Journal of Statistical Planning and Inference 139 (2009) 764 – 779

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Large sample interval mapping method for genetic trait loci in finite regression mixture models

Hong Zhang[a], Hanfeng Chen[b,*], Zhaohai Li[c,d]

[a] *Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, P.R. China*
[b] *Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA*
[c] *Department of Statistics, George Washington University, 2140 Pennsylvania Avenue NW, Washington, DC 20052, USA*
[d] *Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, 6120 Executive Boulevard, EPS, Bethesda, Maryland 20892, USA*

ARTICLE INFO

Article history:
Received 14 November 2006
Received in revised form
21 September 2007
Accepted 12 March 2008
Available online 22 May 2008

MSC:
primary 62F05
62F12
secondary 62P10

Keywords:
Backcross design
Binomial mixture
Generalized linear regression mixture
Identifiability
Intercross design
Likelihood ratio test
Logistic regression mixture
Normal mixture
Poisson mixture

ABSTRACT

This article investigates the large sample interval mapping method for genetic trait loci (GTL) in a finite non-linear regression mixture model. The general model includes most commonly used kernel functions, such as exponential family mixture, logistic regression mixture and generalized linear mixture models, as special cases. The populations derived from either the backcross or intercross design are considered. In particular, unlike all existing results in the literature in the finite mixture models, the large sample results presented in this paper do not require the boundness condition on the parametric space. Therefore, the large sample theory presented in this article possesses general applicability to the interval mapping method of GTL in genetic research. The limiting null distribution of the likelihood ratio test statistics can be utilized easily to determine the threshold values or p-values required in the interval mapping. The limiting distribution is proved to be free of the parameter values of null model and free of the choice of a kernel function. Extension to the multiple marker interval GTL detection is also discussed. Simulation study results show favorable performance of the asymptotic procedure when sample sizes are moderate.

1. Background and model

1.1. Introduction

Interval mapping method, proposed first by Thoday (1960) and developed substantially further by Lander and Botstein (1989), is a fundamental methodology in statistical genetics to systematically map loci underlying a trait in experimental organisms, via a number of genetic markers whose genotypes are observable. The basic idea involved is to consider one marker interval at a time to detect a putative genetic trait locus (GTL) in the flanking marker interval by performing the likelihood ratio test (LRT).

Start with two completely inbred parental population lines, P_1 and P_2, one consisting of identical homozygous individuals with one allele and the other consisting of identical homozygous individuals with another allele; the two populations P_1 and P_2 differ substantially in the trait of interest. Let F_1 be a population derived from a line-cross between P_1 and P_2. The F_1 progeny

* Corresponding author. Tel.: +1 419 372 7463; fax: +1 419 372 6092.
E-mail address: hchen@bgsu.edu (H. Chen).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.03.041

are all identical heterozygotes. If the F_1 individuals are then backcrossed with individuals from a parent population, say P_1 for specific, a backcross population is derived; if the F_1 individuals are selfed or intermated, an intercross population F_2 is induced.

Let Y be the trait of interest, either quantitative or qualitative. In the latter case, assume Y is coded by distinct numbers, say 0 or 1 when Y is a binary trait. In this article, Y is only required to be a random variable. A GTL in either case of Y, quantitative or qualitative, is abbreviated as GTL in this paper. (It is noted that in the quantitative cases, most authors call a trait locus QTL and in the binary trait cases call BTL.)

Let $\mathtt{A}_1$ and $\mathtt{A}_2$ be two markers used to flank a putative GTL $\mathtt{B}$. Denote the gamete recombination fraction between $\mathtt{A}_1$ and $\mathtt{A}_2$ by γ, the recombination fraction between $\mathtt{A}_1$ and $\mathtt{B}$ by r, and the recombination fraction between $\mathtt{B}$ and $\mathtt{A}_2$ by s. Typically and so in this paper, γ is assumed known. It is assumed that there is no interference so that γ and the unknown recombination parameters r and s are associated as $\gamma = r + s - 2rs$. Furthermore, as far as the efficiency of the use of markers for GTL detection is concerned, we can assume the two markers $\mathtt{A}_1$ and $\mathtt{A}_2$ are distinct and linked, i.e., $0 < \gamma < \frac{1}{2}$. Under this assumption there is only one of the two unknown recombination fraction parameters needed to consider, as the other can be implied, say s by $s = (\gamma - r)/(1 - 2r)$.

To describe the model underlying the interval mapping method for GTL detection, consider the backcross design first; the intercross case will be similar and described later in Section 4. Let the individuals of P_1 have the homozygous genotype A_1A_2/A_1A_2 at the markers $\mathtt{A}_1$ and $\mathtt{A}_2$, and those of P_2 have a_1a_2/a_1a_2. In the backcross population derived from a line-cross between P_1 and F_1, there are four different genotypes at the two markers, namely, $A_1A_2/A_1A_2, A_1A_2/A_1a_2, A_1A_2/a_1A_2$ and A_1A_2/a_1a_2, coded by 1, 2, 3, and 4. Let $q(j)$ be the probability that a randomly selected individual from the backcross population has the genotype $J = j$, $j = 1, 2, 3$ and 4. Direct calculation gives

$$q(1) = q(4) = (1 - \gamma)/2, \quad q(2) = q(3) = \gamma/2. \tag{1}$$

Furthermore, denote by $p_r(j)$ the conditional probability that a randomly selected individual has homozygous genotype, say BB, at the putative GTL, given that the individual has the genotype $J = j$ at the two markers. We know

$$\begin{cases} p_r(1) = 1 - p_r(4) = (1 - r)(1 - s)/(1 - \gamma), \\ p_r(2) = 1 - p_r(3) = (1 - r)s/\gamma, \end{cases} \tag{2}$$

with $s = (\gamma - r)/(1 - 2r)$ under the assumption of no interference. The conditional probability of heterozygous genotype Bb is $1 - p_r(j)$.

1.2. Finite regression mixture model

Since the genotype at the putative GTL is unobservable, the backcross population consists of two sub-populations, one being the individuals with homozygous genotype BB at the GTL and the other with heterozygous genotype Bb. (If the individuals of F_1 are backcrossed with the individuals of P_2 rather than P_1, then the backcross progeny have two sub-populations, one with homozygous genotype bb at the putative GTL and the other with heterozygous genotype Bb.) Let a randomly selected individual with the marker genotype $J = j$ give the response $Y = y$, associated with p random covariates, say $X = x \in R^p$. Throughout the paper, it is assumed that J and X are independent. If given $X = x$ the sub-population of Bb's has the conditional probability distribution function $g(y|x; \beta, \mu_1)$ and the other has $g(y|x; \beta, \mu_2)$, where $\beta \in R^{k_1}$ and $\mu_i \in R^{k_2}$, $i = 1, 2$, then the distribution of Y, given $J = j$ and $X = x$, is a finite mixture as follows:

$$f(y|j, x; r, \theta) = (1 - p_r(j))g(y|x; \beta, \mu_1) + p_r(j)g(y|x; \beta, \mu_2), \tag{3}$$

where $g(y|x; \beta, \mu)$ is a specific probability density function of y with respect to a σ-finite measure ν, for any x, β and μ. Here $p_r(j)$ is given in (2), $0 \leqslant r \leqslant \gamma$, and $\theta = (\beta, \mu_1, \mu_2) \in \Theta$, where Θ is an open subset of $R^{k_1+2k_2}$, not necessarily bounded. For convenience, put $k = k_1 + 2k_2$. When a random sample (y_i, j_i, x_i), $i = 1, \ldots, n$, of size n from the backcross population is observed, the log-likelihood function of (r, θ) for statistical inference is

$$l_n(r, \theta) = \sum_{i=1}^{n} \log\{(1 - p_r(j_i))g(y_i|x_i; \beta, \mu_1) + p_r(j_i)g(y_i|x_i; \beta, \mu_2)\}.$$

The maximum likelihood estimates (MLE) can then be obtained by solving the equation $\partial l_n(r, \theta)/\partial(r, \theta) = 0$; the LRT statistic can be used to test whether there is a GTL in the marker interval, i.e., to test $\mathrm{H}_0 : \mu_1 = \mu_2$. The thresholds (critical values) for the LRT are determined asymptotically. In this article, we investigate the large sample behavior of the likelihood-based procedures such as the MLE and LRT in the general regression model (3).

1.3. Remarks

First of all, we would like to remark that the finite mixture model (3) uses a structural component proportion via marker genotypes that is distinct from the common finite mixture models as considered in Chen and Chen (2001). As a result, the current model is identifiable in the proportion parameter r, while the finite mixture model in Chen and Chen (2001) is unidentifiable in