Statistica Sinica **16**(2006), 805-823

TESTING FOR HOMOGENEITY IN GENETIC LINKAGE ANALYSIS

Yuejiao Fu, Jiahua Chen and John D. Kalbfleisch

York University, University of Waterloo and University of Michigan

Abstract: We apply the modified likelihood ratio test to two binomial mixture models arising in genetic linkage analysis. The limiting distribution of the test statistic for both models is shown to be a mixture of chi-squared distributions. A consideration of random family sizes for both models gives similar results. We also explore the power properties under local alternatives. Simulation studies show that the modified likelihood ratio test is more powerful than other methods under a variety of model specifications.

Key words and phrases: Genetic linkage analysis, hypothesis testing, local asymptotic power, mixture models, modified likelihood, random family sizes, recombinant.

1. Introduction

One goal of human genetic linkage analysis is to locate the gene or genes that are responsible for a given disease. This is done by linking the disease gene to genetic markers. When the loci of two genes are close to each other on the same chromosome, the two corresponding maternal (or paternal) alleles tend to stay in the same gamete after meiosis. The closer they are, the smaller the chance of being separated. A gamete with only one of the two maternal (or paternal) alleles at these two loci is called recombinant, and the recombination fraction, denoted as θ, is a useful measure of distance. When θ is close to 0, the two loci are tightly linked and located close to each other on the same chromosome. When two loci are not linked, the recombination fraction takes the maximum possible value, $\theta = 0.5$. Recombination data can be obtained by collecting pedigree information.

A family of three generations with many siblings can provide good information about the recombination fraction between two loci of interest. Nevertheless, except for some animal populations, it is rarely possible to find a sufficiently large and informative pedigree from which to obtain a precise enough estimate of the recombination fraction. Instead, geneticists collect recombination information from a large number of human families that exhibit the specific disease under investigation. In simple situations, for example, the disease has a single genetic cause and the recombination fraction between the disease locus and the marker locus is the same across the entire population; in this case, the number of recombinants in n independent meioses has a binomial distribution. In the presence of genetic heterogeneity (Smith (1963)), however, the disease in some families may be caused by different disease genes at different loci. In this case, the number of recombinants among siblings has a binomial mixture distribution. In more complicated situations, we may not be able to determine which of the two groups of siblings are recombinant. Additional binomial components have to be introduced to the mixture model.

In this paper, we consider a testing problem that is often encountered in linkage analysis, see Shoukri and Lathrop (1993), Lemdani and Pons (1995), Chiano and Yates (1995), Chen (1998), Liang and Rathouz (1999), and Abreu, Hodge and Greenberg (2002). As is well known, the likelihood ratio statistic to test the order of a mixture model does not have the usual asymptotic chi-squared distribution. In particular, for a class of binomial mixture models, Chernoff and Lander (1995) show that the limiting distribution of the likelihood ratio statistic is that of the supremum of a Gaussian process. The null limiting distribution, however, depends on the binomial parameters. Davies (1977, 1987) applies an upper bound to the null tail probabilities of the supremum of the asymptotic Gaussian processes for the purpose of statistical inference. His result has been used in the linkage analysis context, but the asymptotic upper bound depends on family structure and is relatively complicated.

A number of other strategies have also been proposed. Chen and Cheng (1995) Lemdani and Pons (1995) suggest restricting the range of the mixing proportions. This approach restores the chi-squared limiting distribution for many simple mixture models but requires specification of an arbitrary threshold. Liang and Rathouz (1999) define a score function which is sensitive toward a given alternative. This method also has nice mathematical and statistical properties, though choice of the alternative is somewhat arbitrary.

In this paper, we propose the modified likelihood approach discussed in Chen (1998), Chen, Chen and Kalbfleisch (2001, 2004) and Chen and Kalbfleisch (2005). The limiting distribution of the modified likelihood ratio statistic is chi-squared or a mixture of chi-squared distributions for a large variety of mixture models, and simulations suggest that this approach performs well while avoiding many of the drawbacks of other approaches. This gives a natural and quite general approach to testing problems in mixture models.

We consider two types of binomial mixture models, as in Liang and Rathouz (1999), and show that the modified likelihood ratio statistic has a mixture of chi-squared distributions in both cases. This method has better power in detecting linkages than other methods as demonstrated in our theoretical investigation and simulation experiments.

2. Binomial Mixture Models in Pedigree Studies

In human pedigree studies, it is sometimes not possible to ascertain whether or not a child is recombinant. One crucial piece of information is the genotype of the parent who carries the disease gene. We restrict attention to the autosomal genes, so that each individual has two alleles at the disease locus and two alleles at the A marker locus. The parental genotype is called phase known (PK) if, for the parent carrying the disease, it is known which allele of the marker locus shares the same chromosome as the disease allele. This information can sometimes be inferred from the genetic information of the grandparents. If it is not possible to determine this genotyping for the parent, the parental genotype is said to be phase unknown (PU).

Depending on the availability of the phase information, we have two kinds of mixture models.

2.1. Phase known case

Suppose that the linkage between a disease locus and a marker locus is under investigation and that the disease is autosomal dominant with full penetrance. Figure 1 shows an example of pedigree with phase known. In this example, marker A and the disease status of each individual in the three generations are shown. It is seen that the mother inherited both the disease allele D and marker allele A_1 from the grandmother. The disease allele D and A_1 are on the same chromosome and the phase of the mother is known. Hence the mating is of the form $dA_3|dA_4 \times DA_1|dA_2$. Based on this information, the first two offspring are non-recombinant and the third offspring is recombinant. The third child received the disease allele D but not the marker allele A_1.

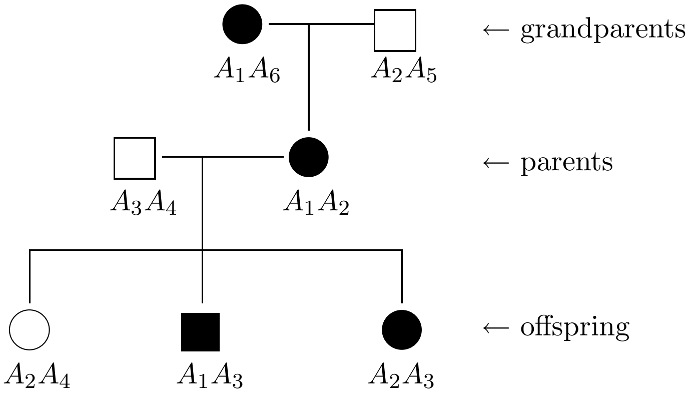

Figure 1. An example of the phase known pedigree: Circles represent females and squares represent males. Solid symbols indicate affected individuals.

When there is a single disease-causing gene and the phase of the parent is known, the number of recombinants among m offspring is observable and has a simple binomial distribution $Bi(m, \theta)$.

In many situations, the disease is complex and can be caused by different disease loci on different chromosomes. Suppose that only for a proportion α of families is the disease locus linked with the marker under consideration. The number of recombinants, Y, from a family with m offspring then has a mixture distribution with two components,

$$\alpha Bi(m, \theta) + (1 - \alpha) Bi(m, 0.5). \tag{1}$$

The corresponding mixing distribution is

$$G(t) = \alpha I(t \geq \theta) + (1 - \alpha) I(t \geq 0.5), \tag{2}$$

where I is an indicator function. The probability function of Y is

$$f_K(y; G) = \binom{m}{y} \left\{ \alpha \theta^y (1 - \theta)^{m-y} + (1 - \alpha)(0.5)^m \right\}.$$

2.2. Phase unknown case

When the phase of the parent is unknown, it is not possible to determine whether an offspring is recombinant. However, it may still be possible to divide the offspring into two groups, one recombinant and one non-recombinant. But, in this case, which group is which is unknown.

Figure 2 shows the same pedigree as Figure 1 except that the grandparental genotypes are unknown. For the mother, the disease allele D could be on the chromosome with either marker A_1 or A_2 and, under linkage equilibrium, these two possibilities would be equally likely. If the disease allele D and marker allele A_1 are on the same chromosome, the first two offspring are non-recombinant and the third is recombinant. In the other case, the first two offspring are recombinant and the third is non-recombinant.

Let Y be the number of offspring in the group identified as recombinant when D and A_1 are assumed to be on the same maternal chromosome. Then Y has a binomial mixture distribution $Bi(m, \theta)/2 + Bi(m, 1 - \theta)/2$. If only a proportion α of families in the population with the disease has the gene linked to this marker, then the distribution of Y is

$$\alpha\{\frac{1}{2} Bi(m, \theta) + \frac{1}{2} Bi(m, 1 - \theta)\} + (1 - \alpha) Bi(m, 0.5). \tag{3}$$

The corresponding mixing distribution and probability function are respectively

$$G(t) = \alpha\{\frac{1}{2} I(t \geq \theta) + \frac{1}{2} I(t \geq 1 - \theta)\} + (1 - \alpha) I(t \geq 0.5) \tag{4}$$

$$f_U(y;G) = \binom{m}{y}[\alpha\{\frac{1}{2}\theta^y(1-\theta)^{m-y} + \frac{1}{2}\theta^{m-y}(1-\theta)^y\} + (1-\alpha)(0.5)^m].$$

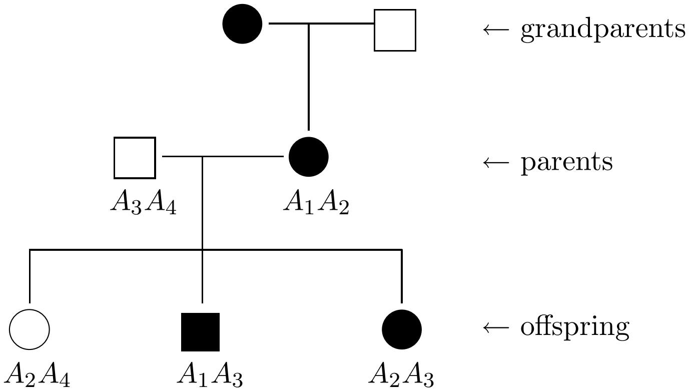

Figure 2. An example of phase unknown pedigree: Circles represent females and squares represent males. Solid symbols indicate affected individuals.

3. Linkage Analysis

We consider testing the hypothesis of no linkage, $H_0 : \alpha = 0$ or $\theta = 0.5$, against the alternative, $H_1 : 0 < \alpha \leq 1$ and $0 \leq \theta < 0.5$.

3.1. Existing methods

For both PK and PU cases, the likelihood function based on independent family data can be easily constructed. However, the likelihood ratio statistic does not have the usual chi-squared limiting distribution. For the PK case with $m \geq 3$, Chernoff and Lander (1995) show that the limiting distribution involves the supremum of a Gaussian process. The conclusion can also be obtained from a general result in Dacunha-Castelle and Gassiat (1999) In order to use this result for the purpose of inference, we need to calculate quantiles of the supremum of the Gaussian process. This could be done by simulations; however, the general analytical result is still unknown (Adler (1990)).

The cause of this non-regular behavior is the loss of identifiability under the null hypothesis; many alternative approaches have been suggested to resolve this non-regularity. In independent papers, Chen and Cheng (1995) and Lemdani and Pons (1995) suggest imposing a restriction $\alpha \geq \gamma$ for some given $\gamma > 0$. Liang and Rathouz (1999) discuss a procedure which initially fixes the parameter value $\alpha = 1$ in the alternative model and develop a score test.

Chen (1998), Chen, Chen and Kalbfleisch (2001, 2004) and Chen and Kalbfleisch (2005) discuss inference in mixture models and, in other contexts, suggest a modified likelihood function in which a penalty function is placed on small and/or large value of α so as to restore identifiability in the model. A resulting advantage is that the modified likelihood ratio test statistic often has a known and simple limiting distribution under the null hypothesis. That is, the limiting distribution does not depend on which particular null distribution is true. Our aim in this paper is to illustrate the modified likelihood approach in the PK and PU cases.

3.2. Modified likelihood method

Let $y_1, \ldots, y_n$ be independent observations, where the density function of y_i is $\alpha f_i(y;\theta) + (1-\alpha) f_i(y;0.5)$, $0 \le \alpha \le 1$ and $0 \le \theta \le 0.5$. The log-likelihood function is

$$l_n(\alpha,\theta) = \sum_{i=1}^{n} \log\{\alpha f_i(y_i;\theta) + (1-\alpha) f_i(y_i;0.5)\}.$$

We consider the modified log-likelihood function $pl_n(\alpha,\theta) = l_n(\alpha,\theta) + C\log(\alpha)$ for some chosen constant C. The main goal of the modified likelihood function is to discourage the fit with α close to 0 and so avoid the alternative representation $\alpha = 0$ of the null hypothesis $\theta = 0.5$. We often take $C = 1$, which has been found to be satisfactory for the data with multinomial component distributions with moderate significance levels, and it works reasonably well for all cases considered in this paper; see the simulation results in Section 4 and Chen (1998). The simulation results in Table 1, however, suggest that $C = 2$ is a better choice for the smaller significance level, 0.5%. For other mixture models, the appropriate choice of C depends on the size of the parameter space. For example, Chen, Chen and Kalbfleisch (2001) suggest $C = \log(M)$ when the parameter space of the component distribution is given by $[-M, M]$, see also Zhu and Zhang (2004). In general, the best choice of C depends on the model and the significance level of interest. Further investigation is needed.

For the PK case, $f_i(y;\theta)$ is $Bi(m,\theta)$, whereas for the PU case, $f_i(y;\theta)$ is $Bi(m,\theta)/2 + Bi(m,1-\theta)/2$, for all i.

The maximum modified log-likelihood estimators $\hat{\alpha}$ and $\hat{\theta}$ are the maximizers of the modified log-likelihood function $pl_n(\alpha,\theta)$. For the problem under consideration, the null model is completely specified as $\theta = 0.5$ with any value of α. If $\theta = 0.5$, $pl_n(\alpha,\theta)$ is maximized at $\alpha = 1$ and, for simplicity, we regard $\alpha = 1$ as the null value of α. The modified likelihood ratio statistic is defined as

$$R_n = 2[l_n(\hat{\alpha},\hat{\theta}) - l_n(1, 0.5)]. \tag{5}$$

In the modified likelihood ratio test (MLRT), the observed value of R_n is assessed against the null distribution or asymptotic distribution of R_n.

Table 1. Simulated null rejection rates (in %) of seven test statistics for detecting linkage: PK case.

		Nominal							
n	m	level%	$T^*_{0.5}$	T^*_1	MR(1)	MR(2)	MR(3)	$L_{0.5}$	$L_{0.0}$
50	2	10	10.16	9.81	10.00	9.82	9.80	10.38	13.71
		5	5.23	5.40	4.90	4.62	4.60	5.45	7.37
		1	1.33	1.17	1.33	1.14	1.10	1.12	1.51
		0.5	0.68	0.51	0.78	0.59	0.57	0.62	0.95
50	4	10	10.30	10.05	9.41	8.99	8.92	10.21	15.90
		5	5.52	5.29	5.90	5.40	5.32	5.86	8.26
		1	1.46	1.10	1.63	1.08	0.98	1.15	2.09
		0.5	0.77	0.59	0.75	0.55	0.44	0.58	0.89
50	8	10	10.18	9.98	11.05	10.59	10.53	11.05	25.5
		5	5.48	5.21	5.68	5.09	5.02	5.59	8.83
		1	1.39	1.08	1.53	1.07	0.97	1.13	2.97
		0.5	0.81	0.60	0.78	0.63	0.53	0.64	1.01
100	2	10	10.55	10.26	9.42	9.34	9.33	10.29	13.75
		5	5.19	5.33	5.53	5.42	5.41	5.63	7.17
		1	1.16	1.03	1.17	0.96	0.96	1.08	1.61
		0.5	0.63	0.53	0.63	0.45	0.45	0.55	0.80
100	4	10	10.42	10.30	11.13	10.85	10.82	11.03	17.28
		5	5.31	5.22	5.45	5.07	5.05	5.45	8.94
		1	1.18	1.12	1.45	1.04	1.00	1.10	1.87
		0.5	0.59	0.60	0.83	0.64	0.60	0.66	0.96
100	8	10	10.13	10.09	10.14	9.60	9.58	10.37	17.45
		5	5.45	5.28	5.65	5.06	5.03	5.51	11.47
		1	1.22	0.99	1.71	1.09	1.06	1.14	2.29
		0.5	0.64	0.54	1.01	0.56	0.52	0.59	1.45

3.3. Large sample properties of the MLRT

We present some asymptotic results for R_n in this section; proofs are given in the Appendix. We first consider the case when all the families have the same number of offspring, m.

Theorem 1. *Suppose $Y_1, \ldots, Y_n$ are independent and identically distributed random variables from either the PK case* (1) *or the PU case* (3). *For fixed $m \geq 1$ in the PK case, or $m \geq 2$ in the PU case, $R_n \xrightarrow{d} \chi_0^2/2 + \chi_1^2/2$ under the null hypothesis, $\theta = 0.5$.*

Note that χ_0^2 is the degenerate distribution at 0, and χ_1^2 is the chi-squared distribution with 1 degree of freedom.

In most applications, geneticists obtain families with various number of offspring. The modified likelihood method turns out to be just as convenient here. Let (M_i, Y_i) represent the data from the ith family, $i = 1, \ldots, n$. Here M_i denotes the number of offspring and Y_i is defined earlier in the PK and PU cases. Conditional on the M_i's, the log-likelihood function is

$$l_n(\alpha, \theta) = \sum_{i=1}^{n} \log P(Y_i = y_i | M_i). \tag{6}$$

The modified log-likelihood function is

$$pl_n(\alpha, \theta) = l_n(\alpha, \theta) + C \log(\alpha), \tag{7}$$

with the modified likelihood ratio statistic

$$R_n = 2[l_n(\hat{\alpha}, \hat{\theta}) - l_n(1, 0.5)]. \tag{8}$$

Suppose (M_i, Y_i), $i = 1, \ldots, n$ are independent and identically distributed, with all phase known, or all phase unknown. Assume also that M_i has an upper bound, then $R_n \xrightarrow{d} \chi_0^2/2 + \chi_1^2/2$ under the null hypothesis, $\theta = 0.5$.

The asymptotic results for the MLRT can also be extended to the situation where there are both PK and PU observations, and the MLRT can still be used in that situation, see Fu (2004). If the numbers of PK and PU observations are comparable, PK observations dominate the asymptotic expansions of the MLRT under the null hypothesis. The asymptotic results under the null hypothesis remain the same. Nonetheless, there can be large efficiency gains in finite observations through adding PU observations to PK observations. Additional discussion will be given in the section of simulation studies.

Theorem 1 provides a convenient tool to determine the critical values for the test. They do not tell us the power properties of the test which can be assessed by simulation. A theoretical approach which gives some insight is to assess the asymptotic power of the test against local alternatives.

Consider the asymptotic power based on observations from the PK case against the following sequence of local alternatives:

$$H_a^n : \alpha = \alpha_0, \quad \theta = 0.5 - n^{-\frac{1}{2}}\tau, \tag{9}$$

where $0 < \alpha_0 \leq 1$ and $\tau > 0$. Note that the value of θ approaches 0.5 as n increases at the rate of $n^{-1/2}$ in H_a^n. This choice of local alternatives is based on the knowledge that the convergence rate of $\hat{\theta}$ is $n^{-1/2}$ in the PK case.

The local alternative H_a^n is contiguous to the null distribution (see Le Cam and Yang (1990)). By Le Cam's contiguity theory, the limiting distribution of R_n under H_a^n can be determined by the null limiting joint distribution of (R_n, Λ_n), where $\Lambda_n = l_n(\alpha_0, 0.5 - n^{-1/2}\tau) - l_n(1, 0.5)$, which is the log-likelihood ratio evaluated at H_a^n. The technical details of deriving the limiting distribution under H_a^n are given in Appendix B. First, we state the result as follows.

Theorem 2. *In the PK case, the limiting distribution of R_n under H_a^n is $\{(Z + \sigma_{12})^+\}^2$, where Z has the standard normal distribution and $\sigma_{12} = 2\tau\alpha_0\sqrt{E(M_1)}$.*

Suppose we have n' out of n observations from families with more than one offspring. If we only use these n' observations to test the local alternative H_a^n, the limiting distribution of R_n becomes $\{(Z + \sigma'_{12})^+\}^2$, where

$$\frac{\sigma'^2_{12}}{\sigma^2_{12}} = \frac{E[M_1 I(M_1 > 1)]}{E(M_1)} < 1.$$

For example, suppose M_1 has the probability function shown below.

m	1	2	3	4	5
$P(M_1 = m)$	0.35	0.35	0.15	0.10	0.05

Note that $E(M_1) = 2.15$ and $E[M_1 I(M_1 > 1)] = 1.8$. This gives a relative efficiency of $\sigma'^2_{12}/\sigma^2_{12} = 83.7\%$. Although families of size one do not by themselves provide information on θ, they do provide information about $\theta\alpha + (1 - \alpha)/2$. Together with larger families, they still provide useful information for detecting linkage.

The problem with the PU case is different. It can be seen that the Fisher information at $\theta = 0.5$ degenerates. It follows that the rate of estimating θ is slower than $n^{-1/2}$ and, in fact, we find that $\hat{\theta}$ converges at rate $n^{-1/4}$. This explains our choice of the local alternatives,

$$H_a^n : \alpha = \alpha_0, \quad \theta = 0.5 - n^{-\frac{1}{4}}\tau, \tag{10}$$

where $0 < \alpha_0 \leq 1$ and $\tau > 0$. In the PU case, the limiting distribution of R_n under H_a^n is $\{(Z + \sigma_{12})^+\}^2$, where $\sigma_{12} = 2\tau^2\alpha_0\sqrt{E\{2M_1(M_1 - 1)\}}$. Here, families of size two but not of size one contribute to the inference.

We have also investigated the asymptotic power of the generalized score test discussed in Liang and Rathouz (1999). The test turns out to have the same asymptotic power as the MLRT under the same local alternatives. The proof is omitted here.

4. Simulation studies

We carry out simulation studies to compare the performance of the MLRT with other testing procedures. The method of Chen and Cheng (1995) and

Lemdani and Pons (1995) requires $\alpha \geq \gamma$ for some $\gamma > 0$. We denote this class of restricted likelihood ratio statistics by $\{L_\gamma, 0 < \gamma < 1\}$. We denote Liang and Rathouz's test statistic as $\{T^*_\lambda, 0 < \lambda \leq 1\}$. The constant λ is often set at 1, but other choices are possible. We use $\{\mathrm{MR}(C),\ C > 0\}$ for the MLRT statistics with the level of modification being C. As discussed earlier, $C = 1$ seems to work well in the context of binomial mixture models with moderate significance levels; increasing C makes the test more conservative.

Samples of size $n(= 50, 100)$ were generated from binomial mixture models (1) or (3) corresponding to the PK and PU situations, and in each case families of size $m(= 2, 4, 8)$ were considered. The parameter values of α and θ were chosen from the combinations of $\alpha(= 0, 0.1, 0.2, 0.5)$ and $\theta(= 0.1, 0.01)$. These combinations cover the null model as well as alternatives with moderate and strong linkages.

For each given sample size n and family size m, the true null rejection rates were estimated using 20,000 replications. The outcomes are presented in Tables 1 and 3. In all cases except L_0, the mixture of chi-squared distributions with zero and one degree of freedom is the null limiting distribution and hence was used to determine the critical values of the tests. It is obvious that the mixture of chi-squared distributions was not a good approximation to the usual likelihood ratio statistic, L_0. In all other cases, the approximations were reasonable.

We generated 10,000 samples from alternative models to evaluate the power of the different testing procedures. To make the power comparison meaningful and fair, the critical values were calculated from the 20,000 samples from the null models, rather than the chi-squared mixture limiting distribution. The results are presented in Tables 2 and 4. It is interesting to note that the simulated null rejection rates of all tests are almost the same when $m = 2$ in the PU case. The binomial mixture model is not identifiable in this case, so that the value of θ or α can be fixed in the tests. It follows that the restrictions in the methods of L_γ and T^*_λ have no effect. In fact, when $m = 3$, the model is also not identifiable in the PU case as shown by Abreu, Hodge and Greenberg (2002).

Overall the MLRT has comparable and higher power under various alternative models. For example, in Table 2, when $n = 50$, $m = 4$, $\theta = 0.01$ and $\alpha = 0.2$, MR(1) improves on competitors $T^*_{0.5}$ by 6%, T^*_1 by 23% and $L_{0.5}$ by 16%.

Table 5 shows the simulated null rejection rates and statistical power for MR(1) and T^*_1 in the PK case with sample size $n = 100$ and a selected family size distribution. All the results were based on 20,000 repetitions. In this case, the simulated null rejection rates approximate the nominal levels satisfactorily and, although the power of MR(1) is comparable to T^*_1 when $\theta = 0.2$, it has considerably larger power when θ is 0.1 or 0.05. The comparison between MR(1) and T^*_1 in the PU case with random family sizes was similar. The results are not presented here.

Table 2. Simulated power (in %) of seven test statistics for detecting linkage: PK case. Upper entry: $\alpha = 0.1$; middle entry: $\alpha = 0.2$; lower entry: $\alpha = 0.5$. The nominal level is 0.005.

n	m	θ	$T^*_{0.5}$	T^*_1	MR(1)	MR(2)	MR(3)	$L_{0.5}$	$L_{0.0}$
50	2	0.10	5	5	6	5	5	5	5
			22	19	23	21	20	21	22
			95	93	95	94	93	94	95
50	2	0.01	8	7	9	8	7	8	8
			38	32	37	34	32	34	37
			100	100	100	100	99	100	100
50	4	0.10	15	11	17	13	12	11	17
			60	46	59	51	48	49	61
			100	100	100	100	100	100	100
50	4	0.01	29	18	35	26	21	20	36
			82	65	88	81	74	72	89
			100	100	100	100	100	100	100
50	8	0.10	45	27	57	49	37	31	58
			92	78	95	92	88	84	95
			100	100	100	100	100	100	100
50	8	0.01	67	41	88	82	72	51	88
			98	90	100	100	100	96	100
			100	100	100	100	100	100	100
100	2	0.10	10	9	9	10	10	9	10
			45	41	43	43	42	42	46
			100	100	100	100	100	100	100
100	2	0.01	16	14	16	16	15	14	17
			69	61	68	65	62	63	71
			100	100	100	100	100	100	100
100	4	0.10	30	21	33	24	21	22	35
			87	76	89	82	78	79	91
			100	100	100	100	100	100	100
100	4	0.01	50	33	64	49	40	37	64
			97	91	99	98	96	94	99
			100	100	100	100	100	100	100
100	8	0.10	69	47	85	77	65	55	86
			99	96	100	100	100	98	100
			100	100	100	100	100	100	100
100	8	0.01	87	67	99	98	95	76	99
			100	99	100	100	100	100	100
			100	100	100	100	100	100	100

Table 3. Simulated null rejection rates (in%) of seven test statistics for detecting linkage: PU case.

		Nominal							
n	m	level%	$T^*_{0.5}$	T^*_1	MR(0.5)	MR(1)	MR(2)	$L_{0.2}$	$L_{0.0}$
50	2	10	10.10	10.10	10.12	10.12	10.12	10.12	10.12
		5	5.90	5.90	5.89	5.89	5.89	5.89	5.89
		1	0.76	0.76	0.76	0.76	0.76	0.76	0.76
		0.5	0.37	0.37	0.37	0.37	0.37	0.37	0.37
50	4	10	10.26	10.10	9.77	9.64	9.64	10.36	11.48
		5	5.60	5.60	4.76	4.46	4.46	5.17	5.58
		1	1.40	1.30	0.99	0.87	0.86	1.04	1.04
		0.5	0.73	0.74	0.49	0.45	0.44	0.54	0.76
50	8	10	10.50	10.40	10.71	9.45	9.24	10.50	13.70
		5	5.60	5.60	6.19	4.72	4.46	5.41	9.30
		1	1.60	1.50	1.36	1.08	0.80	1.05	1.50
		0.5	0.93	0.79	0.94	0.69	0.39	0.52	1.10
100	2	10	9.54	9.54	9.54	9.54	9.54	9.54	9.54
		5	4.46	4.46	4.46	4.46	4.46	4.46	4.46
		1	1.05	1.05	1.05	1.05	1.05	1.05	1.05
		0.5	0.56	0.56	0.56	0.56	0.56	0.56	0.56
100	4	10	10.40	10.15	9.73	9.66	9.66	10.00	13.56
		5	5.17	5.46	4.84	4.73	4.72	5.26	5.74
		1	1.28	1.29	1.07	0.93	0.92	1.09	1.42
		0.5	0.74	0.73	0.56	0.49	0.47	0.59	0.74
100	8	10	10.01	10.02	9.97	9.27	9.21	9.95	13.45
		5	5.60	5.56	5.54	4.67	4.59	5.27	8.28
		1	1.46	1.42	1.59	1.12	0.99	1.17	1.76
		0.5	0.97	0.91	1.11	0.64	0.54	0.66	1.25

We further investigated the informativeness of the samples from families with only one offspring in the PK case, and families with two offspring in the PU case. The MR$(1)^+$ and MR$(1)^-$ denote the MLRT with or without these samples. Table 6 shows the result for a PU example. As expected, the power of the test was improved when including the samples from families with two offspring. The results of the PK cases were similar and therefore omitted.

In applications, the data may contain both PK and PU observations. The presence of PU observations does not change the null limiting distribution of the MLRT but improves the power of the test, sometimes substantially. As shown in Chen (1998), the Kullback-Leibler (KL) information is the determining factor of the testing power. To examine the contribution of the PU observations to the

Table 4. Simulated power (in%) of seven test statistics for detecting linkage: PU case. Upper entry: $\alpha = 0.1$; middle entry: $\alpha = 0.2$; lower entry: $\alpha = 0.5$. The nominal level is 0.005.

n	m	θ	$T^*_{0.5}$	T^*_1	MR(0.5)	MR(1)	MR(2)	$L_{0.2}$	$L_{0.0}$
50	2	0.10	2	2	2	2	2	3	3
			6	6	6	6	6	6	6
			45	45	45	45	45	45	45
50	2	0.01	4	4	4	4	4	4	4
			14	14	14	14	14	14	14
			87	86	87	87	87	87	87
50	4	0.10	8	8	8	8	8	9	9
			33	32	33	32	33	34	34
			98	98	98	98	98	98	98
50	4	0.01	19	18	19	18	18	19	19
			67	65	68	66	66	69	69
			100	100	100	100	100	100	100
50	8	0.10	40	34	48	42	34	42	46
			87	83	89	87	83	88	90
			100	100	100	100	100	100	100
50	8	0.01	70	61	85	82	69	75	82
			98	97	100	100	99	100	100
			100	100	100	100	100	100	100
100	2	0.10	3	3	3	3	3	3	3
			11	10	11	11	11	11	11
			76	75	77	76	76	76	76
100	2	0.01	6	6	6	6	6	6	6
			27	27	28	27	27	27	27
			99	99	100	100	99	99	99
100	4	0.10	16	15	16	15	15	16	16
			63	62	63	62	62	64	64
			100	100	100	100	100	100	100
100	4	0.01	38	36	39	38	36	40	40
			94	93	94	93	93	94	95
			100	100	100	100	100	100	100
100	8	0.10	64	60	72	67	60	67	72
			99	98	100	100	99	100	99
			100	100	100	100	100	100	100
100	8	0.01	91	87	98	97	93	94	98
			100	100	100	100	100	100	100
			100	100	100	100	100	100	100

Table 5. Simulated null rejection rates and statistical power for MR(1) and T_1^*, $\alpha = 0.2$; n=100: PK case.

Family size:	2	3	4	5	6
Probability:	0.35	0.35	0.15	0.10	0.05
		Nominal Levels			
θ	Method	0.10	0.05	0.01	0.005
0.5	MR(1)	0.102	0.053	0.013	0.006
	T_1^*	0.101	0.053	0.013	0.006
0.2	MR(1)	0.795	0.690	0.440	0.368
	T_1^*	0.797	0.685	0.425	0.353
0.1	MR(1)	0.933	0.893	0.762	0.706
	T_1^*	0.926	0.871	0.705	0.637
0.05	MR(1)	0.969	0.952	0.882	0.848
	T_1^*	0.960	0.924	0.804	0.755

Table 6. Simulated null rejection rates and statistical power for $\mathrm{MR}(1)^+$ and $\mathrm{MR}(1)^-$; $\alpha = 0.2$, n=100: PU case.

Family size:	2	3	4	5	6
Probability:	0.35	0.35	0.15	0.10	0.05
		Nominal Levels			
θ	Method	0.10	0.05	0.01	0.005
0.5	$\mathrm{MR}(1)^+$	0.093	0.044	0.009	0.004
	$\mathrm{MR}(1)^-$	0.090	0.046	0.008	0.004
0.2	$\mathrm{MR}(1)^+$	0.538	0.408	0.176	0.135
	$\mathrm{MR}(1)^-$	0.522	0.378	0.171	0.119
0.1	$\mathrm{MR}(1)^+$	0.840	0.751	0.506	0.435
	$\mathrm{MR}(1)^-$	0.811	0.705	0.478	0.390

power of the test, we could directly compute the KL information under various alternative models. We present some simulation results as follows.

For simplicity, we considered the situation of 100 PK and 100 PU families with equal family size 4. Let *pkpu*, *pk* and *pu* denote the tests using both PK and PU observations, only PK, and only PU observations, respectively. Table 7 shows the simulated null rejection rates and power of the three test statistics. The simulated null rejection rates are quite close to the nominal levels. Under the alternative that is relatively close to the null, that is, 20% of the families with recombination fraction $\theta = 0.4$, the power of *pkpu* and *pk* is comparable. However, under the distant alternative, when $(\alpha, \theta) = (0.2, 0.2)$, only using the

PK observations will result in a considerable loss of power. Another interesting observation is the effectiveness of the KL information in determining the power of the test. For example, the total KL information of 100 PK observations when $(\alpha, \theta) = (0.2, 0.3)$ is close to that of 100 PU observations when $(\alpha, \theta) = (0.2, 0.2)$. By inspection, the simulated power of the tests is almost the same.

Table 7. Simulated null rejection rates and statistical power for the tests *pkpu*, *pk* and *pu*.

				Nominal Levels			
α	θ	Method	Total KL	0.10	0.05	0.01	0.005
	0.5	*pkpu*	0	0.100	0.050	0.012	0.006
0.2	0.4	*pkpu*	0.351	0.325	0.204	0.064	0.034
		pk	0.332	0.322	0.196	0.063	0.033
		pu	0.019	0.133	0.067	0.014	0.009
0.2	0.3	*pkpu*	1.777	0.655	0.535	0.272	0.194
		pk	1.482	0.636	0.484	0.244	0.164
		pu	0.295	0.294	0.178	0.049	0.028
0.2	0.2	*pkpu*	5.373	0.909	0.863	0.696	0.628
		pk	3.937	0.866	0.792	0.573	0.481
		pu	1.436	0.639	0.495	0.227	0.159

5. Discussion

For the problems considered here, the modified likelihood approach has the advantage of giving a natural and quite general approach to testing problems in mixture models. As shown in this paper, the MLRT has a simple asymptotic null distribution in both PK and PU cases with fixed or random family sizes. In stating the results in this paper, we assumed that the distribution of the family size was bounded. As can be seen from the Appendix, however, this assumption can be relaxed and we need only assume that $E(s^{M_i})$ is finite for $|s| \leq 2$. The local power properties of the MLRT are identical to those for T^*_λ, but our simulation studies show that the MLRT outperforms other methods against more distant alternatives.

Typically, the modified likelihood can be maximized using an EM type algorithm. The penalty term can be thought to have arisen through an auxiliary experiment in which there are C Bernoulli trials with probability α of success, each trial resulting in a success. With this additional "data", the likelihood is an ordinary likelihood and the EM methods for imputing the missing mixture data can be applied directly. Further discussion of this can be found in

Chen, Chen and Kalbfleisch (2001). Alternatively, general maximization techniques can be implemented.

Acknowledgements

The authors thank the editor, an associate editor and two referees for their insightful comments. The research of Yuejiao Fu and Jiahua Chen was partly supported by the Natural Sciences and Engineering Research Council of Canada. The paper is based on part of Yuejiao Fu's Ph.D. dissertation written at University of Waterloo.

Appendix A. Proof of Theorem 1

To prove Theorem 1, we need the following lemma.

Lemma 1. *Suppose that when $n \to \infty$, $\sup_{\alpha,\theta}\{l_n(\alpha,\theta) - l_n(1, 0.5)\} = O_p(1)$. Then, for the maximum modified likelihood estimator $\hat{\alpha}$, $\log(\hat{\alpha}) = O_p(1)$.*

Proof. This follows since, otherwise, the modified log-likelihood at $\hat{\alpha}$ would diverge to negative infinity which contradicts the definition of $\hat{\alpha}$.

Proof of Theorem 1. It can be seen that the likelihood function in either model satisfies the condition of Lemma 1. Hence, $\log(\hat{\alpha}) = O_p(1)$. Consequently, for any $\epsilon > 0$, there exists $\delta > 0$ such that $P(\hat{\alpha} > \delta) > 1 - \epsilon$ for all n large enough. As a consequence, we can assume $\hat{\alpha} > \delta > 0$ for the purpose of asymptotic derivation. It then follows that $\hat{\theta} \to 0.5$ in probability under the null hypothesis in either model.

Denote $\eta = 2(0.5 - \theta)$. In the PK case,

$$\log f_K(y; \alpha, \theta) - \log f_K(y; 1, 0.5) = \log\{1 + \alpha[(1-\eta)^y(1+\eta)^{m-y} - 1]\}.$$

For all θ close to 0.5, the Taylor expansion about $\eta = 0$ gives

$$\begin{aligned}R_n &= 2[l_n(\hat{\alpha},\hat{\theta}) - l_n(1, 0.5)]\\ &= 2\hat{\alpha}\hat{\eta}\sum_{i=1}^{n}(m - 2y_i) + \hat{\alpha}\hat{\eta}^2\sum_{i=1}^{n}(m^2 - 4my_i - m + 4y_i^2)\\ &\quad -(\hat{\alpha}\hat{\eta})^2\sum_{i=1}^{n}(m - 2y_i)^2 + o_p(n\hat{\eta}^2)\\ &= 2\hat{\alpha}\hat{\eta}\sum_{i=1}^{n}(m - 2y_i) - (\hat{\alpha}\hat{\eta})^2\sum_{i=1}^{n}(m - 2y_i)^2 + o_p(n\hat{\eta}^2).\end{aligned}$$

To get the last equality, note that $E(m^2 - 4mY - m + 4Y^2) = 0$, which implies $\hat{\alpha}\hat{\eta}^2\sum_{i=1}^{n}(m^2 - 4my_i - m + 4y_i^2) = o_p(n\hat{\eta}^2)$. Because $\hat{\alpha}$ and $\hat{\eta}$ maximize the modified log-likelihood with $0 \leq \eta \leq 1$, we must have $\hat{\alpha} = 1 + o_p(1)$ and $\hat{\eta} = \{[\sum_{i=1}^n (m-2y_i)]^+ / \sum_{i=1}^n (m-2y_i)^2\} + o_p(n^{-1/2})$, where $[\cdot]^+$ denotes the positive part of the argument. Thus, the modified likelihood ratio statistic

$$R_n = \frac{\{[\sum_{i=1}^n (m-2y_i)]^+\}^2}{\sum_{i=1}^n (m-2y_i)^2} + o_p(1)$$

has the claimed limiting distribution.

In the PU case, we have

$$\begin{aligned} &\log f_U(y;\alpha,\theta) - \log f_U(y;1,0.5) \\ &= \log[1 + \alpha\{\frac{1}{2}(1-\eta)^y(1+\eta)^{m-y} + \frac{1}{2}(1-\eta)^{m-y}(1+\eta)^y - 1\}]. \end{aligned}$$

Note that this function is symmetric in η and its expansion contains only even terms. The Taylor expansion at $\eta = 0$ is

$$\begin{aligned} R_n &= 2[l_n(\hat{\alpha},\hat{\theta}) - l_n(1,0.5)] \\ &= \hat{\alpha}\hat{\eta}^2 \sum_{i=1}^n (m^2 - 4my_i + 4y_i^2 - m) \\ &\quad -\frac{1}{4}(\hat{\alpha}\hat{\eta}^2)^2 \sum_{i=1}^n (m^2 - 4my_i + 4y_i^2 - m)^2 + o_p(n\hat{\eta}^4). \end{aligned}$$

A similar argument to that for the PK case leads to this result. Note that the assumption $m \geq 2$ is needed or $m^2 - 4my_i + 4y_i^2 - m = 0$ for all possible values of y. This completes the proof of Theorem 1.

The proof for the model with random family size is similar. However, we need to verify that the likelihood ratio statistic is of order 1 in probability. For the PK case, let $X_i(\eta) = \eta^{-1}[(1-\eta)^{Y_i}(1+\eta)^{M_i - Y_i} - 1]$ for $i = 1,\ldots,n$. Then $n^{-1/2}\sum_{i=1}^n X_i(\eta)$ converges to a Gaussian process when the probability generating function of M_i, $E(s^{M_i})$, exists for $|s| \leq 2$.

Since $\log(1+x) \leq x - x^2/4$ when $|x| < 1$, we have

$$\begin{aligned} \sum_{i=1}^n \log(1 + \alpha\eta X_i(\eta)) &\leq \alpha\eta \sum_{i=1}^n X_i(\eta) - \frac{(\alpha\eta)^2}{4} \sum_{i=1}^n X_i^2(\eta) I(X_i(\eta) < 1) \\ &\leq \{n^{-\frac{1}{2}} \sum_{i=1}^n X_i(\eta)\}^2 \{n^{-1} \sum_{i=1}^n X_i^2(\eta) I(X_i(\eta) < 1)\}^{-1} \\ &= O_p(1) \end{aligned}$$

uniformly in α and η.

The rest of the proof follows that of Theorem 1.

Appendix B. Proof of Theorem 2

Proof of Theorem 2. By Le Cam's contiguity theory, the limiting distribution of R_n under H_a^n in (9) is determined by the null limiting joint distribution of (R_n, Λ_n), where

$$\begin{aligned}\Lambda_n &= l_n(\alpha_0, 0.5 - n^{-\frac{1}{2}}\tau) - l_n(1, 0.5)\\ &= \frac{2\tau\alpha_0}{\sqrt{n}}\sum_{i=1}^n (m_i - 2y_i) - \frac{2\tau^2\alpha_0^2}{n}\sum_{i=1}^n (m_i - 2y_i)^2 + o_p(1).\end{aligned}$$

Let $V_n = \sum_{i=1}^n (m_i - 2y_i)/\sqrt{\sum_{i=1}^n (m_i - 2y_i)^2}$. Note that under H_0, $(1/n)\sum_{i=1}^n (m_i - 2y_i)^2 \xrightarrow{p} E(M_1 - 2Y_1)^2 = E(M_1)$. The null limiting joint distribution of (V_n, Λ_n) is bivariate normal:

$$L((V_n, \Lambda_n)^T | H_0) \xrightarrow{d} \mathcal{N}_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}1 & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}\right),$$

where $(\mu_1, \mu_2)^T = (0, -2\tau^2\alpha_0^2 E(M_1))^T$, $\sigma_{12} = \sigma_{21} = 2\tau\alpha_0\sqrt{E(M_1)}$ and $\sigma_{22} = 4\tau^2\alpha_0^2 E(M_1)$. Noting that $\mu_2 = -\sigma_{22}/2$, by Le Cam's third lemma (see Hájek and Šidák (1967)), the limiting distribution of V_n under H_a^n in (9) is $N(\sigma_{12}, 1)$. Since R_n is asymptotically equivalent to $\{V_n^+\}^2$, the limiting distribution of R_n under the local alternatives H_a^n is that of $\{(Z + \sigma_{12})^+\}^2$.

References

Abreu, P. C., Hodge, S. E. and Greenberg, D. A. (2002). Quantification of type I error probabilities for heterogeneity LOD scores. *Genetic Epidemiology* **22**, 156-169.

Adler, R. J. (1990). An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. *IMS Lecture Notes - Monograph Series* **12**. Institute of Mathematical Statistics, Hayward.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. Roy. Statist. Soc. Ser. B* **63**, 19-29.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *J. Roy. Statist. Soc. Ser. B* **66**, 95-115.

Chen, J. (1998). Penalized likelihood ratio test for finite mixture models with multinomial observations. *Canad. J. Statist.* **26**, 583-599.

Chen, J. and Kalbfleisch, J. D. (2005). Modified likelihood ratio test in finite mixture models with a structural parameter. *J. Statist. Plann. Inference* **129**, 93-107.

Chen, J. and Cheng, P. (1995). The Limit distribution of the restricted likelihood ratio statistic for finite mixture models. *Northeast. Math. J.* **11**, 365-374.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J. Statist. Plann. Inference* **43**, 19-40.

Chiano, M. N. and Yates, J. R. W. (1995). Linkage detection under heterogeneity and the mixture problem. *Ann. Human Genetics* **59**, 83-95.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.* **27**, 1178-1209.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33-43.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247-254.

Fu, Y. (2004). Statistical inference for mixture models. Ph.D. Thesis, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics; Some Basic Concepts.* Springer-Verlag, New York.

Lemdani, M. and Pons, O. (1995). Tests for genetic linkage and homogeneity. *Biometrics* **51**, 1033-1041.

Liang, K. Y. and Rathouz, P. (1999). Hypothesis testing under mixture models, application to genetic linkage analysis. *Biometrics* **55**, 65-74.

Shoukri, M. M. and Lathrop, G. M. (1993). Statistical testing of genetic linkage under heterogeneity. *Biometrics* **49**, 151-161.

Smith, C. A. B. (1963). Testing for heterogeneity of recombination fraction values in human genetics. *Ann. Human Genetics* **27**, 175-182.

Zhu, H. T. and Zhang, H. P. (2004). Hypothesis testing in mixture regression models. *J. Roy. Statist. Soc. Ser. B* **66**, 3-16.

Department of Mathematics and Statistics, York University, Toronto, ON, M3J 1P3, Canada.

E-mail: yuejiao@mathstat.yorku.ca

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada.

E-mail: jhchen@uwaterloo.ca

Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, U.S.A.

E-mail: jdkalbfl@umich.edu

(Received November 2004; accepted September 2005)