The Canadian Journal of Statistics
Vol. ?, No. ?, 2006, Pages ???-???
La revue canadienne de statistique

Asymptotic properties of the likelihood ratio test statistics with the possible triangle constraint in Affected-Sib-Pair analysis

Zeny Z. Feng, Jiahua Chen and Mary E. Thompson

Key words and phrases: ASP method; linkage analysis; possible triangle constraint; identical by descent (IBD); likelihood ratio test (LRT).
MSC 2000: Primary 92D10; secondary 92D30.

Abstract: In an Affected-Sib-Pair (ASP) genetic linkage analysis, IBD data for affected sib pairs are routinely collected at a large number of markers along chromosomes. Under very general genetic assumptions, the IBD distribution at each marker satisfies the possible triangle constraint. Statistical analysis of IBD data hence should utilize this information to improve efficiency. At the same time, this constraint renders the usual regularity conditions for likelihood based statistical methods unsatisfied. In this paper, we study the asymptotic properties of the likelihood ratio test under the possible triangle constraint. We derive the limiting distribution of the LRT statistic based on data from a single locus. The precision of the limiting distribution and the power of the test are investigated by simulation. Further, we study the test based on the supremum of the LRT statistics over the markers distributed throughout a chromosome. Instead of deriving a limiting distribution theoretically, we propose to use a mixture of chisquare distributions to approximate the true distribution. The simulation results show that this approach has desirable simplicity and satisfactory precision.

1. INTRODUCTION

The sib-pair method is a non-parametric statistical method for linkage analysis introduced in Penrose (1935). The analysis is based on genetic information on sib pairs and their parents. In comparison, most other linkage analyses need information from families of three or more generations. Thus, the information needed by the sib-pair method is much easier to collect. Compared with a discordant sib pair, a concordant affected sib pair contains more information for the purpose of linkage analysis between a disease-susceptibility gene and a marker under investigation (Suarez et al. 1978 and Liang et al. 2000). Therefore, it is cost efficient to collect genetic information only from families with at least two affected sibs.

In affected-sib-pair analysis, data on alleles shared identical by descent (IBD) at a marker are collected. For a given affected sib pair, the number of alleles IBD at a marker can be 0, 1 or 2. If the marker under investigation is not linked to a disease locus, the IBD distribution at this marker, denoted by (π_0, π_1, π_2), is $(0.25, 0.5, 0.25)$. If the marker is linked to a disease locus, an affected sib pair tends to share more alleles IBD at this marker than under no linkage. Thus, with linkage, the IBD value increases stochastically. A hypothesis test of $H_0 : (\pi_0, \pi_1, \pi_2) = (0.25, 0.5, 0.25)$ versus $H_A : (\pi_0, \pi_1, \pi_2) \neq (0.25, 0.5, 0.25)$ is not statistically efficient as it does not take this genetic fact into account.

Louis, Payami & Thomson (1987) and Holmans (1993) were the first to discover that the IBD distribution at a locus of an affected sib pair satisfies the "possible triangle constraint" under certain conditions. The possible triangle constraint on the IBD distribution is given by: $\pi_1 \leq 1/2, 2\pi_0 \leq \pi_1$. See left panel of Figure 1. The validity of the possible triangle constraint under various genetic models has been intensively investigated. See Holmans (1993) and Farrall (1997). Most recently, Feng, Chen & Thompson (2005) showed that the possible triangle constraint is valid under very general genetic model assumptions. Taking the possible triangle constraint into account in linkage analysis effectively reduces the type I error and improves the power. This is achieved by taking only deviations of the IBD distribution in the direction of the possible triangle constraint as evidence of linkage. For a given size of test and a desired detecting power, the sample size required can be effectively reduced. Thus, it is clearly of great importance to make use of this constraint in linkage analysis.

It is well known that the null limiting distribution of the usual likelihood ratio test (LRT) statistic is chisquare under some regularity conditions. The LRT test statistic for linkage at a given locus based on data from sib pairs with possible triangle constraint is no longer chisquare. Holmans (1993) showed that in this case, the limiting distribution is a mixture of χ^2_1 and χ^2_2, but the mixing proportion was left undetermined. Using an alternative simple and straightforward approach, we obtain the limiting distribution of the LRT statistic under the possible triangle constraint and explicitly determine the mixing proportion. Simulation studies are used to assess the precision of our limiting distribution and to demonstrate the power gain of the test with the possible triangle constraint.

With modern technology, high throughput data are commonly available in which the IBD information is collected on a large number of loci along many segments of chromosomes. Hence, a LRT statistic can be computed at each locus where the IBD information was collected. Consequently, we arrive at a stochastic process made up of LRT statistic values at markers along chromosomes. The supremum of the LRT statistics becomes a natural test statistic for linkage: the test is significant when it exceeds a threshold value determined by its distribution and the significance level. It would hence be ideal to know the limiting distribution of the supremum of the LRT statistics over a chromosome region. As this is a difficult mathematical problem, we instead use simulation to determine a mixture of chisquare distributions that forms a good approximation to the target distribution. We take the degrees of freedom and the mixing proportion to be simple functions of the length of the chromosome under investigation. We find that the quantiles of the resulting chisquare mixture approximate those of the simulated data very well. They thus provide a useful and practical solution for linkage analysis based on affected sib pairs.

The paper is organized as follows. In Section 2, we derive the limiting distribution of the LRT statistic under the constraint at a single locus. Simulation results for assessing the precision and the power follow. In Section 3, we study the empirical distribution of the supremum of LRT statistics.

2. Limiting distribution of LRT statistic

In this section, we derive the limiting distribution of the LRT statistic. The precision of the limiting distribution is examined by a simulation study. The power of the LRT with possible triangle constraint is also investigated via simulation.

2.1 The limiting distribution of LRT statistic under the possible triangle constraint.

Suppose N independent sib pairs affected with a certain disease are recruited and the IBD values at a given marker are collected. Let N_0, N_1 and N_2 be observed frequencies of IBD values equaling 0, 1 and 2. The random variables N_0 and N_1 are jointly multinomially distributed with parameters

π_0, π_1 and $\pi_2 = 1 - \pi_0 - \pi_1$. The log-likelihood function of π_0 and π_1 is

$$l(\pi_0, \pi_1) = N_0 \log \pi_0 + N_1 \log \pi_1 + (N - N_0 - N_1) \log(1 - \pi_0 - \pi_1).$$

It is well known that $\hat{\pi}_0 = \frac{N_0}{N}$ and $\hat{\pi}_1 = \frac{N_1}{N}$ are the unique maximum likelihood estimates (MLE) of π_0 and π_1 with no constraints imposed. To test the null hypothesis of no linkage, the log-likelihood ratio statistic is given by

$$\Lambda_N = 2N[\hat{\pi}_0 \log(4\hat{\pi}_0) + \hat{\pi}_1 \log(2\hat{\pi}_1) + (1 - \hat{\pi}_0 - \hat{\pi}_1) \log(4 - 4\hat{\pi}_0 - 4\hat{\pi}_1)]. \quad (1)$$

By a classical result in Wilks (1938), Λ_N in (1) converges to χ^2_2 in distribution as $N \to \infty$ under the null hypothesis of no linkage.

As we pointed out earlier, the IBD distribution of an affected sib pair satisfies the possible triangle constraint. It is advantageous to reject the null hypothesis only if the deviation of the IBD distribution is in the direction of the triangle region. Thus, the test for linkage becomes a test of

$$H_0 : (\pi_0, \pi_1, \pi_2) = (1/4, 1/2, 1/4)$$
$$\text{vs}$$
$$H_A : 2\pi_0 \le \pi_1 \le 1/2.$$

The problem of the computation of the restricted MLE, denoted as $\tilde{\pi}$, under the possible triangle constraint was studied in Holmans (1993). The numerical solution can be easily obtained following a simple procedure. The likelihood ratio statistic under the possible triangle constraint is given by

$$\tilde{\Lambda}_N = 2N[\hat{\pi}_0 \log(4\tilde{\pi}_0) + \hat{\pi}_1 \log(2\tilde{\pi}_1) + (1 - \hat{\pi}_0 - \hat{\pi}_1) \log(4 - 4\tilde{\pi}_0 - 4\tilde{\pi}_1)]. \quad (2)$$

Due to the violation of regularity conditions, the limiting distribution of $\tilde{\Lambda}_N$ differs from the usual chisquare. We derive its limiting distribution as follows.

For simplicity, we orthogonalize parameters by transforming from the (π_0, π_1) to the (w_0, w_1) space, where

$$w_0 = \sqrt{2}(2\pi_0 + \pi_1 - 1), \qquad w_1 = 2\pi_1 - 1.$$

See Figure 1 for illustration. By this transformation, the MLEs $\hat{\pi}$ and the restricted MLEs $\tilde{\pi}$ are transformed to $\hat{w}$ and $\tilde{w}$ respectively. Under the null hypothesis, it is easy to show that

$$\text{var}(\hat{\pi}_0) = \frac{3}{16N}, \quad \text{var}(\hat{\pi}_1) = \frac{1}{4N}, \quad \text{cov}(\hat{\pi}_0, \hat{\pi}_1) = -\frac{1}{8N}.$$

Consequently,

$$E(\hat{w}_0) = 0, \quad E(\hat{w}_1) = 0,$$

and

$$N\text{var}(\hat{w}_0) = 1, \quad N\text{var}(\hat{w}_1) = 1, \quad \text{cov}(\hat{w}_0, \hat{w}_1) = 0.$$

Here it is seen that, for large N, $(\sqrt{N}\hat{w}_0, \sqrt{N}\hat{w}_1)$ is approximately bivariate $N(\mathbf{0}, I_2)$ distributed, where I_2 is the 2×2 identity matrix. For simplicity, we will treat $(\sqrt{N}\hat{w}_0, \sqrt{N}\hat{w}_1)$ as if they are independent binormally distributed in our later discussion. By applying Chebyshev's inequality, we find

$$\hat{\pi}_j - \pi_j = O_p(N^{-1/2}), \quad j = 0, 1, 2,$$

and

$$\tilde{\pi}_j - \pi_j = O_p(N^{-1/2}), \quad j = 0, 1, 2,$$

under both the null and the alternative models. By Taylor's expansion, the log terms of the likelihood ratio function in (2) can be rewritten as, for example,

$$\begin{aligned} \log(4\tilde{\pi}_0) &= \log\{1 + (4\tilde{\pi}_0 - 1)\} \\ &= (4\tilde{\pi}_0 - 1) - \tfrac{1}{2}(4\tilde{\pi}_0 - 1)^2 + O_p(N^{-3/2}). \end{aligned}$$

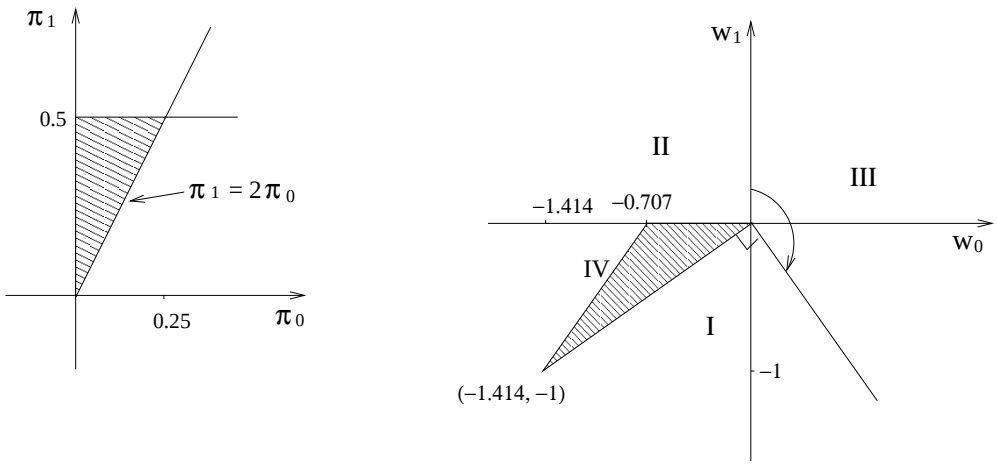


Figure 1: Transformation from (π_0, π_1) plane to (w_0, w_1) plane.

Thus, in summary, we obtain

$$\tilde{\Lambda}_N = N\tilde{w}_0^2 + N\tilde{w}_1^2 + o_p(1). \tag{3}$$

Hence, the null limiting distribution of $\tilde{\Lambda}_N$ is determined by that of $\tilde{w}_0$ and $\tilde{w}_1$.

In Figure 1, the three vertices of the possible triangle on the (π_0, π_1) plane after transformation become:

$$\begin{array}{rcl} (\pi_0, \pi_1) = (\frac{1}{4}, \frac{1}{2}) & \rightarrow & (w_0, w_1) = (0, 0), \\ (\pi_0, \pi_1) = (0, 0) & \rightarrow & (w_0, w_1) = (-\sqrt{2}, -1), \\ (\pi_0, \pi_1) = (0, \frac{1}{2}) & \rightarrow & (w_0, w_1) = (-\frac{1}{\sqrt{2}}, 0). \end{array}$$

The (w_0, w_1) plane is divided into four cones. Note that the possible triangle region in the (w_0, w_1) plane (shaded region) does not cover an entire cone. However, the probability of points $(\hat{w}_0, \hat{w}_1)$ falling outside the triangle but within the fourth cone is $P(\hat{w}_1 > \sqrt{2}\hat{w}_0 + 1)$, which is negligible for large N.

For $x > 0$, we work for an expression of $P(\tilde{\Lambda}_N \geq x)$ under the null hypothesis. Let θ be the angle between the positive w_0 axis and the vector $(\hat{w}_0, \hat{w}_1)$. Note that under the normality assumption on $(\sqrt{N}\hat{w}_0, \sqrt{N}\hat{w}_1)$, θ has a uniform distribution in $[0, 2\pi]$ and is independent of the norm $\rho = \sqrt{N(\hat{w}_0^2 + \hat{w}_1^2)}$. Decomposing the probability by conditioning on the position of θ, we may write

$$\begin{array}{rcl} P(\tilde{\Lambda}_N \geq x) & = & P(\theta \in \text{Cone I})\, P(\tilde{\Lambda}_N \geq x | \theta \in \text{Cone I}) \\ & & + P(\theta \in \text{Cone II})\, P(\tilde{\Lambda}_N \geq x | \theta \in \text{Cone II}) \\ & & + P(\theta \in \text{Cone III})\, P(\tilde{\Lambda}_N \geq x | \theta \in \text{Cone III}) \\ & & + P(\theta \in \text{Cone IV})\, P(\tilde{\Lambda}_N \geq x | \theta \in \text{Cone IV}) \end{array} \tag{4}$$

Suppose $(\hat{w}_0, \hat{w}_1)$ falls in the triangle region. In this case, we have $\tilde{\Lambda}_N = N(\hat{w}_0^2 + \hat{w}_1^2) = \rho^2$. Since ρ and θ are independent, we have

$$P(\tilde{\Lambda}_N \geq x | \theta \in \text{Cone IV}) = P(\chi_2^2 \geq x),$$

where χ_2^2 stands for a random variable with a χ_2^2 distribution. This takes care of the fourth term in (4).

For $(\hat{w}_0, \hat{w}_1)$ falling in Cone I, $(\tilde{w}_0, \tilde{w}_1)$ is the projection of $(\hat{w}_0, \hat{w}_1)$ on the line $w_1 = \frac{\sqrt{2}}{2} w_0$. See Figure 2. Let ϕ be the angle between the line $w_1 = \frac{\sqrt{2}}{2} w_0$ and the vector $(\hat{w}_0, \hat{w}_1)$. Then

$$\tilde{\Lambda}_N = \rho^2 \cos^2 \phi,$$

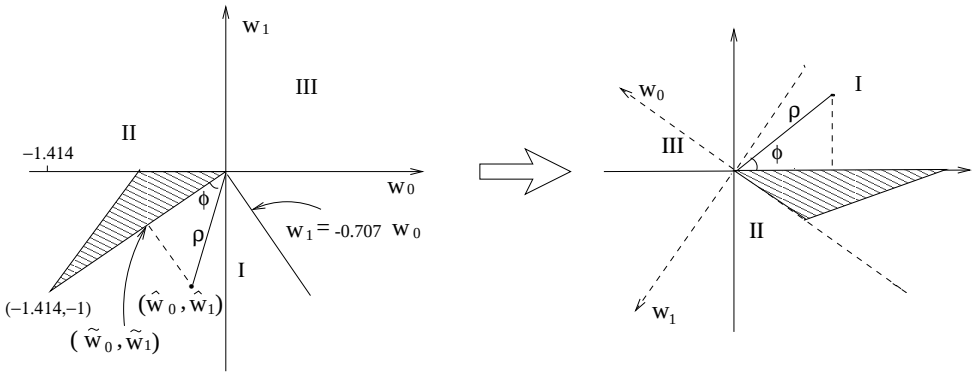


Figure 2: Rotate the (w_0, w_1) plane such that Cone I is the first quadrant.

and $\rho^2 \cos^2 \phi$ has a χ^2_1 distribution.

Rotate the (w_0, w_1) plane as we show in Figure 2, such that Cone I is the first quadrant (Q_1) of the rotated plane. By circular symmetry of the distribution of $(\hat{w}_0, \hat{w}_1)$ and the periodicity properties of $\cos \phi$,

$$P(\tilde{\Lambda}_N \geq x | \theta \in \text{Cone I}) = P(\rho^2 \cos^2 \phi \geq x | \phi \in Q_1) = P(\chi^2_1 \geq x).$$

Similarly, $P(\tilde{\Lambda}_N \geq x | \theta \in \text{Cone II}) = P(\chi^2_1 \geq x)$. For $(\hat{w}_0, \hat{w}_1)$ falling in Cone III, the corresponding $(\tilde{w}_0, \tilde{w}_1) = 0$. Thus we have $P(\tilde{\Lambda}_N \geq x | \theta \in \text{Cone III}) = 0$.

In summary, the expression of $\tilde{\Lambda}_N$ can be written as:

$$\tilde{\Lambda}_N \quad = \quad \begin{cases} N(\sqrt{\frac{2}{3}}\hat{w}_0 + \sqrt{\frac{1}{3}}\hat{w}_1)^2 + o_p(1), & \text{when } (\hat{w}_0, \hat{w}_1) \in \text{ Cone I.} \\ N\hat{w}_0^2 + o_p(1), & \text{when } (\hat{w}_0, \hat{w}_1) \in \text{ Cone II.} \\ o_p(1), & \text{when } (\hat{w}_0, \hat{w}_1) \in \text{ Cone III.} \\ N(\hat{w}_0^2 + \hat{w}_1^2) + o_p(1), & \text{when } (\hat{w}_0, \hat{w}_1) \in \text{ Cone IV,} \end{cases} \tag{5}$$

and the expression in (4) can be simplified as:

$$P(\tilde{\Lambda}_N \geq x) = \frac{1}{2} P(\chi^2_1 \geq x) + \frac{\tan^{-1}(1/\sqrt{2})}{2\pi} P(\chi^2_2 \geq x), \quad \text{for all } x > 0. \tag{6}$$

In Figure 1, we see that the size of the angle of Cone IV is $\tan^{-1}(1/\sqrt{2})$. The weight of the χ^2_2 component is the portion of MLE $(\hat{w}_0, \hat{w}_1)$ that falls into the triangle region. Thus, the weight of the χ^2_2 component is $\frac{\tan^{-1}(1/\sqrt{2})}{2\pi}$. The total portion of $(\hat{w}_0, \hat{w}_1)$ that falls into the Cone I and Cone II is $1/2$, which gives the weight $1/2$ for χ^2_1. Note that the equation (6) does not include the portion where $\tilde{\Lambda}_N = 0$, that is the portion where $(\hat{w}_0, \hat{w}_1)$ falls in Cone III, which has probability 0.402.

In conclusion, the limiting distribution of $\tilde{\lambda}_N$ is given by (6).

2.2 Type I error and power.

We have used simulation to examine the precision of the limiting distribution (6) in approximating the sample distribution of $\tilde{\Lambda}_N$. We generated 10,000 sets of the IBD sample with sample size $N = 100$ under the null distribution. For each sample, we computed $\tilde{\pi}$'s and the values of the corresponding $\tilde{\Lambda}_N$. The sample quantiles of $\tilde{\Lambda}_N$ and the type I errors are computed based on

Table 1: The Null Rejection Rates based on equation (6).

Significance level	1%	5%	10%	25%	50%
Rejection rates of $\tilde{\Lambda}_N$	0.012	0.049	0.101	0.263	0.482
SSE ($\times 10^3$)	1.075	2.154	3.013	4.397	4.997

Table 2: The critical values of LRT with constraint and without constraint.

Size of Test	With constraint	Without constraint
0.1	2.227	4.605
0.05	3.417	5.991
0.01	6.343	9.210
0.005	7.641	10.597
0.001	10.702	13.816
0.0001	15.148	18.421

10,000 repetitions. These values together with the corresponding simulation based standard errors (SSE) are reported in Table 1. It is seen that each empirical null rejection rate approximates each significance level very well.

Since $\tilde{\Lambda}_N$ is the maximum LRT statistic subject to the possible triangle constraint, it is always less than or equal to Λ_N. For any given size of test, the threshold determined by our asymptotic mixture distribution will be smaller than the threshold determined by the χ^2_2 distribution. See Table 2. When the alternative hypothesis is true, the value of $\tilde{\Lambda}_N$ tends to be close to the value of Λ_N. The null hypothesis is hence more likely to be rejected using the constrained likelihood ratio test due to the smaller threshold value. Thus, the test with the constrained likelihood ratio statistic is expected to be more powerful.

The next simulation serves the purpose of illustrating the power gain with the constrained likelihood ratio test. Let (π_0, π_1) be equal to $(0.2, 0.42)$. We generated 10,000 sets of IBD sample with $N = 100$ and 200 under this distribution. For each simulation, we computed the log-likelihood ratio statistics Λ_N without constraint and $\tilde{\Lambda}_N$ with constraint. The powers at a number of significance levels are computed based on the outcomes of 10,000 sets of sample. The powers of the test with possible triangle constraint (restricted) and without possible triangle constraint (unrestricted) are summarized in Table 3. The simulation result reveals the expected power gain by the constrained likelihood ratio test, which is particularly striking when $N = 100$.

Table 3: The power of LRT of restricted and unrestricted model.

	N=100		N=200	
Size of Test	Restricted	Unrestricted	Restricted	Unrestricted
0.05	0.860	0.680	0.988	0.951
0.01	0.669	0.486	0.949	0.859
0.005	0.576	0.399	0.912	0.816
0.001	0.396	0.217	0.816	0.680
0.0001	0.190	0.098	0.621	0.449

3. Approximating the sample distribution of the supremum of LRT Statistics.

With the advance of the modern technology, IBD data of ASPs are usually collected at a large number of markers over stretches of chromosomes. The IBD values are related to each other through a crossover process along the chromosomes. Thus, $\tilde{\Lambda}_N$ values computed at these markers form a stochastic process indexed by the location of these marker on the chromosome. That is, we may write it as $\{\tilde{\Lambda}_N(t), 0 \le t \le T\}$ with t being the marker locus in cM and T being the total genetic length of a chromosome segment under investigation. If a chromosome segment under investigation contains a disease gene, an unusually large value of $\tilde{\Lambda}_N(t)$ is expected at t near the disease locus. Thus, the maximum value of $\tilde{\Lambda}_N(t)$ over $t \in [0, T]$ is a natural test statistic for linkage. The next problem is to determine the sample distribution of the supremum of $\tilde{\Lambda}_N(t)$ under the null hypothesis of no linkage. Given the threshold value, say x, the type I error of the test for each individual chromosome is

$$P_0\{\max_{0\le t\le T} \tilde{\Lambda}_N(t) \ge x\},$$

where the subscript 0 of the probability represents the probability under the null hypothesis.

At each locus t, $\tilde{\Lambda}_N(t)$ is determined by the value of unrestricted process $\hat{w}_0(t)$ and $\hat{w}_1(t)$ defined in (5). The limiting distribution of $\tilde{\Lambda}_N(t)$ is a mixture of χ^2_1 and χ^2_2 and a portion of being 0. It is noted that, for large N, $\sqrt{N}\hat{w}_0(t)$ and $\sqrt{N}\hat{w}_1(t)$ are asymptotically two independent stationary Gaussian processes. Thus, the process of $\tilde{\Lambda}_N$ is built up from the underlying stationary Gaussian Processes. The limiting distribution of $\max_{0\le t\le T} \tilde{\Lambda}_N(t)$ is difficult to obtain theoretically. Thus, we do not have a limiting distribution to be used to compute approximate sample quantiles. Instead, we use simulation to find a simple method to approximate the sample distribution.

We first placed markers at every 1cM grid along a hypothetical chromosome of some length. A crossover process was then simulated as a pure Poisson process according to the Haldane's mapping function (Haldane, 1919). Marker data were consequently determined along the chromosome. We set the the overall chromosome length $T = 150$cM as the longest mean genetic length (mean of male and female) of human chromosome is 150cM. The IBD data at each marker for each sib pair were then obtained with markers assumed fully informative. We set the number of sib pairs at 500 and generated 10,000 sets of samples. In applications, various lengths of chromosome segments might be investigated. We first aim at providing precise approximate sample quantiles of $\max_{0\le t\le T} \tilde{\Lambda}_N(t)$ for $T = 10$cM, 20cM, ..., 150cM. The sample quantile approximation will be discussed later for T values in between. For convenience, we denote $X(T) = \max_{0\le t\le T} \tilde{\Lambda}_N(t)$.

Let $X_i(T)$ be the observed $X(T)$ based on the ith simulated sample. For each given value of T, we may write the probability density function of $X(T)$ as:

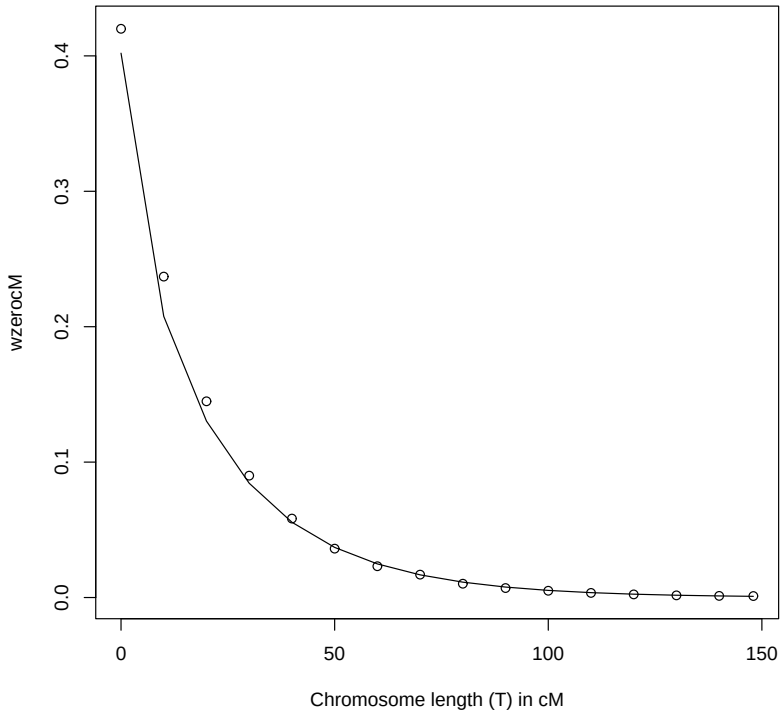
$$g_T(x) = w f_T(x) I_{(x>0)} + (1 - w) I_{(x=0)}, \tag{7}$$

where w is the probability that $X_i(T)$ is greater than 0. As the interval length T increases, w increases. The empty dots in Figure 3 show the empirical proportions of zeroes as a function of T.

Using the simulated 10,000 values of $X_i(T)$ at $T = 10$cM and so on, the corresponding quantiles of the distribution of $X(T)$ can be approximated by the sample quantiles. In applications, we may provide a table of simulated quantiles for a grid of values of T and a number of pre-chosen significance levels. This is not very convenient whenever a new significance level is required. A more useful solution is to find a simple family of parametric distributions that provide accurate approximations to the distributions of $X(T)$ with parameter values being smooth functions of T. Thus, for any value of T in an application and the observed value of $X(T)$, one can use this distribution family to compute an approximate p-value conveniently.

Recall that when $T = 0$, the marginal limiting distribution of the LRT statistic is a mixture of χ^2_1 and χ^2_2 distributions, and a distribution concentrated at 0. Based on simulation results at the

Figure 3: Proportions of zeroes for different size of chromosome interval (T).



selected T values, we propose to use the following distribution family:

$$f(x; p, d_1, d_2, T) = pf(x; d_1, T) + (1-p)f(x; d_2, T),$$

with $f(x; d, T)$ being χ^2_d distributions, and p being the mixing proportion, to approximate the sample distribution of the non-zero portion of $X(T)$. When d is not an integer, we interpret $f(x; d, T)$ as Gamma distribution with $(d/2, 2)$ degrees of freedom. Our task now is to find smooth functions of T for parameters p, d_1 and d_2. Based on the simulated data, with the consideration of simplicity, we propose the following function to be used:

$$\begin{array}{rcl} \text{logit}(p) & = & a_1 + a_2\sqrt{T} + a_3 T, \\ d_1 & = & b_1 + b_2 T, \\ d_2 & = & c_1 + c_2 T, \end{array} \tag{8}$$

where $\text{logit}(p) = \log(p/(1-p))$. The values of $\boldsymbol{a} = (a_1, a_2, a_3), \boldsymbol{b} = (b_1, b_2)$ and $\boldsymbol{c} = (c_1, c_2)$ will be chosen to best fit the simulated sample distributions of $X(T)$. The probability of $w = 1 - P\{X(T) = 0\}$ will be modeled later.

We now regard the simulated nonzero values of $X_i(T)$, $i = 1, 2, \ldots, n = 10,000$ and $T = 0, 20, \ldots, 150$ as a random sample from density $f(x; p, d_1, d_2, T)$ and consider estimating $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$ by maximum likelihood. At each given T, the log-likelihood function is given by:

$$l(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}; T) = \sum_{i=1}^{n} \log\{pf(x_i; d_1, T) + (1-p)f(x_i; d_2, T)\}.$$

Note that p, d_1 and d_2 are functions of $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ for each given T. Due to the very complex relationship between $X(T_1)$ and $X(T_2)$ for any $T_1 \neq T_2$, we maximize, instead of the joint log-likelihood function, the following "pseudo" log-likelihood:

$$l(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \sum_{T} l(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}; T)$$

with summation over $T = 0, 10, \cdots, 150$cM. Since when $T = 0$, the limiting distribution of $X(0)$ is known to be $p = 0.836$, $d_1 = 1$ and $d_2 = 2$, our maximization will be done under the constraints: $a_1 = 1.629$, $b_1 = 1$ and $c_2 = 2$.

We now use the EM algorithm (Dempster, Laird and Rubin, 1977) for the numerical computation. Let Z_i be a latent variable representing the mixture component of the ith observation. The complete log-likelihood function is:

$$\begin{array}{rcl} l_c(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) & = & \displaystyle\sum_{i=1}^{n} z_i \sum_{T} \log f(x_i; d_1; T) + \sum_{i=1}^{n} (1 - z_i) \sum_{T} \log f(x_i; d_2; T) \\ & & \displaystyle + \{\sum_{i=1}^{n} z_i\} \log(p) + \{n - \sum_{i=1}^{n} z_i\} \log(1-p). \end{array}$$

Given initial values of parameters $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$, we compute the conditional expected values of Z_i in the above set up in the E-step. In the M-step, we update the parameter values $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ as the maximizer of the complete likelihood. The E-step and M-step are iterated until convergence.

The E-M algorithm converges at $\hat{a}_2 = -0.551$, $\hat{a}_3 = 0.012$, $\hat{b}_2 = 0.005$ and $\hat{c}_2 = 0.019$. With (8), we find a set of parameters of p, d_1 and d_2, for any given value of T. Naturally, the corresponding distribution may not be a good approximation for the distribution of $X(T)$ when T is much larger than 150cM.

We approximate the proportion of 0 observations $1 - w$ by a function of the form

$$1 - w(T) = 0.402 \exp\{\alpha_1\sqrt{T} + \alpha_2 T\}.$$

This function is motivated by the requirement of $w(0) = 0.598$ (the total weight of the two chisquare components for $T = 0$) and the general trend of the observed proportions. By fitting this model to the observed zero proportions, we get

$$w(T) = 1 - 0.402 \exp\{-0.104\sqrt{T} - 0.033T\}.$$

The solid line in Figure 3 shows the fitted values. Here, we fixed the coefficient of the exponential term to be 0.402, which is the portion of zeroes when $T = 0$ given by our derived marginal limiting distribution of the LRT statistic (6). Figure 3 shows that this simple model gives a very good fit.

In summary, we use the following probability function to approximate the distribution of the supremum statistic for any given T: for any $x \geq 0$,

$$\mathrm{P}\{X(T) < x\} = w[p\mathrm{P}(\chi^2_{d_1} < x) + (1-p)\mathrm{P}(\chi^2_{d_2} < x)] + (1-w), \tag{9}$$

with

$$\begin{array}{rcl} w(T) & = & 1 - 0.402 \exp\{-0.104\sqrt{T} - 0.033T\}; \\ p(T) & = & \exp\{1.629 - 0.551\sqrt{T} + 0.012T\}/[1 + \exp\{1.629 - 0.551\sqrt{T} + 0.012T\}]; \\ d_1(T) & = & 1 + 0.005T; \\ d_2(T) & = & 2 + 0.019T. \end{array}$$

For example, when $T = 10$, we have $w = 0.792$, $p = 0.501$, $d_1 = 1.05$ and $d_2 = 2.19$.

To assess the precision of the approximations, we use Q-Q plots to compare the empirical quantiles with the quantiles from our mixed χ^2 for each size of T. In Figure 4, we give the Q-Q plots for T =20cM, 40cM, 60cM, 80cM, 100cM and 120cM. It is seen that our mixed χ^2 distribution approximates the distribution of the supremum statistics very well for various values of T. Our approximation provides an approximation conveniently for any values of T. To see this, we generated 10,000 sets of sample. For each sample, we took the supremum statistics for T = 15cM, 35cM, 55cM, 75cM, 95cM, 115cM and 135cM. Values of w's, p's, d_1's and d_2's for each T are computed by the equation (9). Figure 5 contains the Q-Q plots between the empirical quantiles and the quantiles from the mixture of χ^2 distributions for each T. It is seen that the quantiles computed from our mixture of χ^2 distributions match the quantiles of the test data very well. Moreover, it can be shown from the simulation results that the distribution approximations give p-values which are accurate when they are near 5%, regardless of the values of T.

In summary, our mixture χ^2 distribution of the form

$$\mathrm{P}\{X(T) < x\} = w[p\mathrm{P}(\chi^2_{d_1} < x) + (1-p)\mathrm{P}(\chi^2_{d_2} < x)] + (1-w),$$

is an adaptable approximation to the limiting distribution of the supremum statistics $\max_{0 \leq t \leq T} \tilde{\Lambda}_N(t)$. With any given size of T within our range of simulation, the parameters p, d_1 and d_2 can be obtained such that the threshold value for a given size of test can be easily determined by our probability function.

4. Discussions

In this paper, the limiting distribution of the LRT statistic at any given marker under the possible triangle constraint is obtained via a technique similar to that used in Chernoff (1954) and Self & Liang (1987) for deriving the limiting distribution of LRT statistics under nonstandard conditions. Our approach is straightforward and simple and consistent with Holmans' result. When it is viewed as a function of the position over a region of chromosome under investigation, the LRT statistic becomes a stochastic process. The inference is often made based on its supremum. The problem of finding the exact limiting distribution of the supremum of the LRT statistic is very interesting, challenging and difficult as well. We propose using a convenient χ^2-mixture distribution

Figure 4: Q-Q plots of approximating mixed χ^2 distributions and empirical distributions of supremum LRT statistics ($\max_{0\leq t\leq T}\tilde{\lambda}_N$) for T =20cM, 40cM, 60cM, 80cM, 100cM and 120cM.

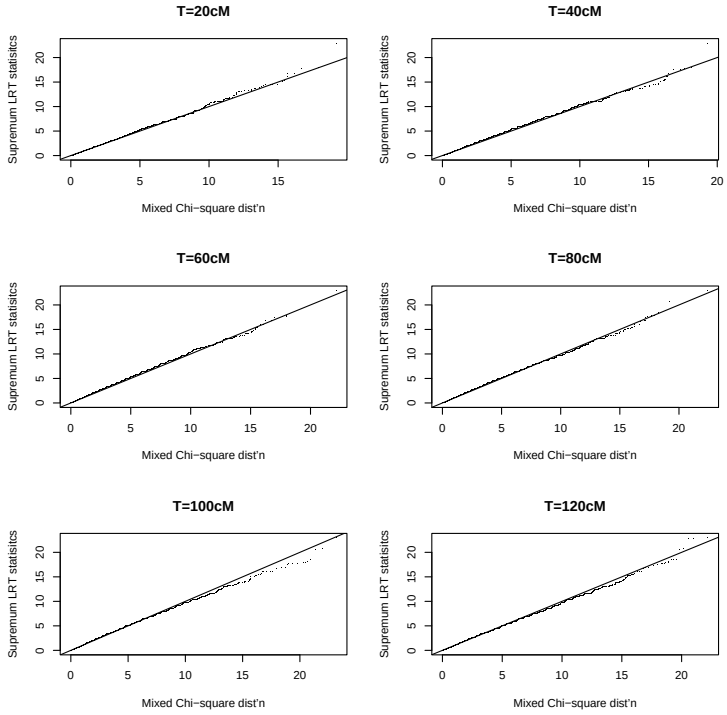
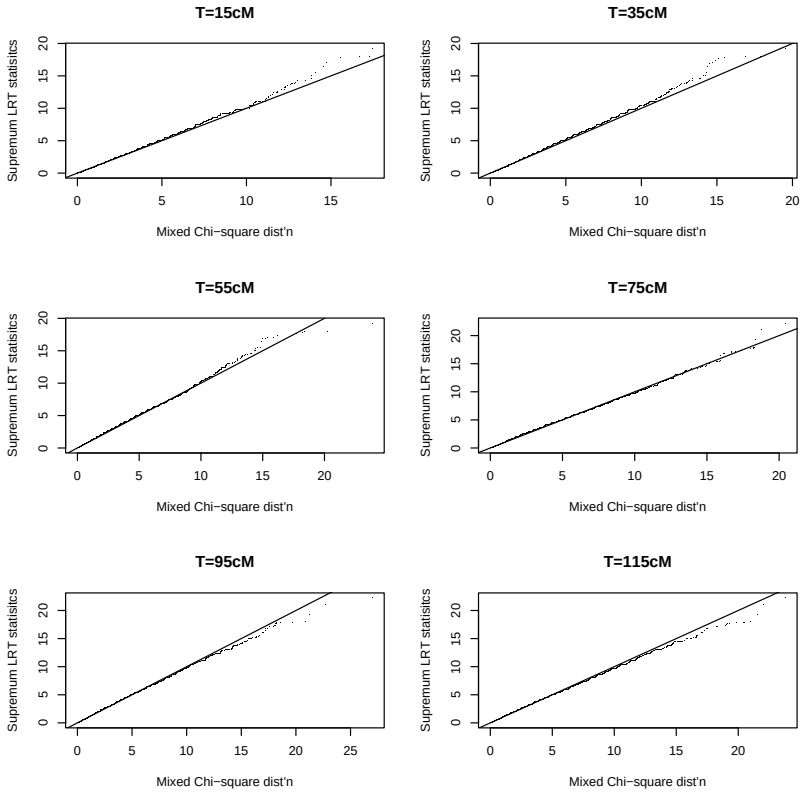


Figure 5: Q-Q plots of approximating mixed χ^2 distributions and the distributions of the new simulated supremum LRT statistics ($\max_{0\leq t\leq T}\tilde{\lambda}_N$) for T =15, 35, 55, 75, 95 and 115cM.



to approximate this supremum statistic under the possible triangle constraint. The simulation result shows the good approximation for a wide range of sizes of chromosome range. We also should notice that the weight of 0 and the weights and the degrees of freedom of each χ^2 mixture may depend on the density of the marker as well. We also conducted simulation study to investigate the effect of the marker density on parameters $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ and the proportion of zeroes. The results show that, if we apply the upper 5% threshold values determined by the fitted model based on a density of 1 marker/cM to the data with a density of 2 markers/cM, the empirical p-values are close to 0.05 for various lengths of T. This indicates that our approximate distributions are not very sensitive to the density of marker as doubling the marker density decreases the accuracy of the p-values only slightly. In the case of Gaussian processes, an upper bound for the tail probabilities has been developed and found to approximate the true value very well. See Feingold, Brown & Siegmund (1993). The techniques employed in those contexts might also be useful for our restricted LRT statistic. It is of interest to see if this problem can be solved.

ACKNOWLEDGMENTS

This work was partially supported by Natural Science and Engineering Research Council of Canada (NSERC).

REFERENCES

H. Chernoff (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25, 573–578.

A. P. Dempster, N. M. Laird & D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39, 1–38.

P. Holmans (1993). Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics*, 52, 362–374.

M. Farrall (1997). Affected sib pair linkage tests for multiple linked susceptibility genes. *Genetic Epidemiology,* 14, 103–115.

E. Feingold, P. O. Brown and D. Siegmund (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *American Journal of Human Genetics*, 53, 234–251.

Z. Feng, J. Chen & M. E. Thompson (2005). The universal validity of the possible triangle constraint for affected sib pairs. *The Canadian Journal of Statistics*, Vol 33, 2, 297–310.

J. B. S. Haldane (1919). The combination of the linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics*, 8, 299–309.

K. Liang, C. Huang & T. H. Beaty (2000). A unified sampling approach for multipoint analysis of qualitative and quantitative traits in sib pairs. *American Journal of Human Genetics*, 66, 1631–1641.

E. J. Louis, H. Payami & G. Thomson (1987). The affected sib pair method. V: Testing the assumptions. *American Journal of Human Genetics*, 51, 75–92.

L. S. Penrose (1935). Genetic data analysis of affected sib pairs. *Annals of Eugenics*, 6, 133–138.

S. G. Self & K. Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 604–610.

B. K. Suarez, J. Rice & T. Reish (1978). The generalized sib pair IBD distribution: its use in the detection of linkage. *Annals of Human Genetics*, 42, 87–94.

S. S. Wilks (1938). The large sample distribution of the likelihood ratio for testing composition hypotheses. *Annals of Mathematical Statistics*, 9, 60–62.

Received ???
Accepted ???

Zeny Z. FENG: zeny.feng@yale.edu
Department of Epidemiology and Public Health, Yale University
New Haven, Conneticut, US and 06511

Jiahua CHEN: jhchen@uwaterloo.ca
Mary E. THOMPSON: methomps@uwaterloo.ca
Department of Statistics and Actuarial Science, University of Waterloo
Waterloo, Ontario, Canada and N2L 3G1