3-15-2010

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

Chong-Zhi Di
Fred Hutchinson Cancer Research Center, cdi@fredhutch.org

Kung-Yee Liang
Johns Hopkins University, kyliang@jhsph.edu

Biometrics 000, 1–24 DOI: 000
000 0000

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 special case of two component mixture model, where one component is indexed by an unknown parameter while the parameter value for the other component is known. It has been 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 χ^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 provide insights on the connection and comparison between LRT and alternative approaches in the literature, mostly modifications of LRT and score tests, including the modified or penalized LRT (Fu et al., 2006). The LRT is an omnibus test that is powerful against general alternative hypothesis. In contrast, alternative approaches may be slightly more powerful against certain type of alternatives, but much less powerful for other types. 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 the focus on genetic linkage analysis. With a kernel distribution $p(y;\gamma)$ indexed by γ, an admixture model has the probability density function of the form

$$f(y\,;\delta,\gamma) = (1-\delta)\,p(y;\gamma_0) + \delta\,p(y;\gamma), \tag{1}$$

where the population is composed of two components from the same parametric family $p(\cdot\,;\gamma)$, with proportions $1-\delta$ and δ, respectively. The first component is indexed by a known parameter value γ_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 $p(\cdot\,;\gamma)$ could be any parametric distribution, such as Gaussian, binomial, exponential or Poisson distributions. We assume that the parameter space for (δ,γ) is $\Omega = [0,1]\times\Gamma$, where $\Gamma \subset R$ is compact and $\gamma_0 \in \Gamma$. 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 $p(\cdot\,;\gamma_0)$ is a particular submodel of the the second component $p(\cdot\,;\gamma)$, and usually has special scientific meanings. One might be interested in testing a simple homogeneous model $p(\cdot\,;\gamma_0)$ versus admixture alternative. With (1), the null hypothesis can be specified as either $\gamma = \gamma_0$, or equivalently $\delta = 0$. Under the specification of $\gamma = \gamma_0$, the parameter δ disappears and any value of δ gives the same null distribution. Similarly, γ disappears under the specification of $\delta = 0$. In other words, each value in the set $\Omega_0 := \{(\delta,\gamma) : \delta = \delta_0 \text{ or } \gamma = \gamma_0\}$ represent exactly the same distribution. Thus, testing $H_0 : \delta = 0$ or $\gamma = \gamma_0$ 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 to Ott (1999), Thomas (2004) and Fu et al. (2006) for detailed descriptions. 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 γ) is the percentage of offsprings that are recombinants, and can take values from 0 to 0.5. The two loci are strongly linked if γ is close to 0, and not linked if $\gamma = 0.5$. Offsprings from a 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 offsprings. If there is possible linkage among all families, Y has a simple binomial distribution,

$$p(y;\gamma) = \Pr(Y = y) = \binom{K}{y}\gamma^y(1-\gamma)^{K-y}, \tag{2}$$

where $\gamma \in [0, 0.5]$ describes the magnitude of linkage. In the PU case, one could only observe that there are two groups of offsprings, 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 50% chance that the first group (with Y offsprings) is recombinant and 50% chance that it is not. Thus, Y follows a mixture of two binomial distributions, i.e.,

$$p(y;\gamma) = \Pr(Y = y) = 0.5\binom{K}{y}\gamma^y(1-\gamma)^{K-y} + 0.5\binom{K}{y}\gamma^{K-y}(1-\gamma)^y. \tag{3}$$

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 $H_0 : \gamma = 0.5$, 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 $\gamma = 0.5$ 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 χ^2 distributions, $0.5\chi_0^2 + 0.5\chi_1^2$ under H_0, where χ_0^2 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 to use an admixture model to account for such hetegeneity. More precisely, the admixture model in the genetic linkage context has the form

$$f(y\,;\delta,\gamma) = (1-\delta)\binom{K}{y}0.5^K + \delta\binom{K}{y}\gamma^y(1-\gamma)^{K-y} \tag{4}$$

for the PK case, and

$$f(y\,;\delta,\gamma) = (1-\delta)\binom{K}{y}0.5^K + \delta\binom{K}{y}\left\{\,0.5\gamma^y(1-\gamma)^{K-y} + 0.5\gamma^{K-y}(1-\gamma)^y\,\right\} \tag{5}$$

for the PU case, where δ 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 $\gamma = 0.5$, and the alternative is $H_1 : 0 < \delta \leqslant 1, 0 \leqslant \gamma < 0.5$. As illustrated in Figure 1 (a), the parameter space is the rectangle $(\delta, \gamma) \in [0, 1] \times [0, 0.5]$ 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 $\gamma = 0.5$. This hypothesis testing problem is nonstandard due to nonidentifiability under the null, in the sense that γ is not identifiable when $\delta = 0$ and δ is not identifiable when $\gamma = 0.5$. In addition, two types of nonstandard situations might occur. First, the null parameter values ($\delta = 0$ and $\gamma = 0.5$) are on the boundary of the parameter space. Second, the Fisher information for γ evaluated at $\gamma = 0.5$ and any δ is always 0 in the PU case.

[Figure 1 about here.]

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 on statistical inference for admixture models, and then present asymptotic results for the LRT.

2.1 *Challenges on inference*

The nonstandard properties of the admixture model under H_0 brings challenges to statistical inference, including both parameter estimation and hypothesis tests. 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, two solid lines ($\delta = 0$ and $\gamma = 0.5$). 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 could 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.

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.

Lemma 1 For the admixture model (1), under the homogeneity null, i.e., $(\delta, \gamma) \in \Omega_0 = \{(\delta, \gamma) : \delta = \delta_0 \text{ or } \gamma = \gamma_0\}$ or equivalently $Y \sim p(\cdot\,; \gamma_0)$, the MLEs $\hat{\delta}$ and $\hat{\gamma}$ satisfy the following.

(1) $(\hat{\delta}, \hat{\gamma})$ does not converge in probability, and may not even be uniquely defined;
(2) $d_{\Omega_0}(\hat{\delta}, \hat{\gamma}) := \inf_{(\delta', \gamma') \in \Omega_0} \|(\hat{\delta}, \hat{\gamma}) - (\delta', \gamma')\|$ converges in probability to 0 as $n \to \infty$;
(3) The estimated density $f(\cdot\,; \hat{\delta}, \hat{\gamma})$ converges to $p(\cdot\,; \gamma_0)$ as $n \to \infty$.

In addition, the asymptotic normality and χ^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 the 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 situation, 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.

2.2 *Two classes of non-identifiability*

There has been much work in the literature on likelihood ratio testing under loss of identifiability. In this paper, we distinguish two classes of non-identifiability problems when the model under consideration involves parameters $(\gamma, \delta) \in \Omega$. In *Class 1*, all parameter values in $\Omega_{01} = \{(\delta, \gamma) : \delta = \delta_0\}$ correspond to the same distribution that depends on a fixed δ_0, and one wish to test this distribution versus others. The null hypothesis is specified via the parameter of interest ($\delta = \delta_0$) while the nuisance parameter γ is not identifiable under the null. Under the null hypothesis, the set of true values is a one-dimensional space Ω_{01}. This class of problems has been studied extensively (e.g., Davies 1977, 1987). In *Class 2*, all parameter values in $\Omega_0 = \{(\delta, \gamma) : \delta = \delta_0 \text{ or } \gamma = \gamma_0\}$ represent the same distribution, and one is interested to distinguish it from other distributions. The null hypothesis can be specified equivalently via each one of the two parameters ($\delta = \delta_0$ or $\gamma = \gamma_0$), and under either specification, the other parameter (γ or δ) is not identifiable. Under the null hypothesis, the set of true values Ω_0 contains the union of two one-dimensional subspaces. These non-identifiability problems are to be contrasted with identifiable *regular class* where the true value (δ_0, γ_0) is unique.

According to our definition, testing homogeneity for admixture models naturally belongs to *Class 2*, because the null hypothesis can be specified via either δ or γ and under either specification the other parameter becomes nonidentifiable. 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 $\gamma = 0.5$ 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 problem could reduce to *Class 1* or *regular class*. For example, if one restricts the parameter space of (δ, γ) 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 $\gamma = 0.5$ and thus the problem reduces to *Class 1*. 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) and will be discussed in details in Section 3.

The asymptotic properties of LRT for *Class 1* problems were well studied in the literature, while those for *Class 2* received less attention. To investigate the latter, we will divide the parameter spaces into a few regions so that the LRT in each region becomes *Class 1*, and then combine all regions. A primary reason for such division is that known results from Class 1 could be utilized conveniently.

2.3 *Asymptotic distribution*

The asymptotic behavior of likelihood ratio tests for some special admixture models have 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 genetic linkage example (4) and (5) as special cases. 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 Ω_0 in large samples. To achieve this, we first choose two small positive numbers ϵ_1 and ϵ_2 and divide the parameter space into four regions: $I - [\epsilon_1, 1] \times [\gamma_0 - \epsilon_2, \gamma_0 + \epsilon_2]$, $II - [0, \epsilon_1) \times \Gamma/[\gamma_0 - \epsilon_2, \gamma_0 + \epsilon_2]$, $III - [0, \epsilon_1) \times [\gamma_0 - \epsilon_2, \gamma_0 + \epsilon_2]$ and $IV - [\epsilon_1, 1] \times \Gamma/[\gamma_0 - \epsilon_2, \gamma_0 + \epsilon_2]$. Figure 1 illustrates such division 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), under the null $H_0 : \delta = 0$ or $\gamma = \gamma_0$, the LRT

statistic converges in distribution to

$$\begin{aligned} \sup_{\gamma\in\Gamma}\{Z^+(\gamma)\}^2\,, & \qquad \text{if } \gamma_0 \text{ is a boundary point of } \Gamma, \text{ and to} \\ \max\left[\sup_{\gamma\in\Gamma}\{Z^+(\gamma)\}^2, Z^2(\gamma_0)\right], & \quad \text{if } \gamma_0 \text{ is an interior point of } \Gamma, \end{aligned}$$

where $Z(\gamma) = \lim_{n\to\infty} Z_n(\gamma)$, $Z_n(\gamma) = \sum_i\{\, p(y_i;\gamma)/p(y_i;\gamma_0)-1\,\}\cdot[\,\sum_i\{\, p(y_i;\gamma)/p(y_i;\gamma_0)-1\,\}^2\,]^{-1/2}$ for $\gamma \neq \gamma_0$ and $Z_n(\gamma_0) = \lim_{\gamma\to\gamma_0} Z_n(\gamma)$. The process $Z(\gamma)$ is a Gaussian process with mean 0, variance 1, and certain autocorrelation function $\rho(\gamma_1,\gamma_2) = \mathrm{cor}\{\, Z(\gamma_1), Z(\gamma_2)\,\}$. Further, if $p'(\cdot\,;\gamma_0) \neq 0$ with positive probability, $Z_n(\gamma_0)$ has the functional form

$$\sum_i \frac{p'(y_i;\gamma_0)}{p(y_i;\gamma_0)} \cdot \left[\sum_i \left\{\frac{p'(y_i;\gamma_0)}{p(y_i;\gamma_0)}\right\}^2\right]^{-1/2}, \tag{6}$$

where $p'(\cdot\,;\gamma_0) = \partial p(\cdot\,;\gamma_0)/\partial\gamma$. If $p'(\cdot\,;\gamma_0) = 0$ almost surely and $p''(\cdot\,;\gamma_0) \neq 0$ with positive probability,

$$Z_n(\gamma_0) = \sum_i \frac{p''(y_i;\gamma_0)}{p(y_i;\gamma_0)} \cdot \left[\sum_i \left\{\frac{p''(y_i;\gamma_0)}{p(y_i;\gamma_0)}\right\}^2\right]^{-1/2}, \tag{7}$$

where $p''(\cdot\,;\gamma_0) = \partial^2 p(\cdot\,;\gamma_0)/\partial\gamma^2$.

The proof of Theorem 1 is provided in the Appendix. 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 $p'(\cdot\,;\gamma_0) = 0$ almost surely. 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 Suppose 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 $\gamma = 0.5$, the LRT statistic converges to the following ,

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

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

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

where $h(y_i, K_i) = K_i - 2y_i$ in the PK case and $h(y_i, K_i) = K_i^2 - 2y_iK_i + 4y_i^2 - K_i$ in the PU case. The process $Z(\gamma)$ is a Gaussian process with mean 0, variance 1, and autocorrelation function $\rho(\gamma_1, \gamma_2) = \text{cor}\{ Z(\gamma_1), Z(\gamma_2) \} = \lim_{n\to\infty} \text{cor}\{ Z_n(\gamma_1), Z_n(\gamma_2) \}$.

The proof of Corollary 1 is 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 $\rho(\gamma_1, \gamma_2)$ in Appendix.

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 $\rho(\gamma_1, \gamma_2)$ of the Gaussian process $Z(\gamma)$, simulate the process $Z(\gamma)$, numerically find the maximum with respect to γ 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 $p(\cdot\,; \gamma_0)$. 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 $\{R_i : i = 1, \cdots, N\}$. This procedure is more computational intensive than Gaussian process based simulations, but might performs 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 $l_n^0 := l_n(\delta, 0.5) = l_n(0, \gamma) = \sum_i \log p(y_i; 0.5)$ for any δ and γ under $H_0 : \delta = 0$ or $\gamma = 0.5$, thus we used these notations exchangeably 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 horizonal 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 $H_0 : \gamma = 0.5$ versus $H_a : 0 \leqslant \gamma < 0.5$, and one could use the test statistic

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

which converges to $0.5\chi_0^2 + 0.5\chi_1^2$ 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 δ, which is equivalent to the simple LRT above to the first order. The simple LRT has a χ^2 type limiting distribution and is convenient to use. However, it requires an arbitrary pre-specification of δ_1, and the power of this test depends on the choice of δ_1. If δ_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 $\gamma = \gamma_1 \neq 0.5$ (corresponding to the vertical dashed lines in Figure 1), and use test statistic

$$LRT^{S,\gamma}(\gamma_1) = LRT(\gamma = \gamma_1) = 2 \sup_{\delta \in [0,1]} \{ l_n(\delta, \gamma_1) - l_n(0, \gamma_1) \},$$

which also converges to $0.5\chi_0^2 + 0.5\chi_1^2$ under the H_0.

The second approach is to restrict the parameter space so that the hypothesis testing problem reduces to *Class 1* in the restricted subspace. As suggested by Lemdani and Pons (1995), one could restrict the parameter space of (δ, γ) 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 $H_0 : \gamma = 0.5$ versus $H_a : 0 \leqslant \gamma < 0.5$, which has now become a *Class 1* problem. Using results in Lemdani and Pons (1995), the restricted LRT statistic satisfies

$$\begin{aligned} LRT^{R,\delta}(\epsilon_1) &= \sup_{\epsilon_1 \leqslant \delta \leqslant 1, \gamma \in [0,0.5]} 2\{ l_n(\delta,\gamma) - l_n(\delta, 0.5) \} \\ &= \sup_{\epsilon_1 \leqslant \delta \leqslant 1} \{ W_1^+(\delta) \}^2 + o_p(1), \end{aligned}$$

where $W_1(\delta)$ is a centered Gaussian process with unit variance. In addition, one can show that $W_1(\delta)$ does not depend on δ, and the process $W_1(\delta)$ reduces to a standard Gaussian random variable. Thus, the test statistic $LRT^{R,\delta}(\epsilon_1)$ converges in distribution to $0.5\chi_0^2 + 0.5\chi_1^2$ for any $0 < \epsilon_1 < 1$, making it convenient to obtain p values. This restricted LRT is designed to detect departures from the null in specific regions, making it more attractive than the simple LRT. On the other hand, it requires an arbitrary choice of ϵ_1. The power of the restricted LRT generally depends on ϵ_1 and the empirical type I error rate also depends on ϵ_1 in finite samples. Further, this test statistic has a peculiar feature that $LRT^{R,\delta}(\epsilon_1)$ is a decreasing function of ϵ_1, yet has the same asymptotic distribution under H_0.

Similarly, one could also restrict the parameter space of (δ, γ) to $[0,1] \times [0, 0.5-\epsilon_2]$ (Regions *II* and *IV* in Figure 1) for some $0 < \epsilon_2 < 0.5$. The restricted LRT statistic in this case satisfies

$$\begin{aligned} LRT^{R,\gamma}(\epsilon_2) &= \sup_{0 \leqslant \delta \leqslant 1, \gamma \in [0,0.5-\epsilon_2]} 2\{ l_n(\delta,\gamma) - l_n(\delta, 0.5) \} \\ &= \sup_{0 \leqslant \gamma \leqslant 0.5-\epsilon_2} \{ W_2^+(\gamma) \}^2 + o_p(1), \end{aligned}$$

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

[Figure 2 about here.]

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

$$\begin{aligned} PLRT(C) &= \sup_{\delta \in [0,1], \gamma \in [0,0.5]} 2\{ l_n(\delta,\gamma) + C \log \delta \} - 2\{ l_n(1, 0.5) + C \log 1 \} \\ &= \sup_{\delta \in [0,1], \gamma \in [0,0.5]} 2\{ l_n(\delta,\gamma) + C \log \delta - l_n(1, 0.5) \} \end{aligned}$$

where a penalty function $C \log \delta$ (where $C > 0$) was added to the ordinary likelihood function. The penalty is heavy when δ is close to 0 and less so when δ approaches 1. Intuitively, as demonstrated by Figure 2 the PLRT is close to the ordinary LRT in region I and IV, but imposes heavy penalty around region 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 region I and IV and $-\infty$ in region 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 to take $C = 1$, 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 focused on comparing the PLRT with the LRT.

Liang and Rathouz (1999) developed a score test procedure which initially fixes the parameter value of δ. 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 $Z(\gamma)$ 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 χ^2

type limiting distribution, while the LRT has a complicated limiting distribution. Actually the proof of Theorem 1 in Appendix sheds lights on how PLRT works for admixture models. More specifically, the LRT over Region *I*&*IV* and *II*&*IV* converges to $\{ W_1^+ \}^2 \xrightarrow{D} 0.5\chi_0^2 + 0.5\chi_1^2$ and $\sup_{\gamma \in \Gamma \setminus (\gamma_0-\epsilon_2, \gamma_0-\epsilon_2)} \{ W_2^+(\gamma) \}^2$, respectively. The PLRT penalizes heavily on Region *II*&*III* by adding penalty $C \log \delta$ on δ, and asymptotically focuses on Region *I*&*IV*. As a result, unsurprisingly the PLRT statistic converges to $0.5\chi_0^2 + 0.5\chi_1^2$. This is one main reason why PLRT penalized on δ instead of γ.

Next, we compare the LRT and PLRT in statistical power to detect alternative hypothesis. In the PK case, as shown in Appendix (Proof of Lemma 4), one has $\hat{\delta}(\hat{\gamma} - \gamma_0) = O_p(n^{-1/2})$. There are three types of local alternatives under such situation, namely,

Type I – $H_{a,1}^n : \ \delta = \delta_a \in (0, 1], \quad \gamma = \gamma_0 - \tau/\sqrt{n},$

Type II – $H_{a,2}^n : \ \delta = \tau/\sqrt{n}, \qquad \gamma = \gamma_a \in [0, 0.5),$

Type III – $H_{a,3}^n : \ \delta = \tau\, n^{-\alpha_1}, \qquad \gamma = \gamma_0 - \tau\, n^{-\alpha_2}$, 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 the type of alternatives in *Region I*, but penalize heavily and thus is not as powerful against Type II and III alternatives. Thus, the PLRT may not be desirable when the proportion of linked families is small, while the LRT is generally powerful. These will be verified in simulation studies (Section 4).

In addition, the PLRT requires specification of the penalty function, which is somewhat arbitrary. Although the asymptotic arguments 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 $0.5\chi_0^2 + 0.5\chi_1^2$ 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, in 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 focused on the admixture model for genetic linkage analysis. We conducted simulations under both PK and PU cases, with sample size $n = 50, 100, 200$ and family size $K = 2, 4, 8$. In each setting, 1000 simulations were used to evaluate type I error or power. The results are reported as follows.

First, we considered two methods to calculate p values in Section 2.4, namely simulating Gaussian process and bootstrap procedures. The former use 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 $Z_n(\gamma)$ is often skewed in small samples for a certain range of γ even though it converges to a Gaussian process asymptotically. In particular, through simulation studies, we find that the process $Z_n(\gamma)$ is very skewed when γ is close to 0 and the Gaussian approximation is poor as a result. Figure 3 displays $Z_n(\gamma)$ in finite samples. But one can see that its empirical distribution could be quite skewed in small samples with medium to large family size ($n = 50, K = 4$, or $n = 50, K = 8$ or $n = 200, K = 8$), especially when γ is around 0. The performance of the Gaussian process depends on both the sample size n and the family size K. 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, especially with small sample size and large family size.

[Figure 3 about here.]

[Figure 4 about here.]

Next, we conduct simulation studies to compare the power of the LRT and PLRT against local alternative hypothesis. Two different types of local alternatives were considered. Type

I local alternatives $H_{1a} : \delta = \delta_1, \gamma = 0.5 - n^{-1/2}\tau$, approach the null from Region I, in the manner that γ approaches to 0.5 while δ is fixed. Type II local alternatives $H_{1a} : \delta = n^{-1/2}\tau, \gamma = \gamma_1$, approach the null from Region II, in the sense that δ approaches to 0 while γ is fixed. Figure 4 displays the power curves of the LRT and PLRT for two types of local alternatives. Both 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. 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.

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 0.042, close to the 0.05 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 0.05 for other choices of C. Thus, for analysis of power, we will drop 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 δ is small, say $\delta = 0.15, \gamma = 0$, the power is 0.483 for the LRT, higher than that of the PLRT, 0.438. When δ is large or γ 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 are generally less than $5 - 10\%$ in this setting. The differences become more noticeable in larger samples. Next, we look at the effect of C on statistical power. If we focus on each row of Table 2, the power of the PLRT decreases with C for certain alternative hypothesis. 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 0.053 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 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 for 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. The MLEs for δ and γ 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 choice 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 shizophrenia study. Figure 5 compared the asymptotic distribution of the LRT (left panel) and the PLRT (right panel) versus their empirical distribution in $1,000$ 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.

[Figure 5 about here.]

6. Discussion

Admixture models are special cases of *Class 2* problems that exhibit non-identifiability features under the null hypothesis. Testing for homogeneity in admixture models have 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, especially the PLRT, in the literature In terms of the choice between the LRT and PLRT, both have their own advantages and drawbacks. The PLRT has a convenient χ^2 type limiting distribution, but requires specification of a somewhat arbitrary penalty function. The choice of penalty affects the 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, by design, is powerful to detect the type of alternatives in region I, but not so powerful against other types of alternatives in region II and III. Thus, the PLRT may not be desirable when the proportion of linked families is small, while the LRT is generally powerful. In practice, one could consider these issues and decide which method is more appropriate for a particular application.

In this paper, we consider admixture models (1) whose first component is totally known. In some applications, there may be additional parameter β that is unknown for both components. The probability density function for such models has the form

$$f(y\,;\delta,\gamma,\beta) = (1-\delta)\,p(y;\gamma_0,\beta) + \delta\,p(y;\gamma,\beta), \tag{9}$$

or even more generally,

$$f(y\,;\delta,\gamma,\beta_1,\beta_2) = (1-\delta)\,p(y;\gamma_0,\beta_1) + \delta\,p(y;\gamma,\beta_2), \tag{10}$$

where β or β_1, β_2 are additional structural parameters. For example, $p(\cdot\,; \gamma, \beta)$ represents a normal distribution with mean γ and variance β and the two components might have equal or unequal variances. It will be interesting to study LRT and PLRT to such more complex admixture settings.

References

Abreu, P., Hodge, S., and Greenberg, D. (2002). Quantification of type I error probabilities for heterogeneity LOD scores. *Genetic epidemiology* **22,** 156–169.

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* pages 687–697.

Chen, H. and Chen, J. (2001). The likelihood ratio test for homogeneity in finite mixture models. *Canadian Journal of Statistics* **29,**.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference* **43,** 19–40.

Chiano, M. and Yates, J. (1995). Linkage detection under heterogeneity and the mixture problem. *Annals of human genetics* **59,** 83–95.

Davies, R. (1977). Hypothesis Testing When a Nuisance Parameter is Present Only Under the Null Hypothesis. *Biometrika* **64,** 247–254.

Faraway, J. (1993). Distribution of the admixture test for the detection of linkage under heterogeneity. *Genetic epidemiology* **10,** 75–83.

Fu, Y., Chen, J., and Kalbfleisch, J. (2006). Testing for Homogeneity in Genetic Linkage Analysis. *Statistica Sinica* **16,** 805–823.

Lemdani, M. and Pons, O. (1995). Tests for Genetic Linkage and Homogeneity. *Biometrics* **51,** 1033–1041.

Lemdani, M. and Pons, O. (1997). Likelihood ratio tests for genetic linkage. *Statistics and Probability Letters* **33,** 15–22.

Liang, K. and Rathouz, P. (1999). Hypothesis Testing Under Mixture Models: Application to Genetic Linkage Analysis. *Biometrics* **55,** 65–74.

Ott, J. (1999). *Analysis of human genetic linkage*. Johns Hopkins Univ Press.

Pulver, A., Karayiorgou, M., Wolyniec, P., Lasseter, V., Kasch, L., Nestadt, G., Antonarakis, S., Housman, D., Kazazian, H., Meyers, D., et al. (1994). Sequential strategy to identify a susceptibility gene for schizophrenia: report of potential linkage on chromosome 22q12-q13. 1: Part 1. *American journal of medical genetics* **54,**.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist* **9,** 225–228.

Rotnitzky, A., Cox, D., Bottai, M., and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli* **6,** 243–284.

Self, S. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82,** 605–610.

Shoukri, M. and Lathrop, G. (1993). Statistical testing of genetic linkage under heterogeneity. *Biometrics* **49,** 151–61.

Smith, C. (1963). Testing for heterogeneity of recombination fraction values in human genetics. *Annals of Human Genetics* **27,** 175–182.

Thomas, D. (2004). *Statistical methods in genetic epidemiology.* Oxford University Press, USA.

Appendix Sketch of Proofs

Proof of Lemma 1

This is a special case of Redner (1981), thus the proof is omitted.

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 Lemma 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

$$T_1(\epsilon_1) = \sup_{(\delta,\gamma)\in[\epsilon_1,1]\times\Gamma} 2\left\{l_n(\delta,\gamma) - l_n^0\right\} = W_{1n}^2 + o_p(1) \xrightarrow{D} \chi_1^2 \tag{A.1}$$

if γ_0 is an interior point of Γ, and

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

if γ_0 is a boundary point of Γ, where $l_n(\delta,\gamma) = \sum_i \log f(y_i;\delta,\gamma)$, $l_n^0 = \sum_i \log p(y_i;\gamma_0)$ and W_{1n} is a random variable with $W_{1n} \xrightarrow{D} N(0,1)$. If $p'(\cdot\,,\gamma_0) = \partial p(\cdot\,;\gamma_0)/\partial\gamma \neq 0$ with positive probability,

$$W_{1n} = \sum_i \frac{p'(y_i;\gamma_0)}{p(y_i;\gamma_0)} \cdot \left[\sum_i \left\{\frac{p'(y_i;\gamma_0)}{p(y_i;\gamma_0)}\right\}^2\right]^{-1/2}.$$

If $p'(\cdot\,;\gamma_0) = 0$ almost surely and $p''(\cdot\,;\gamma_0) = \partial^2 p(\cdot\,;\gamma_0)/\partial\gamma^2 \neq 0$ with positive probability, then

$$W_{1n} = \sum_i \frac{p''(y_i;\gamma_0)}{p(y_i;\gamma_0)} \cdot \left[\sum_i \left\{\frac{p''(y_i;\gamma_0)}{p(y_i;\gamma_0)}\right\}^2\right]^{-1/2}.$$

Proof. In the following, we write $p_i(\gamma) := p(y_i;\gamma)$, $p_i'(\gamma) := p'(y_i;\gamma)$ and $p_i''(\gamma) := p''(y_i;\gamma)$ for notational convenience. We first assume that γ_0 is an interior point of Γ.

For fixed $\delta \in [\epsilon_1, 1]$, the Taylor expansion around γ_0 gives

$$\begin{aligned} 2\{\log_n(\delta,\hat{\gamma}) - l_n^0\} &= 2\{\log_n(\delta,\hat{\gamma}) - l_n(\delta,\gamma_0)\} \\ &= 2\delta(\hat{\gamma}-\gamma_0)\sum_i \frac{p_i'(\gamma_0)}{p_i(\gamma_0)} - \delta^2(\hat{\gamma}-\gamma_0)^2 \sum_i \left\{\frac{p_i'(\gamma_0)}{p_i(\gamma_0)}\right\}^2 + o_p\{n\delta^2(\hat{\gamma}-\gamma_0)^2\} \\ &= \frac{\left\{\sum_i \frac{p'(y_i;\gamma_0)}{p(y_i;\gamma_0)}\right\}^2}{\sum_i \left\{\frac{p'(y_i;\gamma_0)}{p(y_i;\gamma_0)}\right\}^2} + o_p\{n(\hat{\gamma}-\gamma_0)^2\} \end{aligned}$$

where $\hat{\gamma}$ is the MLE for γ with fixed δ.

Case 1. If $p'(\cdot\,,\gamma_0) \neq 0$ with positive probability, then standard asymptotic properties for MLE imply that $\hat{\gamma} = \gamma_0 + O_p(n^{-1/2})$. 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 δ in the eqation above, one obtain the expansion in (A.1).

Case 2. If $p'(\cdot\,;\gamma_0) = 0$ almost surely and $p''(\cdot\,;\gamma_0) \neq 0$ with positive probability, the Fisher information for γ evaluated at γ_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

$$\begin{aligned} 2\{\log_n(\delta,\hat{\gamma}) - l_n^0\} &= 2\{\log_n(\delta,\hat{\gamma}) - l_n(\delta,\gamma_0)\} \\ &= \delta(\hat{\gamma}-\gamma_0)^2\sum_i \frac{p_i''(\gamma_0)}{p_i(\gamma_0)} - \frac{1}{4}\delta^2(\hat{\gamma}-\gamma_0)^4 \sum_i \left\{\frac{p_i''(\gamma_0)}{p_i(\gamma_0)}\right\}^2 + o_p\{n\delta^2(\hat{\gamma}-\gamma_0)^4\} \\ &= \frac{\left\{\sum_i \frac{p''(y_i;\gamma_0)}{p(y_i;\gamma_0)}\right\}^2}{\sum_i \left\{\frac{p''(y_i;\gamma_0)}{p(y_i;\gamma_0)}\right\}^2} + o_p\{n(\hat{\gamma}-\gamma_0)^4\}. \end{aligned}$$

Using results from Rotnitzky et al. (2000), $\hat{\gamma} = \gamma_0 + O_p(n^{-1/4})$ under this situation. Similar to the argument in *Case 1*, one can obtain the expansion in (A.1), except that W_{1n} involves $p''(\cdot;\gamma_0)$ instead of $p'(\cdot;\gamma_0)$.

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 $T_1(\epsilon_1) \xrightarrow{D} \chi_1^2$ for any $0 < \epsilon_1 \leqslant 1$.

Finally, when γ_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 $\gamma \in \Gamma\backslash(\gamma_0-\epsilon_2, \gamma_0-\epsilon_2)$. We further assume the following three conditions,

(i) $p(\cdot;\gamma) \neq p(\cdot;\gamma_0)$ with positive probability for all $\gamma \neq \gamma_0$;

(ii) there exists $\eta > 0$ with $\mathrm{E}_{\gamma_0}\{\, p(\cdot;\gamma)/p(\cdot;\gamma_0) - 1 \,\}^2 \geqslant \eta$ for all $\gamma \in \Gamma\backslash(\gamma_0 - \epsilon_2, \gamma_0 - \epsilon_2)$;

(iii) the process $W_{2n}(\gamma)$ is tight, where

$$W_{2n}(\gamma) = \sum_{i=1}^{n} \left\{ \frac{p(y_i;\gamma)}{p(y_i;\gamma_0)} - 1 \right\} \cdot \left[\sum_i \left\{ \frac{p(y_i;\gamma)}{p(y_i;\gamma_0)} - 1 \right\}^2 \right]^{-1/2}.$$

Then, the LRT statistic over Region II & IV satisfies

$$\begin{aligned} T_2(\epsilon_2) &= \sup_{(\delta,\gamma)\in[0,1]\times\Gamma\backslash(\gamma_0-\epsilon_2,\gamma_0-\epsilon_2)} 2\left\{ l_n(\delta,\gamma) - l_n^0 \right\} \\ &\xrightarrow{D} \sup_{\gamma\in\Gamma\backslash(\gamma_0-\epsilon_2,\gamma_0+\epsilon_2)} \{\, W_2^+(\gamma) \,\}^2, \end{aligned} \tag{A.3}$$

where $W_2(\gamma) = \lim_{n\to\infty} W_{2n}(\gamma)$ is a Gaussian process with mean 0, variance 1 and certain autocorrelation function $\rho(\gamma_1, \gamma_2)$.

Proof. For fixed $\gamma \in \Gamma\backslash(\gamma_0 - \epsilon_2, \gamma_0 - \epsilon_2)$, the Taylor expansion around $\delta = 0$ gives

$$\begin{aligned} 2\{\, \log_n(\hat{\delta},\gamma) - l_n^0 \,\} &= 2\{\, \log_n(\hat{\delta},\gamma) - l_n(0,\gamma) \,\} \\ &= 2\hat{\delta}\sum_i \left\{ \frac{p(y_i;\gamma)}{p(y_i;\gamma_0)} - 1 \right\} - \hat{\delta}^2 \sum_i \left\{ \frac{p(y_i;\gamma)}{p(y_i;\gamma_0)} - 1 \right\}^2 + o_p(n\hat{\delta}^2) \\ &= \frac{\left[\sum_i \left\{ \frac{p(y_i;\gamma)}{p(y_i;\gamma_0)} - 1 \right\} \right]^2}{\sum_i \left\{ \frac{p(y_i;\gamma)}{p(y_i;\gamma_0)} - 1 \right\}^2} + o_p(n\hat{\delta}^2) \\ &= \{\, W_{2n}^+(\gamma) \,\}^2 + o_p(n\hat{\delta}^2), \end{aligned}$$

where $\hat{\delta}$ is the MLE for δ for fixed γ. 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 $\gamma \in \Gamma\backslash(\gamma_0 - \epsilon_2, \gamma_0 - \epsilon_2)$. Taking supremum over γ, one can obtain (A.3). Note that equation (A.3) involves $W_2^+(\gamma)$ instead of $W_2(\gamma)$ 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

$$T_3(\epsilon_1, \epsilon_2) = \sup_{(\delta,\gamma)\in[0,\epsilon_1]\times[\gamma_0-\epsilon_2,\gamma_0-\epsilon_2]} 2\left\{ l_n(\delta,\gamma) - l_n^0 \right\}.$$

If γ_0 is an interior point of Γ,

$$W_{1n}^2 + o_p(1) \leqslant T_3(\epsilon_1, \epsilon_2) \leqslant W_{1n}^2 + (\epsilon_2 + \epsilon_2^2)\, O_p(1), \tag{A.4}$$

and if γ_0 is a boundary point,

$$\{\, W_{1n}^+ \,\}^2 + o_p(1) \leqslant T_3(\epsilon_1, \epsilon_2) \leqslant \{\, W_{1n}^+ \,\}^2 + (\epsilon_2 + \epsilon_2^2)\, O_p(1), \tag{A.5}$$

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

Proof. We provide the proof when γ_0 is an interior point and when $p'(\cdot;\gamma_0) \neq 0$ 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,

$$\begin{aligned} T_3(\epsilon_1, \epsilon_2) &= \sup_{(\delta,\gamma)\in[0,\epsilon_1]\times[\gamma_0-\epsilon_2,\gamma_0-\epsilon_2]} 2\,\left\{ l_n(\delta,\gamma) - l_n^0 \right\} \\ &\geqslant \sup_{(\delta,\gamma)\in\{\epsilon_1\}\times[\gamma_0-\epsilon_2,\gamma_0-\epsilon_2]} 2\,\left\{ l_n(\delta,\gamma) - l_n^0 \right\} \\ &= W_{1n}^2 + o_p(1), \end{aligned}$$

where the last equation is obtained from Lemma 2. Next, we expand $T_3(\epsilon_1, \epsilon_2)$ around $(0, \gamma_0)$,

$$\begin{aligned} T_3(\epsilon_1, \epsilon_2) &= 2\{\, \log_n(\hat{\delta}, \hat{\gamma}) - l_n(0, \gamma_0) \,\} \\ &= 2\hat{\delta}(\hat{\gamma} - \gamma_0) \sum_i \frac{p_i'(\gamma_0)}{p_i(\gamma_0)} + \hat{\delta}(\hat{\gamma} - \gamma_0)^2 \sum_i \frac{p_i''(\gamma_0)}{p_i(\gamma_0)} \\ &\quad + \frac{1}{3}\hat{\delta}(\hat{\gamma} - \gamma_0)^3 \sum_i \frac{p_i'''(\gamma_0)}{p_i(\gamma_0)} - \hat{\delta}^2(\hat{\gamma} - \gamma_0)^2 \sum_i \left\{ \frac{p_i'(\gamma_0)}{p_i(\gamma_0)} \right\}^2 \\ &\quad + o_p\{\, n\hat{\delta}^2(\hat{\gamma} - \gamma_0)^2 \,\}. \end{aligned}$$

Based on this fourth order approximation, one can show that $\hat{\delta}(\hat{\gamma} - \gamma_0) = O_p(n^{-1/2})$ and

$\hat{\gamma} - \gamma_0 = O_p(1)$. Thus,

$$\begin{aligned} T_3(\epsilon_1, \epsilon_2) &= 2\{ \log_n(\hat{\delta}, \hat{\gamma}) - l_n(0, \gamma_0) \} \\ &= 2\,\hat{\delta}(\hat{\gamma} - \gamma_0) \sum_i \frac{p_i'(\gamma_0)}{p_i(\gamma_0)} - \hat{\delta}^2(\hat{\gamma} - \gamma_0)^2 \sum_i \left\{ \frac{p_i'(\gamma_0)}{p_i(\gamma_0)} \right\}^2 \qquad \text{(A.6)} \\ &\quad + \hat{\delta}(\hat{\gamma} - \gamma_0)^2 \sum_i \frac{p_i''(\gamma_0)}{p_i(\gamma_0)} + \frac{1}{3}\hat{\delta}(\hat{\gamma} - \gamma_0)^3 \sum_i \frac{p_i'''(\gamma_0)}{p_i(\gamma_0)} \qquad \text{(A.7)} \\ &\quad + o_p\{ n\hat{\delta}^2(\hat{\gamma} - \gamma_0)^2 \} \\ &\leqslant W_{1n}^2 + (\epsilon_2 + \epsilon_2^2)\, O_p(1), \end{aligned}$$

where (A.6) is equivalent to W_{1n}^2 according to Lemma 2 and (A.7) is bounded because $|\gamma - \gamma_0| \leqslant \epsilon_2$ always hold in Region III. Thus, (A.4) follows.

Lemma 5 The processes W_{1n}, $W_{2n}(\gamma)$, W_1 and $W_2(\gamma)$ satisfy $W_{1n} = \lim_{\gamma \to \gamma_0} W_{2n}(\gamma)$ and $W_1 = \lim_{\gamma \to \gamma_0} W_2(\gamma)$.

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.,

$$LRT = \max\{ T_1(\epsilon_1), T_2(\epsilon_2), T_3(\epsilon_1, \epsilon_2) \}. \qquad \text{(A.8)}$$

Define $Z_n(\gamma) = W_{2n}(\gamma)$, $Z_n(\gamma_0) = W_{1n}$ and $Z(\gamma) = \lim_{n \to \infty} Z_n(\gamma)$, then $Z_n(\gamma)$ and $Z(\gamma)$ 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.

Formulas for $\rho(\gamma_1, \gamma_2)$ in genetic linkage admixture models

PK case.

$$\rho(\gamma_1,\gamma_2) = \begin{cases} \dfrac{\sum_i \phi(\gamma_1,\gamma_2;K_i)}{\sqrt{\sum_i \phi(\gamma_1,\gamma_1;K_i)}\cdot\sqrt{\sum_i \phi(\gamma_2,\gamma_2;K_i)}}, & \text{if } \gamma_1 \in [0,0.5), \gamma_2 \in [0,0.5), \\ \dfrac{(1-2\gamma_2)\cdot\sqrt{\sum_i K_i}}{\sqrt{\sum_i \phi(\gamma_2,\gamma_2;K_i)}}, & \text{if } \gamma_1 = 0.5, \gamma_2 \in [0,0.5), \end{cases}$$

where $\phi(\gamma_1, \gamma_2; K) = \{1 + (1-2\gamma_1)(1-2\gamma_2)\}^K - 1$ and K_i is the size for family i, for $i = 1, 2, \cdots, n$.

PU case.

$$\rho(\gamma_1,\gamma_2) = \begin{cases} \dfrac{\sum_i \phi(\gamma_1,\gamma_2;K_i)}{\sqrt{\sum_i \phi(\gamma_1,\gamma_1;K_i)}\cdot\sqrt{\sum_i \phi(\gamma_2,\gamma_2;K_i)}}, & \text{if } \gamma_1 \in [0,0.5), \gamma_2 \in [0,0.5), \\ \dfrac{(1-2\gamma_2)^2\cdot\sqrt{\sum_i K_i(K_i-1)/2}}{\sqrt{\sum_i \phi(\gamma_2,\gamma_2;K_i)}}, & \text{if } \gamma_1 = 0.5, \gamma_2 \in [0,0.5), \end{cases}$$

where $\phi(\gamma_1, \gamma_2; K) = \left[\{1 + (1-2\gamma_1)(1-2\gamma_2)\}^K + \{1 - (1-2\gamma_1)(1-2\gamma_2)\}^K - 2\right]/2$ and K_i is the size of family i, for $i = 1, 2, \cdots, n$.

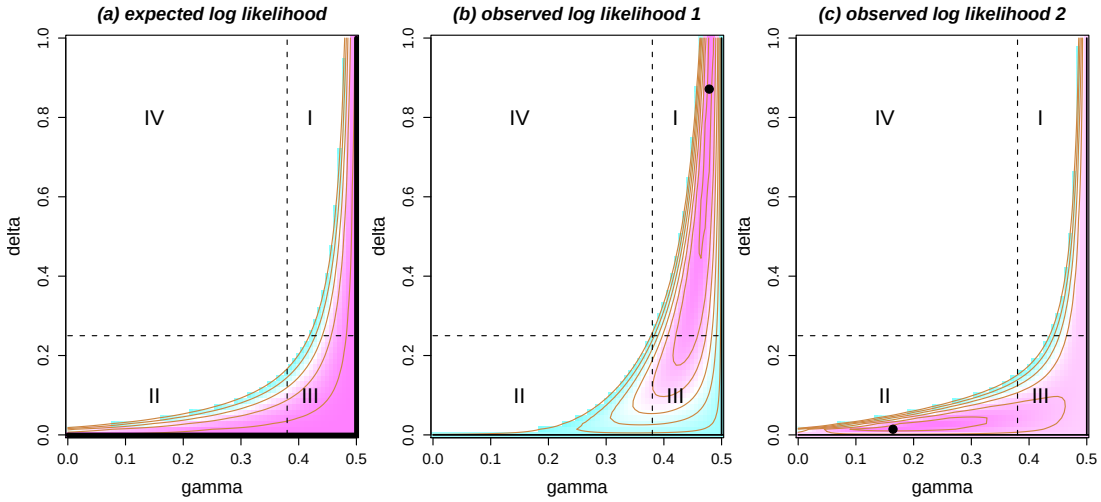

Figure 1. Expected and observed log-likelihood functions for admixture model in genetic linkage analysis. Pink corresponds to large log-likelihood values, and light blue corresponds to small values. Panel (a) displays the expected log-likelihood function, where solid lines represent the set of true values under H_0. Panels (b) and (c) show observed log-likelihood functions for two datasets simulated under H_0.

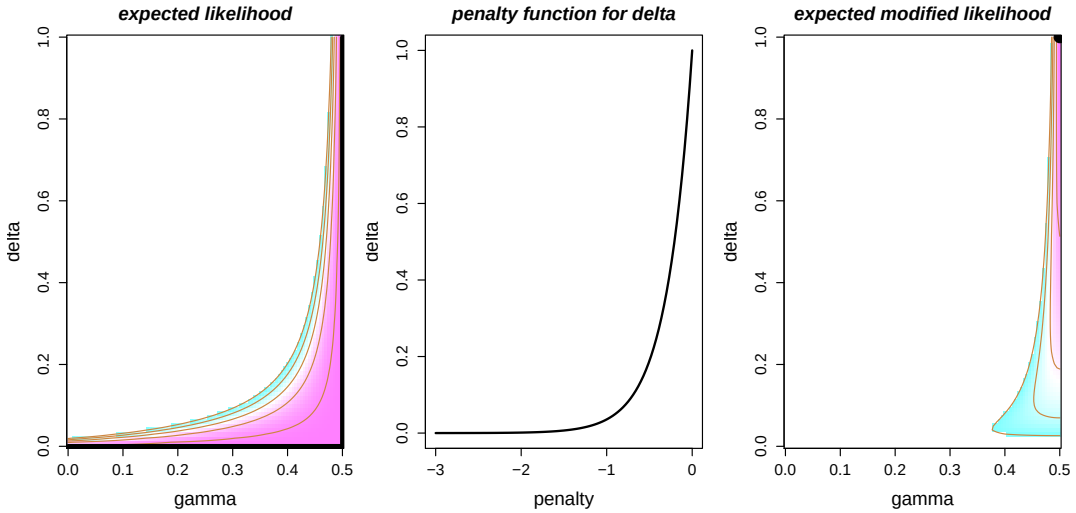

Figure 2. Expected log-likelihood and expected penalized log-likelihood functions for admixture model in genetic linkage analysis.

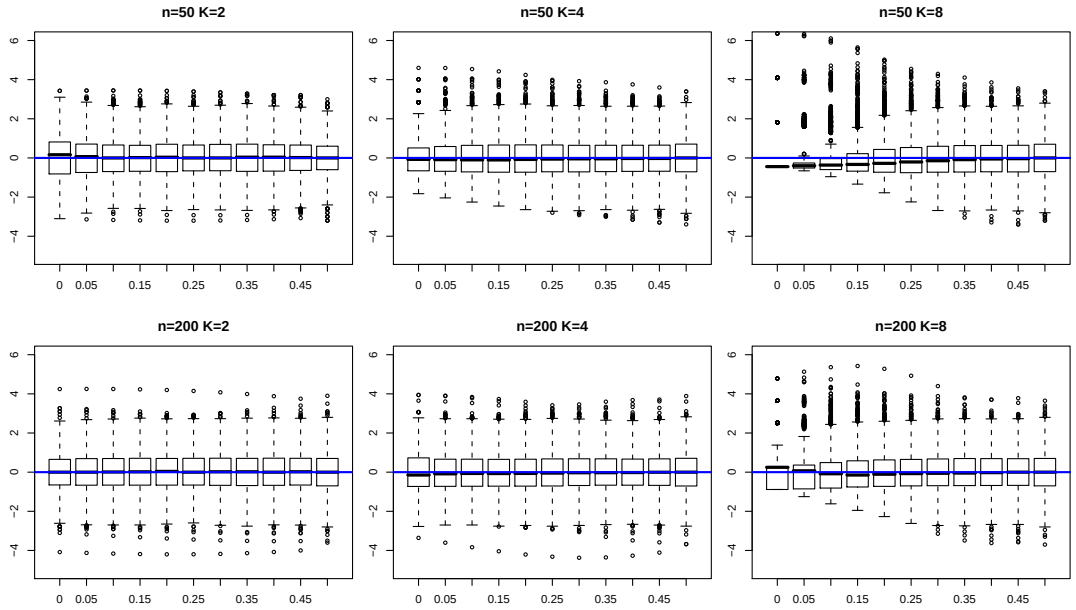

Figure 3. Boxplots of the process $Z(\gamma)$ for different sample size n and family size K. The process $Z(\gamma)$ converges to a Gaussian process in large samples. But in finite samples, it could be quite skewed.

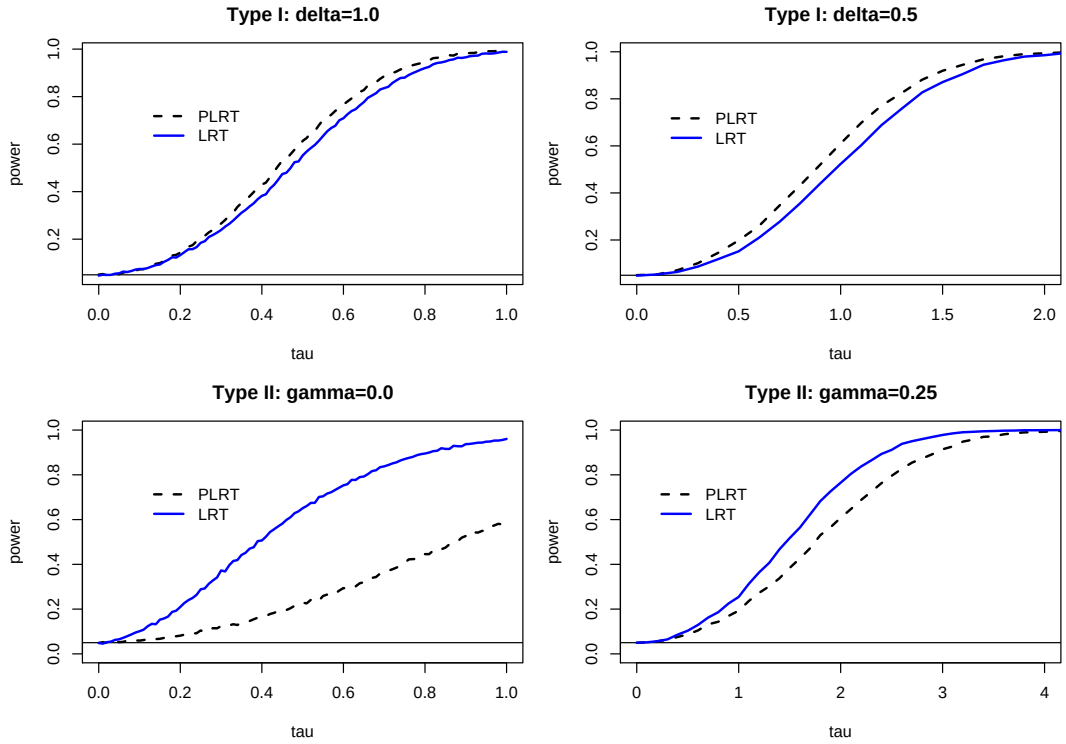

Figure 4. Power curves versus Type I and II local alternatives for the LRT and PLRT.

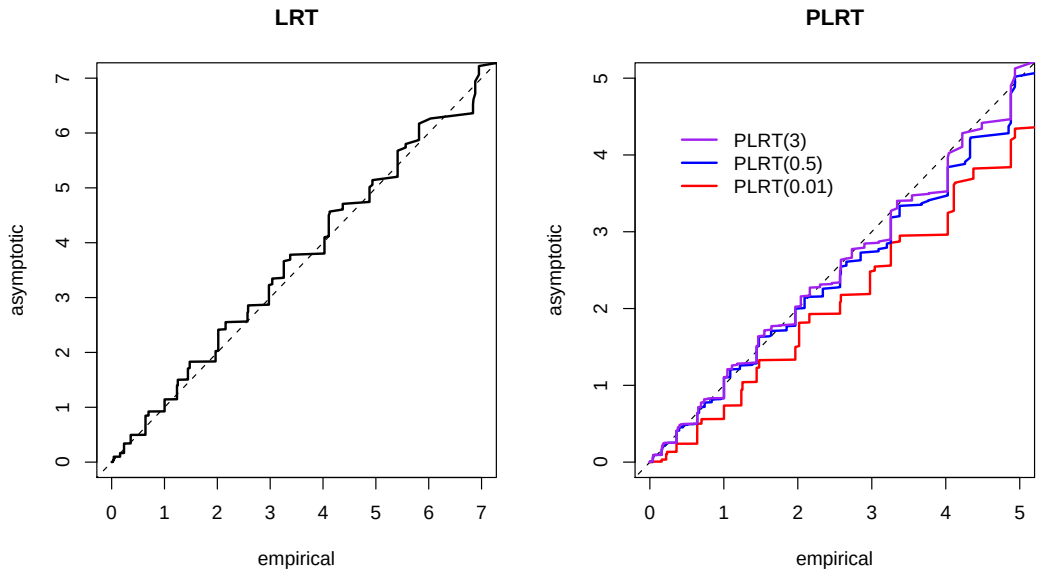

Figure 5. Asymptotic versus empirical distribution of the LRT statistic and the PLRT statistic. The horizonal axis is the empirical distribution from 1,000 simulations. The vertical axis is the asymptotic distribution simulated from Gaussian processes.

Table 1
Comparison of LRT vs alternative methods. The process $Z(\gamma)$ is defined in Theorem 1 in general admixture models, and in Corollary 1 for genetic linkage model. "Turning par." means dependence on specification of a tuning parameter. χ^2 means a $50:50$ mixture of χ^2_0 and χ^2_1.

Method	Test statistic	Tuning parameter	Combined parameter space	Asymptotic null distribution
LRT	LRT	No	$[0,1]\times[0,0.5]$	$\sup_{\gamma\in[0,0.5]}\{Z^+(\gamma)\}^2$
Simple LRT	$LRT^{S,\delta}(\delta_1)$	Yes, δ_1	$\delta_1\times[0,0.5]$	$\{Z^+(0.5)\}^2 \to \chi^2$
	$LRT^{S,\gamma}(\gamma_1)$	Yes, γ_1	$[0,1]\times\gamma_1$	$\{Z^+(\gamma_1)\}^2 \to \chi^2$
Restricted LRT	$LRT^{R,\delta}(\epsilon_1)$	Yes, ϵ_1	$[\epsilon_1,1]\times[0,0.5]$	$\{Z^+(0.5)\}^2 \to \chi^2$
	$LRT^{R,\gamma}(\epsilon_2)$	Yes, ϵ_2	$(0,1]\times[0,0.5-\epsilon_2]$	$\sup_{\gamma\in[0,0.5-\epsilon_2]}\{Z^+(\gamma)\}^2$
PLRT	PLRT(C)	Yes, C	$[\,\epsilon_1(C),1]\times[0,0.5]$	$\{Z^+(0.5)\}^2 \to \chi^2$

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.

δ	γ	LRT	PLRT(C)							
			0.011	0.135	0.368	0.607	1.000	1.649	4.481	148
0	0.5	0.042	0.070	0.053	0.053	0.046	0.044	0.043	0.043	0.043
0.05	0.3	0.063	0.098	0.078	0.078	0.070	0.069	0.068	0.068	0.068
0.05	0.0	0.129	0.189	0.148	0.148	0.129	0.122	0.121	0.120	0.120
0.10	0.3	0.089	0.137	0.112	0.112	0.099	0.097	0.096	0.096	0.096
0.10	0.0	0.288	0.380	0.317	0.317	0.272	0.256	0.251	0.249	0.249
0.15	0.3	0.121	0.193	0.155	0.155	0.142	0.138	0.137	0.137	0.137
0.15	0.0	0.483	0.591	0.515	0.515	0.470	0.438	0.428	0.425	0.425
0.20	0.3	0.174	0.254	0.209	0.209	0.192	0.186	0.184	0.184	0.184
0.20	0.0	0.695	0.781	0.717	0.717	0.668	0.637	0.622	0.616	0.616
0.25	0.3	0.231	0.317	0.269	0.269	0.247	0.241	0.240	0.240	0.240
0.25	0.0	0.842	0.901	0.855	0.855	0.822	0.797	0.784	0.776	0.776
0.30	0.3	0.295	0.391	0.343	0.343	0.320	0.313	0.311	0.311	0.311
0.30	0.0	0.943	0.967	0.947	0.947	0.926	0.908	0.897	0.892	0.892
0.40	0.3	0.447	0.547	0.501	0.501	0.476	0.468	0.465	0.464	0.464
0.40	0.0	0.995	0.997	0.995	0.995	0.991	0.988	0.986	0.984	0.983
0.50	0.3	0.595	0.692	0.651	0.651	0.629	0.621	0.619	0.618	0.618
0.50	0.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
0.70	0.3	0.853	0.900	0.884	0.884	0.872	0.867	0.866	0.866	0.866
0.70	0.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.90	0.3	0.968	0.981	0.978	0.978	0.977	0.976	0.976	0.976	0.976
0.90	0.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.00	0.3	0.987	0.995	0.993	0.993	0.993	0.993	0.993	0.993	0.993
1.00	0.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000