Likelihood Ratio Testing for Admixture Models with Application to Genetic Linkage Analysis

Chong-Zhi Di, Fred Hutchinson Cancer Research Center
Kung-Yee Liang, Johns Hopkins University

Available at: https://works.bepress.com/di/10/
Likelihood Ratio Testing for Admixture Models with Application to Genetic Linkage Analysis

Chong-Zhi Di*
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center
1100 Fairview Ave N, M2-B500, Seattle, WA 98109
*email: cdi@fhcrc.org

and

Kung-Yee Liang**
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe Street, Baltimore, MD 21205
*email: kyliang@jhsph.edu

SUMMARY: We consider likelihood ratio tests (LRT) and their modifications for homogeneity in admixture models. The admixture model is a two-component mixture model, where one component is indexed by an unknown parameter while the parameter value for the other component is known. This model is widely used in genetic linkage analysis under heterogeneity in which the kernel distribution is binomial. For such models, it is long recognized that testing for homogeneity is nonstandard, and the LRT statistic does not converge to a conventional $\chi^2$ distribution. In this paper, we investigate the asymptotic behavior of the LRT for general admixture models and show that its limiting distribution is equivalent to the supremum of a squared Gaussian process. We also discuss the connection and comparison between LRT and alternative approaches such as modifications of LRT and score tests, including the modified LRT (Fu et al., 2006). The LRT is an omnibus test that is powerful to detect general alternative hypothesis. In contrast, alternative approaches may be slightly more powerful to detect certain type of alternatives, but much less powerful for others. Our results are illustrated by simulation studies and an application to a genetic linkage study of schizophrenia.

KEY WORDS: Admixture models; likelihood ratio test; genetic linkage analysis
1. Introduction

1.1 Admixture models

In this paper, we consider likelihood ratio testing for homogeneity in admixture models, with a focus on genetic linkage analysis. With a kernel distribution $p(y; \theta)$ indexed by $\theta$, an admixture model has the probability density function of the form

$$f(y; \delta, \theta) = (1 - \delta) p(y; \theta_0) + \delta p(y; \theta),$$

(1)

where the population comprises two components from the same parametric family $p(\cdot; \theta)$, with proportions $1 - \delta$ and $\delta$, respectively. The first component is indexed by a known parameter value $\theta_0$, which often represents a particular model of interest, while the parameter for the second component is unknown and to be estimated from the data. The kernel function $p(\cdot; \theta)$ could be any parametric distribution, such as Gaussian, binomial, exponential or Poisson distributions. We assume that the parameter space for $(\delta, \theta)$ is $\Omega = [0, 1] \times \Theta$, where $\Theta \subset R$ is compact and $\theta_0 \in \Theta$. The admixture model is different from a typical two component mixture model, which assumes that parameters for both components are unknown.

In admixture models, the first component $p(\cdot; \theta_0)$ is a particular submodel of the second component $p(\cdot; \theta)$, and usually has a special feature of scientific interest. One is often interested in testing a simple homogeneous model $p(\cdot; \theta_0)$ versus admixture alternative. With (1), the null hypothesis can be specified as either $\theta = \theta_0$, or equivalently $\delta = 0$. Under the specification of $\theta = \theta_0$, the parameter $\delta$ disappears and any value of $\delta$ gives the same null distribution. Similarly, $\theta$ disappears under the specification of $\delta = 0$. In other words, each value in the set $\Omega_0 := \{(\delta, \theta) : \delta = 0 \text{ or } \theta = \theta_0\}$ represents exactly the same distribution. Thus, testing $H_0 : \delta = 0 \text{ or } \theta = \theta_0$ is a nonstandard problem that involves nonidentifiability under the null.

Admixture models have been widely used in public health and biomedical studies to account for possible heterogeneity in the population. For example, Davies (1977) considered an admixture model with exponential kernels. Another application is in genetic linkage
analysis, where admixture models with binomial kernels have been used to account for
genetic heterogeneity (Smith, 1963). In the following, we will focus on the admixture model
for linkage analysis, but the arguments and results carry over to general kernels.

1.2 Genetic linkage analysis

We start with a brief introduction of the genetic linkage model, and refer readers to Ott
(1999), Thomas (2004) and Fu et al. (2006) for detailed descriptions. In this paper, we
focused on sibship genetic linkage studies, while linkage studies with fewer but large extended
pedigrees are out of our scope. For simplicity, we focus on two point linkage analysis, which
studies the cosegregation of the disease gene and a genetic marker. Basically, the closer two
genes are on the same chromosome, the less likely that they would be separated during
meiosis. An offspring is called nonrecombinant if it inherits both maternal (or paternal)
alleles at these two loci, and recombinant otherwise. The recombinational fraction (denoted
as $\theta$) is the percentage of offspring that are recombinants, and can take values from 0 to 0.5.
The two loci are strongly linked if $\theta$ is close to 0, and not linked if $\theta = 0.5$. Offspring from one
family can be divided into two groups, recombinants and nonrecombinants, and it is natural
to use a binomial distribution to model such data. However, in human pedigree studies, it
is sometimes not possible to ascertain whether or not a child is recombinant. Depending on
the availability of such information, two different cases occur, phase known (PK) and phase
unknown (PU).

In the PK case, one could record the number of recombinants, $Y$, in a family with $K$
offspring. If there is possible linkage among all families, $Y$ has a simple binomial distribution,

$$p(y; \theta) = \Pr(Y = y) = \binom{K}{y} \theta^y (1 - \theta)^{K-y},$$

where $\theta \in [0, 0.5]$ describes the magnitude of linkage. In the PU case, one could only observe
that there are two groups of offspring, $Y$ and $K - Y$, but could not tell whether each group
is recombinant or not. Under this situation, it is commonly assumed that there is a 50% chance
that the first group (with $Y$ offspring) is recombinant and a 50% chance that it is
not. Thus, \( Y \) follows a mixture of two binomial distributions, i.e.,

\[
p(y; \theta) = \Pr(Y = y) = 0.5 \binom{K}{y} \theta^y (1 - \theta)^{K-y} + 0.5 \binom{K}{y} \theta^{K-y} (1 - \theta)^{y}.
\]

In either PK or PU cases, we are interested in testing whether there is statistical evidence for linkage. In Model (2) for the PK case and Model (3) for the PU case, the null hypothesis of no linkage is specified as \( H_0 : \theta = 0.5 \), under which the probability density function reduces to \( p(y; 0.5) = \binom{K}{y} 0.5^K \). Hypothesis testing problems for these models are regular except that the null value \( \theta = 0.5 \) is on the boundary of the parameter space \([0, 0.5]\). Using the general theory developed by Self and Liang (1987), the LRT converges in distribution to a mixture of \( \chi^2 \) distributions, \( 0.5 \chi^2_0 + 0.5 \chi^2_1 \) under \( H_0 \), where \( \chi^2_0 \) is a point mass at 0.

For complex diseases, however, linkage may exist only in a proportion of families, but not in the remaining families. This phenomenon is known as linkage heterogeneity. Smith (1963) proposed using an admixture model to account for such heterogeneity. More precisely, the admixture model in the genetic linkage context has the form

\[
f(y ; \delta, \theta) = (1 - \delta) \binom{K}{y} 0.5^K + \delta \binom{K}{y} \theta^y (1 - \theta)^{K-y}
\]

for the PK case, and

\[
f(y ; \delta, \theta) = (1 - \delta) \binom{K}{y} 0.5^K + \delta \binom{K}{y} \{ 0.5 \theta^y (1 - \theta)^{K-y} + 0.5 \theta^{K-y} (1 - \theta)^{y} \}
\]

for the PU case, where \( \delta \) is the proportion of families with possible linkage. In these two models, the null hypothesis of no linkage can be specified as \( H_0 : \delta = 0 \) or \( \theta = 0.5 \), and the alternative is \( H_1 : 0 < \delta < 1, 0 \leq \theta < 0.5 \). As illustrated in Figure 1 (a), the parameter space is the rectangle \((\delta, \theta) \in [0,1] \times [0,0.5]\) and under the null hypothesis of no linkage, the set of true values include infinitely many values, on two thick solid lines \( \delta = 0 \) and \( \theta = 0.5 \). This hypothesis testing problem is nonstandard due to nonidentifiability under the null, in the sense that \( \theta \) is not identifiable when \( \delta = 0 \) and \( \delta \) is not identifiable when \( \theta = 0.5 \). In addition, two types of nonstandard situations might occur. First, the null parameter values \((\delta = 0 \text{ and } \theta = 0.5)\) are on the boundary of the parameter space. Second, the Fisher information for \( \theta \) evaluated at \( \theta = 0.5 \) and any \( \delta \) is always 0 in the PU case.
It has been recognized that standard asymptotic results for LRT and score tests do not hold in such nonstandard situations. Various hypothesis testing procedures, including variations of the likelihood ratio tests and score tests, have been studied over the last two decades, for instance, Shoukri and Lathrop (1993), Faraway (1993), Chernoff and Lander (1995), Chiano and Yates (1995), Lemdani and Pons (1995), Lemdani and Pons (1997), Liang and Rathouz (1999), Abreu et al. (2002) and Fu et al. (2006). In this paper, we review existing methods, develop asymptotics for the LRT in general admixture models, and compare the LRT to alternative methods, especially the modified LRT proposed by Fu et al. (2006), in terms of statistical power.

2. Likelihood ratio tests

In this section, we first illustrate challenges of statistical inference for admixture models, and then we present asymptotic results for the LRT.

2.1 Challenges for inference

The nonstandard properties of the admixture model under $H_0$ bring challenges to statistical inference, including both parameter estimation and hypothesis tests. Specifically, the null hypothesis can be specified equivalently via each one of the two parameters ($\delta = \delta_0$ or $\theta = \theta_0$), and under either specification, the other parameter ($\theta$ or $\delta$) is not identifiable. Under the null hypothesis, the set of true values $\Omega_0$ contains the union of two one-dimensional subspaces. These non-identifiability problems are to be contrasted with an identifiable regular class where the true value $(\delta_0, \theta_0)$ is unique.

To illustrate, we consider the PK case of genetic linkage model (2) and explore its likelihood functions. Figure 1 displays the contour plots of the expected log-likelihood function under $H_0$ as well as observed log-likelihood functions for two datasets simulated under $H_0$. Figure 1 (a) shows the expected log-likelihood function, which is maximized at the set of true
values, represented by two solid lines ($\delta = 0$ and $\theta = 0.5$). This gives us an idea of the average shape of log-likelihood functions for observed data. The right panels (b) and (c) show observed log-likelihood functions for two simulated datasets, where the black dots represent the maximum likelihood estimates (MLE). One can see that the overall shape of an observed log-likelihood function is similar to that of the expected log-likelihood function, subject to some random variation. In contrast to the regular case where the likelihood takes large values in a small neighborhood of the unique true value, the likelihood function under non-identifiability generally has large values around the region of true values.

In the genetic linkage context, the parameter space is $[0, 1] \times [0, 0.5]$, and the set of null values is the union of $\delta = 0$ and $\theta = 0.5$ as shown by the thick solid lines in Figure 1 (a). However, if one is willing to restrict the parameter space to a subspace of $[0, 1] \times [0, 0.5]$, the LRT problem may become simpler. For example, if one restricts the parameter space of $(\delta, \theta)$ to $[\epsilon_1, 1] \times [0, 0.5]$ (Regions I and IV in Figure 1) for some $0 < \epsilon_1 < 1$, $H_0$ can only be specified as $\theta = 0.5$. The consequence is similar if one restricts the parameter space to $[0, 1] \times [0, 0.5 - \epsilon_2]$ (Regions II and IV in Figure 1) for some $0 < \epsilon_2 < 0.5$. Such restrictions were considered in Lemdani and Pons (1995). If one is willing to make stronger constrain, e.g., fixing $\delta = \delta_1$, the problem becomes identifiable. In this paper, we focus on the LRT with the natural parameter space $[0, 1] \times [0, 0.5]$.

2.2 Asymptotic consistency

As established in Redner (1981), the MLE under non-identifiability is generally not consistent in the strict sense, but is close to the set of true values in large samples. This can be verified in Figure 1, where the MLEs for two simulated datasets are both close to the region of true values but not close to each other. In the admixture model context, we state the consistency result below. This is a special case of Theorem 5 in Redner (1981), so its proof is omitted.
Lemma 1. For the admixture model (1), assume regularity conditions A1-A5 (Appendix A) for the kernel density $p(\cdot; \theta)$. Under the homogeneity null $H_0 : \delta = 0$ or $\theta = \theta_0$, or equivalently $Y \sim p(\cdot; \theta_0)$, the MLEs $\hat{\delta}$ and $\hat{\theta}$ satisfy the following.

1. $(\hat{\delta}, \hat{\theta})$ does not converge in probability, and may not even be uniquely defined;
2. $d_{\Omega_0}(\hat{\delta}, \hat{\theta}) := \inf_{(\delta', \theta') \in \Omega_0} \| (\hat{\delta}, \hat{\theta}) - (\delta', \theta') \|$ converges in probability to 0 as $n \to \infty$;
3. The estimated density $f(\cdot; \hat{\delta}, \hat{\theta})$ converges to $p(\cdot; \theta_0)$ as $n \to \infty$.

In addition, the asymptotic normality and $\chi^2$ approximation of the LRT statistic do not hold for admixture models due to non-identifiability. The reason is that traditional asymptotics are based on Taylor expansions of likelihood functions in a small neighborhood of the unique true value. When the identifiability condition is violated, however, there are many true values. Under such circumstances, it is not enough to expand the likelihood function around any specific point. Rather, one needs to approximate the likelihood function in a small neighborhood of the region of true values.

2.3 Asymptotic distribution

The asymptotic behavior of likelihood ratio tests for some special admixture models has been investigated in the literature. For example, Lemdani and Pons (1997) derived the asymptotic distribution of the LRT statistic for the genetic linkage model using reparameterizations. However, their results are not generalizable to admixture models with other kernel distributions. In this section, we investigate the limiting distribution of the LRT statistic for the general admixture model (1), which is applicable to the PK genetic linkage example (4) and the PU example (5). In contrast, Lemdani and Pons (1997) is limited to the genetic linkage example with binomial kernels.

Based on Lemma 1, one needs to approximate the log-likelihood function around the region of true values $\Omega_0$ in large samples. To achieve this, we first choose two small positive numbers $\epsilon_1$ and $\epsilon_2$ and divide the parameter space into four regions: $I = [\epsilon_1, 1] \times [\theta_0 - \epsilon_2, \theta_0 + \epsilon_2]$, $II = [0, \epsilon_1) \times \Theta/[\theta_0 - \epsilon_2, \theta_0 + \epsilon_2]$, $III = [0, \epsilon_1) \times [\theta_0 - \epsilon_2, \theta_0 + \epsilon_2]$ and $IV = [\epsilon_1, 1] \times \Theta/[\theta_0 - \epsilon_2, \theta_0 + \epsilon_2]$.
Figure 1 illustrates such divisions for the genetic linkage example. The asymptotic expansions are easier to obtain in each region, and we can then combine all regions. The asymptotic result is summarized in the following theorem.

**Theorem 1.** For the admixture model (1), assume regularity conditions A1-A7 (in Appendix A). In addition, we assume that either A8 or A9 hold. Under the null $H_0: \delta = 0$ or $\theta = \theta_0$, the LRT statistic converges in distribution to

$$
\sup_{\theta \in \Theta} \{ Z^+(\theta) \}^2, \quad \text{if } \theta_0 \text{ is a boundary point of } \Theta,
$$

and to

$$
\max \left( \sup_{\theta \in \Theta} \{ Z^+(\theta) \}^2, Z^2(\theta_0) \right), \quad \text{if } \theta_0 \text{ is an interior point of } \Theta,
$$

where $Z(\theta) = \lim_{n \to \infty} Z_n(\theta)$, $Z_n(\theta) = \sum_i \{ p(y_i; \theta) / p(y_i; \theta_0) - 1 \} \cdot [\sum_i \{ p(y_i; \theta) / p(y_i; \theta_0) - 1 \}^2]^{-1/2}$ for $\theta \neq \theta_0$ and $Z_n(\theta_0) = \lim_{\theta \to \theta_0} Z_n(\theta)$. The process $Z(\theta)$ is a Gaussian process with mean 0, variance 1, and certain autocorrelation function $\rho(\theta_1, \theta_2) = \text{cor} \{ Z(\theta_1), Z(\theta_2) \}$. Further, if $p'(\cdot ; \theta_0) \neq 0$ with positive probability (condition A8), $Z_n(\theta_0)$ has the functional form

$$
\sum_i \frac{p'(y_i; \theta_0)}{p(y_i; \theta_0)} \cdot \left[ \sum_i \left\{ \frac{p'(y_i; \theta_0)}{p(y_i; \theta_0)} \right\}^2 \right]^{-1/2}, \quad (6)
$$

where $p'(\cdot ; \theta_0) = \partial p(\cdot ; \theta_0) / \partial \theta$. If $p'(\cdot ; \theta_0) = 0$ almost surely and $p''(\cdot ; \theta_0) \neq 0$ with positive probability (condition A9),

$$
Z_n(\theta_0) = \sum_i \frac{p''(y_i; \theta_0)}{p(y_i; \theta_0)} \cdot \left[ \sum_i \left\{ \frac{p''(y_i; \theta_0)}{p(y_i; \theta_0)} \right\}^2 \right]^{-1/2}, \quad (7)
$$

where $p''(\cdot ; \theta_0) = \partial^2 p(\cdot ; \theta_0) / \partial \theta^2$.

The proof of Theorem 1 is provided in the Appendix. The idea is to expand and approximate the likelihood function in different regions, through Lemmas 2-4 in Appendix B. More specifically, the LRT over Region I and IV converges to the supremum of a Gaussian process that happens to be degenerate (Lemma 2). Thus the limiting distribution becomes a simple $\chi^2$. The LRT over Region II and IV converges to the supremum of a non-degenerate Gaussian process that cannot be simplified (Lemma 3). Lemma 4 establishes asymptotics in Region III, while Lemma 5 shows connections of the LRT over different regions. Among these, the most interesting property is that the Gaussian process for Region I and IV is
degenerate so that the limiting distribution of the LRT simplifies to $\chi^2$. A few modifications of the LRT that were proposed in the literature, including the restricted LRT and penalized LRT, are asymptotically equivalent to restricting the parameter space to Region $I$ and $IV$. We will elaborate this point in more details in Section 3.

In the asymptotic argument, besides non-identifiability, one has to deal with two other possible violations of typical regularity conditions: parameter value on the boundary of the parameter space and singularity of Fisher information matrix. The former case happens because $\delta = 0$ is on the boundary of its parameter space $[0, 1]$, and we apply the general statistical theory proposed in Self and Liang (1987). The latter case happens when $p'(\cdot; \theta_0) = 0$ almost surely, which occurs for the PU case of genetic linkage example. Rotnitzky et al. (2000) developed asymptotics with singular Fisher information based on higher order Taylor expansions, and we apply their results under such situations.

Applying this result to the genetic linkage example where the kernel distribution is binomial, one can obtain the following result.

**Corollary 1.** Assume that one observe i.i.d. samples $\{(y_i, K_i) : i = 1, \cdots, n\}$ from admixture models for genetic linkage analysis, (4) for the PK case and (5) for the PU case. Under $H_0 : \delta = 0$ or $\theta = 0.5$, the LRT statistic converges to the following,

$$LRT \xrightarrow{D} \sup_{\theta \in [0, 0.5]} \{Z^+(\theta)\}^2,$$

where $Z(\theta) = \lim_{n \to \infty} Z_n(\theta)$, $Z_n(\theta) = \sum_i g(\theta; y_i, K_i) \cdot \{\sum_i g^2(\theta; y_i, K_i)\}^{-1/2}$ for $\theta \neq 0.5$, $Z_n(0.5) = \lim_{\theta \to 0.5-} Z(\theta)$, $g(\theta; y_i, K_i) = 2^{K_i} \theta y_i (1-\theta)^{K_i-y_i} - 1$ for the PK case and $g(\theta; y_i, K_i) = 2^{K_i-1} \theta y_i (1-\theta)^{K_i-y_i} + 2^{K_i-1} \theta y_i (1-\theta)^{y_i} - 1$ for the PU case. One also has the formula

$$Z_n(0.5) = \sum_i h(y_i, K_i) \cdot \left\{\sum_i h^2(y_i, K_i)\right\}^{-1/2}, \quad (8)$$

where $h(y_i, K_i) = K_i - 2y_i$ in the PK case and $h(y_i, K_i) = K_i^2 - 2y_i K_i + 4y_i^2 - K_i$ in the PU case. The process $Z(\theta)$ is a Gaussian process with mean 0, variance 1, and autocorrelation function $\rho(\theta_1, \theta_2) = \text{cor}\{Z(\theta_1), Z(\theta_2)\} = \lim_{n \to \infty} \text{cor}\{Z_n(\theta_1), Z_n(\theta_2)\}$. 

The proof of Corollary 1 is a straightforward application of Theorem 1 and thus is omitted. This result is consistent with Lemdani and Pons (1997). We included analytic formulas for the autocorrelation function \( \rho(\theta_1, \theta_2) \) in the Appendix.

From the proofs of Theorem 1, we also establish some asymptotic properties of the MLEs under the homogeneity null hypothesis. The MLEs may not even be consistent in the strict sense (see Lemma 1), but certain combination of them converges in probability at the typical \( n^{-1/2} \) rate.

**Corollary 2.** For the admixture model (1), assume regularity conditions A1-A7 (in Appendix A). Let \( \hat{\delta} \) and \( \hat{\theta} \) denote the MLEs under \( H_0 : \delta = 0 \) or \( \theta = 0.5 \). If \( p'(\cdot; \theta_0) \neq 0 \) with positive probability (condition A8), then \( \hat{\delta}(\hat{\theta} - \theta_0) = O_p(n^{-1/2}) \). If \( p'(\cdot; \theta_0) = 0 \) almost surely and \( p''(\cdot; \theta_0) \neq 0 \) with positive probability (conditions A9), then \( \hat{\delta}(\hat{\theta} - \theta_0)^2 = O_p(n^{-1/2}) \).

For the genetic linkage analysis, the PK and PU cases correspond to the former and latter, respectively.

### 2.4 Calculating p values

Theorem 1 states that the limiting distribution of the LRT for admixture models is equivalent to that of the supremum of a squared Gaussian process. However, such limiting distribution is often complicated and does not have an analytic form. In practice, one would need simulation or resampling based methods to calculate p values. For the simulation method, one may calculate the autocorrelation \( \rho(\theta_1, \theta_2) \) of the Gaussian process \( Z(\theta) \), simulate the process \( Z(\theta) \), numerically find the maximum with respect to \( \theta \) and obtain an empirical distribution of the LRT statistic. The p value can be calculated accordingly.

An alternative method to obtain p values is a parametric bootstrap procedure. This procedure is similar in spirit to Beran (1988) and Chen and Chen (2001). The first step is to bootstrap \( N \) samples of size \( n \) from the null model \( p(\cdot; \theta_0) \). Next, one calculate the LRT statistic \( R_i \) from the \( i^{th} \) bootstrap sample for \( i = 1, \cdots, N \). The p values can be obtained using the empirical distribution of \( \{ R_i : i = 1, \cdots, N \} \). The bootstrap procedure is easy
to implement since the null distribution $p(\cdot ; \theta_0)$ does not involve unknown parameters. This procedure is more computationally intensive than Gaussian process based simulations, but it may perform better, especially in small samples.

3. Connection and comparison with alternative approaches

A few alternative approaches have been proposed and studied in the literature for admixture models in genetic linkage studies. These are mostly based on modifications of standard LRT or score tests. We now briefly review these methods. Note that $I_n^\delta := l_n(\delta, 0.5) = l_n(0, \theta) = \sum \log p(y_i; 0.5)$ for any $\delta$ and $\theta$ under $H_0 : \delta = 0$ or $\theta = 0.5$, thus we use these notations interchangeably in the following. For convenience of comparison, we restrict our attentions to the genetic linkage admixture models throughout this section.

3.1 Alternative approaches

The first approach is to restrict the parameter space so that the hypothesis testing problem becomes identifiable in the restricted subspace. For example, one could simply fix $\delta = \delta_1 \neq 0$ (corresponding to the horizontal dashed lines in Figure 1), so that there is only one true null value $(\delta_1, 0.5)$ in the restricted subspace $\delta_1 \times [0, 0.5]$. The problem becomes testing $H_0 : \theta = 0.5$ versus $H_a : 0 \leq \theta < 0.5$, and one could use the test statistic

$$
LRT^{S,\delta}(\delta_1) = LRT(\delta = \delta_1) = 2 \sup_{\theta \in [0,0.5]} \{ l_n(\delta_1, \theta) - l_n(\delta_1, 0.5) \},
$$

which converges to $0.5\chi^2_0 + 0.5\chi^2_1$ under $H_0$. We call this test statistic a “simple LRT” in that LRT has been simplified computationally without having to deal with nonstandard situations. Shoukri and Lathrop (1993) considered a score test while fixing $\delta$, which is equivalent to the simple LRT above to the first order. The simple LRT has a $\chi^2$ type limiting distribution and is convenient to use. However, it requires an arbitrary pre-specification of $\delta_1$, and the power of this test depends on the choice of $\delta_1$. If $\delta_1$ is far from the truth, the simple LRT is likely to have very low power to detect the alternative. Similarly, one could also fix $\theta = \theta_1 \neq 0.5$ (corresponding to the vertical dashed lines in Figure 1), and use test
Likelihood Ratio Testing for Admixture Models with Application to Genetic Linkage Analysis

The second approach is to restrict the parameter space so that the hypothesis testing problem becomes simpler to deal with in the restricted subspace. As suggested by Lemdani and Pons (1995), one could restrict the parameter space of \((\delta, \theta)\) to \([\epsilon_1, 1] \times [0, 0.5]\) (Regions I and IV in Figure 1) for some \(0 < \epsilon_1 < 1\). As a consequence, testing linkage can be specified as \(H_0 : \theta = 0.5\) versus \(H_a : 0 \leq \theta < 0.5\). Using results in Lemdani and Pons (1995), the restricted LRT statistic satisfies

\[
LRT_{R, \delta}(\epsilon_1) = \sup_{0 \leq \theta \leq 0.5 - \epsilon_2} \{ W_2^+(\theta) \}^2 + o_p(1),
\]

where \(W_2(\theta)\) is a centered Gaussian process with unit variance and some autocorrelation.
function. The limiting distribution of $LRT^{R, \theta}(\epsilon_2)$ can not be simplified generally. This is also considered by Lemdani and Pons (1995).

The third approach is the penalized or modified likelihood ratio test considered by Fu et al. (2006). We use the term “penalized LRT” (PLRT) instead of “modified LRT” in this paper. The PLRT is defined as

$$PLRT(C) = \sup_{\delta \in [0,1], \theta \in [0,0.5]} \left\{ 2 \left( l_n(\delta, \theta) + C \log \delta \right) - 2 \left( l_n(1, 0.5) + C \log 1 \right) \right\}$$

where a penalty function $C \log \delta$ (where $C > 0$) was added to the ordinary likelihood function. The penalty is heavy when $\delta$ is close to 0 and less so when $\delta$ approaches 1. Intuitively, as demonstrated in Figure 1, the PLRT is close to the ordinary LRT in region $I$ and $IV$, but imposes heavy penalty around regions $II$ and $III$. One could think of the restricted LRT, $LRT^{R, \delta}(\epsilon_1)$, as a special case of PLRT, in which the penalty is 0 in Regions $I$ and $IV$ and $-\infty$ in Regions $II$ and $III$. Thus, it is not surprising that the PLRT has the limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$ for any choice of $C$, which controls the magnitude of the penalty. Actually, the PLRT and restricted LRT are asymptotically equivalent. However, Fu et al. (2006) reported that the PLRT typically performed better in finite samples. As to the choice of $C$, Fu et al. (2006) suggested taking $C = 1$, but did not investigate the effect of $C$ on type I error and power. Furthermore, Fu et al. (2006) compared it with the other two alternative approaches and concluded that the PLRT generally performed as well as, if not better, than the simple LRT and restricted LRT. Thus, in the following, we focus on comparing the PLRT with the LRT.

Liang and Rathouz (1999) developed a score test procedure that initially fixes the parameter value of $\delta$. Fu et al. (2006) showed that this procedure is asymptotically equivalent to the PLRT. Thus we will not discuss the approach of Liang and Rathouz (1999) in detail.

To briefly summarize, modifications of the LRT generally restrict the total parameter space $[0, 1] \times [0, 0.5]$ to a subspace, and the resulting LRT type test statistic in the subspace generally...
has simpler forms. Actually, the asymptotic distributions of each test can be represented using the Gaussian process $Z(\theta)$ defined in Theorem 1 and Corollary 1. Table 1 listed each test procedure with its specified parameter space and asymptotic null distribution. Modifications of the LRT are generally designed to test against alternatives in certain subspaces, but may lose substantial power against other alternatives that are outside of the specified subspace. They also require specification of a tuning parameter. In contrast, the LRT does not need any tuning parameter and is powerful against general alternatives.

[Table 1 about here.]

3.2 LRT vs. PLRT

We now briefly compare the PLRT with the LRT for admixture models from several perspectives. In terms of simplicity, the PLRT has an advantage since it has a convenient $\chi^2$ type limiting distribution, while the LRT has a complicated limiting distribution. The proof of Theorem 1 in Appendix A sheds lights on how PLRT works for admixture models. More specifically, the LRT over Regions $I \& IV$ and $II \& IV$ converges to $\{W^+_1\}^2 \overset{D}{\rightarrow} 0.5\chi^2_0 + 0.5\chi^2_1$ and $\sup_{\theta \in \theta(\theta_0 - \epsilon_2, \theta_0 - \epsilon_2)} \{W^+_2(\theta)\}^2$, respectively. The PLRT penalizes heavily on Regions $II \& III$ by adding penalty $C \log \delta$ on $\delta$, and asymptotically focuses on Regions $I \& IV$. As a result, unsurprisingly the PLRT statistic converges to $0.5\chi^2_0 + 0.5\chi^2_1$. This is one main reason why PLRT penalized on $\delta$ instead of $\theta$.

Next, we compare the LRT and PLRT in statistical power to detect alternative hypothesis. In the PK case, from Corollary 2, one has $\hat{\delta}(\hat{\theta} - \theta_0) = O_p(n^{-1/2})$. There are three types of local alternatives under such situation, namely,

Type I: $H^n_{a,1} : \delta = \delta_a \in (0, 1], \theta = \theta_0 - \tau/\sqrt{n}$,

Type II: $H^n_{a,2} : \delta = \tau/\sqrt{n}, \theta = \theta_a \in [0, 0.5)$,

Type III: $H^n_{a,3} : \delta = \tau n^{-\alpha_1}, \theta = \theta_0 - \tau n^{-\alpha_2}$,

where $0 < \alpha_1, \alpha_2 < 0.5$, $\alpha_1 + \alpha_2 = 0.5$. These correspond to alternatives that approach the null in Regions $I$, $II$, and $III$, respectively. The LRT is capable of picking up evidence in
all regions, and thus is powerful to detect all possible directions of departure from the null. The PLRT, by design, is powerful to detect Type I alternatives, but penalize heavily and thus is not as powerful against Type II and III alternatives. Thus, the PLRT may lose power to detect linkage when the proportion of linked families is small, while the LRT is more powerful under such situation.

More precisely, Figure 2 displays the power curves of the LRT and PLRT for two types of local alternatives. Type I local alternatives $H_{1a} : \delta = \delta_1, \theta = 0.5 - n^{-1/2} \tau$, approach the null from Region I, in the manner that $\theta$ approaches to 0.5 while $\delta$ is fixed as 1 or 0.5. Type II local alternatives $H_{1a} : \delta = n^{-1/2} \tau, \theta = \theta_1$, approach the null from Region II, in the sense that $\delta$ approaches to 0 while $\theta$ is fixed as 0 or 0.25. Both LRT and PLRT have high power to detect Type I alternatives, while the power of PLRT is slightly higher. For Type II alternatives, the LRT has substantially higher power than the PLRT, especially when $\theta$ is fixed at 0. The reason is that the PLRT imposes very heavy penalty in Region II and thus loses capacity to detect departure from the null in this region.

[Figure 2 about here.]

In addition, the PLRT requires specification of the penalty function, which is somewhat arbitrary. Although the asymptotic argument does not depend on the specific functional form of the penalty function and tuning parameter $C$, the finite sample performance does. More specifically, the PLRT is monotonically decreasing with $C$, which means $C$ affects its type I error rate and power in finite samples. If $C$ is too small, the PLRT often has incorrect type I error rates. Under an extreme situation with $C \to 0$, $PLRT(C)$ is approximately the same as LRT, and using $0.5\chi^2_0 + 0.5\chi^2_1$ as a reference distribution would yield incorrect p values. On the other hand, if $C$ is too big, the PLRT is not powerful. When $C \to \infty$, one can show that $PLRT(C)$ is close to the simple LRT, $LRT^{S,\delta}(\delta = 1)$, which is less powerful against alternatives with $\delta \neq 1$. Thus, contrary to Fu et al. (2006)’s arguments that $PLRT(C)$ is not sensitive to $C$, simulation studies suggest that $C$ controls the balance of type I error and
power in finite samples and should be carefully chosen. An optimal choice of $C$ will maximize statistical power under alternatives while maintaining its nominal values under the null.

4. Simulation Studies

In this section, we evaluate the finite sample performance of LRT and PLRT through simulation studies. We focus on the admixture model for genetic linkage analysis and conduct simulations under both PK and PU cases, with sample size $n = 50, 100, 200$ and family size $K = 2, 4, 8$. In each setting, 10,000 simulations were used to evaluate type I error or power.

First, we consider two methods to calculate $p$ values in Section 2.4, namely simulating Gaussian process and bootstrap procedures. The former uses limiting distribution in Theorem 1 directly and simulates its empirical distribution. We found that this approach usually works well in large samples, or with small family sizes. However, when the sample size is small and the family size is large, the Gaussian approximation may not perform as well. The reason is that the process $Z_n(\theta)$ is often skewed in small samples for a certain range of $\theta$ even though it converges to a Gaussian process asymptotically. In particular, through simulation studies, we find that the process $Z_n(\theta)$ is very skewed when $\theta$ is close to 0 and the Gaussian approximation is poor as a result. Figure 3 displays $Z_n(\theta)$ in finite samples. One can see that its empirical distribution could be quite skewed in small samples with medium to large family size (e.g., $n = 50, K = 8$), especially when $\theta$ is around 0. The performance of the Gaussian process depends on both the sample size $n$ and the family size $K$. Based on our simulation experience, when $n = 50$, the Gaussian process method has decent performance when family sizes vary from $K = 2$ to $K = 5$ for sample size of 50, and when family sizes vary from $K = 2$ to $K = 8$ for sample size of 200. On the other hand, the bootstrap procedure is less sensitive to the family size, and is confirmed to perform well even in small to medium samples. Thus, we recommend the use of bootstrap in studies of small sample size and medium to large family size.

[Figure 3 about here.]
Next, Table 2 shows the Type I error rates and power against a variety of alternatives for the LRT and PLRT in finite samples. In the simulations, we choose the sample size $n = 50$, family size $K = 2$, significance level $\alpha = 0.05$ and a wide range of $C$ from 0.011 to 148. The first row of the table gives Type I error rates for the LRT and PLRT. Under the null, the LRT has rejection rate 4.4%, close to the 5% nominal level, while the rejection rates of the PLRT vary with different value of $C$. The rejection rates seems to be too high for $C = 0.011$, but reasonable close to 5% for other choices of $C$. Thus, for analysis of power, we will not consider the column corresponding to $C = 0.011$.

We first compare the LRT to PLRT($C=1$), which was suggested by Fu et al. (2006). When $\delta$ is small, say $\delta = 0.15, \theta = 0$, the power is 50.1% for the LRT, higher than that of the PLRT, 43.84%. When $\delta$ is large or $\theta$ is close to 0.5, the PLRT generally has higher power. These results agree with the previous analysis on statistical power against local alternatives. However, the power differences between the LRT and PLRT($C=1$) are generally less than 5 − 10% in this setting. The differences become more noticeable in larger samples, which we do not report due to space constrains. Next, we look at the effect of $C$ on statistical power. If we focus on each row in Table 2, the power of the PLRT decreases with $C$ for certain alternative hypotheses. Thus, the optimal choice of $C$ would be the smallest $C$ that still provides the correct Type I error rate. From this perspective, the optimal choice of $C$ among those in Table 2 is 0.135, which has type I error 5.33% under the null and highest statistical power under the alternatives. From this simulation study, one can see that the optimal choice of $C$ depends on the balance between Type I error and power. If $C$ is too small, the PLRT might have incorrect Type I error. If $C$ is too large, the PLRT might not be powerful enough to detect alternatives. Thus, we suggest that $C$ should be chosen with caution, perhaps via a small simulation study.

[Table 2 about here.]
5. Application to a schizophrenia study

In this section, we applied the LRT to a genetic linkage study for schizophrenia conducted at the Johns Hopkins School of Medicine. The details of the study design and data collection can be found in Pulver et al. (1994) and Liang and Rathouz (1999). This study included 486 individuals from 54 families with at least two affected relatives. Here “affected” refers to someone who was diagnosed with either schizophrenia or schizoaffective disorder based on the DSM-III-R criteria.

Based on previous studies, one is particularly interested in Marker D22S941 on Chromosome 22. However, it is well known that schizophrenia is prone to heterogeneity. Thus, we use the admixture models to account for the possibility of genetic heterogeneity. We calculated the likelihood ratio test statistic and the p values by simulation methods. The LRT statistic gives rise to 6.86 and the corresponding p value is 0.007. In this example, Gaussian process based simulation method and the bootstrap method give similar p values. The computational time for the former is a few seconds, while that for the latter is approximately 2 hours on a workstation with Intel Pentium M 725 (1.6 GHz) processor and 512MB memory. The MLEs for $\delta$ and $\theta$ are 0.4 and 0.06, respectively. Thus, it suggests that approximately 40% of the families are linked to the marker at Chromosome 22 and that the recombinational fraction is estimated to be 0.06, suggesting a modest evidence of linkage. We also conducted the PLRT, for different choice of $C$. For $C = 3, 0.5$ and 0.01, the PLRT statistics were 5.36, 5.49, 6.84 and the p values were 0.010, 0.009, 0.004, respectively. Obviously, different choices of $C$ gave rise to different p values, and it is not immediately clear which p value one should use for inference.

To assess whether the asymptotic distribution approximates the empirical distribution of the LRT statistic in such small samples, we conducted simulation studies to mimic the data structure of this schizophrenia study. Figure 4 compared the asymptotic distribution of the LRT (left panel) and the PLRT (right panel) versus their empirical distribution in 10,000 simulations. First of all, it suggested that the asymptotic approximation of the LRT
performed reasonably well for such sample sizes. For the PLRT, the asymptotic distribution agreed well with empirical distribution for \( C = 3 \), slightly worse for \( C = 0.5 \), and not so well for \( C = 0.01 \). This suggested that in our application, \( C = 0.01 \) should not be used at all while \( C = 3 \) and \( C = 0.5 \) provide approximately correct p values. Among these two, \( C = 0.5 \) is uniformly more powerful than \( C = 3 \) to detect linkage.

[Figure 4 about here.]

6. Discussion

Testing for homogeneity in admixture models exhibits non-identifiability features under the null hypothesis and thus has received much attention in the literature. In this paper, we consider statistical issues of the LRT, including both asymptotic properties and practical concerns, and compare the LRT to alternative methods, such as the PLRT. We also illustrate these methods in a genetic linkage study of schizophrenia.

We have considered comparison of the LRT vs alternative choices in the literature, especially the PLRT. In terms of the choice between the LRT and PLRT, both have their own advantages and drawbacks. The PLRT has a convenient \( \chi^2 \) type limiting distribution, but requires specification of a somewhat arbitrary penalty function. The choice of penalty affects type I error rates and power of the PLRT in finite samples. The LRT does not depend on any tuning parameter, but has a relatively complex limiting distribution. As for statistical power, the LRT is powerful to detect all possible directions of departure from the null. The PLRT is powerful to detect the type of alternatives in Region I by design, but not as powerful against other types of alternatives in Regions II and III, which happens when the proportion of linked families is small. In practice, one could consider these issues and decide which method is more appropriate for a particular application.

We have focused on the genetic linkage example, which corresponds to admixture models with Binomial kernel. However, our method and results are general and applicable to admixture models with other kernels. For instance, Davies (1977) considered admixture
models with exponential kernel, which is useful for reliability modelling. Good (1979) used admixture models for clinical trials when some participants are non-responders. Our results are applicable to these applications as well.

There are a few possible extensions of our methods for future research. First, in some applications, there is an additional parameter $\beta$ that is unknown for both components. The probability density function for such models has the form

$$f(y; \delta, \theta, \beta) = (1 - \delta) p(y; \theta_0, \beta) + \delta p(y; \theta, \beta),$$

where $\beta$ is an additional structural parameter. For example, $p(\cdot; \theta, \beta)$ represents a normal distribution with mean $\theta$ and variance $\beta$. It will be interesting to study the LRT and PLRT in such complex admixture settings. Second, it is interesting to extend our results to more general semi-parametric mixture models. Recently, Qin and Liang (2010) considered the family of models

$$f(y; \delta, \alpha, \beta) = (1 - \delta) p(y) + \delta \exp(\alpha + \beta y) p(y),$$

where the density functions for two components are $p(y)$ and $\exp(\alpha + \beta y) p(y)$, respectively, with $p(y)$ unspecified. This family of models have various applications, such as clinical trials with nonresponders, case-control studies with contaminated controls, etc. Admixture models can be viewed as specific examples of this semi-parametric family as well. Qin and Liang (2010) proposed a score test for homogeneity, which is in spirit similar to Liang and Rathouz (1999) for admixture models. It would be natural to consider likelihood ratio tests and extend our results to this semi-parametric family.

**Supplementary Materials**

Web Appendices referenced in Sections 2 and 3 are available under the Paper Information link at the Biometrics website [http://www.tibs.org/biometrics](http://www.tibs.org/biometrics).
Acknowledgements

The effort of the authors on this project was partially supported by Grant R01 HL090577 of the National Institutes of Health. The authors thank the editor, associate editor and the referee for their comments and constructive suggestions.

References


Figure 1. Expected and observed log-likelihood functions for admixture model in genetic linkage analysis. Dark color corresponds to large log-likelihood values. Panel (a) displays the expected log-likelihood function, where solid lines represent the set of true values under $H_0$. Panels (b) and (c) show observed log-likelihood functions for two datasets simulated under $H_0$. Panels (d), (e) and (f) show penalized expected and observed log-likelihood functions, by adding penalty $\log(\delta)$ to those in (a), (b) and (c), respectively. This figure appears in color in the electronic version of this article.
Figure 2. Power curves versus Type I and II local alternatives for the LRT and PLRT.
Figure 3. Boxplots of the process $Z(\theta)$ for different sample size $n$ and family size $K$. The process $Z(\theta)$ converges to a Gaussian process in large samples. But in finite samples, it could be quite skewed.
Figure 4. Asymptotic versus empirical distribution of the LRT statistic and the PLRT statistic. The horizontal axis is the empirical distribution from 10,000 simulations. The vertical axis is the asymptotic distribution simulated from Gaussian processes.
Table 1
Comparison of LRT vs alternative methods. The process $Z(\theta)$ is defined in Theorem 1 for general admixture models, and in Corollary 1 for genetic linkage models. “Turning par.” means dependence on specification of a tuning parameter. $\chi^2$ means a 50:50 mixture of $\chi^2_0$ and $\chi^2_1$. For the PLRT, $\epsilon_C$ is a threshold value determined by $C$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test statistic</th>
<th>Tuning parameter</th>
<th>Combined parameter space</th>
<th>Asymptotic null distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRT</td>
<td>LRT</td>
<td>–</td>
<td>$[0, 1] \times [0, 0.5]$</td>
<td>$\sup_{\theta \in [0, 0.5]} {Z^+(\theta)}^2$</td>
</tr>
<tr>
<td>Simple LRT</td>
<td>$LRT^S,\delta(\delta_1)$</td>
<td>$\delta_1$</td>
<td>$\delta_1 \times [0, 0.5]$</td>
<td>${Z^+(0.5)}^2 \rightarrow \chi^2$</td>
</tr>
<tr>
<td></td>
<td>$LRT^S,\theta(\theta_1)$</td>
<td>$\theta_1$</td>
<td>$[0, 1] \times \theta_1$</td>
<td>${Z^+(\theta_1)}^2 \rightarrow \chi^2$</td>
</tr>
<tr>
<td>Restricted LRT</td>
<td>$LRT^R,\delta(\epsilon_1)$</td>
<td>$\epsilon_1$</td>
<td>$[\epsilon_1, 1] \times [0, 0.5]$</td>
<td>${Z^+(0.5)}^2 \rightarrow \chi^2$</td>
</tr>
<tr>
<td></td>
<td>$LRT^R,\theta(\epsilon_2)$</td>
<td>$\epsilon_2$</td>
<td>$(0, 1] \times [0, 0.5 - \epsilon_2]$</td>
<td>$\sup_{\theta \in [0, 0.5 - \epsilon_2]} {Z^+(\theta)}^2$</td>
</tr>
<tr>
<td>PLRT</td>
<td>PLRT($C$)</td>
<td>$C$</td>
<td>$[\epsilon_C, 1] \times [0, 0.5]$</td>
<td>${Z^+(0.5)}^2 \rightarrow \chi^2$</td>
</tr>
</tbody>
</table>
Table 2
Simulated Type I error rates and power (%) for admixture models for genetic linkage studies, with sample size \( n = 50 \), family size \( K = 2 \) and significance level \( \alpha = 0.05 \). The first row of the table shows Type I error rates, and the remaining give statistical power against different alternatives.

<table>
<thead>
<tr>
<th>( \delta )</th>
<th>( \theta )</th>
<th>LRT</th>
<th>PLRT</th>
<th>PLRT(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.50</td>
<td>4.40</td>
<td>7.00</td>
<td>5.33</td>
</tr>
<tr>
<td>0.05</td>
<td>0.30</td>
<td>6.70</td>
<td>9.84</td>
<td>7.82</td>
</tr>
<tr>
<td>0.05</td>
<td>0.00</td>
<td>13.48</td>
<td>18.90</td>
<td>14.84</td>
</tr>
<tr>
<td>0.10</td>
<td>0.30</td>
<td>9.64</td>
<td>13.74</td>
<td>11.16</td>
</tr>
<tr>
<td>0.10</td>
<td>0.00</td>
<td>30.04</td>
<td>37.96</td>
<td>31.70</td>
</tr>
<tr>
<td>0.15</td>
<td>0.30</td>
<td>13.16</td>
<td>19.30</td>
<td>15.52</td>
</tr>
<tr>
<td>0.15</td>
<td>0.00</td>
<td>50.10</td>
<td>59.06</td>
<td>51.50</td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>18.50</td>
<td>25.36</td>
<td>20.88</td>
</tr>
<tr>
<td>0.20</td>
<td>0.00</td>
<td>70.70</td>
<td>78.12</td>
<td>71.70</td>
</tr>
<tr>
<td>0.25</td>
<td>0.30</td>
<td>24.44</td>
<td>31.66</td>
<td>26.92</td>
</tr>
<tr>
<td>0.25</td>
<td>0.00</td>
<td>85.08</td>
<td>90.12</td>
<td>85.48</td>
</tr>
<tr>
<td>0.30</td>
<td>0.30</td>
<td>31.22</td>
<td>39.06</td>
<td>34.30</td>
</tr>
<tr>
<td>( H_A )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.00</td>
<td>94.50</td>
<td>96.70</td>
<td>94.70</td>
</tr>
<tr>
<td>0.40</td>
<td>0.30</td>
<td>46.38</td>
<td>54.74</td>
<td>50.10</td>
</tr>
<tr>
<td>0.40</td>
<td>0.00</td>
<td>99.48</td>
<td>99.72</td>
<td>99.48</td>
</tr>
<tr>
<td>0.50</td>
<td>0.30</td>
<td>61.30</td>
<td>69.18</td>
<td>65.08</td>
</tr>
<tr>
<td>0.50</td>
<td>0.00</td>
<td>99.98</td>
<td>100.00</td>
<td>99.98</td>
</tr>
<tr>
<td>0.70</td>
<td>0.30</td>
<td>86.52</td>
<td>90.00</td>
<td>88.42</td>
</tr>
<tr>
<td>0.70</td>
<td>0.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>0.90</td>
<td>0.30</td>
<td>97.10</td>
<td>98.12</td>
<td>97.80</td>
</tr>
<tr>
<td>0.90</td>
<td>0.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.30</td>
<td>98.84</td>
<td>99.46</td>
<td>99.32</td>
</tr>
<tr>
<td>1.00</td>
<td>0.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Appendices for “Likelihood Ratio Testing for Admixture Models with Application to Genetic Linkage Analysis”

Chong-Zhi Di*

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center
1100 Fairview Ave N, M2-B500, Seattle, WA 98109
*email: cdi@fhcrc.org

and

Kung-Yee Liang**

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe Street, Baltimore, MD 21205
*email: kyliang@jhsph.edu

SUMMARY: This report contains supplementary materials for the paper “Likelihood Ratio Testing for Admixture Models with Application to Genetic Linkage Analysis.” Section A provides regularity conditions. Section B provides proofs for Theorem 1. Section C contains analytic formulas for autocorrelation functions of the genetic linkage example.
For the general admixture model, its probability density function has the form

\[ f(y; \delta, \theta) = (1 - \delta) p(y; \theta_0) + \delta p(y; \theta), \]  

(1)

where the population comprises two components from the same parametric family \( p(\cdot; \theta) \), with proportions \( 1 - \delta \) and \( \delta \), respectively. The parameter space for \((\delta, \theta)\) is \( \Omega = [0, 1] \times \Theta \), where \( \Theta \subset \mathbb{R} \) and \( \theta_0 \in \Theta \). The null parameter space is \( \Omega_0 = \{(\delta, \theta) : \delta = 0 \text{ or } \theta = \theta_0\} \).

A. Regularity Conditions

We list regularity conditions for the asymptotic results listed in the paper.

A1. \( p(\cdot; \theta) \) has support independent of \( \theta \);

A2. \( p(\cdot; \theta) \neq p(\cdot; \theta_0) \) with positive probability for all \( \theta \neq \theta_0 \);

A3. \( p(\cdot; \theta) \) is fourth continuously differentiable with respect to \( \theta \);

A4. For each \( \theta \) and sufficiently small \( r > 0 \),

\[ \int \log \max\{1, p^*(\cdot; \theta, r)\} \cdot p(\cdot; \theta_0) \, d\mu < \infty, \text{ where } p^*(\cdot; \theta, r) = \sup_{\phi \in \{\theta': ||\theta - \theta'|| \leq r\}} p(\cdot; \phi); \]

A5. For \((\theta, \phi) \in \Theta \times \Theta\), \( \int |\log p(\cdot; \theta)| \cdot p(\cdot; \phi) d\mu < \infty \).

A6. There exists \( \eta > 0 \) such that \( \eta \leq E_{\theta_0}\{p(\cdot; \theta)/p(\cdot; \theta_0) - 1\}^2 < \infty \) for all \( \theta \in \Theta \setminus \theta_0 \);

A7. The process \( W_{2n}(\theta) \) is tight, where

\[ W_{2n}(\theta) = \sum_{i=1}^{n} \left\{ \frac{p(y_i; \theta)}{p(y_i; \theta_0)} - 1 \right\} \cdot \left[ \sum_{i=1}^{n} \left\{ \frac{p(y_i; \theta)}{p(y_i; \theta_0)} - 1 \right\}^2 \right]^{-1/2}; \]

A8. \( p'(\cdot; \theta_0) \neq 0 \) with positive probability, or equivalently var\{\( p'(\cdot; \theta)/p(\cdot; \theta_0) \}\} > 0;

A9. \( p'(\cdot; \theta_0) = 0 \) almost surely and \( p''(\cdot; \theta_0) \neq 0 \) with positive probability.

B. Proof of Theorem 1

We first choose fixed \( \epsilon_1 > 0 \) and \( \epsilon_2 > 0 \), and investigate the behavior of the LRT on the different regions. To prove Theorem 1, we first introduce Lemmas 2-5.
Lemma 2 For the general admixture model (1), assume that typical regularity conditions hold for every fixed $\delta \in [\epsilon_1, 1]$. The LRT statistic over Region I & IV satisfies

$$T_1(\epsilon_1) = \sup_{(\delta, \theta) \in \mathcal{Y}_{1,1} \times \Theta} 2 \left\{ l_n(\delta, \theta) - l_n^0 \right\} = W_{1n}^2 + o_p(1) \xrightarrow{D} \chi_1^2$$

(2)

if $\theta_0$ is an interior point of $\Theta$, and

$$T_1(\epsilon_1) = \{ W_{1n}^+ \}^2 + o_p(1) \xrightarrow{D} 0.5 \chi_0^2 + 0.5 \chi_1^2$$

(3)

if $\theta_0$ is a boundary point of $\Theta$, where $l_n(\delta, \theta) = \sum_i \log f(y_i; \delta, \theta)$, $l_n^0 = \sum_i \log p(y_i; \theta_0)$ and $W_{1n}$ is a random variable with $W_{1n} \xrightarrow{D} N(0,1)$. If $p'(\cdot, \theta_0) = \partial p(\cdot, \theta_0) / \partial \theta \neq 0$ with positive probability,

$$W_{1n} = \sum_i \left( \frac{p'(y_i; \theta_0)}{p(y_i; \theta_0)} \right) \left[ \sum_i \left\{ \frac{p'(y_i; \theta_0)}{p(y_i; \theta_0)} \right\}^2 \right]^{-1/2}. $$

If $p'(\cdot, \theta_0) = 0$ almost surely and $p''(\cdot, \theta_0) = \partial^2 p(\cdot, \theta_0) / \partial \theta^2 \neq 0$ with positive probability, then

$$W_{1n} = \sum_i \left( \frac{p''(y_i; \theta_0)}{p(y_i; \theta_0)} \right) \left[ \sum_i \left\{ \frac{p''(y_i; \theta_0)}{p(y_i; \theta_0)} \right\}^2 \right]^{-1/2}. $$

Proof. In the following, we write $p_i(\theta) := p(y_i; \theta)$, $p'_i(\theta) := p'(y_i; \theta)$ and $p''_i(\theta) := p''(y_i; \theta)$ for notational convenience. We first assume that $\theta_0$ is an interior point of $\Theta$.

For fixed $\delta \in [\epsilon_1, 1]$, the Taylor expansion around $\theta_0$ gives

$$2\{ \log l_n(\hat{\delta}, \hat{\theta}) - l_n^0 \} = 2\{ \log l_n(\hat{\delta}, \hat{\theta}) - l_n(\delta, \theta_0) \}$$

$$= 2\{ \log l_n(\hat{\delta}, \hat{\theta}) - l_n(\delta, \theta_0) \}$$

$$= 2\left( \hat{\delta} - \theta_0 \right) \sum_i \frac{p'_i(\theta_0)}{p_i(\theta_0)} - \delta^2(\hat{\theta} - \theta_0)^2 \sum_i \left\{ \frac{p''_i(\theta_0)}{p_i(\theta_0)} \right\}^2 + o_p \left\{ n\delta^2(\hat{\theta} - \theta_0)^2 \right\}$$

$$= \frac{\left\{ \sum_i \frac{p'_i(\theta_0)}{p_i(\theta_0)} \right\}^2}{\sum_i \left\{ \frac{p''_i(\theta_0)}{p_i(\theta_0)} \right\}^2} + o_p \left\{ n(\hat{\theta} - \theta_0)^2 \right\}$$

where $\hat{\theta}$ is the MLE for $\theta$ with fixed $\delta$.

Case 1. If $p'(\cdot, \theta_0) \neq 0$ with positive probability, then standard asymptotic properties for MLE imply that $\hat{\theta} = \theta_0 + O_p(n^{-1/2})$. Thus, the reminder term in the equation above is $o_p(1)$;
indeed it is \( o_p(1) \) uniformly with respect to \( \delta \in [\epsilon, 1] \). Thus, taking the supremum over \( \delta \) in the equation above, one obtains the expansion in (2).

**Case 2.** If \( p'(\cdot; \theta_0) = 0 \) almost surely and \( p''(\cdot; \theta_0) \neq 0 \) with positive probability, the Fisher information for \( \theta \) evaluated at \( \theta_0 \) is 0. Thus, standard first order results for the MLE do not hold, and one needs to further expand the likelihood ratio into the fourth order, which gives

\[
2 \{ \log l_n(\delta, \hat{\theta}) - l^0_n \} = 2 \{ \log l_n(\delta, \hat{\theta}) - l_n(\delta, \theta_0) \} \\
= \delta(\hat{\theta} - \theta_0)^2 \sum_i \frac{p''(\theta_0)}{p_i(\theta_0)} - \frac{1}{4} \delta^2(\hat{\theta} - \theta_0)^4 \sum_i \left\{ \frac{p''(\theta_0)}{p_i(\theta_0)} \right\}^2 + o_p\{ n\delta^2(\hat{\theta} - \theta_0)^4 \}
\]

Using results from Rotnitzky et al. (2000), \( \hat{\theta} = \theta_0 + O_p(n^{-1/4}) \) under this situation. Similar to the argument in **Case 1**, one can obtain the expansion in (2), except that \( W_{1n} \) involves \( p''(\cdot; \theta_0) \) instead of \( p'(\cdot; \theta_0) \).

Under both cases, the numerator of \( W_{1n} \) has mean zero. One can obtain \( W_{1n} \xrightarrow{D} N(0, 1) \) by the central limit theorem and thus \( T_1(\epsilon_1) \xrightarrow{D} \chi^2_1 \) for any \( 0 < \epsilon_1 \leq 1 \).

Finally, when \( \theta_0 \) is a boundary point, we use results of Self and Liang (1987) and replace \( W_{1n} \) by \( W_{1n}^+ \) in the arguments above.

**Lemma 3** For the general admixture model (1), assume that typical regularity conditions hold for every fixed \( \theta \in \theta \setminus (\theta_0 - \epsilon_2, \theta_0 + \epsilon_2) \). Then, the LRT statistic over Region II & IV satisfies

\[
T_2(\epsilon_2) = \sup_{(\delta, \theta) \in [0, 1] \times \theta \setminus (\theta_0 - \epsilon_2, \theta_0 + \epsilon_2)} 2 \left\{ l_n(\delta, \theta) - l^0_n \right\} \\
\xrightarrow{D} \sup_{\theta \in \Theta \setminus (\theta_0 - \epsilon_2, \theta_0 + \epsilon_2)} \left\{ W_{2n}^+ (\theta) \right\}^2, \tag{4}
\]

where \( W_2(\theta) = \lim_{n \to \infty} W_{2n}(\theta) \) is a Gaussian process with mean 0, variance 1 and certain autocorrelation function \( \rho(\theta_1, \theta_2) \).
Proof. For fixed $\theta \in \theta \backslash (\theta_0 - \epsilon_2, \theta_0 - \epsilon_2)$, the Taylor expansion around $\delta = 0$ gives

$$2 \{ \log l_n(\hat{\delta}, \theta) - l^0_n \} = 2 \{ \log l_n(\hat{\delta}, \theta) - l_n(0, \theta) \}$$

$$= 2\hat{\delta} \sum_i \left\{ \frac{p(y_i; \theta)}{p(y_i; \theta_0)} - 1 \right\} - \hat{\delta}^2 \sum_i \left\{ \frac{p(y_i; \theta)}{p(y_i; \theta_0)} - 1 \right\}^2 + o_p(n\hat{\delta}^2)$$

$$= \left[ \sum_i \left( \frac{p(y_i; \theta)}{p(y_i; \theta_0)} - 1 \right) \right]^2 + o_p(n\hat{\delta}^2)$$

$$= \{ W_{2n}^+(\theta) \}^2 + o_p(n\hat{\delta}^2),$$

where $\hat{\delta}$ is the MLE for $\delta$ for fixed $\theta$. Under conditions specified in Lemma 3, one has $\hat{\delta} = O_p(n^{-1/2})$ and that the remainder term above converges to $o_p(1)$ uniformly for $\theta \in \Theta \backslash (\theta_0 - \epsilon_2, \theta_0 - \epsilon_2)$. Taking supremum over $\theta$, one can obtain (4). Note that equation (4) involves $W_{2n}^+(\theta)$ instead of $W_2(\theta)$ because $\delta = 0$ is on the boundary of its parameter space $[0,1]$, see arguments in Self and Liang (1987).

**Lemma 4** For the general admixture model (1), the LRT statistic over Region III is defined as

$$T_3(\epsilon_1, \epsilon_2) = \sup_{(\delta, \theta) \in [0, \epsilon_1] \times [\theta_0 - \epsilon_2, \theta_0 - \epsilon_2]} 2 \{ l_n(\delta, \theta) - l^0_n \}.$$

If $\theta_0$ is an interior point of $\theta$,

$$W_{1n}^2 + o_p(1) \leq T_3(\epsilon_1, \epsilon_2) \leq W_{1n}^2 + (\epsilon_2 + \epsilon^2_2) O_p(1),$$

(5)

and if $\theta_0$ is a boundary point,

$$\{ W_{1n}^+ \}^2 + o_p(1) \leq T_3(\epsilon_1, \epsilon_2) \leq \{ W_{1n}^+ \}^2 + (\epsilon_2 + \epsilon^2_2) O_p(1),$$

(6)

where $W_{1n}$ is defined as in Lemma 2.

Proof. We provide the proof when $\theta_0$ is an interior point and when $p'(:; \theta_0) \neq 0$ with positive probabilities. Extensions to other cases involve either Self and Liang (1987) or higher order
Taylor expansions as in Rotnitzky et al. (2000). These extensions are similar in spirit to those in proofs of Lemma 2 and 3, and thus are omitted.

First,

\[
T_3(\epsilon_1, \epsilon_2) = \sup_{(\delta, \theta) \in [0, \epsilon_1] \times [\theta_0 - \epsilon_2, \theta_0 - \epsilon_2]} 2 \left\{ l_n(\delta, \theta) - l_n^0 \right\} \\
\geq \sup_{(\delta, \theta) \in \{\epsilon_1\} \times [\theta_0 - \epsilon_2, \theta_0 - \epsilon_2]} 2 \left\{ l_n(\delta, \theta) - l_n^0 \right\} \\
= W_{1n}^2 + o_p(1),
\]

where the last equation is obtained from Lemma 2. Next, we expand \( T_3(\epsilon_1, \epsilon_2) \) around \((0, \theta_0)\),

\[
T_3(\epsilon_1, \epsilon_2) = 2 \{ \log l_n(\hat{\delta}, \hat{\theta}) - l_n(0, \theta_0) \} \\
= 2 \hat{\delta}(\hat{\theta} - \theta_0) \sum_i \frac{p_i'(\theta_0)}{p_i(\theta_0)} + \hat{\delta}^2(\hat{\theta} - \theta_0)^2 \sum_i \left\{ \frac{p_i'(\theta_0)}{p_i(\theta_0)} \right\}^2 \\
+ \frac{1}{3} \hat{\delta}^3(\hat{\theta} - \theta_0)^3 \sum_i \frac{p_i''(\theta_0)}{p_i(\theta_0)} - \hat{\delta}^2(\hat{\theta} - \theta_0)^2 \sum_i \left\{ \frac{p_i'(\theta_0)}{p_i(\theta_0)} \right\}^2 \\
+ o_p\{ n\hat{\delta}^2(\hat{\theta} - \theta_0)^2 \}.
\]

Based on this fourth order approximation, one can show that \( \hat{\delta}(\hat{\theta} - \theta_0) = O_p(n^{-1/2}) \) and \( \hat{\theta} - \theta_0 = O_p(1) \). Thus,

\[
T_3(\epsilon_1, \epsilon_2) = 2 \{ \log l_n(\hat{\delta}, \hat{\theta}) - l_n(0, \theta_0) \} \\
= 2 \hat{\delta}(\hat{\theta} - \theta_0) \sum_i \frac{p_i'(\theta_0)}{p_i(\theta_0)} - \hat{\delta}^2(\hat{\theta} - \theta_0)^2 \sum_i \left\{ \frac{p_i'(\theta_0)}{p_i(\theta_0)} \right\}^2 \\
+ \hat{\delta}^2(\hat{\theta} - \theta_0)^2 \sum_i \frac{p_i'(\theta_0)}{p_i(\theta_0)} + \frac{1}{3} \hat{\delta}^3(\hat{\theta} - \theta_0)^3 \sum_i \frac{p_i''(\theta_0)}{p_i(\theta_0)} \\
+ o_p\{ n\hat{\delta}^2(\hat{\theta} - \theta_0)^2 \}
\leq W_{1n}^2 + (\epsilon_2 + \epsilon_2^2) O_p(1),
\]

where (7) is equivalent to \( W_{1n}^2 \) according to Lemma 2 and (8) is bounded because \( |\theta - \theta_0| \leq \epsilon_2 \) always hold in Region III. Thus, (5) follows.
Lemma 5 The processes $W_{1n}$, $W_{2n}(\theta)$, $W_1$ and $W_2(\theta)$ satisfy $W_{1n} = \lim_{\theta \to \theta_0} W_{2n}(\theta)$ and $W_1 = \lim_{\theta \to \theta_0} W_2(\theta)$.

Proof. Based on definitions of these processes, Lemma 5 can be obtained by straightforward limit calculations.

Proof of Theorem 1. For any fixed $\epsilon_1 > 0$ and $\epsilon_2 > 0$, the LRT statistic can be obtained by combining Regions $I$, $II$, $III$ and $IV$, i.e.,

$$LRT = \max\{ T_1(\epsilon_1), T_2(\epsilon_2), T_3(\epsilon_1, \epsilon_2) \}. \quad (9)$$

Define $Z_n(\theta) = W_{2n}(\theta)$, $Z_n(\theta_0) = W_{1n}$ and $Z(\theta) = \lim_{n \to \infty} Z_n(\theta)$, then $Z_n(\theta)$ and $Z(\theta)$ are continuous time processes based on Lemma 5. Based on Lemma 2-5, first let $n \to \infty$ and then let $\epsilon_2 \to 0$, one can obtain Theorem 1.

Remark 1. The asymptotic properties of several modifications of the LRT, e.g., restricted LRT and penalized LRT mentioned in Section 3, can be obtained directly from Lemma 2 and Lemma 3.

Remark 2. The results and proof for Theorem 1 hold for general admixture model. In the genetic linkage context, some results analogous to Lemma 1-5 and Theorem 1 have been obtained before. For example, Lemdani and Pons (1995) considered the restricted LRT and obtained results that are special cases of Lemma 2 and Lemma 3. Lemdani and Pons (1997) considered the LRT and obtained results that are special cases of Theorem 1. We would like to point out that Lemdani and Pons (1997)’s proof utilized re-parameterization specific to binomial kernels and is not easy to generalize. On the other hand, our proof is more general and can potentially generalize beyond admixture models.
C. Formulas for $\rho(\theta_1, \theta_2)$ in genetic linkage admixture models

**PK case.**

$$\rho(\theta_1, \theta_2) = \begin{cases} \frac{\sum_i \phi(\theta_1, \theta_2; K_i)}{\sqrt{\sum_i \phi(\theta_1, \theta_2; K_i)}} \cdot \sqrt{\sum_i \phi(\theta_2, \theta_2; K_i)}, & \text{if } \theta_1 \in [0, 0.5), \theta_2 \in [0, 0.5), \\ (1 - 2\theta_2) \cdot \sqrt{\sum_i K_i} & (1 - 2\theta_2) \cdot \sqrt{\sum_i \phi(\theta_2, \theta_2; K_i)}, & \text{if } \theta_1 = 0.5, \theta_2 \in [0, 0.5), \end{cases}$$

where $\phi(\theta_1, \theta_2; K) = \{1 + (1 - 2\theta_1)(1 - 2\theta_2)\}^K - 1$ and $K_i$ is the size for family $i$, for $i = 1, 2, \cdots, n$.

**PU case.**

$$\rho(\theta_1, \theta_2) = \begin{cases} \frac{\sum_i \phi(\theta_1, \theta_2; K_i)}{\sqrt{\sum_i \phi(\theta_1, \theta_2; K_i)}} \cdot \sqrt{\sum_i \phi(\theta_2, \theta_2; K_i)}, & \text{if } \theta_1 \in [0, 0.5), \theta_2 \in [0, 0.5), \\ (1 - 2\theta_2)^2 \cdot \sqrt{\sum_i K_i(K_i - 1)/2} & (1 - 2\theta_2)^2 \cdot \sqrt{\sum_i \phi(\theta_2, \theta_2; K_i)}, & \text{if } \theta_1 = 0.5, \theta_2 \in [0, 0.5), \end{cases}$$

where $\phi(\theta_1, \theta_2; K) = \left[\{1 + (1 - 2\theta_1)(1 - 2\theta_2)\}^K + \{1 - (1 - 2\theta_1)(1 - 2\theta_2)\}^K\right] / 2$ and $K_i$ is the size of family $i$, for $i = 1, 2, \cdots, n$.

**References**

