpha = 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method</td>
<td>56</td>
<td>0.0107</td>
<td>0.0483</td>
</tr>
<tr>
<td>Gene dropping</td>
<td>5893</td>
<td>0.0099</td>
<td>0.0487</td>
</tr>
<tr>
<td>Parametric bootstrap</td>
<td>3182</td>
<td>0.0100</td>
<td>0.0451</td>
</tr>
<tr>
<td>Permutation</td>
<td>1611</td>
<td>0.0121</td>
<td>0.0547</td>
</tr>
</tbody>
</table>
Asymptotic null distribution of LRTs in multivariate variance components models

previous papers. A device that enabled our simulations to quantify the discrepancies was to separate the mixture distribution into components defined by the number of variance parameters estimated positive. Our geometric arguments help explain how and why the discrepancies occur and give insight into the true nature of the asymptotic distribution.

It is of interest to understand, for various possible derivations that would falsely lead to the previously stated asymptotic distribution, exactly where the reasoning fails. For the simple case of \( k = 2 \) traits, a belief that the asymptotic null distribution is simply \( 0.25 \chi^2_0 + 0.5 \chi^2_1 + 0.25 \chi^2_2 \) would follow from the incorrect reasoning given in Section 4.1. This incorrect reasoning views the parameter space as a quadrant in the 2-dimensional space of possible values for the variance parameters and also makes the further mistake of assuming the information matrix is the identity. Even correcting this reasoning to allow a general information matrix leads only to asymptotic distributions of the form \((\frac{1}{2} - p) \chi^2_0 + \frac{1}{2} \chi^2_1 + p \chi^2_2\); see Case 7 in Self and Liang (1987) for the derivation of this formula. However, this mixture of chi-square distributions is still evidently incorrect, as we have seen that the correct asymptotic distribution has zero mixing probability for the \( \chi^2_1 \) component. All the above reasoning is invalidated by the observation that the parameter space is actually not isomorphic to the first quadrant in 2 dimensions; in fact, for each point \((a_{11}, a_{22})\) in the quadrant, there are 2 different points in the parameter space, with one point having \( a_{12} = \sqrt{a_{11}a_{22}} \) and the other having \( a_{12} = -\sqrt{a_{11}a_{22}} \). Since there are 2 distinct distributions for each point in the quadrant, a reasonable attempt to correct this topological problem is to view the parameter space as 2 quadrants. For example, as in Figure 6, we could consider the parameters to be \( a_1 \) and \( a_2 \) where \( a_1 \) can take any real value (including negative numbers) and \( a_2 \) is constrained to be nonnegative. These 2 parameters determine the covariance \( a_{12} \) by the relationship \( a_{12} = a_1a_2 \). More precisely, to make the problem identifiable, we should remove the ray of points \((a_1, a_2): a_1 < 0, a_2 = 0\) from the parameter space since the 2 points \((a_1, 0)\) and \((-a_1, 0)\) correspond to the same distribution. Although this modified half-plane gives a correct one-to-one correspondence with the set of distributions in our problem, attempting to apply the reasoning of Self and Liang (1987) to this situation (see their Case 6) would lead to a conjectured asymptotic null distribution of \( \frac{1}{2} \chi^2_1 + \frac{1}{2} \chi^2_2 \), which is also incorrect. In fact, the assumptions required for the results of Self and Liang are questionable in our problem in this 2-dimensional formulation since the modified half-plane is not closed and the information matrix at the null hypothesis \((a_1, a_2) = (0, 0)\) is the zero matrix.

In this paper, we focused on the model with the complete pleiotropic constraint. This model has been used in many studies instead of the general model that includes additional correlation parameters (Evans, 2002; Evans and Duffy, 2004; Evans and others, 2004; Loo and others, 2004; Marlow and others, 2003; Monaco, 2007; Vogler and others, 1997). One disadvantage of the general model is that optimization is more difficult (Loo and others, 2004). Also, due to the increased number of parameters, the general model has more degrees of freedom than the complete pleiotropic model. A recent study (Han and others, 2009) showed in a related but different context that using the general model is often less powerful than using the complete pleiotropic model due to the increased degrees of freedom.

Fig. 6. The modified half-plane parameter space.
Although the focus of the current study is on the complete pleiotropic model, our approaches and results can be extended to investigate the general model. The asymptotic null distribution of the LRT in the general model has also been stated to be a mixture of chi-square distributions with binomial mixing probabilities (e.g. Amos and de Andrade, 2001; Amos and others, 2001; SOLAR, 2008). Although the general model does not put explicit constraints on covariance parameters as in the complete pleiotropic model, covariance parameters still have implicit constraints, for example, ensuring that the correlation coefficients lie between $-1$ and $1$. For this reason, applying the binomial mixture of chi-square distributions, which does not take the implicit constraints into account, is incorrect. We are currently studying this issue and preparing a separate paper extending the approaches provided by our present study to the general model.

ACKNOWLEDGMENTS

We are grateful to Dr Elena Grigorenko for inspiring our interest in these tests and for helpful advice and pointers to the literature and to Dr Michael Neale for help with Mx. Conflict of Interest: None declared.

FUNDING

Yale Center for High Performance Computation in Biology and Biomedicine (National Institutes of Health [NIH] grant RR19895-02); NIH (R01 DC007665).

REFERENCES


EVANS, D. M. (2002). The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables. The American Journal of Human Genetics 70, 1599–1602.


Asymptotic null distribution of LRTs in multivariate variance components models


[Received September 15, 2008; revised May 6, 2009; accepted for publication October 27, 2009]