Statistica Sinica **15**(2005), 909-925

ON SOME STATISTICAL ASPECTS OF THE INTERVAL MAPPING FOR QTL DETECTION

Zehua Chen and Hanfeng Chen

National University of Singapore and Bowling Green State University

Abstract: The advent of complete genetic linkage maps of DNA markers has made the systematic study of mapping the quantitative trait loci (QTL) in experimental organisms feasible. In recent years, methodological research on QTL mapping has been extensively carried out. However, some related statistical problems remain unsolved. In this article, we consider these problems for the method of interval mapping proposed by Lander and Botstein (1989). We tackle the intrinsic non-identifiability of the involved irregular statistical models and establish the consistency of the maximum likelihood estimates of the putative QTL effect and position. We derive by a non-standard approach the asymptotic distribution of the likelihood ratio test (LRT) statistic for QTL detection. Our result provides a structure for the asymptotic distribution which enjoys the invariance property of regular models. The applications of the results to the determination of threshold values or p-values of interval mapping for QTL detection are discussed and developed. Simulation studies are performed to compare the new approach with the existing methods. The results are presented only for the backcross model but can be extended easily to the intercross model.

Key words and phrases: Asymptotic distribution, backcross, Gaussian process, identifiability, likelihood ratio test; mixture model, QTL mapping.

1. Introduction

The variations of many quantitative traits in human, plants and animals can be attributed mainly to the segregation of genetic factors. Mapping quantitative trait loci (QTL) is of important scientific and economic value in medical research, in plant and animal breeding. Since the seminal paper of Lander and Botstein (1989) on the method of interval mapping, methodological research on QTL mapping has been extensively carried out in recent years. For instance, Haley and Knott (1992), Zeng (1993) and Haley, Knott and Elsen (1994) explored regression methods for QTL mapping; Jansen (1993), Zeng (1994) and Kao, Zeng and Teasdale (1999) developed variants of interval mapping such as composite interval mapping and multiple interval mapping.

The statistical models involved in QTL mapping are mixture models, which are irregular in terms of identifiability. This renders the classical asymptotic

theory inapplicable and poses great challenges to likelihood-based procedures. For instance, it is a thorny problem to determine the threshold values of interval mapping for QTL detection. Some efforts have been made to cope with this problem. For example, Lander and Botstein (1989, 1994), Feingold, Brown and Siegmund (1993), Rebai, Goffinet and Mangin (1994) and Dupuis and Siegmund (1999) considered using certain stochastic processes for the determination of the threshold values. A stochastic process used in this way is usually considered as an approximation to an empirical process arising from the statistical test procedure. However, no serious effort has been made so far to justify the validity of such approximations. The limiting null distribution of the LRT statistic is still a mystery. Although ad hoc methods such as those proposed by Rebai et al. (1994) and Feingold et al. (1993) have been used for determining threshold values, the rigorous determination of such threshold values remains an unsolved problem.

In this article, we establish the consistency of the maximum likelihood estimates (MLE) and a novel theorem for the asymptotic distribution of the LRT statistic for the interval mapping model. The issues we address are of interest not only in QTL mapping, but also in a more general context of mixture models in statistics. Our results provide the rigorous basis for the determination of threshold values of interval mapping for QTL detection. The applications of the results to the determination of threshold values or p-values are also discussed and developed. Simulation studies are performed to compare the new approach with existing methods proposed by Feingold et al. (1993) (FBS) and Rebai et al. (1994) (RGM). The simulation results demonstrate that our method performs extremely well in general, and performs better than FBS and RGM when the latter approaches are applicable. Furthermore, our method is readily applicable in situations where more than one chromosome is under investigation, while the FBS and RGM methods are restricted to the case of a single chromosome.

The article is organized as follows. In Section 2, some background for interval mapping is described. In Section 3, the consistency of the MLE for the interval mapping model is dealt with. In Section 4, the theorem on the asymptotic distribution of the LRT statistic is presented. In Section 5, the issue of the determination of threshold values for interval mapping is addressed, and simulation results are reported. The technical details are given in the Appendix.

2. Background for Interval Mapping Models

In QTL mapping for experimental organisms, either a backcross or an intercross is arranged for segregating progenies. For simplicity, we elaborate on the backcross model in this article. However, the arguments are also applicable to the intercross model.

Let A and B be two markers and Q be a putative QTL flanked by A and B. Let γ, r and s denote, respectively, the recombination fractions between A and B, between A and Q, and between Q and B, where $0 < \gamma < 1/2$ is known. Note that $\gamma = r + s - 2rs$ under the assumption of no interference, and hence $s = (\gamma - r)/(1 - 2r)$. In the backcross model, the possible genotypes are AA and Aa at A , BB and Bb at B, and QQ and Qq at Q. Let X be a coding variable for the genotypes at the two markers: $X = 1, 2, 3$ or 4 according as the genotypes are $AA/BB, AA/Bb, Aa/BB$ or Aa/Bb. The genotype of the putative QTL cannot be observed but can be inferred from the genotypes of the flanking markers. Given the genotypes of the flanking markers, the conditional probabilities for the putative QTL to take genotype QQ and Qq, i.e., $P(QQ|X = l)$ and $P(Qq|X = l)$, are given in Table 1 below. Also given in Table 1 are the probabilities $P(X = l)$ of the marker genotypes.

Table 1. QTL genotype probabilities.

Marker Genotype	X	$P(X = l)$	$P(QQ\|X = l)$	$P(Qq\|X = l)$
AA/BB	1	$\frac{1-\gamma}{2}$	$\frac{(1-r)(1-s)}{1-\gamma}$	$\frac{rs}{1-\gamma}$
AA/Bb	2	$\frac{\gamma}{2}$	$\frac{(1-r)s}{\gamma}$	$\frac{(1-s)r}{\gamma}$
Aa/BB	3	$\frac{\gamma}{2}$	$\frac{(1-s)r}{\gamma}$	$\frac{(1-r)s}{\gamma}$
Aa/Bb	4	$\frac{1-\gamma}{2}$	$\frac{rs}{1-\gamma}$	$\frac{(1-r)(1-s)}{1-\gamma}$

For convenience, we set $q(l) = P(X = l)$ and $p(l, r) = P(QQ|X = l)$. Let Y be the quantitative trait of concern. It is assumed that, given the genotype Qq of Q, Y follows a normal distribution with mean μ_1 and variance σ^2, and that, given the genotype QQ of Q, Y follows a normal distribution with mean μ_2 and the same variance σ^2. Thus the joint probability density function (pdf) of Y and X is given by

$$f(y, x|\theta) = q(x)f(y|x, \theta), \tag{1}$$

where $\theta = (\sigma, r, \mu_1, \mu_2)$ and

$$f(y|x, \theta) = \{1 - p(x, r)\}\sigma^{-1}\phi\Big\{\frac{y - \mu_1}{\sigma}\Big\} + p(x, r)\sigma^{-1}\phi\Big\{\frac{y - \mu_2}{\sigma}\Big\}. \tag{2}$$

Here $\phi(\cdot)$ denotes the pdf of the standard normal distribution. The parameter space of θ is $\Theta = \{\theta = (\sigma, r, \mu_1, \mu_2)$: $\sigma > 0$, $0 \le r \le \gamma$, $|\mu_1| \le M$, $|\mu_2| \le M\}$, where $M > 0$ is assumed to be finite. The following properties will be useful later. Since $\gamma < 1/2$, we have that for any $0 \le r \le \gamma$,

$$p(1, r) + p(2, r) = \frac{(1 - r)\{\gamma + s(1 - 2\gamma)\}}{\gamma(1 - \gamma)} \ge \frac{\gamma + s(1 - 2\gamma)}{\gamma} \ge 1, \tag{3}$$

and equality holds if and only if $r = \gamma$. Similarly,

$$p(1,r) + p(3,r) = \frac{(1-s)\{\gamma + r(1-2\gamma)\}}{\gamma(1-\gamma)} \geq \frac{\gamma + r(1-2\gamma)}{\gamma} \geq 1 \tag{4}$$

and equality holds if and only if $r = 0$.

Let (Y_i, X_i), $i = 1, \ldots, n$, be the observed quantitative trait values and the flanking marker genotypes of a random sample of size n from the backcross population. The joint pdf of (Y_i, X_i), $i = 1, \ldots, n$, is given by $\prod_{i=1}^{n} q(x_i) f(y_i|x_i, \theta)$. Since $q(x_i)$ does not involve any unknown parameters, the log-likelihood function of θ based on the data (Y_i, X_i) can be written as

$$l_n(\theta) = \sum_{i=1}^{n} \log f(Y_i|X_i, \theta), \tag{5}$$

where $f(y_i|x_i, \theta)$ is given by (2).

3. The MLE's of the QTL Effect and Position

The effect of the putative QTL is represented by the difference $\mu_1 - \mu_2$ and its position is indicated by r, if it exists. In this section, the asymptotic properties of the MLEs of the QTL effect and position are investigated. Note that the distribution family (1) is a mixture model which is not identifiable in the parameter θ. For example, when $\mu_1 = \mu_2$, r can be arbitrary. It has been pointed out by many authors (see, e.g., Chernoff and Lander (1995) and Chen and Chen (2001a)) that the loss of identifiability is the main cause of difficulties with mixture models. Therefore, we first deal with the problem of identifiability.

3.1. Identifiability of the mixture model

Before we proceed, we remark that the identifiability of the distribution $f(y,j|\theta)$ in (1) is equivalent to that of the conditional distribution system $f(y|j,\theta)$, $j = 1, \ldots, 4$. Let $G_j = G(\mu|j, r, \mu_1, \mu_2) = \{1 - p(j,r)\}I\{\mu \geq \mu_1\} + p(j,r)I\{\mu \geq \mu_2\}$, where $I\{\cdot\}$ is the indicator function. Then

$$f(y|j,\theta) = f(y|\sigma, G_j) = \int \sigma^{-1}\phi\Big\{\frac{y-\mu}{\sigma}\Big\} dG(\mu|j, r, \mu_1, \mu_2), \quad j = 1, \ldots, 4. \tag{6}$$

Let $\mathcal{G} = \{(G(\cdot|1, r, \mu_1, \mu_2), \ldots, G(\cdot|4, r, \mu_1, \mu_2)) : 0 \leq r \leq \gamma, |\mu_i| \leq M, i = 1, 2\}$. The parameter space for the new parameterization of the distribution system $f(y|\sigma, G_j)$, $j = 1, \ldots, 4$, is $\Omega = \{(\sigma, g) : \sigma > 0, g \in \mathcal{G}\}$.

Lemma 1. *The distribution family* (6) *is identifiable by the parameter* $(\sigma, g) \in \Omega$. *Furthermore, for a fixed* r, (σ, μ_1, μ_2) *identifies the model* (2); *if the constraint* $\mu_1 \neq \mu_2$ *is imposed,* $(\sigma, r, \mu_1, \mu_2)$ *identifies the distributions* (2).

The moment generating function method, Teicher (1960), can be used to prove Lemma 1. For details, see Chen and Chen (2002).

3.2. The consistency of the MLEs

Let $\hat{\theta} = (\hat{\sigma}, \hat{r}, \hat{\mu}_1, \hat{\mu}_2)$ be the MLE of θ at (5). Let $\hat{G}_j = G(\cdot|j, \hat{r}, \hat{\mu}_1, \hat{\mu}_2)$ be the MLE of G_j. The Lévy distance λ of distributions will be used to measure the discrepancy between two distributions.

Theorem 1.
(a) $\hat{\sigma} \to_p \sigma$, *as* $n \to \infty$.
(b) $\lambda\{\hat{G}_j, G(\cdot|j, r, \mu_1, \mu_2)\} \to_p 0$, *as* $n \to \infty$, $j = 1, \ldots, 4$.

Corollary 1.
(a) $\hat{\mu}_1 \to_p \mu_1$, *and* $\hat{\mu}_2 \to_p \mu_2$, *as* $n \to \infty$.
(b) *If* $\mu_1 \neq \mu_2$, $\hat{r} \to_p r$, *as* $n \to \infty$.

The proof of the above results are given in the appendix. Corollary 1 implies that the MLE of the QTL effect is always consistent and moreover, when the putative QTL does exist, the MLE of its position is also consistent.

4. Asymptotic Distribution of the LRT Statistic for the Significance of QTL Effect

In order to test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$, i.e., to decide whether or not the putative QTL really exists in the interval flanked by the two markers, the likelihood ratio procedure is commonly employed. Let $\hat{\theta}$ and $\hat{\theta}_0$ denote the MLEs of θ under H_1 and H_0, respectively. Note that, under H_0, Y_i is independent of X_i and follows a normal distribution with mean μ and variance σ^2. Since the recombination fraction r does not appear under H_0, we can take $\hat{\theta}_0 = (\hat{\sigma}_0, \gamma, \hat{\mu}_0, \hat{\mu}_0)$, where $\hat{\mu}_0 = \bar{Y}$ and $\hat{\sigma}_0^2 = n^{-1}\sum(Y_i - \bar{Y})^2$ are the MLEs of μ and σ^2 under H_0. The LRT statistic is then given by $T_n = 2[l_n(\hat{\theta}) - l_n(\hat{\theta}_0)]$. The null hypothesis H_0 is rejected if T_n exceeds a pre-specified threshold value.

It is a major challenge to determine the threshold value. The classical asymptotic theory of chi-square approximation does not apply here. There are two complications in deriving the asymptotic null distribution of the LRT statistic. One is due to the presence of the structural parameter σ, see comments in Chen and Chen (2001b and 2003). The other is attributable to the loss of identifiability in the parameter θ under the null hypothesis. The re-parameterization by $(\sigma, G_1, \ldots, G_4)$ does not render the ordinary Taylor expansion analysis possible. A highly technical manipulation in the derivation of the asymptotic null distribution seems inevitable. In the following, we state a theorem which provides an

asymptotic approximation to the null distribution of the LRT statistic. Its proof is in the Appendix.

Theorem 2. *Under H_0, as $n \to \infty$, the LRT statistic T_n is asymptotically distributed as*

$$\sup_{0\le r\le\gamma}\Big[\sum_{j=1}^{4}\sqrt{q(j)}\{p(j,r)-\frac{1}{2}\}Z_j/\tau(r)\Big]^2,$$

where $Z_1,\ldots,Z_4$ are independent $N(0,1)$ and $\tau^2(r)=\sum_{j=1}^4 q(j)\{p(j,r)-1/2\}^2$.

It is worthy remarking that the asymptotic null distribution enjoys the invariance property of regular models, that is, it is invariant in the distributions under the composite null hypothesis. However, this invariance property is not shared by the testing problems under other finite mixture models, see Chen, Chen and Kalbfleisch (2001).

Denote by $U(r)$ the stochastic process in Theorem 2 over which the supremum is taken. For fixed r, $U(r)$ follows a χ^2-distribution with degree of freedom 1. For convenience, we call U a χ^2-process. Let $T_n(r)$ be the LRT statistic for testing H_0 against H_1 when r is fixed, that is, $T_n(r)=2\{l_n(\hat\theta(r))-l_n(\hat\theta_0)\}$ where $\hat\theta(r)$ is the MLE of θ under H_1 when r is treated as fixed. We refer to $T_n(r)$ as the LRT process. Note that $T_n=\sup_{0\le r\le\gamma}T_n(r)$. Theorem 2 can be re-stated as $\sup_{0\le r\le\gamma}T_n(r)\to\sup_{0\le r\le\gamma}U(r)$. This is a weaker result than the convergence of the LRT process to the χ^2 process. However it justifies that, for the purpose of determining threshold values for the likelihood ratio test, we can act as if the LRT process converges to the χ^2 process. In our proof of Theorem 2, we employ a sandwich approach to treat the supremum directly, so that the problem with the convergence of the LRT process is bypassed.

5. Application to Determination of Thresholds of Interval Mapping for QTL Detection

In real interval mapping, the whole genome, or the whole of several chromosomes, is searched for the detection of QTL. This involves a collection of intervals on each of which a LRT is conducted. Suppose a total of m intervals is considered. Let T_{kn} denote the LRT statistic for the kth interval and $L_n=\max_{1\le k\le m}T_{kn}$. To guard against the overall error of false detection of QTL, a threshold value c_α needs to be found such that $P(L_n>c_\alpha)\le\alpha$.

In the case that the intervals are connected and are all located on the same chromosome, Rebai et al. (1994) proposed using the Davies bound to approximate c_α as $c_\alpha=c^2$, with c determined by

$$\frac{\alpha}{2}=\Phi(-c)+\frac{1}{\pi}\exp(\frac{-c^2}{2})\sum_{k=1}^{m}\arctan\Big(\sqrt{\frac{\gamma_k}{1-\gamma_k}}\Big), \tag{7}$$

where Φ is the CDF of the standard normal distribution and γ_k is the interval length of the kth interval expressed in units of centimorgans (cM). Feingold et al. (1993) proposed another approximation that determines c by solving

$$\frac{\alpha}{2} = 1 - \Phi(c) + 2Lc\phi(c)\exp(-1.166c\sqrt{\Delta}), \tag{8}$$

where L is the genetic length of the chromosome and Δ is the average spacing of the markers, both expressed in units of Morgans.

The result in the previous section provides us with another way to approximate c_α. By an extension of Theorem 2 we can obtain that, under the null hypothesis of no QTL,

$$L_n \to \max_{1\le k\le m} \sup_{0\le r_k\le\gamma_k} \Big[\sum_{j=1}^{4} \sqrt{q_k(j)}\{p_k(j,r_k) - \frac{1}{2}\}\frac{Z_{kj}}{\tau_k(r_k)}\Big]^2,$$

where Z_{kj}'s are standard normal with

$$\mathrm{Cov}(Z_{kj}, Z_{lj'}) = \begin{cases} 0, & \text{if } k = l, \\ \frac{P(X_k=j, X_l=j')}{\sqrt{q_k(j)q_l(j')}}, & \text{if } k \neq l. \end{cases}$$

See Apendix A4. Here, the subscripts k and l indicate the corresponding intervals, and $P(X_k = j, X_l = j')$ can be determined from the within and between interval recombination fractions. The details of the above extension are given in the appendix. Based on the above extension, the asymptotic null distribution of L_n can be easily simulated.

In the remainder of this section, we present a simulation study to compare the three approaches described above, namely our new approach (CC), that of Rebai et al. (1994) (RGM) and that of Feingold et al. (1993) (FBS).

In the simulation study, we consider (a) a single interval with lengths 2.5cM, 5cM and 10cM, and (b) a chromosome of length 50cM with 5, 10, 20 and 25 equally-spaced intervals. The threshold values with $\alpha = 0.05, 0.025$ and 0.01 are simulated for the true null distribution of L_n with $n = 200$, and for the asymptotic null distribution, and are calculated by using (7) and (8) for the RGM and FBS approaches, respectively. For the true null distribution, 100,000 replicates are used and, for the asymptotic null distribution, 200,000. The results with $\alpha = 0.05$ are presented in Table 2. The results with other α values have the same features and are not reported here.

The following features manifest themselves in Table 2. (1) In the single interval case, all three approaches provide reasonable approximations to the exact thresholds, though FBS is least satisfactory. (2) In the multiple interval cases, when the markers are sparse (fewer intervals used on a chromosome of the same

length), RGM provides reasonable approximations. The accuracy of the RGM approximation gets worse as the markers get dense. The results with FBS approximation are just the opposite. (3) The new approach CC provides accurate approximations to the exact thresholds and beats the other two in all cases with but one exception.

Table 2. Simulated thresholds of interval mapping for QTL detection by the CC, RGM and FBS approaches. Sample size $n = 200$, m is the number of intervals and δ is the interval length in units of cM. The numbers in parentheses are absolute differences between the simulated asymptotic and exact thresholds.

m	δ	Exact	CC	RGM	FBS
1	2.5	4.23	4.27	4.29	4.10
			(0.04)	(0.06)	(**0.13**)
1	5	4.45	4.47	4.45	4.28
			(0.02)	(0.00)	(**0.18**)
1	10	4.67	4.68	4.65	4.51
			(0.01)	(0.02)	(**0.16**)
5	10	6.38	6.25	6.44	6.14
			(0.13)	(0.06)	(**0.24**)
10	5	6.65	6.64	7.02	6.55
			(0.01)	(**0.37**)	(0.10)
20	2.5	6.73	6.75	7.62	6.88
			(0.02)	(**0.99**)	(0.15)
25	2	7.04	7.05	7.82	6.97
			(0.01)	(**0.78**)	0.07

Besides its superior performance in the above cases, CC can also provide good approximations when more than one chromosome is under investigation. Neither the RGM nor the FBS approach can be applied directly when two or more chromosomes are involved. The formulas (7) and (8) do not distinguish between the case of a single chromosome and the case of multiple chromosomes, while the exact threshold values are quite different in these two cases. For example, consider the case of five intervals of length 10 cM connected one after another and located at a single chromosome, and the case of five intervals of the same length but separately located on five chromosomes. With $\alpha = 0.05$, the exact threshold value in the single chromosome case is 6.38 and is 7.70 in the multiple chromosome case. We end this section by reporting a simulation study to demonstrate the appropriateness of the CC approach in the multiple chromosome case. We considered two extreme unlinked settings: five intervals, each of length 10 cM, located on five chromosomes, and ten intervals, each of length 5 cM, located on ten chromosomes. In each setting, we simulated the

exact thresholds of L_n with $n = 200$, and the asymptotic thresholds by using CC approach. The results are reported in Table 3.

Table 3. Simulated CC and exact thresholds of interval mapping for QTL detection on multiple chromosomes. Sample size $n = 200$.

	Five Intervals		Ten Intervals	
α	Exact	CC	Exact	CC
0.01	10.75	10.66	11.78	11.66
0.025	9.00	8.90	10.01	9.95
0.05	7.70	7.59	8.66	8.63

Appendix. Technical details

A1. Proof of Theorem 1

The following lemma is needed in the proof of Theorem 1. It is similar to (but stronger than) Lemma 1 in Chen and Chen (2001b) and its proof is omitted. For a proof, see Chen and Chen (2002).

Lemma 2. *There are constants* $0 < \delta < \Delta < \infty$ *such that* $\lim_{n\to\infty} P(\delta \le \hat{\sigma} \le \Delta) = 1$.

Proof of Theorem 1. The proof is accomplished by using Wald's (1949) Consistency Theorem. Let $\sigma = \sigma_0$ be the true value of σ. By Lemma 2, without loss of generality, we can confine σ to an interval $[\delta, \Delta]$, where δ and Δ are chosen such that $0 < \delta < \sigma_0 < \Delta < \infty$. The parameter space of the model (6) is reduced to $\bar{\Omega} = \{(\sigma, g) : \sigma \in [\delta, \Delta], g \in \mathcal{G}\}$. Now define a metric on $\bar{\Omega}$ as follows: $d((\sigma_1, g_1), (\sigma_2, g_2)) = |\sigma_1 - \sigma_2| + \sum_{j=1}^{4} \lambda(G_{1j}, G_{2j})$, where $g_k = (G_{k1}, G_{k2}, G_{k3}, G_{k4}), k = 1, 2$, and $\lambda(\cdot, \cdot)$ is the Lévy distance. Note that the convergence of distribution functions in Lévy distance is equivalent to weak convergence. Therefore, the convergence in the d-metric on $\bar{\Omega}$ is equivalent to convergence of the first component in the usual Euclidean metric and the weak convergence of the remaining four components. Note that $\bar{\Omega}$ is compact with metric d (see Chen and Chen (2002) for details).

Denote a point $(\sigma, g) \in \bar{\Omega}$ by ω and let $f(y, x; \omega)$ be the joint pdf of (Y_i, X_i). Let $\omega_0 = (\sigma_0, g_0)$ be the true value of ω. Following Wald (1949), for any ω and for any positive value ρ let $f(y, x; \omega, \rho)$ be the supremum of $f(y, x; \omega')$ with respect to ω' when $d(\omega, \omega') \le \rho$. For any positive τ, let $\psi(y, x; \tau)$ be the supremum of $f(y, x; \omega')$ with respect to ω' when $d(\omega_0, \omega') > \tau$. Let $f^*(y, x; \omega, \rho) = f(y, x; \omega, \rho)$ when $f(y, x; \omega, \rho) > 1$, and 1 otherwise. Similarly, let $\psi^*(y, x; \tau) = \psi(y, x; \tau)$ when $\psi(y, x; \tau) > 1$, and 1 otherwise. It can be seen that the following Wald conditions hold (see Chen and Chen (2002)):(i) $E\{\log(f^*(Y, X; \omega, \rho)\} < \infty$ and

$E\{\log(\psi^*(Y,X;\tau)\} < \infty$, and (ii) if $\lim_{k\to\infty} d(\omega_k,\omega)=0$, then $\lim_{k\to\infty} f(y,x;\omega_k) = f(y,x;\omega)$.

Other conditions of Wald's Consistency Theorem are either trivial or have already been established, such as identifiability (Lemma 1) and compactness of $\bar{\Omega}$. Then the MLE $(\hat{\sigma},\hat{G}_1,\hat{G}_2,\hat{G}_3,\hat{G}_4)$ is consistent with respect to the metric d and the proof is completed.

A2. Proof of Corollary 1

By part (b) of Theorem 1 we have that, for any G-continuous function h (here by a G-continuous function we mean a function which is continuous at $\mu=\mu_1$ and $\mu=\mu_2$), the MLE of $\int h(\mu)dG(\mu|j,r,\mu_1,\mu_2) = \{1-p(j,r)\}h(\mu_1)+p(j,r)h(\mu_2)$ is consistent, i.e.,

$$\{1-p(x,\hat{r})\}h(\hat{\mu}_1)+p(x,\hat{r})h(\hat{\mu}_2) \to \{1-p(x,r)\}h(\mu_1)+p(x,r)h(\mu_2), \tag{9}$$

in probability, for $j=1,\ldots,4$. The above convergence, together with the fact that $p(1,\hat{r})+p(4,\hat{r})=1$, yields that for any G-continuous function h,

$$h(\hat{\mu}_1)+h(\hat{\mu}_2) \to h(\mu_1)+h(\mu_2), \tag{10}$$

in probability.

Now consider the case $\mu_1=\mu_2=\mu_0$ (say) and the case $\mu_1\neq\mu_2$ separately. In the first case, let $h(\mu)=(\mu-\mu_0)^2$ in (10). It follows that $(\hat{\mu}_1-\mu_0)^2+(\hat{\mu}_2-\mu_0)^2 \to 0$ in probability, and hence $\hat{\mu}_1$ and $\hat{\mu}_2$ are consistent.

In the case $\mu_1\neq\mu_2$, consider two situations: $r=\gamma$ and $r<\gamma$. When $r=\gamma$, $s=0$ and $p(1,r)=p(3,r)=1$. Taking $h(\mu)=(\mu-\mu_2)^2$ in (9) yields that for $j=1$ and 3, $\{1-p(j,\hat{r})\}(\hat{\mu}_1-\mu_2)^2+p(j,\hat{r})(\hat{\mu}_2-\mu_2)^2 \to \{1-p(j,r)\}(\mu_1-\mu_2)^2+p(j,r)(\mu_2-\mu_2)^2=0$. Hence $\{p(1,\hat{r})+p(3,\hat{r})\}(\hat{\mu}_2-\mu_2)^2\to 0$, which implies that $\hat{\mu}_2\to\mu_2$ since $p(1,\hat{r})+p(3,\hat{r})\geq 1$ according to (4). Then, by taking $h(\mu)=\mu$, the consistency of $\hat{\mu}_2$ together with (10) implies the consistency of $\hat{\mu}_1$. In this case it is seen from (9) that $p(1,\hat{r})\to 1$ and $p(3,\hat{r})\to 1$, implying $\hat{r}\to\gamma$.

When $r<\gamma$, to fix the point assume $\mu_1>\mu_2$. Let $\bar{\mu}=(\mu_1+\mu_2)/2$. Since $\hat{G}_j(\bar{\mu})\to G(\bar{\mu}|j,r,\mu_1,\mu_2)=p(j,r)$ for all j,

$$\hat{G}_1(\bar{\mu})+\hat{G}_2(\bar{\mu}) \to p(1,r)+p(2,r)\equiv\alpha, \text{ say.} \tag{11}$$

Note that $\hat{G}_1(\bar{\mu})+\hat{G}_2(\bar{\mu})$ assumes only four possible values: $p(1,\hat{r})+p(2,\hat{r})$ if $\bar{\mu}<\hat{\mu}_1$ and $\bar{\mu}\geq\hat{\mu}_2$; $2-\{p(1,\hat{r})+p(2,\hat{r})\}$ if $\bar{\mu}\geq\hat{\mu}_1$ and $\bar{\mu}<\hat{\mu}_2$; 2 if $\bar{\mu}\geq\hat{\mu}_1$ and $\bar{\mu}\geq\hat{\mu}_2$; 0 if $\bar{\mu}<\hat{\mu}_1$ and $\bar{\mu}<\hat{\mu}_2$. But, according to (3), $p(1,\hat{r})+p(2,\hat{r})\geq 1$, $2-\{p(1,\hat{r})+p(2,\hat{r})\}\leq 1$, and $1<\alpha<2$ when $r<\gamma$. Therefore, (11) implies that, with probability approaching one, $\hat{G}_1(\bar{\mu})+\hat{G}_2(\bar{\mu})=p(1,\hat{r})+p(2,\hat{r})$. Hence $\bar{\mu}<\hat{\mu}_1$ and $\bar{\mu}\geq\hat{\mu}_2$ with probability approaching one. It thus follows that for any

bounded G-continuous function h, $(2-\alpha)h(\hat{\mu}_1)+\alpha h(\hat{\mu}_2) \to (2-\alpha)h(\mu_1)+\alpha h(\mu_2)$ in probability. Let $h(\mu) = \mu I(\mu \le \bar{\mu})$. Then $h(\mu)$ is G-continuous and bounded since only $|\mu| \le M$ is relevant. Then $(2-\alpha)h(\hat{\mu}_1) + \alpha h(\hat{\mu}_2) = \alpha\hat{\mu}_2 \to \alpha\mu_2$ in probability, i.e., $\hat{\mu}_2$ is consistent. Similarly, $\hat{\mu}_1$ is consistent. These results in turn imply that $p(1,\hat{r}) \to p(1,r)$ and so $\hat{r} \to r$. The corollary is proved.

A3. Proof of Theorem 2

Suppose the null distribution of the quantitative trait is $N(\mu_0, \sigma_0^2)$. Without loss of generality, take $\mu_0 = 0$ and $\sigma_0 = 1$, since otherwise we can consider transformed data $\tilde{Y}_i = (Y_i - \mu_0)/\sigma_0$ and the re-parameterization $\tilde{\mu}_1 = (\mu_1 - \mu_0)/\sigma_0$, $\tilde{\mu}_2 = (\mu_2 - \mu_0)/\sigma_0$, and $\tilde{\sigma} = \sigma/\sigma_0$ that does not affect the position parameter r. Furthermore, we take $|\sigma - 1| \le \delta$ where $0 < \delta < 1$, since in light of Theorem 1, $\hat{\sigma} \to 1$ in probability under the null hypothesis.

Write $T_n = 2[l_n(\hat{\theta}) - l_n(\theta_0)] - 2[l_n(\hat{\theta}_0) - l_n(\theta_0)] = T_{n1} - T_{n2}$, say, where $\theta_0 = (1, \gamma, 0, 0)$. First we establish asymptotic approximations to T_{n1} and T_{n2}:

$$T_{n1} = \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \sup_{0\le r\le\gamma} \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + o_p(1), \tag{12}$$

$$T_{n2} = n(\bar{Y})^2 + (\frac{1}{2n})(\sum_{i=1}^{n} U_i)^2 + o_p(1), \tag{13}$$

where $U_i = Y_i^2 - 1$ and $V_i(r) = \{p(X_i, r) - 1/2\}Y_i$. Note that $Ep(X_i, r) = 1/2$. Hereafter, by $o_p(1)$ and $O_p(1)$ we mean convergent to zero in probability uniformly in θ and bounded in probability uniformly in θ, respectively. Since (12) and (13) imply

$$T_n = \sup_{0\le r\le\gamma} \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + o_p(1),$$

the theorem follows.

Equation (13) is trivial by a standard analysis. To justify (12), let $R_n(\theta) = 2\{l_n(\theta) - l_n(\theta_0)\}$, so $T_{n1} = R_n(\hat{\theta})$. Write $R_n(\theta) = 2\sum_{i=1}^{n} \log\{1 + \delta_i(\theta)\}$, where

$$\delta_i(\theta) = \{1 - p(X_i, r)\} \left[\sigma^{-1}\phi\{\frac{Y_i - \mu_1}{\sigma}\}/\phi(Y_i) - 1\right] + p(X_i, r) \left[\sigma^{-1}\phi\{\frac{Y_i - \mu_2}{\sigma}\}/\phi(Y_i) - 1\right].$$

We first obtain asymptotic bounds for $R_n(\theta)$. Let

$$\bar{R}_n(\theta) = \sum_{i=1}^{n} \{2\delta_i(\theta) - \delta_i^2(\theta) + \frac{2}{3}\delta_i^3(\theta)\}. \tag{14}$$

Then for any θ, $R_n(\theta) \le \bar{R}_n(\theta)$. Note that

$$\left.\frac{\partial\delta_i(\theta)}{\partial\sigma}\right|_{\theta=\theta_0} = U_i, \quad \left.\frac{\partial\delta_i(\theta)}{\partial\mu_1}\right|_{\theta=\theta_0} = \{1-p(X_i,r)\}Y_i, \quad \left.\frac{\partial\delta_i(\theta)}{\partial\mu_2}\right|_{\theta=\theta_0} = p(X_i,r)Y_i,$$

where $U_i = Y_i^2 - 1$ as defined before. Expanding $\delta_i(\theta)$ at θ_0 gives

$$\delta_i(\theta) = (\sigma-1)U_i + \mu_1\{1-p(X_i,r)\}Y_i + \mu_2 p(X_i,r)Y_i + \epsilon_i(\theta). \tag{15}$$

where $\epsilon_i(\theta) = (\sigma-1)\{U_i(\sigma)-U_i\} + \mu_1\{1-P(X_i,r)\}\{Y_i(\sigma,\mu_1)-Y_i\} + \mu_2 P(X_i,r)\{Y_i(\sigma,\mu_2)-Y_i\}$, with $U_i(\sigma) = (\sigma-1)^{-1}\{\sigma^{-1}\phi(Y_i/\sigma)/\phi(Y_i) - 1\}$ and $Y_i(\sigma,\mu) = \mu^{-1}[\sigma^{-1}\phi\{(Y_i-\mu)/\sigma\} - \sigma^{-1}\phi(Y_i/\sigma)]/\phi(Y_i)$. Re-group the leading terms of $\delta_i(\theta)$ in (15) as a linear combination of uncorrelated random variables as follows:

$$\delta_i(\theta) = a_1(\theta)U_i + a_2(\theta)Y_i + a_3(\theta)V_i(r) + \epsilon_i(\theta), \tag{16}$$

where $V_i(r) = \{p(X_i,r) - 1/2\}Y_i$ as before, $a_1(\theta) = \sigma - 1$, $a_2(\theta) = (\mu_1+\mu_2)/2$, and $a_3(\theta) = \mu_2 - \mu_1$.

Note that under the null hypothesis, for any r, $EU_i = EY_i = EV_i(r) = E\{U_iY_i\} = E\{U_iV_i(r)\} = E\{Y_iV_i(r)\} = 0$. Let

$$L_n(a_1(\theta),a_2(\theta),a_3(\theta);r) = a_1(\theta)\sum_{i=1}^n U_i + a_2(\theta)\sum_{i=1}^n Y_i + a_3(\theta)\sum_{i=1}^n V_i(r), \tag{17}$$

$$Q_n(a_1(\theta),a_2(\theta),a_3(\theta);r) = a_1^2(\theta)\sum_{i=1}^n U_i^2 + a_2^2(\theta)\sum_{i=1}^n Y_i^2 + a_3^2(\theta)\sum_{i=1}^n V_i^2(r). \tag{18}$$

Write $u(\theta) = a_1^2(\theta) + a_2^2(\theta) + a_3^2(\theta)$ and $v(\theta) = \max\{|a_1(\theta)|, |a_2(\theta)|, |a_3(\theta)|\}$.

Lemma 3. *Assume the distribution under the null hypothesis.*

(i) *For $k = 1, 2$ and 3,*

$$\sum_{i=1}^n \delta_i^k(\theta) = a_1^k(\theta)\sum_{i=1}^n U_i^k + a_2^k(\theta)\sum_{i=1}^n Y_i^k + a_3^k(\theta)\sum_{i=1}^n V_i^k(r) + nO_p\{u(\theta)v(\theta)\}.$$

(ii) *The quadratic form $n^{-1}Q_n(x_1,x_2,x_3;r)$ converges to the positive definite form $2x_1^2 + x_2^2 + \tau^2(r)x_3^2$ almost surely and uniformly in $0 \le r \le \gamma$, where $\tau^2(r) = Var\{p(X;r)\}$ as defined in Theorem 2,*

(iii) $\sum \delta_i^3(\theta)/\sum \delta_i^2(\theta) = O_p\{v(\theta)\}$.

In light of Lemma 3, and by (14), (16), (17) and (18), we have

$$\bar{R}_n(\theta) = 2L_n(a_1(\theta),a_2(\theta),a_3(\theta);r) - Q_n(a_1(\theta),a_2(\theta),a_3(\theta);r)[1+O_p\{v(\theta)\}]. \tag{19}$$

Note that for any x_1, x_2, x_3 and $r \in [0, \gamma]$,

$$2L_n(x_1, x_2, x_2; r) - Q_n(x_1, x_2, x_3; r) \le \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)},$$

and equality holds when

$$x_1 = \frac{\sum U_i}{\sum U_i^2},\ x_2 = \frac{\sum Y_i}{\sum Y_i^2},\ x_3 = \frac{\sum V_i(r)}{\sum V_i^2(r)}. \tag{20}$$

Let $\tilde{\theta}(r) = (\tilde{\sigma}, r, \tilde{\mu}_1, \tilde{\mu}_2)$ be the solution to the equations $a_j(\theta) = \tilde{x}_j,\ j = 1, 2, 3$, with the $\tilde{x}_j$'s given by the values in (20). Then (19) implies that

$$\bar{R}_n(\tilde{\theta}(r)) = \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + O_p\{v(\tilde{\theta}(r))\}, \tag{21}$$

and for any θ, $\bar{R}_n(\theta) \le \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \frac{(\sum V_i(r))^2}{\sum V_i^2(r)} + O_p\{v(\theta)\}$. In particular,

$$\begin{aligned} R_n(\hat{\theta}) \le \bar{R}_n(\hat{\theta}) &\le \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \frac{\{\sum V_i(\hat{r})\}^2}{\sum V_i^2(\hat{r})} + O_p\{v(\hat{\theta})\} \\ &\le \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \sup_{0 \le r \le \gamma} \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + O_p\{v(\hat{\theta})\}. \end{aligned}$$

According to Theorem 1 and Corollary 1, $v(\hat{\theta}) = o_p(1)$. Thus,

$$R_n(\hat{\theta}) \le \bar{R}_n(\hat{\theta}) \le \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \sup_{0 \le r \le \gamma} \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + o_p(1). \tag{22}$$

Next we show that the upper bound as given above is attainable. Let $\tilde{\theta}(r)$ be defined as before. Considering a Taylor expansion of $\log(1 + t)$ for $t = \delta_i(\tilde{\theta}(r))$, we have

$$R_n(\tilde{\theta}(r)) = 2\sum_{i=1}^{n} \log\{1 + \delta_i(\tilde{\theta}(r))\} = \bar{R}_n(\tilde{\theta}(r)) - \frac{1}{4}\sum_{i=1}^{n} \{\delta_i^4(\tilde{\theta}(r))/(1 + \eta_i)^4\}, \tag{23}$$

where $|\eta_i| < |\delta_i(\tilde{\theta}(r))|$. We show the remainder in (23) is negligible. Since $\tilde{x}_j = O_p(n^{-1/2})$ uniformly in r, $\tilde{\theta}(r) - \theta_0(r) = O_p(n^{-1/2})$ uniformly in r, where $\theta_0(r) = (1, r, 0, 0)$. Thus, for a generic constant C,

$$\begin{aligned} \sup_{0 \le r \le \gamma} \max_{1 \le i \le n} \{\delta_i^4(\tilde{\theta}(r))\} &\le C \sum_{l+s=4} |\tilde{\sigma} - 1|^l (|\tilde{\mu}_1| + |\tilde{\mu}_2|)^s \sup_{\sigma,\mu} \max_{1 \le i \le n} \{|U_i(\sigma)|^l |Y_i(\sigma, \mu)|^s\} \\ &\le O_p(n^{-2}) o_p(n^{\frac{1}{2}}) = o_p(n^{-\frac{3}{2}}), \end{aligned}$$

where the last inequality follows from

$$|U_i(\sigma)|^l|Y_i(\sigma,\mu)|^s \le [|Y_i(\sigma,\mu)|+|U_i(\sigma)|]^4 \le C \max_{1\le i\le n}[1+|Y_i|^2]^4 \exp\{C|Y_i|\} = o_p(n^{1/2}).$$

See Serfling (1980, p.91). This implies that $\sum_{i=1}^n \sup_{0\le r\le\gamma} \delta_i^4(\tilde\theta(r)) = o_p(n^{-1/2})$. Since $\sup_{0\le r\le\gamma} \max_{1\le i\le n}\{|\eta_i|\} \le \sup_{0\le r\le\gamma} \max_{1\le i\le n}\{|\delta_i(\tilde\theta(r))|\} = o_p(n^{-3/8})$, it is immediate that $(1/4)\sum_{i=1}^n\{\delta_i^4(\tilde\theta(r))/(1+\eta_i)^4\} = o_p(1)$, and hence, by (23),

$$R_n(\hat\theta) \ge \sup_{0\le r\le\gamma} R_n(\tilde\theta(r)) = \sup_{0\le r\le\gamma} \bar R_n(\tilde\theta(r)) + o_p(1). \tag{24}$$

Since $\tilde\theta(r) - \theta_0(r) = O_p(n^{-1/2})$, $v(\tilde\theta(r)) = o_p(1)$ uniformly in r. It is thus seen from (21) that

$$\sup_{0\le r\le\gamma} \bar R_n(\tilde\theta(r)) = \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \sup_{0\le r\le\gamma} \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + o_p(1).$$

So (24) implies

$$R_n(\hat\theta) \ge \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \sup_{0\le r\le\gamma} \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + o_p(1). \tag{25}$$

Combining (22) and (25) yields

$$R_n(\hat\theta) = \frac{(\sum U_i)^2}{\sum U_i^2} + \frac{(\sum Y_i)^2}{\sum Y_i^2} + \sup_{0\le r\le\gamma} \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + o_p(1).$$

Thus (12) is established, and (12) and (13) lead to

$$T_n = \sup_{0\le r\le\gamma} \frac{\{\sum V_i(r)\}^2}{\sum V_i^2(r)} + o_p(1). \tag{26}$$

Note that, uniformly in r, $n^{-1}\sum V_i^2(r) \to \tau^2(r)$, and the process $n^{-1/2}\sum V_i(r)$ approaches weakly the process $\sum_{j=1}^4 \sqrt{q(j)}\{p(j,r) - 1/2\}Z_j$. Theorem 2 is then contingent on Lemma 3.

Proof of Lemma 3. First we see that

$$\sup_{0\le r\le\gamma} n^{-1/2}|\sum_{i=1}^n Y_iV_i(r)| = O_p(1), \tag{27}$$

$$\sup_{0\le r\le\gamma} n^{-1/2}|\sum_{i=1}^n U_iV_i(r)| = O_p(1), \tag{28}$$

from the tightness of the processes $n^{-1/2}\sum Y_iV_i(r)$ and $n^{-1/2}\sum U_iV_i(r)$. (For a proof of the tightness, see Chen and Chen (2002).)

Next we show that, for non-negative integers a, b, c, and d such that $0 \leq a+b+c+d \leq 2$,

$$\sum_{i=1}^{n} \epsilon_i^{1+a}(\theta)\{a_1(\theta)U_i\}^b\{a_2(\theta)Y_i\}^c\{a_3(\theta)V_i(r)\}^d = nO_p\{u(\theta)v(\theta)\}, \qquad (29)$$

and that, for $b+c+d \leq 3$, with at least two of b, c, d being not 0,

$$\sum_{i=1}^{n} \{a_1(\theta)U_i\}^b\{a_2(\theta)Y_i\}^c\{a_3(\theta)V_i(r)\}^d = nO_p\{u(\theta)v(\theta)\}. \qquad (30)$$

By the tightness (Chen and Chen (2001b)) of the derivative process $n^{-1/2}\sum \{U_i(\sigma) - U_i\}/(\sigma-1)$, for $|\sigma - 1| \leq \delta$, $\sup_{|\sigma-1|\leq\delta}\{|\sum_{i=1}^n\{U_i(\sigma) - U_i\}/(\sigma - 1)|\} = O_p(n^{1/2})$. Similar approximations can be obtained for the derivative processes of $n^{-1/2}\sum Y_i(\sigma, \mu)$. Therefore, $\sum_{i=1}^n \epsilon_i(\theta) = \sqrt{n}O_p[(\sigma-1)^2 + |(\sigma - 1)\mu_1| + |(\sigma-1)\mu_2| + |\mu_1|^2 + |\mu_2|^2]$. By Cauchy-Schwards inequality, $|(\sigma-1)\mu_k| \leq a_1^2(\theta) + |\mu_k^2|$ for $k = 1$ and 2, and $|\mu_1|^2 + |\mu_2|^2 \leq C(a_2^2(\theta) + a_3^2(\theta))$ for a constant C. Thus, $\sum \epsilon_i(\theta) = n^{1/2}O_p\{u(\theta)\}$ which, of course, can also be written as $nO_p\{u(\theta)v(\theta)\}$. Other combinations in (29) are easily seen to be of the order required. For example, by the Uniform Strong Law of Large Numbers, $n^{-1}\sup_{|\sigma-1|\leq\delta}\sum_{i=1}^n |\{U_i(\sigma) - U_i\}/(\sigma-1)|^k = O_p(1)$, and similarly for the process $Y_i(\sigma, \mu)$ and others. Thus, $\sum_{i=1}^n \epsilon_i^2(\theta) = nO_p\{|a_1(\theta)|^3 + |a_2(\theta)|^3 + |a_3(\theta)|^3\} = nO_p\{u(\theta)v(\theta)\}$, and $|\sum_{i=1}^n \epsilon_i(\theta)a_1(\theta)U_i| \leq |a_1(\theta)|\sum_{i=1}^n |\epsilon_i(\theta)U_i| = |a_1(\theta)|\, O_p\{nu(\theta)\} \leq nO_p\{u(\theta)v(\theta)\}$. The summations in (30) can be shown similarly, but (27) and (28) may also be used in case of $b+c+d = 2$. Part (i) of the lemma then follows from (27), (28), (29) and (30).

In part (ii), the uniform convergence follows from the Uniform Strong Law of Large Numbers, and the positive-definiteness of the limit quadratic form is guaranteed by the assumption $\gamma < 1/2$.

By parts (i) and (ii), we have

$$\frac{|\sum \delta_i^3(\theta)|}{\sum \delta_i^2(\theta)} \leq v(\theta)O_p\{n^{-1}\sum_{i=1}^n\{|U_i|^3 + |Y_i|^3 + |V_i(r)|^3\}\} = O_p(v(\theta)),$$

which implies part (iii). Lemma 3 is thus proved and hence the proof of Theorem 2 is complete.

A4. Extension of Theorem 2 to Multiple Interval Case

The argument in the proof of Theorem 2 can be carried out for each of the

m intervals. In the proof, we essentially show, see (26), that

$$T_{kn} = \sup_{0\le r_k\le\gamma_k} \frac{\left\{\sum_{i=1}^n \left(p_k(X_{ki},r_k)-\frac{1}{2}\right)Y_i\right\}^2}{\sum_{i=1}^n \left\{p_k(X_{ki},r_k)-\frac{1}{2}\right\}^2 Y_i^2} + o_P(1)$$
$$= \sup_{0\le r_k\le\gamma_k} \frac{\left\{\sum_{j=1}^4 \sqrt{q_k(j)}(p_k(j,r_k)-\frac{1}{2})\frac{1}{\sqrt{nq_k(j)}}\sum_{i=1}^n Y_i I(X_{ki}=j)\right\}^2}{\tau_k^2(r_k)} + o_P(1).$$

Let $U_{nkj} = \{nq_k(j)\}^{-1/2}\sum_{i=1}^n Y_i I\{X_{ki}=j\}$. Then by the Multivariate Central Limit Theorem, $\{U_{nkj}\}\to\{Z_{kj}\}$, jointly in distribution as $n\to\infty$, where Z_{kj} are normal variables with means 0, variances 1 and covariances $\mathrm{Cov}(Z_{kj},Z_{lj'}) = \mathrm{Cov}(U_{nkj},U_{nlj'})$. It is easy to see that when $k=l$ the covariances are zero, and that when $k\neq l$, the covariances are given by

$$\mathrm{Cov}(U_{nkj},U_{nlj'}) = \frac{\mathrm{Cov}(Y_iI\{X_{ki}=j\},Y_iI\{X_{li}=j'\})}{\sqrt{q_k(j)q_l(j')}} = \frac{P(X_{ki}=j,X_{li}=j')}{\sqrt{q_k(j)q_l(j')}}.$$

References

Chen, H. and Chen, J. (2001a). The likelihood ratio test for homogeneity in finite mixture models. *Canad. J. Statist.* **29**, 201-215.

Chen, H. and Chen, J. (2001b). Tests for homogeneity in normal mixtures in presence of a structural parameter: technical details. Working Paper 2001-06 Department of Statistics and Actuarial Science, University of Waterloo.

Chen, H. and Chen, J. (2003). Tests for homogeneity in normal mixtures in presence of a structural parameter. *Statist. Sinica* **13**, 351-365.

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, Z. and Chen, H. (2002). On some statistical aspects of the interval mapping for QTL detection. Technical Report **10/2002** Department of Statistics and Applied Probability, National University of Singapore (http://www.stat.nus.edu.sg /research_treports.htm).

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247-254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33-43.

Dupuis, J. and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**, 373-386.

Feingold, E., Brown, P. O. and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Amer. J. Hum. Genet.* **53**, 234-251.

Haley, C. S. and Knott, S. A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315-324.

Haley, C. S., Knott, S. A. and Elsen, J. M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195-1207.

Jansen, P. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205-211.

Kao, C. H., Zeng, Z. B. and Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-1216.

Lander, E. S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-199.

Rebai, A., Goffinet, B. and Mangin, B. (1994). Approximated thresholds of interval mapping tests for QTL detection. *Genetics* **138**, 235-240.

Rubin, H. (1956). Uniform convergency of random functions with applications to statistics. *Ann. Math. Statist.* **27**, 200-203.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Teicher, H. (1960). Identifiability of mixtures. *Ann. Math. Statist.* **32**, 244-248.

Zeng, Z. B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Acad. Sci. USA* **90**, 10972-10976.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457-1468.

Department of Statistics & Applied Probability, National University of Singapore, 3 Science Drive 2, Singapore 117543.

E-mail: stachenz@nus.edu.sg

Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio 43403, U.S.A.

E-mail: hchen@math.bgsu.edu

(Received January 2004; accepted November 2004)