MAPPING QUANTITATIVE TRAIT LOCI IN HUMANS: SOME STATISTICAL CONTRIBUTIONS

SAURABH GHOSH

Thesis Submitted to the **Indian Statistical Institute**
in Partial Fulfilment of the Requirements for the Award of
the Degree of **Doctor of Philosophy**

Indian Statistical Institute
Calcutta

January 2000

(Revised)

Acknowledgements

Any expression of gratitude towards my supervisor Professor Partha Pratim Majumder would be an understatement. He has been a true friend, philosopher and guide to me, though not necessarily in that order. Having majored in statistics, the field of human genetics was a totally new world to me. Professor Majumder provided me a clear insight into the deep-rooted "linkage" beween statistics and human genetics. I am grateful to him for squeezing out sufficient time for me in the midst of his endless domestic and international commitments, and for patiently listening to my innumerable doubts and queries.

This thesis is a partial assurance to my sceptic parents who still believe that the only job which I can successfully accomplish is to mess up the job itself!

I am thankful to Professor T. Krishnan of the Applied Statistics Unit and Professor Probal Chaudhuri of the Theoretical Statistics and Mathematics Unit for useful statistical discussions. I am indebted to my friends Dr. Diganta Mukherjee and Mr. Analabha Basu for helping me in the preparation of the LaTeX output of this thesis. I am also thankful to Dr. Murari Mitra, Mr. Snigdhansu Bhusan Chatterjee, Mr. Anil Kumar Ghosh, the members of the Human Genetics laboratory of the Anthropology and Human Genetics Unit, the Reprography Unit, the Dean's Office and the Library of the Indian Statistical Institute for helping me during various phases of my tenure as a research fellow. The acknowlegements would remain incomplete if I fail to mention my juniors, who provided me with entertaining relief during my moments of monotony.

Finally I am grateful to many anonymous reviewers, including those of *Genetic Epidemiology*, *American Journal of Human Genetics* and

Advances in Genetics for providing useful comments on chapters and manuscripts that have been accepted for publication in these journals, which have helped me to substantially improve the presentation.

Calcutta, January 2000 Saurabh Ghosh

Contents

Chapter 1

Introduction and Overview of Past Studies

Many quantitative traits — such as, milk yield in cows, blood pressure in humans — are known to be determined primarily, though not exclusively, by inherited genetic factors. It is thus of considerable importance to identify chromosomal locations of the genes that control a quantitative character. Linkage analysis (Ott 1999), which deals with the detection of linkage and estimation of recombination fractions among the loci controlling a qualitative/quantitative character and marker loci whose positions are known *a priori*, is widely used for localization of genes. Although statistical methodologies for mapping genes determining dichotomous qualitative characters in humans are well-developed, the development of such methodologies — especially those that are statistically and computationally efficient — for human quantitative traits is an active area of current research in human genetics. It has been emphasized that many traits that have traditionally been treated as qualitative are inherently quantitative in nature.

Although the idea of mapping quantitative trait loci (QTL mapping) can be traced back to Sax (1923), who studied the nature of association of seed size with seed-coat pattern and pigmentation in beans, the recent development of dense maps of highly polymorphic DNA markers in plants and animals has resulted in a resurgence of interest in QTL mapping. Statistical linkage analysis relies on the nature and extent of co-inheritance of alleles at the trait and marker loci. For many plants and animals experimental

crosses can be set up such that the trait locus genotype of an offspring can be unambiguously inferred. This simplifies the statistical investigation of co-inheritance of alleles at the trait and marker loci. However, it is not possible to set up experimental crosses for humans. Hence, QTL mapping in humans is statistically more difficult than in experimental plants and animals.

In this Chapter, we provide an overview, albeit non-exhaustive, of the different statistical procedures that have been developed for QTL mapping. A quantitative trait (Y) can be modelled in a general way as $Y = G + E$, where G and E are the genetic and environmental contributions to the phenotype, respectively. While this general form of the model can be used in an exploratory way to provide some broad statistical inferences about the quantitative trait, such as heretability of the trait, for making specific inferences or for QTL mapping, it is necessary to formulate a more detailed model. Often models are formulated on the basis of exploratory data analyses.

A quantitative trait may be determined, in addition to an environmental component whose expectation is usually assumed to be zero, by one or more loci, each biallelic or multiallelic, linked or unlinked. There may be dominance effects at various loci, and unlinked loci may also interact epistatically in the determination of the trait values.

For a quantitative trait that is determined by a single biallelic locus, a general model is: $Y|A_1A_1 \sim f_1(\mu_1, {\sigma_1}^2)$, $Y|A_1a_1 \sim f_2(\mu_2, {\sigma_2}^2)$ and $Y|a_1a_1 \sim f_3(\mu_3, {\sigma_3}^2)$, where A_1 and a_1 are the two alleles at the locus, and f_1, f_2 and f_3 are general probability distribution functions with means μ_is and variances ${\sigma_i}^2$s. If allelic effects are assumed to be additive, and there is no dominance, then $\mu_1 = 2\alpha$, $\mu_2 = \alpha + \beta$ and $\mu_3 = 2\beta$, where α and β are the allelic effects of A_1 and a_1, respectively. In the presence of a dominance effect, δ, $\mu_1 = 2\alpha$, $\mu_2 = \alpha + \beta + \delta$ and $\mu_3 = 2\beta$. The above three-parameter model can sometimes, but not always, be reduced to a two-parameter model by appropriate scaling (Mather and Jinks 1982, chapter 4), thereby greatly simplifying statistical treatment. Therefore, the popular two-parameter statistical model is (Haseman and Elston 1972, eqn. 2; Hill 1975, p. 439; Amos and Elston 1989, p. 351; Amos et al. 1989, eqn. 2; Tiwari and Elston 1997, p. 254; Olson 199, p. 2296; Schork et al. 2000): $Y|A_1A_1 \sim f_1(\alpha, \sigma^2)$, $Y|A_1a_1 \sim f_2(\beta, \sigma^2)$ and $Y|a_1a_1 \sim f_3(-\alpha, \sigma^2)$. Often f_1, f_2 and f_3 are assumed to be $N(.,.)$.

We first discuss some of the statistical methods developed for QTL mapping in plants and animals. Jayakar (1970) developed a variance components method for estimating the recombination fraction between a trait locus and a marker locus separately for backcross and intercross parental matings. The idea was to express the recombination fraction as a function of the within and between haplotype-class variances of the quantitative trait. Obviously, this method assumed knowledge of linkage-phase and unambiguous determination of haplotypes of individuals. Weller (1986) presented a maximum likelihood approach of estimating the recombination fraction parameter from a mixture distribution of quantitative trait values within each marker genotype based on data on F_2 progeny of a cross between two species of tomatoes. However, the computational algorithm is quite complicated and it is not possible to assess the efficiency of the relevant estimates. Lander and Botstein (1989) developed a multipoint likelihood-based linkage analysis using data on maize. They used the model:

$$y_j = b_0 + b^* x_j^* + e_j,$$

where y_j is the quantitative trait and x_j^* is an indicator variable depending on the genotypes of flanking markers. The test for linkage is a likelihood ratio test of $b^* = 0$. Haley and Knott (1992) proposed a linear regression of the quantitative trait on an indicator variable depending on the flanking markers. They showed that nearly all the useful information about the QTL are contained in the marker class means. Jansen (1993) proposed a likelihood technique of interval mapping of multiple QTLs, where QTLs were fitted one at a time using backward step-regression. The marker information were used as cofactors in order to eliminate the effects of the other QTLs. Zeng (1994) improved on Lander and Botstein's (1989) method by incorporating a number of non-flanking markers in the model given by:

$$y_j = b_0 + b^* x_j^* + \sum_{k \in \overline{F}} b_k x_k + e_j,$$

where x_k is an indicator variable depending on the genotype of a non-flanking marker and $\overline{F}$ is the set of non-flanking markers . The likelihood ratio test statistic within a given marker interval is unaffected by QTLs outside that interval. Whittaker et al. (1996) showed that in F_2 and backcross

populations, the regression of phenotype on marker genotypes is equivalent to the regression procedures of Haley and Knott (1992) and Zeng (1994). They used the linear model:

$$E(Y) = \beta_0 + \beta_1 E(h|x_i x_{i+1}, r_L) + \sum_{j \in F} b_j x_j,$$

where x_i, x_{i+1} are variables taking values 1, 0, -1 depending on the flanking marker genotypes and h is a variable, also taking values 1, 0, -1 depending on the QTL genotype. Visscher et al. (1996) presented a confidence interval approach for locating a QTL via a bootstrap technique and showed, using simulations, that though the confidence intervals were slightly conservatively biased, the proportion of simulation runs in which the true location of the QTL was contained in the confidence interval was very close to its expected value. Several studies of QTL mapping have been performed using inbred strains in mouse (Berrethini 1993, Fijneman 1995, Schork et al. 1995, van Wezel et al. 1996, Chang 1999). Similar linkage studies have been undertaken using data on flowering plants (Coupland 1995), soybean (Lark et al. 1995), tomatoes (Eshed and Zamir 1996), dairy cattle (Georges et al. 1995) and general experimental organisms (Schork et al. 1996). Tanskley (1993) provides an insight into the development of QTL mapping techniques in plants and animals with the advent of more and more DNA markers. However, as pointed out in the preceding paragraph and discussed in Chapter 4, the major limitation of extending these statistical methods to human linkage analysis is the non-availability of trait locus genotype and haplotype data for humans. Although Goldgar (1990) used multipoint human data on multifactorial traits, the effects of the QTL and additive genetic variance were confounded.

We next provide a detailed overview of various statistical methodologies developed for mapping QTLs in humans. One of the most popular approaches of analyzing human linkage data is based on sib-pairs. Some of the earliest contributions in these studies were made by Penrose. He assessed the efficiency of using concordant and discordant sib-pairs (in terms of quantitative trait values) in studying multifactorial disorders (1937). It was shown by Penrose (1947) based on a linkage study between the loci for phenylketonuria and the presence or absence of the B allele at the ABO

locus, that the efficiency and complexity of detection and estimation of linkage can be increased by distinguishing the two types of identical sib-pairs. Penrose (1953) extended his earlier methods to multiple alleles using data on red-hair and ABO locus restricted to a single generation. An extensive review of Penrose's contributions and the subsequent extensions to QTL mapping procedures using sib-pairs is presented in Edwards (1998).

A model-free linkage method is to utilise the inverse relationship between the difference between trait values of sib-pairs and their marker identity-by-descent (i.b.d.) scores. A pair of related individuals shares an allele i.b.d. if that allele has a common ancestral source. For sib-pairs, the common ancestor are their parents. Haseman and Elston (1972) developed a regression approach of QTL mapping based on squared difference in quantitative trait values of sib-pairs (Y) and their estimated marker i.b.d. scores ($\widehat{\pi}_m$). The basis of the regression is the equation:

$$E(Y|\widehat{\pi}_m) = \alpha + \beta\widehat{\pi}_m, \tag{1.1}$$

where there is no dominance in the trait and β is a one-to-one function of the recombination fraction, θ, between the QTL and the marker locus. The test for no linkage (i.e., $\theta = 0.5$) is equivalent to testing $\beta = 0$ in Equation (1.1). Amos and Elston (1989) extended the above regression procedure to other relative pairs. For each type of relative pair, the regression parameter β is a different function of θ. However, the test for no linkage in each case is equivalent to testing $\beta = 0$. Amos et al. (1989) showed that in presence of dominance in the trait, the least squares estimator of β is biased. They derived the conditional variance of Y given $\widehat{\pi}_m$ as $\alpha_0 + \beta_0\widehat{\pi}_m + \gamma_0\widehat{\pi}_m^2$. The test for linkage is based on the weighted least squares estimators of β_0 and γ_0, and is more powerful than the original Haseman-Elston test (1972). Olson and Wijsman (1993) used generalised estimating equations to combine information from different types of relative pairs in a set of pedigree data. The test for no linkage between the QTL and the marker locus is equivalent to testing $\beta = 0$ where β is the vector of regression coefficients of Ys on $\widehat{\pi}_m$s corresponding to the different types of relative pairs. The test statistic is of the form $\sqrt{N}c'\widehat{\beta}/\{c'\widehat{Var}(\widehat{\beta})c\}^{1/2}$, where c is a vector of weights chosen proportional to $\{\widehat{Var}(\widehat{\beta})\}^{-1}\widehat{\beta}$. Fulker and Cardon (1994) extended the Haseman-Elston (1972) regression equation to the multipoint

case. They proposed an interval mapping method where the i.b.d. scores at the flanking markers (π_{m1} and π_{m2}) are estimated separately using marginal marker information and the trait i.b.d. score (π_t) is estimated using the equation :

$$\widehat{\pi}_t \;=\; \rho_0 \;+\; \rho_1\widehat{\pi}_{m1} \;+\; \rho_2\widehat{\pi}_{m2}.$$

Y is regressed on $\widehat{\pi}_t$ and the approximate position of the QTL is inferred based on the plot of $\widehat{\beta}/\widehat{s.e.}(\widehat{\beta})$, where $\widehat{\beta}$ is the regression estimator of Y on $\widehat{\pi}_t$. Olson (1995a) suggested that in order to obtain maximum information, the marker i.b.d. scores be jointly estimated using all available marker data. She obtained the regression equation:

$$E(Y|\widehat{\pi}_{m1},\widehat{\pi}_{m2}) \;=\; \beta_0 \;+\; \beta_1\widehat{\pi}_{m1} \;+\; \beta_2\widehat{\pi}_{m2},$$

where there is no dominance in the trait loci. Olson (1995b) extended this multipoint regression approach for dichotomous traits using affected sib-pairs. The likelihood function used was:

$$P(\widehat{\pi}_{m1},\widehat{\pi}_{m2}|A) = \frac{P(\widehat{\pi}_{m1},\widehat{\pi}_{m2})}{P(A)}\{\alpha+\beta E(\pi_t|\widehat{\pi}_{m1},\widehat{\pi}_{m2})+\gamma P(\pi_t=\frac{1}{2}|\widehat{\pi}_{m1},\widehat{\pi}_{m2})\},$$

where A is the number of affected sibs in a sib-pair. Tiwari and Elston (1997) extended the original Haseman-Elston (1972) procedure to the case of two unlinked QTLs which might interact epistatically. They showed that under a fairly general model of epistasis, where they assumed that the marginal genotypic effects of the QTLs as well as those of the epistatic interactions are additive, the expectation of Y is a linear function of $\widehat{\pi}_{m1},\widehat{\pi}_{m2},f_1,f_2$ and their pairwise cross-product terms, where f_1 and f_2 are the probabilities that a sib-pair shares 1 and 2 alleles i.b.d., respectively. A relative comparison of the sib-pair based linkage tests and the sampling strategies with respect to affected sibs are presented in Blackwelder and Elston (1985) and Allison et al. (1999). Alcais and Abel (1999) have recently developed a maximum-likelihood- binomial method of mapping QTLs using sibship data. The idea is to introduce a latent binary variable Z which captures linkage information between the QTL and the marker locus. The likelihood is formulated in terms of $P(M_1,M_2|Y) \;=\; \sum_Z P(Z|Y)P(M_1,M_2|Z)$, where Y is the observed phenotype and M_1,M_2 are the alleles at the marker locus. $P(Z|Y)$ is

modelled by a probit distribution and $P(M_1, M_2|Z)$ by a Bernoulli distribution. The test for linkage is based on a likelihood ratio test of the Bernoulli parameter = 0.5.

Parametric methods for mapping QTLs involve parametric models, and thus, are often susceptible to minor distributional assumptions. Some of the non-parametric methods in current use are relatively more robust. Haseman and Elston (1972) proposed a test statistic based on the rank correlation between the absolute differences in trait values of sib-pairs and their estimated marker i.b.d. scores. Kruglyak and Lander (1995a) proposed a Wilcoxon rank sum test based on ranks of squared differences in sib-pair trait values and an indicator variable depending on the marker genotype. A detailed discussion on some of the parametric and non-parametric multipoint sib-pair linkage approaches, which have been implemented in a computer package MAPMAKER/SIBS, is presented in Kruglyak and Lander (1995b).

A recently developed approach has been motivated by the classical LOD score technique used for mapping qualitative trait loci. Page et al. (1998) have proposed a QLOD score statistic for detecting linkage in QTLs, where the traditional critical values of 3 and -2 for the underlying sequential tests were used. Another popular statistical approach for QTL mapping is to assume a random effects model for the quantitative trait (Hill 1975), where sibship data were considered. The model is given by:

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk},$$

where y_{ijk} denotes the quantitative trait, μ, the overall mean, α_i, the sibship effect, β_{ij}, the marker effect nested within sibship and e_{ijk}, the random error. Usual variance component analyses based on σ_β^2, the variance of β_{ij}, have been shown to be equivalent to testing of hypotheses on linkage parameters.

A few mapping techniques of QTLs in humans are based on pedigree data. Almasy and Blangero (1998) have recently proposed a variance components multipoint linkage method using pedigrees of arbitrary size. They developed a general framework of multipoint i.b.d. probability calculations. The correlations in i.b.d. scores were shown to be a function of the chromosomal distances for different relative pairs in a general pedigree. The variance components method considers likelihood of the entire pedigree and has been implemented in a computer package SOLAR. There have been a

few comparative studies between the different statistical techniques for QTL mapping in humans. Lander and Schork (1994) have presented a review of the quantitative linkage studies for the genetic dissection of complex traits and have compared them with other techniques like allele-sharing methods and association studies. A relative comparison between the variance components and sib-pair based linkage methods has been examined recently by Williams and Blangero (1999), where they observed that within single sib-pair and sibship sampling units, the variance components approach gave consistently superior power and efficiency of parameter estimation.

One of the major current challenges in genetic epidemiology is to unravel genetic architectures of complex traits. Quantitative variables, possibly correlated, generally underlie complex traits. Many models and approaches have been developed, including variance components (Lange and Boehnke 1983, Schork 1993), regressive model (Bonney et al. 1988, Moldin and van Eerdewegh 1995), multivariate extension of the Haseman-Elston model (Amos et al. 1990, Amos and Liang 1996) and structural equations model (Eaves et al. 1996, Todorov et al. 1998) to jointly analyze data on several correlated quantitative phenotypes as a single multivariate phenotype. However, the power of a multivariate analysis to detect linkage can be substantially low (Ott and Rabinowitz 1999). Data reduction techniques, such as principal components analysis or factor analysis, (Zlotnik et al. 1983, Hasstedt et al. 1994, Boomsma 1996, Allison and Beasley 1998, Ott and Rabinowitz 1999) help in circumventing this problem of reduced power.

The aim of the present thesis is to propose some computationally simple and statistically efficient QTL mapping techniques based on different types of data. In Chapter 3, we develop a two-stage Expectation-Maximization technique of mapping QTL when data on trait values of both parents and siblings as well as their genotypes at several marker loci are available. In Chapter 4, we discuss two mapping procedures when the quantitative trait is controlled by multiple loci which might interact epistatically. In Chapter 5, we propose a two-stage semi-parametric method of mapping QTLs based on genome-wide scan using sib-pair data. In Chapter 6, we develop a procedure of deciphering the genetic architecture of a multivariate phenotype using the semi-parametric method proposed in Chapter 5. We use simulated data to assess the performance of our proposed procedures.

Chapter 2

Models and General Statistical Methodology

2.1 Genetics of a Quantitative Trait: Models Considered

In this Chapter, we describe the different models underlying the determination of a quantitative trait, that are considered by us. In all our models, we assume that the underlying population is in Hardy-Weinberg equilibrium with respect to each of the quantitative trait loci as well as the marker loci considered.

We assume the classical model for a quantitative trait Y controlled by an autosomal, biallelic locus, which is given in terms of the expectation and variance of Y given the trait genotype.

Suppose A_1 and a_1 denote the alleles at the trait locus. Then,

$$\begin{aligned} E(Y|A_1A_1) &= \alpha,\ \mathrm{Var}(Y|A_1A_1) = \sigma^2; \\ E(Y|A_1a_1) &= \beta,\ \mathrm{Var}(Y|A_1a_1) = \sigma^2; \\ E(Y|a_1a_1) &= -\alpha,\ \mathrm{Var}(Y|a_1a_1) = \sigma^2; \end{aligned}$$

where σ^2 includes the environmental variance and β signifies the dominance parameter at the trait locus. As mentioned in the previous Chapter, in spite of its simplicity, this model continues to be used in contemporary QTL analyses (Tiwari and Elston 1997, Olson 1999, Schork et al. 2000). We

consider models both with and without dominance at the major trait locus. In Chapter 3, we assume that each of the conditional distributions of trait (Y) values given the genotypes is Normal with the parameters as specified above. However, such parametric distributional assumptions are not made in the models considered in other Chapters.

For a trait determined by multiple loci, we assume additivity of locus-specific marginal effects. Suppose the quantitative trait Y is controlled by L autosomal, biallelic loci $(A_1, a_1), (A_2, a_2), \ldots, (A_L, a_L)$. Let the expectation, $E(Y)$, of the quantitative character, Y, given the genotypes of the l^{th} locus be α_l, β_l and $-\alpha_l$ for A_lA_l, A_la_l and a_la_l, respectively. We assume that the variance of X within each single-locus genotype is the same. For the l^{th} locus, let this variance be denoted as ${\sigma_l}^2$. In the absence of epistatic interactions, effects of the loci on $E(X)$ are assumed to be additive. Thus for example, $E(Y|A_1A_1A_2A_2 \ldots A_LA_L) = \sum_{1=1}^{L} \alpha_i$; $E(Y|A_1A_1A_2A_2 \ldots A_{L-1}A_{L-1}a_La_L) = \sum_{i=1}^{L-1} \alpha_i - \alpha_L$, etc., when epistatic interactions are absent. Since the trait loci are assumed to be unlinked, the variance of the trait among individuals of any multilocus genotype is the same, ${\sigma}^2 = {\sigma_G}^2 + {\sigma_E}^2$, where ${\sigma_G}^2 = \sum_{l=1}^{L} {\sigma_l}^2$ and ${\sigma_E}^2 =$ environmental variance.

We also consider the possibility of epistatic interactions among the quantitative trait loci. We use a simple model of additive epistatic interaction at the different trait loci. This model is prompted by experimental observations on some plants and animals (Chang et al. 1999), and has been termed as the digenic interaction model (Kearsey and Pooni 1996) when $L = 2$. We assume that there are epistatic interactions only among homozygotes between pairs of loci. Between loci i and j $(i \neq j = 1, 2, \ldots, L)$, epistatic interaction effects are assumed to be: for $A_iA_iA_jA_j$ and $a_ia_ia_ja_j$ the effect is Δ_{ij}, for $A_iA_ia_ja_j$ and $a_ia_iA_jA_j$ the effect is $-\Delta_{ij}$; the effect for all other two-locus genotypes is 0. Thus, for example, under additive epistatic effects,

$$E(Y|A_1A_1A_2A_2 \ldots A_LA_L) = \sum_{i=1}^{L} \alpha_i + \sum_{i=1}^{L} \sum_{j>i}^{L} \Delta_{ij};$$

$$E(Y|A_1A_1A_2A_2 \ldots A_{L-1}A_{L-1}a_La_L) = \sum_{i=1}^{L-1} \alpha_i - \alpha_L + \sum_{i=1}^{L-1} \sum_{j>i}^{L-1} \Delta_{ij} - \sum_{i=1}^{L-1} \Delta_{iL},$$

etc. For clarity, and to fix ideas, we provide, in Table 2.1, the genotypes (G), their population frequencies and expectation of the quantitative trait

given genotype $[E(Y|G)]$, for $L = 2$ and in the absence of dominance (i.e., $\beta_1 = \beta_2 = 0$) at both the trait loci. When we consider models with two QTLs, we suppress the suffix $_{12}$ and donote Δ_{12} as Δ.

We assume that the QTL is in linkage equilibrium with an autosomal, biallelic and codominant marker locus with alleles M_1 and m_1. The recombination fraction between the QTL and the marker locus is denoted by θ.

Table 2.1. Genotypes (G) of individuals at two autosomal, unlinked, epistatically interacting biallelic loci, relative frequencies of these genotypes in the population and expected values of quantitative trait Y corresponding to these genotypes under the digenic interaction model

G	Relative Frequency	$E(Y\|G)$
$A_1A_1A_2A_2$	${p_1}^2{p_2}^2$	$\alpha_1 + \alpha_2 + \Delta_{12}$
$A_1A_1A_2a_2$	$2{p_1}^2p_2q_2$	α_1
$A_1A_1a_2a_2$	${p_1}^2{q_2}^2$	$\alpha_1 - \alpha_2 - \Delta_{12}$
$A_1a_1A_2A_2$	$2p_1{p_2}^2q_1$	α_2
$A_1a_1A_2a_2$	$4p_1p_2q_1q_2$	0
$A_1a_1a_2a_2$	$2p_1q_1{q_2}^2$	$-\alpha_2$
$a_1a_1A_2A_2$	${p_2}^2{q_1}^2$	$-\alpha_1 + \alpha_2 - \Delta_{12}$
$a_1a_1A_2a_2$	$2p_2{q_1}^2q_2$	$-\alpha_1$
$a_1a_1a_2a_2$	${q_1}^2{q_2}^2$	$-\alpha_1 - \alpha_2 + \Delta_{12}$

2.2 Simulation Methods

In order to assess the efficiencies of our proposed statistical methodologies, we use Monte-Carlo simulations. Our simulation data comprise one of two scenarios: data on nuclear families (i.e., both parental as well as offspring data) or only sib-pair data.

In this section, we provide an outline of our simulation strategies. Specific details are provided in relevant sections of later chapters, wherever necessary.

For two-point mapping based on nuclear family data, it is well-known

(Ott 1999) that at least one of the two parents must be a double heterozygote (at the QT and marker loci) for the family to be informative for linkage. Because it is not generally possible to ensure heterozygosity at the QTL in humans, a prudent strategy is to sample families, genotype the parents at the marker locus, and to proceed to genotype offspring only if at least one of the parents is heterozygous at the marker locus. This obviously provides considerable saving of genotyping costs and efforts. Our simulations are based on this familiar strategy, and we refer to such families in which at least one of the two parents is heterozygous at the marker locus as "informative", although such families, strictly speaking, are informative for linkage only in a restricted sense (because heterozygosity at the QTL is not ensured). We, therefore, present our results only as functions of informative families (in the restricted sense as defined above), since it is known that the prudent strategy outlined above is generally used in practice, although not always explicitly mentioned in publications. Of course, from our results presented later, it is easy to compute the expected number of randomly-sampled families to be sampled to obtain a given number of informative families (as defined above). The probability that a randomly sampled family is informative is $\lambda = 1 - \{\pi^2 + (1-\pi)^2\}$, where π is the frequency of marker allele M_1 at a biallelic marker locus (M_1, m_1). Hence, the expected number of families, N, required to be sampled to obtain n informative families is $N = n/\lambda$. In Chapter 3, we show that analyses based on data generated on an expected number of nuclear families yield estimates with properties identical to those obtained from data where only informative families are generated.

For multipoint mapping, a family in which both parents are homozygous at all marker loci is obviously not informative for multipoint linkage analysis. Therefore, for multipoint linkage analysis, we define "informative" families, in the restricted sense, as those in which at least one of the parents is heterozygous at least at one of the marker loci The probability of an informative family, in this case, is $1 - \prod_{i=1}^{k} \{\pi_i{}^2 + (1-\pi_i)^2\}^2$, where there are k marker loci with alleles (M_1, m_1), $(M_2, m_2), \ldots, (M_k, m_k)$, respectively, and π_i is the allele frequency of M_i, $i = 1, 2, \ldots, k$. Thus, results based on a fixed number of informative families can be easily interpreted in terms of an expected number of nuclear families.

Chapters 3 and 4 deal with linkage analyses, two-point (Chapters 3 and

4) and multipoint (Chapter 3), based on nuclear family data. In these two Chapters, we have generated simulated data on "informative" (restricted-sense) nuclear families using the methodology outlined below. Since we did not generate a random sample of nuclear families, we did not use the conventional "gene-dropping" algorithm in our simulations. However, to provide evidence thast estimates of parameters based on our simulated data of informative families were identical with those obtained on the basis of appropriate expected number of nuclear families (see Table 3.12, Chapter 3), we have also used the "gene-dropping" algorithm as required for these comparisons.

Our simulation algorithm of generating informative nuclear families was as follows. Suppose the allele frequencies of A and a (i.e., the alleles of the trait locus) are p and q respectively. Since we assume the populations to be in Hardy-Weinberg equilibrium, the trait genotype of parents is generated using a trinomial random number generator with cell probabilities p^2, $2pq$ and q^2, corresponding to the trait genotypes A_1A_1, A_1a_1 and a_1a_1 respectively. The trait values of parents as well as offspring is generated from appropriate Normal distributions with means equal to the conditional expectation given the trait genotype as described in the previous Section. As the marker locus is assumed to be biallelic with alleles M_1 and m_1, the marker genotypes of parents are determined using a trinomial random number generator with cell probabilities equal to the probabilities of the three possible genotypes M_1M_1, M_1m_1 and m_1m_1. The marker genotype of an offspring is generated by sampling either from a binomial distribution with success probability $1/2$ for a parental mating in which one parent is homozygous and the other parent is heterozygous at the marker locus (backcross) in which case there are two possible genotypes with equal probabilities, or from a trinomial distribution with cell probabilities $(1/4, 1/2, 1/4)$ for a parental mating in which both parents are heterozygous at the marker locus (intercross), in which case the offspring can be one of two possible homozygotes with probabilities $1/4$ each or a heterozygote with probability $1/2$. Based on the conditional probabilities of offspring trait genotypes given parental trait and mating type as provided in Tables 2.2 (backcross) and 2.3 (intercross), we generate the genotype of the offspring with respect to the trait locus using a trinomial random number generator with cell probabilities equal to

the three possible genotypes A_1A_1, A_1a_1 and a_1a_1.

In chapters 4, 5 and 6, QTL mapping procedures based on data of sib-pairs are discussed. Unlike the case of nuclear families, sib-pair data are generally unselected, that is, all data that are collected are used in analysis. For generating simulated data on sib-pairs, we have used the following algorithm. We first generate the trait identity-by-descent (i.b.d.) scores (π_t), which can assume three possible values $0, 1/2$ or 1, from a trinomial random number generator with cell probabilities $(1/4, 1/2, 1/4)$. The trait genotypes of the sib-pairs (G) are generated from the conditional distribution of sib-pair trait genotype given the trait i.b.d. score as provided in Table 2.4. The trait values of the sib-pairs are generated from a bivariate normal distribution with appropriate mean vectors and dispersion matrices conditioned on the trait genotypes as described in the previous Section. The marker i.b.d. scores of the sib-pairs (π_m), which can assume three possible values $0, 1/2$ or 1, are generated from the conditional distribution of the marker i.b.d. score given the trait i.b.d. score as provided in Table 2.5, using a trinomial random number generator. The estimated marker i.b.d. scores of the sib-pairs ($\widehat{\pi}_m$), which can assume five possible values $0, 1/4, 1/2, 3/4$ or 1 when parental genotype information are available, are generated from the conditional distribution of estimated i.b.d. score given the actual i.b.d. score as provided in Table 2.6, using a 5-nomial random number generator. We note that Tables 2.4, 2.5 and 2.6 are provided in Haseman and Elston (1972). When we consider interval mapping procedures, we generate the marker information sequentially. First the i.b.d. scores at the two markers flanking the trait locus are conditional on the trait i.b.d. score using the trinomial distribution given in Table 2.5, and then the i.b.d. score of each non-flanking marker is generated sequentially conditional on the generated i.b.d. score of the marker flanking it using the same trinomial distribution. We generate multivariate phenotypes from appropriate multivariate normal distributions with parameters depending on the correlation structure of the phenotypes as discussed in Chapter 6.

We assess our proposed procedures in terms of empirical histograms, mean and variance of our estimated recombination fractions, empirical confidence intervals of recombination fractions, power of our tests of linkage, sample size requirement to detect linkage and proportion of correctly iden-

tifying the true interval locations of the QTLs. We also compare our procedures with some existing mapping approaches.

Table 2.2. Trait locus mating types among $M_1M_1 \times M_1m_1$ (backcross) parents, mating probabilities and probabilities of trait locus genotypes among offspring with marker genotype M_1M_1 [1]

g	Mating Type	Probability	π_g		
			A_1A_1	A_1a_1	a_1a_1
1	$A_1A_1 \times A_1A_1$	${p_1}^4$	$\frac{1}{2}$	0	0
2	$A_1A_1 \times A_1a_1$	${p_1}^3p_2$	$\frac{1}{2}(1-\theta)$	$\frac{1}{2}\theta$	0
3	$A_1A_1 \times a_1A_1$	${p_1}^3p_2$	$\frac{1}{2}\theta$	$\frac{1}{2}(1-\theta)$	0
4	$A_1A_1 \times a_1a_1$	$2{p_1}^2{p_2}^2$	0	$\frac{1}{2}$	0
	$a_1a_1 \times A_1A_1$				
5	$A_1a_1 \times A_1A_1$	$2{p_1}^3p_2$	$\frac{1}{4}$	$\frac{1}{4}$	0
	$a_1A_1 \times A_1A_1$				
6	$A_1a_1 \times A_1a_1$	$2{p_1}^2{p_2}^2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}$	$\frac{1}{4}\theta$
	$a_1A_1 \times A_1a_1$				
7	$A_1a_1 \times a_1A_1$	$2{p_1}^2{p_2}^2$	$\frac{1}{4}\theta$	$\frac{1}{4}$	$\frac{1}{4}(1-\theta)$
	$a_1A_1 \times a_1A_1$				
8	$A_1a_1 \times a_1a_1$	$2p_1{p_2}^3$	0	$\frac{1}{4}$	$\frac{1}{4}$
	$a_1A_1 \times a_1a_1$				
9	$a_1a_1 \times A_1a_1$	$p_1{p_2}^3$	0	$\frac{1}{2}(1-\theta)$	$\frac{1}{2}\theta$
10	$a_1a_1 \times a_1A_1$	$p_1{p_2}^3$	0	$\frac{1}{2}\theta$	$\frac{1}{2}(1-\theta)$
11	$a_1a_1 \times a_1a_1$	${p_2}^4$	0	0	$\frac{1}{2}$

[1] Probabilites of trait locus genotypes among offspring with marker genotype M_1m_1 can be obtained by replacing θ by $(1-\theta)$ in this table.

Table 2.3. Trait locus mating types among $M_1m_1 \times M_1m_1$ (intercross) parents, mating probabilities and probabilities of trait locus genotypes among offspring with marker genotype M_1M_1 and M_1m_1[2]

g	Mating Type	Probability	$\pi_g(M_1M_1)$			$\pi_g(M_1m_1)$		
			A_1A_1	A_1a_1	a_1a_1	A_1A_1	A_1a_1	a_1a_1
1	$A_1A_1 \times A_1A_1$	${p_1}^4$	$\frac{1}{4}$	0	0	$\frac{1}{2}$	0	0
2	$A_1A_1 \times A_1a_1$	$2{p_1}^3p_2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}$	$\frac{1}{4}$	0
	$A_1a_1 \times A_1A_1$							
3	$A_1A_1 \times a_1A_1$	$2{p_1}^3p_2$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}$	$\frac{1}{4}$	0
	$a_1A_1 \times A_1A_1$							
4	$A_1A_1 \times a_1a_1$	${p_1}^2{p_2}^2$	0	$\frac{1}{4}$	0	0	$\frac{1}{2}$	0
	$a_1a_1 \times A_1A_1$							
5	$A_1a_1 \times A_1a_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}[1-2\theta(1-\theta)]$	$\frac{1}{2}\theta(1-\theta)$
6	$A_1a_1 \times a_1A_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$
	$a_1A_1 \times A_1a_1$							
7	$a_1a_1 \times A_1a_1$	$2p_1{p_2}^3$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}$	$\frac{1}{4}$
	$A_1a_1 \times a_1a_1$							
8	$a_1A_1 \times a_1A_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}[1-2\theta(1-\theta)]$	$\frac{1}{2}\theta(1-\theta)$
9	$a_1a_1 \times a_1A_1$	$2p_1{p_2}^3$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}$	$\frac{1}{4}$
	$a_1A_1 \times a_1a_1$							
10	$a_1a_1 \times a_1a_1$	${p_2}^4$	0	0	$\frac{1}{4}$	0	0	$\frac{1}{2}$

[2] Probabilites of trait locus genotypes among offspring with marker genotype m_1m_1 can be obtained by replacing θ by $(1-\theta)$ in the block corresponding to the genotype M_1M_1 in this table.

Table 2.4. Conditional Distribution of trait i.b.d. score (π_l) given trait genotypes of sib-pair (G)

G	Conditional Prob. of trait i.b.d. score		
	$\pi_l = 0$	$\pi_l = \frac{1}{2}$	$\pi_l = 1$
$A_1A_1 - A_1A_1$	p^4	p^3	p^2
$a_1a_1 - a_1a_1$	q^4	q^3	q^2
$A_1a_1 - A_1a_1$	$4p^2q^2$	pq	$2pq$
$A_1A_1 - A_1a_1$	$2p^3q$	p^2q	0
$A_1a_1 - A_1A_1$	$2p^3q$	p^2	0
$A_1a_1 - a_1a_1$	$2pq^3$	pq^2	0
$a_1a_1 - A_1a_1$	$2pq^3$	pq^2	0
$A_1A_1 - a_1a_1$	p^2q^2	0	0
$a_1a_1 - A_1A_1$	p^2q^2	0	0

Table 2.5. Joint distribution of trait i.b.d. score (π_l) and marker i.b.d. score (π_m) [3]

π_l	π_m			Total
	0	$\frac{1}{2}$	1	
0	$\frac{\psi^2}{4}$	$\frac{\psi(1-\psi)}{2}$	$\frac{(1-\psi)^2}{4}$	$\frac{1}{4}$
$\frac{1}{2}$	$\frac{\psi(1-\psi)}{2}$	$\frac{(1-2\psi+2\psi^2)}{2}$	$\frac{\psi(1-\psi)}{2}$	$\frac{1}{2}$
0	$\frac{(1-\psi)^2}{4}$	$\frac{\psi(1-\psi)}{2}$	$\frac{\psi^2}{4}$	$\frac{1}{4}$
Total	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

Table 2.6. Joint distribution of marker i.b.d. score (π_m) and estimated marker i.b.d. score ($\widehat{\pi}_m$) in the case of no dominance in the QTL and complete parental information

$\widehat{\pi}_m$	π_m			Total
	0	$\frac{1}{2}$	1	
0	$\frac{p^2q^2}{2}$	0	0	$\frac{p^2q^2}{2}$
$\frac{1}{4}$	$p^3q + pq^3$	$p^3q + pq^3$	0	$2(p^3q + pq^3)$
$\frac{1}{2}$	$\frac{p^4+4p^3q^2+q^4}{4}$	$\frac{p^4+6p^2q^2+q^4}{2}$	$\frac{p^4+4p^2q^2+q^4}{4}$	$p^4 + 5p^2q^2 + q^4$
$\frac{3}{4}$	0	$p^3q + pq^3$	$p^3q + pq^3$	$2(p^3q + pq^3)$
1	0	0	$\frac{p^2q^2}{2}$	$\frac{p^2q^2}{2}$
Total	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

[3] $\psi = \theta^2 + (1-\theta)^2$

Chapter 3

Mapping Quantitative Trait Loci via the EM Algorithm and Bayesian Classification

3.1 Introduction

As mentioned in Chapter 1, many currently used QTL mapping methods, especially those that have been developed in the context of plant genetics or genetics of inbred animals, assume knowledge of linkage phase in individuals which imposes a severe restriction on the applicability of these methods in human genetics. One of the major problems in human QTL mapping is to accurately infer the genotype of an individual at the major locus controlling variation of the quantitative trait, without the knowledge of linkage phase. In this Chapter, we propose a method to estimate, via the expectation maximization (EM) algorithm, the recombination fractions between marker loci and an autosomal major locus controlling a quantitative trait from data on nuclear families without any assumptions on linkage phase and haplotypes. The proposed method is a two-stage strategy. In the first stage, individuals are probabilistically classified into the major locus genotypes. In the second stage, the recombination fractions are estimated using the inferences made in the first stage. The proposed procedure also provides estimates of parameters of the QTL. We examine the efficiency of the estimation procedure

using Monte-Carlo simulations and show that the proposed procedure works very well.

We consider an autosomal biallelic locus, with alleles (A_1, a_1), determining a quantitative trait Y. Suppose the distributions of Y conditioned on the genotypes are:

$$Y|A_1A_1 \sim N(\alpha, \sigma^2)$$

$$Y|A_1a_1 \sim N(\beta, \sigma^2)$$

$$Y|a_1a_1 \sim N(-\alpha, \sigma^2)$$

where $\beta \leq \alpha$ and σ^2 includes the environmental variance.

Suppose the allele frequency of A_1 is p. Then, assuming Hardy-Weinberg equilibrium proportions at the QTL, X has a mixture distribution given by:

$$p^2N(\alpha, \sigma^2) + 2p(1-p)N(\beta, \sigma^2) + (1-p)^2N(-\alpha, \sigma^2).$$

Consider an autosomal biallelic codominant marker locus with alleles (M_1, m_1) possibly linked to the quantitative trait locus (QTL). [Extensions of the proposed method to multiple and multiallelic markers are discussed in later Sections.] Our aim is to estimate the recombination fraction, θ, between the two loci, which are assumed to be in linkage equilibrium.

3.2 Data Description

We consider data on nuclear families. Suppose $\{(y_{i1}, y_{i2}) : i = 1, 2, \ldots, K\}$ are the observed values of the quantitative trait of K pairs of parents such that in each pair, either one parent is M_1M_1 and the other M_1m_1 or both parents are M_1m_1. (Obviously, if neither parent is heterozygous at the marker locus, the family is not informative for linkage.) For the i^{th} pair of parents with n_i offspring, the known trait values will be denoted as $(y_{i3}, y_{i4}, \ldots, y_{in_i+2}); i = 1, 2, \ldots, K$. We further assume that the marker genotype (M_1M_1, M_1m_1 or m_1m_1) of each offspring is known. Thus, the data comprise trait values and marker genotypes of parents and offspring in nuclear families.

3.3 Estimation Procedure

Although our primary aim is to estimate θ, since the trait parameters α, β, σ^2 and p are unknown, we shall estimate these also to facilitate estimation of θ. Knowledge of α, β, σ^2 and p facilitates estimation of θ because using the estimated values of α, β, σ^2 and p, and the observed values of the quantitative trait, we can classify each parent, albeit probabilistically, to a specific trait locus genotype. When trait locus genotypes are known for the parents in a nuclear family, then obtaining an estimate of θ from the remaining data (marker genotypes of parents and offspring, and values of the quantitative trait of the offspring) becomes much simpler. Our estimation procedure is based on this two-stage strategy.

Let, $f_1(x)$, probability density function (p.d.f.) of $N(\alpha, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\alpha)^2}{2\sigma^2}}$,
π_1, prior probability of $f_1 = p^2$,
$f_2(x)$, p.d.f. of $N(\beta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\beta)^2}{2\sigma^2}}$,
π_2, prior probability of $f_2 = 2p(1-p)$,
$f_3(x)$, p.d.f. of $N(-\alpha, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x+\alpha)^2}{2\sigma^2}}$, and
π_3, prior probability of $f_3 = (1-p)^2$
Thus the p.d.f. of y_{ij} $(i = 1, 2, \ldots, K;\ j = 1, 2)$ is given by :

$$f(y_{ij}) = \sum_{n=1}^{3} \pi_n f_n(y_{ij})$$

The parameters to be estimated in this mixture model are α, σ^2 and p. We estimate these parameters by the maximum likelihood method.

The likelihood of the parental data is:

$$L(\alpha, \beta, \sigma^2, p|y_{ij}) = \prod_{i=1}^{K} \prod_{j=1}^{2} \sum_{n=1}^{3} \pi_n f_n(y_{ij})$$

However, a direct analytical maximization of the above function will not yield closed form estimators and iterative numerical maximization procedures, e.g. scoring method (Rao 1973), will involve complicated expressions.

A computationally simpler and more elegant procedure is based on the EM algorithm corresponding to a mixture of normal populations [Dempster,

Laird and Rubin 1977; McLachlan and Krishnan 1997]. A sketch of the algorithm is presented below.

The mixture distribution can be viewed as an "incomplete" set-up in the sense that we have no *a priori* knowledge as to which of the three component distributions any particular observation belongs. The first step (E-step) in this algorithm is, therefore to estimate the probabllities with which an observation may belong to any of the three component distributions. The second step (M-step) uses these estimates to build up the "complete" llkelihood function, which is easily maximized to yield relevant parameter estimates.

Define:

$$\begin{aligned} z_{ijn} &= 1, \text{ if } y_{ij} \text{ is an observation from p.d.f. } f_n, \\ &= 0, \text{ otherwise ;} \end{aligned}$$

$i = 1, 2, \ldots, K;\ j = 1, 2;\ n = 1, 2, 3.$

The introduction of z_{ijn}s thus constitutes the "complete" set-up. However, as z_{ijn}s are unknown, we have to estimate them conditioned on the observations y_{ij}. This is the E-step of the EM algorithm.

$$\begin{aligned} \widehat{z}_{ijn} &= E(z_{ijn}|y_{ij}) \\ &= \frac{\pi_n f_n(y_{ij})}{\sum_{n=1}^{3} \pi_n f_n(y_{ij})}; \end{aligned}$$

$i = 1, 2, \ldots, K;\ j = 1, 2;\ n = 1, 2, 3$. We note that these estimators are Bayes'.

Having obtained the $\widehat{z}_{ijn}$s, we can easily obtain the closed form expressions for the m.l.e. of p, α and σ^2 in the M-step of the algorithm.

$$L(p, \alpha, \beta, \sigma^2 | y_{ij}, \widehat{z}_{ijn}) = \prod_{i=1}^{K} \prod_{j=1}^{2} \prod_{n=1}^{3} \{\pi_n f_n(y_{ij})\}^{\widehat{z}_{ijn}}.$$

The m.l.e.s of the parameters are given by :

$$\widehat{p} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{2} (\widehat{z}_{ij1} + \frac{1}{2}\widehat{z}_{ij2})}{2K},$$

$$\widehat{\alpha} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{2} (\widehat{z}_{ij1} - \widehat{z}_{ij3}) y_{ij}}{\sum_{i=1}^{K} \sum_{j=1}^{2} (\widehat{z}_{ij1} + \widehat{z}_{ij3})},$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{2}\widehat{z}_{ij2}y_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{2}\widehat{z}_{ij2}},$$

$$\widehat{\sigma}^2 = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}\{\widehat{z}_{ij1}(y_{ij}-\widehat{\alpha})^2 + \widehat{z}_{ij2}(y_{ij}-\widehat{\beta})^2 + \widehat{z}_{ij3}(y_{ij}+\widehat{\alpha})^2\}.$$

Thus the l^{th} step of the EM algorithm is :

E-step :

$$\widehat{z_{ijn}}^{(l)} = \frac{\widehat{\pi_n}^{(l-1)}\widehat{f_n(y_{ij})}^{(l-1)}}{\sum_{n=1}^{3}\widehat{\pi_n}^{(l-1)}\widehat{f_n(y_{ij})}^{(l-1)}};$$

$i = 1, 2, \ldots, K;\ j = 1, 2,;\ n = 1, 2, 3.$

M-step :

$$\widehat{p}^{(l)} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{2}(\widehat{z_{ij1}}^{(l)} + \frac{1}{2}\widehat{z_{ij2}}^{(l)})}{2K},$$

$$\widehat{\alpha}^{(l)} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{2}(\widehat{z_{ij1}}^{(l)} - \widehat{z_{ij3}}^{(l)})y_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{2}(\widehat{z_{ij1}}^{(l)} + \widehat{z_{ij3}}^{(l)})},$$

$$\widehat{\beta}^{(l)} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{2}\widehat{z_{ij2}}^{(l)}y_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{2}\widehat{z_{ij2}}^{(l)}},$$

$$\widehat{\sigma^2}^{(l)} = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}\{\widehat{z_{ij1}}^{(l)}(y_{ij}-\widehat{\alpha}^{(l)})^2 + \widehat{z_{ij2}}^{(l)}(y_{ij}-\widehat{\beta}^{(l)})^2 + \widehat{z_{ij3}}^{(l)}(y_{ij}+\widehat{\alpha}^{(l)})^2\}.$$

We require initial estimates of p, α, β and σ^2 $(\widehat{p}^{(0)}, \widehat{\alpha}^{(0)}, \widehat{\beta}^{(0)}, \widehat{\sigma^2}^{(0)})$ to implement this iterative algorithm. The method of moments estimators serve as simple initial choices [see Everitt and Hand 1981].

As an initial approximation of β, we assume that there is no dominance effect, i.e., $\widehat{\beta}^{(0)} = 0$.

Assuming $\beta = 0$, the method of moments yields the following equations:

$$\overline{Y} = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}y_{ij} = \alpha(2p-1),$$

$$s^2 = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}(y_{ij} - \overline{Y})^2 = \sigma^2 + 2p(1-p)\alpha^2.$$

As $0 \leq p \leq 1$, we can fix $\widehat{p}^{(0)} = p_0$ within this interval. Thus,

$$\widehat{\alpha}^{(0)} = \overline{Y}/(2p_0 - 1);$$

$$\widehat{\sigma^2}^{(0)} = s^2 - \frac{2p_0(1-p_0)\overline{Y}^2}{(2p_0-1)^2}.$$

Clearly p_0 cannot be chosen to be 0.5.

Our next stage is to classify the parents (i.e. $\{(y_{i1}, y_{i2}) : i = 1, 2, \ldots, K\}$) into one of the three component distributions. We shall use the usual classification rule given by :

Classify y_{ij} into f_n if and only if

$$\widehat{z}_{ijn} = max_{t=1,2,3}\, \widehat{z}_{ijt} \ ;$$

$i = 1, 2, \ldots, K;\ j = 1, 2;\ n = 1, 2, 3$; the $\widehat{z}_{ijn}$s being the final (converged) values in the above EM algorithm. This is, in fact, the Bayes' classification rule corresponding to the $0-1$ loss function and thus minimises the error in classification under such loss functions [Fergusson 1967].

Having estimated $\alpha, \beta, \sigma^2, p$ and having classified the parents into the trait genotypes, we are now in a position to implement another maximum likelihood procedure to estimate θ. Before describing the actual procedure, let us note a few salient points. Information on θ can be obtained from only those offspring who have at least one of doubly heterozygous (i.e., $A_1a_1M_1m_1$) parent. We shall use the conditional trait distribution of the offspring given the trait genotypes of the parents and the marker genotypes of both parents and the offspring in order to estimate θ. We have already provided these distributions in Tables 2.2 and 2.3.

Let :

M_{ij} = marker genotype of j^{th} individual in i^{th} family,
$i = 1, 2, \ldots, K;\ j = 1, 2, \ldots, n_i + 2$

G_{i1}, G_{i2} = classified trait genotypes of the parents in i^{th} family,

$i = 1, 2, \ldots, K;\ j = 1, 2$

H_{ij} = trait genotype of j^{th} individual [i.e. $(j-2)^{th}$ offspring] in i^{th} family,

$i = 1, 2, \ldots, K;\ j = 3, 4, \ldots, n_i + 2$

P_{ijn} = $P\{H_{ij} = \gamma_n | G_{i1}, G_{i2}, M_{i1}, M_{i2}, M_{ij}\}$, where $\gamma_1 = A_1A_1$, $\gamma_2 = A_1a_1$, $\gamma_3 = a_1a_1; i = 1, 2, \ldots, K;\ j = 3, 4, \ldots, n_i + 2;\ n = 1, 2, 3.$

P_{ijn}s are obviously functions of θ. However, for the same genotype, P_{ijn} may be different for different haplotypes. For example, if $G_{i1} = A_1A_1$, $G_{i2} = A_1a_1$, $M_{i1} = M_1M_1$, $M_{i2} = M_1m_1$, $M_{i3} = M_1M_1$, then $P_{i31} = 1 - \theta$ if the haplotype corresponding to $G_{i2}M_{i2}$ is A_1M_1/a_1m_1 but $P_{i31} = \theta$ if the haplotype is A_1m_1/a_1M_1. Thus, in estimating θ, we have to consider the different possible haplotypes separately for given trait and marker loci genotypes of each parent. We next classify the offspring into their trait genotypes.

Define :

$$\begin{aligned} Q_{ijn} &= P(H_{ij} = \gamma_n | G_{i1}, G_{i2}, M_{i1}, M_{i2}, M_{ij}, y_{ij}) \\ &= \frac{P_{ijn} f_n(y_{ij})}{\sum_{n=1}^{3} P_{ijn} f_n(y_{ij})}, \end{aligned}$$

$i = 1, 2, \ldots, K;\ j = 3, 4, \ldots, n_i + 2;\ n = 1, 2, 3.$

In the computation of Q_{ijn}, we use $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\sigma}^2$ obtained using the EM algorithm described previously.

The usual classification rule is given by :

Classify y_{ij} into f_n if and only if

$$Q_{ijn} = max_{t=1,2,3}\ Q_{ijt}\ ;$$

$i = 1, 2, \ldots, K;\ j = 3, 4, \ldots, n_i + 2,\ n = 1, 2, 3.$

The likelihood of θ is given by :

$$L(\theta) = \prod_{i=1}^{K} L_i(\theta)$$

where $L_i(\theta)$ is the likelihood of the i^{th} family based on the classified genotypes of the n_i offspring of that family. Note that as haplotypic information

is usually unavailable from nuclear family data, $L_i(\theta)$ would be a mixture of the different conditional trait distributions of the offspring corresponding to the different possible haplotypes. For clarity of presentation, let us consider the following example of a nuclear family i with three ($n_i = 3$) offspring. Suppose the parental classified QTL genotypes are A_1A_1 ($= G_{i1}$) and A_1a_1 ($= G_{i2}$). Suppose the marker genotypes of these parents are, respectively, M_1M_1 ($= M_{i1}$) and M_1m_1 ($= M_{i2}$). Then, the possible haplotypes of the doubly heterozygous parent are: $A_1M|a_1m$ and $A_1m|a_1M$. The classification probabilities at the QTL for offspring depend on both marker genotypes of the offspring as also on parental haplotypes. Suppose the marker genotypes of the three offspring are: M_1M_1 ($= M_{i3}$), M_1m_1 ($= M_{i4}$) and M_1M_1 ($= M_{i5}$). Suppose, the classified QTL genotypes of these offspring are, respectively, A_1A_1 ($= H_{i3}$), A_1a_1 ($= H_{i4}$) and A_1a_1 ($= H_{i5}$) when the haplotypic configuration of the doubly heterozygous parent is $A_1M_1|a_1m_1$, and A_1a_1, A_1A_1 and A_1a_1 when the parental haplotypic configuration is $A_1m_1|a_1M_1$. Then,

$$L_i(\theta) = \frac{1}{2}\{\theta(1-\theta)^2 + (1-\theta)^3\} \tag{3.1}$$

In fact $L_i(\theta)$ is a mixture with components of the form $c_{i0}\theta^{i1}(1-\theta)^{i2}$ or $c_{i0}\theta^{i1}(1-\theta)^{i2}\{\theta^2+1-\theta)^2\}^{i3}$ where c_{i0} is some constant. Since a direct analytical maximization procedure is complicated, we implement an EM procedure. For example, the complete likelihood corresponding to (3.1) would be :

$$L_i^*(\theta) = \frac{1}{2}\{\theta(1-\theta)^2\}^m\{(1-\theta)^3\}^{1-m}$$

where $m = \frac{\theta(1-\theta)^2}{\theta(1-\theta)^2+(1-\theta)^3} = \theta$.
Thus, $L_i^*(\theta)$ would be of the form $c_i\theta^{u_i}(1-\theta)^{v_i}$ where c_i is some constant while u_i and v_i are functions of θ. Thus,

$$L^*(\theta) = \{\prod_{i=1}^{K} c_i\}\theta^{\sum_{i=1}^{K} u_i}(1-\theta)^{\sum_{i=1}^{K} v_i}$$

which is easy to maximise giving

$$\widehat{\theta} = \frac{\sum_{i=1}^{K} u_i}{\sum_{i=1}^{K}(u_i+v_i)}$$

Since u_i's and v_i's depend on θ, we need an initial approximation for implementing the EM algorithm. As $0 \leq \theta \leq 0.5$, $\theta = 0.25$ may be used as an initial approximation. If the final (converged) value of $\widehat{\theta}$ exceeds 0.5, we take $\widehat{\theta} = 0.5$.

We finally note that in the first stage of this two-stage procedure, the estimated parameters are α, β, p and σ^2. All these parameters are estimable from a sample of randomly-drawn individuals from the population. If indeed a random sample of individuals are available, then the above parameters can be estimated with trivial changes in the likelihood function derived above. The E and M steps also require trivial changes. Having estimated these parameters, one can sample families and initially classify only the parents into major QTL genotypes using the proposed classification rule (which requires the value of the quantitative trait of the individual to be classified and estimates of the parameters α, β, p and σ^2). Families in which neither parent is classified as a heterozygote at the major QTL can be discarded even before marker-typing because these families will not provide any information for estimating θ. This strategy will be cost-effective.

3.4 Efficiency of the Estimation Procedure

Assessment of the efficiency of the estimation procedure is of obvious interest. For this, we examine the empirical frequency distributions of $\widehat{\theta}$ based on multiple replicates of simulated data. Before providing the results, we describe the simulation procedure for fixed values of $p, \alpha, \beta, \sigma^2$ and θ. In the first step, we randomly generate the trait values of a fixed number ($NOBS$) of pairs of unrelated parents from appropriate (selected randomly using a trinomial random number generator with cell probabilities p^2, $2pq$ and q^2) Normal distributions (see Section 3.1). In the second step, using the data so generated, the trait parameters (α, β, σ^2, p) are estimated using the EM algorithm. (We emphasize that for the purpose of estimating the trait parameters, it is not essential to obtain data on pairs of parents; only data on randomly sampled unrelated individuals suffice.) In the third step, the QTL genotypes of the parents are inferred using the Bayes' rule. For further computations, only those pairs of parents with at least one inferred

QTL heterozygote are retained. In the fourth step, for each parent in the retained pairs, marker genotype is determined using a trinomial random number generator. For subsequent computations, only those parental pairs with at least one double heterozygote are retained. The method of generating the marker and trait genotypes of the offspring conditional on the parental mating types has been discussed in Section 2.2. These steps are repeated until the required number of informative families *(NFAM)* are obtained. Using the data so generated, we again use the EM algorithm to estimate θ. Replication of this procedure a large number of times *(NREP)* yield the empirical frequency distribution. For every set of parameter values, we evaluate the performance of the estimator with 5 offspring per family, $NFAM = 100$ and $NREP = 1000$. We, in a later Section, evaluate the effect of sample size.

3.4.1 Classification of parents with respect to QTL genotypes

As mentioned earlier, in the first stage of the present procedure, parents are classified into genotype classes on the basis of their observed trait values. Success of estimating the recombination fraction accurately by the present procedure critically depends on the performance at the first stage. It is, therefore, important to evaluate how well parents are classified to their true genotypic classes by the present method. Results pertaining to classification of parents to their true genotypes using the proposed algorithm are provided in Figures 3.1 (a)-(c) with $NOBS = 1000$, $NOBS = 250$ and $NOBS = 100$ respectively. We observe that though the classification performance is extremely good for $NOBS = 1000$, the results are sufficiently satisfactory for $NOBS = 250$. We find that when there is no dominance (i.e., $\beta = 0$), between 95% and 99.5% of the parents are correctly classified into their true genotypic classes. The percentage of correct classification increased as p deviated more from 0.5. This is expected because increase in the deviation of p from 0.5 increasingly polarises the distributions corresponding to the genotypes. The percentage of correct classification decreased as the extent of dominance (β) increases. The worst classification arose for $\alpha = 5$ and $\beta = 4$. In this case, the overlap between distributions of the A_1A_1 and

A_1a_1 genotype classes is the largest. Therefore, a non-informative parent (i.e., with true genotype A_1A_1) has a high probability of being classified as informative (i.e., with true genotype A_1a_1) and the vice-versa. However, even in this case, the probability of correct classification is about 80%. We also note that these results are independent of θ. Thus, it is seen that the first stage of the proposed method works extremely well indicating that evaluation of the next stage, in which an estimate of the recombination fraction is obtained, is worthwhile.

3.4.2 Empirical frequency distribution of $\widehat{\theta}$

If indeed the procedure provides a good estimate of the recombination fraction, θ, then one expects that the probability distribution of $\widehat{\theta}$ obtained from multiple replications of simulated data generated using a fixed set of parameter values will be clustered around the true value of θ. Figures 3.2-3.10 depict the frequency distributions of $\widehat{\theta}$ for simulation parameter values of $\theta = 0, 0.1, 0.3$ and 0.5, separately for $p = 0.9,\ 0.7,\ 0.5$ and $\beta = 0,\ 2,\ 4$. The values of the other parameters used in these simulations are: $\alpha = 5$ and $\sigma^2 = 1$. From these figures it is seen that in all cases, except when the trait and marker loci are completely unlinked (i.e., $\theta = 0.5$) and the dominance effect (β) is large [Figures 3.8(d), 3.9(d) and 3.10(d)], the distributions are unimodal and leptokurtic. In these extreme cases, there are higher probabilities of misclassification as has been noted in the previous Subsection. For $\theta = 0$, in 80%-85% of the replications $\widehat{\theta}$ is ≤ 0.08 if $\beta = 0$, while this percentage is between 65%-70% if $\beta = 4$. Similarly for $\theta = 0.3$, in 80%-90% of the replications $\widehat{\theta}$ is in the interval $[0.25, 0.35]$. However, for $\theta = 0.5$, while 95% of the $\widehat{\theta}$ values are between 0.45 and 0.5 for $p = 0.5$, this percentage for $p = 0.9$ is only about 75%. The proportion of $\widehat{\theta}$ values lying close to the true value value of θ decreases as β increases. Thus, it is seen that the procedure provides good estimates in conformity with expectations, unless the degree of dominance (β) is very high. Therefore, if the estimated value of β is close to that of α, the estimate of θ may be inaccurate.

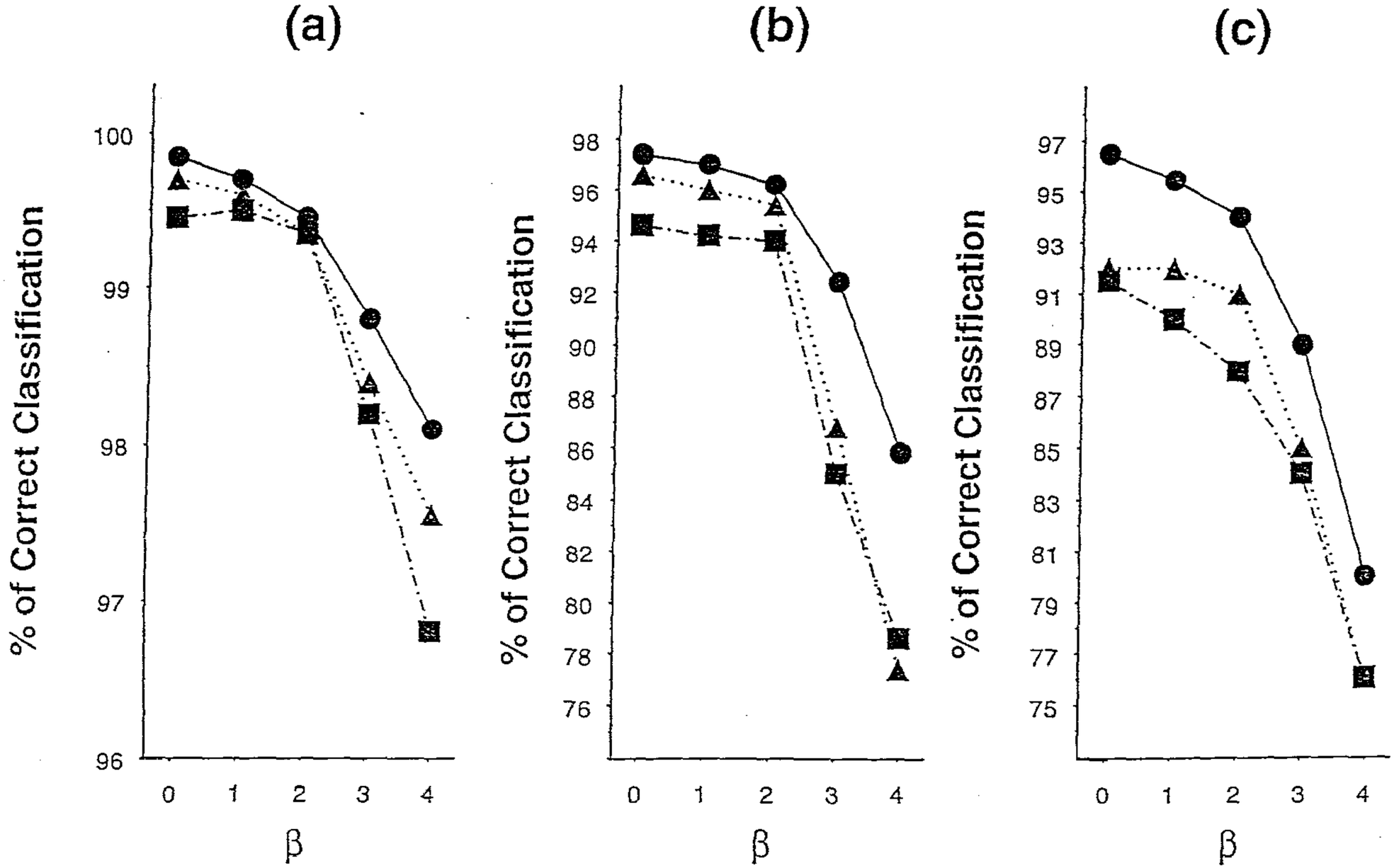


Figure 3.1. Percentage of correct classification of parents for simulation parameter values $\alpha = 5, \beta = 0, 1, 2, 3, 4, \sigma^2 = 1$ and (a) $NOBS = 1000$, (b) $NOBS = 250$ and (c) $NOBS = 100$. Circles correspond to $p = 0.9$, triangles to $p = 0.7$ and squares to $p = 0.5$.

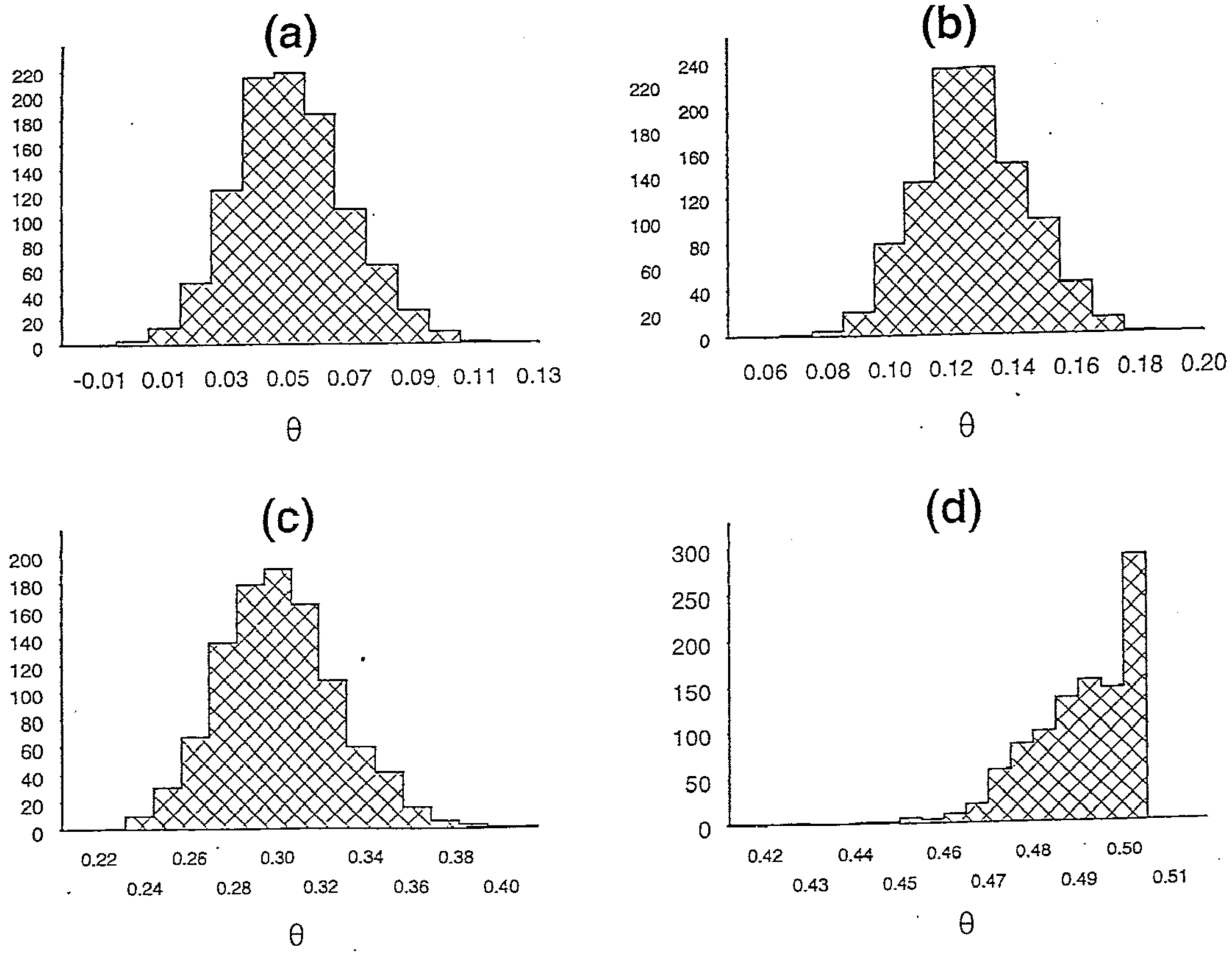
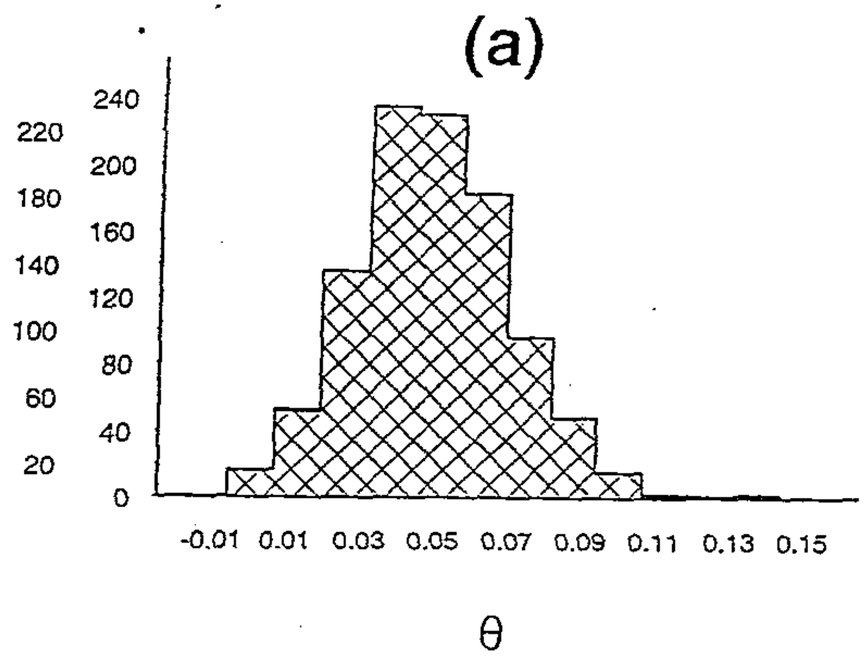
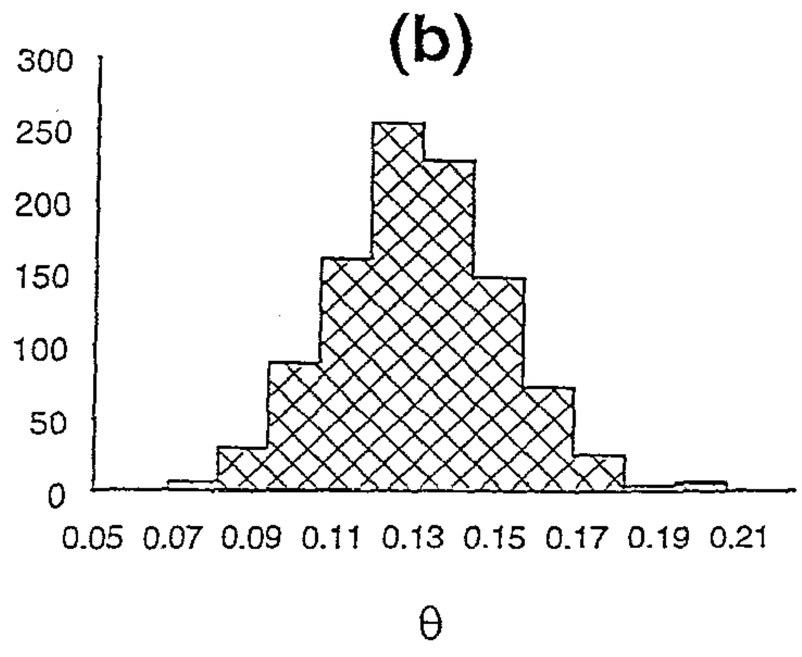
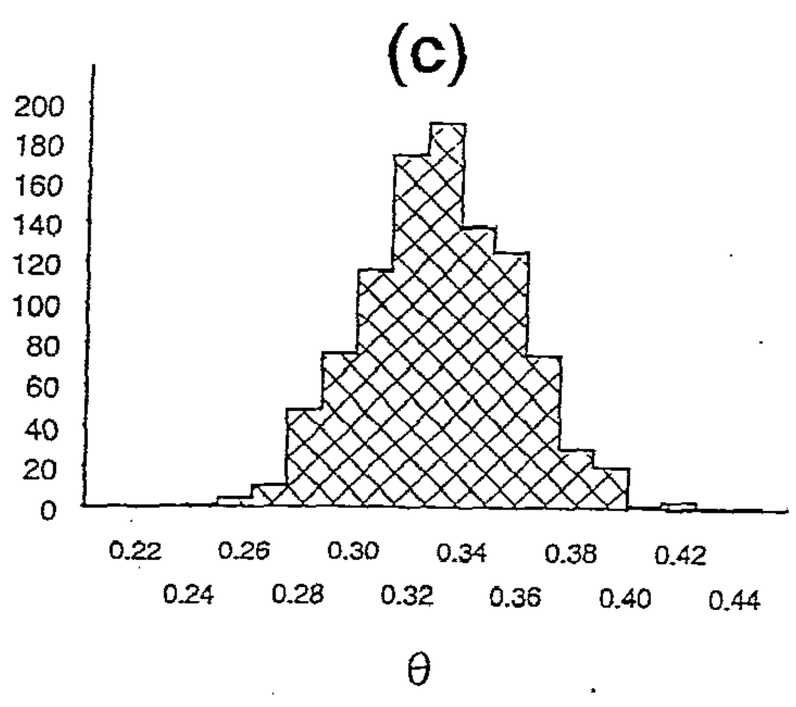


Figure 3.2. Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values $p = .9, \alpha = 5, \beta = 0, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.







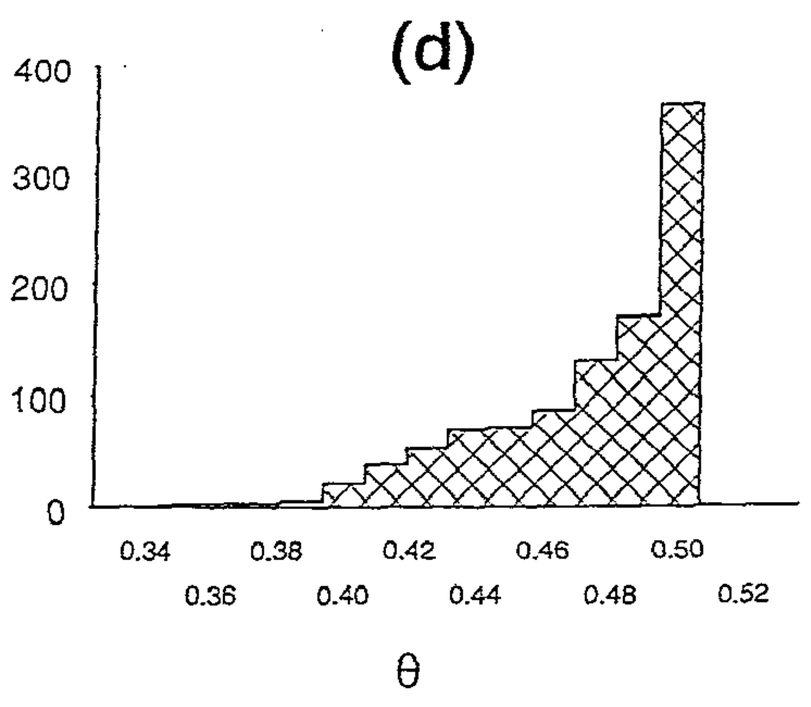


Figure 3.3. Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values $p = .7, \alpha = 5, \beta = 0, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.

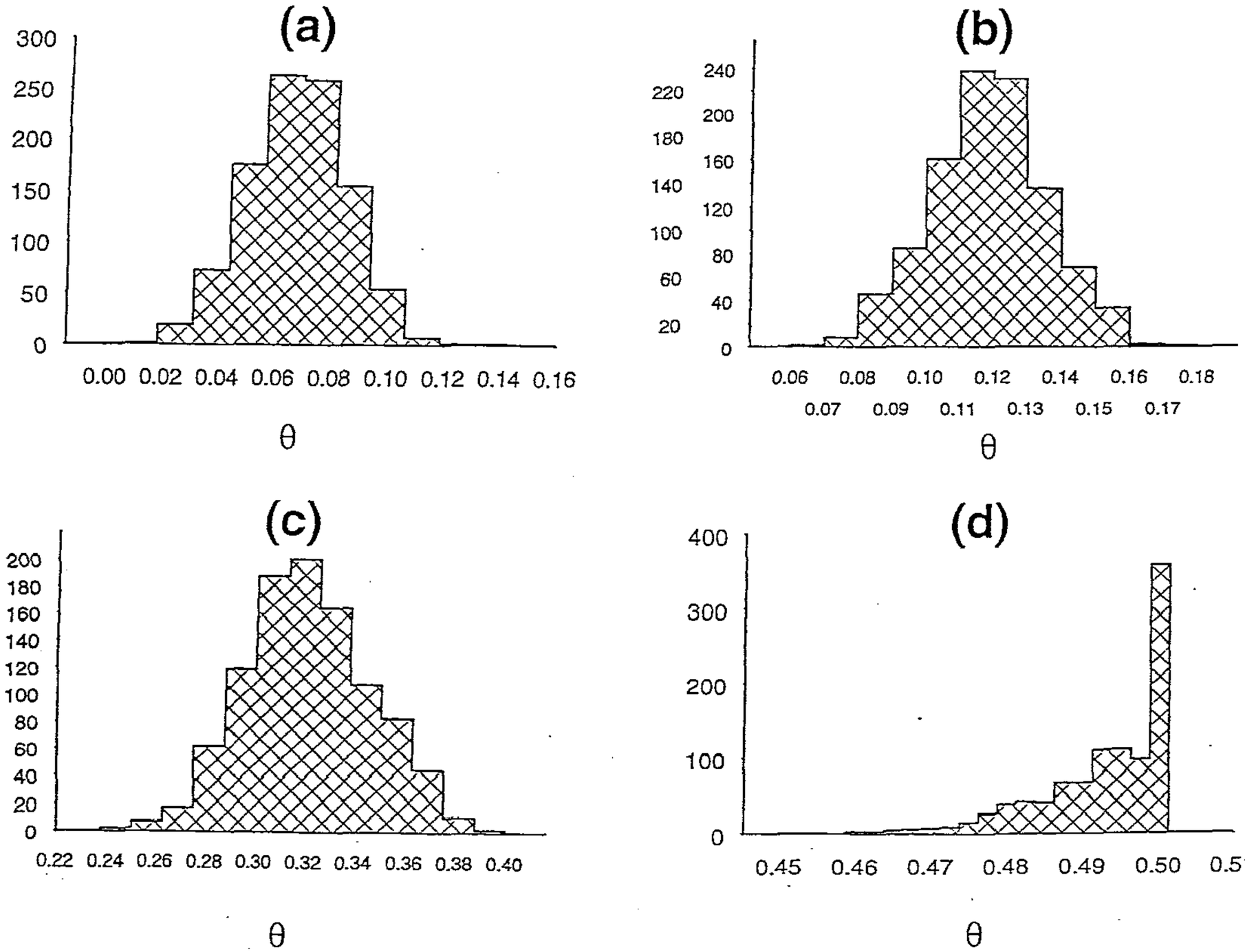
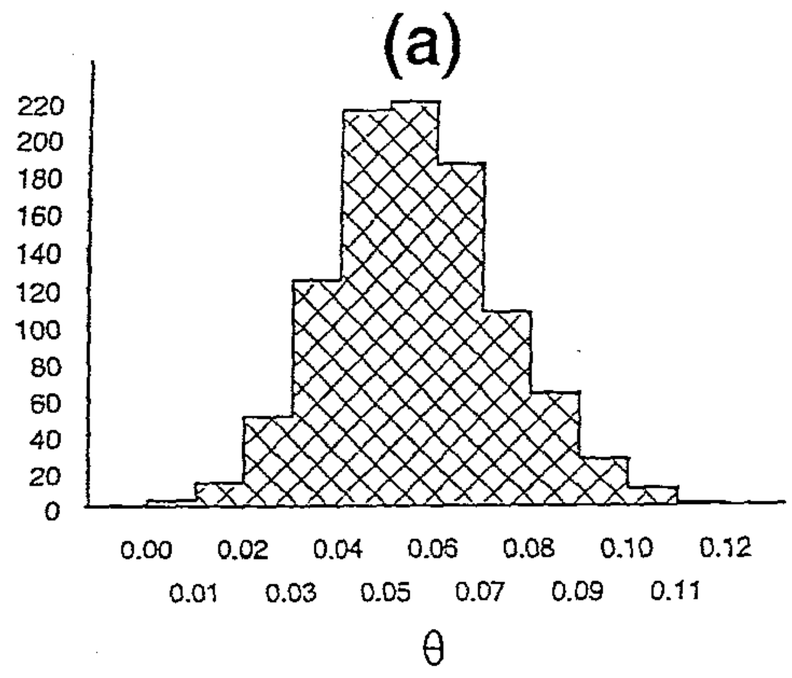
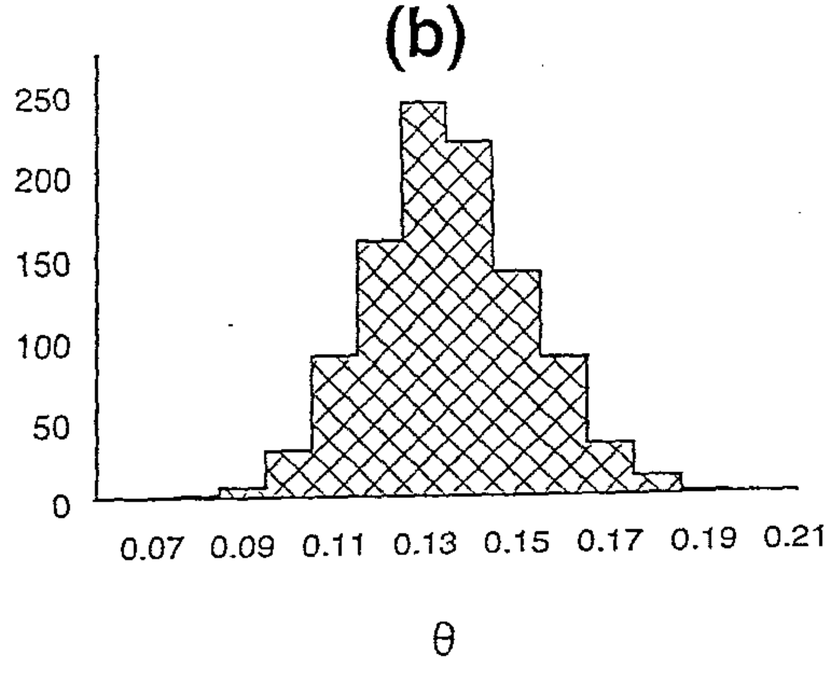
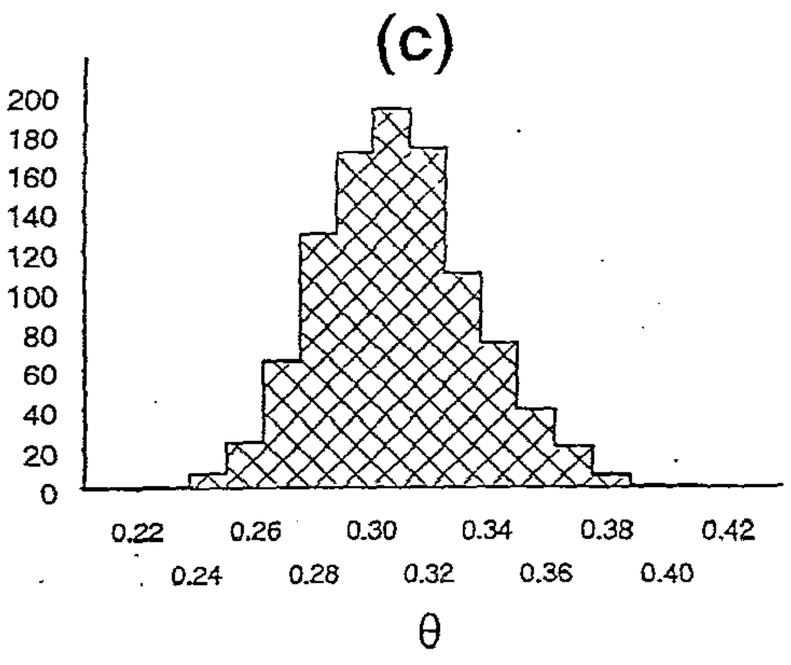


Figure 3.4. Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values $p = .5, \alpha = 5, \beta = 0, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.







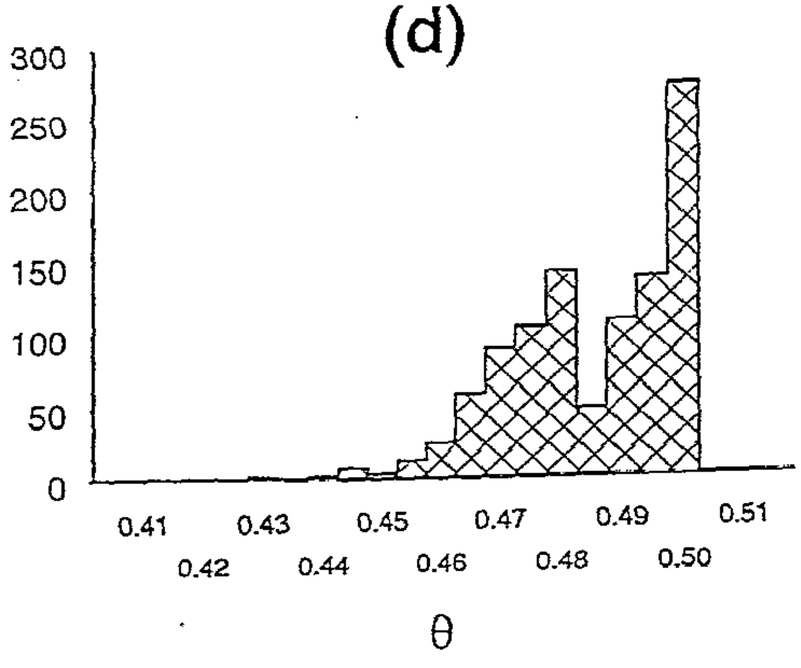


Figure 3.5. Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values $p = .9, \alpha = 5, \beta = 2, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.

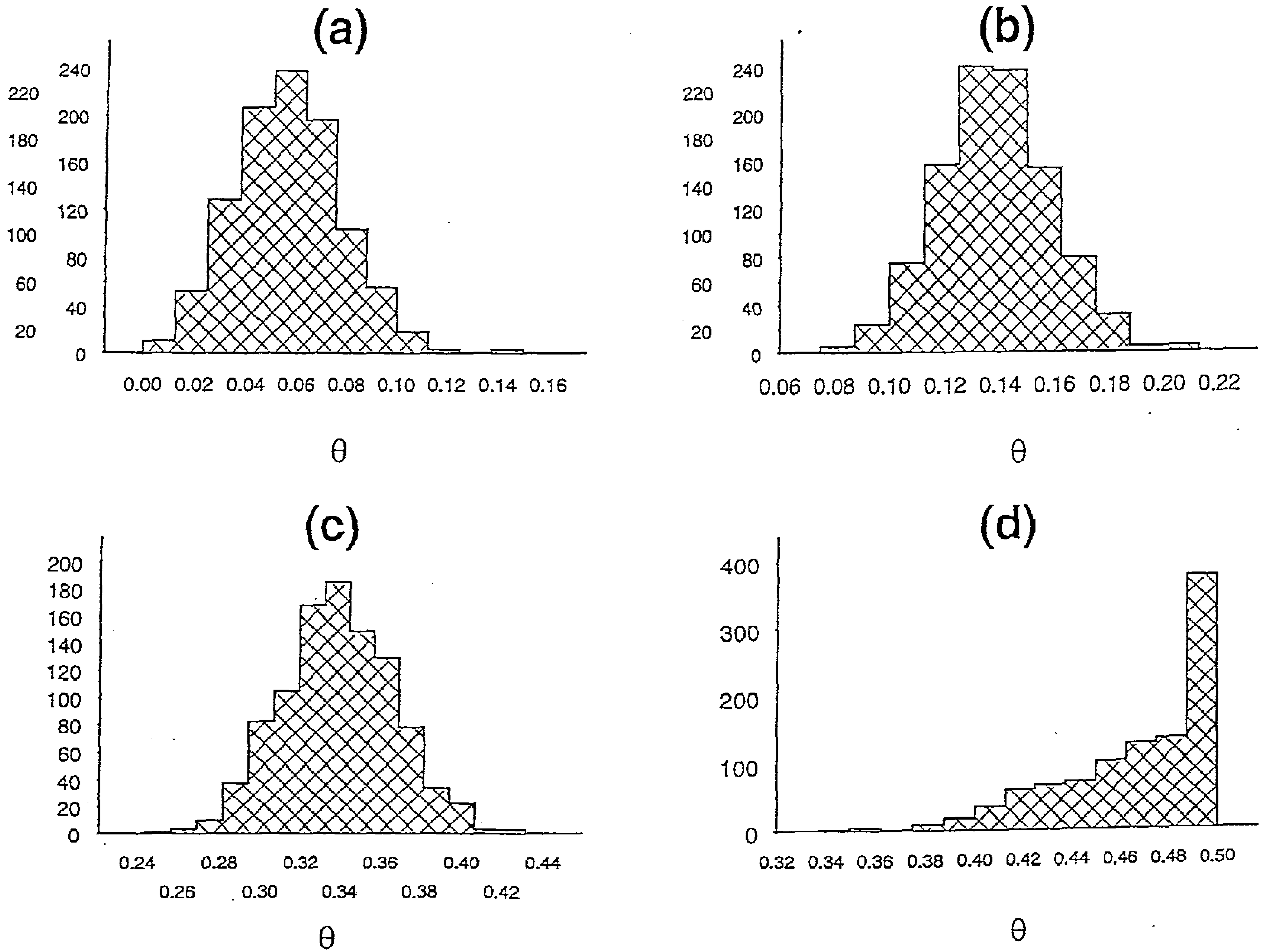


Figure 2.5 Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values

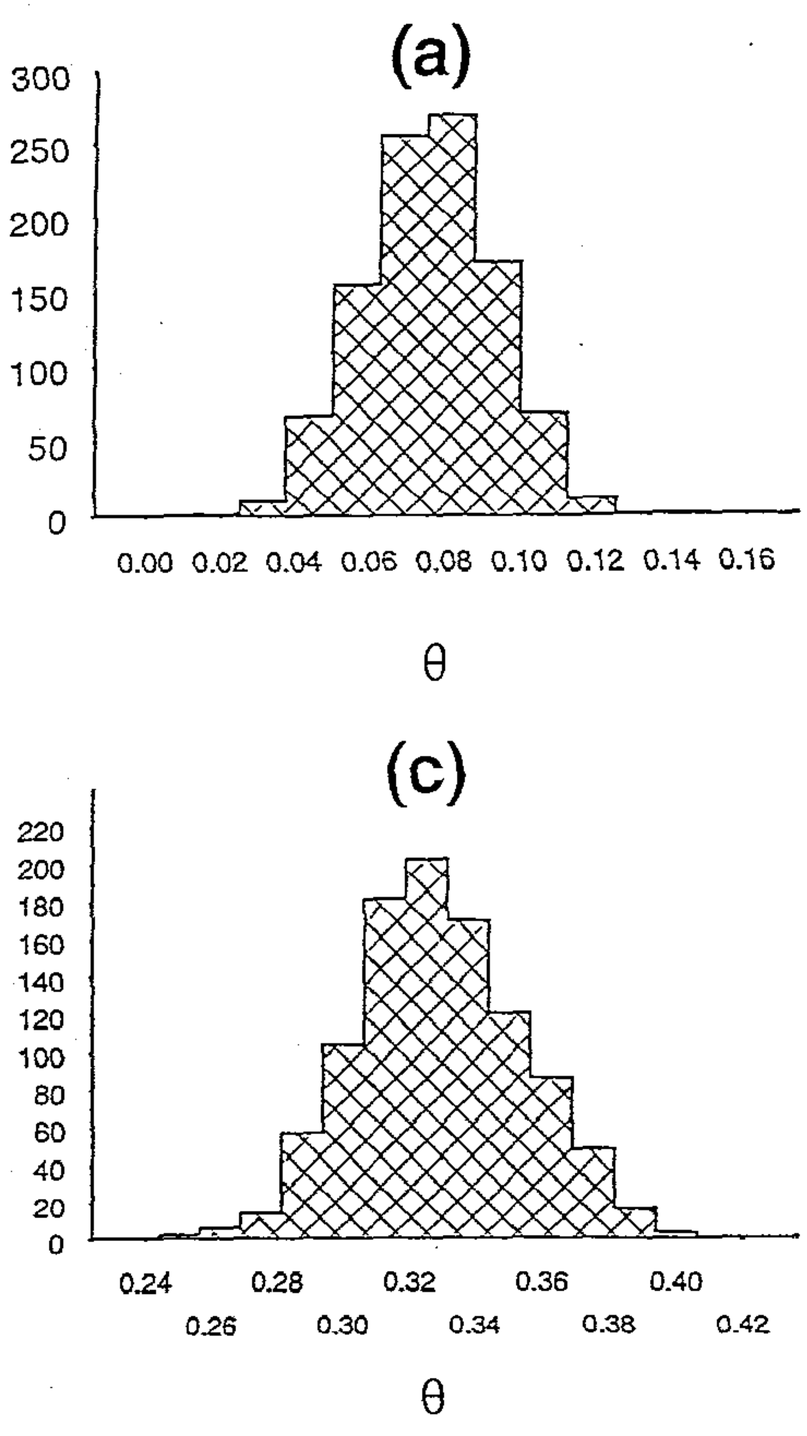


Figure 3.7. Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values $p = .5, \alpha = 5, \beta = 2, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.

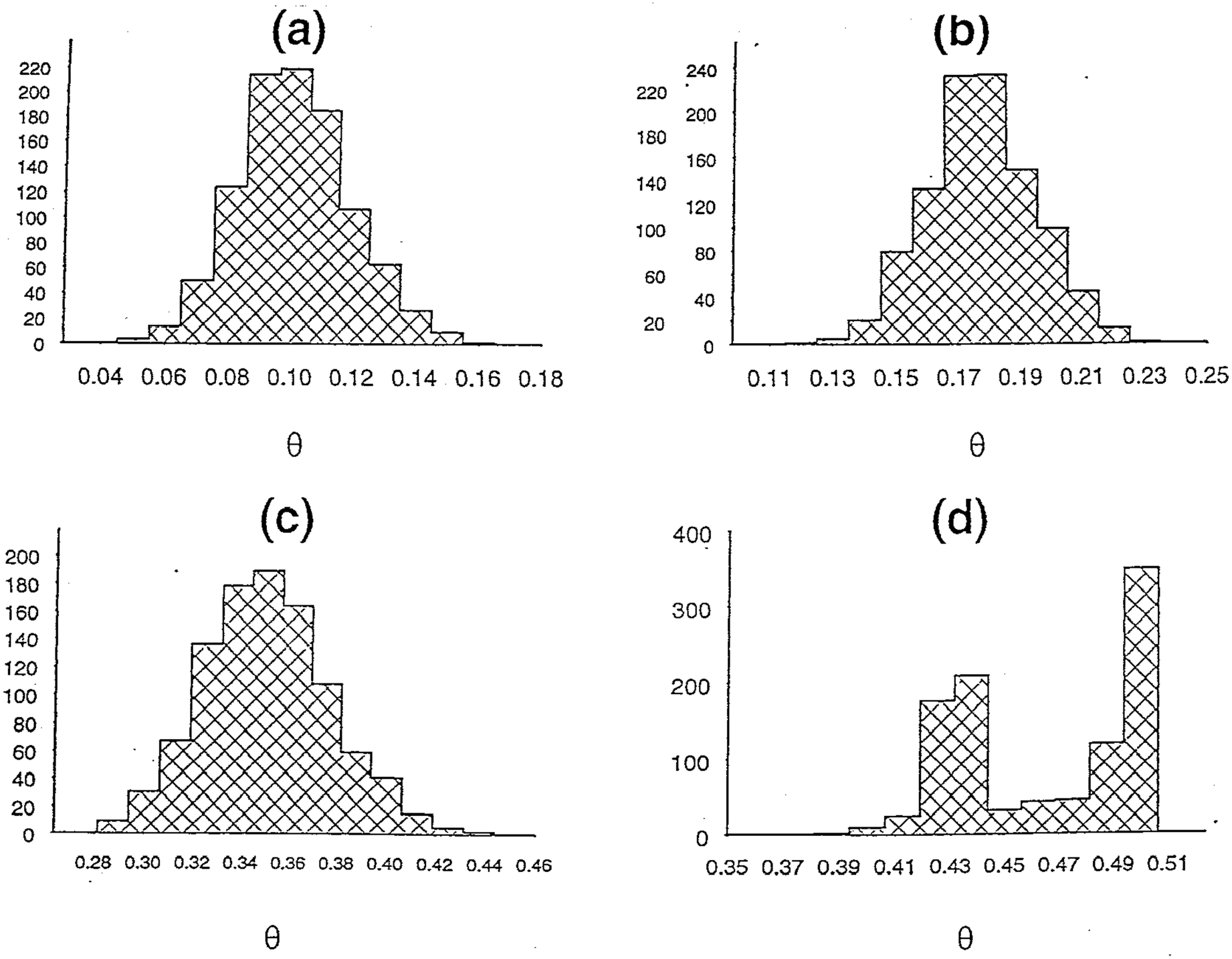
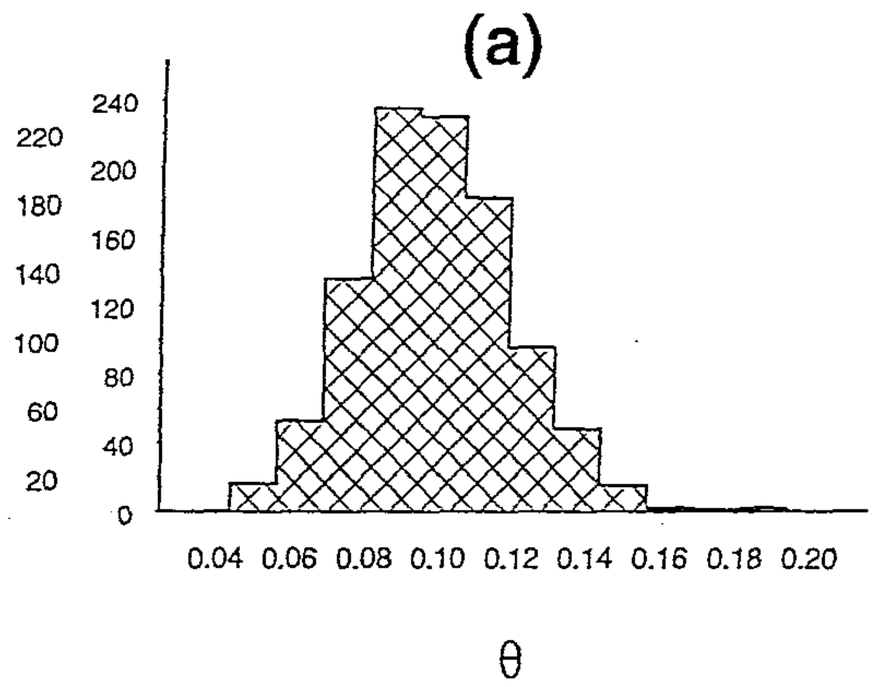
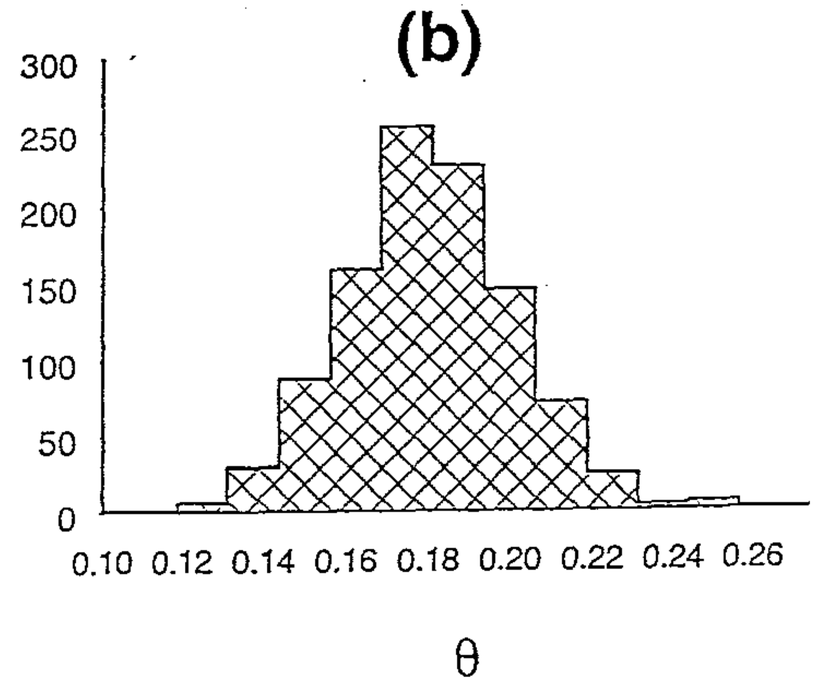
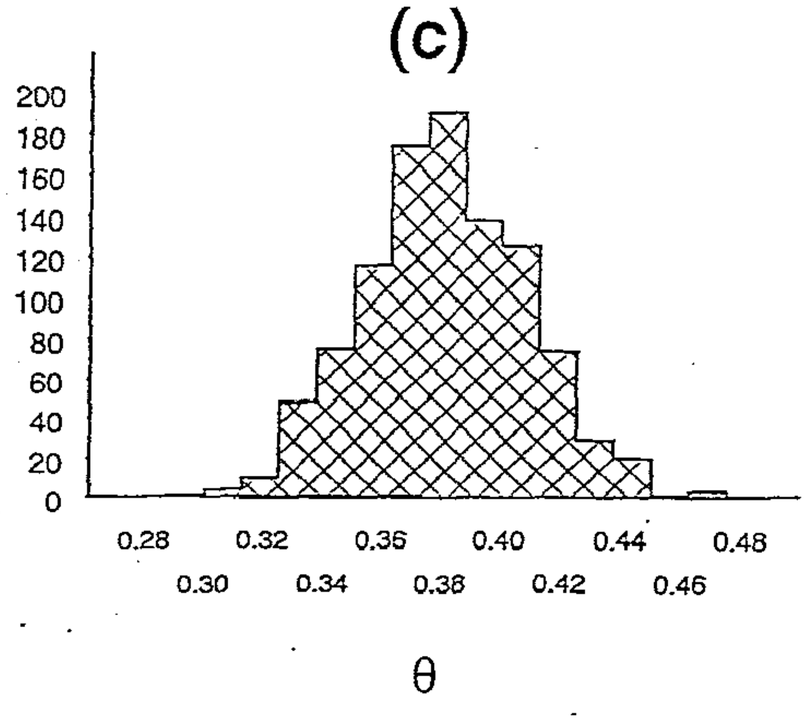


Figure 3.8. Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values $p = .9, \alpha = 5, \beta = 4, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.







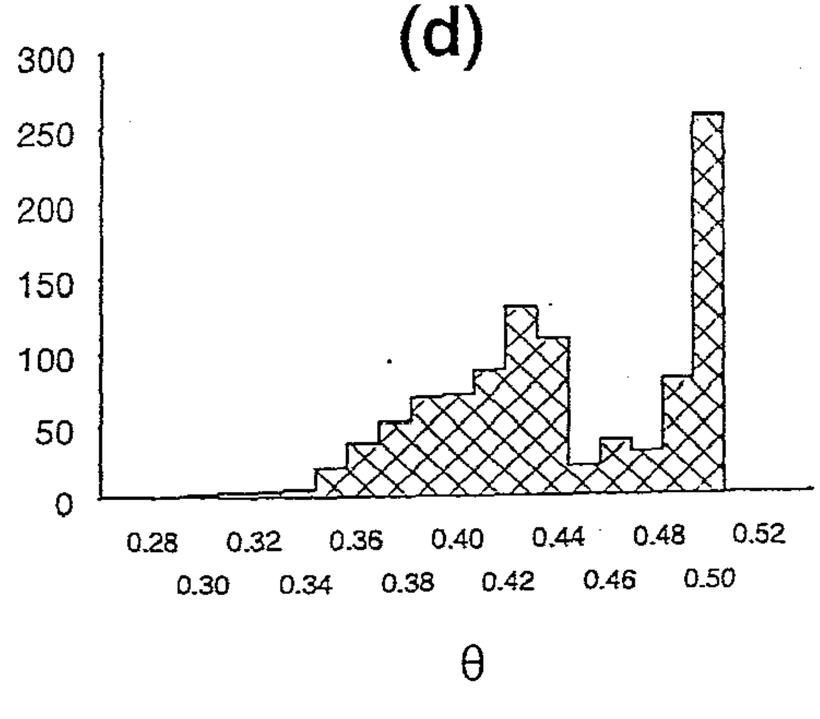


Figure 3.9. Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values $p = .7, \alpha = 5, \beta = 4, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.

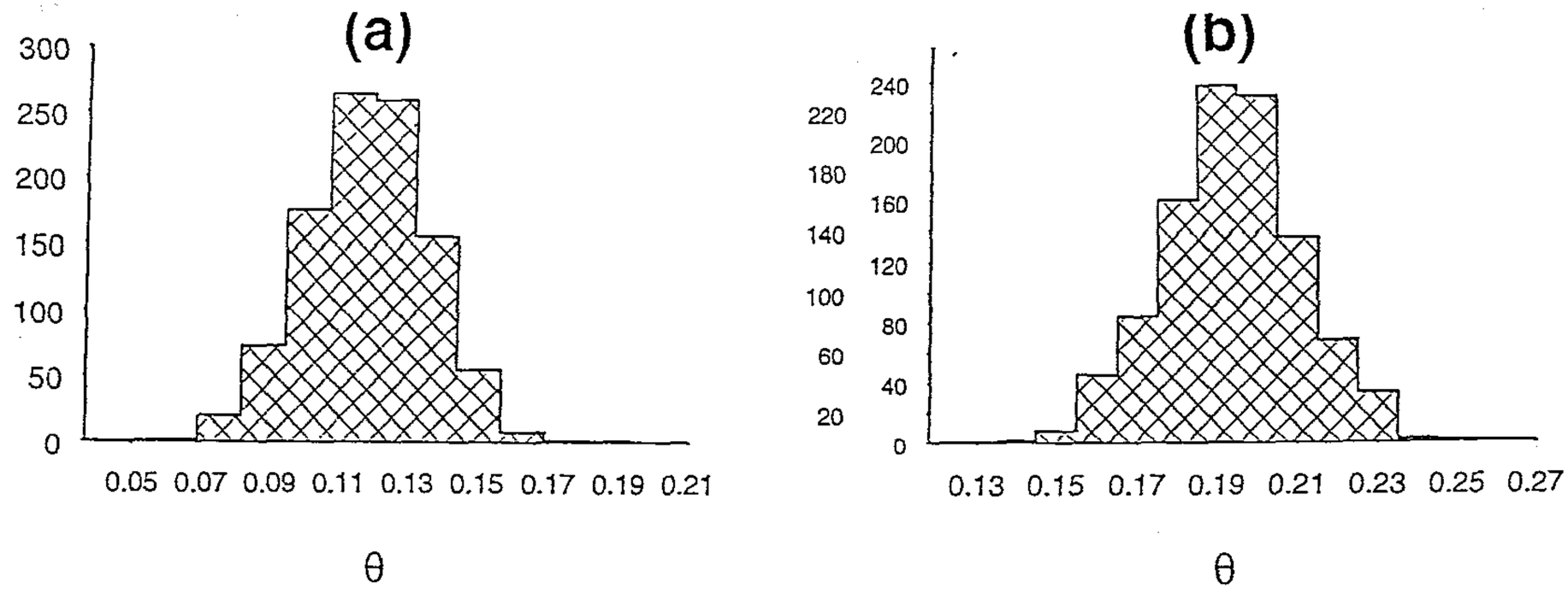


Figure 3.10. Empirical frequency distributions of $\widehat{\theta}$ for simulation parameter values $p = .5, \alpha = 5, \beta = 4, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d)

3.4.3 Mean and Variance of $\widehat{\theta}$ and Confidence Interval for θ

To examine the behavior of the estimator in respect of variation in values of p and β, we perform simulations for fixed parameter values $\alpha = 5$, $\sigma^2 = 1$ and for values of $p = 0.9, 0.7, 0.5$; $\beta = 0, 2, 4$ and $\theta = 0.0.1, 0.3, 0.5$. We evaluate the means and variances of $\widehat{\theta}$ and obtain 95% confidence intervals of θ. These results are given in Table 3.1. It is seen from this table that the true value of θ is always included in the 95% confidence interval of θ. The coefficient of variation of $\widehat{\theta}$ is also $< 0.5\%$. These results indicate that the performance of the proposed estimator is extremely good. It is also seen from Table 3.1 that when p deviates from 0.5, the mean of $\widehat{\theta}$ is closer to the true value of θ and the 95% confidence interval of θ is narrower, unless θ is very close to 0.5. The variance of $\widehat{\theta}$ increases when p deviates more from 0.5. Table 3.1 also shows that estimates of θ deviate more from the true value of θ, their variances increase and the 95% confidence intervals becomes wider as the value of β increases to α. This phenomenon is due to increasing overlap between the two genotypic classes A_1A_1 and A_1a_1 as β becomes closer to α, resulting in errors of parental genotype classification. We also note that for fixed values of α, σ^2, p and θ, the adverse effect of increase of β on estimation of θ is non-linear in β. For the sets of parameter values investigated, the relative error in estimation of θ never exceeds 20%.

Thus, for a fixed value of β, the efficiency of estimation of θ is dependent on both the true value of θ and the trait allele frequency, p, for any finite sample size of families. Since the trait genotype is unknown and needs to be inferred, the inference is better when the value of p is much deviates from 0.5, because in such cases the overlap of trait-value distributions among genotypes is small. However, the efficiency of estimation of θ strongly depends on the "effective" sample size (that is, the number of informative families), especially when the true value of θ is not close to 0. When the value of p deviates from 0.5, the effective sample size decreases, thereby reducing the efficiency of estimation of θ.

Table 3.1. Mean and Variance of $\widehat{\theta}$ and 95% Confidence Interval of θ for $\alpha = 5$, $\sigma^2 = 1$, $p = .9, 0.7, 0.5$; $\beta = 0, 2, 4$; $\theta = 0, 0.1, 0.3, 0.5$.

p	True θ	β	Mean($\widehat{\theta}$)	Var($\widehat{\theta}$)	95% C.I. of θ
0.9	0	0	0.015	0.000174	(0.009 ,0.026)
		2	0.044	0.000432	(0.017 ,0.048)
		4	0.075	0.000695	(0.051 ,0.097)
	0.1	0	0.103	0.000084	(0.099 ,0.114
		2	0.117	0.000277	(0.095 ,0.126)
		4	0.172	0.001008	(0.131 ,0.195)
	0.3	0	0.303	0.000452	(0.291 ,0.311)
		2	0.313	0.000747	(0.286 ,0.328)
		4	0.368	0.001739	(0.345 ,0.401)
	0.5	0	0.478	0.000397	(0.438 ,0.500)
		2	0.471	0.000902	(0.415 ,0.500)
		4	0.409	0.001335	(0.395 ,0.487)
0.7	0	0	0.021	0.000154	(0.019 ,0.041)
		2	0.053	0.000312	(0.023 ,0.057)
		4	0.081	0.000865	(0.063 ,0.101)
	0.1	0	0.107	0.000087	(0.095 ,0.122)
		2	0.122	0.000290	(0.097 ,0.128)
		4	0.182	0.001064	(0.143 ,0.204)
	0.3	0	0.308	0.000497	(0.293 ,0.317)
		2	0.317	0.000683	(0.284 ,0.321)
		4	0.373	0.001867	(0.357 ,0.408)
	0.5	0	0.491	0.000083	(0.477 ,0.500)
		2	0.487	0.000118	(0.472 ,0.500)
		4	0.413	0.001146	(0.401 ,0.494)
0.5	0	0	0.038	0.000186	(0.022 ,0.058)
		2	0.067	0.000299	(0.035 ,0.073)
		4	0.105	0.001018	(0.071 ,0.112)
	0.1	0	0.113	0.000129	(0.097 ,0.123)
		2	0.115	0.000283	(0.089 ,0.124)
		4	0.196	0.001153	(0.162 ,0.208)
	0.3	0	0.314	0.000512	(0.291 ,0.325)
		2	0.321	0.000630	(0.287 ,0.329)
		4	0.381	0.001794	(0.358 ,0.416)
	0.5	0	0.497	0.000056	(0.486 ,0.500)
		2	0.491	0.000068	(0.478 ,0.500)
		4	0.421	0.001062	(0.411 ,0.498)

3.4.4 Sample size effect

We investigate the performance of the proposed estimator when data on fewer offspring are available. Based on simulated data with 3 offspring per family, we obtain the empirical frequency distributions of $\widehat{\theta}$ for different values of θ and $p = 0.5, 0.7, 0.9$. The frequency distributions are marginally less well-behaved compared to those based on 5 offspring per family. The estimated $\widehat{\theta}$ values are also marginally more deviant from thrue θ values with smaller sibship sizes. For example, the mean and variance of $\widehat{\theta}$ for simulation parameter values $\alpha = 5$, $\beta = 2$, $\sigma^2 = 1, p = 0.7$ and $\theta = 0.3$ are 0.342 and 0.00765 respectively based on 100 informative families with 3 offspring per family, while these figures are 0.317 and 0.000683 with 5 offspring per family. For $p = 0.5$ and the remaining simulation parameter values same as above, the mean of $\widehat{\theta}$ is 0.377 with 3 offspring per family, while it is 0.321 with 5 offspring per family. The additional deviation of estimated θ from true θ due to decrease in sibship size from 5 to 3 varies between 8% and 20 %. Thus, while larger data sets are desirable especially when p is close to 0.5, our method continues to perform rather well even with smaller sibship sizes. We also emphasize that although we perform our simulation experiments with fixed sibship sizes of 5 and 3, variable sibship sizes pose no problem. Likelihood equations are easily modified and since data on offspring conditional on parental genotypes are independent, our simulation results are based effectively on 5(or 3)× number of families.

3.4.5 Power of test for linkage detection for varying degrees of major trait locus effect

To assess, more clearly, the efficiency of the proposed estimator $\widehat{\theta}$ at varying degrees of major locus effect, measured as the proportion of variance of QT explained by the major biallelic locus (Δ), we compute the empirical power of the test of hypothesis $H_0 : \theta = \theta_0 < 0.5$ vs $H_1 : \theta = 0.5$ For this assessment, we generate simulated data for different values of Δ. While such data can be generated for various combinations of parameters, we present results of β and σ^2 kept fixed at 0 and 1 respectively, and with α and p varied suitably to attain different values of Δ in the range of 0.2 to 0.9.

A fixed value of $\theta_0 = 0.1$ is used throughout. For each set of simulated data, the empirical 5% cut-off points for rejection for the null hypothesis is determined and the power is estimated as the proportion of replications (out of 1000) with $\theta = 0.5$ in which $\widehat{\theta}$ is greater than the empirical 5% cut-off point. The results are graphically presented in Figure 3.11 from which it is evident that our proposed method performs quite well, at least when the percentage of variance in QT explained by the major locus exceeds 30%. Other combinations of values of parameters $\beta, \sigma^2, \alpha, p$ yielding the same value of Δ result in approximately the same power; estimates are not provided for brevity.

3.4.6 Effect of linkage heterogeneity

Suppose a QT is controlled by a single major biallelic locus which explains Δ % of variance (the remaining variance being environmental), but there is linkage heterogeneity. That is, in a proportion (π) of families, the QT is due to one major locus and in the remaining proportion $(1 - \pi)$ of families, the trait is due to an unlinked major biallelic locus. We assume that the values of Δ for both loci are equal, which in a sense is the worst-case scenario. Suppose, a biallelic marker is linked to the first QTL; that is, linkage is present in only π proportion of families. It is pertinent to examine the performance of the proposed procedure in estimating linkage from the pooled set of families, when one is unaware of the existence of the underlying linkage heterogeneity.

We use simulations to examine this. In generating simulated data, we use $\theta = 0.1$, two values of $\Delta = 80\%$ and 60% (in each case, α, β and σ^2 are kept fixed at 5,0 and 1 respectively; p is varied suitably to attain appropriate values of Δ), and five values of $\pi = 0.9, 0.8, 0.7, 0.6$ and 0.5 for each value of Δ. Results are presented in Table 3.2, from which it is seen that the estimated value of θ is reasonably good unless there is considerable linkage heterogeneity (small π) or the proportion of variance explained by the major locus is small (small Δ). When compared to the results presented in the previous Subection (and Figure 3.11), we see that while in the absence of linkage heterogeneity the proposed method performs quite well even when Δ is as low as 30%, in the presence of linkage heterogeneity the method fails

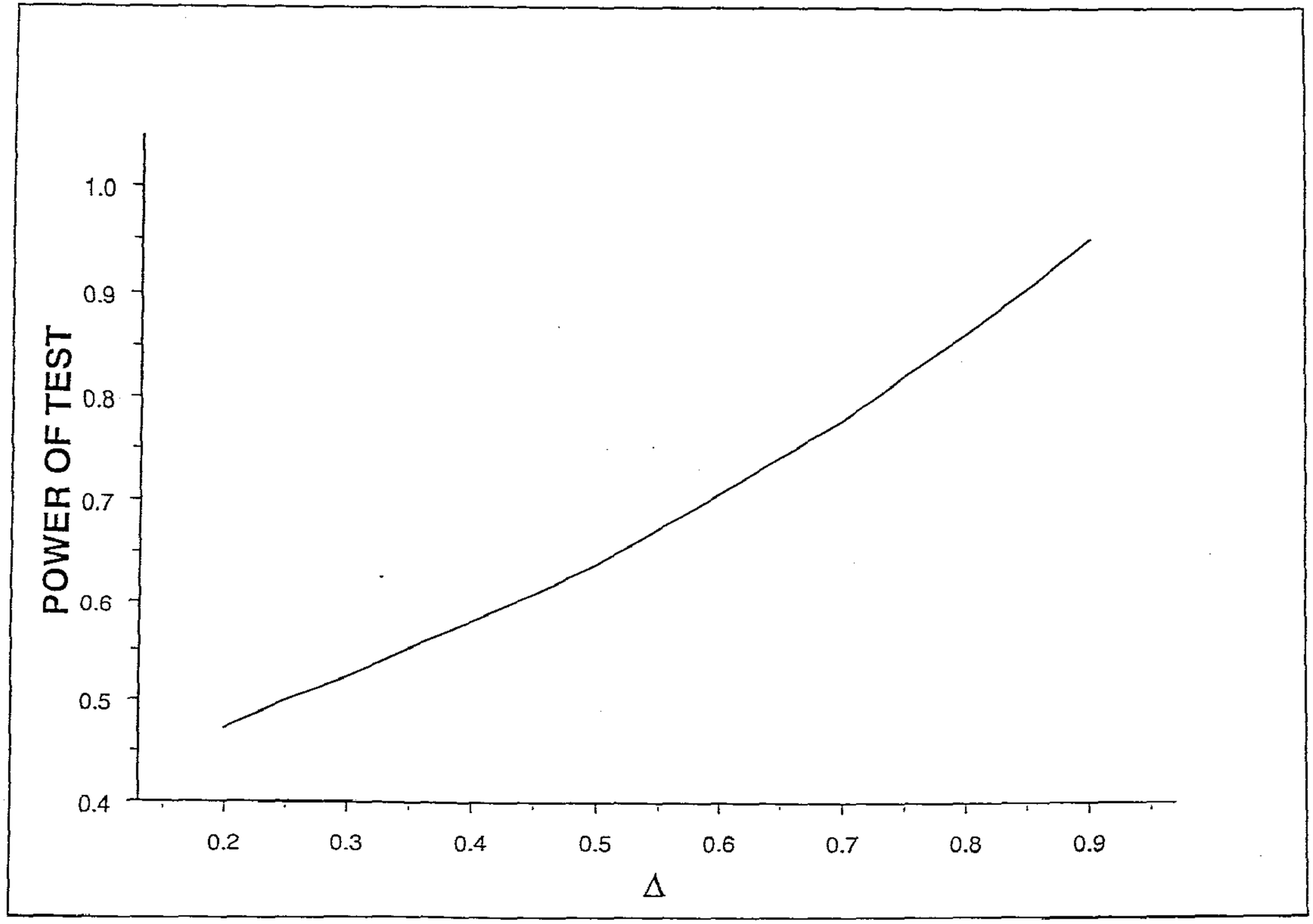


Figure 3.11. Empirical power of the test procedure for detecting linkage for different values of the proportion of variance in QT explained by the major QTL.

to perform well unless Δ is as high as about 80%.

Table 3.2. Mean and Variance of recombination fraction in presence of linkage heterogeneity (π=proportion of linked families) estimated from simulated data sets with differing values of Δ (percentage of variance explained by the major QTL) and recombination fraction=0.1

π	$\Delta = 80$ %		$\Delta = 60$ %	
	Mean($\widehat{\theta}$)	Var($\widehat{\theta}$)	Mean($\widehat{\theta}$)	Var($\widehat{\theta}$)
0.9	0.126	0.0041	0.1466	0.0075
0.8	0.1415	0.0062	0.1683	0.0112
0.7	0.167	0.0091	0.1975	0.0172
0.6	0.194	0.0143	0.223	0.0237
0.5	0.2268	0.0197	0.2549	0.0294

3.4.7 Analyzing a two-locus QT as a single-locus QT: Effect on estimate of θ

Consider a QT that is controlled by two unlinked, biallelic trait loci. Suppose a biallelic marker is linked to one of the two loci. In the absence of knowledge that the QT is controlled by two major loci, it is reasonable to investigate the effect of analyzing data assuming that the QT is controlled by a single major locus, on the estimate of θ. To examine this issue, we generate simulated data sets for different values of α and p. We denote the values of α as α_1 and α_1 for the two trait loci. The values of β and p are held fixed at 0 and 0.7, respectively, for both loci; θ (recombination fraction between the marker and one QT) is taken as 0.1. Having generated replicate data sets for each combination of the above parameters, we use the proposed method for estimating θ assuming that the QT is controlled by a single major locus. The results are presented in Table 3.3. It is seen from this table that when the effect on the QT of the trait locus to which the marker is unlinked (the "unlinked QTL") is small relative to the linked trait locus, the method performs quite well even when the data are incorrectly analyzed assuming that the QT is controlled by a single major locus. However, the estimate of θ, which is always upwardly biased, worsens as the relative effect of the linked trait locus decreases. When the relative effects of the two trait loci are equal, analyzing the two-locus QT data as single-locus data leads to

hopelessly bad estimates of the recombination fraction.

Table 3.3. Effect of analyzing two-locus QT data as a single-locus data, on estimated θ; when its true value is 0.1 (α_1 and α_1 denote the effects of the two QTLs, Δ is the proportion of variance explained jointly by the two QTLs and Δ_1 is the proportion explained by the linked QTL.

α_1	α_1	Δ	Δ_1	Mean($\widehat{\theta}$)	Var($\widehat{\theta}$)
5	1	0.91	0.88	0.1088	0.0007
5	2	0.92	0.79	0.1291	0.0016
5	3	0.93	0.68	0.1516	0.0031
5	4	0.94	0.57	0.1774	0.0058
5	5	0.95	0.47	0.2056	0.0094

3.5 Estimation of θ when the Marker is Multiallelic

The above procedure of estimation of θ can be easily shown to hold in the case of a multiallelic locus. Suppose the marker locus has K alleles denoted as $M_1, M_2, \ldots, M_K$. A mating between a homozygote and a heterozygote will be of the form $M_iM_i \times M_jM_k$, while a mating between two heterozygotes will be of the form $M_iM_j \times M_kM_l$.

A $M_iM_i \times M_jM_k$ mating will produce offspring with marker genotypes M_iM_j and M_iM_k with probability 1/2 each. The probabilities of the trait genotypes of the offspring for varlous parental mating types are identical to those corresponding to marker genotypes M_1M_1 or M_1m_1 given in Table 2.2. In the case of a mating between two heterozygotes, we need to differentiate between matings $M_iM_j \times M_iM_j$ and $M_iM_j \times M_kM_l$ where either $i \neq k$ or $j \neq l$. For $M_iM_j \times M_iM_j$ matings, the distributions of the trait genotypes of the offspring for various parental mating types are identical to those corresponding to marker genotypes M_1M_1, M_1m_1 or m_1m_1 given in Table 2.3. $M_iM_j \times M_kM_l$ ($i \neq k$ or $j \neq l$) matings can produce offspring with marker genotypes M_iM_k, M_jM_k, M_iM_l and M_jM_l with probability 1/4 each. The probabilities of the trait genotypes of the offspring for various parental mating types are given in Table 3.4.

Note that the estimation of the trait parameters α, σ^2 and p does not depend on the marker. Thus the procedure of estimating these parameters in the case of a multiallelic marker is identical to that in the case of a biallelic marker described earlier. While estimating θ, we should consider the appropriate conditional distribution of the trait genotypes of the offspring given in Tables 2.2, 2.3 and 3.4. As all the probabilities in these tables are some multiples of θ, θ^2, $(1-\theta)$, $(1-\theta)^2$, $[\theta^2+(1-\theta)^2]$, we can use the EM procedure described earlier to obtain the m.l.e. of θ.

3.6 The EM Approach in Multipoint Mapping

The proposed EM procedure for mapping a trait locus using two-point linkage can be easily extended to the case of multipoint mapping. For ease of exposition, we consider a three- point mapping set-up. Suppose the trait locus is flanked by two biallelic, codominant marker loci with alleles (M_1, m_1) and (M_2, m_2) respectively such that the recombination fractions between the trait locus and the marker loci are θ_1 and θ_2 respectively. We assume that chromatid interference is absent, and hence, the recombination fraction between the two flanking markers is $\theta = \theta_1 + \theta_2 - 2\theta_1\theta_2$. The conditional probabilities of the trait genotypes of offspring given the parental trait and marker genotypes as well as the offspring marker genotypes are given in Table 3.5 for backcross at both marker loci, in Table 3.6 for intercross at both marker loci and in Table 3.7 for backcross at one marker locus and intercross at the other.

We use simulated data to assess the relative efficiency of multipoint linkage analysis over two-point linkage analysis. The simulation parameter values used are $\alpha = 1$; $\sigma^2 = 1$; $\beta = 0, 2, 4$; $p = 0.9, 0.7, 0.5$ and different values of θ_1 and θ_2. We first note that since the estimation of the trait parameters α, β, σ^2 and p do not involve marker information, the classification of parents into their true genotypes is identical to the case of two-point linkage. The EM algorithm invoked in the second stage of our proposed procedure to estimate θ_1 and θ_2 is also similar to the previous case, except that we need to consider the conditional trait genotypic distribution of the offspring given information at both the marker loci. We assume that the value of θ,

Table 3.4. Trait locus mating types among $M_iM_j \times M_kM_l$ parents, mating probabilities and probabilities of trait locus genotypes among offspring with marker genotype M_iM_k and M_jM_k[1]

g	Mating Type	Probability	$\pi_g(M_iM_k)$			$\pi_g(M_jM_k)$		
			A_1A_1	A_1a_1	a_1a_1	A_1A_1	A_1a_1	a_1a_1
1	$A_1A_1 \times A_1A_1$	${p_1}^4$	$\frac{1}{4}$	0	0	$\frac{1}{4}$	0	0
2	$A_1A_1 \times A_1a_1$	${p_1}^3p_2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0
3	$A_1a_1 \times A_1A_1$	${p_1}^3p_2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0
4	$A_1A_1 \times a_1A_1$	${p_1}^3p_2$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0
5	$a_1A_1 \times A_1A_1$	${p_1}^3p_2$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0
6	$A_1A_1 \times a_1a_1$	$2{p_1}^2{p_2}^2$	0	$\frac{1}{4}$	0	0	$\frac{1}{4}$	0
	$a_1a_1 \times A_1A_1$							
7	$A_1a_1 \times A_1a_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}\theta^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$
8	$A_1a_1 \times a_1A_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$
9	$a_1A_1 \times A_1a_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}\theta^2$
10	$a_1a_1 \times A_1a_1$	$p_1{p_2}^3$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$
11	$A_1a_1 \times a_1a_1$	$p_1{p_2}^3$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$
12	$a_1A_1 \times a_1A_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$
13	$a_1a_1 \times a_1A_1$	$p_1{p_2}^3$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$
14	$a_1A_1 \times a_1a_1$	$p_1{p_2}^3$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$
15	$a_1a_1 \times a_1a_1$	${p_2}^4$	0	0	$\frac{1}{4}$	0	0	$\frac{1}{4}$

[1] Probabilites of trait locus genotypes among offspring with marker genotypes M_jM_l and M_iM_l can be obtained by replacing θ by $(1-\theta)$ in the blocks corresponding to the genotypes M_iM_k and M_jM_k respectively in this table.

Table 3.5. Trait locus mating types among $M_1M_1M_2M_2 \times M_1m_1M_2m_2$ and $M_1M_1M_2m_2 \times M_1m_1M_2M_2$ parents, mating probabilities and probabilities of trait locus genotypes among offspring with marker genotype $M_1M_1M_2M_2$[2]

g	Mating Type	Probability	$\pi_g(M_1M_1M_2M_2 \times M_1m_1M_2m_2)$			$\pi_g(M_1M_1M_2m_2 \times M_1m_1M_2M_2)$		
			A_1A_1	A_1a_1	a_1a_1	A_1A_1	A_1a_1	a_1a_1
1	$A_1A_1 \times A_1A_1$	${p_1}^4$	1	0	0	1	0	0
2	$A_1A_1 \times A_1a_1$	${p_1}^3p_2$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$(1-\theta_1)$	θ_1	0
3	$A_1A_1 \times a_1A_1$	${p_1}^3p_2$	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	θ_1	$(1-\theta_1)$	0
4	$A_1A_1 \times a_1a_1$	${p_1}^2{p_2}^2$	0	1	0	0	1	0
	$a_1a_1 \times A_1A_1$							
5	$A_1a_1 \times A_1A_1$	$2{p_1}^3p_2$	$\frac{1}{2}$	$\frac{1}{2}$	0	$1-\theta_2$	θ_2	0
	$a_1A_1 \times A_1A_1$							
6	$A_1a_1 \times A_1a_1$	$2{p_1}^2{p_2}^2$	$\frac{(1-\theta_1)(1-\theta_2)}{2(1-\theta)}$	$\frac{1}{2}$	$\frac{\theta_1\theta_2}{2(1-\theta)}$	$(1-\theta_1)(1-\theta_2)$	$1-\theta$	$\theta_1\theta_2$
	$a_1A_1 \times A_1a_1$							
7	$A_1a_1 \times a_1A_1$	$2{p_1}^2{p_2}^2$	$\frac{\theta_1\theta_2}{2(1-\theta)}$	$\frac{1}{2}$	$\frac{(1-\theta_1)(1-\theta_2)}{2(1-\theta)}$	$\theta_1\theta_2$	$1-\theta$	$(1-\theta_1)(1-\theta_2)$
	$a_1A_1 \times a_1A_1$							
8	$A_1a_1 \times a_1a_1$	$2p_1{p_2}^3$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$1-\theta_2$	θ_2
	$a_1A_1 \times a_1a_1$							
9	$a_1a_1 \times A_1a_1$	$p_1{p_2}^3$	0	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$(1-\theta_1)$	θ_1
10	$a_1a_1 \times a_1A_1$	$p_1{p_2}^3$	0	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	θ_1	$(1-\theta_1)$
11	$a_1a_1 \times a_1a_1$	${p_2}^4$	0	0	1	0	0	1

[2] Probabilites of trait locus genotypes among offspring with marker genotypes $M_1m_1M_2m_2$, $M_1M_1M_2m_2$ and $M_1m_1M_2M_2$ can be obtained by replacing θ_1 by $(1-\theta_1)$ and θ_2 by $(1-\theta_2)$; θ_2 by $(1-\theta_2)$ and θ by $(1-\theta)$; and θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$ respectively in this table.

Table 3.6. Trait locus mating types among $M_1m_1M_2m_2 \times M_1m_1M_2m_2$ parents, mating probabilities and probabilities of trait locus genotypes among offspring with marker genotypes $M_1M_1M_2M_2$, $M_1M_1M_2m_2$ and $M_1m_1M_2m_2$[3]

g	Mating Type	Probability	$\pi_g(M_1M_1M_2M_2)$			$\pi_g(M_1M_1M_2m_2)$			$\pi_g(M_1m_1M_2m_2)$		
			A_1A_1	A_1a_1	a_1a_1	A_1A_1	A_1a_1	a_1a_1	A_1A_1	A_1a_1	a_1a_1
1	$A_1A_1 \times A_1A_1$	p_1^4	1	0	0	1	0	0	1	0	0
2	$A_1A_1 \times A_1a_1$ $A_1a_1 \times A_1A_1$	$2p_1^3p_2$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$\frac{1}{2}(1-\theta_1)\{\frac{\theta_2}{\theta}+\frac{1-\theta_2}{1-\theta}\}$	$\frac{1}{2}\theta_1\{\frac{1-\theta_2}{\theta}+\frac{\theta_2}{1-\theta}\}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0
3	$A_1A_1 \times a_1A_1$ $a_1A_1 \times A_1A_1$	$2p_1^3p_2$	$\frac{\theta_1\theta_1}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	$\frac{1}{2}\theta_1\{\frac{1-\theta_1}{\theta}+\frac{\theta_1}{1-\theta}\}$	$\frac{1}{2}(1-\theta_1)\{\frac{\theta_2}{\theta}+\frac{1-\theta_2}{1-\theta}\}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0
4	$A_1A_1 \times a_1a_1$ $a_1a_1 \times A_1A_1$	$2p_1^2p_2^2$	0	1	0	0	1	0	0	1	0
5	$A_1a_1 \times A_1a_1$	$p_1^2p_2^2$	$\frac{(1-\theta_1)^2(1-\theta_2)^2}{(1-\theta)^2}$	$\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{(1-\theta)^2}$	$\frac{\theta_1^2\theta_2^2}{(1-\theta)^2}$	$\frac{(1-\theta_1)^2\theta_2(1-\theta_2)}{\theta(1-\theta)}$	$\frac{\theta_1(1-\theta_1)\{1-2\theta_2(1-\theta_2)\}}{\theta(1-\theta)}$	$\frac{\theta_1^2\theta_2(1-\theta_2)}{\theta(1-\theta)}$	$\frac{2\theta_1(1-\theta_1)\theta_2(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$1-\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$
6	$a_1A_1 \times A_1a_1$ $A_1a_1 \times a_1A_1$	$2p_1^2p_2^2$	$\frac{\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{(1-\theta)^2}$	$1-\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{(1-\theta)^2}$	$\frac{\theta_1\theta_2(1-\theta_1)1-\theta_3)}{(1-\theta)^2}$	$\frac{\theta_1(1-\theta_1)\{1-2\theta_2(1-\theta_2)\}}{\theta(1-\theta)}$	$\frac{\theta_2(1-\theta_2)\{1-2\theta_1(1-\theta_1)\}}{\theta(1-\theta)}$	$\frac{\theta_1(1-\theta_1)\{1-2\theta_2(1-\theta_2)\}}{\theta(1-\theta)}$	$1-\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$1-\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$
7	$a_1A_1 \times a_1A_1$	$p_1^2p_2^2$	$\frac{\theta_1^2\theta_2^2}{(1-\theta)^2}$	$\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{(1-\theta)^2}$	$\frac{(1-\theta_1)^2(1-\theta_2)^2}{(1-\theta)^2}$	$\frac{\theta_1^2\theta_2(1-\theta_2)}{\theta(1-\theta)}$	$\frac{\theta_1(1-\theta_1)\{1-2\theta_2(1-\theta_2)\}}{\theta(1-\theta)}$	$\frac{(1-\theta_1)^2\theta_2(1-\theta_2)}{\theta(1-\theta)}$	$\frac{2\theta_1(1-\theta_1)\theta_2(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$1-\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$
8	$A_1a_1 \times a_1a_1$ $a_1a_1 \times A_1a_1$	$2p_1p_2^3$	0	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$\frac{1}{2}(1-\theta_1)\{\frac{\theta_2}{\theta}+\frac{1-\theta_2}{1-\theta}\}$	$\frac{1}{2}\theta_1\{\frac{1-\theta_2}{\theta}+\frac{\theta_2}{1-\theta}\}$	0	$\frac{1}{2}$	$\frac{1}{2}$
9	$a_1a_1 \times a_1A_1$ $a_1A_1 \times a_1a_1$	$2p_1p_2^3$	0	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	$\frac{1}{2}\theta_1\{\frac{1-\theta_2}{\theta}+\frac{\theta_2}{1-\theta}\}$	$\frac{1}{2}(1-\theta_1)\{\frac{\theta_2}{\theta}+\frac{1-\theta_2}{1-\theta}\}$	0	$\frac{1}{2}$	$\frac{1}{2}$
10	$a_1a_1 \times a_1a_1$	p_2^4	0	0	1	0	0	1	0	0	1

[3] Probabilites of trait locus genotypes among offspring with marker genotype $M_1M_1m_2m_2$, $m_1m_1M_2M_2$ and $m_1m_1m_2m_2$ can be obtained by replacing θ_2 by $(1-\theta_2)$ and θ by $(1-\theta)$; θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$; and θ_1 by $(1-\theta_1)$ and θ_2 by $(1-\theta_2)$ in the block corresponding to genotype $M_1M_1M_2M_2$ and those of marker genotype $m_1m_1M_2m_2$, $M_1m_1M_2M_2$ and $M_1m_1m_2m_2$ can be obtained by replacing θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$; θ_1 by θ_2 and θ_2 by θ_1; and θ_1 by $(1-\theta_2)$, θ_2 by $(1-\theta_1)$ and θ by $(1-\theta)$ respectively in the block corresponding to genotype $M_1M_1M_2m_2$ in this table.

Table 3.7. Trait locus mating types among $M_1M_1M_2m_2 \times M_1m_1M_2m_2$ parents, mating probabilities and probabilities of trait locus genotypes among offspring with marker genotypes $M_1M_1M_2M_2$ and $M_1M_1M_2m_2$[4]

g	Mating Type	Probability	$\pi_g(M_1M_1M_2M_2)$			$\pi_g(M_1M_1M_2m_2)$		
			A_1A_1	A_1a_1	a_1a_1	A_1A_1	A_1a_1	a_1a_1
1	$A_1A_1 \times A_1A_1$	${p_1}^4$	1	0	0	1	0	0
2	$A_1A_1 \times A_1a_1$	${p_1}^3p_2$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$1-\theta_1$	θ_1	0
3	$A_1A_1 \times a_1A_1$	${p_1}^3p_2$	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	θ_1	$1-\theta_1$	0
4	$A_1A_1 \times a_1a_1$	${2p_1}^2{p_2}^2$	0	1	0	0	1	0
	$a_1a_1 \times A_1A_1$							
5	$A_1a_1 \times A_1A_1$	${p_1}^3p_2$	$1-\theta_2$	θ_2	0	$\theta(1-\theta_2)+\theta_2(1-\theta)$	$\theta\theta_2+(1-\theta)(1-\theta_2)$	0
6	$a_1A_1 \times A_1A_1$	${p_1}^3p_2$	θ_2	$1-\theta_2$	0	$\theta\theta_2+(1-\theta)(1-\theta_2)$	$\theta(1-\theta_2)+th_2(1-\theta)$	0
7	$A_1a_1 \times A_1a_1$	${p_1}^2{p_2}^2$	$\frac{(1-\theta_1)(1-\theta_2)^2}{1-\theta}$	$\frac{\theta_2(1-\theta_2)}{1-\theta}$	$\frac{\theta_1{\theta_2}^2}{1-\theta}$	$2(1-\theta_1)\theta_2(1-\theta_2)$	$1-2\theta_2(1-\theta_2)$	$2\theta_1\theta_2(1-\theta_2)$
8	$a_1A_1 \times A_1a_1$	${p_1}^2{p_2}^2$	$\frac{(1-\theta_1)\theta_2(1-\theta_2)}{1-\theta}$	$1-\frac{\theta_2(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2(1-\theta_2)}{1-\theta}$	$\frac{1}{2}\theta_1\{1-\theta_2(1-\theta_2)\}$	$2\theta_2(1-\theta_2)$	$\frac{1}{2}(1-\theta_1)\{1-\theta_2(1-\theta_2)\}$
9	$A_1a_1 \times a_1A_1$	${p_1}^2{p_2}^2$	$\frac{\theta_1\theta_2(1-\theta_2)}{1-\theta}$	$1-\frac{\theta_2(1-\theta_2)}{1-\theta}$	$\frac{(1-\theta_1)\theta_2(1-\theta_2)}{1-\theta}$	$\frac{1}{2}\theta_1\{1-\theta_2(1-\theta_2)\}$	$2\theta_2(1-\theta_2)$	$\frac{1}{2}(1-\theta_1)\{1-\theta_2(1-\theta_2)\}$
10	$a_1A_1 \times a_1A_1$	${p_1}^2{p_2}^2$	$\frac{\theta_1{\theta_2}^2}{1-\theta}$	$\frac{\theta_2(1-\theta_2)}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)^2}{1-\theta}$	$2\theta_1\theta_2(1-\theta_2)$	$1-2\theta_2(1-\theta_2)$	$2(1-\theta_1)\theta_2(1-\theta_2)$
11	$A_1a_1 \times a_1a_1$	$p_1{p_2}^3$	0	$1-\theta_2$	θ_2	0	$\theta(1-\theta_2)+\theta_2(1-\theta)$	$\theta\theta_2+(1-\theta)(1-\theta_2)$
12	$a_1A_1 \times a_1a_1$	$p_1{p_2}^3$	0	θ	$1-\theta_2$	0	$\theta\theta_2+(1-\theta)(1-\theta_2)$	$\theta(1-\theta_2)+th_2(1-\theta)$
13	$a_1a_1 \times A_1a_1$	$p_1{p_2}^3$	0	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$1-\theta_1$	θ_1
14	$a_1a_1 \times a_1A_1$	$p_1{p_2}^3$	0	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	θ_1	$1-\theta_1$
15	$a_1a_1 \times a_1a_1$	${p_2}^4$	0	0	1	0	0	1

[4] Probabilites of trait locus genotypes among offspring with marker genotype $M_1M_1m_2m_2$, $M_1m_1M_2M_2$ and $M_1m_1m_2m_2$ can be obtained by replacing θ_2 by $(1-\theta_2)$ and θ by $(1-\theta)$; θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$; and θ_1 by $(1-\theta_1)$ and θ_2 by $(1-\theta_2)$ respectively in the block corresponding to the genotype $M_1M_1M_2M_2$ and that of marker genotype $M_1m_1M_2m_2$ can be obtained by replacing θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$ in the block corresponding to the genotype $M_1M_1M_2m_2$ in this table.

i.e., the recombination fraction between the two marker loci is known a priori. Based on our simulated data, we find that the histograms of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ (i.e., the estimated values of θ_1 and θ_2) are much more well behaved and concentrated than in the case of two-point linkage, especially when a marker locus is unlinked to the trait locus (i.e., when θ_1 or θ_2 is 0.5). Two representative histograms are presented in Figures 3.12 and 3.13. We also find that using three-point linkage analysis, the mean of $\widehat{\theta}_1$ (or $\widehat{\theta}_2$) is more close to the true value of θ_1 (or θ_2) than in the case of two-point linkage. The variances of the estimates in the case of multipoint linkage are also lower than those in the case of two-point linkage. Relevant statistics are provided in Table 3.8. The relative efficiency of the three-point linkage analysis over the two-point linkage analysis (defined as the ratio of the variance of the estimate in the case of two-point linkage to that in the case of three-point linkage) is found to be about 1.3.

Table 3.8. Mean and Variance of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ and 95 % C.I. of θ_1 and θ_2 for $\alpha = 5$ and $\sigma^2 = 1$

p	β	θ_1	θ_2	M($\widehat{\theta}_1$)	V($\widehat{\theta}_1$)	95% C.I. of θ_1	M($\widehat{\theta}_2$)	V($\widehat{\theta}_2$)	95% C.I. of θ_2
.9	0	0	.1	.013	.000156	(.007,.023)	.102	.000067	(.097,.115)
.7	2	.1	.1	.113	.000238	(.098,.124)	.115	.000242	(.096,.126)
.9	4	.1	.3	.170	.000945	(.132,.194)	.362	.001583	(.341,.395)
.5	2	.3	.3	.316	.000586	(.287,.328)	.314	.000588	(.288,.325)
.5	4	.3	.5	.372	.001658	(.359,.416)	.444	.000971	(.420,.498)
.7	0	.5	.1	.492	.000065	(.481,.500)	.101	.000068	(.096,.119)

3.7 Comparisons With MAPMAKER/QTL

It is of obvious interest to compare our proposed method for estimating θ with some currently-used estimation procedures. One of the most popular estimation procedures is incorporated in the MAPMAKER/QTL package (Lincoln, Daly and Lander 1993). MAPMAKER/QTL utilizes marker genotype data and values of the underlying quantitative trait of offspring in nuclear families and uses a LOD score approach (Ott 1999) to provide the

most likely position of the QTL in terms of map distances from a marker locus considered. However, unlike our proposed procedure which can utilise backcross and intercross families jointly in the analysis, MAPMAKER/QTL analyzes data separately for backcross and intercross families. Since our procedure provides estimates in terms of recombination fractions and MAPMAKER/QTL yields estimates in terms of map distances, a direct comparison of these estimates is not possible. To circumvent this problem, we convert the map distances provided by MAPMAKER/QTL to recombination fractions using the Haldane map function (1919) given by:

$$x = -\frac{1}{2} \ln (1 - 2\theta).$$

We have used the Haldane map function, which uses the no-interference assumption, because in generating our simulated data sets, we did not use any interference parameter.

Using the simulation algorithm described in Section 3.4 we generate data on backcross and intercross families separately. Note that the simulation algorithm differs only in the sense that since data on backcrosses and intercrosses are separately generated, we do not need to choose between backcross and intercross families using a random number generator as we had described in Section 3.4. We present our results based on 100 replications. The mean and variance of $\widehat{\theta}$ and the 95% confidence intervals of θ under the two estimation procedures are provided in Tables 3.9 and 3.10 for backcross and intercross families, respectively. We find that while our proposed procedure performs better for low values of dominance at the QTL, MAPMAKER/QTL performs increasingly better as the degree of dominance increases. For $\beta = 0$ and 2, our procedure yields estimates which are closer to the true value of θ and have less variances. Moreover, the confidence intervals for θ are less wide compared to those obtained using MAPMAKER/QTL. However, for $\beta = 4$, MAPMAKER/QTL provides more efficient estimates in terms of all three criteria discussed above. To summarize, our classification-EM procedure is more efficient than the procedure incorporated in MAPMAKER/QTL for both backcross and intercross data, unless the degree of dominance at the QTL is very high.

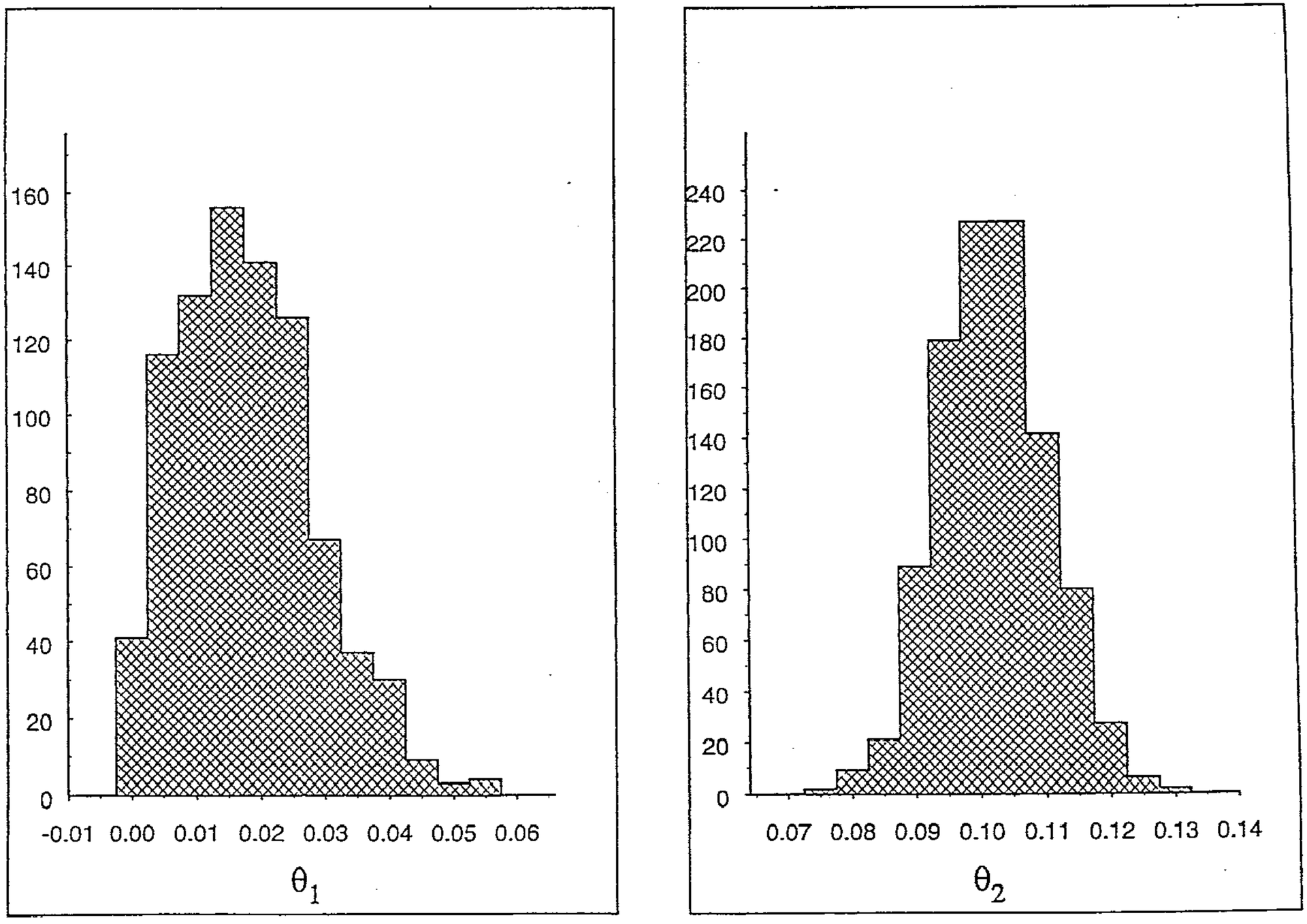
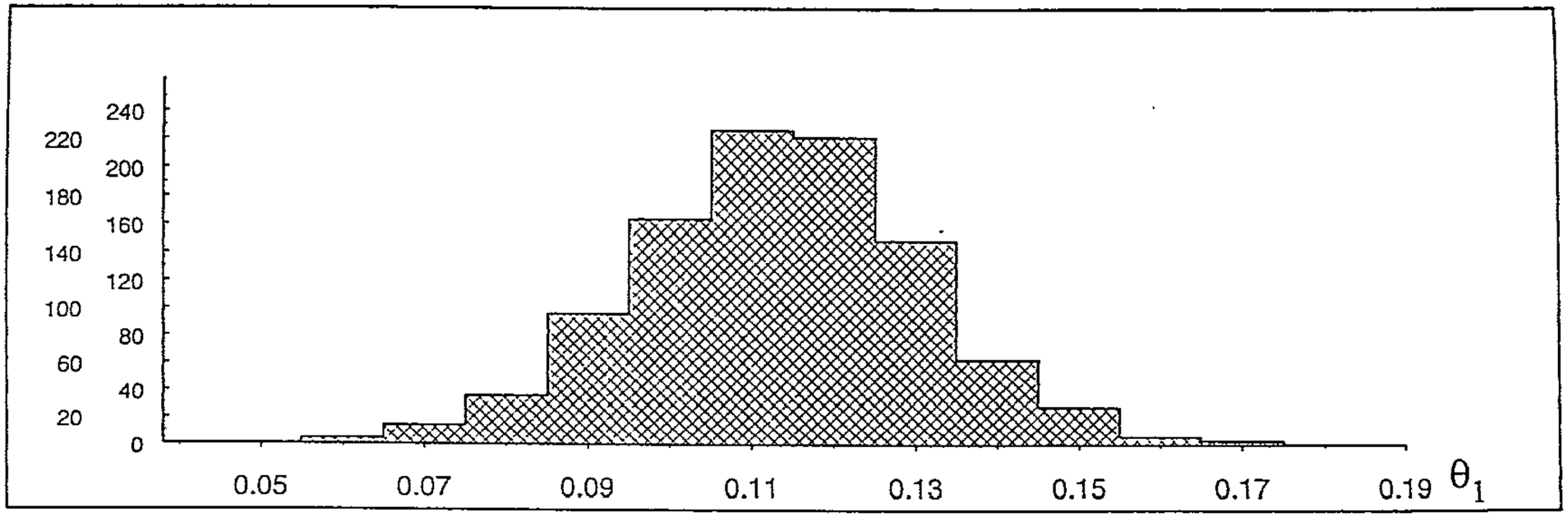


Figure 3.12. Empirical frequency distributions of $\widehat{\theta_1}$ and $\widehat{\theta_2}$ for simulation parameter values $p = .9, \alpha = 5, \beta = 0, \sigma^2 = 1, \theta_1 = 0, \theta_2 = 0.1$.



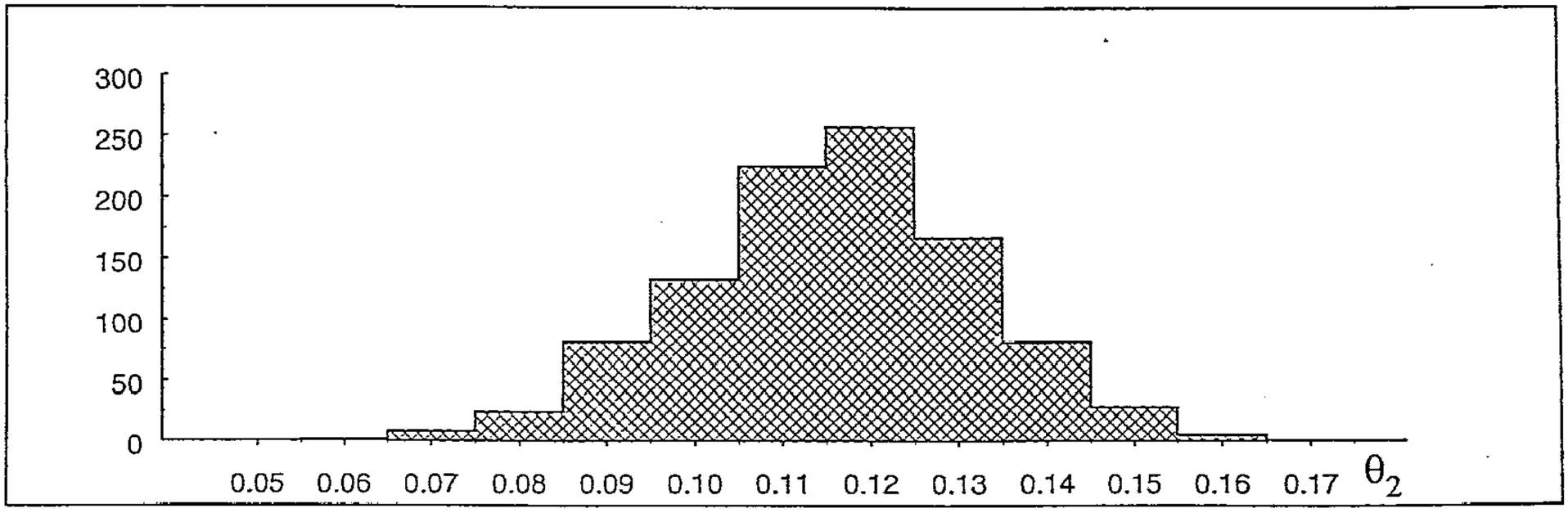


Figure 3.13. Empirical frequency distributions of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for simulation parameter values $p = .7, \alpha = 5, \beta = 2, \sigma^2 = 1, \theta_1 = 0.1, \theta_2 = 0.1$.

Table 3.9. Comparison between the EM procedure and MAPMAKER/QTL in terms of the Mean and Variance of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ and 95 % C.I. of θ_1 and θ_2 for $\alpha = 5$ and $\sigma^2 = 1$ for backcross families.

p	β	θ_1	θ_2	EM						MAPMAKER/QTL					
				$M(\widehat{\theta}_1)$	$V(\widehat{\theta}_1)$	95% C.I. of θ_1	$M(\widehat{\theta}_2)$	$V(\widehat{\theta}_2)$	95% C.I. of θ_2	$M(\widehat{\theta}_1)$	$V(\widehat{\theta}_1)$	95% C.I. of θ_1	$M(\widehat{\theta}_2)$	$V(\widehat{\theta}_2)$	95% C.I. of θ_2
.7	0	.05	.1	.053	.000099	(.046,.062)	.105	.000108	(.096,.110)	.043	.000305	(.034,.053)	.094	.000381	(.077,.104)
.9	0	.1	.1	.104	.000106	(.098,.106)	.102	.000102	(.097,.107)	.093	.000428	(.068,.105)	.093	.000430	(.069,.102)
.5	0	.1	.3	.106	.000108	(.096,.112)	.310	.000437	(.295,.320)	.095	.000304	(.082,.108)	.294	.000582	(.267,.305)
.9	0	.3	.3	.303	.000397	(.295,.310)	.303	.000402	(.295,.311)	.290	.000711	(.256,.307)	.292	.000704	(.260,.306)
.7	0	.5	.1	.495	.000076	(.492,.500)	.104	.000106	(.096,.109)	.490	.000157	(.467,.498)	.090	.000378	(.076,.103)
.5	0	.3	.5	.308	.000443	(.294,.318)	.498	.000049	(.495,.500)	.294	.000589	(.266,.310)	.491	.000098	(.466,.495)
.7	2	.05	.1	.058	.000308	(.051,.068)	.108	.000302	(.096,.113)	.038	.000328	(.031,.048)	.093	.000376	(.072,.105)
.9	2	.1	.1	.106	.000224	(.098,.110)	.105	.000231	(.098,.111)	.090	.000432	(.068,.103)	.088	.000437	(.066,.105)
.5	2	.1	.3	.109	.000217	(.095,.118)	.315	.000525	(.293,.323)	.096	.000295	(.080,.110)	.292	.000612	(.264,.306)
.9	2	.3	.3	.305	.000586	(.291,.315)	.307	.000566	(.293,.315)	.289	.000703	(.252,.305)	.290	.000715	(.255,.308)
.7	2	.5	.1	.493	.000102	(.489,.500)	.107	.000298	(.094,.112)	.490	.000166	(.457,.496)	.092	.000388	(.070,.105)
.5	2	.3	.5	.317	.000538	(.295,.324)	.495	.000057	(.492,.500)	.293	.000601	(.260,.313)	.492	.000105	(.460,.497)
.7	4	.05	.1	.081	.000852	(.065,.094)	.158	.000965	(.145,.171)	.036	.000338	(.030,.051)	.095	.000392	(.073,.112)
.9	4	.1	.1	.144	.000897	(.131,.156)	.140	.000913	(.128,.151)	.092	.000461	(.065,.105)	.091	.000455	(.066,.108)
.5	4	.1	.3	.169	.001054	(.142,.190)	.381	.001277	(.353,.405)	.092	.000321	(.077,.114)	.287	.000620	(.261,.306)
.9	4	.3	.3	.362	.001196	(.345,.396)	.364	.001205	(.344,.394)	.287	.000718	(.250,.307)	.287	.000692	(.253,.311)
.7	4	.5	.1	.423	.001089	(.408,.490)	.162	.000983	(.139,.175)	.492	.000164	(.458,.495)	.087	.000401	(.068,.107)
.5	4	.3	.5	.383	.001254	(.355,.404)	.435	.001008	(.422,.496)	.290	.000619	(.255,.316)	.488	.000113	(.458,.498)

Table 3.10. Comparison between the EM procedure and MAPMAKER/QTL in terms of the Mean and Variance of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ and 95 % C.I. of θ_1 and θ_2 for $\alpha = 5$ and $\sigma^2 = 1$ for intercross families.

p	β	θ_1	θ_2	EM						MAPMAKER/QTL					
				M($\widehat{\theta}_1$)	V($\widehat{\theta}_1$)	95% C.I. of θ_1	M($\widehat{\theta}_2$)	V($\widehat{\theta}_2$)	95% C.I. of θ_2	M($\widehat{\theta}_1$)	V($\widehat{\theta}_1$)	95% C.I. of θ_1	M($\widehat{\theta}_2$)	V($\widehat{\theta}_2$)	95% C.I. of θ_2
.7	0	.05	.1	.055	.000108	(.044,.064)	.105	.000101	(.096,.112)	.041	.000311	(.032,.053)	.094	.000359	(.075,.106)
.9	0	.1	.1	.103	.000113	(.098,.108)	.102	.000109	(.097,.110)	.093	.000425	(.067,.105)	.091	.000442	(.072,.103)
.5	0	.1	.3	.105	.000122	(.094,.114)	.312	.000452	(.292,.324)	.094	.000318	(.080,.113)	.289	.000576	(.266,.305)
.9	0	.3	.3	.304	.000413	(.293,.312)	.303	.000419	(.292,.314)	.291	.000674	(.252,.306)	.294	.000727	(.263,.310)
.7	0	.5	.1	.493	.000091	(.488,.500)	.106	.000112	(.096,.113)	.489	.000163	(.465,.498)	.093	.000364	(.073,.105)
.5	0	.3	.5	.311	.000476	(.292,.326)	.497	.000058	(.494,.500)	.286	.000600	(.268,.313)	.490	.000103	(.470,.496)
.7	2	.05	.1	.058	.000316	(.050,.071)	.110	.000322	(.095,.117)	.038	.000335	(.029,.050)	.092	.000366	(.076,.108)
.9	2	.1	.1	.112	.000307	(.096,.118)	.113	.000312	(.097,.120)	.089	.000444	(.066,.103)	.086	.000439	(.068,.107)
.5	2	.1	.3	.115	.000357	(.094,.120)	.318	.000566	(.292,.326)	.090	.000326	(.082,.107)	.290	.000605	(.264,.305)
.9	2	.3	.3	.308	.000543	(.292,.315)	.307	.000553	(.293,.316)	.290	.000722	(.254,.308)	.288	.000731	(.250,.306)
.7	2	.5	.1	.491	.000114	(.482,.500)	.110	.000335	(.093,.118)	.490	.000168	(.462,.497)	.090	.000379	(.072,.107)
.5	2	.3	.5	.319	.000571	(.291,.326)	.494	.000062	(.491,.500)	.286	.000608	(.263,.313)	.491	.000104	(.466,.496)
.7	4	.05	.1	.093	.000965	(.061,.105)	.172	.001104	(.142,.174)	.039	.000340	(.030,.054)	.094	.000364	(.075,.109)
.9	4	.1	.1	.153	.001006	(.128,.160)	.159	.001018	(.127,.157)	.094	.000438	(.067,.103)	.091	.000448	(.068,.105)
.5	4	.1	.3	.186	.001249	(.148,.199)	.392	.001422	(.356,.410)	.097	.000325	(.076,.115)	.288	.000613	(.257,.311)
.9	4	.3	.3	.375	.001371	(.352,.402)	.378	.001397	(.356,.403)	.288	.000730	(.253,.310)	.285	.000733	(.252,.308)
.7	4	.5	.1	.415	.001116	(.408,.485)	.174	.001106	(.139,.177)	.494	.000166	(.459,.497)	.091	.000370	(.070,.105)
.5	4	.3	.5	.396	.001415	(.353,.409)	.428	.001034	(.420,.494)	.288	.000615	(.258,.315)	.490	.000109	(.468,.497)

Table 3.11. Mean and Variance of $\widehat{\theta}$ and 95% Confidence Interval of θ Using Posterior Probabilities for $\alpha = 5$, $\sigma^2 = 1$, $p = .9, 0.7, 0.5$; $\beta = 0, 2, 4$; $\theta = 0, 0.1, 0.3, 0.5$.

p	True θ	β	Mean($\widehat{\theta}$)	Var($\widehat{\theta}$)	95% C.I. of θ
0.9	0	0	0.018	0.000182	(0.011 ,0.028)
		2	0.040	0.000234	(0.021 ,0.044)
		4	0.053	0.000316	(0.038 ,0.067)
	0.1	0	0.104	0.000091	(0.098 ,0.116
		2	0.116	0.000274	(0.096 ,0.124)
		4	0.131	0.000457	(0.115 ,0.143)
	0.3	0	0.305	0.000471	(0.294 ,0.315)
		2	0.310	0.000619	(0.292 ,0.323)
		4	0.331	0.000715	(0.316 ,0.347)
	0.5	0	0.484	0.000384	(0.458 ,0.500)
		2	0.477	0.000353	(0.443 ,0.500)
		4	0.465	0.000505	(0.432 ,0.500)
0.7	0	0	0.014	0.000106	(0.008 ,0.024)
		2	0.025	0.000165	(0.017 ,0.032)
		4	0.036	0.000255	(0.021 ,0.052)
	0.1	0	0.102	0.000082	(0.098 ,0.110)
		2	0.111	0.000227	(0.097 ,0.120)
		4	0.119	0.000336	(0.109 ,0.130)
	0.3	0	0.302	0.000404	(0.295 ,0.311)
		2	0.311	0.000508	(0.294 ,0.322)
		4	0.320	0.000609	(0.310 ,0.335)
	0.5	0	0.495	0.000084	(0.485 ,0.500)
		2	0.490	0.000281	(0.474 ,0.500)
		4	0.479	0.000362	(0.455 ,0.500)
0.5	0	0	0.010	0.000082	(0.005 ,0.018)
		2	0.017	0.000112	(0.010 ,0.025)
		4	0.023	0.000194	(0.014 ,0.038)
	0.1	0	0.102	0.000075	(0.098 ,0.107)
		2	0.105	0.000186	(0.098 ,0.115)
		4	0.111	0.000265	(0.102 ,0.120)
	0.3	0	0.300	0.000257	(0.297 ,0.308)
		2	0.305	0.000338	(0.296 ,0.315)
		4	0.313	0.000426	(0.301 ,0.326)
	0.5	0	0.498	0.000064	(0.491 ,0.500)
		2	0.494	0.000167	(0.485 ,0.500)
		4	0.491	0.000245	(0.477 ,0.500)

3.8 Effect of Using Posterior Probabilities at the Second Stage

In our proposed EM algorithm, we classify each parent into a most likely trait genotype using Bayes' 0-1 classification rule (see Section 3.3). As we note from our simulation results (Section 3.4), the performance of our estimator is strongly dependent on the percentage of correct genotypic classification of the parents. The estimator does not perform well for high degrees of dominance in the trait.

An alternative approach, although computationally more heavy, is to incorporate the actual posterior probabilities of the various genotypes for each parent at the second stage of estimation of θ, instead of classifying a parent into that QTL genotype for which the posterior probability is the maximum [which is equivalent to using one of posterior probability distributions (1,0,0),(0,1,0) or (0,0,1), corresponding to the three QTL genotypes]. We investigate whether the performance of our estimator improves by using this strategy.

For these investigations, we cannot use the classification rule given by Equation 3.1. We note that the posterior probability of the j^{th} parent of the i^{th} family belonging to the t^{th} trait genotype is given by $\widehat{z_{ijt}}$, $i = 1, 2, \ldots, K$; $j = 1, 2$, $t = 1, 2, 3$, which will be used in the second stage of our estimation procedure.

In the present set-up, we need to redefine G_{i1}, G_{i2} and P_{ijn} as:

$$\begin{aligned} G_{i1}, G_{i2} &= \text{trait genotypes of parents in the } i^{th} \text{ family.} \\ {P_{ijn}}^{l,m} &= P(H_{ij} = \gamma_n | G_{i1} = \gamma_l, G_{i2} = \gamma_m, M_{i1}, M_{i2}, M_{ij}), \\ &\quad \text{where } \gamma_1 = A_1A_1,\ \gamma_2 = A_1a_1,\ \gamma_3 = a_1a_1. \end{aligned}$$

Similarly, Q_{ijn} has to be redefined as:

$$\begin{aligned} {Q_{ijn}}^{l,m} &= P(H_{ij} = \gamma_n | G_{i1} = \gamma_l, G_{i2} = \gamma_m, M_{i1}, M_{i2}, M_{ij}, y_{ij}) \\ &= \frac{{P_{ijn}}^{l,m} f_n(y_{ij})}{\sum_{n=1}^{3} {P_{ijn}}^{l,m} f_n(y_{ij})}, \end{aligned}$$

Thus, at the trait genotype classification stage of each offspring, we need to classify the offspring for every possible trait genotype combination of the parents (i.e., for each combination of $(l, m), l, m = 1, 2, 3$). The likelihood

function $L(\theta)$ is identical to Equation 3.2 except that each $L_i(\theta)$ comprises more complex mixture components than in the example given in Section 3.3, with the mixture proportions being functions of the product $(\widehat{z_{i1l}} \times \widehat{z_{i2m}})$, for each combination of (l, m), i.e., the posterior trait genotype probabilities of the parents in the i^{th} family.

We use simulated data with the same sets of trait and linkage parameters as in Section 3.4 to compare the performances of the estimators under the two strategies. The results based on the present strategy is given in Table 3.11. Comparing this table with Table 3.1, we find that mean of the estimates of θ are, in general, more close to the true values of θ and have less variance compared to our earlier procedure based on parental classification. Moreover, the confidence intervals of θ are less wide under this srategy. The performance of the two procedures are similar when the proportion of homozygotes is high and dominance at the trait locus is low. However, as the proportion of heterozygotes or the degree of dominance at the trait locus increases, the performance of this procedure becomes increasingly better. This is due to the fact that unlike our proposed procedure, this procedure does not depend on the performance of parental trait genotype classification. Thus, the performance of this procedure is not affected by parameters which increase the misclassification probabilities like trait locus heterozygosity and dominance. The estimation procedure using posterior probabilities, therefore, has more desirable statistical properties than our previous estimation procedure using classification of parents into a specific QTL genotype. Data analysis using this new strategy is, however, computationally more complex.

Table 3.12. Mean and Variance of $\widehat{\theta}$ and 95% Confidence Interval of θ based on 133 randomly sampled nuclear families for simulation parameter values $a = 5$, $\sigma^2 = 1$, $\theta = 0.1$. [5]

p	MIF	SDIF	β	Mean$(\widehat{\theta})$	Var$(\widehat{\theta})$	95% C.I. of θ
0.9	100.048	5.003	4	0.178	0.000952	(0.134 ,0.199)
0.7	100.023	5.016	2	0.126	0.000334	(0.096 ,0.124)
0.5	99.989	5.008	0	0.110	0.000122	(0.098 ,0.121)

[5]MIF and SDIF denote the empirical mean and standard deviation of the number of informative families

3.9 Evidence that analyzing data on restricted-sense "informative" nuclear families is equivalent to analyzing data on a larger number, predictable *a priori*, of randomly-sampled families

In our simulations, we have generated data on nuclear families which are "informative" in the sense that at least one of the parents is heterozygous at the marker locus. These data were then used for understanding the performance of the estimator. In Section 2.2, we have argued that this strategy is equivalent to analyzing data on a larger number of randomly-sampled families; that is, the performance of the estimator will be the same whether we use *NFAM* "informative" families or use the larger number of randomly-sampled families required to produce *NFAM* "informative" families. Formulas for calculating the expected number of families to be randomly sampled for this purpose have also been provided in Section 2.2. We now provide empirical evidence to confirm our claim that these two sampling procedures are indeed equivalent for understanding the behaviour of the estimator.

With 0.5 as the frequencies of the alleles M_1 and m_1 at a biallelic marker locus as used in our simulations, the expected number of nuclear families to generate *NFAM* informative families (*NFAM* is the number of replicate families used in our simulations) is $NFAM \times 4/3$. [Analytical formula has been provided in Section 2.2.] Thus, we generate data on 133 randomly chosen nuclear families in each replication so that the average number of informative families is 100 (= *NFAM*). The results based on 133 randomly chosen nuclear families are given in Table 3.12 for different sets of parameter values. We first note that for these sets of parameter values, the average number of "informative" families is very close to 100 with a very small standard deviation (≈ 5). Second, we find from Table 3.12 that the means and variances of $\widehat{\theta}$ based on 133 randomly chosen families are very close to those (Table 3.1) obtained from data on 100 informative families. Moreover, the confidence intervals of θ are also similar. Thus, the properties of the estimator are equivalently revealed either by analyzing data on 100 "informative"

families or on 133 randomly chosen families. We have also investigated the performance of the estimator for other sets of parameter values. The results, as expected, were analogous to those presented above; details are not provided for brevity. This provides an empirical evidence of our theoretical claim made in Section 2.2.

3.10 Discussion and Overview

The proposed method of linkage detection exploits the fact that knowledge of parental genotypes at the QTL greatly eases statistical estimation of θ. Since for a quantitative character, the QTL genotype of an individual cannot be inferred with certainty because of intrinsic variability within genotype classes, we have used the EM algorithm coupled with a Bayes' classification procedure to classify parents into QTL genotype classes. A similar EM approach to estimate the trait parameters was used by Kao and Zeng (1997) in mapping a quantitative trait locus in an interval flanked by two markers. However their procedure was based on inherent knowledge of haplotype information which is not readily available in human genetic studies. Moreover the effect of marker genotype on trait value was assumed to be linear. The procedure proposed by us does not use these assumptions. In our procedure, estimates of trait parameters and recombination fraction are obtained. The estimates of trait parameters are used in inferring the parental QTL genotypes. The estimation of trait parameters, in the first stage of the proposed two-stage procedure, can be based either on data of a random sample of individuals or on data of parents (assumed to be unrelated) in families. The first stage of our procedure does not use marker genotype information. Thus, even if families are sampled, it will be prudent to initially obtain only measurements of the quantitative trait on parents. The EM algorithm implemented in the first stage will provide estimates of trait parameters. Having obtained these estimates, we can classify parents in families into major QTL genotypes, using the proposed classification rule. This enables identification of potentially informative families (i.e., at least one parent heterozygous at the major QTL). Then, the investigator can obtain genotype information at marker loci on both parents and measurements

of quantitative trait on offspring in those families in which at least one parent is doubly heterozygous. Thus, the proposed two-stage procedure provides cost effectiveness in terms of data collection. In the second stage, based on data on only informative families, the proposed EM algorithm provides the maximum likelihood estimate of the recombination fraction between a marker locus and the major QTL. We note that in the context of human pedigree analysis, the EM algorithm was first applied by Ott [1977].

We have shown that our proposed method results in virtually error free classification of parental QTL genotypes, unless the dominance effect is very large. We have also shown, using simulations, that for a wide range of parameter values, that corresponds to widely different values of the proportion of variance in QT explained by the major locus, the estimates of recombination fractions and the power to detect linkage are quite good for reasonable sample sizes. We have shown that our method performs more efficiently when data on multiple markers flanking the trait locus are used. One major advantage of the proposed method is that the estimation of recombination fractions is not as strongly tied to estimates of QTL parameters as in lod-score analysis. Through the use of EM algorithm, the present procedure extracts appropriate information from the quantitative data and then uses a Bayesian classification rule to transform the QT data to qualitative genotypes before estimating θ from the transformed data by a likelihood-based method. This reduces the impact of error in estimating trait parameters on the estimate of θ. Because of the weak dependence of $\widehat{\theta}$ on estimates of trait parameters, and because no separate segregation and linkage analyses need to be performed in the present approach, the earlier observations that model misspecification can seriously affect estimates of trait parameters [Dizier et al. 1993; Atwood et al. 1995] and that prior segregation analysis can reduce the power to detect linkage [Atwood and Slifer 1997] become less relevant. Further this approach leads to a considerable reduction in computational load, which in usual parametric segregation and linkage analyses of a QT can be sufficiently heavy to require the use of a supercomputer [Atwood and Slifer 1997]. Joint segregation and linkage is computationally even more expensive, although there are some indications that it may be more powerful [Gauderman et al. 1997] than separate segregation and linkage analyses. We have shown that for reasonable levels of linkage heterogeneity, the proposed

method performs quite well. Model misspecification, treating a two-locus QT as a single locus QT, even though yields biased estimates of θ, leads to gross errors in inference only when both trait loci have nearly equal effects. Thus, the present method appears to be quite useful and robust for QTL mapping. Compared to numerical maximization of the likelihood of parental and offspring data, on all families jointly with respect to all parameters (recombination fraction, trait parameters and allele frequencies), the proposed stagewise procedure using the EM algorithm is computationally much more efficient and provides reduction of data collection costs.

We have compared our Classification-EM procedure with a currenly-used QTL mapping procedure as implemented in the MAPMAKER/QTL (Ver. 1.1) package (Lincoln et al. 1993). We have shown that our proposed procedure outperforms the procedure implemented in this package, unless the degree of dominance at the trait locus is very high. We further note that while in our procedure data on backcross and intercross families can be analyzed jointly, data on these two types of families need to be analyzed separately using MAPMAKER/QTL.

Since it was found that our Classification-EM procedure does not perform well when the degree of dominance is high, we investigated the performance of another related, but computationally more heavy, procedure. In the first stage of our Classification-EM procedure, the posterior probabilities of QT genotypes of each parent are calculated and the parent is classified into that QT genotype class for which the posterior probability is the maximum. These "classified" parental data are then used in the second stage of estimation of θ. Instead of classifying each parent into an inferred QT genotype class, one may use the three values of the posterior probabilities corresponding to the three QT genotypes that are estimated in the first stage, to build up likelihood function used in the second stage. The likelihood function, when the values of posterior probabilities are used, obviously involves a greater number of mixture terms than when "classified" parental genotype data are used. Using simulated data, and the method in which the estimated values of the posterior probabilities are used, we find significant improvements in the efficiency of estimation of the recombination fraction, even when the degree of dominance at the QTL is high.

Chapter 4

Mapping in the Presence of Epistatic Interactions

4.1 Introduction

Experimental studies on quantitative characters in plants (see, e.g., Tanksley 1993), dairy cattle (Georges et al. 1995), mice (Berrethini et al. 1994), rat (Schork et al. 1995), etc. have revealed that quantitative traits may often be determined by multiple loci. There is also increasing evidence (Lark et al. 1995; Coupland 1995; Fijneman et al. 1996; van Wezel et al. 1996, Chang 1999) that alleles at the loci determining a quantitative trait may interact epistatically. In Chapter 3, we have primarily considered the quantitative trait as being determined by a single major locus. Although, a quantitative trait may be determined by multiple loci, often the effects of the loci on the trait are highly variable, so that it may suffice to consider only those loci which have large effects, which are generally few in number, and are, therefore, more easily mapped than those loci with small effects. However, in view of the experimental observations cited above, it is necessary to consider multiple loci and the possibility of epistatic interactions among the loci. Statistical methods for mapping quantitative trait loci have generally ignored epistatic interactions (Frankel and Schork 1996). In this Chapter, we propose two computationally simple statistical techniques, by extending some traditional techniques, for mapping QTLs when the trait is actually

determined by a set of unlinked, autosomal, epistatically interacting loci. The two methodologies pertain to two different types of data; nuclear families and sib-pairs. We also elaborately examine the performance of these modified and extended methodologies from various statistical considerations which results in insights on the performances of these methodologies under various scenarios.

We first consider parental and offspring data separately on families in which only one parent is heterozygous at the marker locus and those in which both parents are heterozygous and suitably modify the estimator proposed by Jayakar (1970) based on variance components. We show, based primarily on the widths of confidence intervals, that for a wide range of parameter values the proposed estimator is quite efficient. Additionally, we suggest a non-parametric procedure for testing null hypotheses regarding θ and show that the power function of the test has desirable statistical properties. We also show that analyses of data ignoring epistatic interactions, when in fact these are present, may lead to grossly inaccurate inferences about linkage. However, the variance of the proposed estimator is found to be larger than that of the maximum likelihood estimator (m.l.e.). Our results provide statistical insights on the major reasons why Jayakar's (1970) estimators do not perform well in practice and are, therefore, not used.

Our second data type includes quantitative trait values of sib-pairs and their estimated marker i.b.d. scores. One of the popular statistical techniques to analyze such data is based on the regression of squared difference in trait values of sib pairs on their estimated marker i.b.d. scores. Under a very general setup, even in the presence of dominance and epistatic effects, Tiwari and Elston (1997) have extended the classical regression method for QTL mapping when the trait is controlled by two unlinked, autosomal, biallelic loci. Since this general model involves too many parameters, insights into effects of variation of individual parameters on the performance of the method were difficult to obtain. We, therefore, examine the performance of the method under the specific digenic interaction model. We also extend the method to the case of a quantitative trait that is controlled by multiple unlinked loci. The competing strategies of analyzing the data by simultaneous, as opposed to sequential, consideration of the markers are quantitatively assessed using simulation studies. As is intuitively expected, the simultaneous

strategy is found to be more optimal and cost-effective.

4.2 Modification of Jayakar's (1970) Procedure for Nuclear Family Data

Based on observations on members of nuclear families, that is observations on parents and their offspring, Jayakar (1970) derived an estimator of θ, the recombination fraction between the putative trait locus (a single locus is assumed to determine the quantitative trait) and a marker locus, as a function of the variances of the quantitative trait in the population and among offspring of specified marker genotypes within and across various parental mating types. In this Section, we modify Jayakar's estimator of θ to the case of a quantitative trait being determined by multiple, unlinked, epistatically interacting, loci. We also examine the statistical properties of the modified estimator.

We assume that a quantitative trait is controlled by L autosomal biallelic loci. Let A_l and a_l denote the alleles at the l^{th} locus, $l = 1, 2, \ldots, L$. We assume that the loci are mutually unlinked and that the population is in Hardy-Weinberg equilibrium in respect of each of these loci. Let the allele frequencies at the l^{th} locus be denoted as p_l and $q_l = 1 - p_l$. Let the expectation, $E(Y)$, of the quantitative character, Y, given the genotypes of the l^{th} locus be $\alpha_l, 0$ and $-\alpha_l$ for A_lA_l, A_la_l and a_la_l, respectively, i.e., there is no dominance in any of the QTLs. We assume that the variance of Y within each single-locus genotype is the same. The epistatic interaction effects between the QTLs have been described in Section 2.1. The conditional expectation and variance of Y given the genotypes at all the L QTLs have also been discussed in the same Section.

We assume, without loss of generality, that the trait locus (A_1, a_1) is linked to an autosomal biallelic codominant marker locus with alleles M_1 and m_1. These two loci are assumed to be in linkage equilibrium. Let the recombination fraction between the loci be denoted as θ. Our purpose is to estimate θ from observations on the quantitative trait and the genotypes at the marker locus on members of families.

To ensure informativeness for linkage, it is necessary to only consider

matings for which at least one parent is heterozygous at the marker locus. We shall distinguish, at the marker locus, the two types of families, backcross ($M_1M_1 \times M_1m_1$) and intercross ($M_1m_1 \times M_1m_1$). It is obvious that $m_1m_1 \times M_1m_1$ families can be handled in the same manner as $M_1M_1 \times M_1m_1$ families by relabelling alleles. Families in which neither parent is heterozygous at the marker locus are excluded from analyses.

Since only the (A_1, a_1) locus is linked to (M_1, m_1), information on θ is contained only in two locus gametotypes obtainable upon joint consideration of these two loci only. The probabilities of offspring genotypes at the (A_1, a_1) locus for various backcross parental genotypic matings has been provided in Table 2.2.

For any particular parental genotypic mating $g(= 1, 2, \ldots, 11)$, let π_{gi} and Y_{gi} denote the probability and value of the quantitative trait, respectively, for an offspring of type i; $i = 1 = A_1A_1M_1M_1$, $i = 2 = A_1a_1M_1M_1$, $i = 3 = a_1a_1M_1M_1$, $i = 4 = A_1A_1M_1m_1$, $i = 5 = A_1a_1M_1m_1$ and $i = 6 = a_1a_1M_1m_1$. Let,

$T_g = Var\{Y_{g1}, Y_{g2}, \ldots, Y_{g6}\}$ = Variance of trait values among all offspring;

$V_{g1} = Var\{Y_{g1}, Y_{g2}, Y_{g3}\}$ = Variance of trait values among offspring of marker genotype M_1M_1;

$V_{g2} = Var\{Y_{g4}, Y_{g5}, Y_{g6}\}$ = Variance of trait value among offspring of marker genotype M_1m_1;

$V_g = V_{g1} + V_{g2}$; and

V_p = Variance of the trait value Y in the whole population.

For the model considered,

$$
\begin{aligned}
T = E_g(T_g) \;=\; & \sigma^2 + p_1q_1\{\alpha_1 + \sum_{j=2}^{L}\Delta_{1j}(p_j - q_j)\}^2 + 2\sum_{j=2}^{L} p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}^2 \\
& + 2\sum_{i=2}^{L}\sum_{j>i}^{L}\Delta_{ij}{}^2\{(p_i{}^2 + q_i{}^2)(p_j{}^2 + q_j{}^2) - (p_i - q_i)^2(p_j - q_j)^2\} \\
& + 4p_1q_1\sum_{j=2}^{L}\Delta_{1j}{}^2p_jq_j + 4\sum_{i=2}^{L}\sum_{j\neq i}^{L}\Delta_{ij}(p_i - q_i)p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}.
\end{aligned}
$$

$$
V = E_g(V_g) \;=\; 2\sigma^2 + p_1q_1\{1 + 4\theta(1-\theta)\}\{\alpha_1 + \sum_{j=2}^{L}\Delta_{1j}(p_j - q_j)\}^2
$$

$$+4\sum_{j=2}^{L} p_j q_j \{\alpha_j + \Delta_{1j}(p_1 - q_1)\}^2 + 8p_1 q_1 \sum_{j=2}^{L} {\Delta_{1j}}^2 p_j q_j$$

$$+4\sum_{i=2}^{L}\sum_{j>i}^{L} {\Delta_{ij}}^2 \{({p_i}^2 + {q_i}^2)({p_j}^2 + {q_j}^2) - (p_i - q_i)^2(p_j - q_j)^2\}$$

$$+8\sum_{i=2}^{L}\sum_{j\neq i}^{L} \Delta_{ij}(p_i - q_i) p_j q_j \{\alpha_j + \Delta_{1j}(p_1 - q_1)\}.$$

The variance of the trait value Y, in the whole population is:

$$\begin{aligned} V_p &= \sigma^2 + 2p_1 q_1 \{\alpha_1 + \sum_{j=2}^{L} \Delta_{1j}(p_j - q_j)\}^2 + \sum_{j=2}^{L} p_j q_j \{\alpha_j + \Delta_{1j}(p_1 - q_1)\}^2 \\ &\quad +2\sum_{i=2}^{L}\sum_{j>i}^{L} {\Delta_{ij}}^2 \{({p_i}^2 + {q_i}^2)({p_j}^2 + {q_j}^2) - (p_i - q_i)^2(p_j - q_j)^2\} \\ &\quad +4p_1 q_1 \sum_{j=2}^{L} {\Delta_{1j}}^2 p_j q_j + 4\sum_{i=2}^{L}\sum_{j\neq i}^{L} \Delta_{ij}(p_i - q_i) p_j q_j \{\alpha_j + \Delta_{1j}(p_1 - q_1)\}. \end{aligned}$$

From the above equations, we get:

$V_p - T = p_1 q_1 \{\alpha_1 + \sum_{j=2}^{L} \Delta_{1j}(p_j - q_j)\}^2$

and, $2T - V = p_1 q_1 \{\alpha_1 + \sum_{j=2}^{L} \Delta_{1j}(p_j - q_j)\}^2 (1 - 2\theta)^2$

Hence, $(1 - 2\theta)^2 = \frac{2T - V}{V_p - T}$

or,

$$\theta = \frac{1}{2}[1 - \sqrt{\frac{2T - V}{V_p - T}}] \tag{4.1}$$

The estimate of θ is obtained from the above equation by plugging in observed values of T, V and V_p. We note that although Equation (4.1) is independent of the parameters underlying the model governing the trait and marker loci (i.e., αs, Δs, σ^2 and allele frequencies), the sampling distribution of the proposed estimator of θ is a function of these parameters as is evident from the expressions of T, V and V_p.

When the family is an intercross ($M_1 m_1 \times M_1 m_1$), the probabilities of different offspring types for various parental genotypic matings are given in Table 2.3. As in the case of backcross, let, for any particular genotypic

mating $g(=1,2,\ldots,10)$, π_{gi} and Y_{gi} denote the probability and quantitative trait value, respectively, for an offspring of type i; $i = 1 = A_1A_1M_1M_1$, $i = 2 = A_1a_1M_1M_1, \ldots, i = 9 = a_1a_1m_1m_1$.

Let,

$V_{g1} = Var\{Y_{g1}, Y_{g2}, Y_{g3}\}$

$V_{g2} = Var\{Y_{g4}, Y_{g5}, Y_{g6}\}$

$V_{g3} = Var\{Y_{g7}, Y_{g8}, Y_{g9}\}$

If $V_1 = E_g(V_{g1}), V_2 = E_g(V_{g2})$ and $V_3 = E_g(V_{g3})$, we have:

$$\begin{aligned} V_1 = E_g(V_{g1}) &= \sigma^2 + 4p_1q_1\theta(1-\theta)\{\alpha_1 + \sum_{j=2}^{L}\Delta_{1j}(p_j - q_j)\}^2 \\ &\quad +2\sum_{j=2}^{L}p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}^2 + 4p_1q_1\sum_{j=2}^{L}{\Delta_{1j}}^2p_jq_j \\ &\quad +2\sum_{i=2}^{L}\sum_{j>i}^{L}{\Delta_{ij}}^2\{({p_i}^2 + {q_i}^2)({p_j}^2 + {q_j}^2) - (p_i - q_i)^2(p_j - q_j)^2\} \\ &\quad +4\sum_{i=2}^{L}\sum_{j\neq i}^{L}\Delta_{ij}(p_i - q_i)p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}, \end{aligned}$$

$$\begin{aligned} V_2 = E_g(V_{g2}) &= \sigma^2 + p_1q_1\{\alpha_1 + \sum_{j=2}^{L}\Delta_{1j}(p_j - q_j)\}^2 + 2\sum_{j=2}^{L}p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}^2 \\ &\quad +2\sum_{i=2}^{L}\sum_{j>i}^{L}{\Delta_{ij}}^2\{({p_i}^2 + {q_i}^2)({p_j}^2 + {q_j}^2) - (p_i - q_i)^2(p_j - q_j)^2\} \\ &\quad +4p_1q_1\sum_{j=2}^{L}{\Delta_{1j}}^2p_jq_j + 4\sum_{i=2}^{L}\sum_{j\neq i}^{L}\Delta_{ij}(p_i - q_i)p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}. \end{aligned}$$

$$V_3 = E_g(V_{g3}) = V_1$$

Then, $2V_2 - (V_1 + V_2) = 2p_1q_1\{\alpha_1 + \sum_{j=2}^{L}\Delta_{1j}(p_j - q_j)\}^2(1-2\theta)^2$

Hence, $(1-2\theta)^2 = \frac{2V_2 - (V_1+V_3)}{2(V_p - V_2)}$

or,

$$\theta = \frac{1}{2}[1 - \sqrt{\frac{2V_2 - (V_1 + V_3)}{2(V_p - V_2)}}] \qquad (4.2)$$

Before proceeding further, we wish to note that these computationally simple estimators are analogous, but not identical, to the estimators obtained by Jayakar (1970). The equations corresponding to equations (4.1) and (4.2) derived by Jayakar (1970) for a single quantitative trait locus, has in the denominator the term $V_p - {\sigma_1}^2$. These equations fail to hold when ${\sigma_1}^2$ is replaced by $\sum_{l=1}^{L} {\sigma_l}^2$, if there are L trait loci, even in the absence of any interactions.

4.2.1 Test procedure and evaluation of power

Having estimated θ, one is obviously interested in testing the null hypothesis $\theta = 0.5$. We suggest a non-parametric test procedure that is analogous to the permutation test. For the observed values of the marginal totals of offspring, one can generate simulated data under the null hypothesis $\theta = 0.5$. Using the simulation procedure described in Section 2.2, the required number of offspring ($NOFF$) are obtained. Based on the simulated data, one obtains an estimate of θ by equation (1) or (2). When the simulation is replicated a large number of times ($NREP$), an empirical probability distribution of θ can be obtained and empirical cut-off point(s), for a predetermined level of significance, determined. An inspection of whether the observed value of θ is outside the interval determined by the empirical cut-off point(s) provides the decision on rejection of the null hypothesis.

For the case of two epistatically interacting loci, we evaluate the efficiency of the proposed estimator of θ by examining the empirical frequency distributions of $\widehat{\theta}$ based on multiple replicates of simulated data. A similar method is also used to obtain the power function of the test procedure.

For obtaining the power at $\theta = \theta_1$, the simulation is carried out with $\theta = \theta_1^{'}$ and at each replication, a check is made whether the estimated value of θ lies outside of the interval defined by the empirical cut-off points determined earlier (using $\theta = \theta_0 = 0.5$, say). If the number of replications is n_1, then the power at θ_1 is a/n_1, where a is the number of replications for which $\theta = \theta_0$ is rejected. For every set of parameter values, these evaluations are performed with $NOFF = 1000$ and $NREP = 10000$. We emphasize that $NOFF$ is the total number of offspring in the pooled data set of a particular mating type (backcross or intercross). If each family comprises 4 offspring,

we are in effect dealing with 250 families.

Empirical frequency distribution of $\widehat{\theta}$

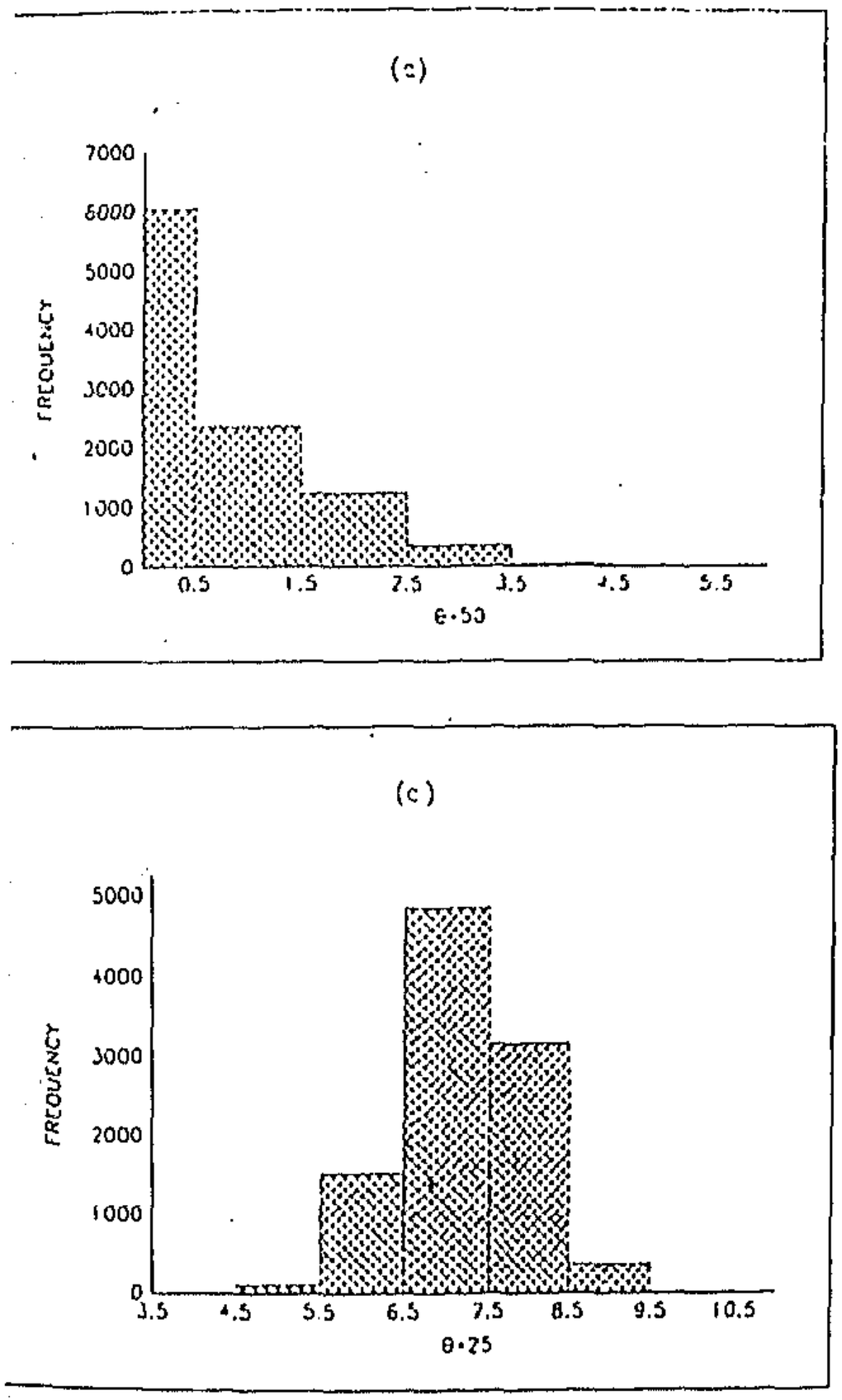
Figures 4.1 and 4.2 depict the frequency distributions of θ for simulation parameter values of $\theta = 0, 0.1, 0.3$ and 0.5, separately for backcross (Figure 4.1) and intercross (Figure 4.2) matings. The values of the other parameters used in these simulations are : $p_1 = 0.5$, $p_2 = 0.5$, $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta_{12} = 1$ and $\sigma^2 = 1$. From these figures it is seen that in all cases, the distributions are unimodal and leptokurtic. For $\theta = 0$, in 50%-60% of the replications $\widehat{\theta}$ is ≤ 0.1. Similarly for $\theta = 0.3$, in 50%-60% of the replications $\widehat{\theta}$ is in the interval $[0.25, 0.35]$. However, for $\theta = 0.5$, while over 45% of the $\widehat{\theta}$ values are between 0.45 and 0.5 for intercross matings, this percentage for backcross matings is only about 7%. For backcross matings, at the simulated value of $\theta = 0.5$, about 90% of the $\widehat{\theta}$ values are in the interval $[0.38, 0.46]$. The nature of the empirical frequency distributions of $\widehat{\theta}$ for other sets of parameter values are similar; figures are, however, not presented for brevity.

Mean and Variance of $\widehat{\theta}$

To examine the behavior of the estimators in respect of variations of values of the underlying parameters $(p_1, p_2, \alpha_1, \alpha_2, \Delta_{12}, \sigma^2)$, we perform simulations for different sets of values of the parameters and evaluate the means and variances of $\widehat{\theta}$. These results are given in Table 4.1. It is seen from this table that the mean of $\widehat{\theta}$ for both backcross and intercross matings deviate more from the true value of θ and the variances of $\widehat{\theta}$ increase with increase in the value of the interaction parameter, Δ_{12}. Similar deviations in Mean$(\widehat{\theta})$ and similar increases in Var$(\widehat{\theta})$ are observed when (a) the variance, σ^2, of the quantitative trait increases, (b) the expected value of the quantitative trait given the genotype of the trait locus A_2, α_2, increases, and (c) when p_1 deviates from 0.5. Variation in p_2 has virtually no effect on Mean(θ) or Var$(\widehat{\theta})$. Although for brevity, results for only a selected number of sets of parameter values are provided in Table 4.1, we verify the above facts for a large number of sets of parameter values.

Table 4.1. Means and variances of estimated values of recombination fraction, θ, each based on 10,000 replications of data simulated at given sets of values of underlying parameters for backcross and intercross families

θ	p_1	p_2	α_1	α_2	Δ_{12}	σ^2	Backcross		Intercross	
							Mean($\hat{\theta}$)	Var($\hat{\theta}$)	Mean($\hat{\theta}$)	Var($\hat{\theta}$)
.00	.50	.50	5.00	1.00	1.00	1.00	.0121	.00027	.0180	.00049
.00	.50	.50	5.00	1.00	2.00	1.00	.0122	.00028	.0186	.00058
.00	.50	.50	5.00	1.00	3.00	1.00	.0128	.00032	.0214	.00080
.00	.50	.50	5.00	1.00	4.00	1.00	.0132	.00037	.0251	.00114
.00	.50	.50	5.00	1.00	5.00	1.00	.0136	.00040	.0301	.00169
.00	.50	.50	5.00	1.00	1.00	1.00	.0121	.00027	.0180	.00049
.00	.50	.50	5.00	1.00	1.00	5.00	.0139	.00040	.0236	.00097
.00	.50	.50	5.00	1.00	1.00	10.00	.0158	.00055	.0326	.00196
.00	.50	.50	5.00	1.00	1.00	1.00	.0121	.00027	.0180	.00049
.00	.50	.50	5.00	5.00	1.00	1.00	.0165	.00062	.0332	.00213
.00	.50	.50	5.00	10.00	1.00	1.00	.0214	.00131	.0811	.01463
.00	.50	.10	5.00	1.00	1.00	1.00	.0121	.00027	.0175	.00050
.00	.50	.20	5.00	1.00	1.00	1.00	.0123	.00029	.0181	.00051
.00	.50	.30	5.00	1.00	1.00	1.00	.0122	.00028	.0183	.00053
.00	.50	.40	5.00	1.00	1.00	1.00	.0117	.00026	.0179	.00050
.00	.50	.50	5.00	1.00	1.00	1.00	.0121	.00027	.0180	.00049
.00	.10	.50	5.00	1.00	1.00	1.00	.0178	.00053	.0300	.00119
.00	.20	.50	5.00	1.00	1.00	1.00	.0135	.00033	.0215	.00067
.00	.30	.50	5.00	1.00	1.00	1.00	.0124	.00029	.0184	.00051
.00	.40	.50	5.00	1.00	1.00	1.00	.0120	.00026	.0182	.00049
.00	.50	.50	5.00	1.00	1.00	1.00	.0121	.00027	.0180	.00049
.10	.50	.50	5.00	1.00	1.00	1.00	.0983	.00081	.1028	.00166
.30	.50	.50	5.00	1.00	1.00	1.00	.2888	.00084	.2982	.00582
.50	.50	.50	5.00	1.00	1.00	1.00	.4228	.00072	.4135	.00750



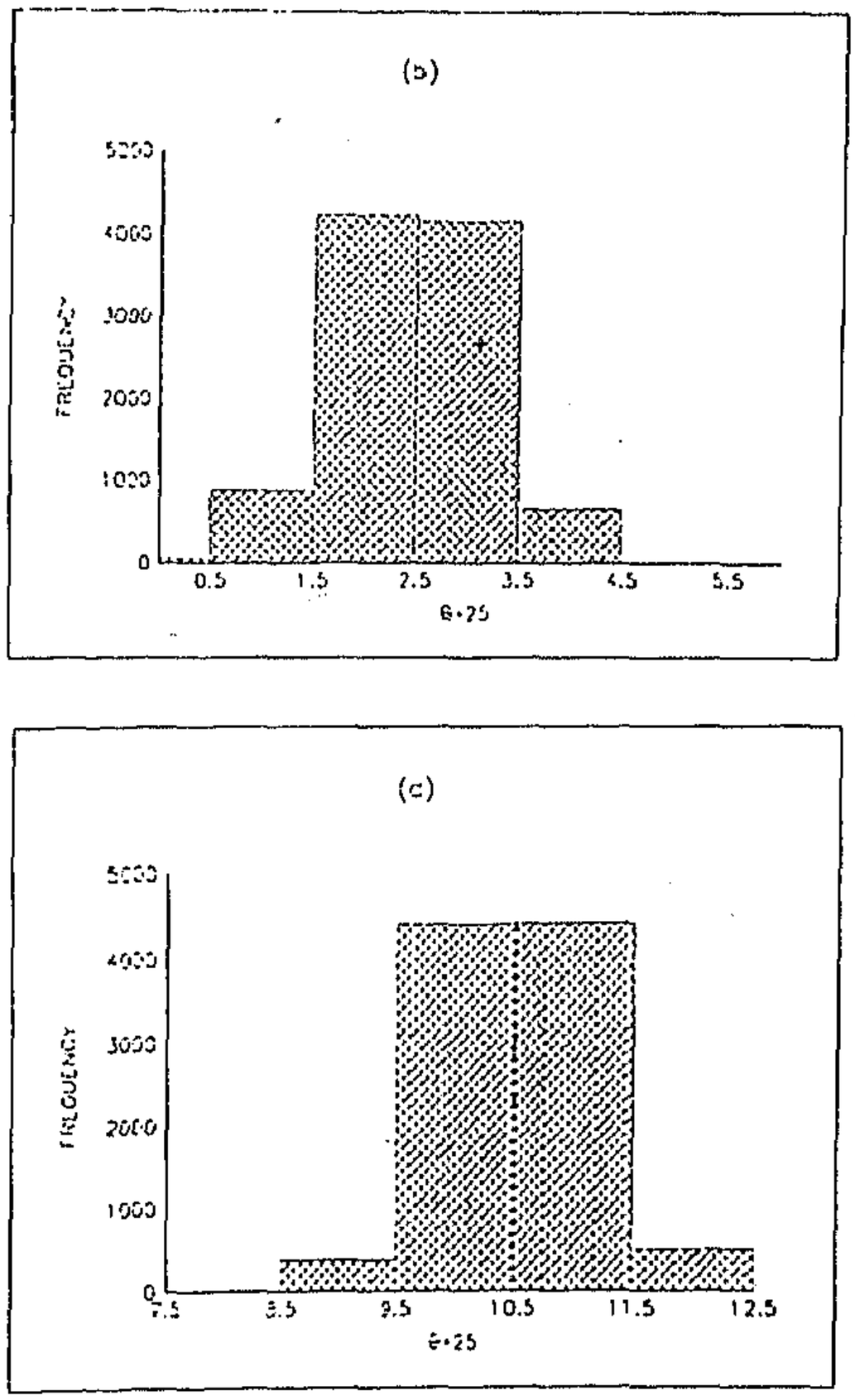


Figure 4.1. Empirical frequency distributions of $\widehat{\theta}$ for backcross families (with $NOFF = 1000$) at simulation parameter values $p_1 = p_2 = .5, \alpha_1 = 5, \alpha_2 = 1, \Delta_{12} = 1, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.

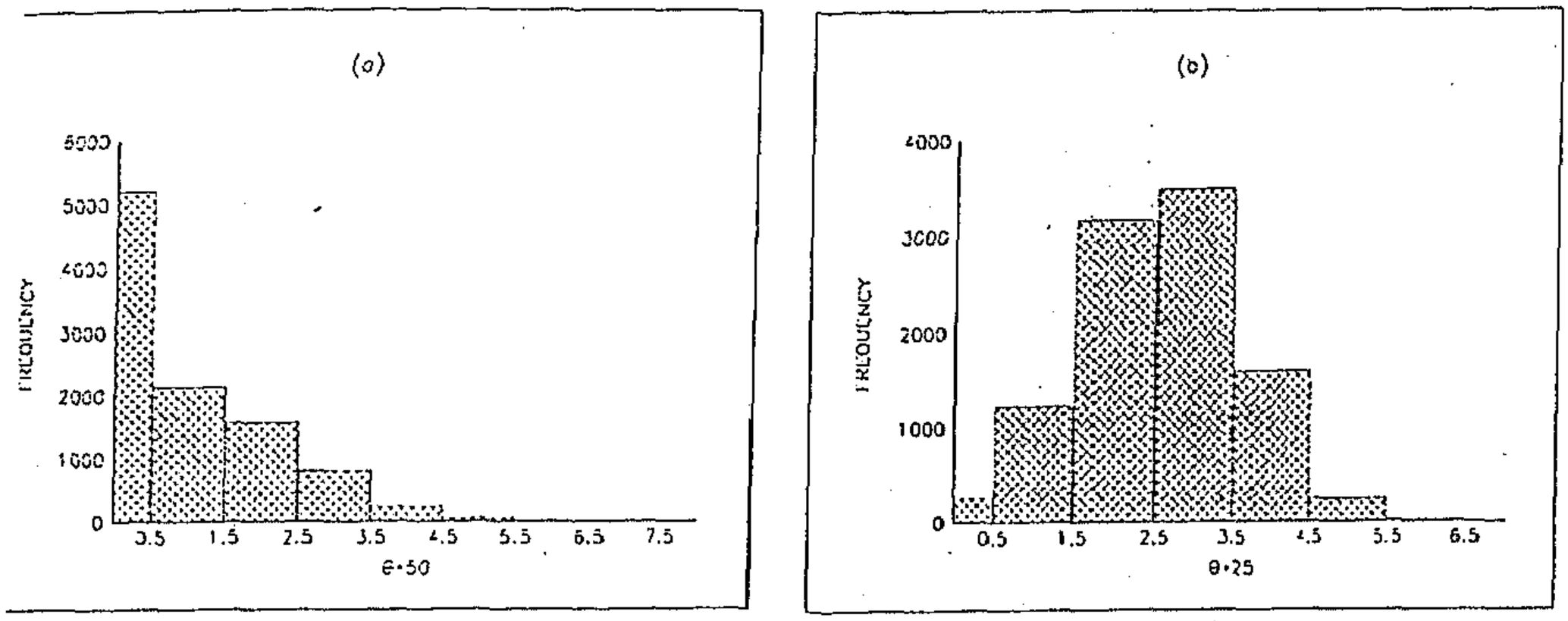


Figure 4.2. Empirical frequency distributions of $\widehat{\theta}$ for intercross families (with $NOFF = 1000$) at simulation parameter values $p_1 = p_2 = .5, \alpha_1 = 5, \alpha_2 = 1, \Delta_{12} = 1, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.

Power function

The power functions for different values of θ, separately for backcross and intercross cases are depicted in *Figures* 4.3 and 4.4, respectively, for a single set of parameter values $p_1 = 0.5$, $p_2 = 0.5$, $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta_{12} = 1$ and $\sigma^2 = 1$. It is seen from this figure that while the power functions are very well-behaved for values of θ in the range $0 < \theta < 0.5$. For $\theta = 0$ or 0.5, the powers are rather high even for values of θ quite close to that specified under the null hypothesis. However, from a practical viewpoint, this undesirable fact for the two extreme values of θ may not imply a serious limitation of the test procedure. As we have already noted in an earlier section, for both backcross and intercross matings, at these two extreme values of θ, in the vast majority of replications the estimated $\widehat{\theta}$ is quite close to the true θ. We have evaluated the power functions for many other sets of parameter values; the results are not provided for brevity since the general feature described above is true for the other sets of values.

Effect of ignoring epistatic interactions

We investigate the effect of ignoring epistatic interactions when in fact these are present, using the following strategy. For a set of parameter values $\theta, p_1, p_2, \alpha_1, \alpha_2, \sigma^2$ and $\Delta_{12} = 0$, we first obtain, based on 10,000 simulation replications, the 95% confidence interval of θ using the procedure outlined earlier. Then, for the same fixed values of $\theta, p_1, p_2, \alpha_1, \alpha_2$ and σ^2, but with $\Delta_{12} = \Delta_0 \neq 0$, we generate 1000 simulated data sets. For each such simulated data set, we estimate θ using equation (1) or (2), as appropriate, and check whether $\widehat{\theta}$ is included in the confidence interval obtained earlier (with $\Delta_{12} = 0$). Inclusion of $\widehat{\theta}$ in the confidence interval implies that the estimate of θ is not significantly adversely affected in spite of ignoring the effect of epistatic interaction when in fact it is present.

For several sets of parameter values, we find, using this procedure, that for most sets of parameter values, the percentage of inclusion of $\widehat{\theta}$ in the appropriate confidence interval varies from about 40% to about 60%. For example, for backcross families with $p_1 = 0.5$, $p_2 = 0.5$, $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta_{12} = 1$ and $\sigma^2 = 1$, this value is 47.3%. Thus, there is a strong adverse

effect of ignoring epistatic interactions for estimating θ when in fact such interactions are present.

Sample size effect

For the proposed analysis, data on all offspring of a particular mating type, backcross or intercross, may be pooled. In our simulation results presented above, we assume that the data set of a particular mating type comprises data on 1000 (= *NOFF*) offspring. We investigate the performance of the proposed estimators when data on fewer offspring are available. Based on simulated data with θ=0.3 and *NOFF*=300, we obtain the power functions for different values of θ separately for backcross and intercross families. The power curves are depicted in Figure 4.5, which as expected, are not as well-behaved as with *NOFF* = 1000. However for the backcross, the power function is reasonably well-behaved, although for the intercross, the power is not very high even for values distant from 0.3. Thus, while for the backcross, data on even 300 offspring may suffice, for the intercross larger data sets are desirable especially when the true θ is large.

4.2.2 Estimation of θ when the marker locus is multiallelic

The above procedure of estimation of θ can be easily shown to hold in the case of a multiallelic locus. Suppose the marker locus has K alleles denoted by $M_1, M_2, \ldots, M_K$. A backcross mating will be of the form $M_iM_i \times M_jM_k$, while an intercross mating will be of the form $M_iM_j \times M_kM_l$.

A mating backcross will produce offspring with marker genotypes M_iM_j and M_iM_k with probability 1/2 each. The probabilities of the trait genotypes of the offspring for various parental mating types are identical to those corresponding to marker genotypes M_1M_1 or M_1m_1 in Table 2.2. Thus Equation (4.1) holds for any type of backcross mating and pooled data on offspring from all backcross families can be used to estimate θ.

In the intercross case, we need to differentiate between matings $M_iM_j \times M_iM_j$ and $M_iM_j \times M_kM_l$ where either $i \neq k$ or $j \neq l$. For $M_iM_j \times M_iM_j$ matings, the distributions of the trait genotypes of the offspring for various parental mating types are identical to those corresponding to the marker

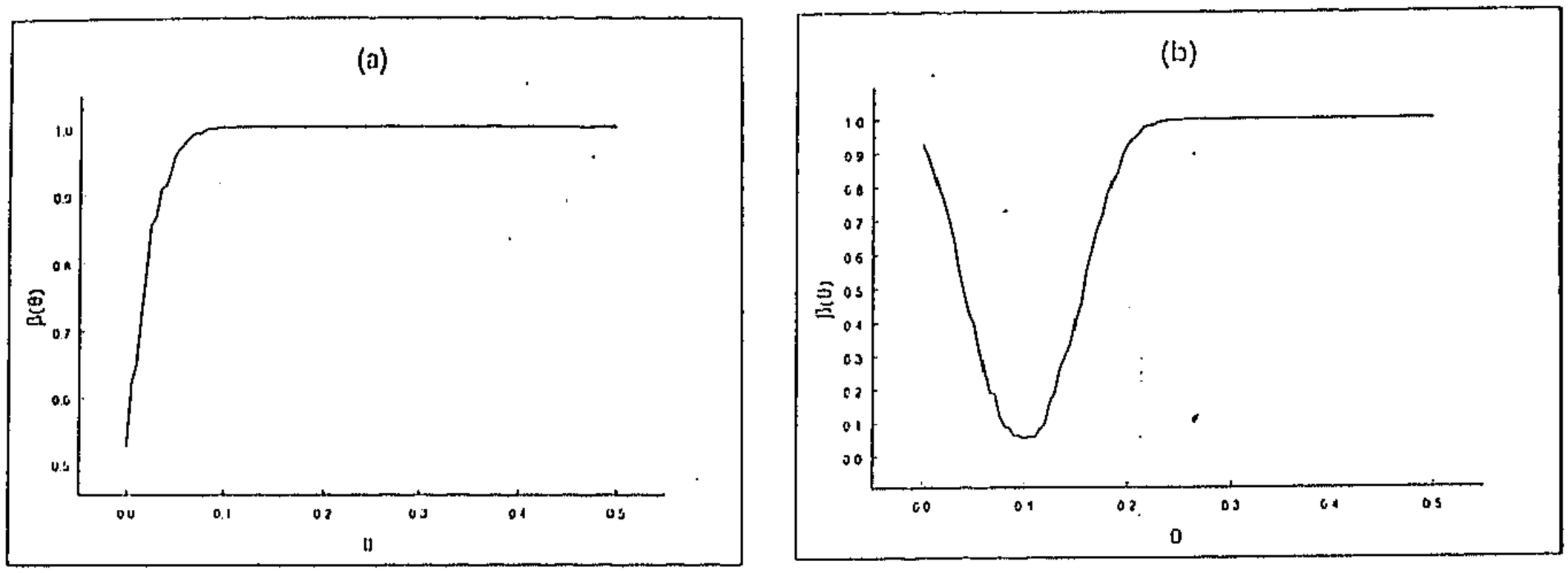


Figure 4.3. Power functions of the proposed test procedure for backcross families (with $NOFF = 1000$) at simulation parameter values $p_1 = p_2 = .5, \alpha_1 = 5, \alpha_2 = 1, \Delta_{12} = 1, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.

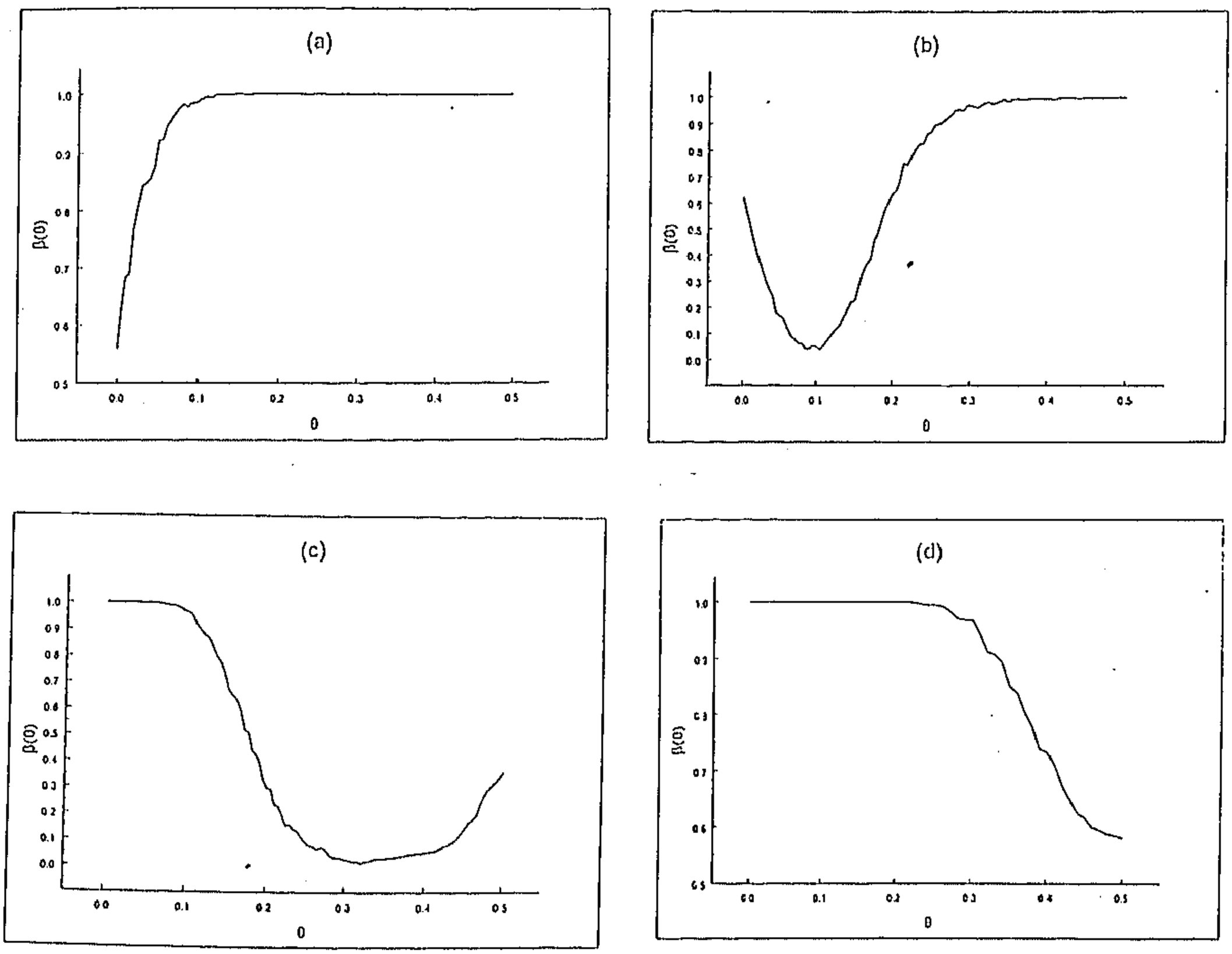


Figure 4.4. Power functions of the proposed test procedure for intercross families (with $NOFF = 1000$) at simulation parameter values $p_1 = p_2 = .5, \alpha_1 = 5, \alpha_2 = 1, \Delta_{12} = 1, \sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$ and (d) $\theta = .5$.

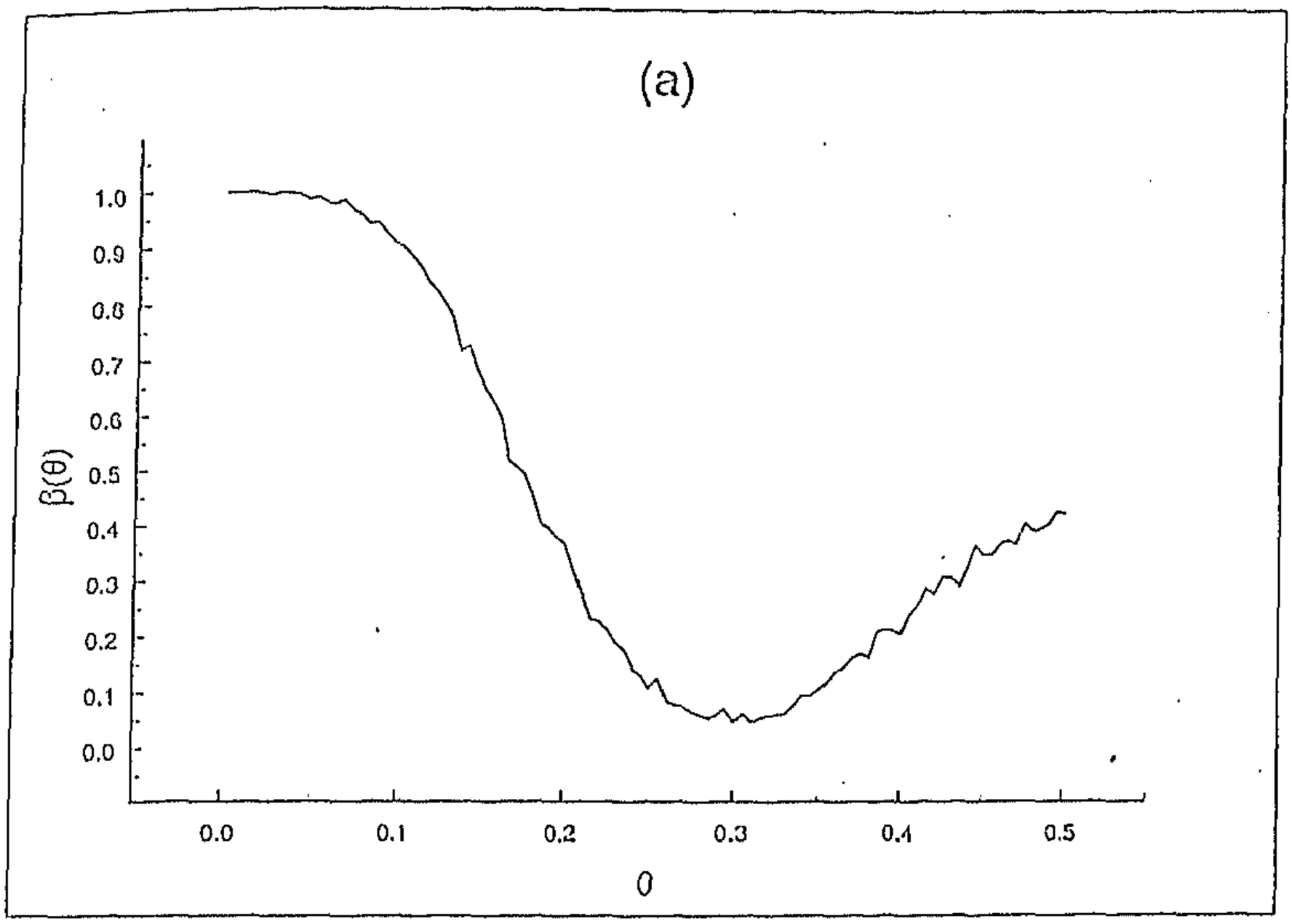
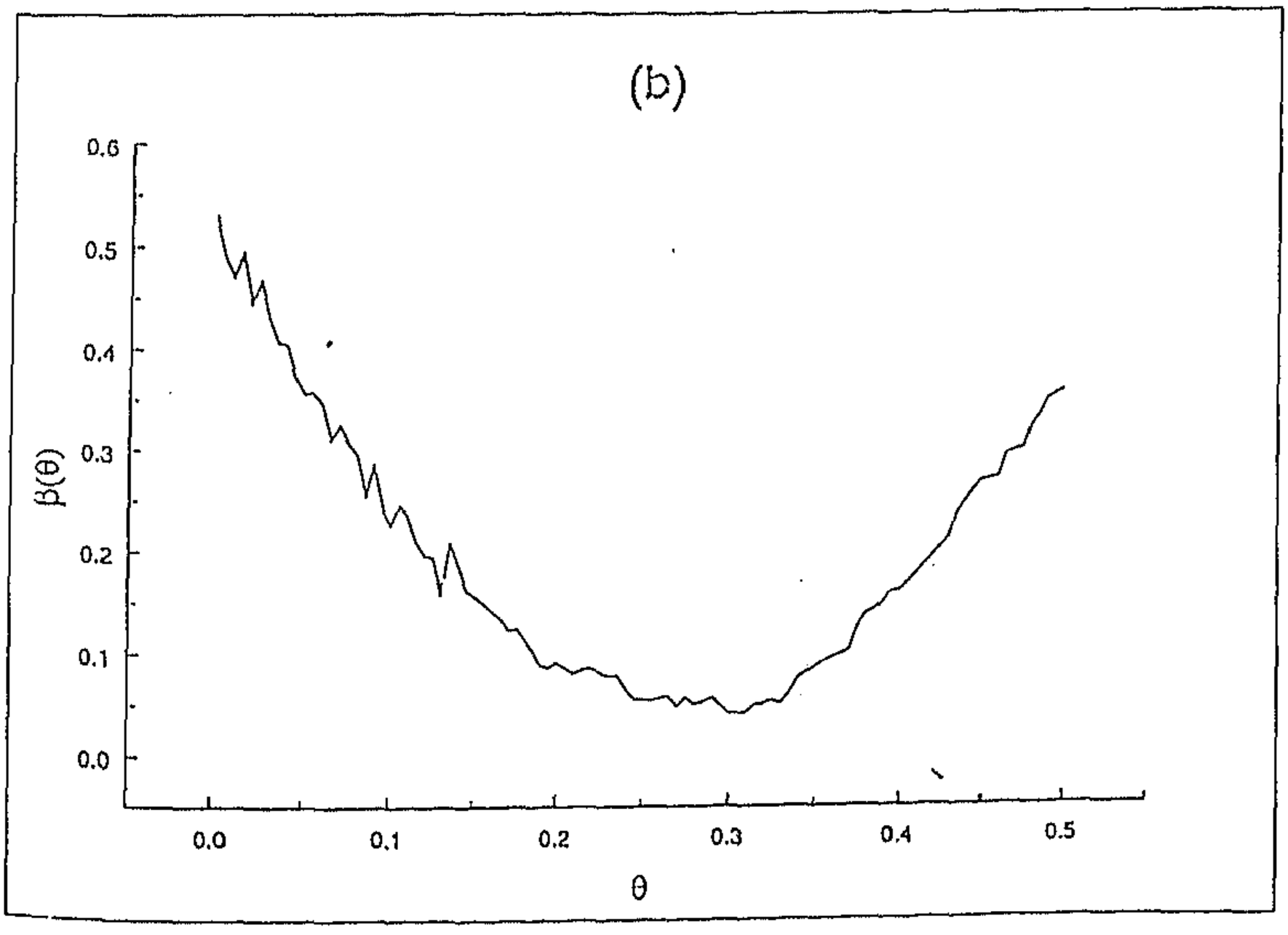


Figure 4.5. Power functions of the proposed test procedure (with $NOFF = 1000$) at simulation parameter values $p_1 = p_2 = .5, \alpha_1 = 5, \alpha_2 = 1, \Delta_{12} = 1, \sigma^2 = 1, \theta = .3$ for (a) *backcross families* and (b) **intercross families**.



genotypes M_1M_1, M_1m_1 or m_1m_1 in Table 2.3. Thus for all matings of this type, Equation (4.2) can be used to estimate θ. $M_iM_j \times M_kM_l$ $(i \neq k \, or \, j \neq l)$ matings can produce offspring with marker genotypes M_iM_k, M_jM_k, M_iM_l and M_jM_L with probability 1/4 each. The probabilities of the trait genotypes of the offspring for various parental mating types are given in Table 3.4. Equation (4.2) does not hold for such matings. In order to derive an estimator of θ, as before define T to be the expected variance of the trait values among all offspring and V_1, V_2, V_3, V_4, respectively to be the expected variances of the trait values among offspring of the four marker genotypes M_iM_k, M_jM_k, M_iM_l, M_jM_l.

Then:

$$\begin{aligned} T &= \sigma^2 + p_1q_1\{\alpha_1 + \sum_{j=2}^{L}\Delta_{1j}(p_j - q_j)\}^2 + 2\sum_{j=2}^{L} p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}^2 \\ &\quad + 2\sum_{i=2}^{L}\sum_{j>i}^{L}\Delta_{ij}{}^2\{(p_i{}^2 + q_i{}^2)(p_j{}^2 + q_j{}^2) - (p_i - q_i)^2(p_j - q_j)^2\} \\ &\quad + 4p_1q_1\sum_{j=2}^{L}\Delta_{1j}{}^2p_jq_j + 4\sum_{i=2}^{L}\sum_{j\neq i}^{L}\Delta_{ij}(p_i - q_i)p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}, \end{aligned}$$

$$\begin{aligned} V_1 &= \sigma^2 + 4p_1q_1\theta(1-\theta)\{\alpha_1 + \sum_{j=2}^{L}\Delta_{1j}(p_j - q_j)\}^2 \\ &\quad + 2\sum_{j=2}^{L} p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}^2 + 4p_1q_1\sum_{j=2}^{L}\Delta_{1j}{}^2p_jq_j \\ &\quad + 2\sum_{i=2}^{L}\sum_{j>i}^{L}\Delta_{ij}{}^2\{(p_i{}^2 + q_i{}^2)(p_j{}^2 + q_j{}^2) - (p_i - q_i)^2(p_j - q_j)^2\} \\ &\quad + 4\sum_{i=2}^{L}\sum_{j\neq i}^{L}\Delta_{ij}(p_i - q_i)p_jq_j\{\alpha_j + \Delta_{1j}(p_1 - q_1)\}, \\ V_2 &= V_3 = V_4 = V_1. \end{aligned}$$

Let $V = V_1 + V_2 + V_3 + V_4$.

Then, $4T - V = 2p_1q_1\{\alpha_1 + \sum_{j=2}^{L}\Delta_{1j}(p_j - q_j)\}^2(1-2\theta)^2$.

Hence, $(1-2\theta)^2 = \frac{4T-V}{4(V_p-T)}$

or, $$\theta = \frac{1}{2}\left[1 - \sqrt{\frac{4T - V}{4(V_p - T)}}\right]$$

This estimator is structurally similar to that derived for backcross matings [Equation (4.1)]. Thus, the methodology and estimators proposed for a biallelic marker extend straightforwardly to a multiallelic marker.

4.2.3 Modification of the estimators in the presence of dominance

Suppose dominance is present in the trait loci, that is, $E(Y|A_lA_l) = \beta_l \neq 0$ for $l = 1, 2, \ldots, L$. Then the proposed estimators in Section 4.3 (backcross and intercross) are not valid and need to be suitably modified. For brevity and to avoid complicated expressions, we shall derive the estimators for $L = 2$. We shall show that under suitable assumptions, the proposed methods yield approximate estimators of θ.

Recalling the notations T, V and V_p in the backcross case, we have:

$$\begin{aligned} T &= \sigma^2 + p_1q_1[\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2 + (1 - p_1q_1){\beta_1}^2 - 2(p_1 - q_1)\alpha_1\beta_1] \\ &\quad + 2p_2q_2\{\alpha_2 + \Delta_{12}(p_1 - q_1)\}^2 - 8p_1q_1\beta_1\beta_2 + 4p_1q_1p_2q_2{\Delta_{12}}^2 \\ &\quad + 2p_2q_2{\beta_2}^2 - 4{p_2}^2{q_2}^2{\beta_2}^2 - 4(p_2 - q_2)p_2q_2\alpha_2\beta_2; \end{aligned}$$

$$\begin{aligned} V &= 2\sigma^2 + p_1q_1[\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2\{1 + 4\theta(1 - \theta)\} + {\beta_1}^2\{1 + 4\theta(1 - \theta) \\ &\quad - 8p_1q_1\theta(1 - \theta)\} - 2(p_1 - q_1)\{1 + 4\theta(1 - \theta)\}\alpha_1\beta_1] + 4p_2q_2\{\alpha_2 + \Delta_{12}(p_1 - q_1)\}^2 \\ &\quad - 16p_1q_1\beta_1\beta_2 + 8p_1q_1p_2q_2{\Delta_{12}}^2 + 4p_2q_2{\beta_2}^2 - 8{p_2}^2{q_2}^2{\beta_2}^2 - 8(p_2 - q_2)p_2q_2\alpha_2\beta_2; \end{aligned}$$

$$\begin{aligned} V_p &= \sigma^2 + 2p_1q_1[\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2 + (1 - 2p_1q_1){\beta_1}^2 - 2(p_1 - q_1)\alpha_1\beta_1] \\ &\quad + 2p_2q_2\{\alpha_2 + \Delta_{12}(p_1 - q_1)\}^2 - 8p_1q_1\beta_1\beta_2 + 4p_1q_1p_2q_2{\Delta_{12}}^2 \\ &\quad + 2p_2q_2{\beta_2}^2 - 4{p_2}^2{q_2}^2{\beta_2}^2 - 4(p_2 - q_2)p_2q_2\alpha_2\beta_2. \end{aligned}$$

From the above equations, we get:

$$V_p - T = p_1q_1\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2 + {\beta_1}^2(1 - 3p_1q_1) - 2(p_1 - q_1)\alpha_1\beta_1.$$

$$2T - V = p_1q_1[\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2 + \beta_1{}^2(1 - 2p_1q_1) - 2(p_1 - q_1)\alpha_1\beta_1](1 - 2\theta)^2.$$

Hence, $\frac{2T-V}{2(V_p-T)} \approx (1 - 2\theta)^2$, assuming $p_1{}^2q_1{}^2\beta_1{}^2$ to be negligible compared to $V_p - T$.

Similarly, in the intercross case, we have:

$$\begin{aligned} V_1 = V_3 = {} & \sigma^2 + 4p_1q_1[\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2\theta(1 - \theta) + \beta_1{}^2\{(1 - 2p_1q_1)\theta(1 - \theta) \\ & + 2\{1 - 2\theta(1 - \theta)\}\theta(1 - \theta)p_1q_1\} - 2(p_1 - q_1)\alpha_1\beta_1\theta(1 - \theta)] \\ & + 2p_2q_2\{\alpha_2 + \Delta_{12}(p_1 - q_1)\}^2 + 4p_1q_1p_2q_2\Delta_{12}{}^2 - 8p_1q_1p_2q_2\beta_1\beta_2 \\ & + 2p_2q_2\beta_2{}^2 - 4p_2{}^2q_2{}^2\beta_2{}^2 - 4p_2q_2(p_2 - q_2)\alpha_2\beta_2; \end{aligned}$$

$$\begin{aligned} V_2 = {} & \sigma^2 + p_1q_1[\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2 + \beta_1{}^2\{(1 - 2p_1q_1) \\ & + 8\{1 - 2\theta(1 - \theta)\}\theta(1 - \theta)p_1q_1\} - 2(p_1 - q_1)\alpha_1\beta_1\theta(1 - \theta)\}] \\ & + 2p_2q_2\{\alpha_2 + \Delta_{12}(p_1 - q_1)\}^2 + 4p_1q_1p_2q_2\Delta_{12}{}^2 - 8p_1q_1p_2q_2\beta_1\beta_2 \\ & + 2p_2q_2\beta_2{}^2 - 4p_2{}^2q_2{}^2\beta_2{}^2 - 4p_2q_2(p_2 - q_2)\alpha_2\beta_2; \end{aligned}$$

$$\begin{aligned} T = {} & \sigma^2 + p_1q_1[\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2 + \beta_1{}^2\{(1 - p_1q_1) \\ & - 2(p_1 - q_1)\alpha_1\beta_1\}] + 2p_2q_2\{\alpha_2 + \Delta_{12}(p_1 - q_1)\}^2 \\ & + 4p_1q_1p_2q_2\Delta_{12}{}^2 - 8p_1q_1p_2q_2\beta_1\beta_2 + 2p_2q_2\beta_2{}^2 \\ & - 4p_2{}^2q_2{}^2\beta_2{}^2 - 4p_2q_2(p_2 - q_2)\alpha_2\beta_2. \end{aligned}$$

Then,

$$2V_2 - (V_1 + V_3) = 2p_1q_1[\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2 + \beta_1{}^2(1 - 2p_1q_1) - 2(p_1 - q_1)\alpha_1\beta_1](1 - 2\theta)^2.$$

$$V_p - T = p_1q_1\{\alpha_1 + \Delta_{12}(p_2 - q_2)\}^2 + \beta_1{}^2(1 - 3p_1q_1) - 2(p_1 - q_1)\alpha_1\beta_1.$$

Hence, $\frac{2V_2-(V_1+V_3)}{2(V_p-T)} \approx (1 - 2\theta)^2$, assuming $p_1{}^2q_1{}^2\beta_1{}^2$ to be negligible compared to $V_p - T$.

Thus, in order to use the above estimators, we need to verify the condition that ${p_1}^2{q_1}^2{\beta_1}^2$ is negligible compared to $V_p - T$. Since $p_1 q_1 \leq 1/4$, it is sufficient to verify whether β_1 is small compared to $4\sqrt{V_p - T}$. This can be done by plugging in observed values of V_p and T and estimating β_1 by the mean of the trait values of those offspring with trait genotype $A_1 a_1$.

4.2.4 Comparison with the maximum likelihood estimator

Since the character Y is controlled by L trait loci, the number of possible trait genotypes of an individual is 3^L. Let the probability density function of Y given these trait genotypes be $f_1, f_2, \ldots, f_{3^L}$ respectively. Then the likelihood of the offspring data given the parental data is:

$$L(\theta) = \prod_{j=1}^{n} f(Y_j)\pi_j,$$

where n is the number of observations, f takes values $f_1, f_2, \ldots$ or f_{3^L} and π_j is the probability of the quantitative trait.

We note that, since the only trait locus linked to the marker is (A_1, a_1), π_j can be interpreted as the probability of the quantitative trait with respect to this locus only. As mentioned in Section 4.2, $f(Y_j)$ is independent of θ and thus the m.l.e. of θ turns out to be a simple function of the number of observations in those genotypic classes for which π_{gi} is not independent of θ.

We compare the efficiency of the proposed variance method in a two trait loci set-up with that of the maximum likelihood approach through simulation studies. We perform simulations for parameter values of $\alpha_1 = 5$; $\alpha_2 = 1$; $\sigma^2 = 1$; $\Delta_{12} = 1$; $p_1 = 0.5, 0.3, 0.1$; $p_2 = 0.5$ and $\theta = 0, 0.1, 0.3, 0.5$ separately for backcross and intercross matings. The results are given in Table 4.2. It is seen that the m.l.e. of θ has more precision than the modified-Jayakar estimator in terms of variance of $\widehat{\theta}$. At the boundary values of θ, that is, $\theta = 0$ and $\theta = 0.5$, the mean of the m.l.e. is also closer to the true value of θ than the modified-Jayakar estimator. We note that the m.l.e. of θ is independent of the trait values of individuals, while Jayakar's estimator is not. However, in spite of using the additional information on trait values, the relative efficiency of the modified-Jayakar estimator is much lower than the m.l.e.

Table 4.2. Comparison between means and variances of estimated values of recombination fraction, θ, each based on 10,000 replications of data simulated at parameter values $\alpha_1 = 5, \sigma^2 = 1$ for backcross and intercross families in the two trait loci set-up using Jayakar's approach and maximum likelihood approach [1]

θ	p_1	Backcross				Intercross			
		$M(\widehat{\theta}_J)$	$V(\widehat{\theta}_J)$	$M(\widehat{\theta}_M)$	$V(\widehat{\theta}_M)$	$M(\widehat{\theta}_J)$	$V(\widehat{\theta}_J)$	$M(\widehat{\theta}_M)$	$V(\widehat{\theta}_M)$
0	.50	.0121	.00027	.00003	.00001	.0180	.00049	.00003	.00001
	.30	.0122	.00028	.0001	.00002	.0186	.00058	.0001	.00002
	.10	.0128	.00032	.0001	.00004	.0214	.00080	.0001	.00004
.10	.50	.0983	.00081	.0997	.00005	.1028	.00166	.1248	.00004
	.30	.1029	.00105	.0994	.00012	.1041	.00210	.1165	.00007
	.10	.1045	.00117	.1013	.00016	.1056	.00229	.1187	.00009
.30	.50	.2888	.00084	.3001	.00012	.2984	.00582	.3312	.00008
	.30	.3101	.00118	.3013	.00023	.2970	.00639	.3282	.00011
	.10	.3138	.00142	.3067	.00049	.3026	.00725	.3304	.00015
.50	.50	.4228	.00072	.04953	.00005	.4135	.00750	.4964	.00003
	.30	.4183	.00154	.04929	.00006	.4102	.00848	.4951	.00005
	.10	.4117	.00202	.04908	.00008	.4036	.00971	.4926	.00006

4.3 Extension of Haseman-Elston (1972) Procedure for Sib-pair Data

In this Section, we extend Haseman and Elston's (1972) regression-based QTL mapping method using sib-pair data. We start with a simple model of the QT being determined by two unlinked, autosomal, epistatically interacting biallelic loci, and then extend it further to multiple QT loci. We note that for two loci, our model is a special case of the more general model considered by Tiwari and Elston (1997). However, as mentioned earlier, our model of epistasis is prompted by experimental observations, and the small number of parameters in our model enables clearer evaluation of the marginal effects of different trait and linkage parameters on the sample size requirement to detect linkage.

[1] $\widehat{\theta}_J$ refers to Jayakar's estimator and $\widehat{\theta}_M$ refers to maximum likelihood estimator.

The digenic-interaction model considered here is given in Section 2.1. We assume that the trait locus (A_l, a_l) is linked to an autosomal, bial-lelic codominant marker locus with alleles M_l and m_l; $l = 1, 2$. The loci (A_l, a_l) and (M_l, m_l) are assumed to be in linkage equilibrium. Our aim is to make inferences on θ_l, the recombination fraction between (A_l, a_l) and (M_l, m_l); $l = 1, 2$, based on data on the quantitative trait values of sib-pairs.

Suppose $\{(y_{j1}, y_{j2}) : j = 1, 2, \ldots, n\}$ are the observed values of the quantitative trait of n independent sib-pairs. We assume that $(y_{j1}, y_{j2})s$ are distributed with an identical covariance structure given by

$$\sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Let π_{j1} and π_{j2} be the proportions of alleles shared i.b.d. at the loci (A_1, a_1) and (A_2, a_2), respectively, for the j^{th} sib-pair. These proportions can assume values $0, \frac{1}{2}, 1$. The conditional probabilities of genotypes of sib-pairs with respect to the locus (A_l, a_l) given π_{jl} are provided in Ta-ble 2.4. As the loci (A_1, a_1) and (A_2, a_2) are unlinked, the joint condi-tional probability of the trait locus genotypes given the trait i.b.d. scores is the product of the marginal conditional probability of each trait lo-cus genotype given the corresponding trait i.b.d. score. For example, $P(\text{Sib 1} = A_1A_1A_2A_2 \text{ and Sib 2} = A_1A_1A_2A_2 | \pi_{j1} = \frac{1}{2}, \pi_{j2} = 1) = {p_1}^3{p_2}^2$.

Define $Y_j = (Y_{j1} - Y_{j2})^2$, $j = 1, 2, \ldots, n$; i.e., Y_j denotes the squared pair difference in the trait values for the j^{th} sib-pair.

Note that $V(Y_{j1} - Y_{j2}) = 2\sigma^2(1 - \rho) = \phi^2, \forall j = 1, 2, \ldots, n$. Now,

$$\begin{aligned} & E(Y_j|\pi_{j1}, \pi_{j2}) \\ &= V(Y_{j1} - Y_{j2}|\pi_{j1}, \pi_{j2}) + \{E(Y_{j1} - Y_{j2}|\pi_{j1}, \pi_{j2})\}^2 \\ &= \phi^2 + \{E(Y_{j1} - Y_{j2}|\pi_{j1}, \pi_{j2})\}^2 \end{aligned}$$

The conditional expectations of Y_j for the different values of π_{j1} and π_{j2} can be summarised by the following relation:

$$E(Y_j|\pi_{j1}, \pi_{j2}) = \alpha_0 + \alpha_1\pi_{j1} + \alpha_2\pi_{j2},$$

where:

$$\alpha_0 = \phi^2 + 4p_1q_1\{{\alpha_1}^2 + \Delta^2({p_2}^2 + {q_2}^2) + 2\alpha_1\Delta(p_2 - q_2)\} + 4p_2q_2\{{\alpha_2}^2 + \Delta^2({p_1}^2 + {q_1}^2) + 2\alpha_2\Delta(p_1 - q_1)\};$$

$$\alpha_1 = -4p_1q_1\{{\alpha_1}^2 + \Delta^2({p_2}^2 + {q_2}^2) + 2\alpha_1\Delta(p_2 - q_2)\};$$

$$\alpha_2 = -4p_2q_2\{{\alpha_2}^2 + \Delta^2({p_1}^2 + {q_1}^2) + 2\alpha_2\Delta(p_1 - q_1)\}. \qquad (4.3)$$

For clarity, the derivation of $E(Y_j|\pi_{j1} = \frac{1}{2}, \pi_{j2} = 1)$ is given in Appendix 4.1.

4.3.1 Derivation of the regression equation

Let π_{jm_1} and π_{jm_2} denote the proportions of alleles shared i.b.d. at the marker loci (M_1, m_1) and (M_2, m_2) respectively for the j^{th} sib-pair. Let ${f_{ji}}^{(l)}$ denote the probability that the j^{th} sib-pair has i alleles shared i.b.d. at the marker locus $(M_l, m_l), i = 0, 1, 2; l = 1, 2$. Then the estimator of π_{jm_l} is given by $\widehat{\pi}_{jm_l} = {f_{j2}}^{(l)} + \frac{1}{2}{f_{j1}}^{(l)}; l = 1, 2$. Haseman and Elston (1972) have explicitly calculated ${f_{ji}}^{(l)}$ for different mating types and in the case of missing parental information, they have suggested an algorithm considering phenosets (Cotterman 1969).

Suppose now we are interested in evaluating $E(Y_j|\widehat{\pi}_{jm_1}, \widehat{\pi}_{jm_2})$. In order to compute this expectation, we further condition Y_j on π_{j1}, π_{j2}, π_{jm_1}, π_{jm_2}.

$$\begin{aligned}
&E(Y_j|\widehat{\pi}_{jm_1}, \widehat{\pi}_{jm_2}) \\
&= \textstyle\sum_{\pi_{j1}} \sum_{\pi_{j2}} E(Y_j|\pi_{j1}, \pi_{j2}) P(\pi_{j1}, \pi_{j2}|\widehat{\pi}_{jm_1}, \widehat{\pi}_{jm_2}) \\
&= \textstyle\sum_{\pi_{j1}} \sum_{\pi_{j2}} \sum_{\pi_{jm_1}} \sum_{\pi_{jm_2}} E(Y_j|\pi_{j1}, \pi_{j2}) P(\pi_{j1}, \pi_{j2}|\pi_{jm_1}, \pi_{jm_2}) P(\pi_{jm_1}, \pi_{jm_2}|\widehat{\pi}_{jm_1}, \widehat{\pi}_{jm_2}) \\
&= \sum_{\pi_{j1}} \sum_{\pi_{j2}} \sum_{\pi_{jm_1}} \sum_{\pi_{jm_2}} E(Y_j|\pi_{j1}, \pi_{j2}) P(\pi_{j1}|\pi_{jm_1}) P(\pi_{j2}|\pi_{jm_2}) P(\pi_{jm_1}|\widehat{\pi}_{jm_1}) P(\pi_{jm_2}|\widehat{\pi}_{jm_2})
\end{aligned} \qquad (4.4)$$

In the above expressions, $\sum_{\pi_{jl}}$, etc, denote summations taken over all possible values of π_{j1}, etc. The conditional distributions of $\pi_{jl}|\pi_{jm_l}$ and $\pi_{jm_l}|\widehat{\pi}_{jm_l}$ are provided in Tables 2.5 and 2.6, respectively. Thus using Equations (4.3) and (4.4), we can obtain expressions for $E(Y_j|\widehat{\pi}_{jm_1}, \widehat{\pi}_{jm_2})$. For example, $E(Y_j|\widehat{\pi}_{jm_1} = 1, \widehat{\pi}_{jm_2} = 0) = \phi^2 + 4p_1q_1\{{\alpha_1}^2 + \Delta^2({p_2}^2 + {q_2}^2) + 2\alpha_1\Delta(p_2 - q_2)\}(1 - \psi_1) + 4p_2q_2\{{\alpha_2}^2 + \Delta^2({p_1}^2 + {q_1}^2) + 2\alpha_2\Delta(p_1 - q_1)\}\psi_2$ where $\psi_l = {\theta_l}^2 + (1 - \theta_l)^2;\ l = 1, 2$. For brevity, the conditional expectations of Y_j for other values of $\widehat{\pi}_{jm_1}$ and $\widehat{\pi}_{jm_2}$ are not presented.

Combining the different values of $\widehat{\pi}_{jm_1}$ and $\widehat{\pi}_{jm_2}$, we obtain the relation:

$$E(Y_j|\widehat{\pi}_{jm_1}, \widehat{\pi}_{jm_2}) = \beta_0 + \beta_1\widehat{\pi}_{jm_1} + \beta_2\widehat{\pi}_{jm_2},$$

where:

$$\beta_0 = \phi^2 + 4p_1q_1\{{\alpha_1}^2 + \Delta^2({p_2}^2 + {q_2}^2) + 2\alpha_1\Delta(p_2 - q_2)\}(1 - 2\theta_1 + 2{\theta_1}^2)$$
$$+ 4p_2q_2\{{\alpha_2}^2 + \Delta^2({p_1}^2 + {q_1}^2) + 2\alpha_2\Delta(p_1 - q_1)\}(1 - 2\theta_2 + 2{\theta_2}^2);$$

$$\beta_1 = -4p_1q_1\{{\alpha_1}^2 + \Delta^2({p_2}^2 + {q_2}^2) + 2\alpha_1\Delta(p_2 - q_2)\}(1 - 2\theta_1)^2;$$

$$\beta_2 = -4p_2q_2\{{\alpha_2}^2 + \Delta^2({p_1}^2 + {q_1}^2) + 2\alpha_2\Delta(p_1 - q_1)\}(1 - 2\theta_1)^2. \qquad (4.5)$$

This provides the motivation to set up the linear model:

$$Y_j = \beta_0 + \beta_1\widehat{\pi}_{jm_1} + \beta_2\widehat{\pi}_{jm_2} + e_j,\ j = 1, 2, \ldots, n$$

where $e_j s$ are i.i.d. $N(0, \tau^2)$.

We now note that for $l = 1, 2; \beta_l = 0 \Leftrightarrow \theta_l = 0.5$ and $\beta_l < 0 \Leftrightarrow \theta_l < 0.5$ as β_l is an increasing $1 - 1$ function of θ_l. Thus, a test for linkage at the l^{th} locus (i.e., $H_0 : \theta_l = 0.5$ vs $H_1 : \theta_l < 0.5$), it is equivalent to testing for $H_0 : \beta_l = 0$ vs $H_1 : \beta_l < 0$ in the above linear model. The test statistic is given by $T_l = \frac{\widehat{\beta}_l}{\widehat{s.e}(\widehat{\beta}_l)}$ where $\widehat{\beta}_l$ is the least squares estimator of β_l. In order to compute the standard error of $\widehat{\beta}_l$, consider the design matrix X given by:

$$\begin{pmatrix} \widehat{\pi}_{1m_1} & \widehat{\pi}_{1m_2} & 1 \\ . & . & . \\ . & . & . \\ . & . & . \\ \widehat{\pi}_{nm_1} & \widehat{\pi}_{nm_2} & 1 \end{pmatrix}$$

Let $S = (X'X)$. Then $\widehat{s.e.}(\widehat{\beta}_l) = \sqrt{\frac{s^{ll}{R_0}^2}{n-3}}$ where $s^{ll} = (S^{-1})_{ll}$ and ${R_0}^2 =$ residual sum of squares $= \sum_{j=1}^{n}\left(Y_j - \widehat{\beta}_0 - \widehat{\beta}_1\widehat{\pi}_{jm_1} - \widehat{\beta}_2\widehat{\pi}_{jm_2}\right)^2$. Under H_0, T_l follows a t-distribution with $(n - 3)$ degrees of freedom. Thus the critical region for a level α test is given by: $T_l < t_{n-3,1-\alpha}$.

If n is sufficiently large, by the Central Limit Theorem (C.L.T.), we can approximate the critical region by: $T_l < z_{1-\alpha}$ where z_p is the $(1 - p)^{th}$ quantile of a standard normal variate.

The power function $P(\beta_l)$ is given by:

$$P(\beta_l) = P_{\beta_l}\{T_l < z_{1-\alpha}\}$$

$$= P_{\beta_l}\{\frac{\widehat{\beta}_l}{\widehat{s.e.}(\widehat{\beta}_l)} < z_{1-\alpha}\}$$

$$= P_{\beta_l}\{\frac{\widehat{\beta}_l - \beta_l}{\widehat{s.e.}(\widehat{\beta}_l)} < z_{1-\alpha} - \frac{\beta_l}{\widehat{s.e}(\widehat{\beta}_l)}\}$$

$$= \Phi\{z_{1-\alpha} - \frac{\beta_l}{\widehat{s.e}(\widehat{\beta}_l)}\}(\text{using C.L.T.})$$

where Φ is the c.d.f. of $N(0,1)$.

4.3.2 Determination of sample size required to detect linkage

Having derived the power function of the proposed test, one is obviously interested in determining the minimum sample size required to detect linkage at the l^{th} locus; $l = 1, 2..$ In order for the test to have a power β at β_l (which is a $1-1$ increasing function of θ_l), we require the condition:

$$\Phi\{z_{1-\alpha} - \frac{\beta_l}{\sqrt{\frac{s^{II}{R_0}^2}{n_l-3}}}\} = \beta$$

$$\Rightarrow \quad z_{1-\alpha} - \frac{\beta_l}{\sqrt{\frac{s^{II}{R_0}^2}{n_l-3}}} = z_{1-\beta}$$

$$\Rightarrow \quad n_l = \frac{(z_{1-\alpha}-z_{1-\beta})^2 s^{II}{R_0}^2}{{\beta_l}^2} + 3 \qquad (4.6)$$

Thus the required sample size to detect linkage at both the loci is given by $n = max(n_1, n_2)$. In order to examine the effects of different trait parameters and linkage parameters on sample size requirement, we provide the following proposition.

Proposition: The sample size (n_l) required to detect linkage at the l^{th} locus is:

(i) an increasing function of θ_l;

(ii) an increasing function of p_l;

(iii) a decreasing function of $p_i, i \neq l$;

(iv) a decreasing function of α_l;

(v) a decreasing function of Δ;

(vi) independent of $\alpha_i, i \neq l$;

(vii) independent of σ^2 and ρ.

Proof: Without loss of generality, we assume that $p_l \geq q_l$, $l = 1,2$. Equation (4.5) implies that β_l is an increasing function of θ_l and p_l and a decreasing function of α_l, Δ and p_i $(i \neq l)$. Now, $\beta_l < 0 \Rightarrow {\beta_l}^2$ is a decreasing function of β_l. Considering equation (4.6), $(i) - (v)$ follow immediately.

Again, equations (4.5) and (4.6) are both independent of α_i $(i \neq l), \sigma^2$ and ρ. Thus, (vi) and (vii) are obviously true.

Hence, as intuitively expected, if the strength of linkage between a trait locus and a marker locus is higher, a smaller sample size suffices to detect linkage. Moreover, if a locus is controlled by several loci with comparable effects, then the sample size required for mapping the QTL with the highest level of heterozygosity is the smallest. Further, if among several QTLs, the marginal effect of one QTL increases, then smaller sample sizes are required to map that locus. Thus, among several QTLs, the QTLs with major effects are easiest to map. Moreover, if two QTLs have equal effects, then smaller sample sizes are required to map them if they epistatically interact than if they do not. A similar result holds if there are multiple loci even with unequal effects.

4.3.3 Simultaneous detection vs. sequential detection as strategies to reduce sample size

One interesting question that may arise in the determination of sample size to detect linkage at both the loci is whether it is more optimal to analyse the data by considering both the markers simultaneously (as illustrated in the previous Subsection) or by considering them sequentially, one by one. In order to resolve this problem, let us first obtain expressions for $E(Y_j|\widehat{\pi}_{jm_l})$; $l = 1,2$. Using the conditioning arguments as before, we can easily show that:

$$E(Y_j|\widehat{\pi}_{jm_l}) = \beta_0 + \beta_l \widehat{\pi}_{jm_l},\ l = 1,2,$$

where:

$$\begin{aligned}\beta_0 &= \phi^2 + 4p_lq_l\{{\alpha_l}^2 + \Delta^2({p_{3-l}}^2 + {q_{3-l}}^2) + 2\alpha_l\Delta(p_{3-l} - q_{3-l})\}(1 - 2\theta_l + 2{\theta_l}^2) \\ &\quad + 4p_{3-l}q_{3-l}\{{\alpha_{3-l}}^2 + \Delta^2({p_l}^2 + {q_l}^2) + 2\alpha_{3-l}\Delta(p_l - q_l)\};\ \text{and}\end{aligned}$$

$$\beta_l = -4p_lq_l\{{\alpha_l}^2 + \Delta^2({p_{3-l}}^2 + {q_{3-l}}^2) + 2\alpha_l\Delta(p_{3-l} - q_{3-l})\}(1 - 2\theta_l)^2.$$

We note that β_l is a $1-1$ increasing function of θ_l and $\beta_l = 0 \Leftrightarrow \theta_l = 0.5$ while $\beta_l < 0 \Leftrightarrow \theta_l < 0.5$. Thus, in order to detect linkage at the l^{th} locus, we use the test statistic $T_l = \frac{\widehat{\beta}_l}{\widehat{s.e.}(\widehat{\beta}_l)}$ where

$$\widehat{s.e.}(\widehat{\beta}_l) = \frac{\sum_{j=1}^{n}(y_j - \widehat{\beta}_0 - \widehat{\beta}_l \widehat{\pi}_{jm_l})^2}{(n-2)\sum_{j=1}^{n}(\widehat{\pi}_{jm_l} - \overline{\widehat{\pi}_{m_l}})^2}.$$

T_l follows a $t-$distribution with $(n-2)$ degrees of freedom under the null hypothesis of no linkage at the l^{th} locus. If n is sufficiently large, the power function, $P(\beta_l)$, of the above test can be approximated by:

$$P(\beta_l) = \Phi\{z_{1-\alpha} - \frac{\beta_l}{\widehat{s.e.}(\widehat{\beta}_l)}\}$$

The minimum sample size to detect linkage at the l^{th} locus (i.e., to attain a power of β at β_l) is given by:

$$N_l = \frac{(z_{1-\alpha} - z_{1-\beta})^2}{\beta_l^2} \frac{\sum_{j=1}^{n}(y_j - \widehat{\beta}_0 - \widehat{\beta}_l \widehat{\pi}_{jm_l})^2}{(n-2)\sum_{j=1}^{n}(\widehat{\pi}_{jm_l} - \overline{\widehat{\pi}_{m_l}})^2} + 2.$$

Thus the minimum sample size to detect linkage at both the loci is given by $N = max(N_1, N_2)$. Though N (as given above) and n (as determined in the previous section) cannot be compared analytically, we show through simulation studies that n is, in general, smaller than N, that is, we require a smaller sample size to detect linkage at both the loci if we analyse the data by considering both the markers simulateneously as opposed to considering them sequentially, one by one.

We note here that since the two QTLs are unidentifiable, in the sequential strategy, evidence of linkage would imply that the chosen marker locus is linked to one of the two trait loci. As the chosen markers are themselves unlinked, evidence of linkage for two different markers would indicate that we are able to map both the QTLs. Moreover the sequential analysis of markers is equivalent to analysing the data under the misspecified model of a single QTL. We thus observe that under this misspecified model, linkage of the chosen marker locus with the QTL can be correctly detected, but it requires a larger sample to map that QTL.

4.3.4 Extension of the regression procedure when the quantitative trait is controlled by more than two loci

The regression procedure described above can be easily extended when the quantitative trait Y is controlled by k autosomal,biallelic, unlinked loci $(A_1, a_1), (A_2, a_2), \ldots, (A_k, a_k)$. The generalized epistatic interaction model considered in the case of multiple loci has been described earlier in Section 2.1.

Suppose the trait locus (A_l, a_l) is in linkage equilibrium with an autosomal, biallelic, codominant marker locus (M_l, m_l), $l = 1, 2, \ldots, k$, and the recombination fraction between these two loci is θ_l. Our aim is to make inferences on θ_l based on observations on the quantitative trait of n sib-pairs given by $\{(y_{j1}, y_{j2}) : j = 1, 2, \ldots, n\}$.

Suppose $\{\pi_{jl} : l = 1, 2, \ldots, k\}$ and $\{\pi_{jm_l} : l = 1, 2, \ldots, k\}$ denote the proportions of alleles shared i.b.d. at the l^{th} trait locus $\{(A_l, a_l) : l = 1, 2, \ldots, k\}$ and the l^{th} marker locus $\{(M_l, m_l) : l = 1, 2, \ldots, k\}$ respectively. π_{jm_l} can be estimated by $\widehat{\pi}_{jm_l} = {f_{j2}}^{(l)} + \frac{1}{2}{f_{j1}}^{(l)}$, where ${f_{ji}}^{(l)}$ is the probabillty that the j^{th} sib-pair has i alleles shared i.b.d.at the marker locus (M_l, m_l), $i = 0, 1, 2$; $l = 1, 2, \ldots, k$.

Defining $Y_j = (Y_{j1} - Y_{j2})^2$ and using the same conditioning argument as in the two loci case, we can show that:

$$E(Y_j|\widehat{\pi}_{jm_1}, \ldots, \widehat{\pi}_{jm_k}) = \beta_0 + \sum_{l=1}^{k} \beta_l \widehat{\pi}_{jm_l},$$

where β_l is a $1-1$ increasing function of θ_l and $\theta_l = 0.5 \Leftrightarrow \beta_l = 0$ and $\theta_l < 0.5 \Leftrightarrow \beta_l < 0$.

Thus in the linear model:

$$Y_j = \beta_0 + \sum_{l=1}^{k} \beta_l \widehat{\pi}_{jm_l} + e_j, \; j = 1, 2, \ldots, n$$

where $e_j s$ are i.i.d. $N(0, \tau^2)$, testing for $H_0 : \beta_l = 0$ vs $H_1 : \beta_l < 0$ is equivalent to testing for $H_0 : \theta_l = 0.5$ vs $H_1 : \theta_l < 0.5$.

The test statistic given by $T_l = \frac{\widehat{\beta_l}}{\widehat{s.e.(\beta_l)}}$ follows t_{n-k-1} under H_0. Note that $\widehat{s.e}(\widehat{\beta}_l)$ is given by $sqrt\frac{s^{ll}{R_0}^2}{n-k-1}$ where ${R_0}^2 = \sum_{j=1}^{n}(Y_j - \widehat{\beta}_0 - \sum_{l=1}^{k}\widehat{\beta}_l\widehat{\pi}_{jm_l})^2$ and s^{ll} is computed from the design matrix X given by:

$$\begin{pmatrix} \widehat{\pi}_{1m_1} & \cdot & \cdot & \cdot & \widehat{\pi}_{1m_k} & 1 \\ \cdot & & & & \cdot & \cdot \\ \cdot & & & & \cdot & \cdot \\ \cdot & & & & \cdot & \cdot \\ \widehat{\pi}_{nm_1} & \cdot & \cdot & \cdot & \widehat{\pi}_{nm_k} & 1 \end{pmatrix}$$

If n is sufficiently large, the power function, $P(\beta_l)$, of the above test becomes:

$$P(\beta_l) = \Phi\{z_{1-\alpha} - \frac{\beta_l}{\widehat{s.e}(\widehat{\beta}_l)}\}$$

The minimum sample size required to detect linkage at the l^{th} locus (i.e., to attain a power β at β_l) is given by:

$$n_l = \frac{(z_{1-\alpha} - z_{1-\beta})^2 s^{ll} {R_0}^2}{{\beta_l}^2} + (k+1).$$

Thus the minimum sample size required to detect linkage in all the k loci is given by $n = max(n_1, n_2, \ldots, n_k)$. As in the two loci set-up, simultaneous analysis of the k markers reduces the sample size requirement to detect linkage at all the k loci as compared to sequential analysis of the markers, one by one.

4.3.5 Simulation results

In order to assess the performance of our proposed regression strategy, we generate data on trait values of sib-pairs and estimated marker i.b.d. scores for different parameter values. The different steps of the simulation algorithm have already been described in Section 2.2.

Having generated the required data on 100 sib-pairs, we regress the squared difference in trait values on the different estimated marker i.b.d. scores. Based on the regression coefficients obtained, we evaluate the sample size requirements for detecting linkage for different values of recombination fractions. We perform the regression analysis both by considering the two markers simultaneously as well as sequentially, one by one. In each case we determine the sample size requirement (i.e., n and N) and compared them by an 'Efficiency' ratio $E = N/n$.

In our simulation examples, we assume the quantitative trait to be controlled by two autosomal loci and thus consider two marker loci which are

in linkage equilibrium with the trait loci. Table 4.3 provides the results of the regression of squared difference in trait values on the two estimated marker i.b.d. scores and Tables 4.4 (a)-(d) provide the sample sizes necessary to detect linkage at the two trait loci for simulation parameter values of $\alpha_1 = 5, \alpha_2 = 1, \Delta = 1, \sigma^2 = 1$ and different parameter values of p_1, p_2, ρ, θ_1 and θ_2. We perform the tests of linkage at 5% level of significance and determine the sample size requirements to attain a power of 0.9 for each test.

In each of the four cases, we find that the regression procedure detects linkage quite efficiently and the sample size requirements are in accordance with our proposition stated in an earlier Subsection. [i.e., n_l increases with p_l and decreases with α_l, Δ and p_i $(i \neq l)$]. The significance of the regression coefficients $\widehat{\beta}_1$ and $\widehat{\beta}_2$ (i.e., the extent of linkage at the two loci) depends not only on θ_1 and θ_2, but also on α_1 and α_2 (i.e., on the effect of each trait locus on the quantitative trait). When this effect is small, the corresponding regression coefficient tends to be less significant even if the trait locus is actually tightly linked to the marker locus [as in case (c)]. Similarly when the effect is large, the corresponding regression coefficient tends to be less insignificant even if the trait locus is actually unlinked to the marker locus [as in case (d)].

We assess the nature of sample size requirements under various scenarios assuming that the QT is controlled by two loci. First, in the absence of epistatic interaction ($\Delta = 0$), if both loci have equal effects ($\alpha_1 = \alpha_2 = \alpha$), then the sample size required to map the first (or, the second) trait locus decreases with heterozygosity at that locus increases (Figure 4.6). The rate of decrease, however, is greater when the locus has a smaller effect on the QT. Second, in the absence of epistatic interaction ($\Delta = 0$), the sample size required to map the locus which has a greater effect on the QT decreases as its relative effect increases [Figure 4.7(a)]. However although the rate of decrease in sample size depends largely on the heterozygosity of the locus with the greater effect, it also varies with the heterozygosity of the second locus. For example, we find that while the sample size requirement to map the first locus in the case $p_1 = p_2 = 0.75$ is more than in the case $p_1 = p_2 = 0.5$ when the marginal effects of the two loci do not differ significantly, it requires a smaller sample to do so in the case $p_1 = p_2 = 0.75$ if the marginal

Table 4.3. Regression and multiple correlation coefficients in the two types of regression analyses (simultaneous and sequential) for different sets of trait parameter values.[2]

(a) Parameter values: $p_1 = 0.7, p_2 = 0.5, \rho = 0.6, \theta_1 = 0$ and $\theta_2 = 0.5$

Type of Analysis	R^2	β_0			β_1			β_2		
		Est	S.E.	t-val	Est	S.E.	t-val	Est	S.E.	t-val
Simult	0.87	19.49	5.32	3.67	-15.30	6.66	-2.3*	-1.85	8.23	-0.22
Seq using:										
Marker 1	0.79	23.90	6.21	3.85	-18.43	7.63	-2.42*			
Marker 2	0.10	30.28	9.58	3.16				-1.00	7.55	-0.13

(b) Parameter values: $p_1 = 0.7, p_2 = 0.9, \rho = 0.3, \theta_1 = 0.4$ and $\theta_2 = 0.1$

Type of Analysis	R^2	β_0			β_1			β_2		
		Est	S.E.	t-val	Est	S.E.	t-val	Est	S.E.	t-val
Simult	0.72	11.07	9.76	1.13	-10.95	6.54	-1.68*	-13.50	7.12	-1.90*
Seq using:										
Marker 1	0.25	13.93	8.10	1.72	-12.68	7.01	-1.81*			
Marker 2	0.67	14.01	9.20	1.52				-15.55	7.96	-1.95*

(c) Parameter values: $p_1 = 0.9, p_2 = 0.9, \rho = 0.8, \theta_1 = 0.1$ and $\theta_2 = 0.1$

Type of Analysis	R^2	β_0			β_1			β_2		
		Est	S.E.	t-val	Est	S.E.	t-val	Est	S.E.	t-val
Simult	0.94	10.19	4.05	2.51	-14.67	7.63	-1.92*	-4.04	2.28	-1.70*
Seq using:										
Marker 1	0.74	12.18	6.69	1.82	-16.73	8.06	-2.07*			
Marker 2	0.63	11.74	5.06	2.32				-7.83	4.36	-1.80*

(d) Parameter values: $p_1 = 0.5, p_2 = 0.5, \rho = 0.1, \theta_1 = 0.5$ and $\theta_2 = 0.5$

Type of Analysis	R^2	β_0			β_1			β_2		
		Est	S.E.	t-val	Est	S.E.	t-val	Est	S.E.	t-val
Simult	0.24	11.07	6.24	1.77	-2.70	8.55	-0.32	-1.96	8.69	-0.23
Seq using:										
Marker 1	0.12	13.43	7.00	1.92	-3.61	5.83	-0.63			
Marker 2	0.08	14.08	8.72	1.62				-2.36	6.95	-0.34

[2] * significant at 5% level

Table 4.4(a). Efficiency of the simultaneous strategy over the sequential strategy for simulation parameter values $p_1 = 0.7, p_2 = 0.5, \theta_1 = 0, \theta_2 = 0.5$.

θ_1	θ_2	n_1	n_2	N_1	N_2	E
0	0	37	31	43	40	1.16
	0.1	39	76	43	98	1.29
	0.2	40	105	43	132	1.26
	0.3	36	168	43	203	1.21
	0.4	41	203	43	262	1.29
0.1	0	82	34	108	40	1.32
	0.1	88	80	108	98	1.23
	0.2	86	112	108	132	1.18
	0.3	92	181	108	203	1.12
	0.4	93	217	108	262	1.21
0.2	0	108	36	153	40	1.42
	0.1	110	87	153	98	1.39
	0.2	118	117	153	132	1.30
	0.3	114	188	153	203	1.08
	0.4	116	224	153	262	1.17
0.3	0	179	34	226	40	1.26
	0.1	176	90	226	98	1.28
	0.2	182	115	226	132	1.25
	0.3	185	191	226	203	1.18
	0.4	184	223	226	262	1.17
0.4	0	226	38	303	40	1.34
	0.1	228	88	303	98	1.33
	0.2	225	123	303	132	1.35
	0.3	231	195	303	203	1.31
	0.4	234	241	303	262	1.26

Table 4.4(b). Efficiency of the simultaneous strategy over the sequential strategy for simulation parameter values $p_1 = 0.7, p_2 = 0.9, \theta_1 = 0.4, \theta_2 = 0.1$.

θ_1	θ_2	n_1	n_2	N_1	N_2	E
0	0	40	50	47	61	1.22
	0.1	42	107	47	119	1.11
	0.2	42	153	47	164	1.07
	0.3	43	201	47	233	1.16
	0.4	44	258	47	312	1.21
0.1	0	103	52	110	61	1.07
	0.1	107	112	110	119	1.06
	0.2	104	152	110	164	1.08
	0.3	106	204	110	233	1.14
	0.4	108	264	110	312	1.18
0.2	0	151	55	157	61	1.04
	0.1	149	115	157	119	1.05
	0.2	152	155	157	164	1.06
	0.3	153	207	157	233	1.13
	0.4	155	270	157	312	1.16
0.3	0	203	57	229	61	1.13
	0.1	203	116	229	119	1.13
	0.2	204	156	229	164	1.12
	0.3	205	216	229	233	1.08
	0.4	208	281	229	312	1.11
0.4	0	272	58	307	61	1.13
	0.1	278	117	307	119	1.10
	0.2	274	159	307	164	1.12
	0.3	277	220	307	233	1.11
	0.4	275	286	307	312	1.09

Table 4.4(c). Efficiency of the simultaneous strategy over the sequential strategy for simulation parameter values $p_1 = 0.9, p_2 = 0.9, \theta_1 = 0.1, \theta_2 = 0.1$.

θ_1	θ_2	n_1	n_2	N_1	N_2	E
0	0	63	68	72	76	1.12
	0.1	65	123	72	143	1.16
	0.2	64	176	72	205	1.16
	0.3	68	254	72	293	1.15
	0.4	68	298	72	351	1.18
0.1	0	112	70	124	76	1.11
	0.1	114	126	124	143	1.14
	0.2	117	173	124	205	1.18
	0.3	115	258	124	293	1.14
	0.4	120	303	124	351	1.16
0.2	0	176	70	187	76	1.06
	0.1	178	128	187	143	1.05
	0.2	176	181	187	205	1.13
	0.3	180	260	187	293	1.13
	0.4	179	305	187	351	1.15
0.3	0	251	72	262	76	1.04
	0.1	254	131	262	143	1.03
	0.2	253	185	262	205	1.04
	0.3	256	264	262	293	1.11
	0.4	259	312	262	351	1.13
0.4	0	310	73	326	76	1.05
	0.1	312	135	326	143	1.04
	0.2	308	188	326	205	1.06
	0.3	311	267	326	293	1.05
	0.4	313	316	326	351	1.11

Table 4.4(d). Efficiency of the simultaneous strategy over the sequential strategy for simulation parameter values $p_1 = 0.5, p_2 = 0.5, \theta_1 = 0.5, \theta_2 = 0.5$.

θ_1	θ_2	n_1	n_2	N_1	N_2	E
0	0	27	30	36	38	1.27
	0.1	28	75	36	84	1.12
	0.2	30	107	36	120	1.12
	0.3	30	158	36	185	1.17
	0.4	32	199	36	232	1.17
0.1	0	72	32	78	38	1.09
	0.1	74	77	78	84	1.09
	0.2	74	109	78	120	1.10
	0.3	73	159	78	185	1.16
	0.4	75	203	78	232	1.14
0.2	0	104	34	112	38	1.08
	0.1	103	77	112	84	1.09
	0.2	105	111	112	120	1.08
	0.3	107	162	112	185	1.14
	0.4	109	208	112	232	1.12
0.3	0	159	35	168	38	1.06
	0.1	161	79	168	84	1.04
	0.2	162	114	168	120	1.04
	0.3	164	167	168	185	1.11
	0.4	165	210	168	232	1.10
0.5	0	200	36	211	38	1.06
	0.1	198	81	211	84	1.07
	0.2	202	117	211	120	1.04
	0.3	203	170	211	185	1.04
	0.4	205	214	211	232	1.08

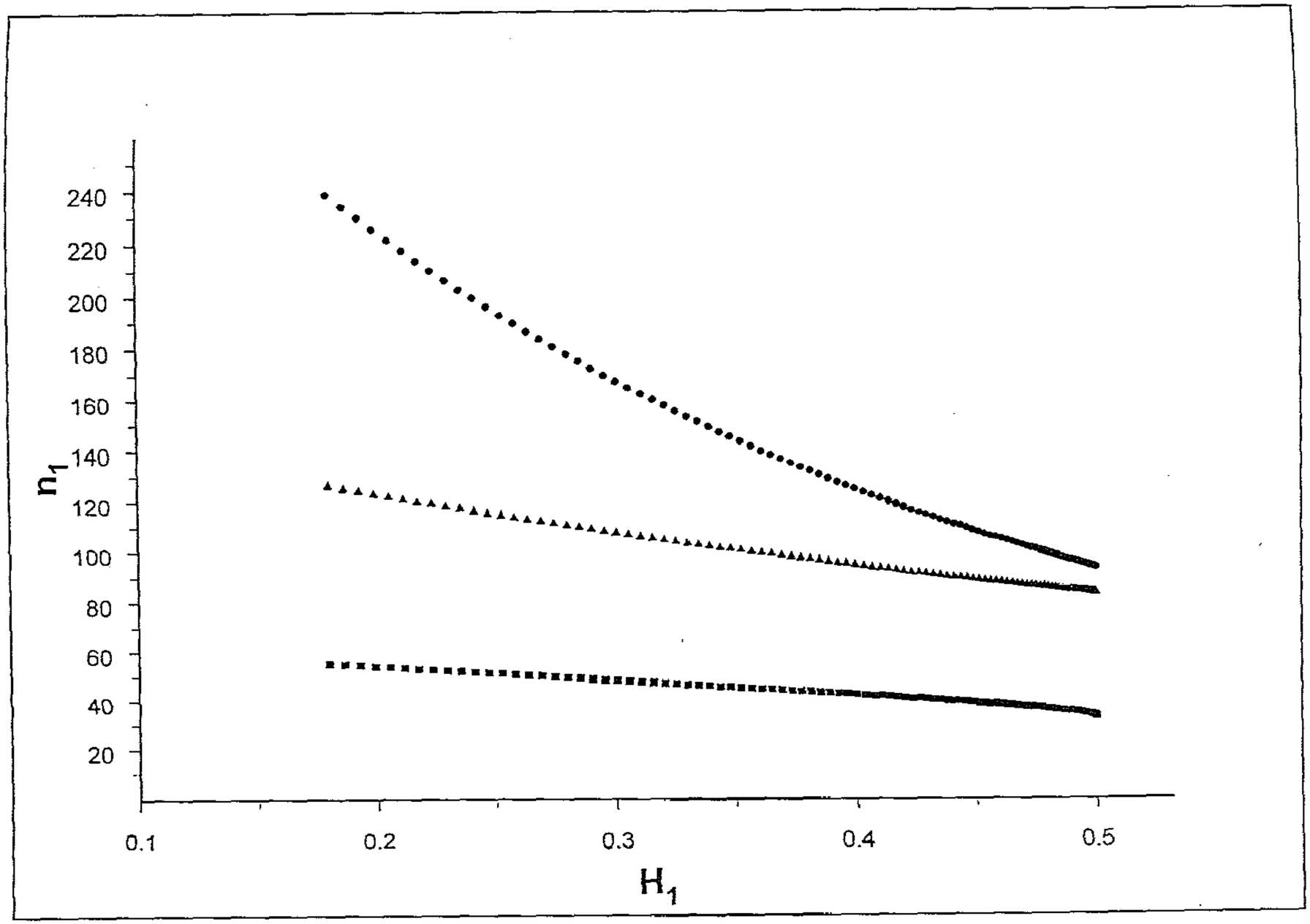


Figure 4.6. Sample size reqirement to map the first trait locus for simulation parameter values $p_2 = 0.5$ and $\Delta = 0$. Circles correspond to $\alpha_1 = \alpha_2 = 2$, triangles to $\alpha_1 = \alpha_2 = 5$ and squares to $\alpha_1 = \alpha_2 = 10$.

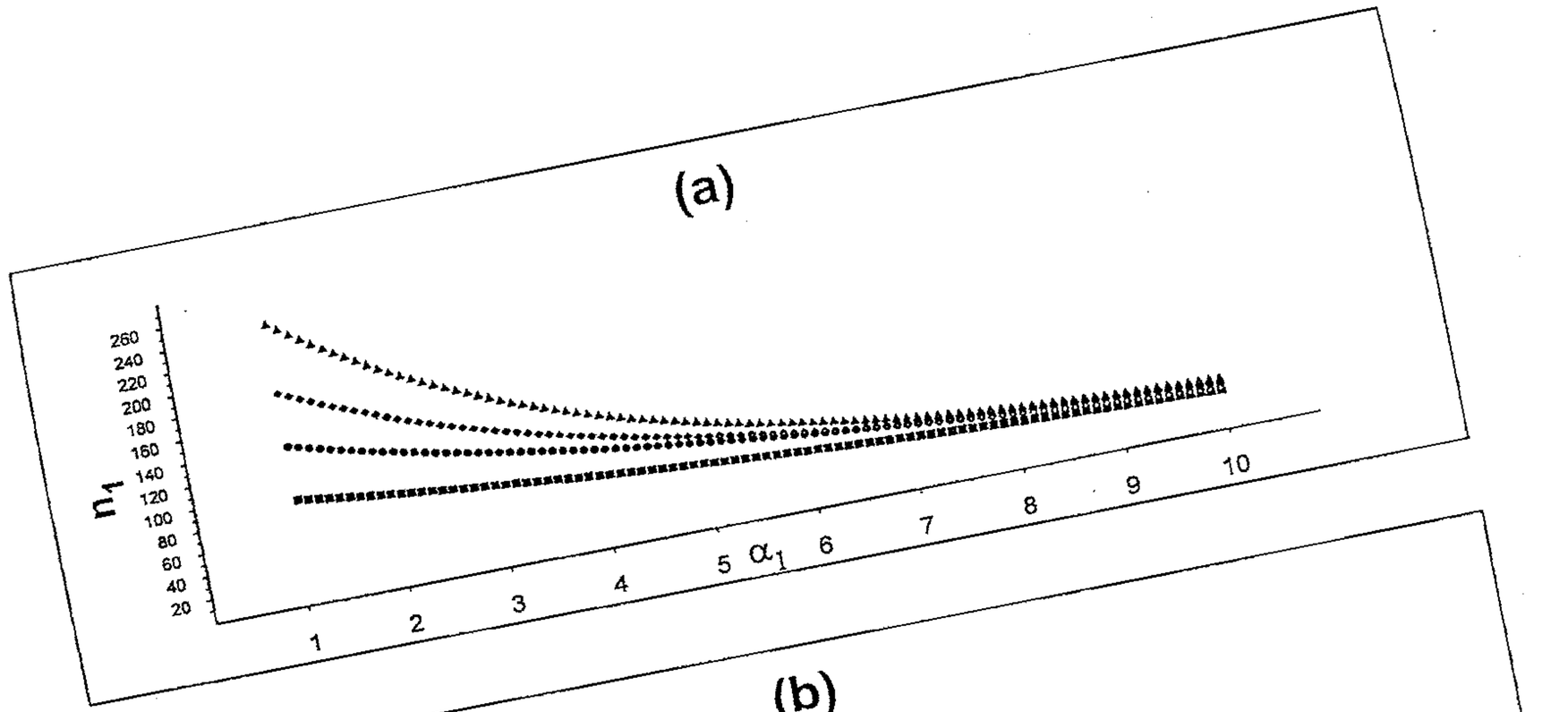


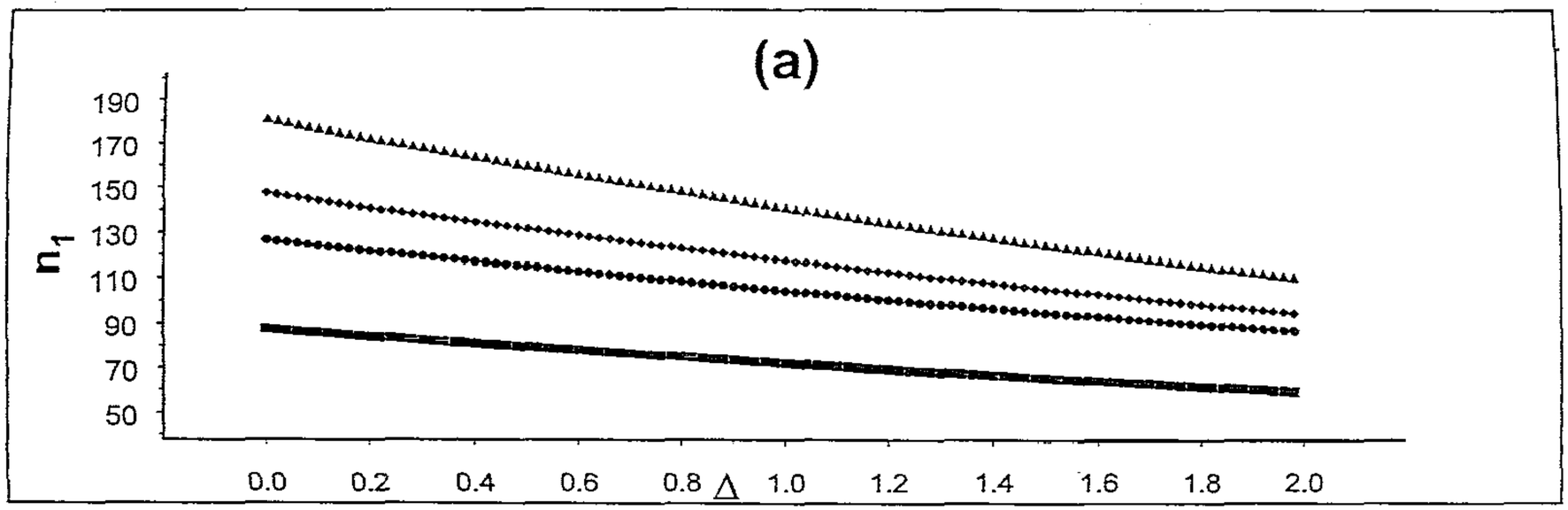
Figure 4.7. Sample size requirement to map (a) the first trait locus and (b) both the trait loci for simulation parameter values $\alpha_2 = 1$ and $\Delta = 0$. Triangles correspond to $p_1 = 0.75, p_2 = 0.5$, diamonds to $p_1 = 0.75, p_2 = 0.75$, circles to $p_1 = 0.5, p_2 = 0.5$ and squares to $p_1 = 0.5, p_2 = 0.75$.

effect of the first locus is quite large compared to that of the second locus. Moreover, while it requires a smaller sample to detect linkage at the first trait locus for $p_1 = p_2 = 0.5$, $p_1 = p_2 = 0.75$ and $p_1 = 0.5, p_2 = 0.75$ whenever the marginal effect of the first locus is higher, the sample size requirement in the case $p_1 = 0.75, p_2 = 0.5$ is more for the first locus if the marginal effects of the two loci do not differ significantly, but is less if the marginal effect of the first locus is very large compared to that of the second locus [Figure 4.7(b)]. Next, we find that the sample size required to map either of the two loci decreases as the degree of epistatic interaction (Δ) between the two loci increases (Figures 4.8-4.10). Moreover in the presence of epistatic interaction, it requires a smaller sample to map the locus with a greater marginal effect on the QT [Figures 4.8(b), 4.9(b), 4.10(b)].

Comparing the simultaneous strategy with the sequential strategy [the results are presented in Tables 4.4 (a)-(d)], we find that the sample size requirement is, in general, less when we analyze the data by considering the two markers simultaneously. The 'Efficiency' ratio E (defined earlier in this Subsection) is found to be greater than 1 in all our simulation studies.

4.3.6 Comparison with the Tiwari-Elston (1997) Method

As we noted earlier, the digenic interaction model is a special case of the more general epistasis model assumed by Tiwari and Elston (1997), in which the epistasis parameters can vary arbitrarily. However, as their model involves a larger number of regressors, the tests for linkage (i.e., $H_0 : \theta_i = 0.5$ vs $H_1 : \theta_i < 0.5$) are much more conservative. We compare the powers of the two procedures under our proposed model using simulated data. We generate data sets with simulation parameter values of $\alpha_1 = 5, \alpha_2 = 1, \Delta = 1, \sigma^2 = 1, \theta_1 = 0.5$ and different values of p_1, p_2, θ_2 for varying sample sizes. We perform 100 replications of regression using each set of parameter values and evaluated the power of the test $H_0 : \theta_1 = 0.5$ vs $H_1 : \theta_1 < 0.5$ at $\theta_1 = 0.1$. The average power of the 100 replications for each set of parameter values is presented in Table 4.5. If our model is indeed true, we find that the tests for linkage under their regression set-up are less powerful especially if the sample size is small. Moreover, in that case, their regression equation is a gross overfit to our model.



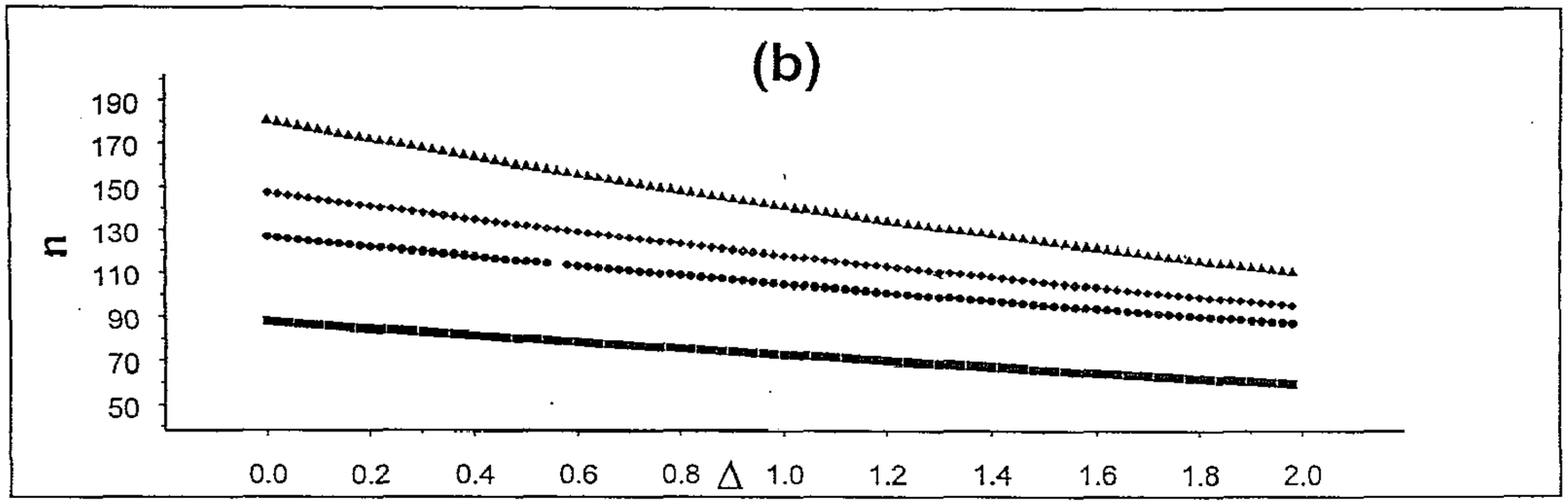
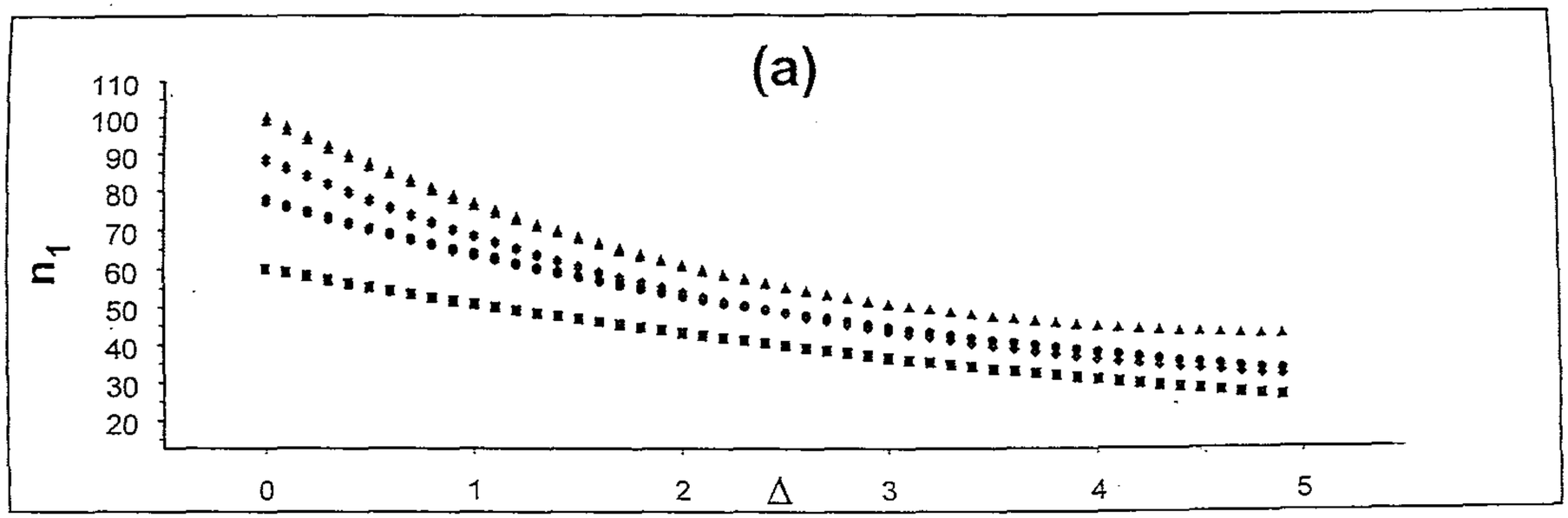


Figure 4.8. Sample size requirement to map (a) the first trait locus and (b) both the trait loci for simulation parameter values $\alpha_1 = 2$ and $\alpha_2 = 1$. Triangles correspond to $p_1 = 0.75, p_2 = 0.5$, diamonds to $p_1 = 0.75, p_2 = 0.75$, circles to $p_1 = 0.5, p_2 = 0.5$ and squares to $p_1 = 0.5, p_2 = 0.75$.



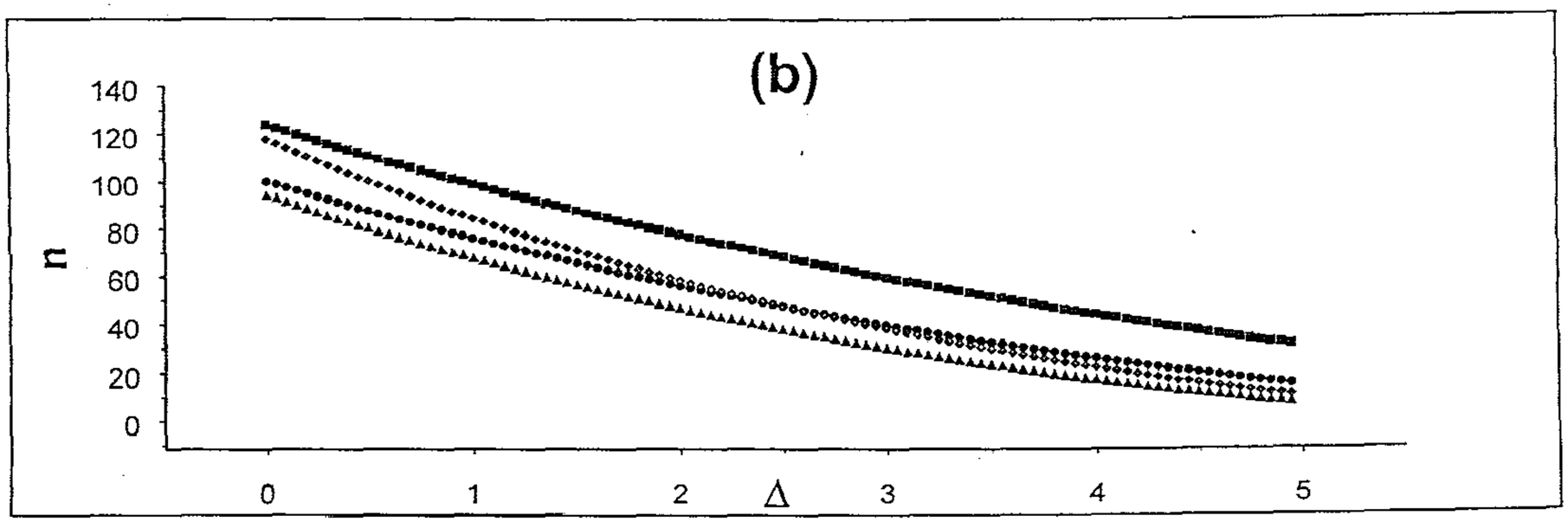


Figure 4.9. Sample size requirement to map (a) the first trait locus and (b) both the trait loci for simulation parameter values $\alpha_1 = 5$ and $\alpha_2 = 1$. Triangles correspond to $p_1 = 0.75, p_2 = 0.5$, diamonds to $p_1 = 0.75, p_2 = 0.75$, circles to $p_1 = 0.5, p_2 = 0.5$ and squares to $p_1 = 0.5, p_2 = 0.75$.

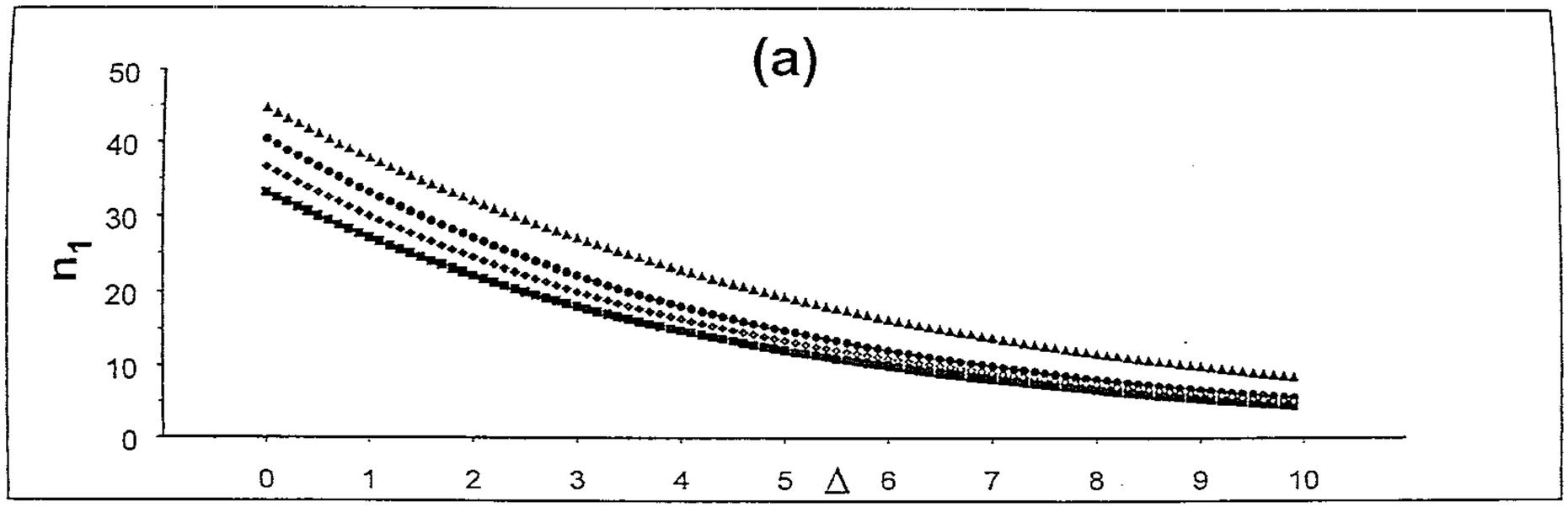


Figure 4.10. Sample size requirement to map (a) the first trait locus and (b) both the trait loci for simulation parameter values $\alpha_1 = 10$ and $\alpha_2 = 1$. Triangles correspond to $p_1 = 0.75, p_2 = 0.5$, diamonds to $p_1 = 0.75, p_2 = 0.75$, circles to $p_1 = 0.5, p_2 = 0.5$ and squares to $p_1 = 0.5, p_2 = 0.75$.

Table 4.5. Comparison of the powers of the digenic interaction model and the Tiwari-Elston model at $\theta_1 = 0.1$ for simulated parameter value of $\theta_1 = 0.5$ and different values of p_1, p_2, θ_2 and number of sib-pairs $n = 50, 100, 200$. [3]

p_1	p_2	θ_1	θ_2	$n = 50$		$n = 100$		$n = 200$	
				P_{DI}	P_{TE}	P_{DI}	P_{TE}	P_{DI}	P_{TE}
0.7	0.5	0.5	0	0.75	0.67	0.83	0.80	0.91	0.90
0.7	0.9	0.5	0.1	0.81	0.74	0.88	0.85	0.95	0.94
0.9	0.9	0.5	0.3	0.72	0.64	0.81	0.79	0.88	0.88
0.5	0.5	0.5	0.5	0.78	0.72	0.85	0.82	0.90	0.89

4.4 Discussion and Overview

Since there is increasing evidence of epistatic interactions among the loci determining a quantitative trait (Lark et al. 1995, Coupland 1995, Fijneman et al. 1996, van Wezel et al. 1996, Chang 1999), we have attempted to devise an efficient estimator of the recombination fraction, θ, between a quantitative trait locus and a marker locus in the presence of such interactions. We have used a simple model of interaction among homozygotes at the different trait loci. This model is one of the basic models used in the study of epistatic interactions (Kearsey and Pooni 1996) and is helpful for capturing some essential features and complexities that underline QTL mapping in presence of epistatic interactions. Using an approach originally proposed by Jayakar (1970), we have proposed separate, computationally simple, estimators for families in which only one parent is heterozygous at the marker locus (backcross type families) and those in which both parents are heterozygous (intercross type families). We have studied the efficiencies of these estimators when there are two trait loci and have shown that for a wide range of parameter values the estimators are quite efficient. We have proposed a non-parametric procedure for testing null hypotheses regarding θ and have shown that the power function of the test has desirable properties. We have also shown that analyses of data ignoring epistatic interactions when in fact these are present may lead to grossly inaccurate inferences about linkage. Although most of our results pertain to the case

[3] P_{DI} and P_{TE} denote the powers of the digenic interaction model and the Tiwari-Elston model at $\theta_1 = 0.1$.

of the marker locus being biallelic, we have theoretically shown that extension to a multiallelic marker locus is straightforward. However, we have found that the estimators obtained by this approach is not as efficient as the maximum likelihood estimators. We also note that our procedure does not provide simultaneous estimates of recombination and other parameters (i.e., quantitative trait locus effects, epistatic interaction effects, etc.). Independent estimates of these other parameters have to be obtained to estimate the recombination fraction and test hypotheses concerning this parameter. The upshot is that although the modified-Jayakar estimator proposed by us is computationally simple and enjoys some desirable statistical properties, in practice it is preferable to use the maximum likelihood estimator in view of its superior performance over a wider range of scenarios and parameter values.

For sib-pair data, we have extended the regression procedure proposed by Haseman and Elston (1972) to map a single QTL, to the case of mapping two unlinked QTLs in the presence of epistatic interactions. Our proposed procedure provides a test for detecting linkage between a trait locus and a marker locus but fails to provide, as in the Haseman-Elston approach, an estimate of the recombination fraction between the two loci. We have derived expressions for the sample size requirements to map the two QTLs. The marginal effects of the different trait and linkage parameters on the sample size requirements have been analysed theoretically. In particular, we have shown that the sample size requirement for mapping a QTL is smaller if its marginal effect and heterozygosity are larger. Moreover, the presence of epistatic interactions reduces the sample size requirement compared to the situation in which the marginal effects are same, but epistatic interactions are absent. We have also assessed the nature of dependence of sample size requirements on different trait parameters considered simultaneously. We have shown through simulation studies that the simultaneous analysis of markers reduce the sample size requirements and thus is more cost effective compared to the sequential analysis. This equivalently implies that under a misspecified model of a single QTL, we would require a larger sample to map the QTL. The proposed regression approach has been extended to the case of multiple QTLs with a typical epistatic interaction structure. The results are similar to the case of two QTLs with digenic epistatic interaction.

Appendix 4.1

Derivation of $E(Y_j|\pi_{j1} = \frac{1}{2}, \pi_{j2} = 1)$

Given below is the table showing the squared difference in trait values of the j^{th} sib pair with i.b.d score at the first trait locus $= \frac{1}{2}$ and that at the second trait locus $= 1$:

	$A_2A_2 - A_2A_2$	$A_2a_2 - A_2a_2$	$a_2a_2 - a_2a_2$
$A_1A_1 - A_1A_1$	${e_j}^2$	${e_j}^2$	${e_j}^2$
$A_1a_1 - A_1a_1$	${e_j}^2$	${e_j}^2$	${e_j}^2$
$a_1a_1 - a_1a_1$	${e_j}^2$	${e_j}^2$	${e_j}^2$
$A_1A_1 - A_1a_1$	$(\alpha_1 + \Delta + e_j)^2$	$(\alpha_1 + e_j)^2$	$(\alpha_1 - \Delta + e_j)^2$
$A_1a_1 - A_1A_1$	$(-\alpha_1 - \Delta + e_j)^2$	$(-\alpha_1 + e_j)^2$	$(-\alpha_1 + \Delta + e_j)^2$
$A_1a_1 - a_1a_1$	$(\alpha_1 + \Delta + e_j)^2$	$(\alpha_1 + e_j)^2$	$(\alpha_1 - \Delta + e_j)^2$
$a_1a_1 - A_1a_1$	$(-\alpha_1 - \Delta + e_j)^2$	$(-\alpha_1 + e_j)^2$	$(-\alpha_1 + \Delta + e_j)^2$

where $e_j = (e_{j1} - e_{j2})$, e_{j1} and e_{j2} being the random errors associated with y_{j1} and y_{j2} respectively.

$$
\begin{aligned}
E(Y_j|\pi_{j1} = \frac{1}{2}, \pi_{j2} = 1) &= E({e_j}^2) + {p_2}^2{\alpha_1}^2(2{p_1}^2q_1 + 2p_1{q_1}^2) + {q_2}^2{\alpha_1}^2(2{p_1}^2q_1 + 2p_1{q_1}^2) \\
&\quad + 2p_2q_2{\alpha_1}^2(2{p_1}^2q_1 + 2p_1{q_1}^2) + {p_2}^2\Delta^2(2{p_1}^2q_1 + 2p_1{q_1}^2) \\
&\quad + {q_2}^2\Delta^2(2{p_1}^2q_1 + 2p_1{q_1}^2) + 2{p_2}^2\alpha_1\Delta(2{p_1}^2q_1 + 2p_1{q_1}^2) \\
&\quad - 2{q_2}^2\alpha_1\Delta(2{p_1}^2q_1 + 2p_1{q_1}^2) \\
&= \phi^2 + 2p_1q_1\{{\alpha_1}^2 + \Delta^2({p_2}^2 + {q_2}^2) + 2\alpha_1\Delta(p_2 - q_2)\}
\end{aligned}
$$

(Note that $E(e_j) = 0$ and $Var(e_j) = \phi^2$)

Chapter 5

A Two-Stage Variable Stringency Semi-Parametric Method for Mapping Based on Genome-Wide Scan Data of Sib-pairs

5.1 Introduction

Genome-wide scans are a powerful approach for mapping genes (Collins 1995, Lander 1996) and have already proven successful (Wyst et al. 1999, Elbein et al. 1999, Niu et al. 1999, Krushkal et al. 1999). In this approach, in addition to collecting data on the trait/disease of interest, genotype data are generated on a large number of markers spread, preferably evenly, across the entire genome. Since the collection of pedigree data is difficult, a popular approach is to collect data on sib-pairs and analyze the data using appropriate statistical techniques (Haseman and Elston 1972, Blackwelder and Elston 1985, Amos and Elston 1989, Amos et al 1989, Lander and Botstein 1989, Goldgar 1990, Haley and Knott 1992, Jansen 1993, Olson and Wijsman 1993, Fulker and Cardon 1994, Olson 1995a,b, Page et al. 1998, Alcais et

al. 1999, Allison et al. 1999). While for qualitative traits in humans various statistical methods, both parametric and non-parametric, for linkage analysis have been proposed and their relative efficiencies extensively tested, for human quantitative traits, such methods are still being developed (Almasy et al. 1998, Olson 1995a, Page et al. 1998, Alcais et al. 1999, Allison et al. 1999) and compared (Williams et al. 1999). Parametric methods for mapping quantitative trait loci (QTL) involve parametric models, and thus, are susceptible to minor deviations in distributional assumptions. The non-parametric methods in current use (Haseman and Elston 1972, Kruglyak and Lander 1995) are relatively more robust, but require specification of the expectation and variance of trait values for each QT genotype, although no assumption regarding the probability distribution of the trait values is made. Further, inferences based on the proposed statistics rely on asymptotic distributions. In this Chapter, we propose a two-step method of locating the most likely position of a QTL on a chromosome given trait values and marker genotypes trait values for a set of sib-pairs. We first consider the trait as being determined by a single QTL with environmental effects, and then extend our procedure to the possibility of the trait being determined by multiple QTLs. When genome-wide scans that involve a large number of markers are performed, a preferred strategy is to use a set of low density markers (say, 5-10cM apart) to identify the region(s) in which the QTL(s) may be located, and then to saturate these identified regions with high density markers (say, 1-5cM apart) to fine-map the QTL. This two-stage approach is cost-effective, both in terms of genotyping and computations. Our proposed two-stage protocol is meant for analyzing sib-pair data data so generated. We use variable stringencies at the two steps of our procedure. A low-stringency is used in the first step in order to reduce the possibility of missing any marker interval that may contain the trait loci. At the second stage of fine-mapping, we use a higher stringency to reduce the false positive error probability. We, however, note that the second step of our procedure can also be directly used for analyzing sib-pair data, although the computational cost will be higher. In any case, from a study-design point of view, the two-stage strategy of data generation and analysis is logically more preferable than a one-step strategy. In the first step, we identify the subset of markers which is linked to the QTL using a test statistic based on rank correlation of es-

timated marker identity-by-descent (i.b.d.) scores and squared difference of sib-pair trait values. In the second step, we perform a non-parametric regression of squared sib-pair trait difference on estimated i.b.d. scores for the different possible pairs of flanking markers using kernel smoothing (Silverman 1986). We term our procedure as semi-parametric, even though we use non-parametric data-analytic procedures at both stages, because of certain underlying model parameters and assumptions (e.g., allele frequency, Hardy-Weinberg equilibrium). We compare our semi-parametric procedure with the parametric regression procedure proposed by Olson (1995a) and show, using Monte-Carlo simulation, that while the parametric method is marginally more efficient than our semi-parametric method when there is no dominance effect at the trait locus (loci), the proposed method is much more efficient in presence of dominance and/or epistasis.

5.2 Model

We assume that a quantitative trait Y is controlled by an autosomal biallelic locus with alleles A_1 and a_1. The expectations of Y conditional on the three genotypes A_1A_1, A_1a_1 and a_1a_1 are assumed to be α, β and $-\alpha$, respectively. The variance of Y within each genotype is assumed to be equal, σ^2. No assumption is made on the shape of the probability distribution of the trait values. The underlying population is assumed to be in Hardy-Weinberg equilibrium with respect to the trait locus. We assume that the trait locus is in linkage equilibrium with a pair of autosomal, biallelic, codominant flanking marker loci.

Suppose $\{(y_{j1}, y_{j2}) : j = 1, 2, \ldots, n\}$ are the observed values of the quantitative trait of n independent sib-pairs. We assume that the expectation of the correlation coefficient between the trait values of any sib-pair is equal, ρ. Let $\pi_{j1}, \pi_{j2}, \ldots, \pi_{jk}$ denote the proportions of alleles shared i.b.d. at k ordered marker loci located on the same chromosome, for the j^{th} sib-pair. The estimation of π_{jl} has been discussed in Section 4.3.1.

Given data on the quantitative trait of the sib-pairs and the estimated i.b.d. scores at the k ordered marker loci, our aim is to determine the most

likely interval in which the trait locus is located.

We define $y_j = (y_{j1} - y_{j2})^2$, $j = 1, 2, \ldots, n$; i.e., y_j denotes the squared pair difference in the trait values for the j^{th} sib-pair.

5.3 Coarse-Mapping Based on Rank Correlation

The first step is to analyze data generated from a genome-wide scan using coarsely-spaced (5-10cM) markers and test whether the tralt locus is at all linked to any of the k ordered marker loci considered. When a trait locus and a marker locus are linked, it is expected that siblings with similar trait values will exhibit considerable sharing of alleles at the marker locus. If the trait and the marker loci are unlinked, then in spite of a significant sharing of alleles i.b.d. between a pair of siblings, their trait values may be largely dissimilar. Thus, a natural test for linkage between the trait locus and the l^{th} marker locus ($l = 1, 2, \ldots, k$), is a test for the strength of correlation between y_js and $\hat{\pi}_{jl}$s. A non-parametric technique of testing for no correlation between y_js and $\hat{\pi}_{jl}$s is based on Spearman's rank correlation (see Randles and Wolfe, 1979).Since $\hat{\pi}_{jl}$ can assume only 5 distinct values (i.e., $0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$), it is expected that there will be many ties in $\hat{\pi}_{jl}$ values. Thus, we need to use Spearman's rank correlation formula for the case of ties, which is given by:

$$R_n = \frac{\frac{n^2-1}{12} - \frac{T_u+T_v}{2} - \frac{1}{2n}\sum_{j=1}^{n} {d_j}^2}{\sqrt{(\frac{n^2-1}{12} - T_u)}\sqrt{(\frac{n^2-1}{12} - T_v)}}$$

where:

d_j= rank(y_j) - rank($\hat{\pi}_{jl}$);

$T_u = \sum_{i=1}^{p}({u_i}^3 - u_i)/12n$;

$T_v = \sum_{i=1}^{q}({v_i}^3 - v_i)/12n$;

there being p ties in y_js of lengths $u_1, u_2, \ldots, u_p$ and q ties in $\hat{\pi}_{jl}$s of lengths $v_1, v_2, \ldots, v_q$.

The test statistic is $\sqrt{n-1}R_n$ which is asymptotically distributed as $N(0,1)$ under the null hypothesis of no correlation. Thus for a level α test, the critical region is given by: $\sqrt{n-1}|R_n| > z_{\alpha/2}$ where z_m is the $(1-m)^{th}$ quantile of a standard normal variate.

If the null hypothesis of no correlation is accepted for all the k

marker loci (the level of significance adjusted to α/k to account for the multiple tests), then our conclusion is that the trait locus is most probably not located on the same chromosome as the k marker loci.

Using the above test procedure, we select those marker loci for which the null hypothesis of no correlation between y_js and $\widehat{\pi}_{jl}$s is rejected, i.e., those marker loci which show evidence of linkage with the trait locus. In the next Section, we consider two such consecutive marker loci as candidate markers flanking the trait locus.

5.4 Fine-Mapping Based on Non-parametric Regression

Since at the first-stage of the genome-wide scan, the marker-spacing is coarse, the distance between the two markers found to provide the highest evidence of linkage (highest value of the rank correlation) with the QTL is 5-10cM. This genomic region/interval is, at the second stage, covered with densely-spaced markers and the data thus generated are analyzed for the purpose of fine-mapping the QTL. Let us assume that this region/interval is covered with M densely-spaced markers. Consider, without loss of generality, the ordered consecutive densely-spaced markers 1 and 2. We propose a non-parametric additive regression model given by:

$$y_j = \psi_1(\widehat{\pi}_{j1}) + \psi_2(\widehat{\pi}_{j2}) + e_j;\ j = 1, 2, \ldots, n;$$

where ψ_1, ψ_2 are real valued functions of $\widehat{\pi}_1$ and $\widehat{\pi}_2$, respectively, and e_js are random errors. The regression model is motivated by the fact that the estimated i.b.d. scores of siblings at both marker loci 1 and 2 are found to be individually significantly correlated with the squared difference of the trait values (y). However the nature of dependence of the estimated i.b.d. scores $\widehat{\pi}_{j1}$ and $\widehat{\pi}_{j2}$ on y_j is a function of the recombination distances between the marker and trait loci and other biological parameters, such as interference and dominance at the trait locus. Hence, we do not assume any specific form of the functions ψ_1 and ψ_2, but only general functional forms to model the nature of dependence between $(\widehat{\pi}_{j1}, y_j)$ and $(\widehat{\pi}_{j2}, y_j)$. The functional forms are estimated from the data. Estimates of ψ_1 and ψ_2 are obtained in steps and iteratively using kernel-smoothing techniques (see Silverman, 1986). In

this technique of non-parametric regression, the domains of the explanatory variables are divided into a number of windows. Local smoothing is carried out within each window and appropriate adjustments are made to ensure continuity at window boundaries. In step 1, we perform a non-parametric regression analysis of y on $\widehat{\pi}_1$ (details given later) and obtain $\widehat{\psi}_1$, an estimate of ψ_1. In step 2, we replace y by $y^* = y - \widehat{\psi}_1(\widehat{\pi}_1)$. In step 3, we regress y^* on $\widehat{\pi}_2$ to obtain $\widehat{\psi}_2$, an estimate of ψ_2. In step 4, we compute the residual sum of squares given by $\sum_{j=1}^{n} \{y_j - \widehat{\psi}_1(\widehat{\pi}_{j1}) - \widehat{\psi}_2(\widehat{\pi}_{j2})\}^2$. We then restart the process at step 1 and perform a regression analysis of $y^{**} = y - \widehat{\psi}_2(\widehat{\pi}_2)$ on $\widehat{\pi}_1$. We continue to iterate until $\widehat{\psi}_1$ and $\widehat{\psi}_2$ stabilize reasonably, i.e., the residual sum of squares differ negligibly ($< \epsilon$, a small predetermined positive real number) in two successive iterations. The stringency parameter, ϵ, is obviously variable. Let the final residual sum of squares obtained be denoted by $CV(1,2)$, and in general, by $CV(l, l+1)$ when the l^{th} and $(l+1)^{th}$ marker loci are considered. The most likely position of the trait locus is given by the interval flanked by the i^{th} and $(i+1)^{th}$ marker loci where i corresponds to:

$$CV(i, i+1) \;=\; min_l \;\; CV(l, l+1)$$

In order to regress y on $\widehat{\pi}_1$, the range of $\widehat{\pi}_1$ is divided into windows of length h. The kernel function used is:

$$\kappa(t) = \;\; \tfrac{3}{4}(1-t^2), \;\; \text{if } |t| < 1;$$
$$0, \text{ otherwise}$$

The kernel estimator of ψ_1 is given by:

$$\widehat{\psi}_1(x) = \frac{\sum_{j=1}^{n} \kappa(\frac{x - \widehat{\pi}_{j1}}{h}) y_j}{\sum_{j=1}^{n} \kappa(\frac{x - \widehat{\pi}_{j1}}{h})}$$

Since non-parametric regression tends to overfit data (Silverman 1986), we use the "leave-one-out technique", i.e., leave out the observation $(y_j, \widehat{\pi}_{j1})$ in order to predict y_j. The predictor of y_j is given by:

$$\begin{aligned} \widehat{y}_j &= \widehat{\psi}_1(\widehat{\pi}_{j1}) \\ &= \frac{\sum_{i \neq j} \kappa(\frac{\widehat{\pi}_{j1} - \widehat{\pi}_{i1}}{h}) y_i}{\sum_{i \neq j} \kappa(\frac{\widehat{\pi}_{j1} - \widehat{\pi}_{i1}}{h})} \end{aligned}$$

For the given window length h, the total error in prediction is given by $R_h = \sum_{j=1}^{n} (y_j - \widehat{y}_j)^2$. The process is repeated for different window lengths. The optimal window length h^* is given by that h for which R_h is minimum.

5.5 A Linear Regression Strategy in Current Use

Suppose A_1 and a_1 denote the alleles at the trait locus. Let the conditional expectation, $E(Y)$, of the quantitative character Y given the genotypes at the trait locus be $\alpha, 0$ and $-\alpha$ for A_1A_1, A_1a_1 and a_1a_1 respectively. If $\widehat{\pi}_1$ and $\widehat{\pi}_2$ denote the estimated i.b.d. scores at the two marker loci flanking the trait locus, Olson (1995a) showed that:

$$E(y_j|\widehat{\pi}_{j1}, \widehat{\pi}_{j2}) = \beta_0 + \beta_1\widehat{\pi}_{j1} + \beta_2\widehat{\pi}_{j2} \tag{5.1}$$

for some constants β_0, β_1 and β_2.

Thus, a strategy of determining the location of the trait locus is based on linear regression of y_js on i.b.d. scores of possible pairs of flanking markers. If the l^{th} and $(l+1)^{th}$ marker loci are considered, y_j is predicted by $\widehat{y}_j = \widehat{\beta}_0 + \widehat{\beta}_1\widehat{\pi}_{jl} + \widehat{\beta}_2\widehat{\pi}_{j(l+1)}$ where $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\beta}_2$ are the least squares estimators of β_0, β_1 and β_2 respectively. Tests for linkage are equivalent to tests of β_1 and β_2 which involve parametric test statistics. The error sum of squares is given by $E(l, l+1) = \sum_{j=1}^{n} (y_j - \widehat{y}_j)^2$. The most likely interval of the trait locus location is given by that flanked by the i^{th} and $(i+1)^{th}$ markers if and only if:

$$E(i, i+1) = min_l \;\; E(l, l+1)$$

Using Monte-Carlo simulations, we examine the relative efficiencies of the proposed non-parametric procedure and the parametric method developed by Olson (1995a). In regression analysis, to avoid regressional overfits to data, it is statistically desirable to use the "leave-one-out" technique for predicting y_j, which is what we prescribe and use for our semi-parametric regression procedure. However, in Olson's (1995a) parametric regression procedure, this was not prescribed and is perhaps not used in practice. For purposes of comparing our proposed method with Olson's (1995a) method, we use the leave-one-out technique for both methods. We also compute

and compare the error sum of squares without using the leave-one-out technique for Olson's (1995a) method, although such comparisons are not strictly valid, because it is expected *a priori* that the error sum of squares without leave-one-out will be smaller than with leave-one-out in view of overfitting.

We note that equation (5.1) is valid only when there is no dominance at the trait locus. When there is dominance, the conditional expectation on the left hand side of equation (5.1) is not a linear function of $\widehat{\pi}_{j1}$ and $\widehat{\pi}_{j2}$. Hence, the use of the linear regression given in equation (5.1) may yield incorrect inferences.

5.6 Simulation

In order to assess the performance of our proposed non-parametric regression strategy, and to compare it with the parametric regression strategy described in the previous Section, we generate data on trait values of sib-pairs and estimated marker i.b.d. scores for different sets of parameter values. The different steps of the simulation algorithm have been described in Section 2.2.

Having generated the required data on n independent sib-pairs, we use the proposed test of linkage based on rank correlation in order to select the possible pairs of flanking markers. We then perform both the non-parametric and parametric regressions to determine the most likely position of the trait locus. For the non-parametric regression, the stringency parameter, c is kept fixed at 0.001.

5.7 Results

In what follows, we denote the trait parameters as :

Effect of the genotype A_1A_1 on trait values=α.

Dominance effect of the trait locus=β.

Frequency of allele A_1=p.

Variance of the trait values within any trait genotype=σ^2.

Correlation coefficient between the trait values of any sib-pair=ρ.

5.7.1 Identifying probable interval locations of the QTL

In order to assess the performance of the rank correlation statistic to identify the interval location of the QTL, we generate data on 100 ordered, equispaced markers, such that the recombination fraction between any two consecutive markers is 0.05. Simulated data are generated assuming that the trait locus is flanked by the 24^{th} and 25^{th} markers and that the recombination fraction between the trait locus and the 24^{th} marker is 0.02. The trait parameter values used in the simulation are $\alpha = 5; \beta = 0, 2, 4; p = 0.7; \sigma^2 = 1; \rho = 0.6$ (or higher). The nature of the absolute rank correlation between the different markers and the squared difference in trait values of the sib-pairs are presented in Figures 5.1(a)-(c), for $\beta = 0, 2$ and 4 , respectively. From the figures, we find that the absolute rank correlation increases with the proximity of the considered marker to the trait locus. The peak is at the 24^{th} marker, correctly indicating the approximate location of the trait locus. Though with increase in β, (i.e., the dominance effect) the peak becomes less pronounced, the approximate position of the trait locus is fairly clear even for high dominance effect.

To investigate the effect of changing α, we present in Figures 5.2(a)-(c), graphs similar to those as Figures 5.1(a)-(c) but with $\alpha = 3$ and $\beta = 0, 1$ and 2, respectively. As evident from these figures, although the mean values of the rank correlation become slightly smaller, the nature of the graphs and hence qualitative inferences remain unchanged.

The variation in the values of the rank correlation across the 1000 simulation replications is extremely small, for every set of parameter values. We present in Figure 5.3, the empirical 95% confidence band for a section of the graph presented in Figure 5.1(a). [The empirical confidence bands are so narrow that these are not clearly presentable on Figures 5.1(a)-(c).] This indicates another desirable statistical property of our proposed method.

5.7.2 Finer localization of the QTL

Having identified the interval in which the QTL may be located, in practice one saturates this interval with more dense markers to arrive at a finer localization of the QTL. To simulate this practice, we consider data on

multiple markers that are more densely located within the coarse interval identified at the previous stage. In our simulations, we generated data on a set of M ordered markers. We use the following notations:

θ_2, θ_3 =recombination fraction between the trait locus and the nearest flanking markers 2 and 3, respectively.

θ_1=recombination fraction between markers 1 and 2.

θ_4= recombination fraction between markers 3 and 4

θ_5= recombination fraction between markers 4 and 5.

We use simulation parameter values of $M = 5; \alpha = 5; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01; \beta = 0, 2, 4$ and different parameter values of p and ρ such that the proportion of variance in the trait explained by the QTL varies between 85-95 %. For each set of parameter values, we perform 1000 iterations. The results are given in Table 5.1. In all the cases, the five markers are found to be linked to the trait locus at 1% level of significance. Thus we have 4 candidate intervals (i.e., those flanked by markers 1 and 2, 2 and 3, 3 and 4, 4 and 5) in which the trait locus may be located.

When $\beta = 0$ (i.e., there is no dominance effect), equation (5.1) holds. Thus, it is expected that the parametric approach will be more efficient. We find that in almost all replications, both the methods correctly identify the interval in which the QTL is located. Though the parametric regression has a smaller error in prediction, the error in the non-parametric regression is not much larger. The error in prediction is lowest for the parametric regression without leave-one-out (P2). This, as mentioned earlier, is not unexpected because without leave-one-out there is obviously overfitting of the regression model to the data. Since we, therefore, recommend and use the leave-one-out technique, the apropriate comparison of prediction errors with our non-parametric approach and the parametric approach should be between columns NP and P1. (P2 is presented for completeness as the leave-one-out technique may not be used in practice, even though it should be used to avoid false inferences from model overfits.) When $\beta = 2$ or 4, equation (5.1) does not hold. In presence of dominance, while the non-parametric approach identifies the correct interval in 91% of the cases when $\beta = 2$, the parametric approach does so in only 83-84% of the cases. The non-parametric approach has a smaller error in prediction. When $\beta = 4$,

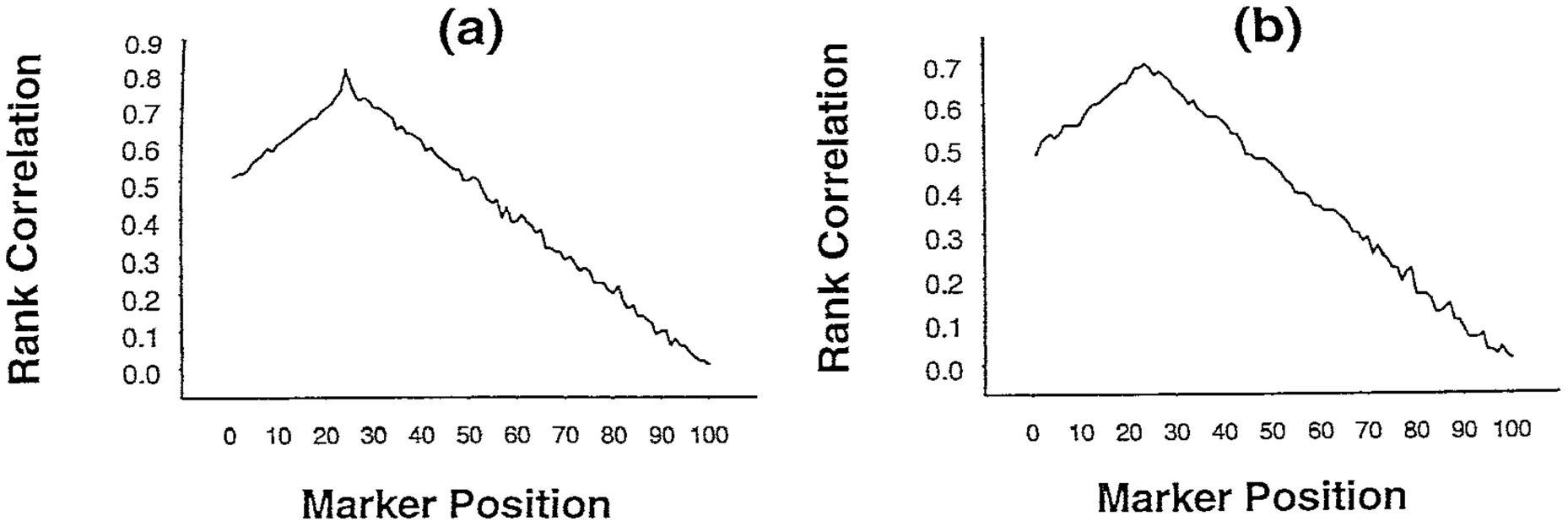


Figure 5.1. Mean rank correlation, based on 1000 replications, between squared difference of trait values of a sib-pair and estimated i.b.d. scores at 100 ordered markers with simulation parameter values $\alpha = 5, \sigma^2 = 1, p = 0.7, \rho = 0.6$ and (a) $\beta = 0$; (b) $\beta = 2$; (c) $\beta = 4$ based on 100 sib-pairs.

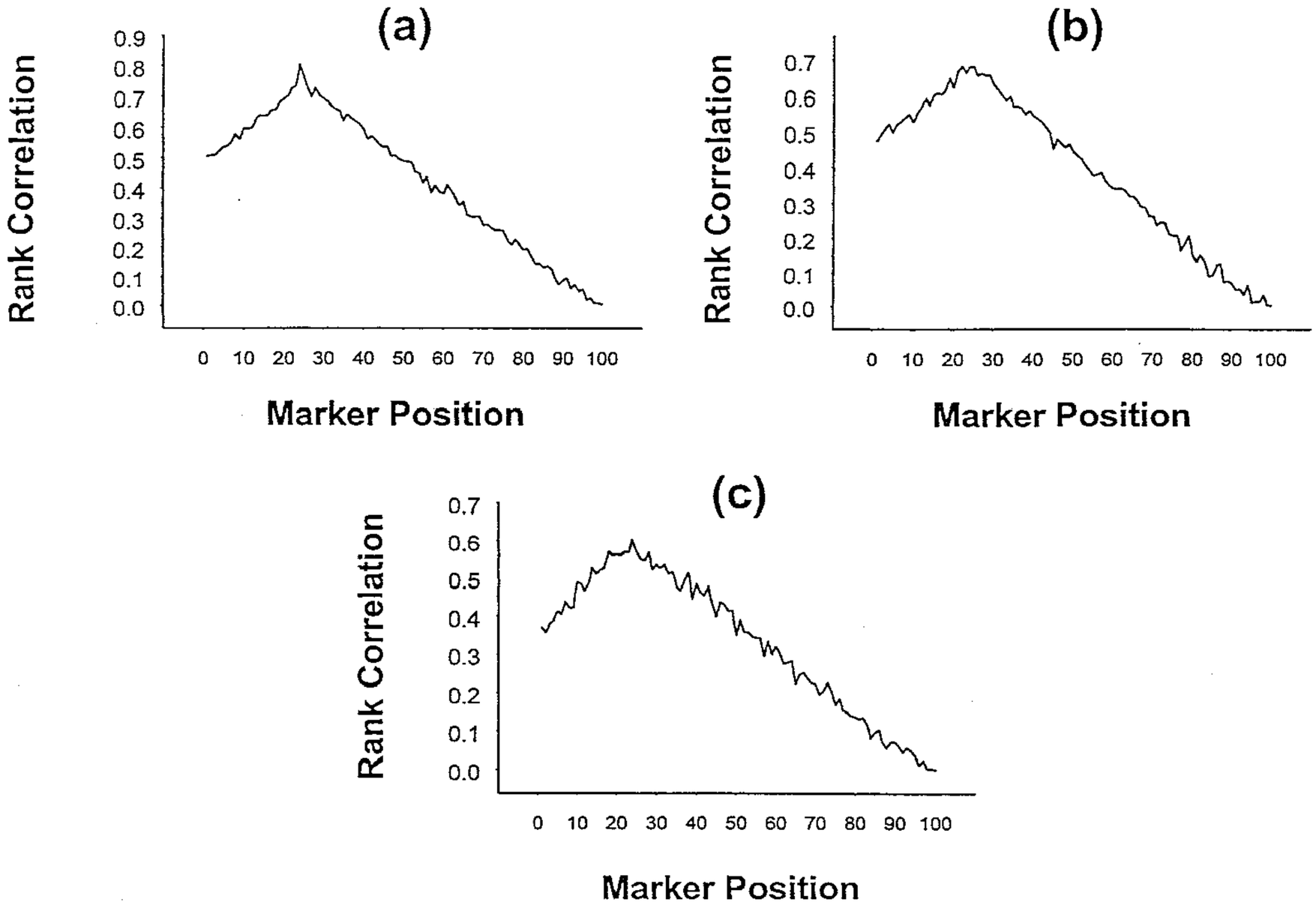


Figure 5.2. Mean rank correlation, based on 1000 replications, between squared difference of trait values of a sib-pair and estimated i.b.d. scores at 100 ordered markers with simulation parameter values $\alpha = 3, \sigma^2 = 1, p = 0.7, \rho = 0.6$ and (a) $\beta = 0$; (b) $\beta = 1$; (c) $\beta = 2$ based on 100 sib-pairs.

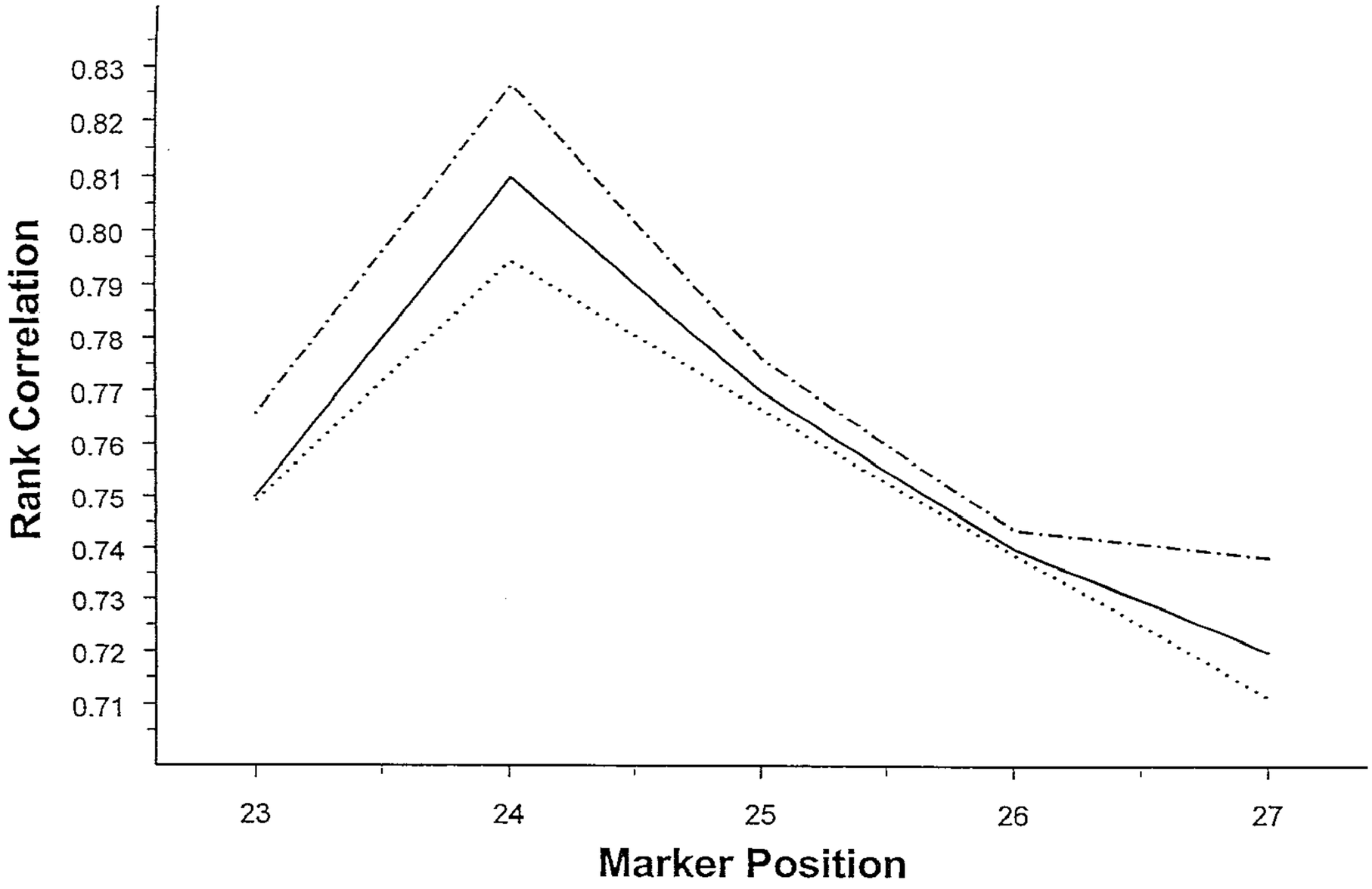


Figure 5.3. Mean rank correlation (solid line), based on 1000 replications, between squared difference of trait values of a sib-pair and estimated i.b.d. scores at markers around the true QTL location and empirical 95% confidence band (dotted lines) for simulation parameter values $\alpha = 5, \beta = 0, \sigma^2 = 1, p = 0.7, \rho = 0.6$.

Table 5.1. Comparison between the non-parametric and parametric regressions based on average prediction error (residual sums of squares averaged over 1000 replications) in the case of a single QTL using 100 Sib-pairs with simulation parameter values of $\alpha = 5; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$. [1]

(a) $\beta = 0, p = 0.5, \rho = 0.8$

Candidate Interval	Error in Prediction		
	NP (97.2%)	P1 (98.5%)	P2 (98.9%)
(1,2)	100.56	95.46	92.71
(2,3)	87.65	74.72	70.62
(3,4)	104.29	99.55	97.68
(4,5)	117.03	110.84	107.27

(b) $\beta = 2, p = 0.9, \rho = 0.7$

Candidate Interval	Error in Prediction		
	NP (90.7%)	P1 (82.5%)	P2 (84.1%)
(1,2)	152.76	157.63	155.35
(2,3)	143.37	148.54	146.72
(3,4)	152.90	160.81	157.64
(4,5)	166.29	173.06	171.18

(c) $\beta = 4, p = 0.7, \rho = 0.5$

Candidate Interval	Error in Prediction		
	NP (75.8%)	P1 (43.0%)	P2 (51.5%)
(1,2)	182.45	196.74	193.27
(2,3)	180.34	194.68	191.93
(3,4)	185.74	194.52	193.22
(4,5)	190.59	207.02	203.51

[1] 'NP' = Non-parametric; P1 = Parametric with leave-one-out; P2 = Parametric without leave-one-out, that is, standard parametric regression. Figures in parentheses denote the percentages of correct identification of true interval location.

(i.e., when there is a high dominance effect), the performance of the parametric approach is very poor compared to the non-parametric approach. While the percentage of correct interval identification using the parametric regression is only 43-51%, that using the non-parametric regression is 76%. The average prediction error under this scenario is also much higher for the parametric, than the non-parametric, regression method. Thus, we find that while the non-parametric approach performs almost as efficiently as the parametric approach when there is no dominance effect; it performs increasingly better than the parametric approach as the dominance effect increases.

We also investigate the effect of changing the values of the parameters α and β. In Table 5.2, we present the results similar to those in Table 5.1, but with $\alpha = 3$ and $\beta = 0, 1, 2$. Qualitatively, the inferences are similar to those derived from Table 5.1; the parametric regression performs better than the non-parametric regression in the absence of dominance, but the converse is true in the presence of dominance. We find that the percentages of correct identification and/or the prediction errors presented in Table 5.2 are higher than those in Table 5.1. This is because the proportion of trait variance explained by the QTL is a function of α and β in addition to other parameters; this proportion decreases with reduction in α for fixed values of β and other parameters. In other words, there is decrease in efficiencies of performance of both non-parametric and parametric procedures as the proportion of trait variance explained by the QTL decreases.

We, likewise, investigate the effect of changes in trait allele frequencies for fixed values of α, β and other parameters. Results are presented in Table 5.3. We find that as p deviates from 0.5 (for fixed values of the other parameters), the percentage of correct interval identification decreases and the error in prediction increases for both non-parametric and parametric regression methods. This, as explained in the preceding paragraph, is not unexpected because for fixed values of the other parameters, the proportion of trait variance explained by the QTL decreases as p deviates from 0.5. With dominance, the non-parametric method performs better than the parametric method for all values of the trait allele frequency.

Table 5.2. Comparison between the non-parametric and parametric regressions based on average prediction error (residual sums of squares averaged over 1000 replications) in the case of a single QTL using 100 sib-pairs with simulation parameter values of $\alpha = 3; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$. [2]

(a) $\beta = 0, p = 0.5, \rho = 0.8$

Candidate Interval	Error in Prediction		
	NP (95.2%)	P1 (97.0%)	P2 (97.8%)
(1,2)	104.72	98.44	95.61
(2,3)	92.83	78.69	74.25
(3,4)	106.29	101.54	99.02
(4,5)	122.18	114.84	110.49

(b) $\beta = 1, p = 0.9, \rho = 0.7$

Candidate Interval	Error in Prediction		
	NP (88.4%)	P1 (80.6%)	P2 (82.9%)
(1,2)	162.26	167.05	165.11
(2,3)	147.75	154.68	151.83
(3,4)	164.90	168.32	165.17
(4,5)	179.44	188.69	184.72

(c) $\beta = 2, p = 0.7, \rho = 0.5$

Candidate Interval	Error in Prediction		
	NP (71.5%)	P1 (40.4%)	P2 (43.7%)
(1,2)	196.65	211.76	203.38
(2,3)	188.07	200.55	197.63
(3,4)	199.19	213.01	206.05
(4,5)	212.92	225.47	218.86

[2] NP = Non-parametric; P1 = Parametric with leave-one-out; P2 = Parametric without leave-one-out, that is, standard parametric regression. Figures in parentheses denote the percentages of correct identification of true interval location.

Table 5.3. Comparison between the non-parametric and parametric regressions based on average prediction error (residual sums of squares averaged over 1000 replications) for different allele frequencies of the QTL in the case of a single QTL using 100 sib-pairs with simulation parameter values of $\alpha = 5; \beta = 2; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$. [3]

(*a*) $p = 0.9$

Candidate Interval	Error in Prediction		
	NP (90.7%)	P1 (82.5%)	P2 (84.1%)
(1,2)	152.76	157.63	155.35
(2,3)	143.37	148.54	146.72
(3,4)	152.90	160.81	157.64
(4,5)	166.29	173.06	171.18

(*b*) $p = 0.7$

Candidate Interval	Error in Prediction		
	NP (92.7%)	P1 (85.3%)	P2 (87.8%)
(1,2)	146.52	153.67	150.29
(2,3)	135.44	143.48	140.03
(3,4)	150.41	155.13	152.45
(4,5)	166.22	176.52	171.38

(*c*) $p = 0.5$

Candidate Interval	Error in Prediction		
	NP (94.5%)	P1 (87.0%)	P2 (89.7%)
(1,2)	139.80	146.26	142.97
(2,3)	123.04	137.51	131.58
(3,4)	141.36	150.75	144.84
(4,5)	157.59	168.22	164.17

[3] NP = Non-parametric; P1 = Parametric with leave-one-out; P2 = Parametric without leave-one-out, that is, standard parametric regression. Figures in parentheses denote the percentages of correct identification of true interval location.

5.7.3 Assessment of Type I Error

To determine the efficacy of a statistical procedure, it is imperative that the type I error rate be assessed. In the present context, type I error probability is the probability of rejection of the null hypothesis of no linkage between the QTL and any of the markers considered when actually the QTL is unlinked to the markers. To assess this, we generate the trait values from the underlying distribution, details of which have been provided earlier. The i.b.d. scores of the sib-pairs at the various marker loci are generated from a trinomial distribution, independent of the trait i.b.d. scores. This ensures that the QTL is unlinked to any of the markers considered. Such data are generated on 100 sib-pairs for each replication; 1000 replications are performed.

These data are then analyzed using the rank correlation statistic, as prescribed for the first-stage of our two-stage procedure. The values of the rank correlation, averaged over 1000 replicatons, for the set of 100 ordered markers are graphically presented in Figure 5.4. It is observed that the mean rank correlation values are all small, and are statistically non-significant. This inference holds at all levels of dominance at the trait locus. Thus, the empirical estimate of the type I error probability is zero.

In practice, since a fine-mapping protocol is undertaken only when some "probable" intervals are identified at the first-stage based on statistically significant values of the rank correlation, in the present case, there is no need for further investigation as the null hypothesis is accepted for all the markers considered.

5.7.4 Effect of sample size

In order to assess the effect of reducing the sample size on our proposed procedure, we simulated the required data on 50 and 25 sib-pairs with varying dominance effect on the trait. The nature of the absolute rank correlations between the trait value and the estimated i.b.d. scores at the 100 generated markers are presented in Figures 5.5(a)-(c) and 5.6(a)-(c) for sample sizes 50 and 25 respectively. Compared to the rank correlations based on 100 sib-pairs [as illustrated in Figures 5.1(a)-(c)], the rank correla-

tions, in general, decrease with decrease in sample size. However, the peak at the 24^{th} marker is prominent even with 25 sib-pairs. Thus the approximate position of the trait locus is indicated correctly even for small sample sizes. The effect of dominance on the rank correlations is identical to that discussed in an earlier Subsection with 100 sib-pairs.

We repeat the non-parametric regression using 50 and 25 sib-pairs with the same set of parameter values and 5 markers as before. The results are presented in Tables 5.4 and 5.5, respectively. We find that the percentage of correct identification of flanking markers decreases with decrease in sample size both for the parametric and the non-parametric regression procedures. The rate of decrease is more when the dominance effect is high (i.e., $\beta = 4$). As is observed with 100 sib-pairs, with smaller sample sizes too, we find that while the performance of the non-parametric regression approach is similar to the parametric regression approach when there is no dominance effect; the performance of the non-parametric regression procedure is significantly better when the degree of dominance in the trait is high. Further, the non-parametric method performs increasingly better with decreasing sample size, in the presence of dominance effects.

5.7.5 Effect of deviation from Normality

Non-parametric statistical procedures are usually less sensitive to minor deviations in distributional assumptions. Both the linear regression procedure (Olson 1995a) and our proposed non-parametric regression procedure are expected to be robust with respect to the underlying trait distribution of the sib-pairs. We note that the test procedure in Olson's method (1995a) is based on distributional assumptions. Thus it is of considerable interest to assess the performance of both the procedures when there is deviation from the assumed trait distribution. One of the existing methods of evaluating the effect of deviation is to introduce local perturbations in the original distribution. In our previous simulation examples, we had generated the trait values of the sib-pairs from a bivariate normal distribution. In order to assess the effect of the trait distribution deviating from normal on the identification of interval location of the QTL, we perturb the relevant bivariate normal distributions with an exponential distribution with mean 1. In order

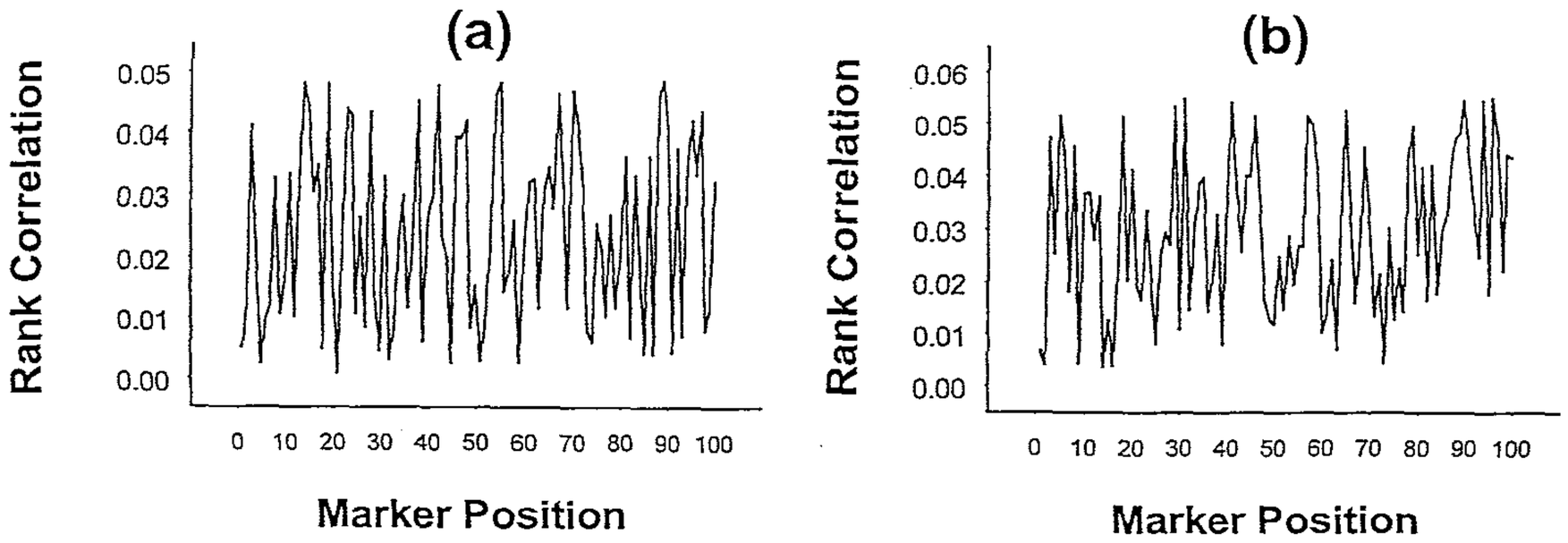


Figure 5.4. Mean rank correlation, based on 1000 replications, between squared difference of trait values of a sib-pair and estimated i.b.d. scores at 100 ordered markers, all unlinl. d to the QTL, with simulation parameter values $\alpha = 5, \sigma^2 = 1, p = 0.7, \rho = 0.6$ and (a) $\beta = 0$; (b) $\beta = 2$; (c) $\beta = 4$ based on 100 sib-pairs.

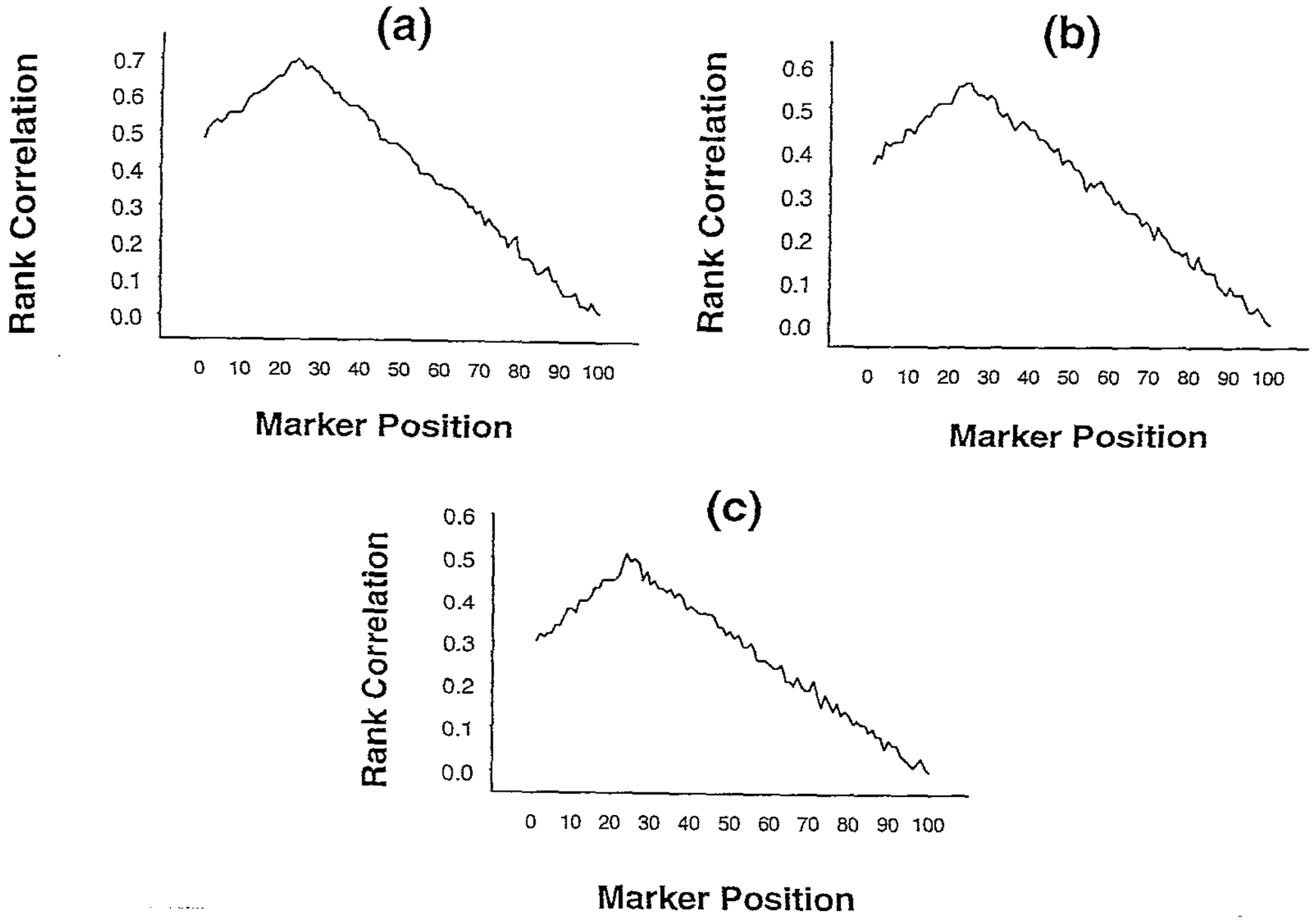


Figure 5.5. Mean rank correlation, based on 1000 replications, between squared difference of trait values of a sib-pair and estimated i.b.d. scores at 100 ordered markers with simulation parameter values $\alpha = 5, \sigma^2 = 1, p = 0.7, \rho = 0.6$ and (a) $\beta = 0$; (b) $\beta = 2$; (c) $\beta = 4$ based on 50 sib-pairs.

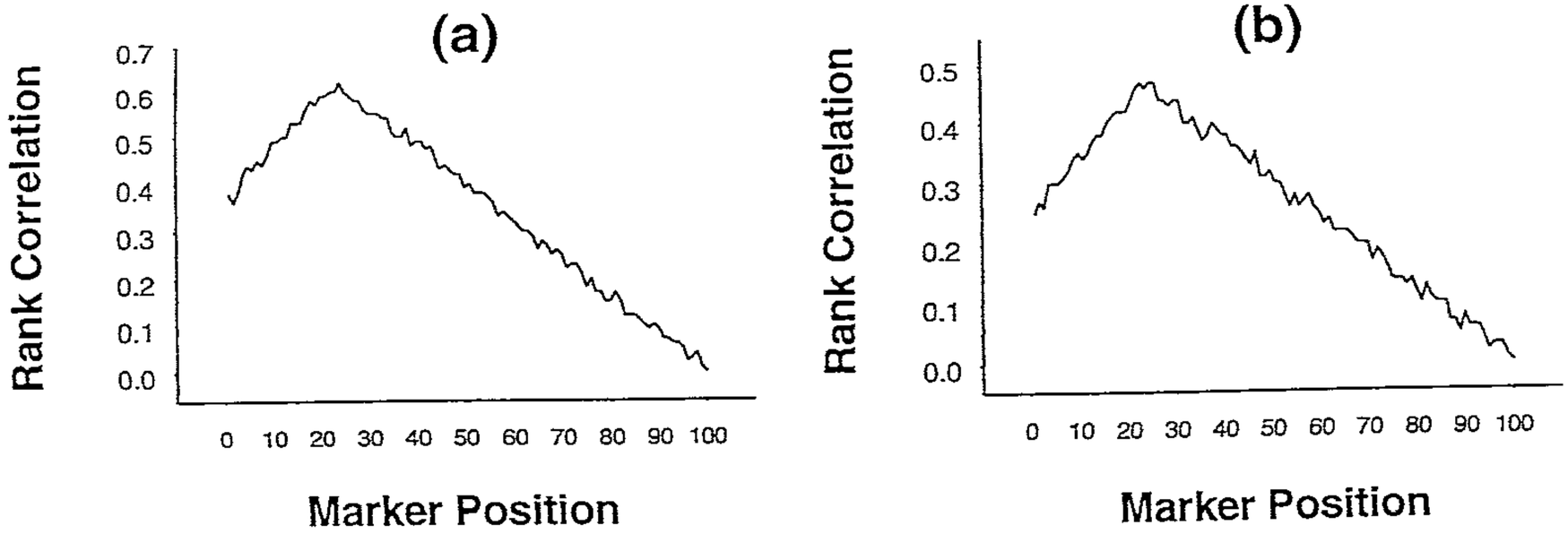


Figure 5.6. Mean rank correlation, based on 1000 replications, between squared difference of trait values of a sib-pair and estimated i.b.d. scores at 100 ordered markers with simulation parameter values $\alpha = 5, \sigma^2 = 1, p = 0.7, \rho = 0.6$ and (a) $\beta = 0$; (b) $\beta = 2$; (c) $\beta = 4$ based on 25 sib-pairs.

Table 5.4. Comparison between the non-parametric and parametric regressions based on average prediction error (residual sums of squares averaged over 1000 replications) in the case of a single QTL using 50 sib-pairs with simulation parameter Values of $\alpha = 5; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$. [4]

(a) $\beta = 0, p = 0.5, \rho = 0.8$

Candidate Interval	Error in Prediction		
	NP (95.3%)	P1 (97.6%)	P2 (98.0%)
(1,2)	111.45	107.56	104.87
(2,3)	103.40	100.48	97.84
(3,4)	112.83	109.58	105.53
(4,5)	122.17	118.97	116.04

(b) $\beta = 2, p = 0.9, \rho = 0.7$

Candidate Interval	Error in Prediction		
	NP (84.7%)	P1 (80.2%)	P2(81.3%)
(1,2)	167.93	170.56	168.01
(2,3)	160.26	165.02	161.36
(3,4)	169.88	172.64	169.90
(4,5)	184.71	191.39	188.55

(c) $\beta = 4, p = 0.7, \rho = 0.5$

Candidate Interval	Error in Prediction		
	NP (70.7%)	P1 (38.8%)	P2(40.7%)
(1,2)	212.68	216.44	214.85
(2,3)	207.79	215.75	210.26
(3,4)	210.92	214.50	213.13
(4,5)	221.36	229.23	226.39

[4] NP = Non-parametric; P1 = Parametric with leave-one-out; P2 = Parametric without leave-one-out, that is, standard parametric regression. Figures in parentheses denote the percentages of correct identification of true interval location.

Table 5.5. Comparison between the non-parametric and parametric regressions based on average prediction error (residual sums of squares averaged over 1000 replications) in the case of a single QTL using 25 sib-pairs with simulation parameter values of $\alpha = 5; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$. [5]

(a) $\beta = 0, p = 0.5, \rho = 0.8$

Candidate Interval	Error in Prediction		
	NP (93.1%)	P1 (95.4%)	P2 (96.2%)
(1,2)	126.04	122.76	119.64
(2,3)	118.48	115.57	112.05
(3,4)	128.16	121.35	120.03
(4,5)	143.74	137.43	134.68

(b) $\beta = 2, p = 0.9, \rho = 0.7$

Candidate Interval	Error in Prediction		
	NP (82.6%)	P1 (75.3%)	P2 (77.0%)
(1,2)	171.28	176.55	174.16
(2,3)	164.09	171.63	167.32
(3,4)	173.37	178.80	175.09
(4,5)	188.48	198.06	195.45

(c) $\beta = 4, p = 0.7, \rho = 0.5$

Candidate Interval	Error in Prediction		
	NP (65.5%)	P1 (34.2%)	P2 (35.9%)
(1,2)	229.53	240.08	237.62
(2,3)	220.49	238.16	233.44
(3,4)	226.86	237.61	233.38
(4,5)	243.35	258.77	255.26

[5] NP = Non-parametric; P1 = Parametric with leave-one-out; P2 = Parametric without leave-one-out, that is, standard parametric regression. Figures in parentheses denote the percentages of correct identification of true interval location.

to preserve the original mean vector and dispersion matrix of $(y_{i1}, y_{i2})s$ (i.e., $\alpha = 5, \sigma^2 = 1, \rho = 0.7$ and $\beta = 0, 2, 4$), suitable shifts in location are made. We consider two different perturbations with different intensities. In the first case, we consider a mixture of 80 % of the original bivariate normal distribution and 20 % exponential distribution with mean 1. In the second case, the mixture comprise 50 % of each of the above distribution. The other parameters (i.e., recombination fractions) remaining same, we perform both the non-parametric and parametric regressions to identify the most likely position of the QTL. The results on percentages of correct identification of flanking interval are given in Table 5.6. When these percentages are compared with those presented in Table 5.1, we find that perturbation added to the normal distribution has very marginal effect on the ability to correctly identify the QTL interval location, even when the amount of perturbation is as high as 50 %. As was seen in the previous cases, while the non-parametric regression procedure performs almost as well as the parametric regression procedure when there is no dominance; it preforms increasingly more efficiently as the dominance effect increases.

Table 5.6. Comparison between the non-parametric and parametric regressions in the case of a single QTL when the trait distribution is perturbed with exponential distribution with simulation parameter values of $\alpha = 5; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01, \rho = 0.7$. [6]

Degree of Dominance (β)	p	20 % perturbation			50 % perturbation		
		PCI(NP)	PCI(P1)	PCI(P2)	PCI(NP)	PCI(P1)	PCI(P2)
0	0.5	95.1	98.3	98.9	94.8	98.1	98.6
2	0.9	91.2	81.7	83.2	88.0	81.3	84.0
4	0.7	73.6	48.5	50.6	71.7	46.4	48.8

[6]NP = Non-parametric; P1 = Parametric with leave-one-out; P2 = Parametric without leave-one-out, that is, standard parametric regression. Figures in parentheses denote the percentages of correct identification of true interval location.

5.8 A Comparison With Olson's (1995a) Estimator In Presence Of Dominance

As we noted in Section 5.5, equation (5.1) is valid only when there is no dominance at the trait locus. Olson (1995a) showed that in presence of dominance, equation (5.1) needs to be modified as:

$$E(y_j|\widehat{\pi}_{j1},\widehat{\pi}_{j2}) = \beta_0+\beta_1\widehat{\pi}_{j1}+\beta_2\widehat{\pi}_{j2}+\beta_3\widehat{\pi}_{j1}\widehat{\pi}_{j2}+ \beta_4 f_{1.}+\beta_5 f_{.1}+\beta_6 f_{11}$$

where, β_0, β_1, β_2, β_3, β_4, β_5 and β_6 are some constants and $f_{11} = P\{\pi_1 = 1/2, \pi_2 = 1/2|\widehat{\pi}_1,\widehat{\pi}_2\}$; $f_{1.} = \sum_{i=0}^{2} f_{1i}$ and $f_{.1} = \sum_{i=0}^{2} f_{i1}$. Since this equation provides the exact expectation, while equation (5.1) provides only an approximate expectation, in the presence of dominance, it is of interest to compare our non-parametric strategy with the parametric linear regression based on the above equation. If we compare the two procedures using a parametric test based on prediction errors, we would encounter a "degrees of freedom" problem because of unequal number of explanatory variables in the two regression models. However, we can use the percentage of correct identification of true interval location of the QTL as a criterion to compare the two procedures.

Simulated data were generated using the methodology described in Sections 2.2 and 5.6. The only additional point that needs to be mentioned is that f_{ij}s were obtained as the sample proportions of the marker i.b.d. scores conditional on the estimated marker i.b.d. scores. Based on simulated data for different sets of parameter values and the "leave-one-out" technique, we present the results of the present parametric regression in Table 5.7. Comparing this table with Tables 5.1, 5.4 and 5.5, we find that there is a large increase in the percentages of correct interval location of the QTL using the present linear regression compared to the linear regression discussed in Section 5.5. However, the corresponding percentages using our proposed non-parametric procedure are, in general, higher than those using the present parametric regression, particularly with increase in the degree of dominance at the QT. When $\beta = 0$, the parametric strategy performs slightly better, but as β increases, the relative performance of our

proposed procedure becomes increasingly better. These observations are consistent with different sample sizes ($n = 100, 50, 25$). We, however, note that the average prediction error is less in the present parametric regression compared to the non-parametric strategy. We emphasize again that comparisons based on prediction errors may be inappropriate because of a larger number of regressors in the Olson's (1995a) parametric regression model, than in our non-parametric regression model. In sum, the proposed semi-parametric method outperforms Olson's (1995a) parametric method in presence of dominance at the QT.

5.9 Detection of Multiple QTLs

When the trait is controlled by multiple loci, the proposed procedure for detection of a QTL using flanking markers can be easily extended. Suppose the quantitative trait is determined by two biallelic trait loci (A_1, a_1) and (A_2, a_2). Let the marginal expectations of trait values for individuals of genotypes A_1A_1, A_1a_1 and aa be α_1, β_1 and -α_1, respectively, and for individuals of genotypes A_2A_2, A_2a_2 and a_2a_2 be α_2, β_2 and -α_2, respectively. We assume that the conditional expectation of the trait given the genotypes at the two QTLs are additive. Thus, for example, the expected trait value for an individual of genotype $A_1A_1A_2A_2$ is $\alpha_1+\alpha_2$; for an individual of genotype $A_1a_1A_2a_2$ is β_1-α_2, etc. For ease of exposition and simulation, we assume that the unlinked QTLs are actually on different chromosomes. Further, the QTLs are separately assumed to be in linkage equilibrium with a pair of flanking markers. Based on data on trait values of n independent sib-pairs and the estimated i.b.d. scores of two sets of ordered markers on two different chromosomes, our aim is to detect both the QTLs by identifying the closest pair of flanking markers on each chromosome. Using the rank correlation statistic, we can identify the possible pairs of candidate flanking markers on each chromosome and then invoke the parametric or the non-parametric regression procedure to select the most likely intervals where the two QTLs are located.

We perform simulations to assess the performance of the rank correlation statistic when there are two QTLs and compare the performances

Table 5.7. Results of Olson's (1995a) parametric regression in presence of dominance, based on average prediction error (residual sums of squares averaged over 1000 replications) in the case of a single QTL for various sib-pair sizes with simulation parameter values of $\alpha = 5; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$. [7]

(*a*) $\beta = 0, p = 0.5, \rho = 0.8$

Candidate Interval	Error in Prediction		
	n=100 (99.4%)	n=50 (98.7%)	n=25 (97.0%)
(1,2)	89.76	97.36	113.62
(2,3)	66.38	90.43	104.35
(3,4)	91.05	98.20	114.59
(4,5)	100.81	108.94	128.17

(*b*) $\beta = 2, p = 0.9, \rho = 0.7$

Candidate Interval	Error in Prediction		
	n=100 (86.8%)	n=50 (82.7%)	n=25 (80.3%)
(1,2)	142.06	160.68	167.41
(2,3)	135.63	152.33	159.55
(3,4)	144.92	162.79	170.02
(4,5)	159.68	176.04	182.38

(*c*) $\beta = 4, p = 0.7, \rho = 0.5$

Candidate Interval	Error in Prediction		
	n=100 (70.6%)	n=50 (67.2%)	n=25 (65.1%)
(1,2)	170.88	203.45	218.56
(2,3)	157.54	197.16	206.23
(3,4)	166.37	203.95	216.44
(4,5)	188.12	212.73	231.37

[7] n denotes the number of sib-pairs considered and figures in parentheses denote the percentages of correct identification of true interval location.

of the parametric and the non-parametric regression procedures in locating the flanking intervals correctly. In order to study the nature of rank correlations, we generate data on 100 sib-pairs as before. We consider 100 ordered markers on each of the two chromosomes with the recombination fraction between successive markers equal to 0.05. The first QTL is assumed to be located between the 24^{th} and 25^{th} markers on the first chromosome and the second QTL between the 60^{th} and 61^{st} markers on the second chromosome. Two sets of trait parameter values are chosen for generating simulated data. α_1 is chosen to be 5 in both sets, and the other parameters are chosen such that in the *first case* there is no dominance at either QTL and the first QTL explains 80% of the variance in the trait, while, in the *second case* there is dominance effect only at the first QTL and it explains 60% of the variance in the trait. The nature of the rank correlations are presented in Figures 5.7(a)-(d). Though the magnitudes of the rank correlations are, in general, less than in the case of a single QTL, we find that in both cases, peaks are prominent at the 24^{th} marker on the first chromosome and the 60^{th} marker on the second chromosome, thus correctly identifying the approximate positions of the QTLs.

In order to compare between the parametric and the non-parametric regression strategies in the case of two QTLs, we generate data on 5 markers on each of the two chromosomes. The two sets of simulation trait parameter values are chosen as mentioned in the preceding paragraph. The percentages of correct identification of flanking markers on each chromosome are given in Table 5.8. We find that in the first case, where there is no dominance effect at either QTL, the percentage of correct identification for both QTLs is, as expected, slightly higher in the parametric procedure. However, the percentage of correct identification of both QTLs by the non-parametric procedure is as high as 93.2% and, for all practical purposes, is almost as efficient as the parametric procedure. In the second case, where there is dominance at the major QTL, the percentage of correctly locating both the QTLs is substantially higher (88.2%)in the non-parametric procedure. While the parametric procedure locates the second QTL (which has no dominance effect) in about 90% of the simulation replications, the first QTL is located correctly in only 61-73% of the replications. The corresponding figures for the non-parametric procedure are 92% and 87.5% respectively.

Thus, we find that the non-parametric procedure performs more efficiently even when there is dominance in one of the two QTLs. We note that in our simulations, whenever the flanking interval is incorrectly identified, the QTL is identified in an adjacent interval. Thus, the error in identification may not be of any major practical consequence. We also note that for given values of the proportions of trait variance explained by the QTLs, there may be several possible combinations of trait parameter values (αs,βs,ps). An obvious question is whether the performance of the procedures differ for such different combinations of trait parameter values that correspond to the same proportions of trait variance explained by the QTLs. We investigate this problem, and find that different trait parameter values conforming to the same proportion of variance explained by the major QTL yields almost identical results in terms of percentage of correct identification of interval location.

In the above, we have ignored the possibility of epistatic interactions between the two QTLs. Epistatic interactions can be parametrized in a multitude of ways (Kearsey and Pooni 1996). However, to perform some preliminary investigations on the effect of epistatic interactions on the perfomance of our proposed method, we consider the digenic interaction model given in Table 2.1.

Simulated data under this digenic interaction model are generated as described earlier. The results of the first-stage of our procedure are graphically depicted in Figures 5.8(a)-(d), with $\Delta = 1$, $\alpha_1 = 5$ and other sets of parameter values chosen such that the first locus without any dominance effect explains 80% [Figures 5.8(a) and (b)] and with dominance effect explains 60% [Figures 5.8(c) and (d)] of the total variation in Y. It is observed that in the presence of epistatic interaction, the magnitudes of the rank correlations are slightly lower than if epistatic interaction are absent. The peaks are pronounced at the right locations of the QTLs. The results of the second-stage of our procedure are provided in Table 5.9, and show that the qualitative inferences are identical with those in the absence of epistasis, but the percentages of correct interval identification are marginally lower. Thus, it is clear that our proposed procedure performs well at both stages even in the presence of reasonable levels of epistatic interaction between the QTLs.

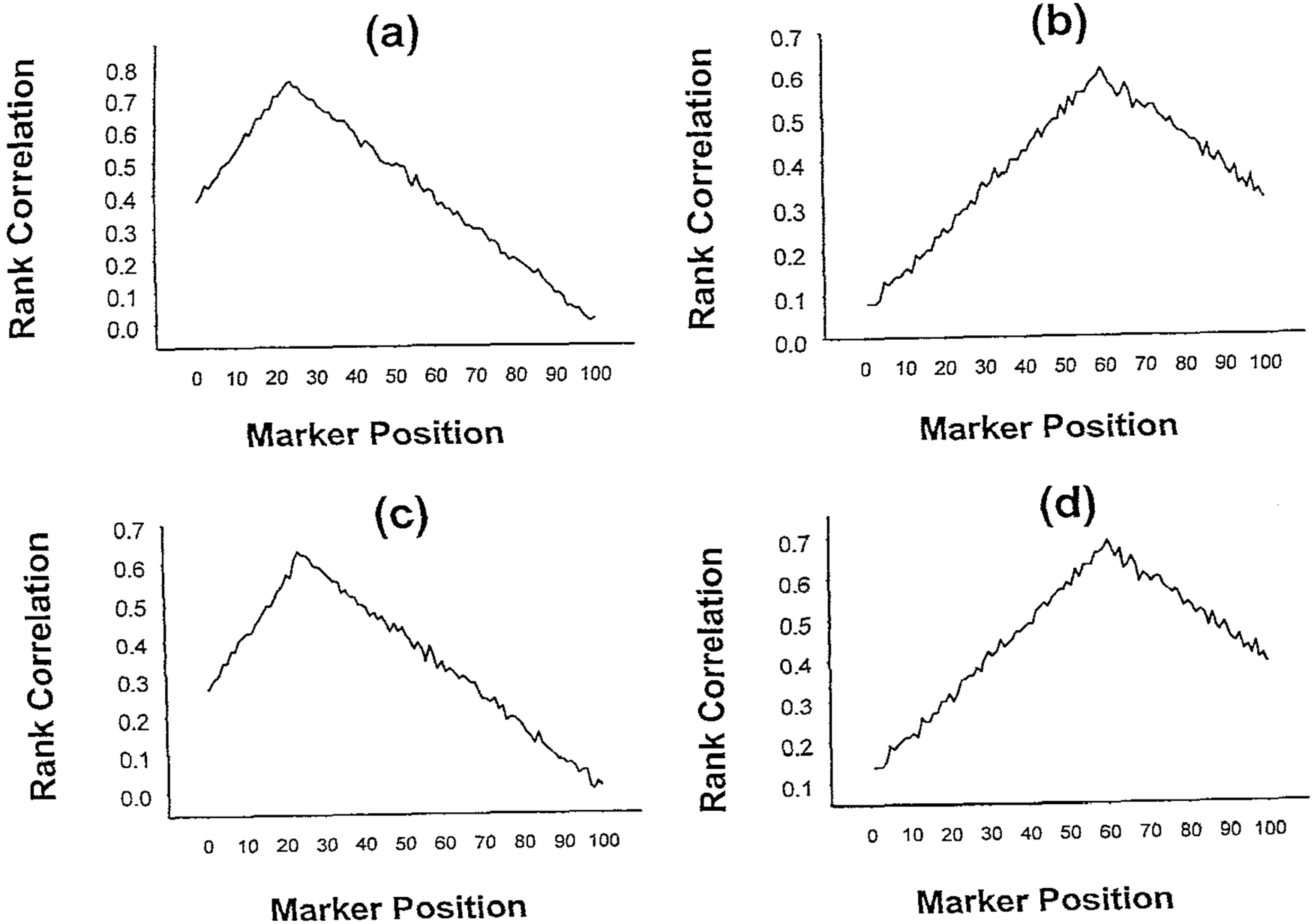


Figure 5.7. Mean rank correlation, based on 1000 replications, between squared difference of trait values of a sib-pair and estimated i.b.d. scores at 100 ordered markers. (a) and (b) pertain to the first and second loci, respectively, when the first locus, without dominance, explains 80 % of the variation in trait values; (c) and (d) pertain to the first and second loci, respectively, when the first locus, with dominance, explains 60 % of the variation in trait values.

Table 5.8. Comparison between the non-parametric and parametric regressions in the case of two QTLs in absence of epistasis using 100 sib-pairs with simulation parameter values of $\alpha = 5; \sigma^2 = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$ and 1000 replications. [8]

(*a*) No dominance effect at either QTL and 80% of trait variance explained by the first QTL.

Type of Identification*	Percentage		
	NP	P1	P2
CC	93.2	96.5	97.4
CI	6.8	3.5	2.6
IC	0	0	0
II	0	0	0

(*b*) Dominance effect at first QTL only and 60% of trait variance explained by the first QTL.

Type of Identification*	Percentage		
	NP	P1	P2
CC	82.2	65.7	69.6
CI	5.3	3.0	3.4
IC	9.8	23.6	21.8
II	2.5	6.7	5.2

5.10 Discussion and Overview

Recent developments in molecular genetics have resulted in the increasing use of genome-wide scans for mapping traits. Genome-wide scans yield huge data sets that require analyses using efficient and robust statistical methods. In this Chapter, we have proposed a semi-parametric strategy of interval mapping of quantitative trait loci. Given trait values of sibs and estimated i.b.d. scores of a set of ordered markers on a chromosome, we have developed a two-step multipoint linkage method. We first reduce our data on

[8] NP = Non-parametric; P1 = Parametric with leave-one-out; P2 = Parametric without leave-one-out, that is, standard parametric regression. C and I denote, respectively, correct and incorrect identification of interval location of the two QTLs (e.g., CI denotes that the first QTL's interval is correctly identified, but the second QTL's interval location is not correctly identified.)

markers to only those which provide indications of linkage (coarse-mapping) to the QTL using rank correlation. Then, we fine-map the QTL location. The two-step approach was prompted by cost-benefit considerations of genotypic data generation and statistical analyses. While the adoption of a set of high-density markers in genome-wide scans may provide maximal information, it is often prohibitively expensive. A statistically and logically more sound, as also cost-effective strategy is to initially use low-density markers (perhaps, 5-10cM apart) and identify a set of probable marker intervals in which the QTL(s) may be located. Then, one can saturate these "probable intervals" with higher-density (say, 1-5cM) markers and localize the QTL(s) to finer intervals. In fact, such a strategy has recently been adopted in a sib-pair linkage study of schizophrenia (Williams et al. 1999). The investigators performed a two-stage genome-wide scan. In the first stage, the average density of markers used was 17.26cM. In the second stage, intervals identified in stage 1 were saturated with markers with an average density of 5-10cM. Our proposed protocol uses a computationally easy, low-stringency, statistical criterion based on rank correlation for analyzing low-density marker data on sib-pairs. For analyzing high-density marker data, that is, for fine-mapping, we have proposed a method that is capable of identifying even small "signals" of linkage evidence, because it does not use assumed functional forms for the nature of dependence between squared difference of sib-pair trait values and estimated i.b.d. scores. In fact, in the presence of dominance effects at the trait loci, which may be the rule rather than the exception, functional forms are difficult to algebraically derive. Further, since local smoothing is performed, the efficiency of detecting evidence of linkage in small marker intervals is higher and variations in values of trait parameters keeping the proportion of trait variance explained by the QTL(s) at the same level. We note that a similar two-stage procedure, in the context of qualitative traits, was proposed by Elston et al. (1996). We have compared our procedure with a currently-used parametric regression procedure (Olson 1995a) and have shown that the efficiency of our procedure in correctly identifying interval locations increases with increase in the degree of dominance at the trait locus. Moreover, in our procedure, the percentage of correct identification of flanking markers is not significantly adversely affected with

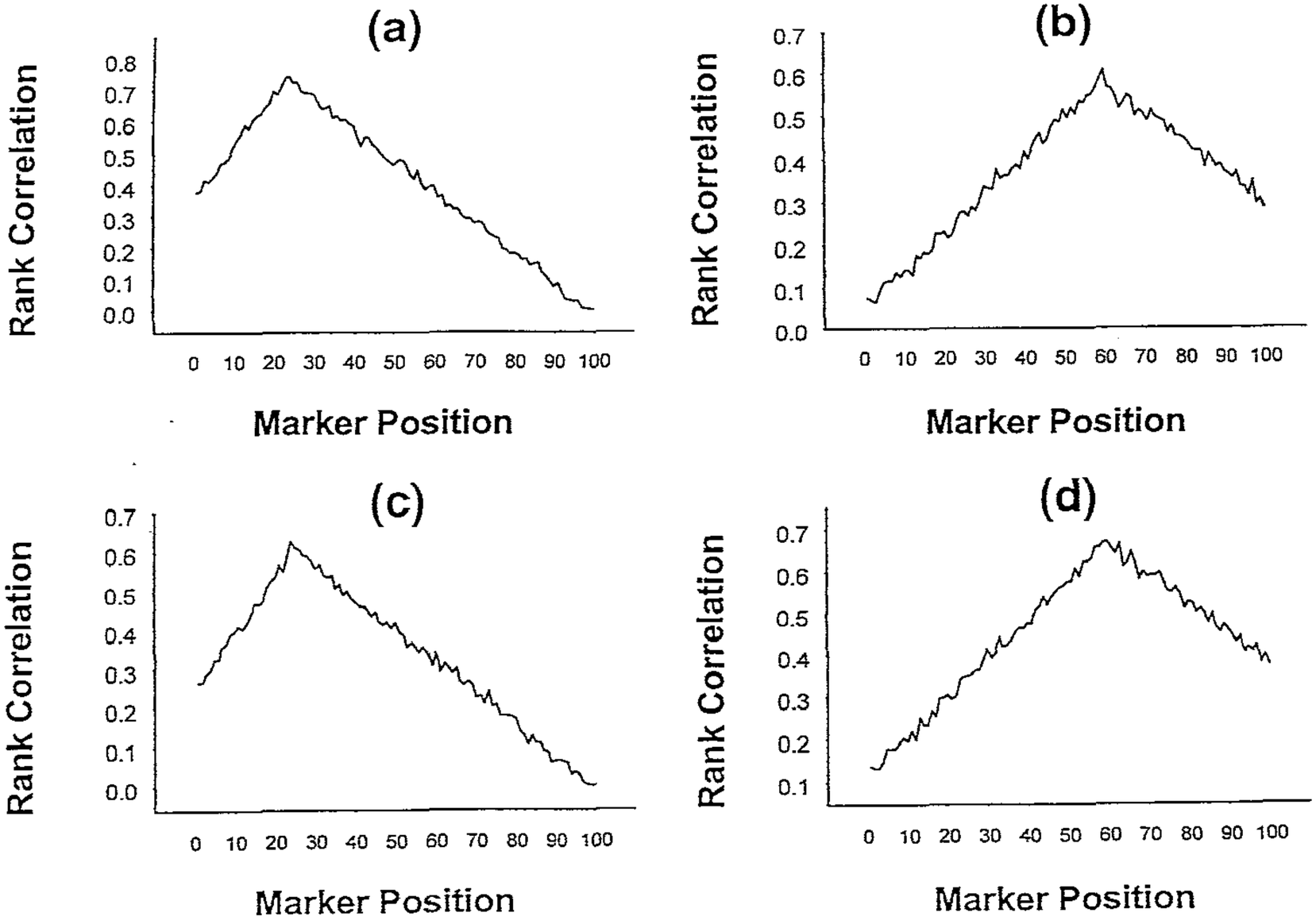


Figure 5.8. Mean rank correlation, based on 1000 replications, between squared difference of trait values of a sib-pair and estimated i.b.d. scores at 100 ordered markers. (a) and (b) pertain to the first and second loci, respectively, when the first locus, without dominance but epistatically interacting with the second locus, explains 80 % of the variation in trait values; (c) and (d) pertain to the first and second loci, respectively, when the first locus, with dominance and epistatically interacting with the second locus, explains 60 % of the variation in trait values.

Table 5.9. Comparison between the non-parametric and parametric regressions in the case of two QTLs in presence of epistasis (Δ) using 100 sib-pairs with simulation parameter values of $\alpha = 5; \sigma^2 = 1; \Delta = 1; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$ and 1000 replications. [9]

(a) No dominance effect at either QTL and 80% of trait variance explained by the first QTL.

Type of	Percentage		
Identification*	NP	P1	P2
CC	91.4	94.5	95.6
CI	8.6	5.5	4.4
IC	0	0	0
II	0	0	0

(b) Dominance effect at first QTL only and 60% of trait variance explained by the first QTL.

Type of	Percentage		
Identification*	NP	P1	P2
CC	78.1	60.8	64.8
CI	6.5	3.3	4.9
IC	12.3	26.4	23.0
II	3.1	9.5	7.3

reasonable reductions in sample sizes. We have also shown that the procedure is robust with respect to distributional assumptions.

We emphasize that if one wishes to perform a one-step genome scan, the data can be analyzed using either our procedure based on rank correlation (computationally cheap) or using non-parametric regression (computationally more expensive). A major advantage of the proposed procedure is that, unlike parametric linkage methods, it does not involve modeling of epistasis and other trait parameters, and hence, is much more robust with respect to distributional assumptions.

[9] NP = Non-parametric; P1 = Parametric with leave-one-out; P2 = Parametric without leave-one-out, that is, standard parametric regression. C and I denote, respectively, correct and incorrect identification of interval location of the two QTLs (e.g., CI denotes that the first QTL's interval is correctly identified, but the second QTL's interval location is not correctly identified.)

Chapter 6

Deciphering the Genetic Architecture of a Multivariate Phenotype

6.1 Introduction and Objective

One of the major current challenges in genetic epidemiology is to unravel genetic architectures of complex traits. Quantitative variables, possibly correlated, generally underlie complex traits. Often, a dichotomous trait definition is adopted for such traits based on cut-off points defined on suitable functions of the underlying quantitative variable(s). Examples are diabetes, hypertension, schizophrenia, etc. Such dichotomization often leads to loss of power in estimating genetic and environmental contributions to such traits, and in mapping the loci controlling such traits. Further, this approach to defining a phenotype may lead to inconsistencies in inferences across studies. Therefore, it is desirable to use the information on the set of underlying multivariate phenotypes. Often, individual components of the multivariate phenotype vector are analyzed separately, both in order to estimate genetic and environmental contributions and for gene-mapping. This approach has many obvious pitfalls, including the statistical problem of multiple comparisons, especially when genome-wide scans are performed.

It has been emphasized that the genetic dissection of complex traits and

diseases may require study designs and statistical methods that are more sophisticated than those used in the analysis of simple Mendelian genetic traits and diseases (Lander and Schork 1994). There is currently a major interest in using data on multivariate phenotypes for genetic epidemiological analysis of complex traits. Methodologies have been developed and there have been attempts to jointly analyze data, of sib pairs or of other sets of family members, on several correlated quantitative phenotypes as a single multivariate phenotype. Many models and approaches have been used, including variance components (Lange and Boehnke 1983, Schork 1993), regressive model (Bonney et al. 1988, Moldin and van Eerdewegh 1995), multivariate extension of the Haseman-Elston model (Amos et al. 1990, Amos and Liang 1996) and structural equations model (Eaves et al. 1996, Todorov et al. 1998). It has been noted that with a large number of components in a multivariate phenotype vector, the power of a multivariate analysis to detect linkage can be substantially lower than the power of an analysis applied to a "genetically relevant" phenotype (Ott and Rabinowitz 1999).

To circumvent the above mentioned problem of power reduction, one approach that has been adopted is the application of data reduction techniques, such as principal components analysis or factor analysis, by which the dimension of the original multivariate phenotype vector is reduced and subsequent analyses are performed on a lower dimensional vector of a few linear combinations of the original phenotypes (Zlotnik et al. 1983, Hasstedt et al. 1994, Boomsma 1996, Allison and Beasley 1998, Ott and Rabinowitz 1999). While this approach may overcome the problem of loss of statistical power to a certain extent, it is important to realize that unless the choice of variables from the vector of variables to be combined as a new quantitative phenotype is made judiciously, using certain statistical and genetic principles, inferences may be grossly incorrect. Hasstedt et al. (1994) and Ott and Rabinowitz (1999) have emphasized that those principal components which have high heritabilities be chosen in the final analysis. Majumder et al. (1998) have emphasized that an initial correlational analysis be performed using individual components of the multivariate phenotype vector, and only that subset of variables which show high correlations within individuals in families be chosen for further data reduction. They have also suggested that only those principal components whose coefficients show consistency across

family members be chosen for final analysis.

Since the multivariate phenotype underlying a complex trait may be controlled by more than one locus, it is unclear whether an initial analysis and examination of the correlation structure of the variables should be carried out to identify subsets of variables, within each of which data reduction may be performed. The purpose of this study is precisely to examine this issue in the context of gene-mapping using data on sib-pairs. The overarching goal of this study is to propose a methodology for analysis of a multivariate phenotype for the purpose of mapping the underlying loci controlling the phenotype.

6.2 Scenarios and Models

We assume that we have a phenotype vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$. A set of genetic loci pleiotropically control individual components, Y_i, of this phenotype vector. Since we assume the existence of pleiotropic effects, the number of loci, l, will necessarily be much smaller than p. We also entertain the possibility that some of the individual components of the phenotype vector may not be under any genetic control, and may be solely determined by environmental effects. Further, even when an individual component is under genetic control, we accommodate the possibility of environmental effects on this component. For purposes of illustration and the simulation studies described subsequently, we consider three simple scenarios; Cases (1) - (3) depicted in Figure 6.1. We consider a multivariate phenotype vector comprising 7 individual components. In Case (1), the component phenotype Y_1 is under the control of an autosomal biallelic locus, which also pleiotropically controls the component phenotypes Y_2 and Y_3. The component phenotype Y_4 is under the control of another autosomal biallelic locus, unlinked to the first locus; Y_5, Y_6 and Y_7 are pleiotropically controlled by this second locus. In Case (2), Y_1, Y_2, Y_3 are controlled similarly as in Case (1), but Y_4, Y_5, Y_6 and Y_7 are not under any genetic influence, but are only influenced by environmental factors. In Case (3), both loci have direct effects on some of the component phenotypes as depicted in Figure 6.1.

The model that we consider for Case (1) is that Y_i, $i = 1, 2, 3$, is

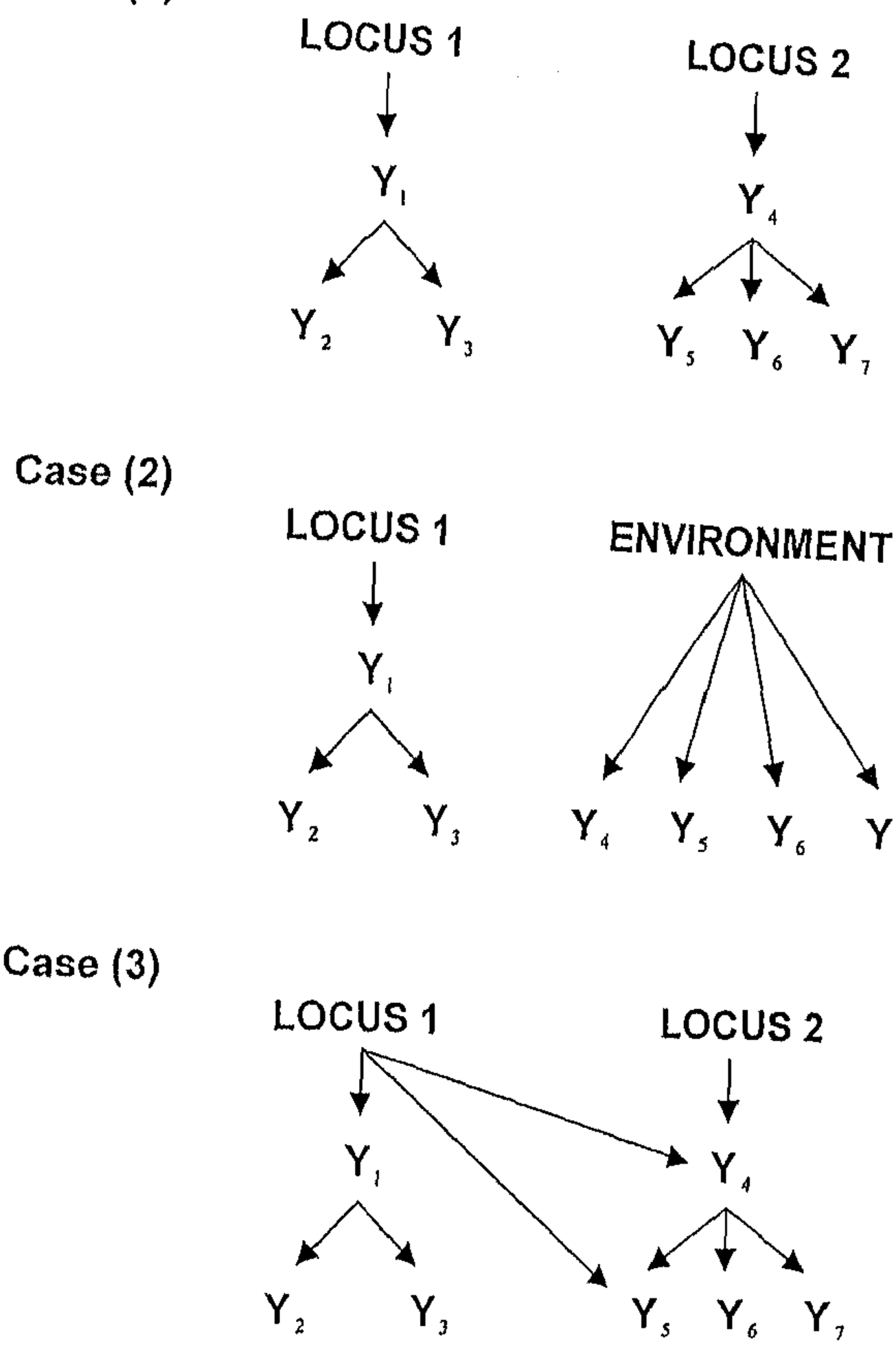


Figure 6.1. Three models of a multivariate phenotype $Y = (Y_1, Y_2, ..., Y_7)$ considered.

distributed as normal with mean $\alpha_i, \beta_i, -\alpha_i$ and variance ${\sigma_i}^2$ according as the genotype at the first locus is A_1A_1, A_1a_1 or a_1a_1, where A_1 and a_1 are the alleles at that locus. Similarly, Y_i, $i = 4, 5, 6, 7$, is distributed as normal with mean $\alpha_i, \beta_i, -\alpha_i$ and variance ${\sigma_i}^2$ according as the genotype at the second locus is A_2A_2, A_2a_2 or a_2a_2, where A_2 and a_2 are the alleles at that locus. In Case (2), the model for Y_i, $i = 1, 2, 3$, is identical to that in Case (1). However, as Y_i, $i = 4, 5, 6, 7$, is influenced only by environmental factors, the underlying distribution is normal with the same mean α_i and variance ${\sigma_i}^2$. The model for Case (3) is identical to that for Case (1), except that Y_4 and Y_5 are also influenced by the first locus (see Figure 6.1), and so their means depend on the genotype at that locus. Additive genotypic effects model is assumed for these two phenotypic components.

Under the three scenarios considered, the 7-dimensional phenotype vector really comprise two subvectors (Y_1,Y_2,Y_3) and (Y_4,Y_5,Y_6,Y_7). Our problem is to decipher the genetic architecture of the 7-dimensional phenotype. That is, under Case (1), we would like to collect appropriate data and analyze the data to be able to map both loci; under Case (2), we should be able to identify and map one locus, etc.

If we denote the expected correlation matrix of $\mathbf{Y}$ as:

$$\rho = \begin{pmatrix} \rho_{11\ 3\times3} & \rho_{12\ 3\times4} \\ & \rho_{22\ 4\times4} \end{pmatrix}$$

then under the models specified above, the elements of the submatrices ρ_{11}, ρ_{12} and ρ_{22}; $i, j = 1, 2$, will follow certain patterns and constraints under the three scenarios. These are:

Case (1) --- The elements of ρ_{12} are all small in magnitude and are less than the elements of ρ_{11} and ρ_{22}. We note that if common environmental or other common small genetic effects are absent, then the elements of ρ_{12} are expected to be all zero.

Case (2)--- As for Case (1) elements of ρ_{12} are expected to be close to zero. Further, the elements of ρ_{22} are also expected to be smaller in magnitude than the elements of ρ_{11}.

Case (3) — Because certain phenotypic components are controlled by more than one locus (Figure 6.1), the correlation matrix of $\mathbf{Y}$ can be further partitioned as explained below and in Table 6.1.

$$\rho = \begin{pmatrix} \rho_{11\ 3\times3} & \rho_{12\ 3\times2} & \rho_{13\ 3\times2} \\ & \rho_{22\ 2\times2} & \rho_{23\ 2\times2} \\ & & \rho_{33\ 2\times2} \end{pmatrix}$$

The elements of ρ_{13} are expected to be close to zero. Further, the elements of ρ_{12}, ρ_{22} and ρ_{23} are expected to be smaller in magnitude than the elements of ρ_{11} and ρ_{33}.

6.3 Methodology

6.3.1 Data reduction

The problem that we seek to examine under these models and expectations is compare the efficiencies of statistically deciphering the genetic architecture of the multivariate phenotype with or without ignoring the expected correlation structure under the three scenarios listed above. We used the principal components approach. Ignoring the correlation structure implies that principal components are extracted using the observations on the 7-dimensional phenotype vector. The first two principal components extracted from these data are denoted as *PC-1* and *PC-2*, respectively. For taking the correlation structure into account, we have, for Cases (1) and (2), extracted principal components based on the observations of the sub-vectors (Y_1,Y_2,Y_3) and (Y_4,Y_5,Y_6,Y_7) separately. For Case (3), principal components based on the observations of the sub-vectors (Y_1,Y_2,Y_3,Y_4,Y_5) and (Y_4,Y_5,Y_6,Y_7) are extracted. We note that two component phenotypes Y_4 and Y_5 are common to both sub-vectors. This is because both loci have effects on Y_4 and Y_5. The first principal components extracted from observations of these sub-vectors are denoted $PC^{(1)}$ and $PC^{(2)}$, respectively. For linkage analysis, we use *PC-1*, *PC-2*, $PC^{(1)}$ and $PC^{(2)}$. We note that while for Cases (1) and (2), the expected correlation between $PC^{(1)}$ and $PC^{(2)}$ is zero, for Case (3) it is positive. However, this poses no problem in linkage analysis as our method does not, at any stage, consider $PC^{(1)}$ and $PC^{(2)}$ jointly.

Table 6.1. Simulation Parameter Values of α_i, β_i and σ_i^2 for the Different Components of the Multivariate Phenotype [1]

Case	Phenotype	α_i	β_i	σ_i^2
1	Y_1	5	2	1
	Y_2	10	3	3
	Y_3	35	10	10
	Y_4	2	0	0.1
	Y_5	20	5	5
	Y_6	30	8	10
	Y_7	100	20	15
2	Y_1	5	2	1
	Y_2	10	3	3
	Y_3	35	10	10
	Y_4	2		0.1
	Y_5	20		5
	Y_6	30		10
	Y_7	100		15
3	Y_1	5	2	1
	Y_2	10	3	3
	Y_3	35	10	10
	Y_4^*	2	0	0.1
		23	4	7
	Y_5^*	20	5	5
		45	12	12
	Y_6	30	8	10
	Y_7	100	20	15

[1]Since Y_4 and Y_5 are controlled by both loci, the two sets of values pertain to the effects of the two unlinked loci of these components. The genotypic effects of the two loci are assumed to be additive.

We note that because of the specific scenarios considered by us, partitioning of the 7-dimensional phenotype vector into sub-vectors occurs naturally. In practice, it will be necessary to try different permutations of the rows and columns of the correlation matrix, examine and perform tests of hypotheses on the structures of the submatrices, to determine the most appropriate partitioning of the phenotype vector. This is not done in the present study because the purpose of this study is to examine the effect of ignoring the correlation structure among the phenotypic variables on the efficiency of gene-mapping.

6.3.2 Mapping QTLs

We use the semi-parametric method of quantitative trait locus (QTL) mapping proposed in Chapter 5. The data comprise observations on the principal components on pairs of siblings. We assume that a genome-wide scan has been performed and that genotype data at the various marker loci are available on these pairs of siblings. As described in Chapter 5, a two-stage variable-stringency strategy is used.

6.3.3 Simulation

Our simulation procedure comprises generation of trait values of siblings under the models considered and also identity-by-descent (i.b.d.) scores based on marker genotype data. This is done in several steps; the number of steps is variable for the different scenarios considered by us.

Case (1)

In the first step of our simulation method, we generate the genotypic mating type of parents at each of the two unlinked trait loci from 6-nomial distributions with the cell probabilities being the probabilities of the different mating types. In the second step, we generate the genotypes of the sib-pair at each of the trait loci from trinomial distributions with cell probabilities being the conditional probabilities of the different trait genotypes given parental genotype information. In the third step, we generate, for both sibs, the three phenotypic values controlled by the first trait locus from a trivariate normal distribution with appropriate mean vector and dispersion matrix as described in Section 6.2. In the fourth step, we generate, for both sibs, the four phenotypic values controlled by the second trait locus from a 4-variate normal distribution with suitable mean vector and dispersion matrix described earlier. The method of generation of estimated i.b.d. scores at different marker loci has been described in Section 2.2.

Case (2)

In the first step of our simulation method, we generate the genotypic mating type of parents at the trait locus from a 6-nomial distribution with the cell probabilities being those of the different mating types. In the second step, we generate the genotypes of the sib-pair at the trait locus from a trinomial distribution with cell probabilities being the conditional probabilities of the different trait genotypes given parental genotype information. In the third step, we generate, for both sibs, the three phenotypic values controlled by the trait locus from a trivariate normal distribution with appropriate mean vector and dispersion matrix as described in Section 6.2. In the fourth step, we generate, for both sibs, the four phenotypic values that are environmental in nature from a 8-variate normal distribution with mean vector and dispersion matrix described earlier. The method of generation of i.b.d. scores at different marker loci is identical to *Case 1*

Case (3)

The method of generation of the genotypes of the sib-pairs at the two trait loci conditional on the parental genotypes at the two loci is identical to *Case 1*. Next, we generate, for both sibs, the seven phenotypic values sequentially (the three phenotypes controlled solely by the first trait locus, then the two phenotypes controlled jointly by the first and the second trait loci and finally the two phenotypes controlled solely by the second trait locus) from a 7-variate normal distribution with appropriate mean vector and dispersion matrix described in Section 6.2. (The mean of each phenotypic value jointly controlled by the two trait loci is assumed to be the sum of the marginal means of the phenotypic value at each trait locus.) The method of generation of i.b.d. scores at different marker loci is identical to *Case 1*.

The parameter values used in the simulations are presented in Table 6.1. All results are based on 100 simulation runs for each case. For linkage analysis, each simulation run is based on 100 sib pairs.

6.4 Results

As mentioned earlier, after generating data on the multivariate phenotype vector, we extract the first principal component from the data on each of the two sub-vectors (Y_1,Y_2,Y_3) and (Y_4,Y_5,Y_6,Y_7); these are denoted $PC^{(1)}$ and $PC^{(2)}$, respectively. We also extract the first two principal components, *PC-1* and *PC-2*, from data on the complete 7-dimensional phenotype vector.

We present in Tables 6.2,6.3 and 6.4, the coefficients of the various principal

components for the two siblings, and the percentages of variance explained (PVE) by the principal components, for 5 simulation runs, for each of the three Cases considered. Consistency in the signs and magnitudes of the coefficients of the various principal components across simulation runs for each Case is obvious. Both $PC^{(1)}$ and $PC^{(2)}$ explain between 85% and 90% of the variance under Case (1). For this Case, *PC-1* and *PC-2*, on the other hand, each explain about 40%-55% of the variance. As is intuitively expected, for Case (2), while the percentages of variance explained by *PC-1* and *PC-2* remain roughly the same as for Case (1), there is a large decrease in this percentage for $PC^{(2)}$ and a large increase for $PC^{(1)}$. For Case (3), the percentages of variance explained by $PC^{(1)}$ *and* $PC^{(2)}$ are about the same as in Case (1), while those explained by $PC-1$ and $PC-2$ are about 5% higher and lower, respectively, than in Case (1).

For the Cases (1) and (3), correlations of values of principal components between siblings are similar whether or not the correlation structure of the phenotypic variables is taken into account before data reduction (Table 6.5). However, for the Case (2), *in view of the fact that the phenotypic components* (Y_4,Y_5,Y_6,Y_7) are not influenced by any genetic factor, the sib-sib correlation for $PC^{(2)}$ is very small (Table 6.5). Thus, sib-sib correlation values do not provide much clue about the underlying genetic architecture of the multivariate phenotype, *except to help identify* those phenotypic components that may not be under any major genetic control.

In order to assess the performance of the rank correlation statistic to identify the interval location of the QTL, we generate data on two unlinked sets of 100 ordered, equispaced markers, such that the recombination fraction between any two consecutive markers is 0.05. Simulated data are generated assuming that the two trait loci are unlinked. For ease of presentation, we shall assume that they are on two separate chromosomes and arbitrarily call them as Chromosomes 1 and 2. Each chromosome, as mentioned earlier, is saturated by a set of 100 markers. We arbitrarily assume that first trait locus is flanked by the 24^{th} and 25^{th} markers on Chromosome 1, and the second trait locus is flanked by the 60^{th} and 61^{st} markers on Chromosome 2. We further assume that the recombination fraction between the first trait locus and the 24^{th} marker of the first set is 0.02, while that between the second trait locus and the 60^{th} marker of the second set is 0.03. The nature of the absolute rank correlation between the different markers and the squared difference in the selected principal components of the sib-pairs mentioned above are presented in Figures 6.2 and 6.3 for Case (1); Figure 6.4 for Case (2); and Figures 6.5 and 6.6 *for Case* (3).

Table 6.2. Coefficients of various principal components and percentages of variance explained by the components for a pair of siblings for 5 independent simulation runs for Case (1)

Type of PC	Run No.	PVE	Coefficients of Sib 1							PVE	Coefficients of Sib 2						
$PC^{(1)}$	1	0.85	-0.55	-0.58	-0.61					0.87	-0.55	-0.58	-0.60				
	2	0.87	-0.56	-0.58	-0.60					0.86	-0.56	-0.59	-0.59				
	3	0.84	-0.55	-0.59	-0.61					0.88	-0.54	-0.58	-0.61				
	4	0.86	-0.55	-0.58	-0.61					0.85	-0.55	-0.57	-0.62				
	5	0.88	-0.56	-0.58	-0.60					0.85	-0.55	-0.58	-0.61				
$PC^{(2)}$	1	0.85	-0.40	-0.53	-0.53	-0.53				0.9	-0.45	-0.51	-0.52	-0.52			
	2	0.85	-0.42	-0.52	-0.52	-0.53				0.88	-0.43	-0.52	-0.52	-0.53			
	3	0.86	-0.43	-0.50	-0.52	-0.52				0.86	-0.44	-0.51	-0.51	-0.54			
	4	0.85	-0.40	-0.51	-0.53	-0.53				0.87	-0.42	-0.50	-0.50	-0.55			
	5	0.88	-0.41	-0.53	-0.53	-0.53				0.89	-0.45	-0.51	-0.51	-0.52			
$PC-1$	1	0.50	0.18	0.15	0.18	-0.38	-0.51	-0.50	-0.51	0.53	0.14	0.17	0.17	-0.44	-0.49	-0.50	-0.50
	2	0.52	0.16	0.14	0.19	-0.38	-0.48	-0.52	-0.52	0.54	0.16	0.18	0.18	-0.42	-0.48	-0.50	-0.51
	3	0.52	0.17	0.16	0.18	-0.36	-0.50	-0.51	-0.52	0.50	0.12	0.16	0.16	-0.46	-0.49	-0.49	-0.50
	4	0.51	0.17	0.15	0.19	-0.40	-0.49	-0.50	-0.51	0.51	0.16	0.16	0.17	-0.41	-0.51	-0.51	-0.51
	5	0.52	0.16	0.17	0.20	-0.37	-0.51	-0.51	-0.51	0.54	0.13	0.15	0.15	-0.43	-0.50	-0.51	-0.52
$PC-2$	1	0.35	-0.52	-0.56	-0.58	-0.14	-0.14	-0.15	-0.15	0.37	-0.54	-0.55	-0.57	-0.1	-0.18	-0.14	-0.15
	2	0.38	-0.50	-0.57	-0.58	-0.11	-0.13	-0.17	-0.17	0.38	-0.51	-0.54	-0.55	-0.12	-0.20	-0.15	-0.15
	3	0.37	-0.54	-0.56	-0.57	-0.12	-0.12	-0.18	-0.18	0.35	-0.53	-0.57	-0.59	-0.08	-0.16	-0.14	-0.16
	4	0.35	-0.52	-0.55	-0.59	-0.14	-0.15	-0.15	-0.15	0.38	-0.50	-0.55	-0.58	-0.14	-0.17	-0.13	-0.14
	5	0.36	-0.53	-0.55	-0.58	-0.13	-0.14	-0.16	-0.16	0.37	-0.54	-0.54	-0.56	-0.11	-0.19	-0.15	-0.15

Table 6.3. Coefficients of various principal components and percentages of variance explained by the components for a pair of siblings for 5 independent simulation runs for Case (2)

Type of PC	Run No.	PVE	Coefficients of Sib 1							PVE	Coefficients of Sib 2						
$PC^{(1)}$	1	0.90	-0.58	-0.60	-0.55					0.88	-0.58	-0.60	-0.55				
	2	0.91	-0.56	-0.61	-0.56					0.92	-0.58	-0.59	-0.57				
	3	0.90	-0.57	-0.59	-0.54					0.90	-0.58	-0.58	-0.56				
	4	0.88	-0.56	-0.60	-0.56					0.87	-0.58	-0.61	-0.56				
	5	0.88	-0.58	-0.59	-0.55					0.89	-0.58	-0.59	-0.56				
$PC^{(2)}$	1	0.70	-0.30	-0.51	-0.55	-0.59				0.75	-0.1	-0.57	-0.58	-0.58			
	2	0.71	-0.32	-0.50	-0.57	-0.58				0.73	-0.12	-0.56	-0.57	-0.58			
	3	0.73	-0.31	-0.51	-0.56	-0.58				0.69	-0.09	-0.55	-0.58	-0.60			
	4	0.68	-0.30	-0.50	-0.56	-0.59				0.71	-0.1	-0.57	-0.57	-0.58			
	5	0.70	-0.32	-0.50	-0.55	-0.58				0.74	-0.11	-0.58	-0.58	-0.59			
$PC-1$	1	0.47	-0.38	-0.40	-0.40	-0.19	-0.40	-0.40	-0.43	0.46	0.30	0.31	0.28	-0.10	-0.49	-0.49	-0.49
	2	0.45	-0.36	-0.40	-0.41	-0.18	-0.42	-0.38	-0.42	0.48	0.33	0.32	0.29	-0.12	-0.47	-0.48	-0.49
	3	0.50	-0.38	-0.39	-0.40	-0.17	-0.37	-0.40	-0.45	0.46	0.30	0.30	0.26	-0.14	-0.47	-0.47	-0.48
	4	0.48	-0.36	-0.38	-0.43	-0.17	-0.38	-0.40	-0.42	0.50	0.32	0.33	0.27	-0.08	-0.49	-0.49	-0.50
	5	0.47	-0.39	-0.40	-0.42	-0.16	-0.36	-0.43	-0.44	0.47	0.30	0.30	0.28	-0.12	-0.48	-0.49	-0.49
$PC-2$	1	0.32	-0.45	-0.44	-0.37	0.26	0.31	0.38	0.39	0.34	-0.50	-0.51	-0.47	0.02	-0.30	-0.30	-0.30
	2	0.34	-0.42	-0.44	-0.38	0.24	0.30	0.38	0.41	0.36	-0.52	-0.53	-0.45	0.03	-0.31	-0.31	-0.28
	3	0.33	-0.43	-0.46	-0.35	0.27	0.30	0.37	0.39	0.35	-0.50	-0.52	-0.46	0.01	-0.33	-0.32	-0.30
	4	0.32	-0.44	-0.44	-0.38	0.28	0.30	0.38	0.38	0.32	-0.50	-0.52	-0.46	0.02	-0.32	-0.32	-0.30
	5	0.36	-0.44	-0.45	-0.35	0.28	0.31	0.39	0.39	0.34	-0.51	-0.51	-0.47	0.01	-0.33	-0.31	-0.28

Table 6.4. Coefficients of various principal components and percentages of variance explained by the components for a pair of siblings for 5 independent simulation runs for Case (3)

Type of PC	Run No.	PVE	Coefficients of Sib 1							PVE	Coefficients of Sib 2						
$PC^{(1)}$	1	0.88	-0.60	-0.60	-0.39	-0.21	-0.34			0.88	0.55	0.55	-0.36	-0.19	-0.39		
	2	0.86	-0.60	-0.61	-0.37	-0.22	-0.36			0.87	0.55	0.58	-0.38	-0.22	-0.42		
	3	0.87	-0.59	-0.60	-0.39	-0.22	-0.32			0.85	0.56	0.56	-0.34	-0.21	-0.40		
	4	0.90	-0.61	-0.61	-0.37	-0.23	-0.33			0.88	0.55	0.57	-0.35	-0.22	-0.38		
	5	0.87	-0.59	-0.61	-0.38	-0.22	-0.34			0.86	0.55	0.56	-0.37	-0.19	-0.39		
$PC^{(2)}$	1	0.96	-0.50	-0.49	-0.50	-0.50				0.94	-0.51	-0.48	-0.50	-0.51			
	2	0.96	-0.50	-0.48	-0.50	-0.51				0.95	-0.51	-0.48	-0.50	-0.51			
	3	0.95	-0.50	-0.50	-0.50	-0.50				0.96	-0.51	-0.49	-0.50	-0.50			
	4	0.95	-0.49	-0.49	-0.50	-0.52				0.95	-0.50	-0.50	-0.50	-0.50			
	5	0.96	-0.50	-0.49	-0.50	-0.51				0.94	-0.50	-0.49	-0.50	-0.51			
$PC-1$	1	0.55	0.04	0.02	0.01	-0.51	-0.48	-0.50	-0.51	0.58	0.24	0.24	0.03	-0.48	-0.43	-0.48	-0.48
	2	0.52	0.04	0.03	0.02	-0.52	-0.48	-0.51	-0.51	0.56	0.24	0.25	0.03	-0.46	-0.45	-0.48	-0.49
	3	0.54	0.04	0.03	0.01	-0.52	-0.47	-0.50	-0.50	0.56	0.23	0.23	0.03	-0.48	-0.45	-0.47	-0.50
	4	0.55	0.03	0.03	0.03	-0.51	-0.49	-0.50	-0.51	0.57	0.26	0.25	0.04	-0.47	-0.44	-0.48	-0.49
	5	0.55	0.04	0.02	0.02	-0.51	-0.48	-0.50	-0.51	0.58	0.25	0.24	0.03	-0.47	-0.46	-0.49	-0.49
$PC-2$	1	0.31	-0.65	-0.63	-0.39	-0.01	-0.14	0.04	0.01	0.27	0.59	0.56	0.44	0.15	0.31	0.08	0.09
	2	0.33	-0.63	-0.62	-0.37	-0.01	-0.16	0.06	0.02	0.28	0.60	0.57	0.44	0.17	0.28	0.06	0.11
	3	0.32	-0.65	-0.64	-0.38	-0.02	-0.15	0.04	0.01	0.30	0.59	0.57	0.45	0.16	0.30	0.08	0.10
	4	0.32	-0.64	-0.63	-0.37	-0.01	-0.14	0.05	0.01	0.27	0.59	0.58	0.44	0.13	0.30	0.07	0.09
	5	0.31	-0.64	-0.62	-0.36	-0.01	-0.16	0.05	0.02	0.28	0.60	0.55	0.43	0.15	0.31	0.06	0.10

Table 6.5. Means and standard deviations of correlation coefficients between values of various principal components for a Pair of Siblings under the three scenarios considered

Case 1

Type of PC	Sib-Sib Correlation	
	Mean	s.d.
$PC^{(1)}$	0.53	0.03
$PC^{(2)}$	0.55	0.05
$PC-1$	0.55	0.08
$PC-2$	0.54	0.08

Case 2

Type of PC	Sib-Sib Correlation	
	Mean	s.d.
$PC^{(1)}$	0.55	0.04
$PC^{(2)}$	0.12	0.02
$PC-1$	0.46	0.07
$PC-2$	0.44	0.08

Case 3

Type of PC	Sib-Sib Correlation	
	Mean	s.d.
$PC^{(1)}$	0.61	0.05
$PC^{(2)}$	0.58	0.05
$PC-1$	0.58	0.07
$PC-2$	0.57	0.07

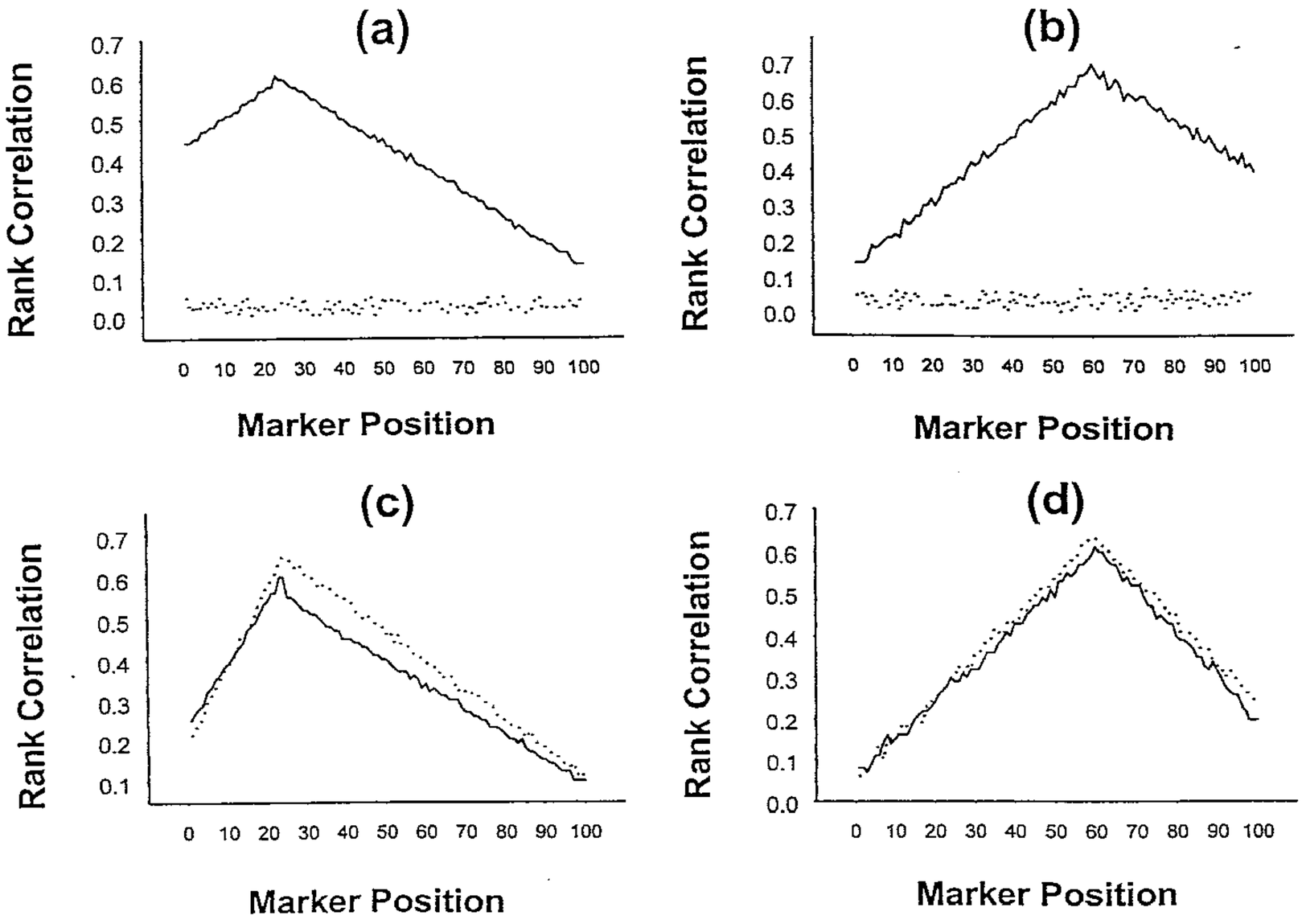


Figure 6.2. Rank correlations between squared difference of values of principal components and i.b.d. scores estimated from genotype data at various ordered marker loci along a chromosome based on a sample of 100 sib-pairs for Case (1).

(a) along markers on Chromosome 1; solid and dotted lines corresopond to $PC^{(1)}$ and $PC^{(2)}$, respectively

(b) along markers on Chromosome 2; solid and dotted lines correspond to $PC^{(2)}$ and $PC^{(1)}$, respectively

(c) along markers on Chromosome 1; solid and dotted lines correspond to $PC-1$ and $PC-2$, respectively

(d) along markers on Chromosome 2; solid and dotted lines correspond to $PC-1$ and $PC-2$, respectively.

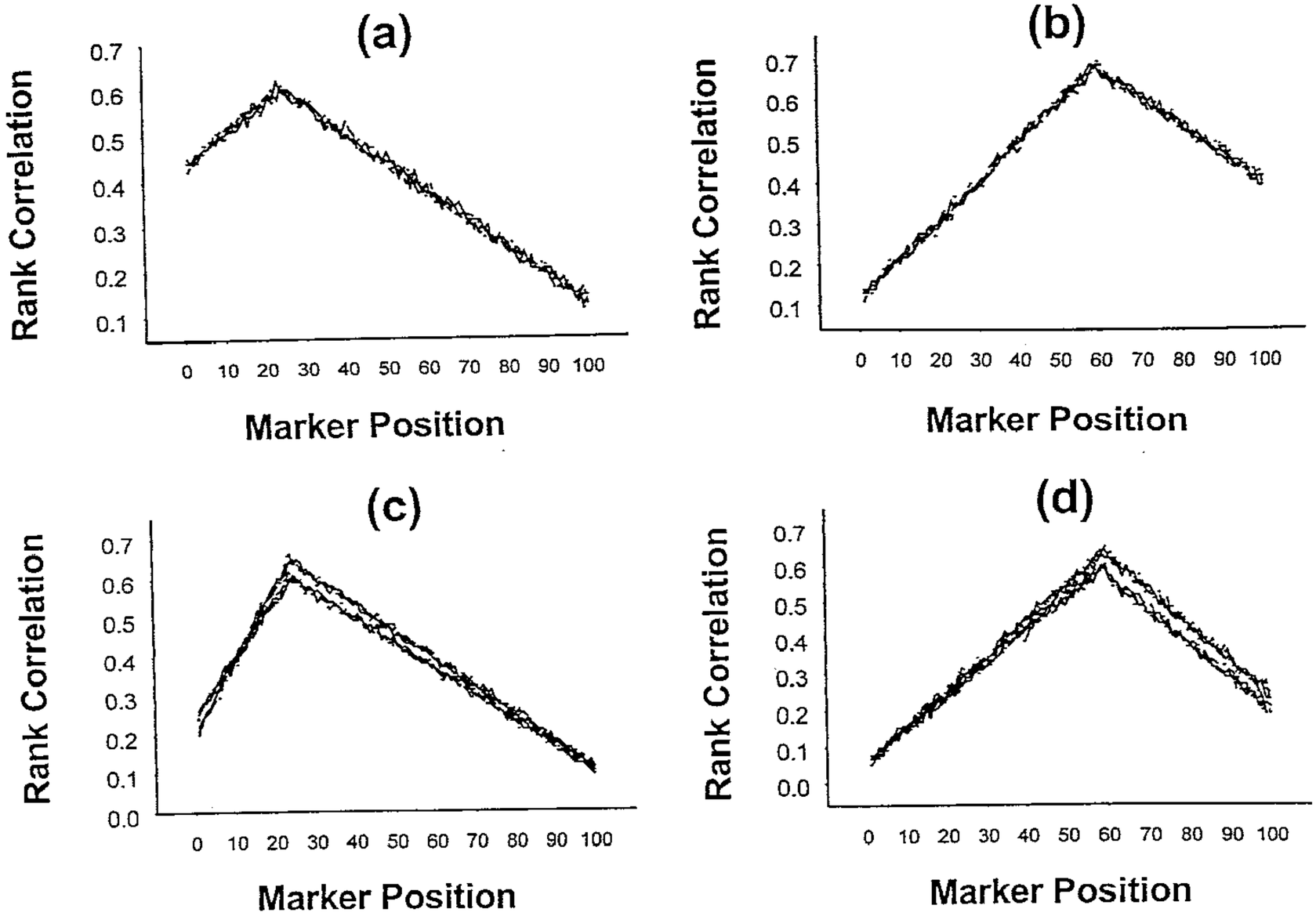


Figure 6.3. Rank correlations between squared difference of values of principal components and i.b.d. scores estimated from genotype data at various ordered marker loci along a Chromosome based on a sample of 100 sib-pairs for Case (1): results of multiple simulation runs

(a) along markers on Chromosome 1 for $PC^{(1)}$

(b) along markers on Chromosome 2 for $PC^{(2)}$

(c) along markers on Chromosome 1 for $PC-1$ and $PC-2$ (the two bands of lines correspond to these two principal components)

(d) along markers on Chromosome 2 for $PC-1$ and $PC-2$ (the two bands of lines correspond to these two principal components).

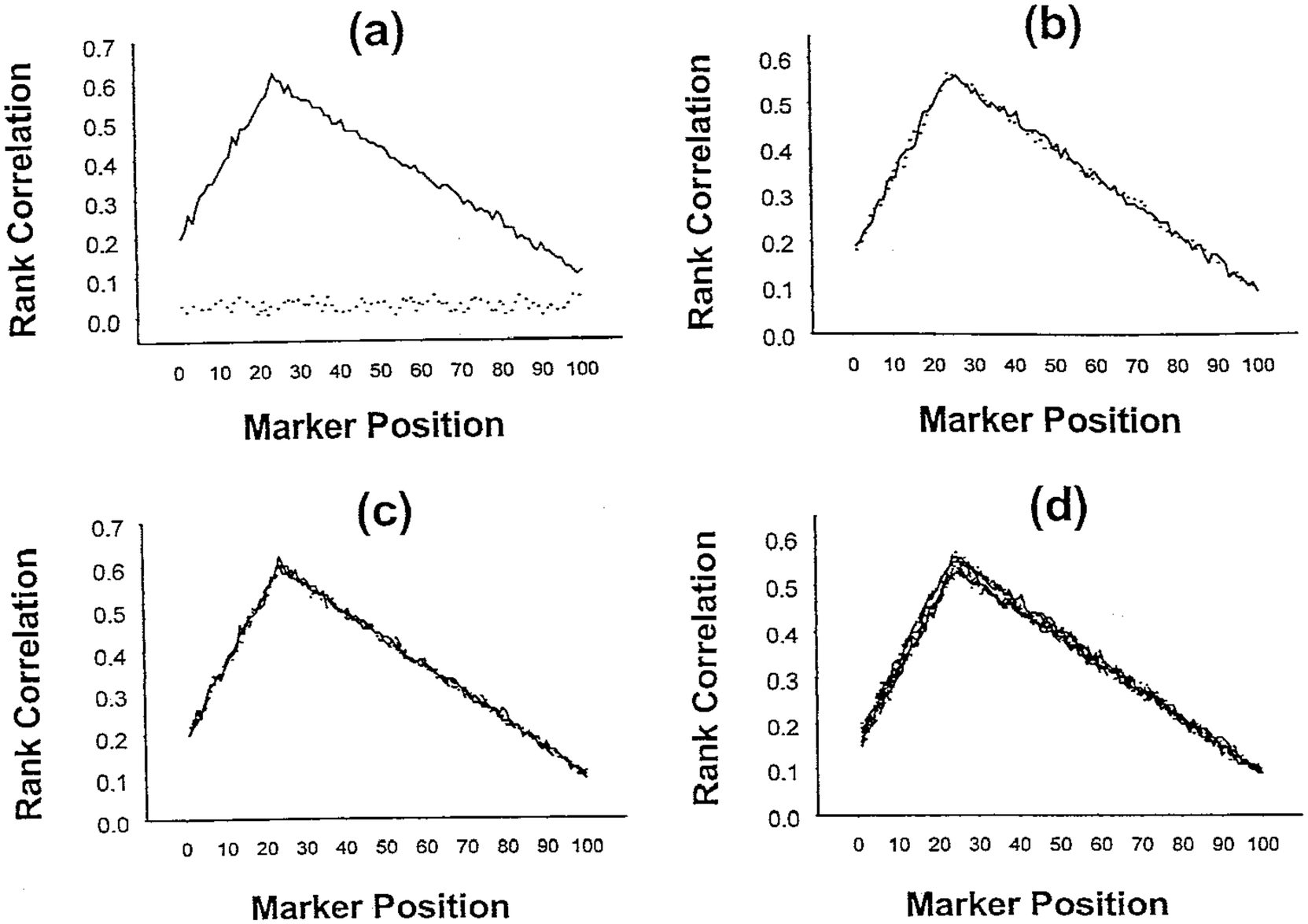


Figure 6.4. Rank correlations between squared difference of values of principal components and i.b.d. scores estimated from genotype data at various ordered marker loci along Chromosome 1 based on a sample of 100 sib-pairs for Case (2)

(a) solid and dotted lines correspond to $PC^{(1)}$ and $PC^{(2)}$, respectively

(b) solid and dotted lines correspond to $PC-1$ and $PC-2$, respectively

(c) results of multiple runs for $PC^{(1)}$

(d) results of multiple runs for $PC-1$ and $PC-2$ (the two bands of lines correspond to these two principal components).

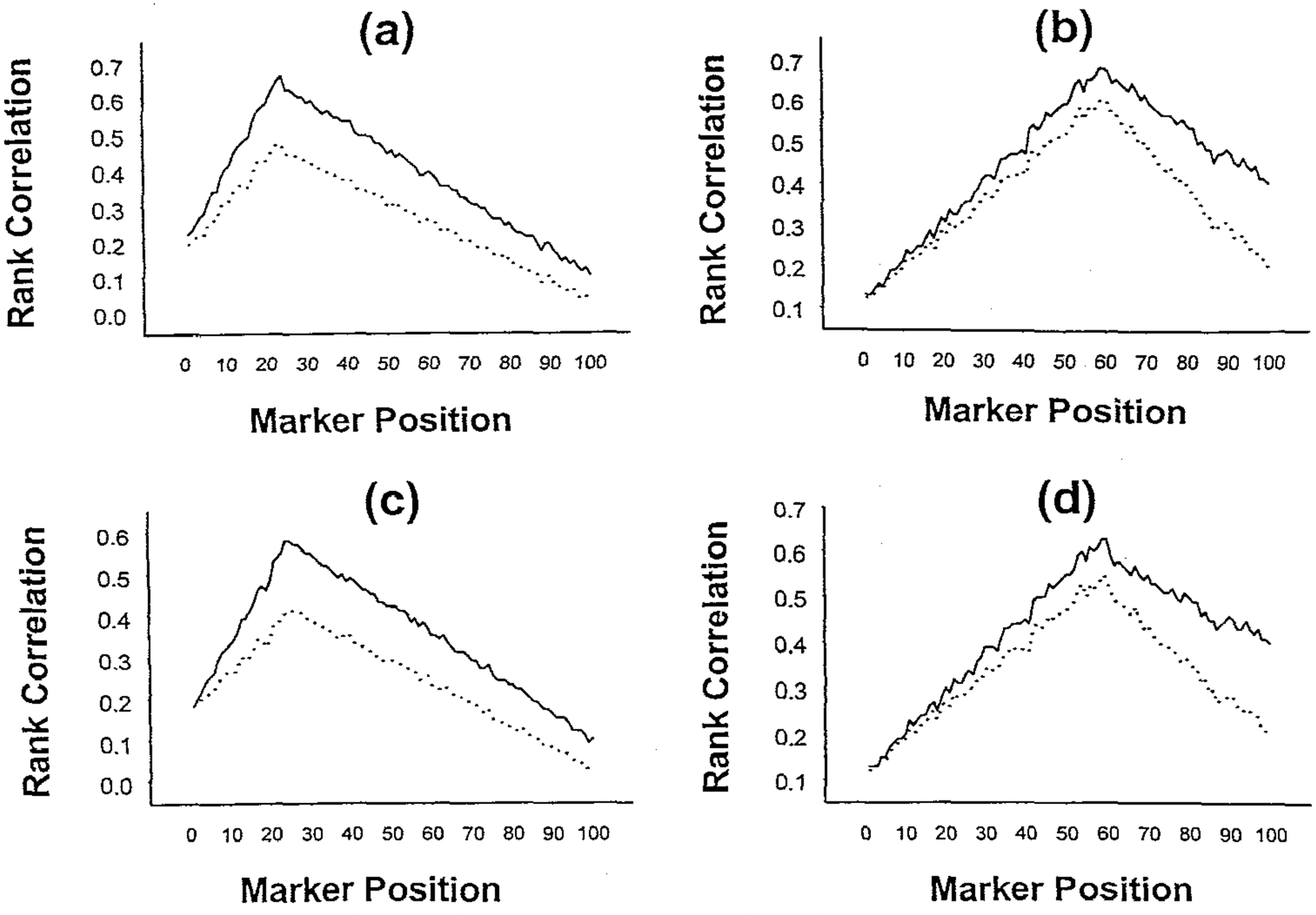


Figure 6.5. Rank correlations between squared difference of values of principal components and i.b.d. scores estimated from genotype data at various ordered marker loci along a chromosome based on a sample of 100 sib-pairs for Case (3).

(a) along markers on Chromosome 1; solid and dotted lines correspond to $PC^{(1)}$ and $PC^{(2)}$, respectively

(b) along markers on Chromosome 2; solid and dotted lines correspond to $PC^{(2)}$ and $PC^{(1)}$, respectively

(c) along markers on Chromosome 1; solid and dotted lines correspond to $PC-1$ and $PC-2$, respectively

(d) along markers on Chromosome 2; solid and dotted lines correspond to $PC-1$ and $PC-2$, respectively.

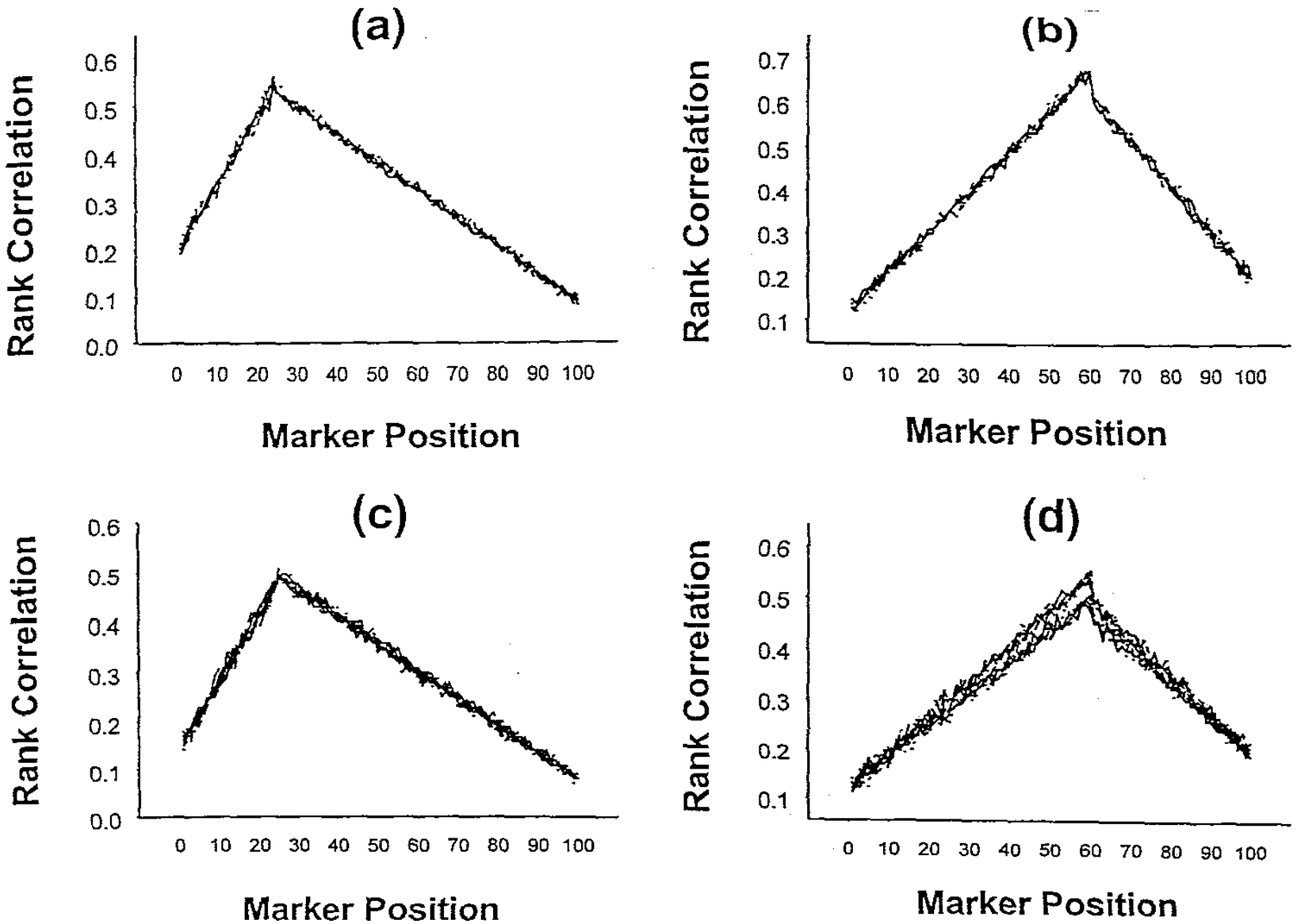


Figure 6.6. Rank correlations between squared difference of values of principal components and i.b.d. scores estimated from genotype data at various ordered marker loci along a Chromosome based on a sample of 100 sib-pairs for Case (3): results of multiple simulation runs

(a) along markers on Chromosome 1 for $PC^{(1)}$

(b) along markers on Chromosome 2 for $PC^{(2)}$

(c) along markers on Chromosome 1 for $PC-1$ and $PC-2$ (the two bands of lines corresponding to these two principal components are completely overlapping and, therefore, not distinguishable)

(d) along markers on Chromosome 2 for $PC-1$ and $PC-2$ (the two bands of lines correspond to these two principal components).

From Figure 6.2(a), we find that the absolute rank correlation increases with the proximity of the considered marker to the first trait locus when $PC^{(1)}$ is considered. Moreover, the peaks are at the 24^{th} marker on Chromosome 1, correctly indicating the approximate location of the first trait locus. However, the absolute rank correlations are uniformly very low for all the 100 markers when $PC^{(2)}$ is considered. This is expected as $PC^{(2)}$ is a function of Y_4, Y_5, Y_6 and Y_7, which are not controlled by the first trait locus on Chromosome 1. Similarly, we find from Figure 6.2(b) that the absolute rank correlation increases with the proximity of the considered marker to the second trait locus on Chromosome 2 when $PC^{(2)}$ is considered. The peaks are correctly detected at the 60^{th} marker. The absolute rank correlations are also, as desirable, uniformly very low for all the 100 markers when $PC^{(1)}$ is considered. Each of $PC-1$ and $PC-2$, which are the first two principal components constructed on the basis of all the 7 component phenotypes, are, on the other hand, expected to detect both loci and they indeed do so as evident from Figures 6.2(c) and 6.2(d). Thus, under the scenario that all component variables are genetically controlled, ignoring the correlation structure ofthe multivariate phenotype has no major effect on the ability of correctly identify the coarse (5 cM) marker interval in which the QTLs are located. In fact, as the graphs in Figures 6.3(a)-(d) show, there is very little variation across simulation runs, which indicates the high efficiency of the method.

Having identified the interval in which the QTL may be located, in practice one saturates this interval with more dense markers (say, at 1 cM density) to obtain a finer location of the QTL. To simulate this practice, we consider data on multiple markers that are more densely located within the coarse interval identified at the previous stage. In our simulations, we generate data on a set of 5 ordered markers. We use the following notations:

θ_2, θ_3 =recombination fraction between the trait locus and the nearest flanking markers 2 and 3, respectively.

θ_1=recombination fraction between markers 1 and 2.

θ_4= recombination fraction between markers 3 and 4

θ_5= recombination fraction between markers 4 and 5.

We use simulation parameter values of $M = 5; \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0.01$. For each set of parameter values, we perform 100 replications; each

replication is based on data on 100 sib pairs. Fine-mapping is performed using the non-parametric kernel-smoothing regression technique, described in Section 5.4. The results are given in Table 6.6.

It is seen from Table 6.6 that for the first scenario, while $PC^{(1)}$ identifies the correct fine-interval of the first locus in 86% of the runs, the corresponding percentages for $PC-1$ and $PC-2$ are 43% and 51% respectively. As expected, $PC^{(2)}$ does not detect the correct interval. Similarly, $PC^{(2)}$ locates the second trait locus correctly in 92% of the cases, while the corresponding percentages for $PC-1$ and $PC-2$ are much lower; 51% and 58%, respectively. $PC^{(1)}$ did not at all identify the correct location of the second QTL. Moreover, we find that, whenever $PC^{(1)}$ and $PC^{(2)}$ yield incorrect fine-interval identifications for the first and second QTLs, respectively, they identify the corresponding trait locus in an adjacent interval. However, the *same is not true when* $PC-1$ *and* $PC-2$ *are used.*

The conclusion, therefore, is that while ignoring the correlation structure among component phenotypes results in no serious loss of efficiency in localizing the QTLs in coarse marker intervals when the multivariate phenotype is primarily controlled by major loci with large effects, for the purpose of fine-mapping of the QTLs, there is, however, a serious loss of efficiency.

Under the second scenario, Case (2), in which the multivariate phenotype is partially controlled by QTLs with major effects, we find that the detection of peaks using rank correlation for the first (and only) trait locus gives identical results as Case (1) [Figure 6.4(a)]. $PC^{(2)}$ did not falsely detect any peak on this chromosome as it is solely influenced by environmental components [Figure 6.4(a)]. Both $PC-1$ and $PC-2$ also correctly detected [Figure 6.4(b)] the interval location of the trait locus controlling (Y_1, Y_2, Y_3). Further, variation across simulation runs is small [Figures 6.4(c) and 6.4(d)]. Under this scenario, at the fine-mapping stage, the percentage of correct identification of locus 1 using $PC^{(1)}$ is 83%, while those using $PC-1$ and $PC-2$ are 70% and 60% respectively [Table 6.6, Case (2)]. Since $PC^{(2)}$ is a function of purely environmental components, clearly it is irrelevant for localizing the first locus; therefore, it is not considered. Thus we find that, although ignoring the correlation structure of the multivariate phenotype did not affect the efficiency of localizing the underlying trait locus in a coarse marker interval even when some of the components of the multivariate

phenotype are not under genetic control, there is a serious adverse effect when fine-mapping is performed.

Under the third scenario, Case (3), we find that the results based on rank correlation are identical to Case (1) [Figures 6.5 (a)-(d)], except that $PC^{(2)}$ and $PC^{(1)}$ also detected the peaks at the 24^{th} marker on Chromosome 1 and 60^{th} marker on Chromosome 2, respectively [Figures 6.5(c) and (d)], although the absolute rank correlations are lower than those when the other three principal components are considered. Since Y_4 and Y_5 are controlled by both the trait loci, and $PC^{(1)}$ and $PC^{(2)}$ are functions of Y_4 and Y_5, these principal components locate the trait loci, based on rank correlations, fairly well. Moreover, the magnitudes and trends of the rank correlations over the various marker intervals are consistent across simulation runs [Figures 6.6(a)-(d)]. At the fine-mapping stage, we find that the percentages of runs correctly localizing the trait loci are, in general, slightly smaller [Table 6.6, Case (3)] than the corresponding percentages for Case (1). However, $PC^{(1)}$ is also able to detect the correct interval of the second locus in 27% of the cases and $PC^{(2)}$ the first locus in 18% of the runs. As explained earlier, this is due to the fact that Y_4 and Y_5 are controlled by both the trait loci.

6.5 Discussion and Overview

We have investigated the method of statistical treatment of sib-pair data on a multivariate phenotype for deciphering its genetic architecture, specifically for fine localization of the underlying loci, if any. We have considered three scenarios in which (i) disjoint components of the multivariate phenotype are pleiotropically controlled by a set of major loci, (ii) a subset of the components of the multivariate phenotype is pleiotropically controlled by a major locus, but the remaining subset is influenced only by environmental factors, and (iii) the multivariate phenotype is pleiotropically controlled by a set of major loci, some of which can jointly influence some component phenotypes. These three scenarios yield different expected correlation structures of the multivariate phenotype. We have examined the importance of taking these correlation structures into account for statistical data analyses.

Table 6.6. Percentages of correct identification of the two QTLs using various principal components under the three scenarios considered

Case 1

Type of PC	% of Correct Identification	
	Locus 1	Locus 2
$PC^{(1)}$	86	0
$PC^{(2)}$	0	92
$PC-1$	43	52
$PC-2$	51	58

Case 2

Type of PC	% of Correct Identification of Locus 1
$PC^{(1)}$	83
$PC-1$	70
$PC-2$	66

Case 3

Type of PC	% of Correct Identification	
	Locus 1	Locus 2
$PC^{(1)}$	79	27
$PC^{(2)}$	18	74
$PC-1$	52	47
$PC-2$	45	41

We have proposed, in keeping with a widely-adopted strategy (Zlotnik et al. 1983, Hasstedt et al. 1994, Boomsma 1996, Allison and Beasley 1998, Ott and Rabinowitz 1999), that the multivariate phenotype be reduced to a smaller number of phenotypes by the use of principal components. In extracting the principal components, often the underlying correlation structure is ignored, and the first few principal components are used for deciphering the genetic architecture of the multivariate phenotype. However, we posited that ignoring the underlying correlation structure may lead to loss in efficiency of mapping the loci controlling the multivariate phenotype. We have, therefore, performed simulations under the three scenarios described above, and have extracted principal components with and without taking

the underlying correlation structure into account.

We have found that ignoring the underlying correlation structure of the multivariate phenotype has no major effect on the ability to map the loci controlling the phenotype, using principal components, in low-density (wide) marker intervals. However, for the purpose of fine-mapping the quantitative trait loci, that is, localizing them in narrow (say, 1 cM) marker intervals, there is considerable loss in statistical efficiency unless the underlying correlation structure is taken into account at the data-reduction stage, that is, for extraction of principal components. We, therefore, suggest that the empirical correlation matrix of the components of a multivariate phenotype be critically examined, using suitable row/column permutations of the correlation matrix and appropriate tests of hypotheses pertaining to various submatrices of the correlation matrix, to identify patterns of relationships among the components. These identified patterns then should be used at the data-reduction stage. If principal components are extracted without consideration of the underlying correlation structure of the multivariate phenotype, mapping the loci controlling the phenotype cannot be done efficiently and may require huge sample sizes.

Chapter 7

Summary and Conclusions

In this Chapter, we provide a summary and highlight the main conclusions of the present study. The overarching goal of this study was to examine the various currently-used statistical procedures for human QTL mapping and to develop more efficient procedures, based primarily, but not exclusively, on sib-pair data, under more relaxed sets of assumptions.

In Chapter 1, we have provided a critical overview, albeit non-exhaustive, of the different statistical procedures that have been developed for QTL mapping. Starting with statistical methods for QTL mapping in plants and animals (Jayakar 1970, Haley and Knott 1992, Coupland 1995, Georges et al. 1995, Lark et al. 1995), we have focused primarily on the methods developed for human QTL mapping. In particular, we have provided a detailed overview of various QTL mapping methods — based on regression, likelihood and variance-components — using data on sib-pairs (Penrose 1937,1947,1953, Haseman and Elston 1972, Hill 1975, Blackwelder and Elston 1985, Lander and Botstein 1989, Amos et al. 1989, Olson 1995*a*, *b*, Almasy and Blangero 1999, Williams and Blangero 1999).

In Chapter 2, we have described the different models, which we have examined, underlying the determination of a quantitative trait. We have considered models both with and without dominance at the major trait locus. For a trait determined by multiple loci, we have assumed additivity of locus-specific marginal effects. We have also considered the possibility of epistatic interactions among the quantitative trait loci. We have used

a simple model of additive epistatic interaction among homozygotes at the different trait loci. This model was prompted by experimental observations on some plants and animals (Chang et al. 1999), and has been termed as the digenic interaction model (Kearsey and Pooni 1996).

In order to assess the efficiencies of our proposed statistical methodologies, based primarily on sib-pair data, we have used Monte-Carlo simulations. We have generated the trait values from normal distributions with appropriate mean vectors and dispersion matrices consistent with our assumed models. We have also argued why the simulation methods used by us are in agreement with 'traditional' simulation methods used in genetic epidemiology. In Chapter 3, we have provided empirical evidence in support of this fact. For generating identity-by-descent (i.b.d.) scores of sib-pairs, we have used the marginal distributions of trait i.b.d. scores and the conditional distributions of marker i.b.d. scores given the trait i.b.d. scores. The methodology for generating simulated data have also described in this Chapter.

In Chapter 3, trait values and genotypes at several codominant genetic markers with known genomic locations were collected from members of families and statistically analyzed in order to map a locus controlling a quantitative genetic trait to a specific genomic region. The parameter of interest was the vector of recombination fractions, θ, between the putative quantitative trait loci and the genomic markers. One of the major complications in estimating θ for a quantitative trait in humans is the lack of haplotype information on members of families. We have proposed a computationally simple and efficient method of estimation of θ in the absence of haplotypic information.

Our method was a two-stage estimation procedure using the expectation-maximization (EM) algorithm (Dempster et al. 1977). In the first stage, parameters of the QTL were estimated from data on a sample of unrelated individuals. From estimates thus obtained, we have used a Bayesian rule to infer QTL genotypes of parents in families. Finally, in the second stage of the procedure, we have proposed an EM algorithm for obtaining the maximum likelihood estimate of θ based on data of informative families (which are identified upon inferring parental QTL genotypes performed in the first stage).

We have tested our proposed methodology using simulated data. We initially considered a single major trait locus and performed two-point linkage analysis. We then extended our procedure to multiple trait loci and multipoint linkage analysis. We have shown, using simulated data, that our proposed procedure was cost-effective, computationally simple and statistically efficient. We have found that even in the presence of very high dominance effect, our procedure gave 80% correct classification of parental trait genotypes, which was essential for obtaining accurate estimates of θ. Moreover, our method performed quite efficiently even when the major trait locus explained only 30% of the variation in the QT. We have compared our proposed method with the popular method as implemented in MAPMAKER/QTL and have shown that our method performs better than that in MAPMAKER/QTL for the entire range of parameter values, except when the degree of dominance at the trait locus is very high.

We have also proposed a modification of our Classification-EM procedure by using the estimated posterior probabilities of parental QT genotypes in the second (estimation of θ) stage of our two-stage procedure, without actually classifying the parents into specific QT genotypes. This resulted in a more complex likelihood function at the second stage and resultantly computations were heavier. However, we have shown that this procedure performs better than our Classification-EM procedure, even when the degree of dominance is high, because genotypic misclassification of the parents is avoided in this procedure.

Many quantitative traits are determined jointly by multiple loci. Allelic effects at these loci are often not independent. That is, there are epistatic interactions among these loci. In Chapter 4, we have devised two computationally simple statistical procedures to detect and estimate the recombination fraction, θ, between quantitative trait loci and marker loci in the presence of additive epistatic digenic interactions. Our proposed methodologies are based on two different types of data.

We have first considered parental and offspring data separately on families in which only one parent was heterozygous at the marker locus and those in which both parents were heterozygous and have suitably modified the estimator proposed by Jayakar (1970) based on variance components. We have shown, based primarily on the widths of confidence intervals, that for

a wide range of parameter values the proposed estimator was quite efficient. Additionally, we have suggested a non-parametric procedure for testing null hypotheses regarding θ and have shown that the power function of the test had desirable statistical properties. We have also shown that analyses of data ignoring epistatic interactions, when in fact these were present, may lead to grossly inaccurate inferences about linkage. We found that if epistatic interactions were present, then only 50% of the simulation runs yielded estimates of θ which were included in the 95% empirical confidence intervals estimated by assuming that epistatic interactions were absent. However, the variance of the proposed estimator was found to be larger than that of the maximum likelihood estimator (m.l.e.). Our results provide insights into the major reasons why Jayakar's (1970) estimators do not perform well in practice and are, therefore, not used.

Our second data type included quantitative trait values of sib-pairs and their estimated marker i.b.d. scores. One of the popular statistical techniques to analyze such data is based on the regression of squared difference in trait values of sib pairs on their estimated marker i.b.d. scores. Under a very general setup, even in the presence of dominance and epistatic effects, Tiwari and Elston (1997) have extended the classical regression method for QTL mapping when the trait is controlled by two unlinked, autosomal, biallelic loci. Since this general model involved too many parameters, insights into effects of variation of individual parameters on the performance of the method were difficult to obtain. We have, therefore, examined the performance of the method under the specific digenic interaction model. We have also extended the method to the case of a quantitative trait that is controlled by multiple unlinked loci. We have shown that the sample size requirement for mapping a QTL was smaller if its marginal effect and heterozygosity were larger. Moreover, the presence of epistatic interactions reduced the sample size requirement compared to the situation in which the marginal effects were same, but epistatic interactions were absent. The competing strategies of analyzing the data by simultaneous, as opposed to sequential, consideration of the markers have been quantitatively assessed using simulation studies. As intuitively expected, the simultaneous strategy was found to be more optimal and cost-effective.

Genome-wide scans for mapping loci have proved to be extremely power-

ful and popular. In Chapter 5, we have presented a semi-parametric method of mapping quantitative trait loci using data generated from a two-stage genomic scan using sib-pairs. In a two-stage genomic scan, the entire genome or a large portion of the genome is saturated with low-density markers (> 5cM, say) at the first stage. At the second stage, the intervals that are identified as probable locations of the trait loci by analyzing data of the first stage, are then saturated with higher-density markers (1cM, say). These data are then analyzed for fine mapping the loci.

Our statistical strategy for analyzing the first-stage data was a low stringency semi-parametric method based on rank correlation of squared trait difference values of sib pairs and the estimated i.b.d. scores at the marker loci. We have suggested a low stringency method at the first stage to save on computational time and to identify all possible intervals that may contain the trait loci using a simple test statistic which asymptotically has a standard normal distribution. For analyzing the second-stage data, we have developed a non-parametric regression approach based on kernel smoothing (Silverman 1986), which we have shown to be more powerful than a currently-used linear regression method (Olson 1995a) for QTL mapping using sib pairs, particularly in the presence of dominance effects at the trait loci. When the dominance effect was small, both the parametric and the non-parametric regression procedures correctly identified the interval-location of the trait locus in 85-95% of the cases. However, when the degree of dominance was high, while the non-parametric method identified the correct interval in about 75% of the cases, the parametric method did so in only 40% of the cases. These and other results showed that the performance of our method was superior to the widely-used method of Olson (1995a), particularly in the presence of dominance and epistatic effects.

In Chapter 6, we have considered a *multivariate* quantitative trait and have suggested statistical methodologies for deciphering its underlying genetic architecture. A heritable multivariate quantitative phenotype comprises several correlated component phenotypes that are usually pleiotropically controlled by multiple loci and environmental factors. One approach to decipher the genetic architecture of a multivariate phenotype, in particular, to map the underlying loci, is to reduce the dimensionality of the data by a data-reduction technique, such as principal component analysis.

The extracted principal components can then be analyzed in conjuction with marker data to map the underlying loci. Ott and Rabinowitz (1999) have suggested the use of those principal components that have high heritabilities. We have used the two-stage variable-stringency semi-parametric method proposed in Chapter 5 based on data on squared differences in the extracted principal components of sib-pairs and their identity-by-descent scores on several marker loci.

We have examined the efficiency of this approach with and without taking into account the correlation structure of the multivariate phenotype when extracting principal components. We have assumed that genome-wide scan data on sib-pairs are available for low-density (widely-spaced) and high-density markers. We have assumed three models of multivariate phenotypes — in the first model, there were two uncorrelated sets of phenotypes, each set being controlled by a different locus; in the second model also, there were two uncorrelated sets of phenotypes, but the first set was controlled by a major locus, while the second set was determined solely by environmental factors; the third model was similar to the first model, except that some of the phenotypes were controlled by both loci. Using extensive simulations, based on the above three models of the multivariate phenotype, we have shown that although the ignoring of the correlation structure of the multivariate phenotype did not have any serious impact on the efficiency of mapping the underlying trait loci in wide marker intervals, there was a significant adverse effect of this for fine-mapping. When the correlation structure was taken into consideration, the correct intervals of the trait loci were identified in about 80-90% of the cases, but when it was ignored, the corresponding percentage was only about 50%. We have, therefore, recommended that the correlation structure of the multivariate phenotype be carefully examined to decide on the strategy of extracting principal components for deciphering the genetic architecture of the multivariate phenotype. Our results provide a framework and a general statistical approach to the problem of mapping loci that underlie a multivariate quantitative trait.

REFERENCES

1. Alcais A, Abel L (1999) Maximum-likelihood-binomial method for genetic model-free linkage analysis of quantitative traits in sibships. Genet Epidemiol 17:102-117.

2. Allison DB, Beasley M (1998) A method and computer program for controlling the family-wise alpha rate in gene association studies involving multiple phenotypes. Genet Epidemiol 15:87-101.

3. Allison DB, Heo M, Kaplan N, Martin ER (1999) Sibling-based tests of linkage and association for quantitative traits. Am J Hum Genet 64:1754-1763.

4. Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198-1211.

5. Amos CI, Elston RC (1989) Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol 6:349-360.

6. Amos CI, Elston RC, Bonney GE, Keats BJB, Berenson GS (1990) A multivariate approach for detecting linkage, with applIcation to a pedigree with an adverse lipoprotein phenotype. Am J Hum Genet 47:247-254.

7. Amos CI, Elston RC, Wilson AF, Bailey-Wilson JE (1989) A more powerful robust sib-pair test linkage for quantitative trait. Genet Epidemiol 6:435-449.

8. Amos CI, Liang AE (1996) A comparison of univariate and multivariate tests for genetic linkage. Genet Epidemiol 10:671-676.

9. Atwood LD, Mitchell BD, Stowell NC (1995) Segregation and linkage analysis of the complex trait Q1. Genet Epidemiol 12:713-718.

10. Atwood LD, Slifer SH (1997) Prior segregation analysis and the power to detect linkage. Genet Epidemiol 14:755-760.

11. Bonney GE, Lathrop GM, Lalouel J -M (1988). Combined linkage and segregation analysis using regressive models. Am J Hum Genet 43:29-37.

12. Boomsma DI (1996). Using multivariate genetic modeling to detect pleiotropic quantitative loci. Behav Genet 26:161-166.

13. Berrethini WH, Ferraro TN, Alexander RC, Buchberg AM, Vogel WH (1994) Quantitative trait loci mapping of three loci controlling morphine preference using inbred mouse strains. Nat Genet 7:54-58.

14. Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85-97.

15. Chang B, Smith RS, Hawes NL, Anderson MG, Zabaleta A, Saninova O, Roderick TH, Heckenlively JR, Davisson MT, John SW (1999) Interacting loci cause iris atrophy and glaucoma in DBA/2J mice. Nature Genet 21:405-409.

16. Collins FS (1995) Positional cloning moves from perditional to traditional Nature Genet 9:347-350.

17. Cotterman CW (1969) Factor union phenotype system. Computer Appl in Genet. University of Hawaii Press, Honululu 1-19.

18. Coupland G (1995) Genetic and environmental control of flowering time in *Arabidopsis*. Trends Genet 11:393-397.

19. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39:1-38.

20. Dizier M, Bonaiti-Pellie C, Clerget-Darpoux F (1993) Conclusions of segregation analysis for family data generated under two locus models. Am J Hum Genet 53:1338-1346.

21. Eaves LJ, Neale MC, Maes H (1996). Multivariate multipoint linkage analysis of quantitative trait loci. Behav. Genet 26:519-525.

22. Edwards JH (1998) Penrose and sib-pairs. Ann Hum Genet 62:365-377.

23. Elbein SC, Hoffman MD, Teng K, Leppert MF, Hasstedt SJ (1999) A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. Diabetes 48:1175-1182.

24. Elston RC, Guo X, Williams LV (1996) Two-stage global search designs for linkage analysis using pairs of affected relatives. Genet Epiodemiol 13:535-558.

25. Eshed Y, Zamir D (1996) Less-than-additive epistatic interactions of quantitative trait loci in tomato. Genet 143:1807-1817.

26. Everitt BS, Hand DJ (1981) Finite mixture distributions. Chapman and Hall, London.

27. Fergusson TS (1967) Mathematical Statistics: A Decision-theoretic Approach. Academic Press, New York.

28. Fijneman RJA, de Vries SS, Jensen RC, Demant P (1996) Complex interactions of new quantitative trait loci it Sluc 1, Sluc 2, Sluc 3 and *Sluc 4*, that influence the susceptibility to lung cancer in the mouse. Nat Genet 9:465-467.

29. Frankel WN, Schork NJ (1996) Who's afraid of epistasis? Nat Genet 9:371-373

30. Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait data. Am J Hum Genet 54:1092-1103.

31. Gauderman WJ, Faucett CL, Morrison JL, Carpenter CL (1997) Joint segregation and linkage analysis of a quantitative trait compared to separate analyses. Genet Epidemiol 14:993-998.

32. Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, *et al.* (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genet 139:907-920.

33. Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. Am J Hum Genet 47:957-967.

34. Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315-324.

35. Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet 8:299-309

36. Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behaviour Genet 2:3-19.

37. Hasstedt SJ, Hunt SC, Wu LL (1994) Evidence for multiple genes determining sodium transport. Genet Epidemiol 11:553-568.

38. Hill A (1975) Quantitative linkage: a statistical procedure for its detection and estimation(1975) Ann of Human Genet 38:439-449.

39. Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genet 135:205-211.

40. Jayakar SD (1970) On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. Biometrics 26:451-464.

41. Kao CH, Zeng ZB (1997) General formulas for observing the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics 53:653-665.

42. Kearsey MJ, Pooni HS (1996) The Genetical Analysis of Quantitative Traits. Chapman and Hall, London.

43. Kruglyak L, Lander ES (1995a) A nonparametric approach for mapping quantitative trait loci. Genet 139: 1421-1428.

44. Kruglyak L, Lander ES (1995b) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439-454.

45. Krushkal J, Ferrell R, Mockrin SC, Turner ST, Sing CF, Boerwinkle E (1999) Genome-wide linkage analysis of systolic blood pressure using highly discordant siblings. Circulation 99:1407-1410.

46. Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genet 121:185-199.

47. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037-2048.

48. Lange K, Boehnke M (1983) Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. Am J Med Genet 14:513-524.

49. Lark KG, Chase, Adler F, Mansur LM, Orf JH (1995) Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele at another. Proc Natl Acad Sci USA 92:4656-4660.

50. Lincoln SE, Daly MJ and Lander ES (1993) MAPMAKER/QTL Version 1.1. http://www.genome.wi.mit.edu.

51. Majumder PP, Moss HB, Murrelle L (1998) Familial and nonfamilial factors in the prediction of disruptive behaviors in boys at risk for substance abuse. J Child Psychol Psychiat 39:203-213.

52. Mather K and Jinks JL (1982) Introduction to Biometrical Genetics. Chapman and Hall, London.

53. Moldin SO, van Eerdewegh P (1995) Multivariate genetic analysis of oligogenic disease. Genet Epidemiol 12:801-806.

54. McLachlan GJ and Krishnan T (1997) The EM algorithm and extensions. New York, John Wiley and Sons.

55. Niu T, Chen C, Cordell H, Yang J, Wang B, Fang Z, Schork NJ et al. (1999) A genome-wide scan for loci linked to forearm bone mineral density. Hum Genet 104:226-233.

56. Olson JM (1995a) Robust multipoint linkage analysis: an extension of the Haseman-Elston method. Genet Epidemiol 12:177-193.

57. Olson JM (1995b) Multipoint linkage analysis using sib-pairs: an interval mapping approach for dichotomous outcomes. Am J Hum Genet 56: 788-798.

58. Olson JM (1999) Linkage Analysis, Model-Free. Encyclopaedia of Biostatistics, edited by Armitage P. and Colton T. John Wiley & Sons, Chichester.

59. Olson JM, Wijsman EM (1993) Linkage between quantitative trait and marker loci: methods using all relative pairs. Genet Epidemiol 10:87-102.

60. Ott J (1977) Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. Ann Hum Genet 40:443-454.

61. Ott J (1999) Analysis of Human Genetic Linkage. 3^{rd} edition. Johns Hopkins University Press, Baltimore.

62. Ott J, Rabinowitz D (1999) A principal-components approach based on heritability for combining phenotype information. Hum Hered 49:106-111.

63. Page GP, Amos CI, Boerwinkle E (1998) A quantitative lod score test statistic and sample size for exclusion and linkage of quantitative traits in human sibships. Am J Hum Genet 62:962-968.

64. Penrose LS (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugen 8:133-138.

65. Penrose LS (1947) A further note on the sib-pair linkage method. Ann Eugen 13:25-29.

66. Penrose LS (1953) The general purpose of sib-pair linkage test. Ann Eugen 18:120-124.

67. Randles RH, Wolfe DA (1979) Introduction to the theory of nonparametric statistics. John Wiley & Sons.

68. Rao CR (1973) Linear Statistical Inference and its applications. Second Edition, New York: Wiley.

69. Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. Genet 8:552-560.

70. Schork NJ (1993) Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. Am J Hum Genet 53:1306-1319.

71. Schork NJ, Kreiger MR, Trolliet KG, Franchini G, Kolke EM, Kreiger VJ, Lander ES *et al.* (1995) A biometrical genome search in rats reveals the mutigenic basis of blood pressure variation. Genome Res. 5:164-172.

72. Schork NJ, Nath SK, Lindpainter K, Jacob HJ (1996) Extensions to quantitative trait locus mapping in experimental organisms. Hypertension 28:1104-1111.

73. Schork NJ, Nath SK, Fallin D and Chakravarti A (2000) Linkage disequilibrium analysis of bi-allelic DNA markers, human quantitative trait loci and threshold-defined cases and controls. *Submitted.*

74. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall.

75. Tanksley SD (1993) Mapping polygenes. Annu Rev Genet 27:207-233.

76. Tiwari HK, Elston RC (1997) Linkage of multilocus components of variance to polymorphic markers. Ann Hum Genet 61:253-261.

77. Todorov AA, Volger GP, Gu C, Province MA, Li Z, Heath AC, Rao DC (1998) Testing causal hypotheses in multivariate linkage analysis of quantitative traits: general formulation and application to sib-pair data. Genet Epidemiol 15:263-278.

78. van Wezel T, Stassen APM, Moen CJA, Hart AAM, van der Valk MA, Demant P (1996) Gene interactions and single gene effects in colon tumour susceptibility in mice. Nat Genet 9:468-470.

79. Visscher PM, Thompson R, Haley CS (1996) Confidence intervals in QTL mapping by bootstrapping. Genet 143:1013-1020.

80. Weller JI (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42:627-640.

81. Whittaker JC, Thompson R, Visscher, PM (1995) On the mapping of QTL by regression of phenotype on markertype. Heredity 77:23-32.

82. Williams JT, Blangero J (1999) Comparison of variance components and sibpair-based approaches to quantitative trait linkage analysis in unselected samples. Genet Epidemiol 16:113-134.

83. Williams NM, Rees MI, Holmes P, Norton N, Cardno AG, Jones LA, Murphy KC et al. (1999) A two-stage genome scan for schizophernia susceptibility genes in 196 affected sibling pairs. Hum Mol Genet 8:1729-1739.

84. Wyst M, Fisher G, Immervoll T, Jung M, Saar K, Rueschendorf F, Reis A et al. (1999) A genome-wide search for linkage to asthma. Genomics 58:1-8.

85. Zeng ZB (1994) Precision mapping of quantitative trait loci. Genet 136:1457-1468.

86. Zlotnik LH, Elston RC, Namboodiri KK (1983) Pedigree discriminant analysis: a method to identify monogenic segregation. Am J Med Genet 15:307-313.