MULTIFACTORIAL MODELS AND LIKELIHOODS FOR THE SEGREGATION ANALYSIS OF QUANTITATIVE TRAITS

by

G. Jay Graepel

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1361

September 1981

MULTIFACTORIAL MODELS AND LIKELIHOODS FOR THE SEGREGATION ANALYSIS OF QUANTITATIVE TRAITS

by

G. Jay Graepel

A Dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1981

Approved by:

R. C. Elston.

Adviser

L L Kupper

Reader

K K Nz Broodie

Reader

ABSTRACT

GEORGE JAY GRAEPEL. Multifactorial Models and Likelihoods for the Segregation Analysis of Quantitative Traits. (Under the direction of Robert C. Elston.)

A multifactorial model for the segregation analysis of quantitative traits in pedigrees is presented. The model includes both polygenic and monogenic effects. The model also allows for two types of environmental correlation, a within sibship correlation and a within nuclear family correlation.

Methods for the approximation of the true likelihood for this model are presented. These models are derived from the methods for estimating the parameters of a mixture of normal distributions. The estimation methods are those equating moments, maximum likelihood and two methods of least squares estimation. One least squares method involves minimizing the sum of squared differences between the exact likelihood function and the approximation function; the other method minimizes the sum of squared differences between the moment generating function of the exact likelihood function and the moment generating function of the approximating function. Mixtures of up to three distributions are used in the approximation. The adequacy of each approximation is studied on simulated nuclear family data of size eight, as well as on a subset of real data containing nuclear families of varying size. The results of the study indicate that under suitable conditions the

approximations are adequate.

Segregation analysis of IgE levels in 173 nuclear families is done. The results of the analysis confirm the presence of a major gene segregating in these data.

ACKNOWLEDGMENTS

I would like to express my appreciation to my adviser, Dr. R. C. Elston, for his initial suggestion of the topic and more importantly for his infectious enthusiasm and continuous support. Gratitude is also expressed to Drs. K. K. Namboodiri, M. J. Symons, N. J. Johnson, L. L. Kupper and N. R. Mendell, who served on my advisory committee.

Ellen Kaplan did the computer programming. I thank her for her technical skills which were irreplaceable and her patience and energy which were limitless.

Only those who have seen my handwriting can appreciate the work of Marie Dominelli. I have. I do.

The financial support of the Genetics Curriculum and the Department of Biostatistics is gratefully acknowledged.

Last but not least I would like to thank my friends and family for their support in helping me maintain a semblance of sanity which made this work all possible.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iii

LIST OF TABLES vi

LIST OF FIGURES vii

Chapter

I. INTRODUCTION. 1

II. THE GENETIC MODELS. 4

2.1 The Morton-MacLean model for nuclear families. 4
2.1.1 The likelihood 6
2.1.2 Evaluation of the likelihood 7
2.1.3 Hypothesis testing 8

2.2 The mixed model for small pedigrees. 10
2.2.1 Evaluation of the likelihood 12

2.3 The general model of Elston and Stewart for pedigrees. 14
2.3.1 The likelihood computation algorithm for a simple pedigree. 14
2.3.2 Likelihood computation for complex pedigrees without loops. 18
2.3.3 The single locus model 20
2.3.3.1 Hypothesis testing under the single locus model. 22
2.3.4 The polygenic model 25
2.3.4.1 Extensions to the polygenic model. . . 27

2.4 The mixed model. 28

III. THE METHODS FOR APPROXIMATION 33

3.1 Mixtures of normal distributions 34

Chapter Page

3.1.1 Introduction 34
3.1.2 Methods of estimating parameters . . . 35
3.1.3 Summary. 45

3.2 Methods of approximation 45
3.2.1 Method of moments. 48
3.2.2 Method of maximum likelihood 48
3.2.3 Method of least squares. 49
3.2.4 Method using moment generation function 49
3.2.5 Method of fitting a single distribution 50

IV. EMPIRICAL STUDY OF THE APPROXIMATION METHODS . . . 52

4.1 Description of the data. 52
4.2 Description of methods of approximation. . . . 55
4.3 Methods of comparison. 56
4.4 Results of comparisons on simulated data . . . 57
4.5 Summary of results from simulated data 62
4.6 Comparisons on real data 91
4.7 Results of approximations on real data 92

V. THE SEGREGATION ANALYSIS OF IgE 108

5.1 Description of the data. 108
5.2 Summary of previous analyses 109
5.3 Analysis 110
5.3.1 Sample size. 111
5.3.2 Numerical accuracy 112
5.3.3 Joint vs conditional likelihood. . . . 114

VI. SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH. . . . 116

BIBLIOGRAPHY. 120

LIST OF TABLES

Table		Page
2.1	Transmission probabilities p_{stu} for two allele autosomal locus.	23
4.1	Simulated data for comparing approximation methods	54
4.2	Distribution of accuracy (d_i) for SD1 and SD2 methods	63
4.3	Distribution of accuracy (d_i) for MM method . . .	64
4.4	Distribution of accuracy (d_i) for ML2 and ML3 methods	65
4.5	Distribution of accuracy (d_i) for LS2 and LS3 methods	66
4.6	Distribution of accuracy (d_i) for MG2 and MG3 methods	67
4.7	Summary of accuracy (d_i) and time for all methods	68
4.8	Distribution of accuracy for each approximation on all 75 likelihoods	94
4.9	Log likelihood for first 25 families of IgE data set for 3 sets of parameters for 7 numerical methods	95
5.1	Log likelihood and resulting χ^2 for hypothesis q=t=d=0	113

LIST OF FIGURES

Figure			Page
2.1		Hypothetical examples to illustrate the the notation for pedigrees.	16
2.2		Hypothetical complex pedigree	19
4.1	a.	Fit of method SD1 on model 1 data	69
	b.	Fit of method SD1 on model 2 data	70
	c.	Fit of method SD1 on model 3 data	71
	d.	Fit of method SD2 on model 1 data	72
	e.	Fit of method SD2 on model 2 data	73
4.2	a.	Fit of method MM on model 1 data.	74
	b.	Fit of method MM on model 2 data.	75
4.3	a.	Fit of method ML2 on model 1 data	76
	b.	Fit of method ML2 on model 2 data	77
	c.	Fit of method ML2 on model 3 data	78
	d.	Fit of method ML3 on model 1 data	79
	e.	Fit of method ML3 on model 2 data	80
4.4	a.	Fit of method LS2 on model 1 data	81
	b.	Fit of method LS2 on model 2 data	82
	c.	Fit of method LS2 on model 3 data	83
	d.	Fit of method LS3 on model 1 data	84
	e.	Fit of method LS3 on model 2 data	85
4.5	a.	Fit of method MG2 on model 1 data	86
	b.	Fit of method MG2 on model 2 data	87
	c.	Fit of method MG3 on model 1 data	88
	d.	Fit of method MG3 on model 2 data	89
4.6		Distribution of accuracy. Percentiles (97.5, 95., 83.3, 50, 16.7, 5., 2.5).	90

Figure Page

4.7 a. Fit of method SD1 on IgE data 96

b. Fit of method SD2 on IgE data 97

c. Fit of method ML2 on IgE data 98

d. Fit of method LS2 on IgE data 99

e. Fit of method MG2 on IgE data 100

f. Fit of MIXMOD on IgE data 101

4.8 a. Accuracy of SD1 method vs family size on IgE data. 102

b. Accuracy of SD2 method vs family size on IgE data. 103

c. Accuracy of ML2 method vs family size on IgE data. 104

d. Accuracy of LS2 method vs family size on IgE data. 105

e. Accuracy of MG2 method vs family size on IgE data. 106

f. Accuracy of MIXMOD vs family size on IgE data. 107

CHAPTER I

INTRODUCTION

The study of traits in families or pedigrees requires special statistical methods since the observations on individuals are not independent. One of the statistical methodologies which deals with this problem is called segregation analysis. The purpose of segregation analysis is to determine the genetic mechanisms underlying traits with heritable components. If there is a heritable component we would like to be able to determine what the genetic mechanism is, and we would like to be able to classify individuals by this genetic component. Early methods of segregation analysis were primarily restricted to qualitative data or dichotomous traits from samples of nuclear families, and are discussed by Elandt-Johnson (1971). The work here will concentrate on one aspect of segregation analysis, the 'mixed model' (a term used by Morton and MacLean (1974)) for pedigrees (more than two generational data) and particularly with regard to quantitative traits.

A mixed model in segregation analysis is considered to be a model which allows for both a major gene (monogenic) effect and a polygenic effect. A major gene effect is an

effect which is the result of segregation at one locus, with this locus contributing a large portion of the variability of a trait in a particular sample. Polygenic effects are considered to be the result of an indefinite number of additive unlinked loci with small, approximately equal effects. The sum of these effects is considered to be normally distributed and is called the polygenic effect. As demonstrated by Elston (1980), this assumption of normality is reasonable for as few as three equal and additive loci.

The mixed model was originally proposed by Elston and Stewart (1971); Morton and MacLean (1974) have also included a common within sibship environmental effect in their model. Allowance might be made for other random effects, including other environmental effects and effects due to assortative mating (Boyle and Elson, 1979). These other random effects can also generally be considered to be normally distributed. Much theoretical work has been done within the mixed model framework, but at present these theoretical constructs have not been applied to the analysis of large or even moderate sized pedigrees for reasons to be discussed later.

Chapter II presents the genetic models for quantitative traits in detail. It discusses the virtues and limitations of the various models. One of the limitations of the model that we are most interested in, is the practical difficulty of computing the likelihood for even moderate size samples, since the time required to calculate a likelihood is a rapidly

increasing exponential function of the sample size.

Chapter III presents a possible solution to these computational difficulties. It includes a review of the literature on methods for estimating parameters for mixtures of normal distributions. It offers proposals on how to use these methods to approximate the required likelihood and therefore resolve the computational problems.

Chapter IV is a study of the proposed approximations given in Chapter III. It is an empirical study and is based on both simulated and observed data. The simulated data are data generated under three genetic models. The observed data are a subset of a larger data set which is the subject of Chapter V.

Chapter V discusses the segregation analysis of IgE, a serum protein associated with allergenic responses. The data set has 173 nuclear families with a total of 781 individuals. It has been analyzed previously by two groups of investigators and conflicting conclusions were reached. The purpose of Chapter V is to attempt to resolve this conflict.

The final chapter summarizes the results of the previous chapters. It also includes suggestions for further research and study.

CHAPTER II

THE GENETIC MODELS

In this chapter various genetic models will be presented, together with the associated likelihoods, which have been used to study quantitative traits. The first section will discuss a model for nuclear families which is parameterized in such a way that a major gene, polygenic effect and an effect due to common environment in a sibship are accommodated. The subsequent sections will present more general models, for more than two-generational data, with both similar and different parameter schemes.

2.1. The Morton-MacLean Model for Nuclear Families

Morton and MacLean (1974) developed a mixed model for nuclear families. Their model incorporates environmental effects in addition to a major gene and polygenic effects. The environmental effect E is distributed $N(0,\sigma_E^2)$ over the population. For offspring, the environmental effect is partitioned into two effects, a within sibship effect C and a random effect R, which are both normal, independent and additive, with corresponding variances σ_C^2 and σ_R^2. Since

they are independent the total environmental variance is $\sigma_E^2 = \sigma_C^2 + \sigma_R^2$. The polygenic effect G is distributed $N(0,\sigma_G^2)$, and is also partitioned for offspring into two independent effects B and Y. B is the midparental breeding value and is distributed $N(0,\sigma_G^2/2)$, and Y is the individual deviation around this midparental value, also distributed $N(0,\sigma_G^2/2)$. The major gene effect M is the result of one of 3 possible outcomes or genotypes AA, Aa or aa. Its distribution is defined as

Genotype	AA	Aa	aa
Frequency	q^2	$2q(1-q)$	$(1-q)^2$
Effect	z+t	z+td	z

where z is the mean of genotype aa

t is the displacement of the major gene, $t>0$

d is the degree of dominance, $0 \leq d \leq 1$

and q is the population frequency of allele A.

The mean of M is given by $\mu = z+q^2t+2q(1-q)td$, and the variance is given by $\sigma_M^2 = q^2(z+t)^2+2q(1-q)(z+td)^2+(1-q)^2z-^2\mu^2$. The model for the phenotype X is then, using the subscript i to indicate the i-th individual,

$$X_i = M_i + G_i + E_i \quad \text{for a parent,}$$

$$X_i = M_i + B + Y_i + C + R_i \quad \text{for an offspring.}$$

It should be noted B and C are effects dependent only on the sibship and are not specific to an individual. The mean of X is given by μ, and, since all the effects are

independent, the variance of X is given by $\sigma_X^2 = \sigma_G^2 + \sigma_M^2 + \sigma_E^2$.

2.1.1. The Likelihood

Morton bases inference on this model and the likelihood of the sibship's phenotypes given the phenotypes of the parents. The likelihood is derived for various situations, i.e. when both parents' phenotypes are known, only one is known, or neither is known. Letting s and t index the major genotypes of these parents, u the major genotypes of the sibs, $V = B + C$, x_i the observed phenotypes of the i-th sib, and x_s and x_t the observed phenotypes of the parents, the general likelihood as given by Boyle and Elston (1979) is

$$L = \int_V \sum_s \sum_t f(V, M_s, M_t, x_s, x_t) \prod_i \sum_u f(x_i|V, M_u)\, f(M_u|M_s, M_t) \quad (1)$$

where the symbol $\int_V$ is used to mean that everything following to the right is to be integrated with respect to V from minus infinity to plus infinity. Each term to the right of the product sign is the conditional likelihood of observing a sib's phenotype given the genotypes of the parents and V. The terms to the left of the product sign in (1) are the parental probabilities. Using the notation $\phi(x,y) = \frac{1}{\sqrt{2\pi y^2}} \exp\{-x^2/2y^2\}$, the function $f(x_i|V, M_u)$ is given by

$$\phi(x_i - M_u - V, \sigma_G^2/2 + \sigma_R^2).$$

The function $f(M_u|M_s, M_t)$ involves the Mendelian probabilities of transmission from parents to offspring. The function

$f(V, M_s, M_t, x_s, x_t)$ can factored as

$$f(V, M_s, M_t, x_s, x_t) =$$

$$f(M_s)\ f(M_t)\ f(x_s|M_s)\ f(x_t|M_t)\ f(V|x_s, x_t, M_s, M_t).$$

The functions $f(M_s)$, $f(M_t)$ were defined earlier in the table, and

$$f(x|M) = \phi(x-M, \sigma_G^2 + \sigma_E^2).$$

After some manipulation Boyle and Elston have shown that

$$f(V|x_s,x_t, M_s, M_t) =$$

$$\phi(V - \sigma_G^2[(x_s + x_t) - (M_s + M_t)]/[2(\sigma_G^2 + \sigma_E^2)],$$

$$\sigma_C^2 + \sigma_G^2 \cdot \sigma_E^2/[2(\sigma_G^2 + \sigma_E^2]).$$

As mentioned earlier, Morton and MacLean use the conditional likelihood of observing the sibship given the parents phenotype,

$$L' = L/(f(x_s)\ f(x_t)), \text{ where } f(x_s)\ f(x_t) =$$

$$\sum_{st}\sum f(M_s)f(M_t)\ \phi(x_s - M_s, \sigma_G^2 + \sigma_E^2)\ \phi(x_t - M_t, \sigma_G^2 + \sigma_E^2).$$

2.1.2. Evaluation of the Likelihood

Although an analytical calculation of the likelihood is possible, Morton and MacLean have chosen to approximate the integration by estimating the area under parts of the curve and summing over these parts. They divide the domain of the function into equal-length intervals and then calculate estimates of the area bounded by these intervals as the product of the length of the interval and the value of the function at the midpoint of the interval. By summing over all intervals an approximation to the integral is obtained. They suggest intervals of length .25 standard deviations,

with the extreme intervals bounded by 4 standard deviations. It is not mentioned in the article, but the term standard deviations presumably refers to the phenotypic standard deviations, and the range of the approximation is the interval of the phenotypic mean $\pm$ 4 standard deviations.

2.1.3. Hypothesis Testing

Hypothesis testing under this model is discussed e.g. by Gerrard et al. (1978). They use the likelihood ratio test. Before actually performing tests of hypotheses using this model it is necessary to transform the data to reduce any skewness, as under this model skewness can simulate the effect of a major locus. The first step in the transformation procedure they adopt is to standardize the observations within generations by subtracting the mean of the respective generation from each observation and then dividing the difference by the standard deviation of the respective generation. The second step is to use the transformation,

$$Y = \frac{r}{p}\left[\left(\frac{X}{r} + 1\right)^{p} - 1\right]$$

discussed by MacLean et al. (1976), where X represents the standardized observation, Y represents the corresponding tranformed observation, r represents an arbitrary constant such that $\frac{X}{r} - 1 > 0$, for all X, and p is a parameter to be estimated. By using maximum likelihood procedures, is estimated under the assumption that Y is distributed as a mixture of either one, two or three normal distributions.

For each of the assumptions there is a corresponding estimate of p and in each case the within-distribution skewness is reduced. By comparing the likelihoods under the three different models it is possible to determine which model best describes the data and consequently which estimate of is best. This method yields information on the possibility and nature of a genetic mechanism as well as finding the appropriate value of p for reducing skewness; this is because a mixture of 2 or 3 distributions is more suggestive of a major gene influence then a single distribution is. Using the transformed values, hypotheses can be tested by the methods of segregation analysis.

The following are suitable hypotheses to test.

1) H_o: $q = t = d = 0$.

This is a test to determine if there is a major gene effect.

2) H_o: $\sigma_G^2/\sigma_X^2 = 0$.

This is a test to determine if there is a polygenic effect.

3) H_o: $d = 0$; $\sigma_C^2/\sigma_X^2 = 0$.

This is a joint test of whether there is complete recessiveness at the major locus and whether there is environmentally caused sibling correlation.

4) H_o: $d = 1$ and/or H_o: $d = .5$.

These correspond to testing whether there is complete dominance or additivity at the major locus.

Using the Morton-MacLean approach to segregation analysis, one should also test for homogeneity among mating types. Since the likelihood used for testing is conditioned on the parents' phenotypes, analysis of different parental mating types should yield the same results. If there is consistency over mating types, one can conclude there is no evidence for heterogeneity. To test for such consistency the phenotypic scale is polychotomized and each resulting group is defined as a mating class. As a simple example, consider using a dichotomy with individuals whose phenotype is less than the median value being in the L class and those individuals with phenotypes greater than the median value being the H class. The previous tests are then repeated on three separate sets of data, those with mating type H x H, H x L or L x L. If the results are similar for the three sets of data, an overall conclusion of a major gene is supported.

2.2. The Mixed Model for Small Pedigrees

Ott (1979) presented the same model as the Morton-MacLean model, but for a pedigree. Ott's model included the same effects as the Morton-MacLean model, but with a slightly different parameterization. He also used the likelihood of the entire pedigree, and not a likelihood conditional on a subset of the individuals in the pedigree.

Suppose there are n individuals in the pedigree with x_i being the phenotype of the ith individual (i=1,...,n),

and let $\underset{\sim}{x} = (x_1,\ldots,x_n)$ be the vector of phenotypes. Let g_i be the major genotype of the ith individual. For a two allele system g_i is one of three values, say AA, Aa, aa. Let $\underset{\sim}{g} = (g_1,\ldots,g_n)$ be the corresponding vector of major genotypes. Since there are three possible genotypes for each individual there are 3^n different vectors possible. The probability of any one of these vectors, $P(\underset{\sim}{g})$, is a function of the population genotypic frequencies and the Mendelian probabilities. For some vectors the genotypic configurations are incompatible, for example the mating AA x AA would not yield any aa offspring; therefore the probability associated with such vectors would equal zero.

The likelihood can then be expressed as a mixture of multivariate normal density functions, i.e. it is of the form

$$L = \sum_{\underset{\sim}{g}} f(\underset{\sim}{x}|\underset{\sim}{g})\ P(\underset{\sim}{g}),$$

where $f(\underset{\sim}{x}|\underset{\sim}{g})$ is a multivariate normal distribution.

The structure of the variance matrix is the same for each distribution and is determined by the particular family structure. It is given by

$$V(\underset{\sim}{x}) = \sigma_a^2 A + \sigma_c^2 C + \sigma_e^2 I.$$

The polygenic variance is represented by σ_a^2 and A is a matrix of coefficients which when multiplied by σ_a^2 gives the covariance of the polygenic effect between relatives in the pedigree. The variance due to common sibship environment is given by σ_c^2 and the matrix C has elements

c_{ij} = 1 if i = j or if the i th and j th individuals are in the same sibship, and c_{ij} = 0 otherwise. The random environment variation is given by σ_e^2 and the matrix I is an identity matrix. Ott has shown that these variances are linear functions of the variances of the Morton-MacLean model.

2.2.1. Evaluation of the Likelihood

For $n \leq 10$ a program is available from Ott which evaluates the likelihood by this method of summation, but it is relatively slow. Even after the elimination of those vectors whose probability is equal to zero, the number of terms is prohibitively large for pedigrees with $n > 10$. For nuclear families the number of vectors with positive probability is given by $4 + 2^n + 3^{n-2}$, which, although it is smaller than 3^n, is still an unmanageable number of terms.

Ott has suggested another way of evaluating this likelihood. He proposed sampling from all possible genotypic configurations. The likelihood can then be estimated by

$$L = \frac{1}{N} \sum_{i=1}^{N} f(\underset{\sim}{x} | \underset{\sim}{g}_i),$$

where N is the number of vectors sampled.

Lee (1978) estimated a similar likelihood by estimating the most probable major genotypic vector or configuration, and using this configuration as known, calculated one term of the complete likelihood. Inference is then based on the

likelihood conditional on this most probable major genotypic configuration. He estimated this most probable genotypic configuration by finding the most probable genotype for each individual, given the phenotypes of his relatives and assuming the hypothesis of monogenic Mendelian inheritance is true. The following is a brief discussion of how this most probable genotypic configuration was obtained.

Let H_o denote the hypothesis of Mendelian inheritance, and $\underset{\sim}{\theta}$ be the unknown parameters in the model, e.g. means and regression coefficients. We can write the likelihood of observing a set of pedigree data $\underset{\sim}{x}$, under Mendelian inheritance, as

$$L(\underset{\sim}{x}|H_o,\underset{\sim}{\theta}).$$

This likelihood can be written as the sum of three joint likelihoods, each being the likelihood of the pedigree and that a given individual i has a particular genotype g_i, (g_i = $\underline{AA}$, $\underline{Aa}$, $\underline{aa}$), ie.

$$L(\underset{\sim}{x}|H_o,\underset{\sim}{\theta}) = \sum_i L(x,\ g_i|H_o,\underset{\sim}{\theta}).$$

The posterior probability that individual i has genotype t_i, given what is known about his relatives, is estimated by the ratio

$$R = L(x,\ t_i|H_o,\hat{\underset{\sim}{\theta}}) / \sum_i L(x,\ g_i|H_o,\hat{\underset{\sim}{\theta}}),$$

where $\hat{\theta}$ represents the ML estimates of unknown parameters $\underset{\sim}{\theta}$.

The genotype t_i for which the ratio is a maximum is the most probable genotype for individual i. The most probable genotype for each individual is then the corresponding element of the vector $\underset{\sim}{g}$. Lee also generated other major genotypic vectors. He did this by perturbing his estimates of $\underset{\sim}{\theta}$ and then calculating the corresponding values of R. He had four parameters in the model and he perturbed estimates of each of these by $\pm$ 1.5 standard deviations, giving him 16 other genotypic configurations. The results from using a likelihood with just the most probable genotypic vector were similar to the results using any of the other 16 likelihoods. This similarity suggested the method was reasonable.

2.3. The General Model of Elston and Stewart for Pedigrees

2.3.1. The Likelihood Computation Algorithm for a Simple Pedigree

Elston and Stewart (1971) proposed a general model for the analysis of pedigree data. They derived the likelihood that a particular set of data is observed in a pedigree. Their derivation assumed random mating, one set of original parents, no environmental correlations and no consanguineous matings. After constructing this likelihood in a general case, they presented some specific cases under different genetic models. Below is a discussion of how this likelihood is constructed.

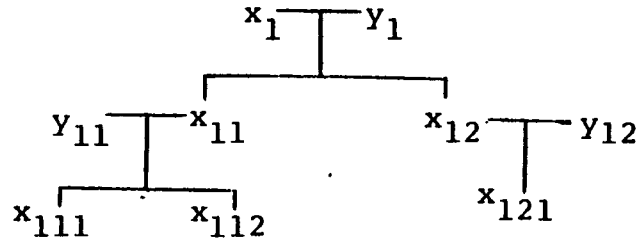
The notation used here is slightly different from

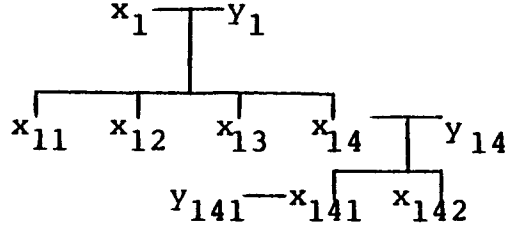
what Elston and Stewart used, but it will be consistent with that used in the remainder of this discussion. Let a measure on an individual who is related to someone in a previous generation be denoted by x, a measure on an individual 'marrying into' the pedigree be denoted by y, and for the original parents arbitrarily let one be denoted by x and the other by y. The observations need to be indexed, so let the observations of the original parents be given by x_1, y_1, and the observation on the i-th child of these parents be given by x_{1i}. The subscripts continue with the position of the subscript denoting the generation until all the members are indexed. Measures on spouses of individuals in the pedigree are denoted by y, with the same subscripts for both spouses. The notation is best understood by an example. (See figure 2.1 which is copied from the paper by Elston and Stewart). Let k equal the number of different genotypes that cause variation for the trait of interest. For example, assuming one locus and two possible alleles at that locus, there are three possible genotypes, say AA, Aa, and aa. With each possible genotype there is associated a probability density function, say $g_u(x)$ $(u=1,\ldots,k)$. This density function is the density function of the phenotype x given the genotype u. Let p_{stu}, the transmission probability, equal the probability that an individual has genotype u given his parents genotypes are s and t.

From the above we can construct the likelihood L of

FIGURE 2.1

HYPOTHETICAL EXAMPLES TO ILLUSTRATE THE NOTATION FOR PEDIGREES





observing a set of data from a sibship of size N given the parents' genotypes are s and t:

$$L = \prod_{i=1}^{N} \sum_{u=1}^{k} p_{stu}\, g_u(x_i).$$

If we let ψ_v be the probability that a person from the population has genotype v, then the likelihood of observing a particular phenotype on a random person is

$$L = \sum_{v=1}^{k} \psi_v\, g_v(y).$$

Therefore the likelihood of the phenotypes observed in a sibship and the spouses of the sibship, given the parents genotypes, is, under random mating,

$$L = \prod_{i=1}^{N} \sum_{u=1}^{k} p_{stu}\, g_u(x_i) \sum_{v=1}^{k} \psi_v g_v(y_i). \quad (1)$$

This is a function of s and t; but s and t of this generation are u and v of the previous generation. This likelihood can thus be written as

$$\Gamma_j = \prod_{i=1}^{N_j} \sum_{s_j=1}^{k} p_{s_{j-1}t_{j-1}s_j}\, g_{s_j}(x_i) \sum_{t_j=1}^{k} \psi_{t_j}\, g_{t_i}(y_i), \quad (2)$$

where Γ_j is a function of s_{j-1} and t_{j-1}; N_j is the size of a particular sibship in the j-th generation. By starting at the most recent generation and successively moving up the pedigree with this operator, one can obtain the likelihood for the entire pedigree, provided at $j=1$ $p_{s_{j-1}t_{j-1}s_j}$ is set equal to ψ_{s_j}.

Under different models the p_{stu} will vary as well as the $g_u(x)$. In their paper, Elston and Stewart present these p_{stu} in matrix notation for one autosomal locus, one X-linked locus, multiple linked and unlinked loci, polygenic models and the mixed major gene polygenic model. In the polygenic model the summation signs are replaced by integral signs and the p_{stu} are normal density functions.

2.3.2. Likelihood Computation for Complex Pedigrees Without Loops

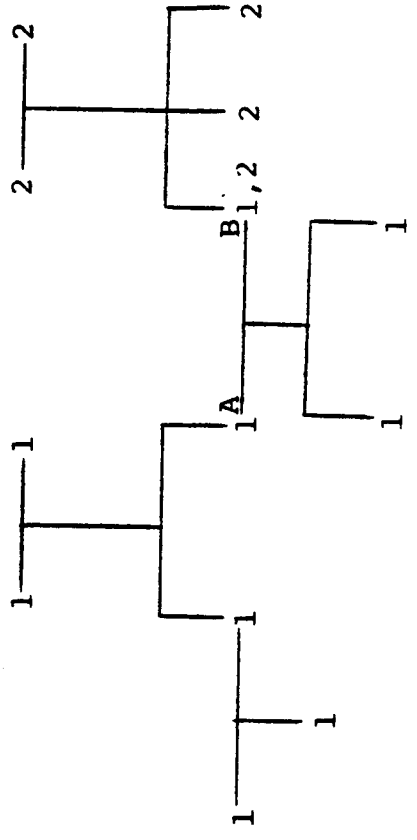
Lange and Elston (1975) discussed the calculation of the likelihood in more detail and with more extensive pedigree structures. They used a graph theoretic approach, where the pedigree is considered to be a finite, connected graph. Their algorithm allows for both multiple birth and consanguineous marriages. It allows for pedigrees where more than a single pair of original parents exist. It does not allow for assortative mating, environmental correlations or any polygenic effects.

Cannings, et al. (1976, 1978) also used a graph to represent the pedigree. They discussed the likelihood calculations of pedigrees when a great deal of consanguineous mating has taken place. Their methods are similar to those of Lange and Elston's work, but they also allow for going down from the original parents when calculating the likelihood.

The extension of the Elston-Stewart algorithm to

FIGURE 2.2

HYPOTHETICAL COMPLEX PEDIGREE



complex pedigrees without loops (i.e. pedigrees with more than one set of original parents) is not difficult, and is most easily demonstrated by an example. Figure 2.2 illustrates this type of pedigree. The mode of calculating the likelihood is to split the pedigree into simple pedigrees. This is done by selecting the individual(s) that link simple pedigrees together; in figure 2 this is either individual A or B. If we treat either of these two individuals (for concreteness, say B) as two separate but genotypically and phenotypically identical individuals we can calculate the likelihood for each simple pedigree (1 and 2) given B is a particular genotype: $L(1|B=v)$, $L(2|B=v)$, where v indexes all possible genotypes. The likelihood of the entire pedigree is then given by

$$L = [\sum_{v} L(1|B=v)\ L(2|B=v)]/g_B(x),$$

where $g_B(x)$ is the likelihood of observing individual B. Cannings, et al., have generalized this procedure so that the complete likelihood can be calculated starting from any arbitrary subset of the original pedigree.

2.3.3. The Single Locus Model

The simplest model using the Elston-Stewart approach is that where a single autosomal locus is assumed to be the only genetic effect and the rest of the variability in the trait of interest can be accounted for by random variability caused by the environment. The model for an individual i

is then

$$x_i = \mu_i + e_i,$$

where x_i is the observed value of the trait, μ_i is the major gene effect and e_i is the effect due to environment. In a two allele system (the simplest system) there are three possible genotypes which we identify by AA, Aa and aa. The distribution of this single gene effect is the population is defined as

	GENOTYPE		
	AA	Aa	aa
Genotype Index v	1	2	3
Frequency	ψ_1	ψ_2	ψ_3
Effect	μ_1	μ_2	μ_3

where μ_v = the mean of genotype v and ψ_v = the genotypic frequency of genotype v, $\sum_{v=1}^{3} \psi_v = 1$. The mean of the distribution is given by

$$E(\mu) = \sum_{v=1}^{3} \psi_v \mu_v.$$

The variance is given by

$$V(\mu) = \sum_{v=1}^{3} \psi_v \mu_v^2 - (E(\mu))^2.$$

If we assume Hardy-Weinberg equilibrium then, $\sqrt{\psi_1} = 1-\sqrt{\psi_3}$. This is a reasonable assumption if it assumed there is random mating, eg. mating choices are made independently of genotype.

The p_{stu} function, as defined by Elston-Stewart is given in Table 2.1. Each entry corresponds to the vector $(p_{st1}\ p_{st2}\ p_{st3})$. The environmental effect is defined to be normally distributed with mean zero and variance σ_e^2. For this model the general likelihood follows from equation (2) of section 2.3.1 and can be written as

$$\Gamma_j = \prod_{i=1}^{N_j} \sum_{s_j=1}^{3} p_{s_{j-1}t_{j-1}s_j} \phi(x_i-\mu_{s_j}, \sigma_e^2) \sum_{t_j=1}^{3} \psi_{t_j} \phi\ (y_i-\mu_{t_j},\ \sigma_e^2)$$

2.3.3.1. Hypothesis Testing Under the Single Locus Model

The approach to testing under this model is different in two major respects from that of the Morton-MacLean approach. One difference is that the Morton-MacLean approach uses the conditional likelihood of a sibship given the phenotype of the parents, whereas the Elston-Stewart approach uses the joint likelihood. The most obvious reason for this difference is that the Morton-MacLean model is for nuclear families, where a conditioned likelihood is not impractical; for pedigrees, a simple answer as to what conditional likelihood is best is not available. A second reason for this difference in approach has to do with sampling and bias considerations as well as philosophical differences between the investigators. Go, et al. (1978) have done simulation studies which suggest that, at least for the single gene models, the unconditional likelihoods are better than the conditional likelihoods for the purposes of estimation when the single gene model is simulated.

TABLE 2.1

TRANSMISSION PROBABILITIES P_{stu} FOR TWO ALLELE AUTOSOMAL LOCUS

s	t		
	1 = AA	2 = Aa	3 = aa
1 = AA	(1 0 0)	(½ ½ 0)	(0 1 0)
2 = Aa	(½ ½ 0)	(¼ ½ ¼)	(0 ½ ½)
3 = aa	(0 1 0)	(0 ½ ½)	(0 0 1)

The second major difference in approach results from the fact that the transmission probabilities p_{stu} are parameters to be estimated in the Elston-Stewart model, while in the Morton-MacLean model they are assumed to be Mendelian. By considering them as parameters in the model, the difficulty of environmentally caused skewness mimicking a major gene effect is removed. Removal of this difficulty is important, as was demonstrated by simulation studies done by MacLean et al. (1975). For their studies, they used sample sizes of at least 200 families with four sibs each. Even if all three effects, major gene, polygenic and common sibship effect, were included in the model and the test of heterogeneity of mating types was included in the analysis, the model was not robust against skewness in the data.

By reparameterizing the p_{stu} values hypothesis testing is made simpler. Define the transmission probabilities

τ_t = probability that an individual with genotype t transmit gene $\underline{A}$ to offspring,

where t = 1, 2 or 3, corresponding to $\underline{AA}$, $\underline{Aa}$, $\underline{aa}$. It follows that $1-\tau_t$ is the probability that an individual with genotype t transmit gene $\underline{a}$ to offspring. If we let s subscript τ of the second parent, it can be shown that the vectors

$$(p_{st1}, p_{st2}, p_{st3}), \text{ and}$$

$$(\tau_t\tau_s, \tau_s(1-\tau_t) + \tau_t(1-\tau_s), (1-\tau_s)(1-\tau_t))$$

are equivalent when there is Mendelian segregation.

The outline for testing the single gene model is given by Elston, et al. (1975). As in the Morton-MacLean approach, the preliminary analysis consists of determining whether a mixture of two normal distributions fits the data significantly better than a single normal distribution using an appropriately chosen transformation. If a mixture fits the data significantly better, then it is appropriate to test the following hypotheses:

1) H_o: $\psi_2 = 2\sqrt{\psi_1\psi_3}$

Rejection of this hypothesis suggests the population is not in H-W equilibrium.

2) H_o: $\tau_1 = 1, \tau_2 = .5, \tau_3 = 0$

Nonsignificant departures from this hypothesis are supportive of Mendelian inheritance.

3) H_o: $\tau_1 = \tau_2 = \tau_3$

Rejection of this hypothesis is supportive of a genetic hypothesis.

4) H_o: $\mu_1 = \mu_2$ or $\mu_2 = \mu_3$

These tests are appropriate when testing for dominance or recessivity.

2.3.4. The Polygenic Model

In the simple polygenic model we hypothesize that the trait of interest is dependent on two independent and additive effects; the polygenic effect g and the random environment effect e. For individual i, the model can be

written as

$$x_i = \mu + g_i + e_i$$

where x_i is the observed trait and μ is the overall mean for the trait. As in the Morton-MacLean model, the polygenic effect can be considered to be normally distributed with mean zero and variance σ_g^2. If we let a_j represent the polygenic effect of the x parent and b_j represent the polygenic effect of the y parent, then the polygenic effect of any of the offspring is distributed normally with mean $(a_j + b_j)/2$ and variance $\sigma_g^2/2$. The environmental effect is defined as it was in the single locus model.

To construct the likelihood, instead of summing over all possible genotypes we integrate over all possible polygenic effects. Equation (2) of section 2.3.1 of the general likelihood can be written for the polygenic model as

$$\Gamma_j = \prod_{i=1}^{N_j} \int_{a_j} \phi(a_j-(a_{j-1}+b_{j-i})/2,\ \sigma_g^2/2)\,\phi(x_i-\mu-a_j,\sigma_e^2)$$
$$\int_{b_j} \phi\ (b_j,\sigma_g^2)\,\phi(y_i-\mu-b_j,\sigma_e^2),$$

where $\int_c$ means that everything to the right of it is to be integrated over c from minus infinity to plus infinity.

This integral can be quickly evaluated by an algorithm presented by Elston and Stewart. For any given set of parameter values the following is true:

$$\int_s \prod_i \phi(s + A_i t + B_i,\ C_i) = \kappa\ \phi(t + \nu,\ \tau^2),$$

where

$$\tau^2 = \left[\Sigma \frac{A_i^2}{C_i} - \left(\Sigma \frac{A_i}{C_i} \right) \left(\Sigma \frac{1}{C_i} \right)^{-1} \right]^{-1}$$

$$\nu = \left[\Sigma \frac{A_i B_i}{C_i} - \left(\Sigma \frac{A_i}{C_i} \right) \left(\Sigma \frac{B_i}{C_i} \right) \left(\Sigma \frac{1}{C_i} \right) -1 \right] \tau^2$$

and

$$\kappa = (2\pi)^{(1-n/2)} \left(\Pi C_i^{-1/2} \right) \left(\Sigma \frac{1}{C_i} \right)^{-1/2} \tau$$

$$\cdot e^{-1/2 \left\{ \Sigma \frac{B_i^2}{C_i} - \left(\Sigma \frac{B_i}{C_i} \right)^2 \left(\Sigma \frac{1}{C_i} \right)^{-1} - \nu^2 \tau - 2 \right\}}$$

2.3.4.1. Extensions to the Polygenic Model

Boyle and Elston (1979) extended the simple polygenic model to include various environmental effects. These effects are the following:

1) an effect due to common sibship in generation j, c_j, distributed $N(0,\sigma_c^2)$

2) an effect due to common nuclear family made up of parents in generation j and their offspring, $c_{j,j+1}$, distributed $N(0,\sigma_n^2)$

3) an effect due to common nuclear family made up of a sibship in generation j and their parents, $c_{j,j-1}$, distributed $N(0,\sigma_n^2)$

4) an effect due to assortative mating, m_j, distributed $N(0,\sigma_m^2)$

5) an effect due to random environment, r distributed $N(0,\sigma_r^2)$.

The model is discussed in detail in the paper; here we will present all but the effect for assortative mating, since it is not easily incorporated into a mixed model.

Again all the effects are considered to be independent and additive. We can partition the random environment effect of the simple polygenic model as

$$e_i = c_j + c_{j,j+1} + c_{j,j-1} + r_i$$

for the i-th individual in the j-th generation. For individuals who have either no sibs, no offspring and/or no parents the corresponding effects are unique. Equation (2) of section 2.3.1 can then be written as

$$\Gamma_j = \int_{c_j} \int_{c_{j,j+1}} \phi(c_j,\sigma_c^2)\phi(c_{j,j+1},\sigma_n^2)$$

$$\prod_{i=1}^{N_j} \int_{a_j} \varphi(a_j-(a_{j-1}+b_{j-1})/2,\ \sigma_g^2/2)\phi(x_i-\mu-a_j-c_j-c_{j,j+1}-c_{j,j-1},\ \sigma_r^2)$$

$$\int_{b_j} \phi(b_j,\sigma_g^2)\phi(y_i-\mu-b_j-c_{j,j+1},\sigma_r^2).$$

2.4. The Mixed Model

At present there is no operational model which includes both a single gene and polygenic effect that can make use of the Elston-Stewart algorithm. What is needed is an algorithm to calculate the likelihood for a general pedigree which incorporates both genetic effects and an effect for common sibship environment, and which is robust against environmentally caused skewness.

Simulation studies suggest the need for such a model. MacLean et al. (1975) found by simulation studies that such a model is robust against parent offspring environmental correlation, assortative mating and sporadic outliers.

Go et al. (1978) found that depending on the details of the null hypothesis, the single gene model was not sufficient when polygenic, environmental correlations within sibships and environmentally caused skewness were simultaneously present in the data.

If we assume all the effects are independent and additive, the mixed model follows directly from the single gene and polygenic models. The simple polygenic model is, as before,

$$x_i = \mu + g_i + e_i.$$

If we let μ vary and subscript it by u(u=1, 2 or 3 corresponding to AA, Aa or aa) and define its distribution as was done in the single gene model,then the mixed model is given by

$$x_i = \mu_u + g_i + e_i.$$

The likelihood for a sibship also follows without difficulty from the simpler models. We must sum the likelihood of the simple polygenic model over all possible effects, μ_u. If we change the subscript u to s_j and t_j, where appropriate, the general likelihood can be written as

$$\Gamma_j = \prod_{i=1}^{N_j} \sum_{s_j=1}^{3} p_{s_{j-1}t_{j-1}s_j}$$

$$\int_{a_j} \phi(a_j-(a_{j-1}+b_{j-1})/2,\ \sigma_g^2/2)\ \phi(x_i-\mu_{s_j}-a_j,\sigma_e^2)$$

$$\int_{b_j} \sum_{t_j=1} \psi_{t_j}\ \phi(b_j,\sigma_g^2)\ \ \phi(y_i-\mu_{t_j}-b_j,\sigma_e^2)$$

If we eliminate the y individuals in this likelihood we have

$$\Gamma_j' = \prod_{i=1}^{N_j} \sum_{s_j=1}^{3} p_{s_{j-1}t_{j-1}s_j}$$

$$\int_{a_j} \phi(a_j-(a_{j-1}+b_{j-1})/2,\ \sigma_g^2/2)\ \phi(x_i-\mu_{s_j}-a_j,\sigma_e^2)\ . \qquad (1)$$

Using the Elston-Stewart algorithm for evaluating integrals described in section 2.3.4 we then have

$$\Gamma_j' = \prod_{i=1}^{N_j} \sum_{s_j=1}^{3} p_{s_{j-1}}\ t_{j-1}s_j\ \kappa_{is_j}\ \phi\ (a_{j-1}+b_{j-1}+\nu,\tau^2) \qquad (2)$$

where

$$\tau = 2(\sigma_g^2 + 2\ \sigma_e^2),$$

$$\nu = 2(\mu_{s_j} - x_i) \text{ and}$$

$$\kappa_{is_j} = 2.$$

It can be seen that for $N_j=1$ the function is a sum of three normal density functions. In general there are 3^{N_j} terms in this expression. As this operator moves up the pedigree this exponential increase in terms still holds. For even moderate sized N_j the function is too large to for our present computer facilities. One possible solution to this problem is to approximate this weighted average of 3^{N_j} terms by a function of 2 or 3 such terms. This approximation will be the subject of Chapter 3.

As with the polygenic model the random environmental variance can be partitioned into various effects due to

common environment. It is possible to include both an effect due to common sibling environment and an effect due to common nuclear family environment. Using the same notation and effects as in section 2.3.4.1. the general likelihood is given by

$$\Gamma_j = \int_{c_j} \int_{c_{j,j+1}} \phi\,(c_j, \sigma_c^2)\ \phi\,(c_{j,j+1}, \sigma_n^2) \prod_{i=1}^{N_j} \sum_{s_j=1}^{3} p_{s_{j-1}t_{j-1}s_j} \quad (3)$$

$$\int_{a_j} \phi\,(a_j - (a_{j-1}+b_{j-1})/2, \sigma_g^2)\ \phi\,(\kappa i - \mu_{s_j} - a_j - c_j - c_{j,j+1} - c_{j,j-1}, \sigma_r^2)$$

$$\int_{b_j} \sum_{t_j=1}^{3} \psi_{t_j}\ \phi\,(y_i - \mu_{t_j} - b_j - c_{j,j+1}\ \sigma_r^2)$$

Again if we eliminate for the sake of simplicity the y individuals, by the use of the Elston-Stewart algorithm for evaluating integrals three times we can simplify the operator as follows:

First integrate out the a_j as before, to obtain

$$\Gamma_j' = \int_{c_j} \int_{c_{j,j+1}} \phi\,(c_j, \sigma_c^2)\ \phi\,(c_{j,j+1}, \sigma_n^2) \cdot$$

$$\prod_{i=1}^{N_j} \sum_{s_j=1}^{3} p_{s_{j-1}t_{j-1}s_j}\ \kappa_{is_j} \cdot \quad (4)$$

$$\phi\,(a_{j-1}+b_{j-1}-c_j-c_{j,j+1}-c_{j,j-1}+\nu, \tau^2)$$

where

$$\tau = 2(\sigma_g^2 + 2\,\sigma_e^2),$$

$$\nu = 2(\mu_{s_j} - x_i) \text{ and}$$

$$\kappa_{is_j} = 2.$$

Now the second and third lines of equation (4) are essentially the sum of 3^{N_j} terms, where each term is the product of N_j $p_{s_{j-1}t_{j-1}s_j}$ factors, N_j κ_{is_j} factors (each equal to 2), and N_j $\phi(a_{j-1}+b_{j-1}-c_j-c_{j,j+1}-c_{j,j-1}+\nu,\tau^2)$ factors. Call a representative of this sum $p_k\kappa_k\phi_k$. We then have

$$\Gamma_j' = \int_{c_j} \int_{c_{j,j+1}} \phi\,(c_j,\sigma_c^2)\ \phi\,(c_{j,j+1},\sigma_n^2)\cdot \sum_{k=1}^{3^{N_j}} p_k\kappa_k\ \phi_k$$

Using the Elston-Stewart algorithm twice on each of these 3^{N_j} terms would reduce Γ_j' to a function of a_{j-1}, b_{j-1} and $c_{j,j-1}$ and numerical constants.

CHAPTER III

THE METHODS FOR APPROXIMATION

In the preceding chapter various genetic models were presented. The operator for calculating the likelihood function for one of the models, the mixed model, without environmental correlations is of the form

$$L = \sum_{i=1}^{3^N} p_i \, \phi \, (x_i, y_i)$$

where N is the number of observations in the sibship. We propose to approximate this function by the function

$$L^* = A \sum_{j=1}^{K} p_j \, \phi \, (x_j, y)$$

where $\sum_{j=1}^{K} p_j = 1$, K = 1, 2 or 3 and A is a scale parameter related to the area under the function.

In section 3.1 methods for estimating parameters of functions such as L* are reviewed. The remaining sections of the chapter will discuss how such methods can be incorporated into the algorithm for calculating the likelihood of a pedigree.

3.1. Mixtures of Normal Distributions

3.1.1. Introduction

The problem of fitting mixtures of distributions, particularly mixtures of 2 normal distributions, has stimulated a great deal of research interest. The interest has centered around the problem of estimating the parameters of such mixtures. Both graphical and numerical techniques have been proposed. Preston (1953) and Bhattacharya (1967) have discussed some of the graphical techniques for fitting mixtures of normal distributions to sample data. The numerical techniques have included maximum likelihood, minimum χ^2, the method of moments, and least squares methodology.

The papers that have been published in this area fall into two basic categories; those that have presented methods for estimation and those that have examined the relative effectiveness of these methods. The work on moment estimates has centered around the original work of Pearson (1894). Rao (1948) and Cohen (1967) presented modifications which simplified Pearson's solution, while Day (1969) generalized the technique to the multivariate problem. Hasselblad (1966) was the first to publish methods of estimating the parameters of a mixture using maximum likelihood techniques. Studies on the relative accuracy of these methods have been done by Robertson and Fryer (1970, 1972), Hosmer (1973a, 1973b), Dick and Bowden (1973), and Tan and Chang (1970). Approaches using least square methodology have been presented by

Bartlett and MacDonald (1968), and Quandt and Ramsey (1978). The least conventional technique has been presented by Gregor (1969); it is an iterative procedure which assumes the number of components in the mixture is not known. By using Fourier theory and a Kolmogorov-Smirnov test, both the number of components and the parameters of each component are estimated.

3.1.2. Methods of Estimating Parameters

Pearson (1894) presented a general theory for determining the parameters of a distribution when the distribution is believed to be a mixture of normal distributions. He did this by equating the known moments of a sample to the corresponding moments of the mixture. By generating as many equations as there are unknowns, one could solve these equations and obtain estimates of the parameter values.

For example, if the distribution were a mixture of two normals, six parameters would need to be estimated: μ_1, μ_2, σ_1^2, σ_2^2, p and q, where μ_1 and μ_2 are the two means, σ_1^2 and σ_2^2 are the two variances and p and q are the corresponding proportions of the component distributions. Therefore six equations would need to be solved, or five moment equations would be necessary, since $p + q = 1$ would be one of the equations. Pearson's solution to these six equations requires the solution of a nonic equation. From each real root of this nonic there is a possible solution. Cohen (1967) rederived Pearson's method. The following

discussion is a brief sketch of this derivation.

Letting m_1 and m_2 represent the distances from the subpopulation mean to the grand mean ($m_1<0<m_2$), σ_1^2 and σ_2^2 represent the corresponding population variances, p represents the weight of population 1 and ν_i represent the ith central moment, the five equations to be solved are

$$\begin{aligned}
pm_1 + (1-p)m_2 &= 0 \\
p[\sigma_1^2 + m_1^2 - \nu_2] + (1-p)[\sigma_2^2 + m_2^2 - \nu_2] &= 0 \\
p[3\sigma_1^2 m_1 + m_1^3 - \nu_3] + (1-p)[3\sigma_2^2{}_2 + m_2^3 - \nu_3] &= 0 \qquad (1) \\
p[3\sigma_1^4 + 6m_1^2\sigma_1^2 + m_1^4 - \nu_4] + (1-p)[3\sigma_2^4 + 6m_2^2\sigma_2^2 + m_2^4 - \nu_4] &= 0 \\
p[15\sigma_1^4 m_1 + 10\sigma_1^2 m_1^3 + m_1^5 - \nu_5] + (1-p)[15\sigma_2^4 m_2 + 10\sigma_2^2 m_2^3 + m_2^5 - \nu_5] &= 0
\end{aligned}$$

The first step is to estimate p from the equations to reduce the number of equations to four. With some algebraic manipulations, introduction of the sample cumulants, k_i, and the substitution

$$\begin{aligned}
\beta &= (\sigma_1^2 - \nu_2 + m_1^2)/m_1 \\
\beta &= (\sigma_2^2 - \nu_2 + m_2^2)/m_2, \qquad (2)
\end{aligned}$$

the system of equation can be reduced to three.

With more algebra and the substitutions

$$R = m_1 + m_2$$

$$V = m_1 + m_2$$

$$W = RV \quad \text{and}$$

$$Z = W + \nu_3$$

a ninth degree polynomial in V is obtained with coefficients a_i, where i corresponds to the degree of the term, given

as follows

$$a_9 = 24, \qquad a_4 = 444k_4\nu_3^2 - 18k_5^2,$$
$$a_8 = 0, \qquad a_3 = 228\nu_3 - 108\nu_3 k_4 k_5 + 27k_4^3,$$
$$a_7 = 84k_4, \qquad a_2 = -(63\nu_3^2 k_4^2 + 72\nu_3^3 k_5),$$
$$a_6 = 36\nu_3^2, \qquad a_1 = -96\nu_3^4 k_4,$$
$$a_5 = 90k_4^2 + 72k_5\nu_3 \qquad a_0 = -24\nu_3^6. \qquad (4)$$

There is at least one negative real root of this polynomial, with each negative real root providing a potential solution. If v* is a real root of the nonic, then an estimator of R, say r*, can be obtained from the equation

$$r^* = \frac{-8\nu_3 v^{*3} + 3k_5 v^{*2} + 6\nu_3 k_4 v^* + 2\nu_3^3}{v^*(2v^{*3} + 3k_4 v^* + 4\nu_3^2)}, \qquad (5)$$

which follows from the earlier transformations. Estimates of m_1 and m_2, say m_1^* and m_2^* can be obtained from r* and v*. This relationship can be expressed by the quadratic equation

$$M^2 - r^*M + v^* = 0; \qquad (6)$$

m_1^* and m_2^* are the roots of this equation. This equation need not have real roots, so even though a real root from the nonic exists there is no guarantee of real estimates for m_1 and m_2. (Clark, 1978).

The estimates of the parameters are then given by

$$\hat{\mu}_1 = m_1^* + \bar{x}$$
$$\hat{\mu}_2 = m_2^* + \bar{x}$$
$$\hat{p} = m_2^*/(m_2^* - m_1^*)$$

$$\begin{aligned} \hat{\sigma}_1^2 &= m_1^*(2r^* - \nu_3/v^*)/3 + \nu_2 - m_1^{*2} \\ \hat{\sigma}_1^2 &= m_2^*(2r^* - \nu_3/v^*)/3 + \nu_2 - m_2^* \end{aligned} \qquad (7)$$

When more than one real root of the nonic equation is possible, Pearson suggested using the solution which gives the estimates of the sixth moment closest to the sample sixth moment; others have suggested using a χ^2 goodness of fit test and then using the solution which gives the best fit.

Rao (1948) has given a simpler solution under the assumption that the two subpopulations have equal variances. There are then only 4 equations to be solved and the fifth equation in (1) is not necessary. The substitution of

$$\begin{aligned} R &= m_1 + m_2 \quad \text{and} \\ V &= m_1 m_2 = \sigma^2 - \nu_2 \end{aligned} \qquad (8)$$

can reduce the system of equations to the cubic equation

$$2V^3 + k_4 V + \nu_3^2 = 0. \qquad (9)$$

There is one negative root, which is the required solution v^*. The estimate r^* can then be obtained from its relationship to v^* given by

$$r^* = -\nu_3/v^*. \qquad (10)$$

With r^* and v^*, the estimates of the parameters are obtained as before in the general case. Since there can be only one suitable root from the cubic equation (9), the estimates of the parameters are unique.

The accuracy and bias of the moment estimators have been discussed by Robertson and Fryer (1970). They have presented a method for calculating the bias and variance of the moment estimates to order n^{-2}. In cases where more than one real solution is possible their methods supply an over-estimate of the true variance.

For a mixture of three normal distributions, 9 parameters must be estimated; therefore equations involving the first eight moments are required. Some simplifying assumptions can reduce the number of equations to be solved. For example, if the components of the mixture are assumed to have equal variances then only 7 parameters need be estimated.

Hasselblad (1966) discussed the estimation of the parameters for k (k > 1) subpopulations when the data are grouped. He used maximum likelihood methods with the areas under the curve being approximated by the product of class width times the mid-interval ordinates. Let p_j be the weight of the j-th subpopulation (j = 1, k), μ_j and σ_j^2 be the corresponding momements of the subpopulation, and f_i be the number of observations in the i-th interval (i = 1,..N) where N is the number of intervals for which the data are groups. Using the mid-interval ordinate

$$q_{ij} = \frac{1}{\sqrt{2\pi\sigma_j^2}} \quad \exp\left[-\frac{1}{2}(x_i - \mu_j)^2/\sigma_j^2\right]$$

and letting

$$Q_i = \sum_{i=1}^{k} q_{ij} p_j ,$$

an approximation of the log likelihood function is given by

$$L = \sum_{i=1}^{N} f_i \ln(Q_i) .$$

The estimates of the unknown parameters which maximize this function are the approximate ML estimates. He has discussed the general case as well as the case when the variances of each subpopulation are assumed equal. The likelihood equations are solved using either the Newton-Raphson algorithm or the method of steepest ascent. He has presented results for the theoretical asymptotic variance of these maximum likelihood estimates of the parameters of a mixture of three normals. The results suggest that the estimation of parameters in such a problem would be difficult when the means of the subpopulations are separated by less than 2 standard deviations.

Fryer and Robertson (1972) have compared the Pearson estimates with those of Hasselblad, as well as with the mimimum χ^2 estimates. They have compared the methods on nine populations of two component mixtures. They presented no conclusive evidence for the superiority of any method, but it was suggested that all the methods benefited from a large separation of the components. In a population where the subpopulations had equal variances, all the estimates derived

under the assumption of equal variance were much more accurate than those estimates which were made without this assumption.

The estimation problem for two component distributions when each component is a multinormal has been discussed by Day (1969). He has presented methods for obtaining both the moment and maximum likelihood estimators. He has presented the results of simulation studies which indicate that in all but the univariate case the maximum likelihood estimates are superior to the moment estimates. Day did not publish the relevant results, but did indicate that in the univariate case the minimum χ^2 estimates are also satisfactory. He also discussed problems with the maximum likelihood method. It seems that there often will be a number of local maxima, therefore the search must be done from various points on the surface to insure that the supremum is obtained. There is also the the problem that there is a singularity associated with each sample point. Murphy and Bolling (1967) have demonstrated this problem. Suppose the population is a mixture of two normal densities with corresponding weights p_1 and p_2. The likelihood for random sample of size n can then be written as

$$L = \prod_{i=1}^{n} [p_1\phi_1(x_i-\mu_1,\sigma_1^2) + p_2\phi_2(x_i-\mu_2,\sigma_2^2)].$$

Without loss of generality, factor out the first factor of the product

$$L = [p_1\phi_1(x_2-\mu_1,\sigma_1^2) + p_2\phi_2(x_1-\mu_2,\sigma_2^2)] \cdot$$

$$\prod_{i=2}^{n} [p_1\phi_1(x_i-\mu_1,\sigma_1^2) + p_2\phi_2(x_i\mu_2,\sigma_2^2)],$$

$$\geq p_1\phi_1(x_i-\mu_1,\sigma_1^2) \prod_{i=2}^{n} (p_1\phi_1(x_i-\mu_1,\sigma_1^2) + p_2\phi_2(x_i-\mu_2,\sigma_2^2)).$$

Now

$$\phi_1(x_i-\mu_1,\sigma_1^2) = \sqrt{2\pi\sigma_1^2}^{-1} \exp\{-\frac{1}{2}(x_i - \mu_1)^2/\sigma_1^2\}.$$

If we set $x_i = \mu_1$ then as σ_1^2 decreases the entire likelihood will increase without limit. A possible solution to this problem is to restrict the values of σ^2. Bryant (1978) mentioned that it appears from practice that as long as $np \geq 1$, where n is the sample size and p is the weight of the smaller component, the problem of singularities should not cause difficulties.

Hosmer (1973a, 1973b) and Dick and Bowden (1973) have presented Monte Carlo studies for mixtures of two distributions, with unequal variances, for small and moderate size samples using maximum likelihood estimates. These studies suggest that large sample size or large separation of the components is required for reliable estimates. In lieu of these conditions, independent information about one of the components, or other information on the mixture, can improve the reliability of the estimates.

The relative efficiency of the moment estimators to those of maximum likelihood for mixtures of two normals with common variance has been calculated by Tan and Chang (1972). They have done this by calculating the asymptotic variance

matrix of the moment estimator and then comparing this to the information matrix of the ML estimators. They found the relative efficiency was over 75% when the 2 subpopulations are of equal size and are separated by 1 or more standard deviations. The efficiency decreases as the difference in the subpopulations size grows larger.

Quandt and Ramsey (1978) have used least squares to solve for estimates of the parameters of a mixture. Their estimates are derived by minimizing the sum of the squared differences between the sample moment generating function and the theoretical moment generating function. The moment generating function of the mixture of k normal density functions is given by

$$G(v,t) = \sum_{i=1}^{k} p_i \exp(\mu_i t + \tfrac{1}{2}\sigma_i^2 t^2),$$

where v represents the unknown parameters. Given a sample $x_1,\ldots,x_n$, we can estimate the expected value of $e^{t_j x}$, where t_j is an arbitrary value of t, by

$$y_j = \frac{1}{n} \sum_{i=1}^{n} \exp\, t_j x_i.$$

If we choose m such t_j, we then want to minimize

$$S(v,t) = \sum_{j=1}^{m} (y_j - G(v,t_j))^2.$$

The t_j should be chosen on some small interval around zero and the value of m should be at least as great as the number of unknown parameters. The values of m and t_j must be chosen so that the solution of the equations is nonsingular for almost

all sequences of $\{x_n\}$ and so that the computations are not intractable. Values of t_j too close to zero or too large can cause difficulty in computation. The estimates derived from this method are asymptotically normal and consistent. Quandt and Ramsey compared the moment estimates with these estimates, using simulation on seven different populations, and found these estimates superior. Hosmer (1978) has compared this method by simulation to the maximum likelihood method and found the moment generating function method superior to it, at least for small samples. Quandt and Ramsey mentioned two problems with their method. One is a problem of numerical convergence; the second is that, as yet, the optimum values of t_j for evaluating the moment generation function are not known and the variance of the estimates are dependent on these values. Clark and Heathcote (1978) suggest this problem could be avoided by instead of evaluating $S(v,t)$ at discrete values of t, treating $S(v,t)$ as a continuous function of t. If t were continuous it would be possible to integrate it out of the expression.

Paulson, et al., (1975) and Heathcote (1977) have discussed using the characteristic function for the general problem of estimation of parameters, and this method parallels the moment generating function method.

An earlier form of least squares estimates was proposed by Bartlett and MacDonald (1968). Their solution minimized the integral

$$\int (dF_s - dF)^2/dG,$$

where F_s is the empirical cumulative distribution function and F is the theoretical distribution function. dG is an appropriately chosen weighting function; if it were discrete the integration would be replaced by summation.

3.1.3. Summary

Although a great deal of interest has been shown in the various methods of parameters estimation there appears to be no 'best' method of fitting mixtures of normal distributions. The moment estimators are the simplest conceptually, but they do not enjoy the asymptotic properties of the maximum likelihood estimators. Estimators using least squares methodology may prove to be suitable estimators, but litte work has been done in determining how well they compare with the other modes of estimation. Most of the work has been done with mixtures of two distributions, and very little with three; our primary interest lies with mixtures of three distributions.

3.2. Methods of Approximation

From section 2.4 (equation (2)) we can see that the function we want to approximate for the simple mixed model, without sibling environmental correlation and without the y individuals, is given by

$$\Gamma'_{j+1} = \prod_{i=1}^{N_j} \sum_{s_{j+1}=1}^{3} p_{s_j t_j s_{j+1}} \kappa_{i s_{j+1}} \phi(a_j + b_j + \nu_{i s_{j+1}}, \tau^2_i, s_{j+1})$$

where p, κ, ν, τ^2 are numbers which are functions of the initial parameter estimates and the observations. The symbols

a_j and b_j are dummy variables of integration which are integrated out in subsequent steps in the likelihood computation procedure. If we let $\underset{\sim}{\theta}$ represent the values of p, κ, and τ^2 then we can write

$$\Gamma'_{j+1} = f(a_j+b_j, \underset{\sim}{\theta}).$$

This function is a mixture of 3^N normal distributions, although the area under the function is no longer equal to 1. We wish to approximate this function by a mixture of no more than three normal distributions.

The function we wish to fit is given by

$$g(a_j+b_j, \underset{\sim}{\theta}^*) = A \sum_{i=1}^{K} p_i \phi (a_j+b_j-\mu_i, \sigma^2)$$

where A, p_i, μ_i and σ^2 are the parameters to be estimated and are represented by $\underset{\sim}{\theta}^*$ and K = 1, 2 or 3. To do this we can evaluate the function $f(a_j + b_j, \underset{\sim}{\theta})$ at n points and using one of the methods described earlier find the function $g(a_j + b_j, \underset{\sim}{\theta}^*)$ which best fit these n points. We then have

$$\Gamma'_{j+1} \tilde{=} g(a_j+b_j, \underset{\sim}{\theta}^*)$$

This function is then the function which is used in the recursive calculation of the entire likelihood. In the recursive calculation an approximation is made each time the number of terms becomes unmanageable, i.e. each time six to eight individuals are incorporated into the function.

For a nuclear family we know from section 2.3.1 that the joint likelihood is given by $\Gamma_1(\Gamma'_2)$. Therefore the

approximation for a nuclear family is given by

$$\int_{a_1} \sum_{s_1=1}^{3} \psi_{s_1} \phi(a_1, \sigma_g^2) \phi(a_1 + \mu_{s_1} - x_1, \sigma_e^2) \cdot$$

$$\int_{b_1} \sum_{t_1=1}^{3} \psi_{t_1} \phi(b_1, \sigma_g^2) \phi(b_1 + \mu_{t_1} - y_1, \sigma_e^2) \cdot$$

$$g(a_1+b_1, \underset{\sim}{\Theta}^*).$$

Since $g(a_1+b_1, \underset{\sim}{\Theta}^*)$ is a sum of K normal distributions we can integrate out a_1 and b_1 analytically as before and the entire likelihood is then the sum of $9 \cdot K$ terms.

We can treat the sum $a_j + b_j$ as a random variable with mean equal to zero and variance equal to $2\sigma_g^2$. We are interested in evaluating $f(a_j+b_j, \underset{\sim}{\Theta})$ on the interval $\pm d\sqrt{2\sigma_g^2}$ where d is the number of standard deviations we choose and its value is dependent on the particular set of data being used. We select n (assume n is odd) equally spaced values on this interval. As an examples assume d = 1 and n = 5, we then evaluate $f(a_j+b_j, \underset{\sim}{\Theta})$ at the values

$(a_j+b_j)_1 = -\sqrt{2\sigma_g^2}$, $(a_j+b_j)_2 = -.5\sqrt{2\sigma_g^2}$, $(a_j+b_j)_3 = 0$, $(a_j+b_j)_4 = .5\sqrt{2\sigma_g^2}$, $(a_j+b_j)_5 = \sqrt{2\sigma_g^2}$.

From this framework the general methods for estimating parameters of mixtures discussed earlier can be applied to the estimation of the parameters p_i, μ_i and σ^2. Since we are not dealing with distributions, the parameter A, the

area under the function $f(a_j + b_j, \underset{\sim}{\Theta})$ must also be estimated. This is done by the use of Simpson's rule, integrating over the interval $\pm d\sqrt{2\sigma_g^2}$. It is important to note that although the methods discussed in the following section are valid, the associated statistical properties are not necessarily valid since we are not dealing with sample data.

3.2.1. Method of Moments

As was discussed earlier the method of moments equates the sample moments to the moments of the distribution whose parameters we wish to estimate. If we treat the $(a_j+b_j)_i$ as sample values and $f(a_j+b_j)_i, \underset{\sim}{\Theta})$ as corresponding frequencies, we can estimate the sample moments by the usual methods for grouped data. If we assume the approximating function is a mixture of two normal with common variance, Rao's method of estimation which was described earlier is suitable. Moment estimates are not practical for mixtures of three normals. There is apparently no simple solution to the six simultaneous equations which need to be solved for such a problem.

3.2.2. Method of Maximum Likelihood

The methods of maximum likelihood can also be used to arrive at estimates of the approximating mixture. The methods used in this problem are identical to those of maximum likelihood; the properties of these estimates are not those of maximum likelihood. Analogous to the log-likelihood for grouped data we can minimize the function

$$Q = \sum_{i=1}^{n} f((a_j+b_j)_i, \underset{\sim}{\Theta}) \ln g((a_j+b_j)_i, \underset{\sim}{\Theta}^*).$$

The function can be maximized by using programs such as MAXLIK (Kaplan and Elston, 1978).

3.2.3. Method of Least Squares

For this method we wish to estimate the parameters that minimize the sum of squared differences between the function Γ_{j+1} and the approximating mixture; using the same notation as before we wish to minimize

$$\sum_{i=1}^{n} (f((a_j+b_j)_i, \underset{\sim}{\Theta}) - g((a_j+b_j)_i, \underset{\sim}{\Theta}^*))^2.$$

Again we can use the program MAXLIK to estimate the value of $\underset{\sim}{\Theta}^*$ which minimizes this function.

3.2.4. Method Using Moment Generating Function

A second approach using least squares is to minimize the sum of squared differences between the moment generating function of Γ_{j+1} and the moment generating function of the approximating mixture. An estimate of the moment generating function of Γ_{j+1} is given by

$$y(t_s) = \frac{\sum_{i=1}^{n} f((a_j+b_j)_i, \underset{\sim}{\Theta}) \cdot \exp(a_j+b_j)_i \, t_s}{\sum_{i=1}^{n} f((a_j+b_j)_i, \underset{\sim}{\Theta})}$$

The moment generating function of the approximating mixture

is given by

$$h(t_s) = \sum_{\ell=1}^{K} p_\ell \exp \mu_\ell t_s + \frac{1}{2}\sigma^2 t_s^2,$$

where $\sum_{\ell=1}^{K} p_\ell = 1$ and p_ℓ, μ_ℓ and σ^2 are the parameters to be estimated.

The least squares estimates are then given by the values of p_ℓ, μ_ℓ and σ^2 which minimize

$$\sum_{s=1}^{R} (y(t_s) = h(t_s))^2,$$

where $R \geq$ the number of parameters to be estimated. The values of t_s are chosen on the interval (-1,1) and only trial and error can tell which values are best. Again this function can be minimized by using a program such as MAXLIK.

3.2.5. Method of Fitting a Single Distribution

Although it is expected that more than two parameters will be required to obtain accurate estimates, we wish to investigate the simplest case as well. The estimates of the two parameters of a single normal distribution are obtained by the analogs of the moment estimators for grouped data. They are given as follows:

Let

$$S = \sum_{i=1}^{n} f((a_j+b_j)_i, \underset{\sim}{\Theta})$$

$$S_1 = \sum_{i=1}^{n} f((a_j+b_j)_i, \underset{\sim}{\Theta}) \ (a_j+b_j)_i$$

and

$$S_2 = \sum_{i=1}^{n} f((a_j+b_j)_i, \underset{\sim}{\theta}) \ (a_j+b_j)_i^2$$

then

$$\hat{\mu} = S_1/S \quad \text{and}$$

$$\hat{\sigma}^2 = S \cdot S_2 - S_1^2/S^2.$$

CHAPTER IV

EMPIRICAL STUDY OF THE APPROXIMATION METHODS

In this chapter the various methods of approximation defined in Chapter III are compared. Both simulated and real data are used to make comparisons on the accuracy of the approximations. The real data is a subset of a larger dataset; the segregation analysis of the entire dataset is discussed in Chapter V. The following section will describe the simulated data. The subsequent sections will describe the actual methods studied, the means of comparing these methods and the results of these comparisons

4.1. Description of the Data

To compare the previously defined methods of approximation to the exact likelihood, as well as to each other, we have chosen to simulate data from three genetic models. The models in the Morton-MacLean notation are the following:

Model	d	t	σ^2_G
1	1	1	1
2	1	1	.5
3	.5	3	1

For all models $z = 2.0$, $q = .75$ and $\sigma_E^2 = 1$.

Each of these models was simulated on four nuclear families of eight members each (see Table 4.1). The reason for this choice of size was that eight individuals is the largest size family that can be effectively handled by the program calculating the exact likelihood. Only two generational data is used because, at present, the program for calculating the approximation is written for this type of data. There is no reason to suspect that more generations would adversely affect the approximation, since the functions are structurally the same. When a more general program is written, it would be set up to make an approximation each time six to eight individuals are incorporated into the recursive calculation, not necessarily for each sibship. The generalization of the program to more than two generations is not inherently difficult, merely time consuming, and not necessary for this study.

The parameter values for which the likelihoods are evaulated were selected to give a wide spread in the value of the likelihood. By choosing values for which the likelihood indicated poor as well as good fits, the robustness of the approximations can be studied. Unfortunately not all the methods were evaluated at all the parameter values, or for all the families, but it is a fact of life that neither computer time nor resources are unlimited.

TABLE 4.1

SIMULATED DATA FOR COMPARING APPROXIMATION METHODS

Model	Family	Parents		Sibs					
		1	2	1	2	3	4	5	6
1	1	2.798	3.171	-1.203	-.885	3.173	3.321	3.042	2.494
	2	2.822	3.747	-.226	-.237	2.425	3.476	3.469	4.213
	3	2.392	2.770	.370	-.799	3.195	3.635	3.388	3.038
	4	2.012	2.224	-.269	-.712	2.503	1.414	1.935	2.384
2	1	.242	3.693	.359	-1.727	1.494	2.844	3.617	.507
	2	4.221	2.701	.435	.782	3.260	3.784	5.134	2.458
	3	3.618	1.777	-1.056	1.241	2.731	.921	.561	2.474
	4	3.599	1.799	-.940	.832	2.522	2.013	4.990	3.905
3	1	1.757	1.715	-.762	-.025	2.085	.894	1.585	2.686
	2	2.351	1.364	-.213	.021	1.593	1.784	1.567	2.071
	3	2.467	1.616	.000	-.365	1.355	1.849	1.793	.947
	4	2.514	1.856	-.404	.029	1.583	1.465	2.466	2.056

4.2. Description of Methods of Approximation

The estimation of the single distribution was done two ways. The first way (SD1) was to evaluate the function $f(a_j+b_j,\underset{\sim}{\theta})$ over the interval $\pm$ 4 σ_G at 33 equally spaced points; the second method (SD2) was to evaluate the function over the interval $\pm$ 5 σ_G at 201 equally spaced points.

Rao's method of moments (MM) was used to obtain parameter estimates for the mixture of two distributions which fit the moment equations best. The $(a_j+b_j)_i$ were selected on interval $\pm$ 4 σ_G at 33 equally spaced points. The cubic equation requiring solution was solved using the program MAXLIK. More efficient subroutines exist for solving such equations, but the MAXLIK subroutine was more easily incorporated into the present program.

Both mixtures of two and three distributions were estimated using maximum likelihood and least squares techniques. The abbreviations for these methods are ML2, ML3, LS2 and LS3. For all four of the methods the function $(f(a_j+b_j,\underset{\sim}{\theta})$ was evaluated on the interval $\pm$ 4 σ_G at 33 equally spaced points.

For the moment generating function method both mixtures of two and three distributions were also estimated on the interval $\pm$ 4 σ_G at 33 equally spaced points (MG2 and MG3). For this method it is also necessary to choose values of t_s as discussed in section 3.2.4. The values selected were -.75, -.5, -.25, .1, .25, .5 and .75. Other values can be selected and investigated but this did not seem fruitful.

As was discussed earlier there are, at present, no criteria published for selecting good values. Rather than attempting to find better values with no way of knowing where to begin the search, the decision was made to use only these values of t_s in the evaluation here.

4.3. Methods of Comparison

To illustrate the accuracy of the approximations a number of techniques are used. The primary data are the pairs of values (x_i, y_i) where x_i is the value of the exact negative log likelihood at the set of parameters i and y_i is the corresponding value for the approximation. The quantity of interest is the difference between the two values, $d_i = x_i - y_i$. Tables 4.2-4.6 give the descriptive statistics for these d_i values for each of the five proposed methods. Table 4.7 summarizes the information on all the methods. The median, along with the mean, is given because it is probably a better estimator of the center of the accuracy distribution. This is because in general the approximations appear to behave consistently but there are enough outliers to warrant the more robust estimator. A t-test of $\bar{d}=0$ can be construed as a test of overall bias of the approximation method. The correlation given in the tables is indicative of bias as a function of the fit of the parameters.

The suitability of the approximations at different values of the log likelihood is demonstrated by a series of

plots (figures 4.1-4.5). These plots are the plots of the (x_i,y_i) pairs for each approximation method. The solid line in the plots is that of the line y=x. Obviously the closer the individual points are to this line the better the approximation is. Figure 4.6 is a representation of the distribution of the d_i values for all the methods. For each method the 2.5, 5.0, 16.7, 50, 66.7, 95, and 97.5 percentiles are plotted. By viewing this figure it is possible to compare the distribution of the accuracy of the various proposed methods.

The calculations for the approximation were all done in double precision accuracy. For the purposes of the comparison, however, each log likelihood was rounded to four decimal places. Therefore for any (x_i,y_i) pair for which x_i is identical to y_i to four decimal places the values are considered to be equal.

The computer time required can only be discussed in relative terms. At present the programs are not set up to be the most efficient time wise. Numerous things can be done to make the program more efficient, since there are many checks and I/O options present which were required for this study but need not be included in a final program.

4.4. Results of Comparisons on Simulated Data

The first methods used in approximating the exact likelihood were the methods which fit single distributions, SD1 and SD2. The SD1 method was used on all three models,

while the SD2 method was used on only the first two models. Figures 4.1a-e illustrate that these methods generally work very well. There is no apparent relationship between the accuracy and the magnitude of the log likelihood. This assertion is supported by the statistics in Table 4.2. There is a nonsignificant overall correlation between d_i values (accuracy) and the exact log likelihood for either SD1 or SD2. A t test of $\bar{d}_i=0$ can be construed as a test of whether the methods are biased. Neither method is found to be significantly biased by this measure. Neither method appears to be obviously superior to the other method numerically. As would be expected, the CPU time is about five times faster for SD1 than for SD2 for this size family. The SD1 method took approximately .5 seconds of CPU time, SD2 about 2.5 seconds, and the exact method 1.5 seconds. This improvement of SD1 over the exact method for such small datasets is important and should be noted.

Models 1 and 2 were used to compare the approximation method MM to the exact value. From figures 4.2a-b it appears the method performs well. There is no apparent relationship between the accuracy and the magnitude of the log likelihood, the correlation being a nonsignificant .21 (see Table 4.3). There is also no significant bias toward either under or over estimating the true log likelihood. This method is extremely quick, as a result of the fact that only the solution of the cubic equation requires an iterative process and this equation

has only one suitable root. In CPU time it is twice as fast as the exact calculation. One difficulty with the method is that it always estimates a mixture of two distributions. The flexibility of fitting one or three distributions, which is possible with the other iterative approximation methods, is not available.

Comparisons between the ML2 method and the exact method were done on the data from all three simulated models; comparisons for the ML3 method were done on only the first two datasets. Both methods, as demonstrated in figures 4.3a-e did a very good job of approximating the likelihood functions. Neither method exhibited any form of bias (see Table 4.4). The ML2 method has approximately the same variance as the SD1, SD2 and MM methods. The ML3 method had more than a order of magnitude smaller variance than these methods. This reduction in variance is a strong vote for the efficacy of the ML3 method. Sixty-one likelihoods were estimated by both the ML2 and ML3 methods. Of these, sixty-one commonly estimated sets, the ML3 method gave identical approximations to the ML2 method four times and fit the likelihood better 36 times which is a significant (α=.05) improvement. As expected, the ML3 method is quite a bit slower than the ML2 method or the other methods mentioned previously. It is twice as slow as the ML2 method, which is nearly four times as slow as the SD1 method. For both methods

the program was set up to allow up to twenty iterations. It is felt even if convergence was not obtained in twenty iterations, the approximation at that point would be sufficiently accurate to represent the true function. This censorship did not prove to be needed with the ML2 method. With the ML3 method, approximately 25% of the approximations had not converged in twenty iterations. This may be too high, but the overall improvement in accuracy suggests censorship at twenty iterations is not deleterious to accuracy.

The accuracy of the least squares method (LS2 and LS3) was generally not as good as any of the previously mentioned methods. This can be seen in figures 4.4a-e where there are more points off the line of equality then have been seen previously. In figure 4.4b there is a point which is the worst fitting approximation for all the methods. The reason for failure at this point is not clear. Using a t-test on the hypothesis $\bar{d}_i=0$ (Table 4.5) suggests there is no bias in either method, but this is not true: given the great variability in the accuracy, bias was not apparent with this approach. Of the fifty likelihoods that were estimated by both the LS2 and LS3 methods, the LS2 method overestimated the true value thirty seven times while the LS3 method overestimated it thirty eight times. Both these figures are significantly higher than would be expected if there were no bias ($\alpha=.05$). A good explanation for this general bias is not apparent; a possibility is that the method of

unweighted least squares was used. The method weights the squared differences between the fitted function and the actual function at the tails equally with those at the center of the domain of the function; this equal emphasis at the tails may cause the distortion.

The MG2 and MG3 methods were studied on data from Models 1 and 2. Although both the MG2 and MG3 methods are quicker than the corresponding methods using maximum likelihood, neither method was as accurate. As can be seen from Table 4.7 the standard deviation of the MG2 methods was twice that of the ML2 method, and that of the MG3 method was an order of magnitude greater than that of the ML3 method. (See figures 4.5a-e). The MG3 method offered no improvement over the MG2 method in terms of accuracy either. Of the sixty-four commonly fitted likelihoods the MG2 method actually fitted the data better on thirty nine of the fits. For both the MG2 and MG3 methods there was no significant bias or correlation with the exact log likelihood (see Table 4.6). The reason for the relative inaccuracy is probably the same reason which makes these methods relatively quick. As presented here, the methods required the solution of a least squares equation with only seven points. Computationally this should be a very fast task but the paucity of points does not lead to a good overall fit. Quandt and Ramsey (1978), who presented this method, only used as many points as there were parameters to be estimated. This may be

adequate with great discrimination between the mixed distributions, but it does not appear to be sufficient here.

4.5. Summary of Results from Simulated Data

In general none of the proposed methods of approximation were abject failures and they all could be used in a likelihood computation algorithm to estimate likelihoods. The method of fitting a single distribution (SD1) and the maximum likelihood methods seem to offer the best combination for likelihood evaluation. The quickness, relative accuracy and lack of bias speak well for the SD1 method. The apparent improvement in accuracy with the maximum likelihood methods makes these methods attractive. The ML methods are the slowest, however. Given the tradeoffs between the two types of approximation, a strategy for likelihood evaluation can be proposed. When the likelihood surface for a given pedigree is initially being investigated, the SD1 method may serve as a good exploratory tool at either single points on the surface or within an iterative scheme. Once reasonably good estimates are found, the ML methods could be used to improve the accuracy until final estimates are reached. With this combination it should be possible to have a relatively fast and accurate method for evaluating likelihoods under mixed genetic models.

TABLE 4.2

DISTRIBUTION OF ACCURACY (d_i) FOR SD1 AND SD2 METHODS

Model	N	Mean	Standard Deviation	Median	Minimum	Maximum	Minimum $\lvert d_i \rvert$	Correlation of d_i With Exact Log Likelihood
SD1								
1	64	-.0029	.1150	-.0019	-.2123	.6010	.0000	.39
2	64	-.0014	.1049	-.0105	-.4355	.3606	.0006	-.21
3	28	-.0174	.0538	-.0046	-.2843	.0103	.0000	-.71*
OVERALL	156	-.0037	.1123	-.0052	-.4355	.601	.0000	.18
SD2								
1	40	-.0212	.0707	-.0006	-.2350	.1217	.0000	-.48*
2	36	-.0400	.1747	-.0087	-.6609	.3249	.0003	.16
OVERALL	76	-.0302	.1302	-.0048	-.6609	.3249	.0000	-.05

*Significantly different from zero ($\alpha = .05$)

TABLE 4.3

DISTRIBUTION OF ACCURACY (d_i) FOR MM METHOD

Model	N	Mean	Standard Deviation	Median	Minimum	Maximum	Minimum $\lvert d_i \rvert$	Correlation of d_i With Exact Log Likelihood
1	51	-.0016	.1009	.0090	-.2074	.5706	.0003	.18
2	24	.0607	.1924	-.0006	-.5000	.4225	.0096	.01
OVERALL	75	.0115	.1398	-.0090	-.5000	.5706	.0003	.21

TABLE 4.4

DISTRIBUTION OF ACCURACY (d_i) FOR ML2 AND ML3 METHODS

Model	N	Mean	Standard Deviation	Median	Minimum	Maximum	Minimum $\lvert d_i \rvert$	Correlation of d_i With Exact Log Likelihood
ML2								
1	52	.0034	.0275	.0022	-.1212	.0755	.0000	.01
2	36	-.0014	.1506	.0011	-.5067	.3602	.0003	.17
3	32	-.0378	.0664	-.0120	-.2839	.0016	.0000	-.42*
OVERALL	120	-.0090	.0919	-.0007	-.5067	.3602	.0000	.05
ML3								
1	37	.0095	.0185	.0027	-.0153	.088	.0005	.19
2	24	.0072	.0300	.0063	-.0709	.1106	.0001	.01
OVERALL	61	.0086	.0235	.0430	-.0709	.1106	.0001	.05

*Significantly different from zero ($\alpha = .05$)

TABLE 4.5

DISTRIBUTION OF ACCURACY (d_i) FOR LS2 AND LS3 METHODS

Model	N	Mean	Standard Deviation	Median	Minimum	Maximum	Minimum $\lvert d_i \rvert$	Correlation of d_i With Exact Log Likelihood
LS2								
1	50	.0638	.2062	.0015	-.0143	1.3727	.0000	.48*
2	24	.0207	.2811	.0178	-.9789	.3913	.0000	-.03
3	28	-.0073	.0627	.0004	-.1979	.1248	.0000	.19
OVERALL	102	.0341	.2016	.0029	-.9789	1.3727	.0000	.20
LS3								
1	30	.0149	.0333	.0010	-.0081	.1470	.0000	.18
2	20	-.0256	.2698	.0164	-.9466	.3220	.0000	-.02
OVERALL	50	-.0013	.1711	.0046	-.9466	.3220	.0000	-.07

*Significantly different from zero ($\alpha = .05$)

TABLE 4.6

DISTRIBUTION OF ACCURACY (d_i) FOR MG2 AND MG3 METHODS

Model	N	Mean	Standard Deviation	Median	Minimum	Maximum	Minimum $\lvert d_i \rvert$	Correlation of d_i With Exact Log Likelihood
MG2								
1	36	.0514	.2500	-.0020	-.3732	.7238	.0000	.10
2	28	.0575	.1729	.0420	-.4187	.3707	.0098	-.04
OVERALL	64	.0540	.2180	.0143	-.4187	.7238	.0000	.03
MG3								
1	36	.0194	.2203	.0139	.3397	.4677	.0027	-.29
2	28	-.0058	.2691	-.0379	-.5117	.6857	.0127	.11
OVERALL	64	.0084	.2412	.0143	-.5117	.6857	.0027	-.08

TABLE 4.7

SUMMARY OF ACCURACY (d_i) AND TIME FOR ALL METHODS

Method	N	Mean	Standard Deviation	Minimum	Maximum	Mean CPU Time ± Standard Deviation*
SD1	156	-.0037	.1123	-.4355	.601	.5 ± .1
SD2	76	-.0302	.1301	-.6609	.3249	2.5 ± .1
MM	75	.0115	.1398	-.5000	.5706	.5 ± .1
ML2	120	-.0090	.0919	-.5067	.3602	7.3 ± .6
ML3	61	.0086	.0235	-.0709	.1106	15.8 ± 4.1
LS2	102	.0341	.2016	-.9789	1.3727	6.2 ± .6
LS3	50	-.0013	.1711	-.9466	.3220	13.8 ± 2.6
MG2	64	.0540	.2180	-.4187	.7238	1.5 ± .1
MG3	64	.0084	.2412	-.5117	.6857	2.4 ± .2

*Mean CPU for exact = 1.5 ± .1

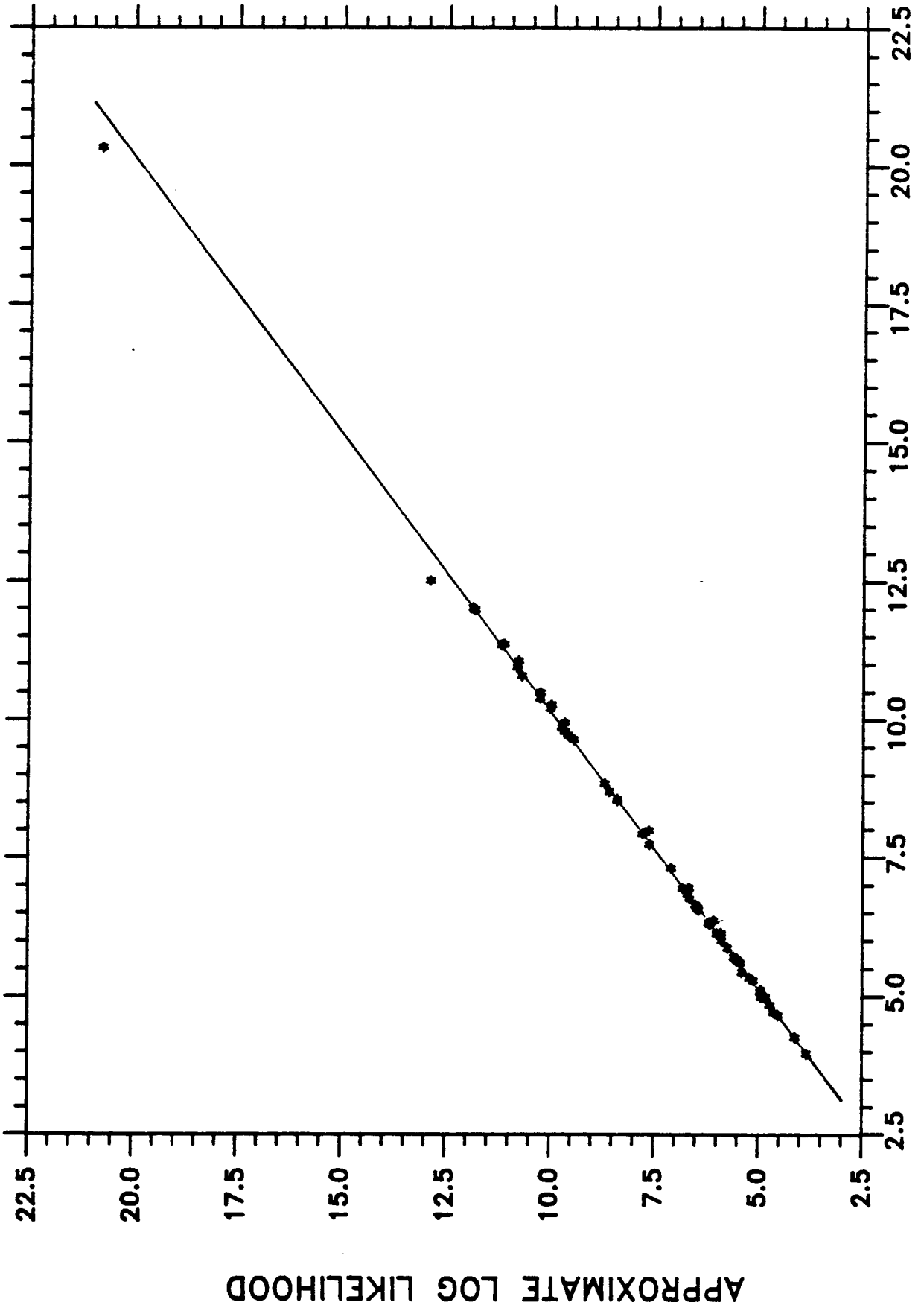
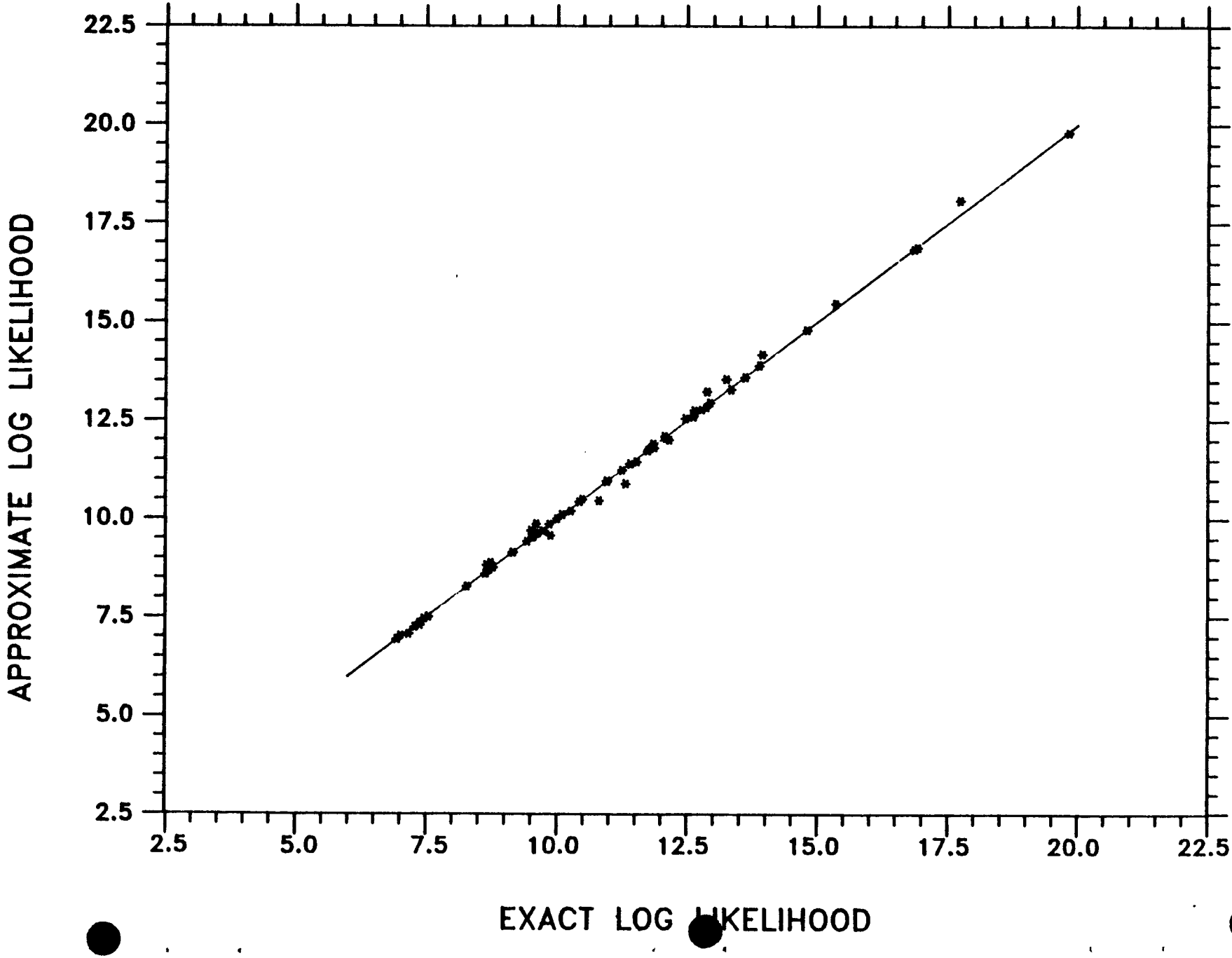


FIGURE 4.1a - FIT OF METHOD SD1 ON MODEL 1 DATA

FIGURE 4.1b - FIT OF METHOD SD1 ON MODEL 2 DATA



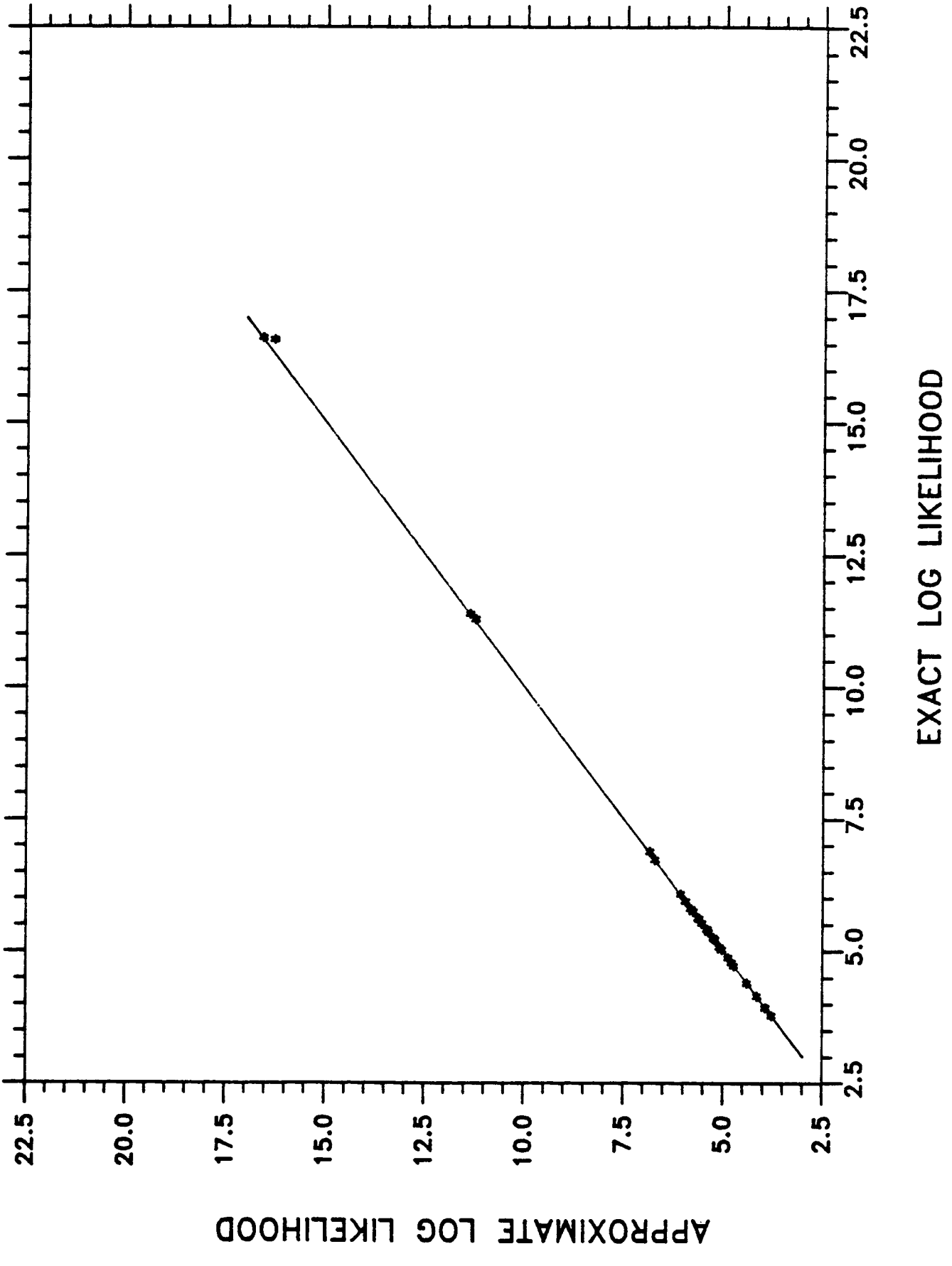
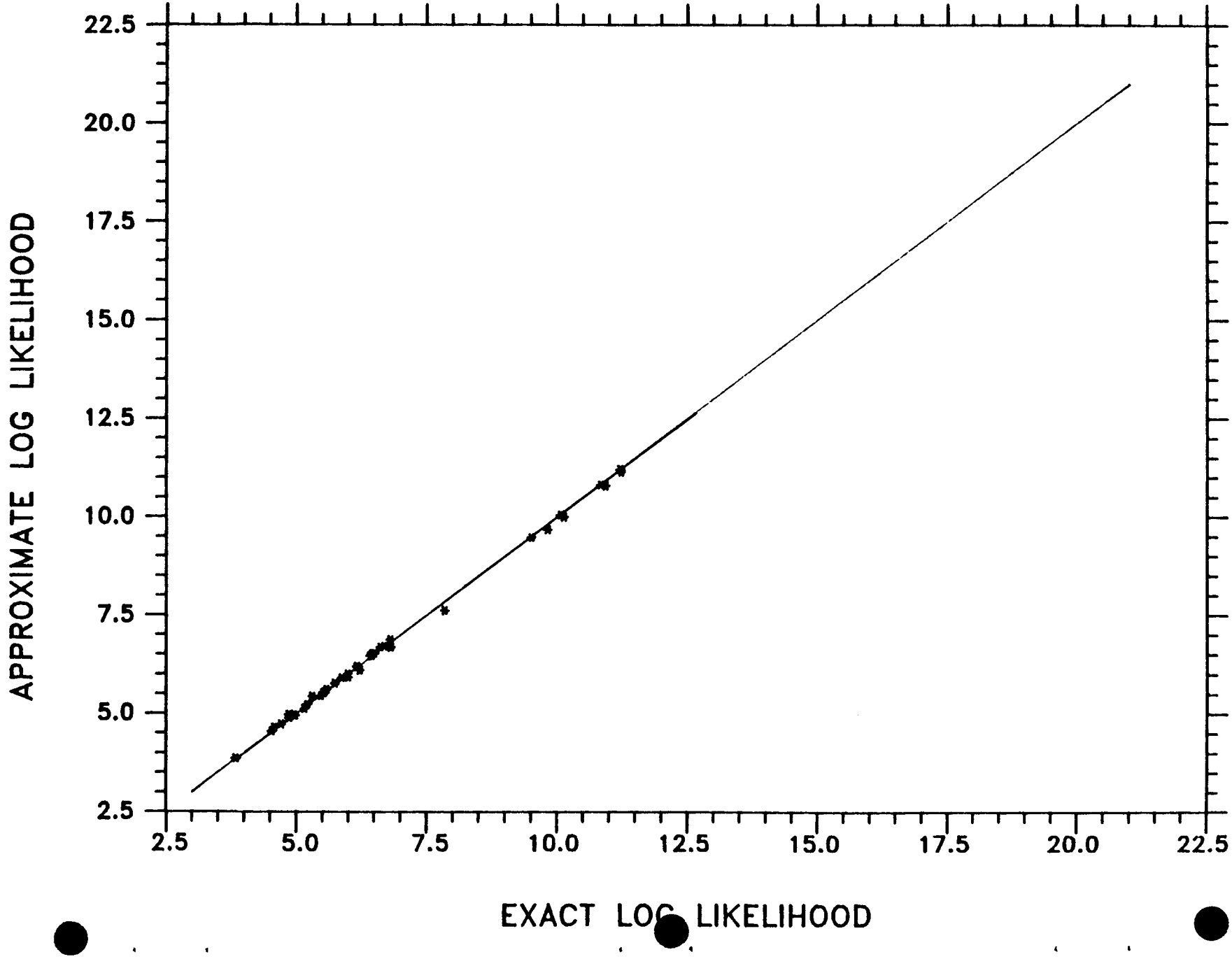


FIGURE 4.1c - FIT OF METHOD SD1 ON MODEL 3 DATA

FIGURE 4.1d - FIT OF METHOD SD2 ON MODEL 1 DATA



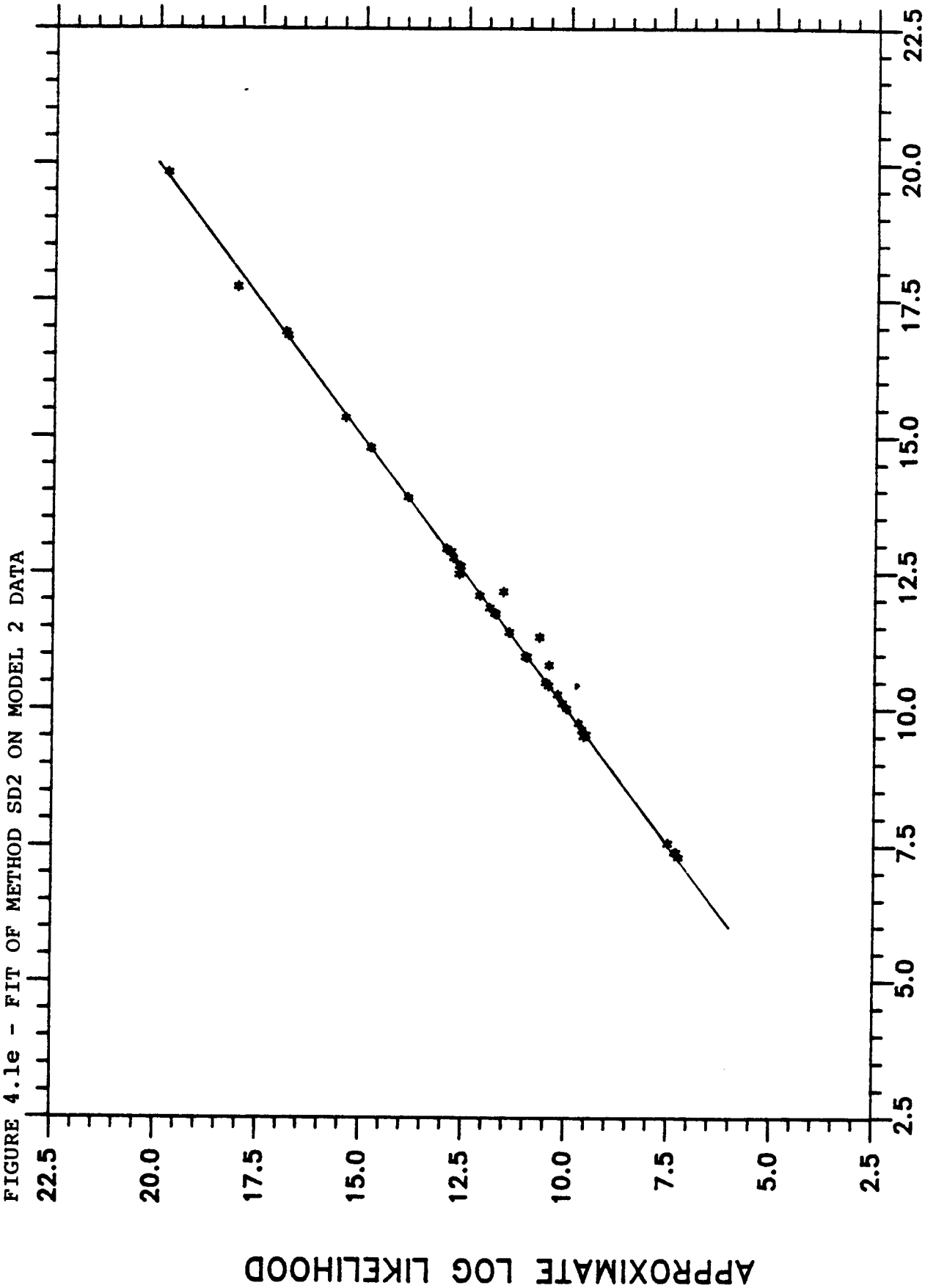
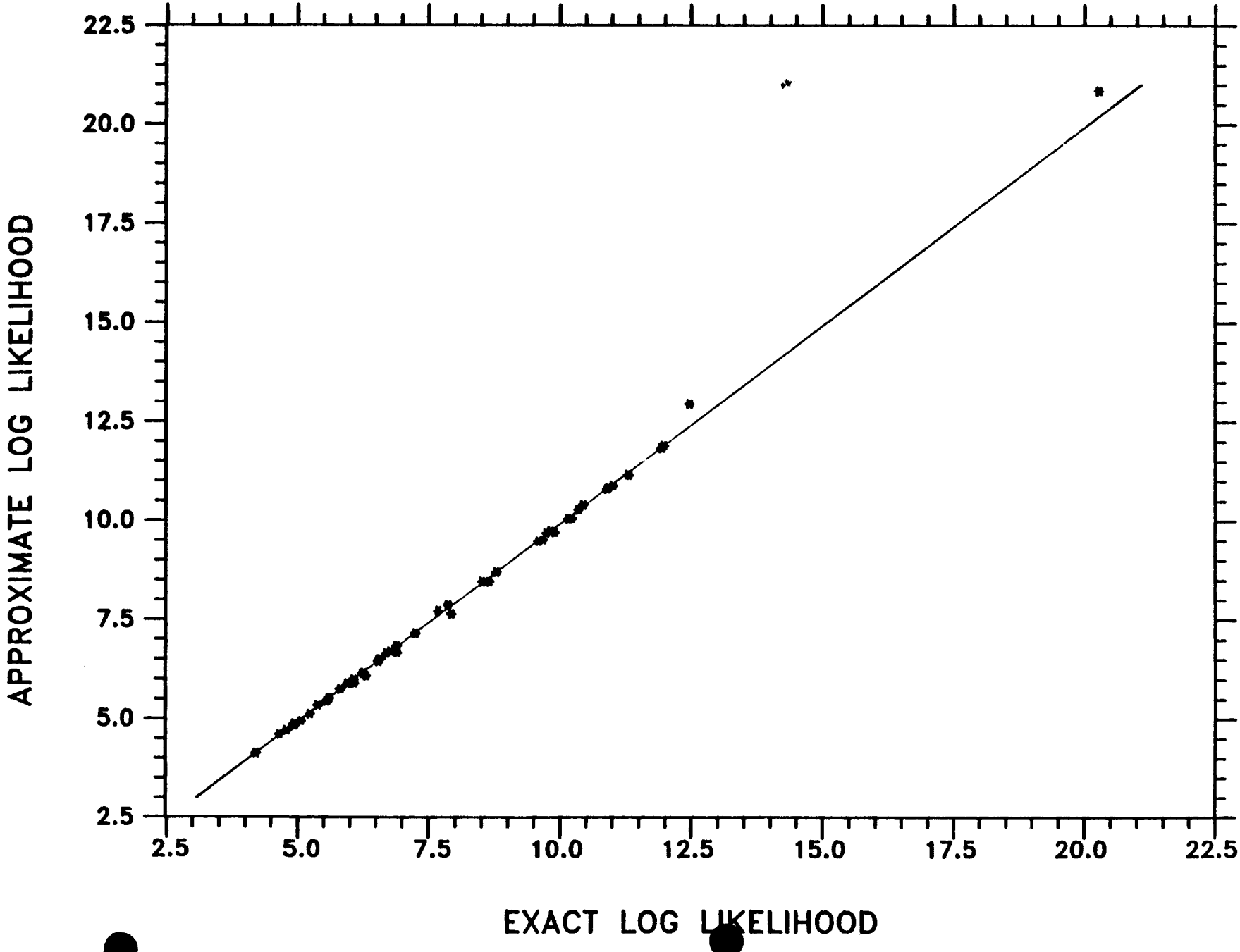


FIGURE 4.1e - FIT OF METHOD SD2 ON MODEL 2 DATA

FIGURE 4.2a - FIT OF METHOD MM ON MODEL 1 DATA



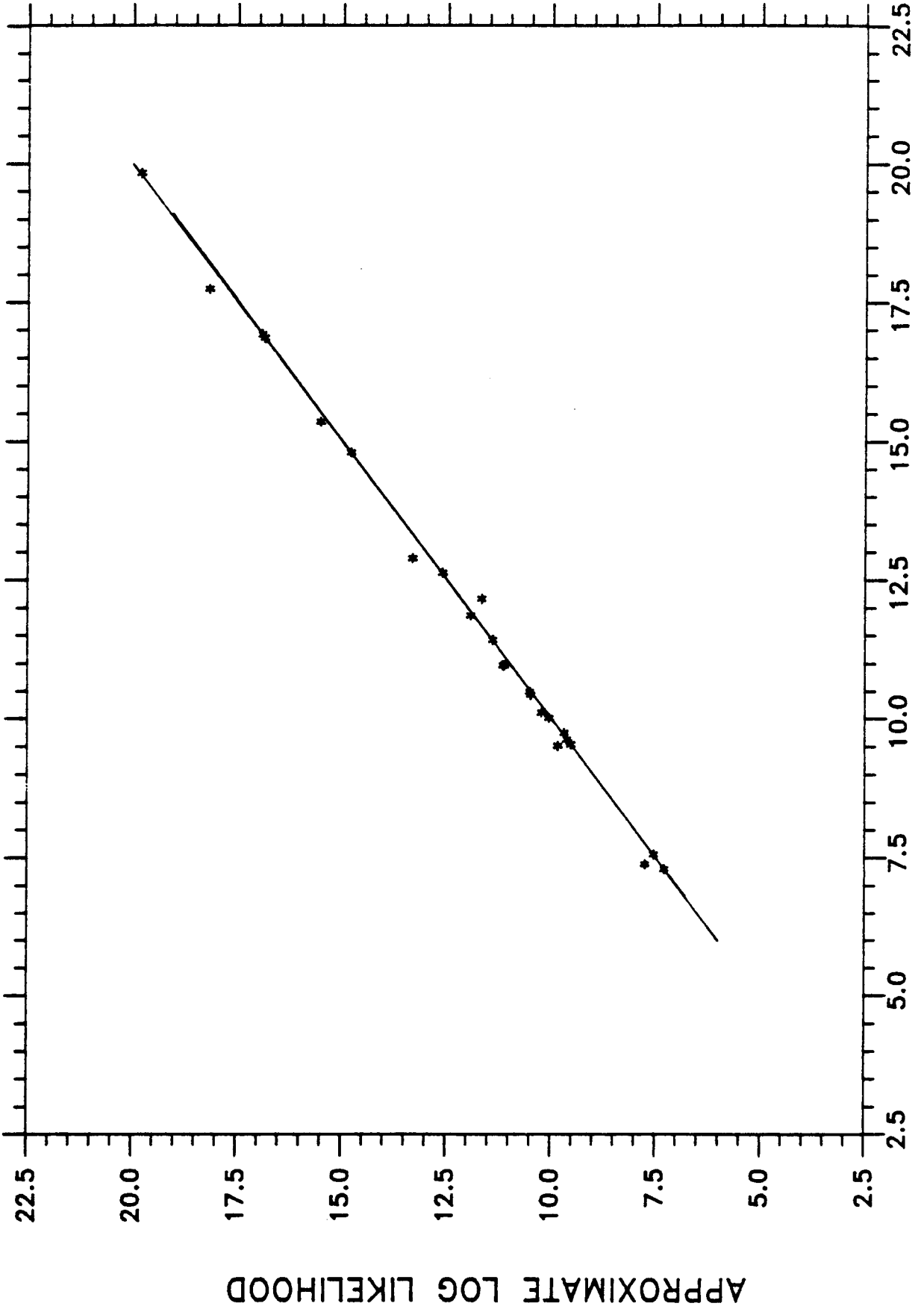
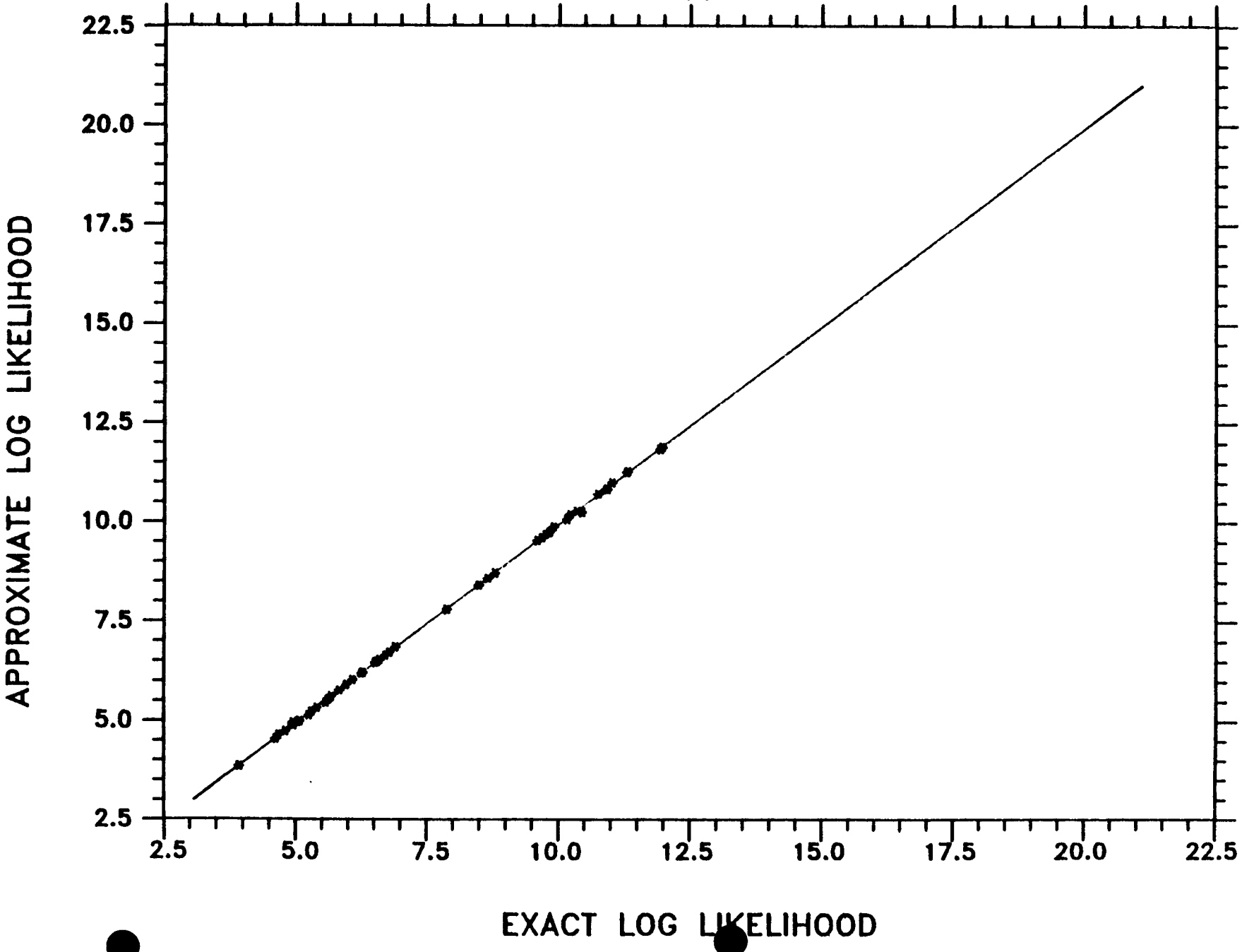


FIGURE 4.2b - FIT OF METHOD MM ON MODEL 2 DATA

FIGURE 4.3a - FIT OF METHOD ML2 ON MODEL 1 DATA



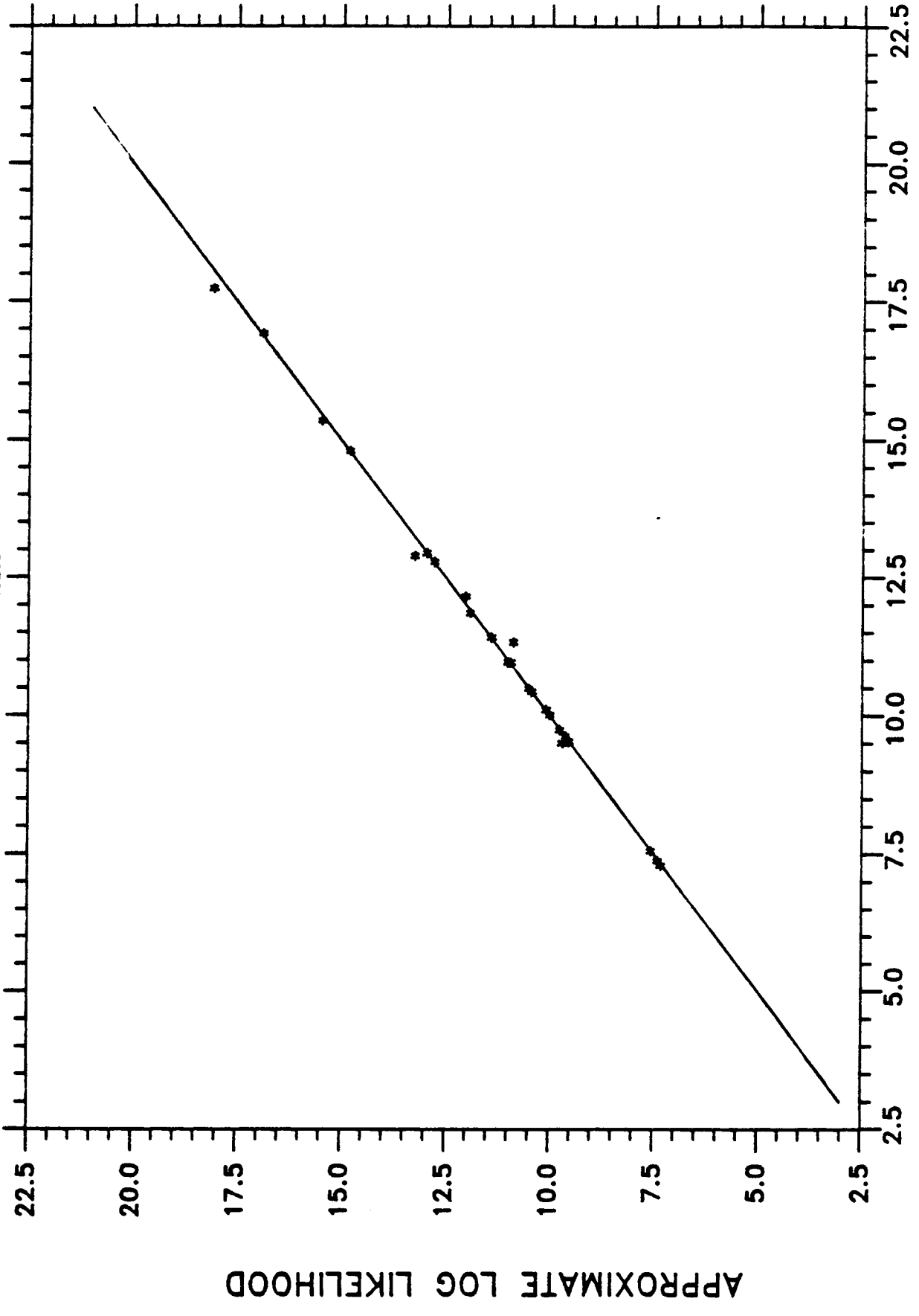


FIGURE 4.3b - FIT OF METHOD ML2 ON MODEL 2 DATA

FIGURE 4.3c - FIT OF METHOD ML2 ON MODEL 3 DATA

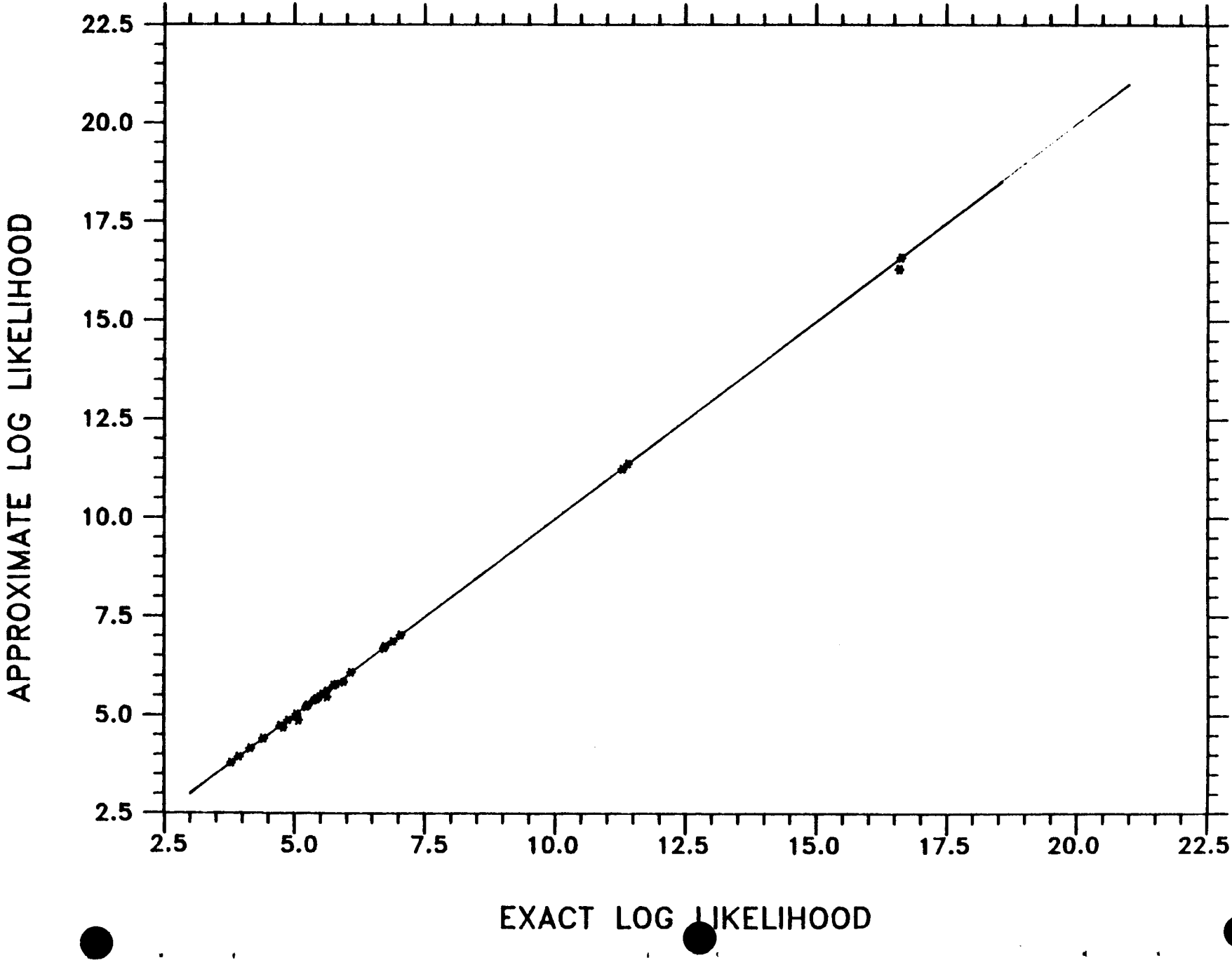


FIGURE 4.3d - FIT OF METHOD ML3 ON MODEL 1 DATA

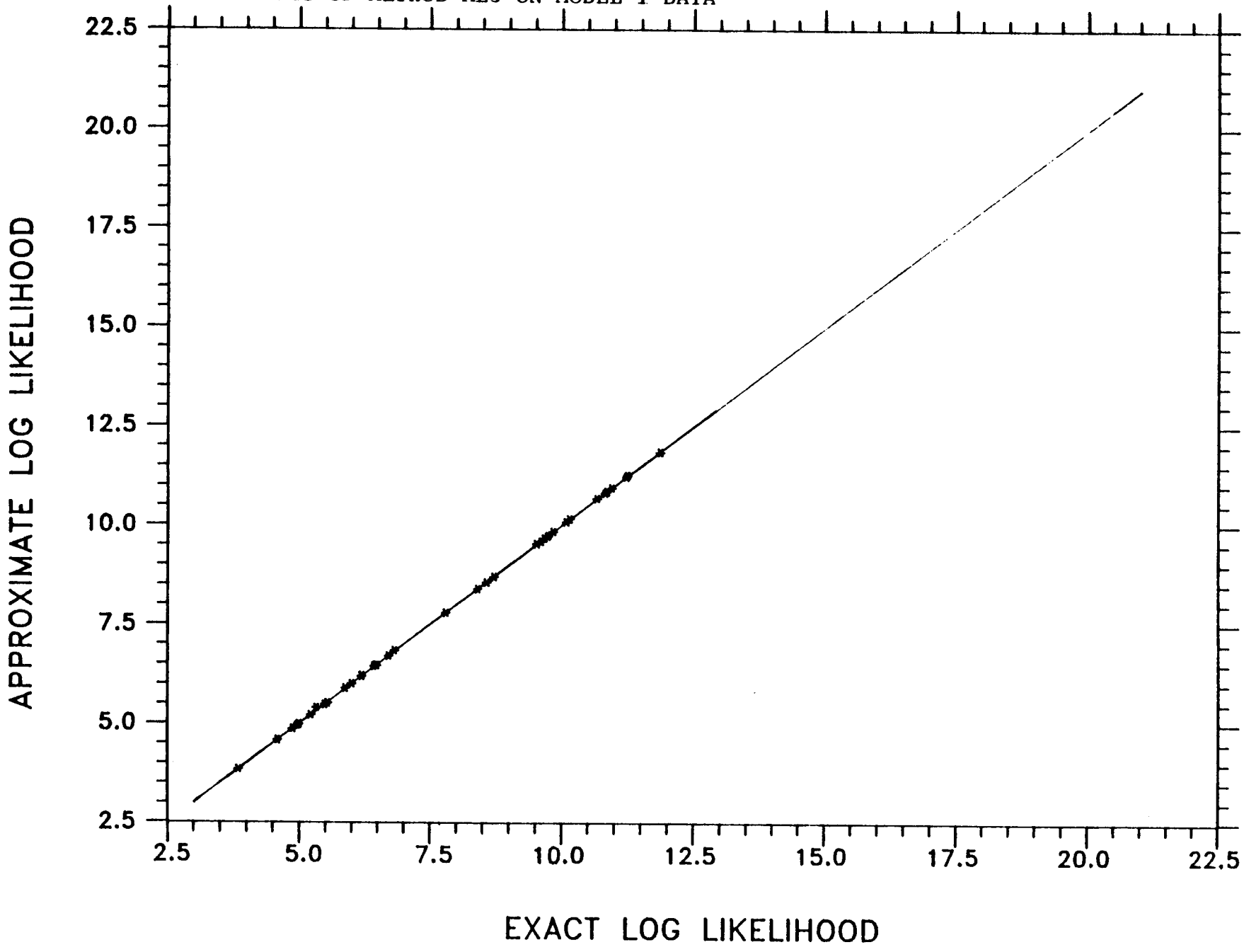


FIGURE 4.3e - FIT OF METHOD ML3 ON MODEL 2 DATA

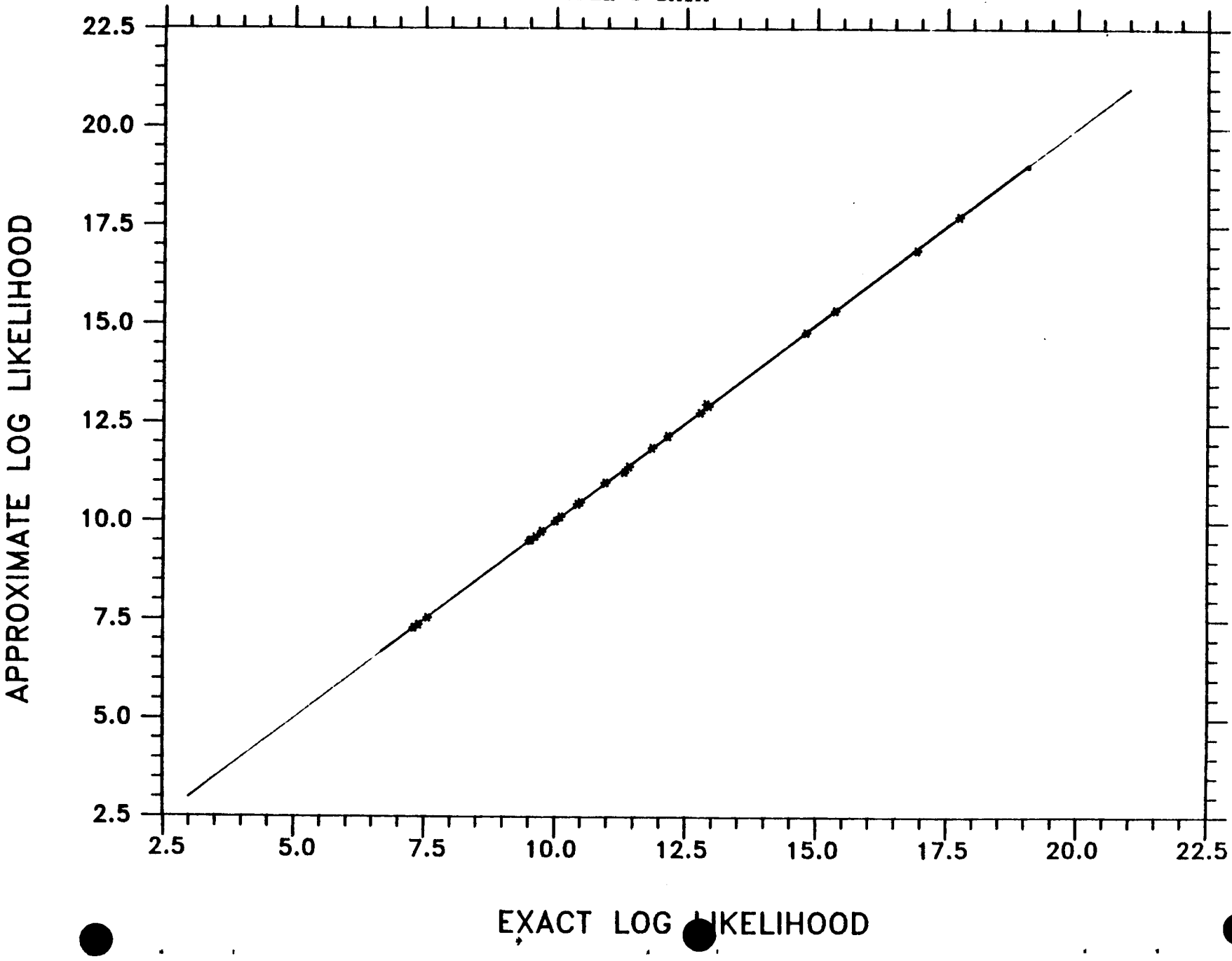


FIGURE 4.4a - FIT OF METHOD LS2 ON MODEL 1 DATA

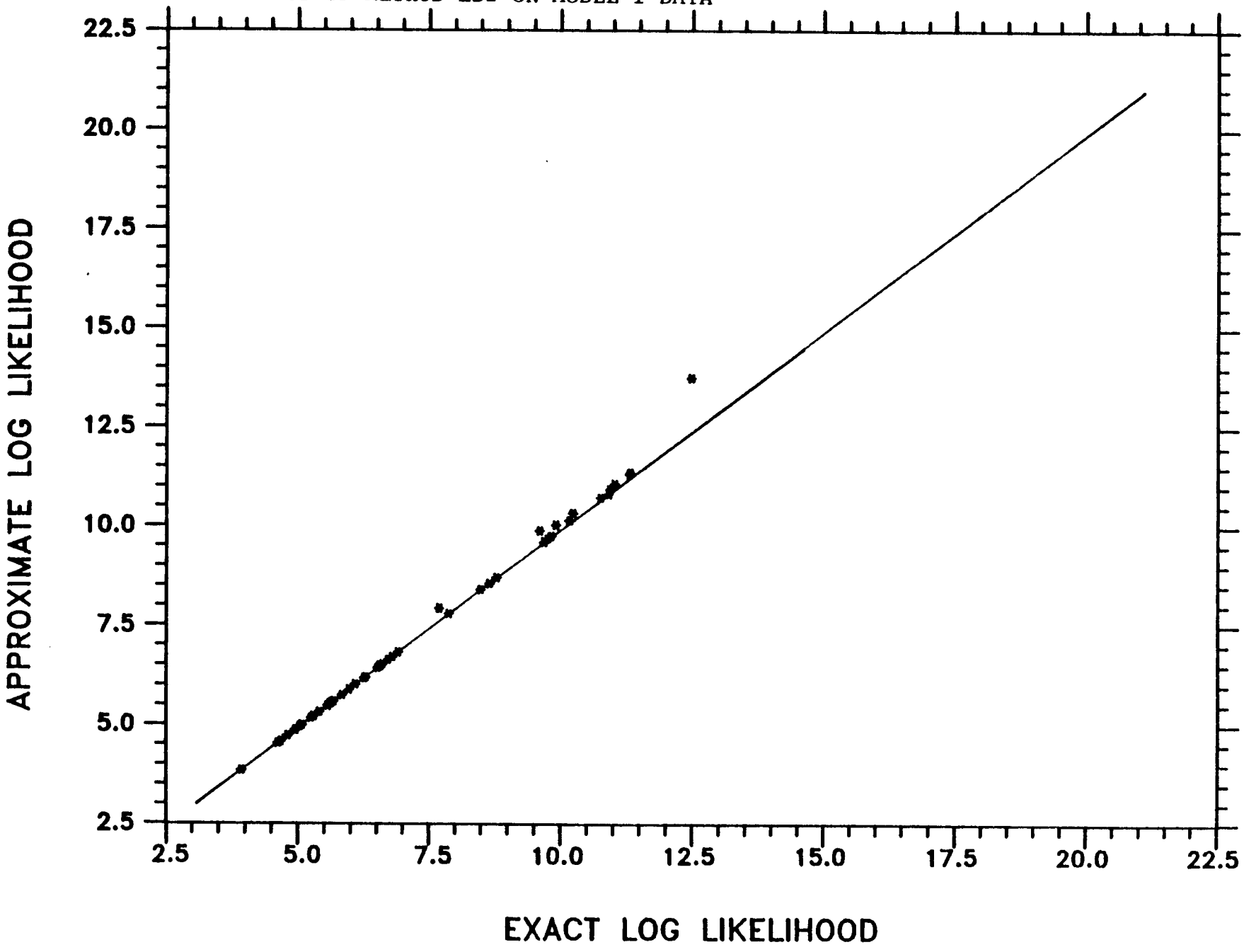


FIGURE 4.4b - FIT OF METHOD LS2 ON MODEL 2 DATA

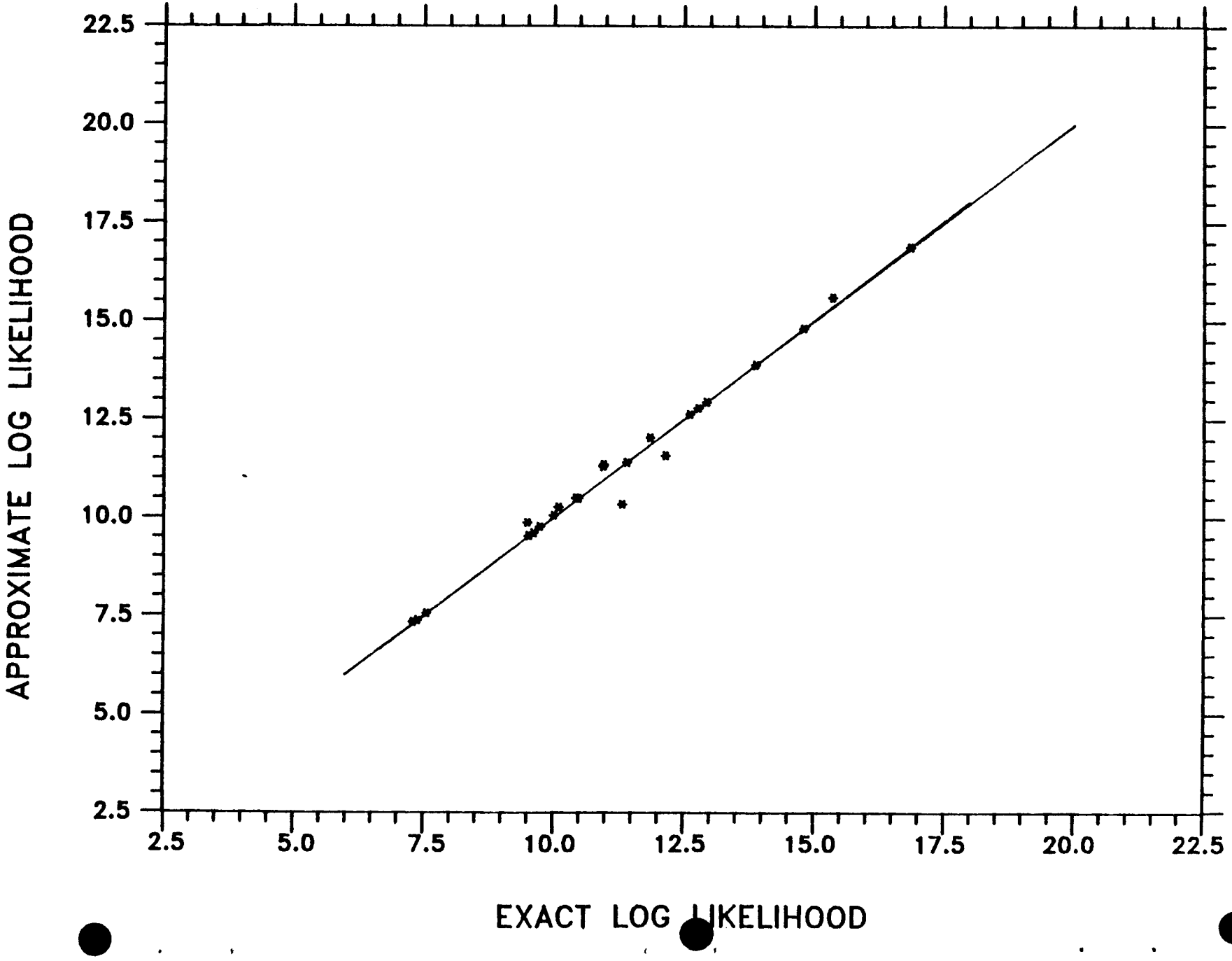


FIGURE 4.4c - FIT OF METHOD LS2 ON MODEL 3 DATA

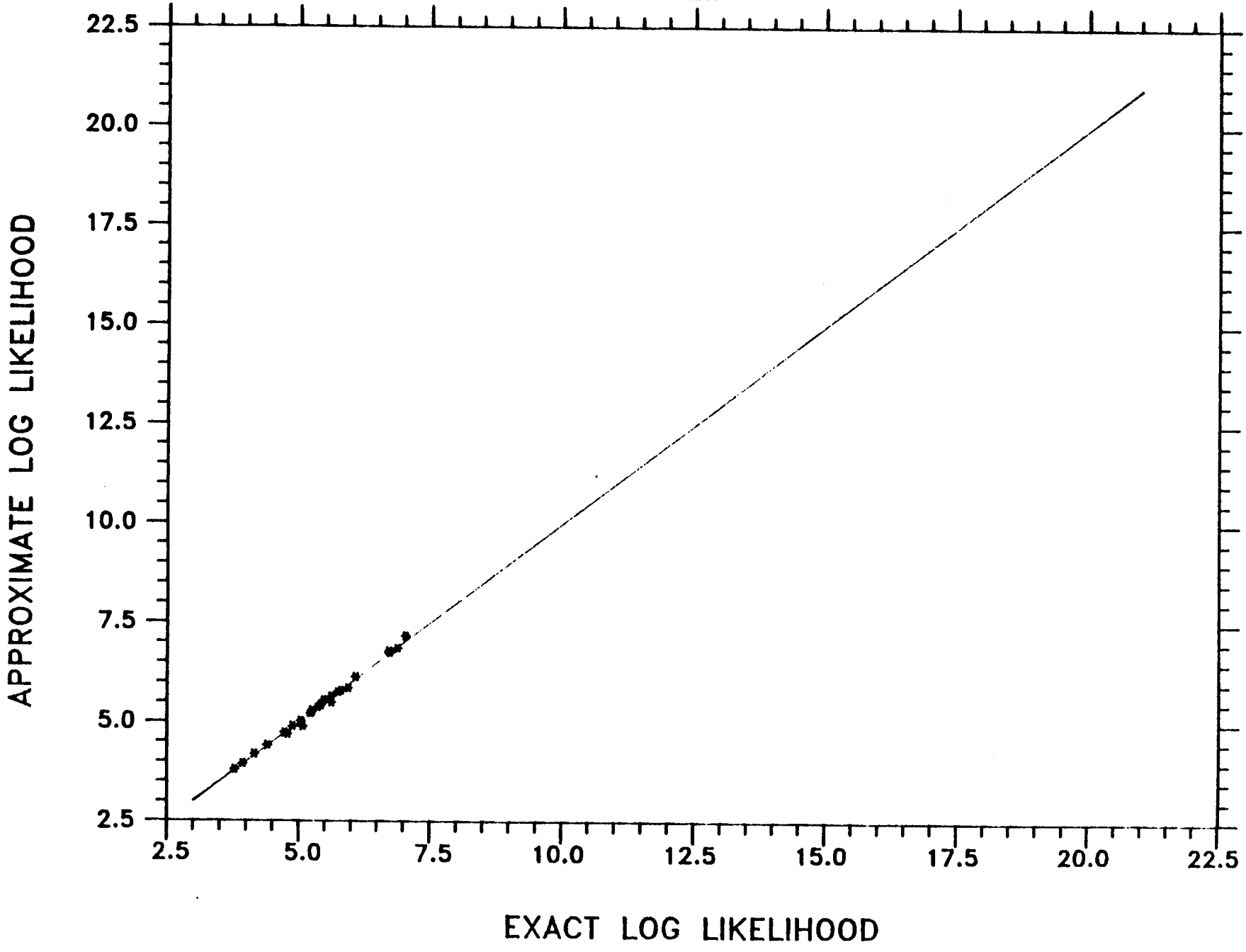


FIGURE 4.4d - FIT OF METHOD LS3 ON MODEL 1 DATA

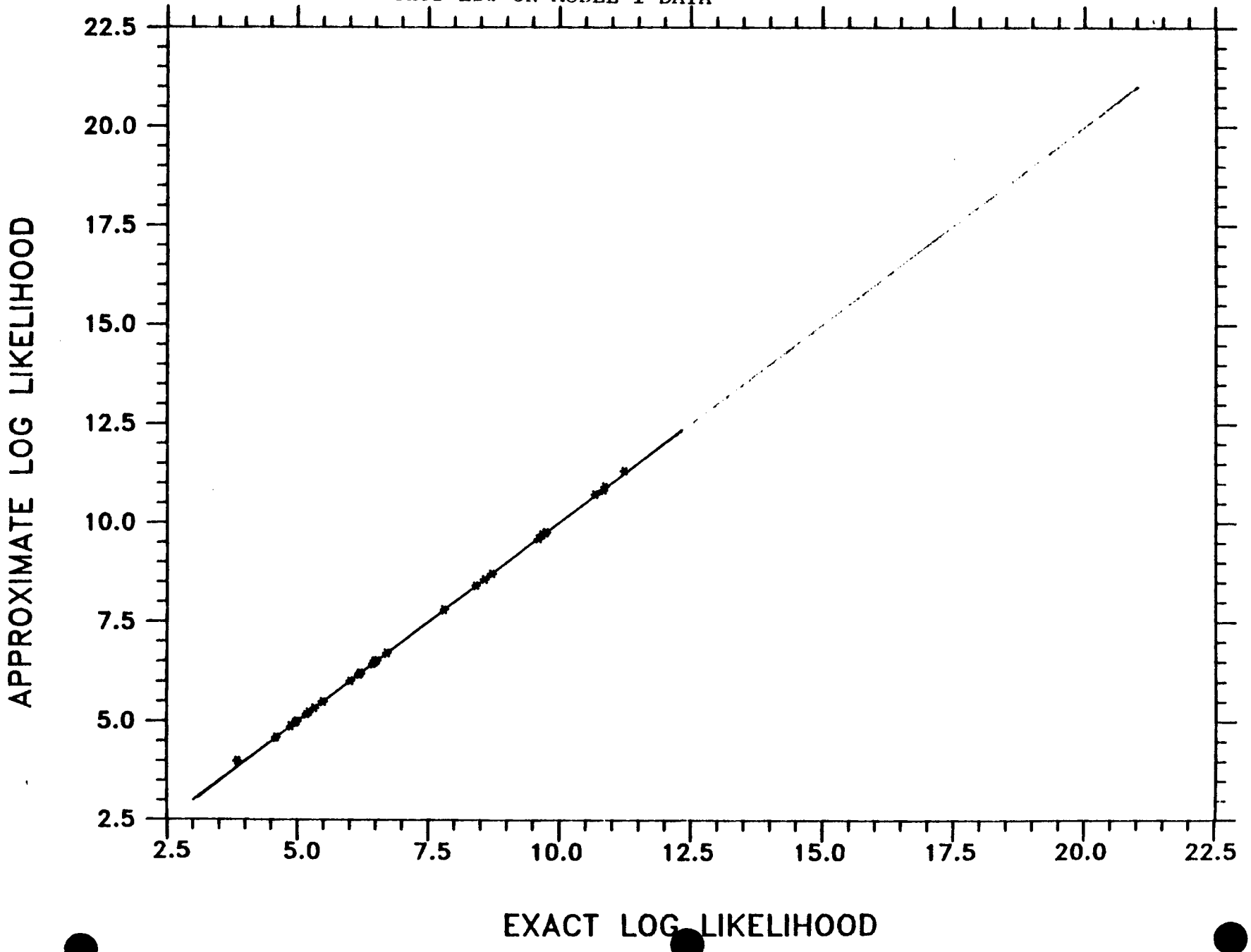


FIGURE 4.4e - FIT OF METHOD LS3 ON MODEL 2 DATA

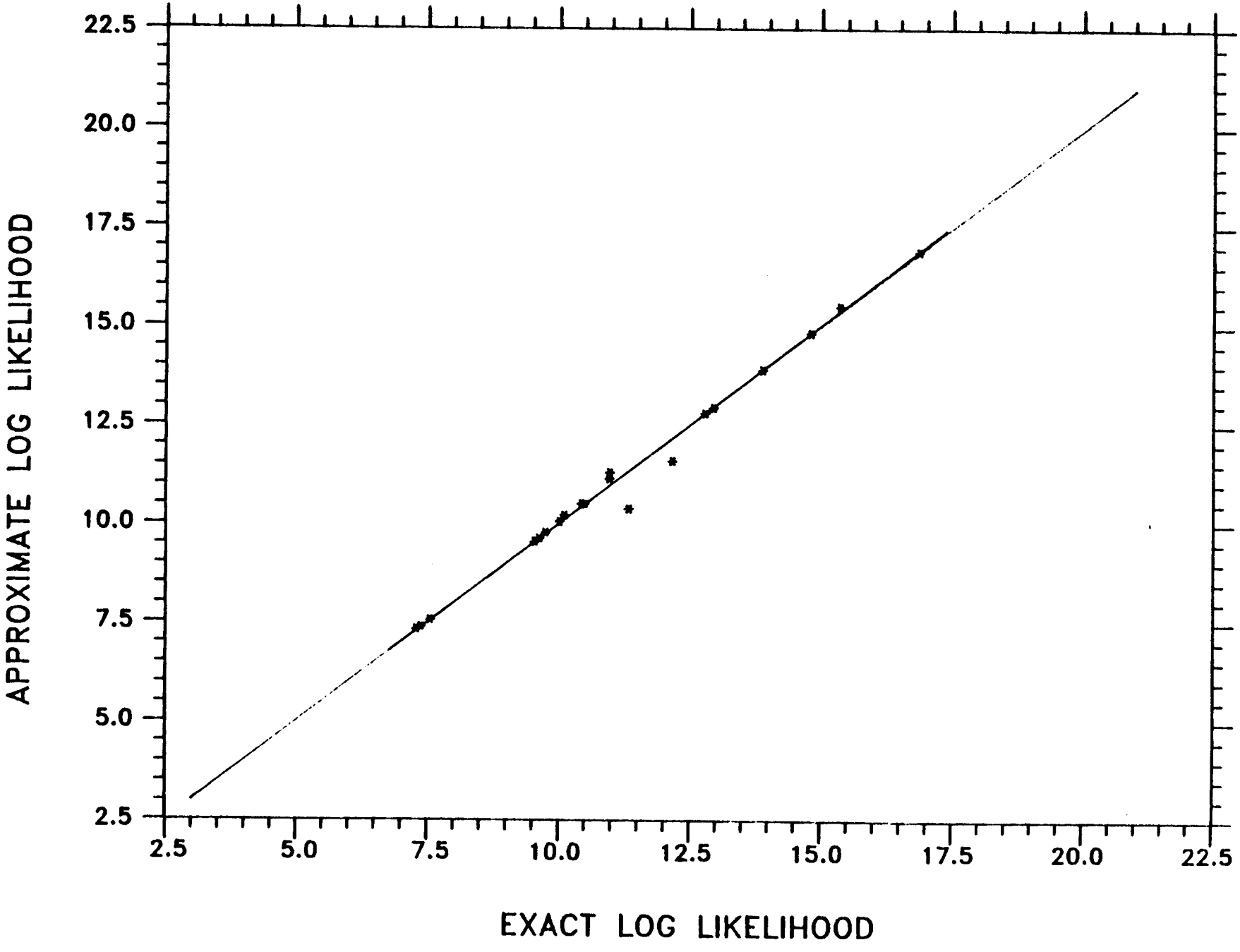


FIGURE 4.5a - FIT OF METHOD MG2 ON MODEL 1 DATA

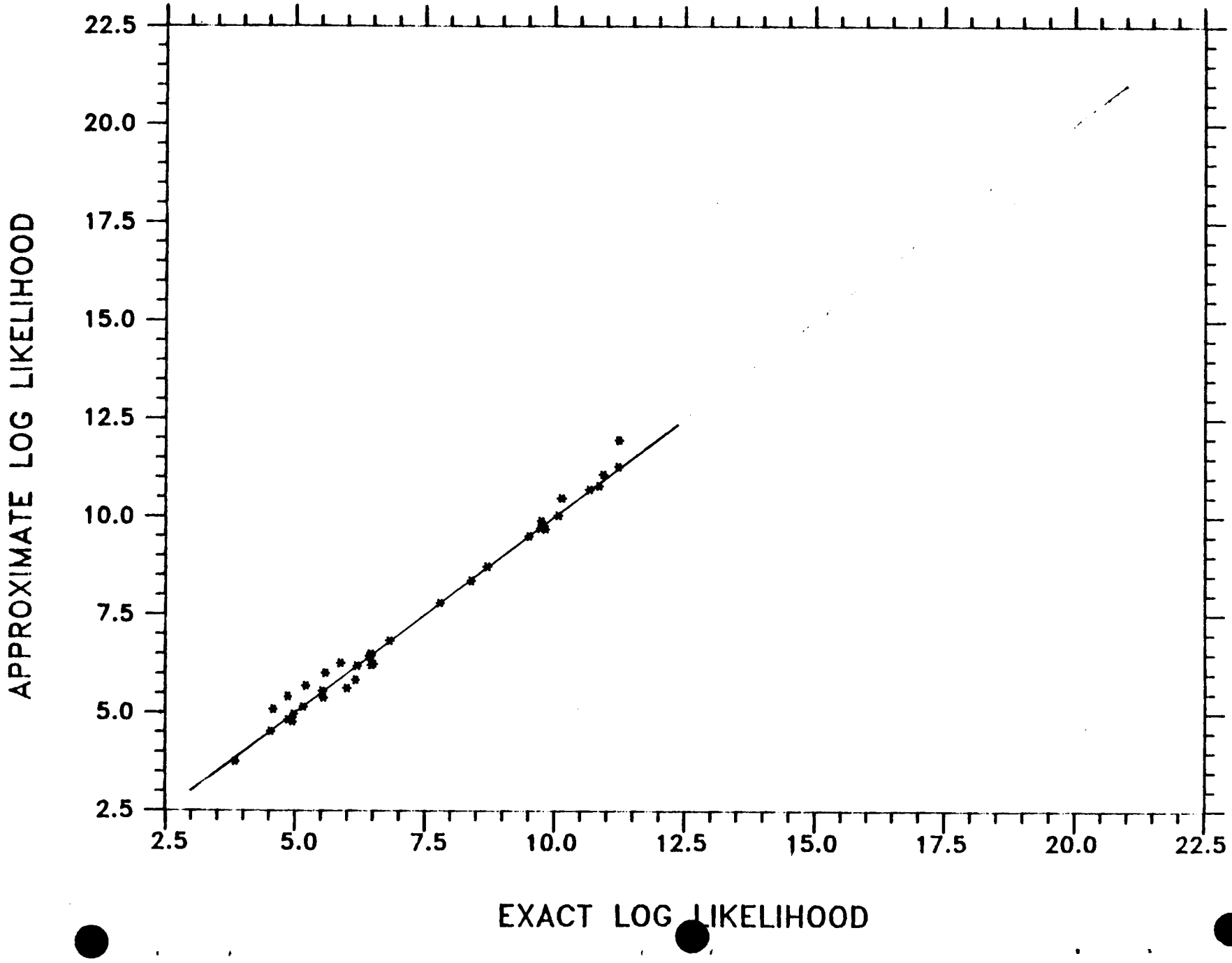


FIGURE 4.5b - FIT OF METHOD MG2 ON MODEL 2 DATA

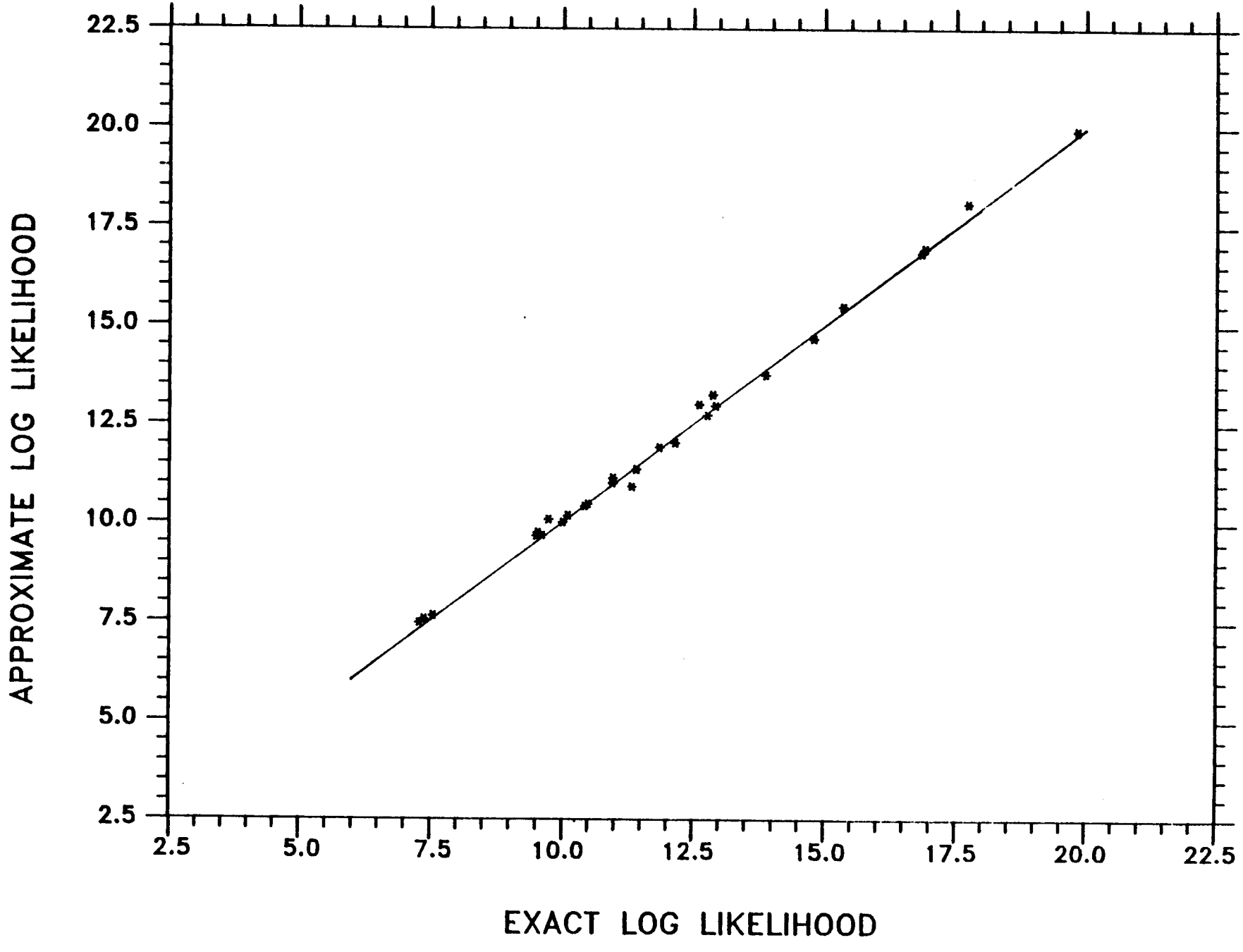


FIGURE 4.5c - FIT OF METHOD MG3 ON MODEL 1 DATA

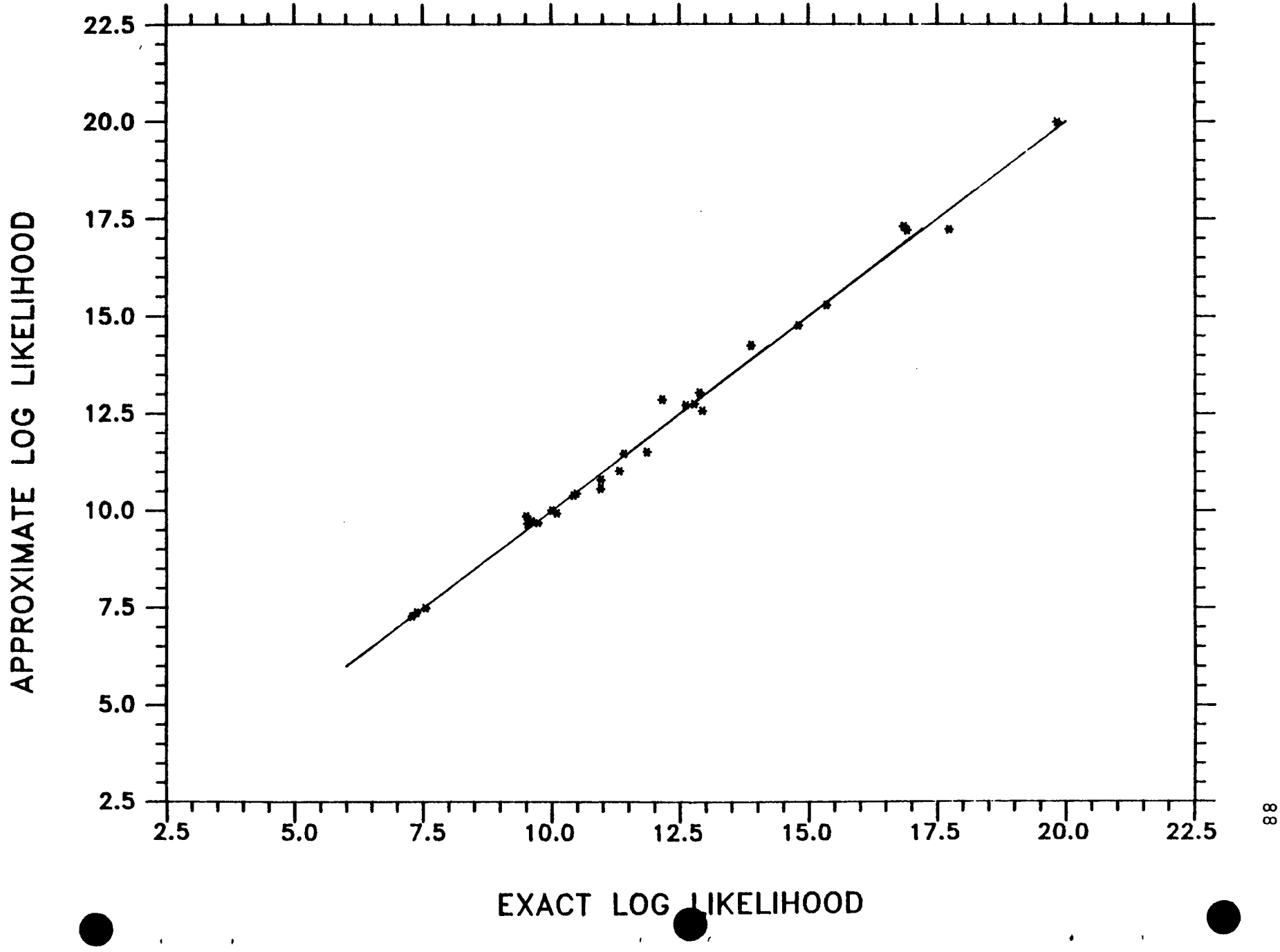


FIGURE 4.5d - FIT OF METHOD MG3 ON MODEL 2 DATA

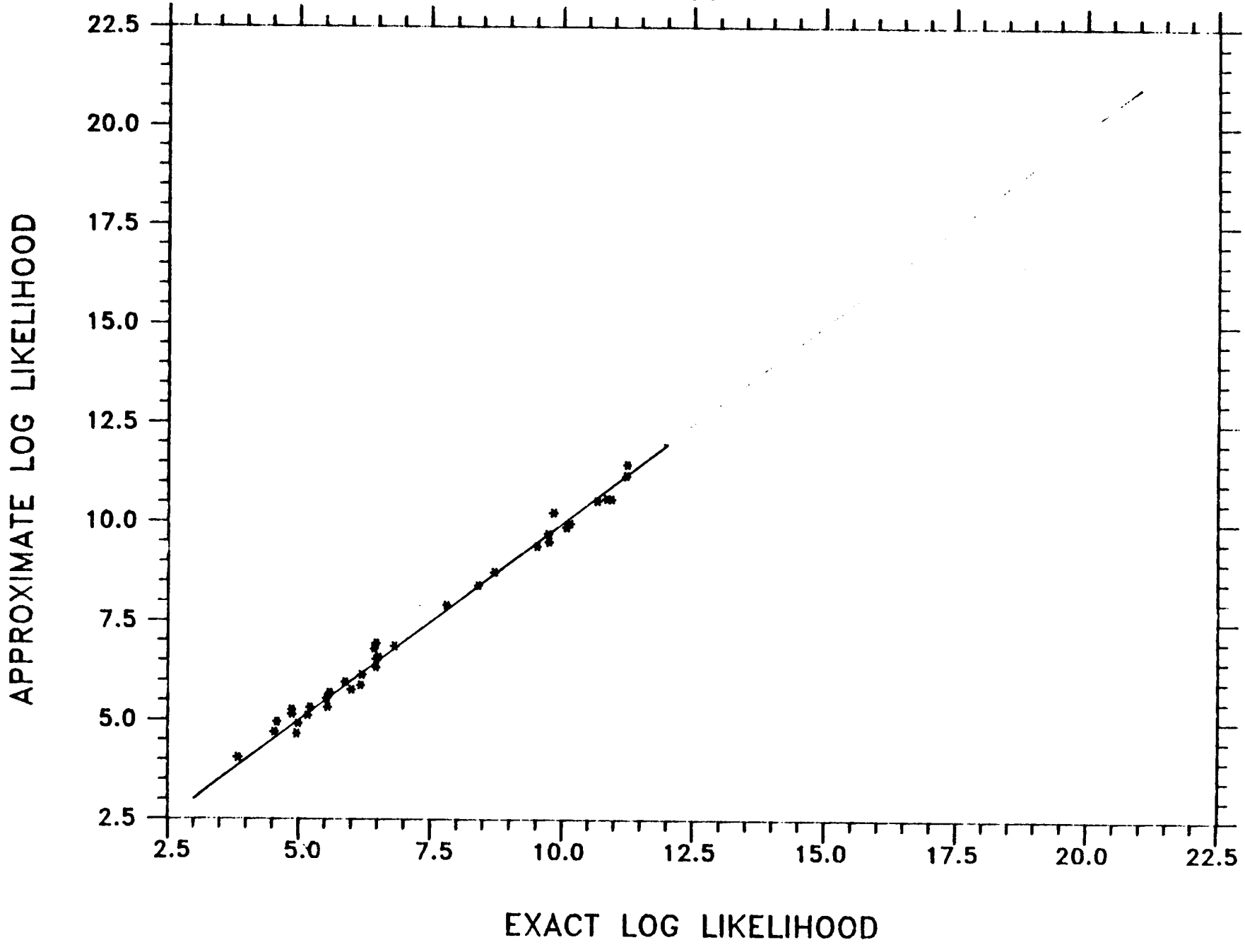
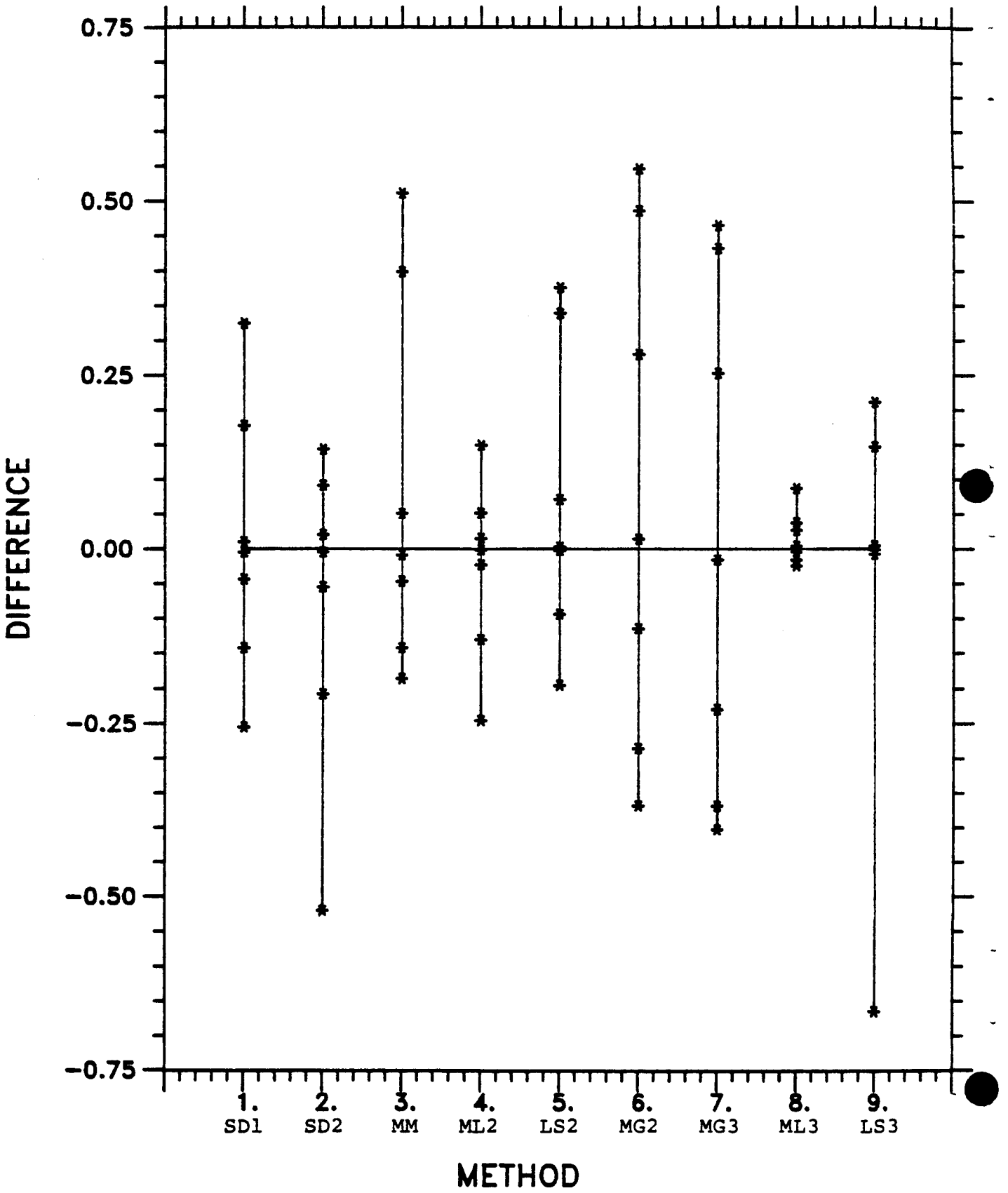


Figure 4.6

DISTRIBUTION OF ACCURACY [90]

Percentiles (97.5, 95., 83.3, 50, 16.7, 5., 2.5)



4.6. Comparisons on Real Data

To study the accuracy of the approximations on real data the first twenty-five families of the dataset described in Chapter V were chosen. Maximum likelihood estimates for the entire dataset were calculated for three genetic models by the program MIXMOD (Spence, et al. (1979)),a program that calculates likelihoods for mixed genetic models for nuclear families. Using these estimates obtained from the entire dataset, it is possible to calculate the likelihood for each family as well as the total likelihood. As a result it is possible to observe the cumulative effect of using an approximation as well as to study the effect of individual families as has been previously discussed. Since the families vary in size there is also the possibility of deciding whether the accuracy of the approximations is influenced by the size of the family. This dataset presents a "worst case" test for the accuracy of the approximations. The median number of individuals for which an approximation is made is only two, and it is expected that the accuracy increases as more individuals are present. For such small families the approximation will also inevitably be slower computationally than the exact methods.

The estimates used are those ML estimates of the complete dataset on the 2N(0,1) transformed data; this is a transformation described in Chapter V. The estimates used are those under the restrictions d=1, those under the

restrictions q=t=d=0 and those under the unrestricted model. An attempt was made to calculate the ML estimates for the subset of twenty-five families, but convergence was difficult to obtain even without the approximations. Apparently the likelihood surface is flat over the range of reasonable parameter values. This circumstance emphasizes the importance of large sets of data when attempting to discriminate between different genetic models.

4.7. Results of Approximations on Real Data

Table 4.8 gives the descriptive statistics on the accuracy for each of the approximation methods for 75 log likelihoods (25 families x 3 sets of parameters). Figures 4.7a-f are constructed in an identical manner to figures 4.1-4.5. Figures 4.8a-f are plots of the d_i values versus family size.

As expected, the approximations did not perform nearly as well on these small families as they have on the larger families. It can be seen by comparing Tables 4.7 and 4.8 that the mean inaccuracy ($\bar{d}_i$) for all the methods is generally an order of magnitude larger in the dataset with the small families than with the dataset with large families. These means are also significantly different from zero, which indicates the approximations were biased as well. This bias is illustrated in figures 4.7a-e, where for all but the LS2 method most of the points were below the line of equality and

for the LS2 method most of the points were above this line. This bias can be seen to be related to family size, as figure 4-8a-e show. For all but the ML2 method there is an obvious relationship between family size and bias. If we disregard the three points from one family of size six, there appears to be a strong improvement in overall accuracy with larger families. The reason for the poor accuracy at one particular family is not known, but it is important to note that the MIXMOD program also poorly estimates the likelihoods for this family. (See figures 4.7f and 4.8f.) This trend to better accuracy with larger families is encouraging, and suggests the approximations can work well if the approximation is done on large families.

The cumulative effect of this bias is illustrated in Table 4.9, which gives the total log likelihood for each of the sets of parameters as well as the difference between the two restricted models and the unrestriced model. None of the methods reproduce the results of exact calculation, not even MIXMOD. Given the bias present within each of the families, however, it is not surprising that the cumulative performance is less than stellar.

TABLE 4.8

DISTRIBUTION OF ACCURACY FOR EACH APPROXIMATION ON ALL 75 LIKELIHOODS

Method	Mean	Standard Deviation	Median	Minimum	Maximum	Minimum $\lvert d_i \rvert$	Correlation of d_i With Exact Log Likelihood	Correlation of d_i With Size
SD1	-.0560*	.0649	-.0553	-.1656	.2086	.0002	.41*	.66*
SD2	-.0425*	.0686	-.0384	-.1721	.2723	.0000	.38*	.64*
ML2	-.0310*	.0497	-.0271	-.1555	.1317	.0000	.16	.18
LS2	.1031*	.2180	-.0636	.0000	1.4200	.0000	.29*	.24*
MG2	-.0546*	.0615	-.0531	-.1754	.1994	.0002	.39*	.67*
MIXMOD	-.0057	.0636	-.0057	-.2891	.1328	.0000	.05	.07

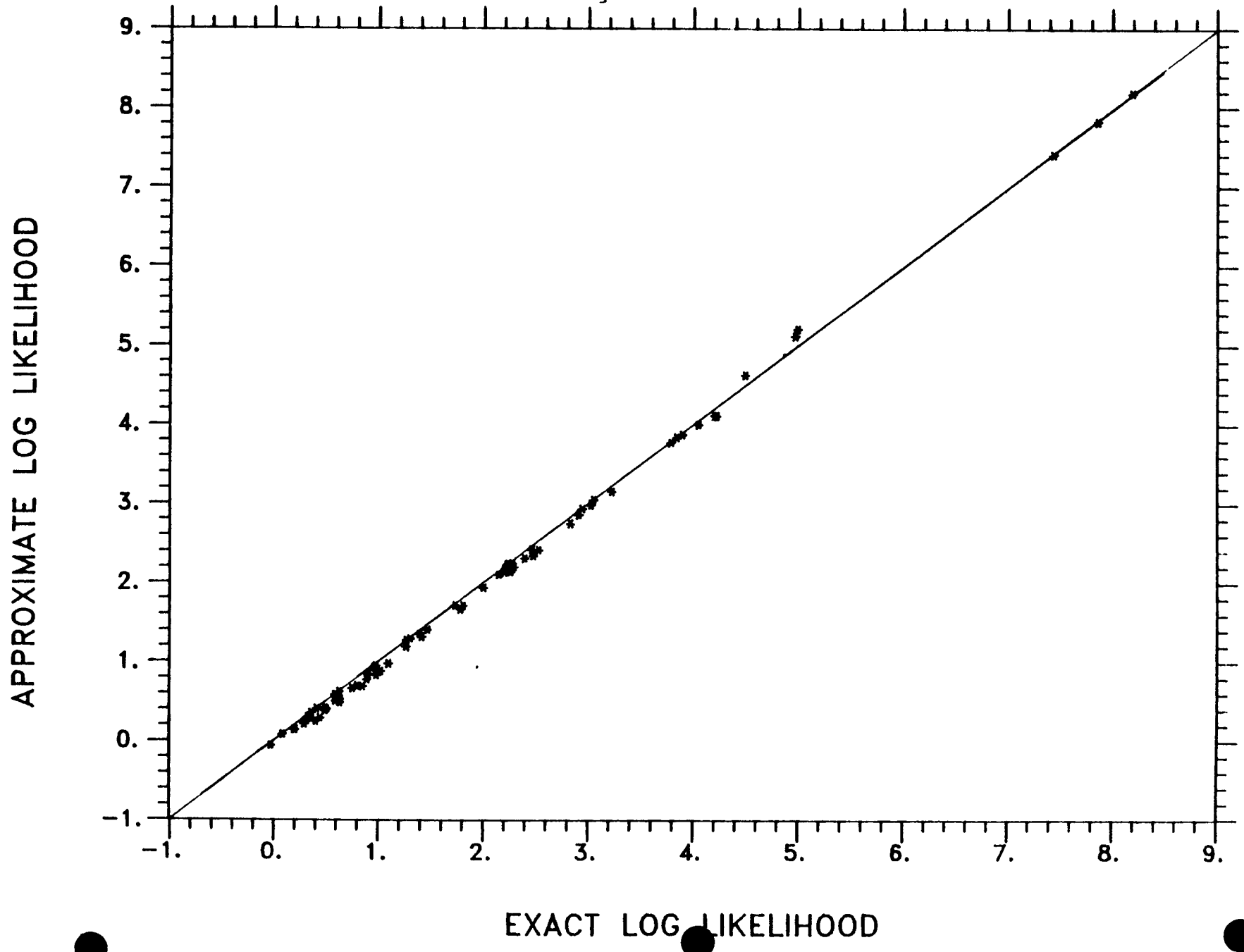
*Significantly different from zero ($\alpha = .05$)

TABLE 4.9

LOG LIKELIHOODS FOR FIRST 25 FAMILIES OF IgE DATA SETS FOR 3 SETS OF PARAMETERS FOR 7 NUMERICAL METHODS

Method	Unrestricted Log Likelihood	Log Likelihood With Restriction d=1	Log Likelihood With Restriction q=t=d=0	Column 1 - Column 2	Column 1 - Column 3
Exact	150.43	150.09	149.29	1.14	.34
MIXMOD	150.18	149.86	149.34	.32	.84
SD1	148.37	148.35	148.89	.02	-.52
SD2	148.79	148.82	149.10	-.03	-.31
ML2	149.18	149.39	148.90	-.21	-.72
LS2	153.68	153.53	150.33	.15	3.35
MG2	148.59	148.44	148.75	.15	-.16

FIGURE 4.7a - FIT OF METHOD SD1 ON IgE DATA



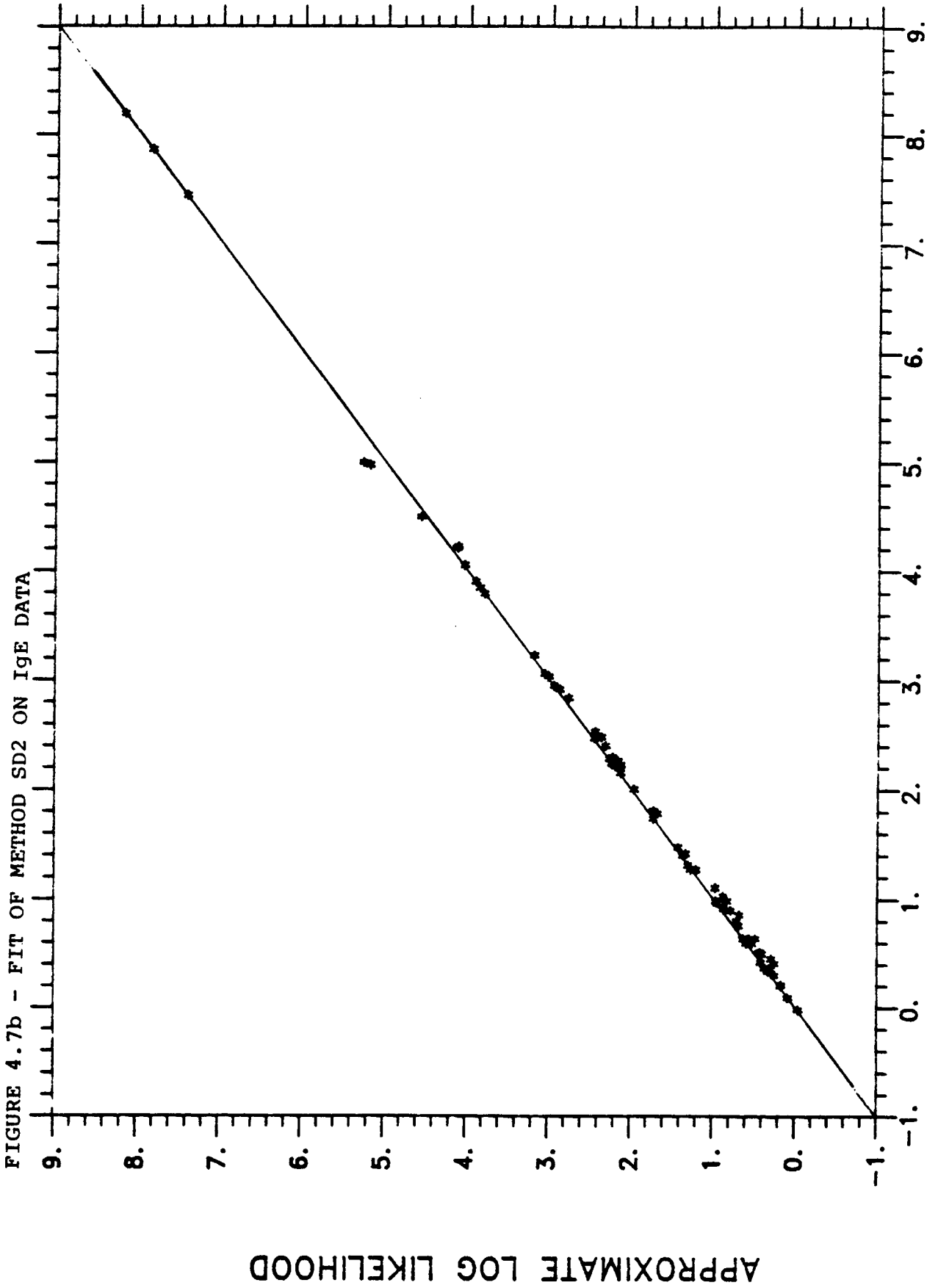


FIGURE 4.7b - FIT OF METHOD SD2 ON IgE DATA

FIGURE 4.7c - FIT OF METHOD ML2 ON IgE DATA

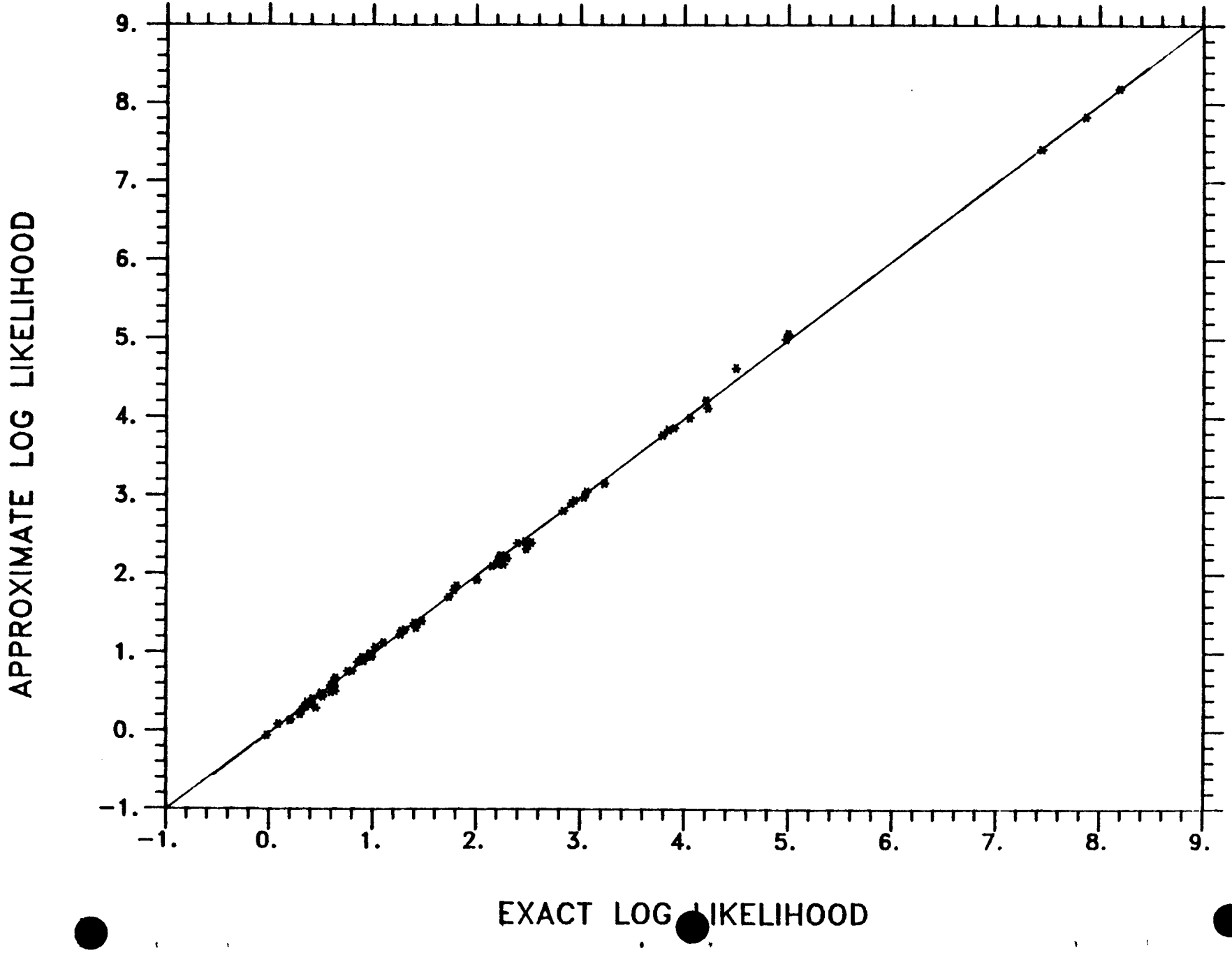


FIGURE 4.7d - FIT OF METHOD LS2 ON IgE DATA

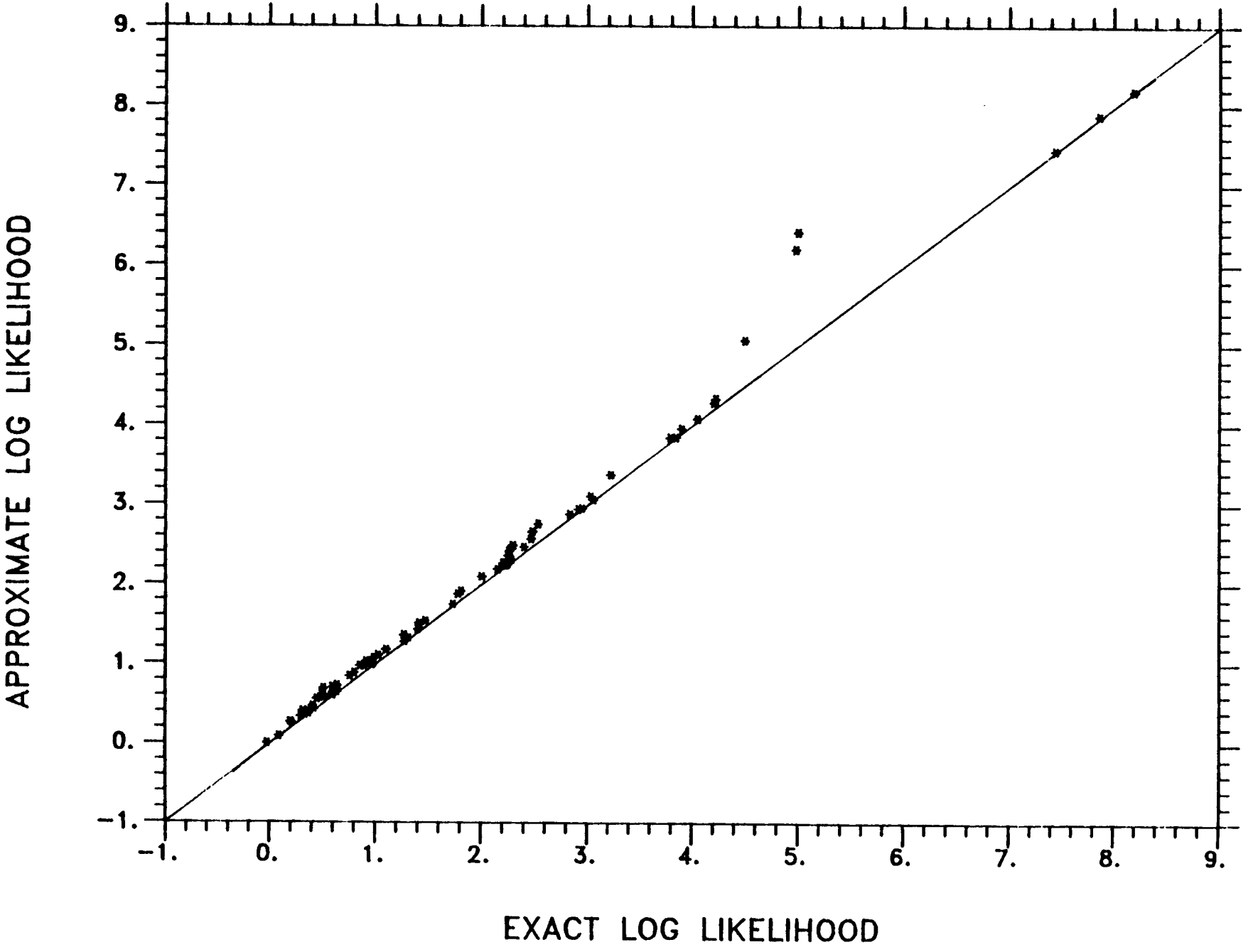
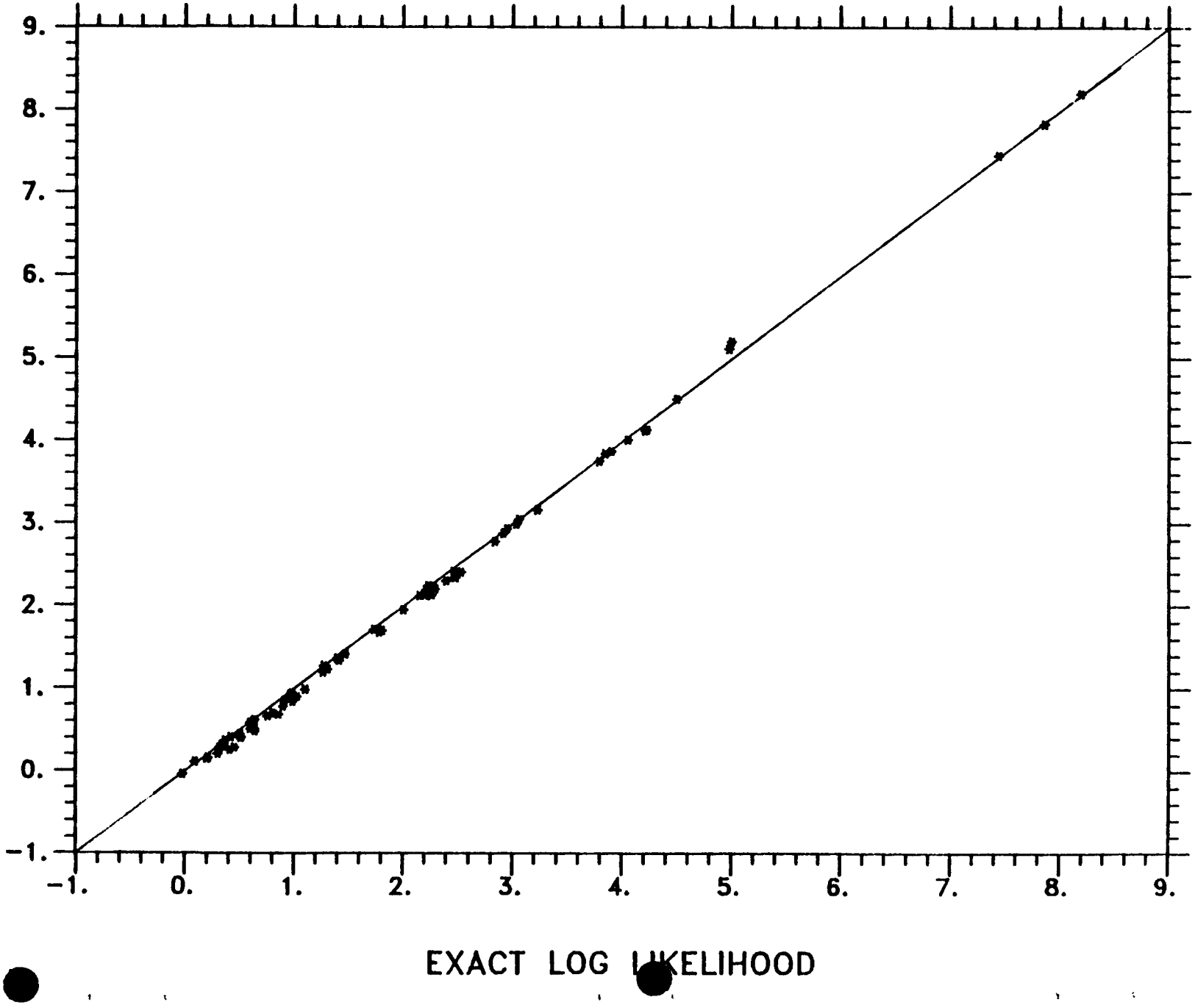


FIGURE 4.7e - FIT OF METHOD MG2 ON IgE DATA



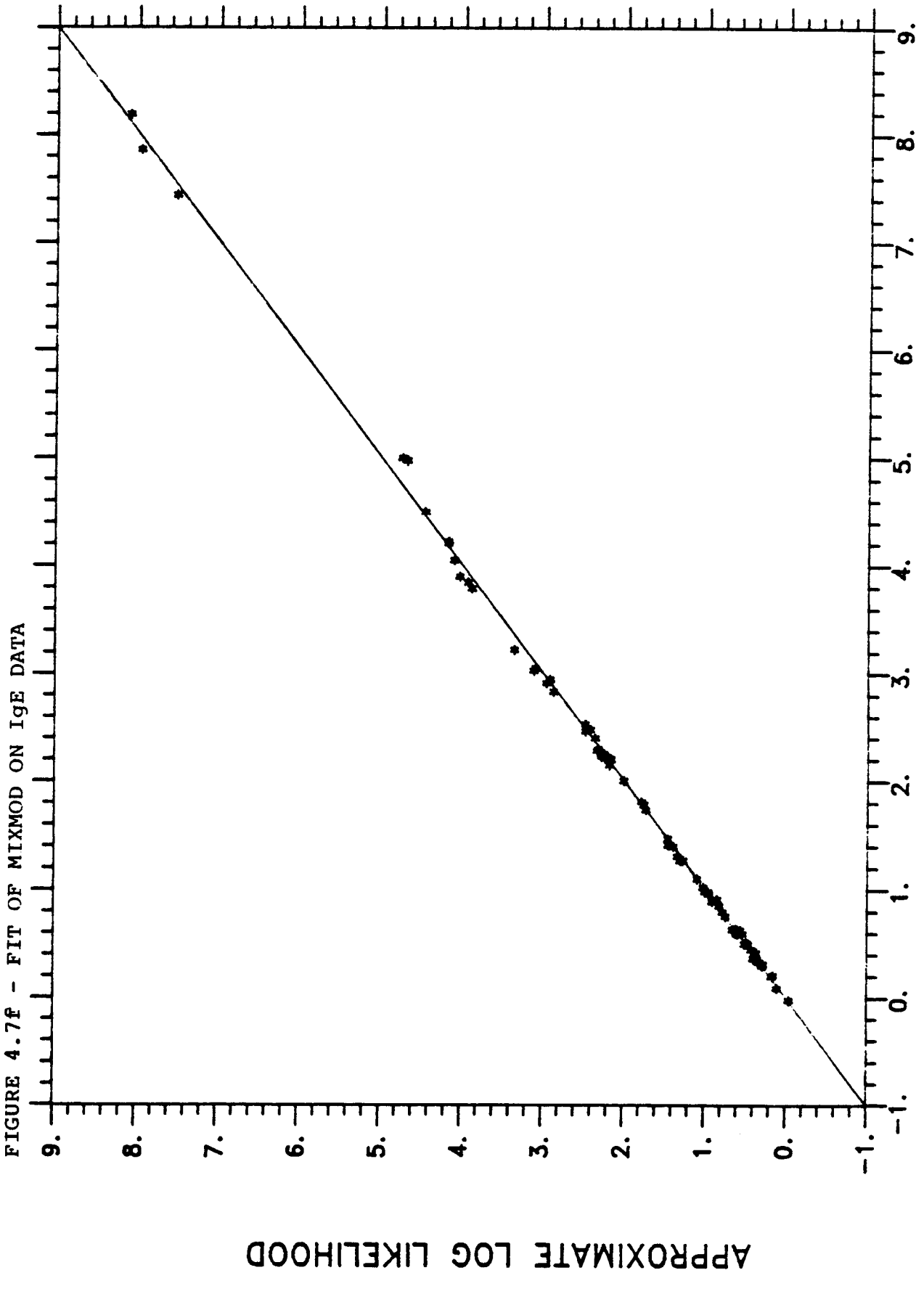


FIGURE 4.7f - FIT OF MIXMOD ON IgE DATA

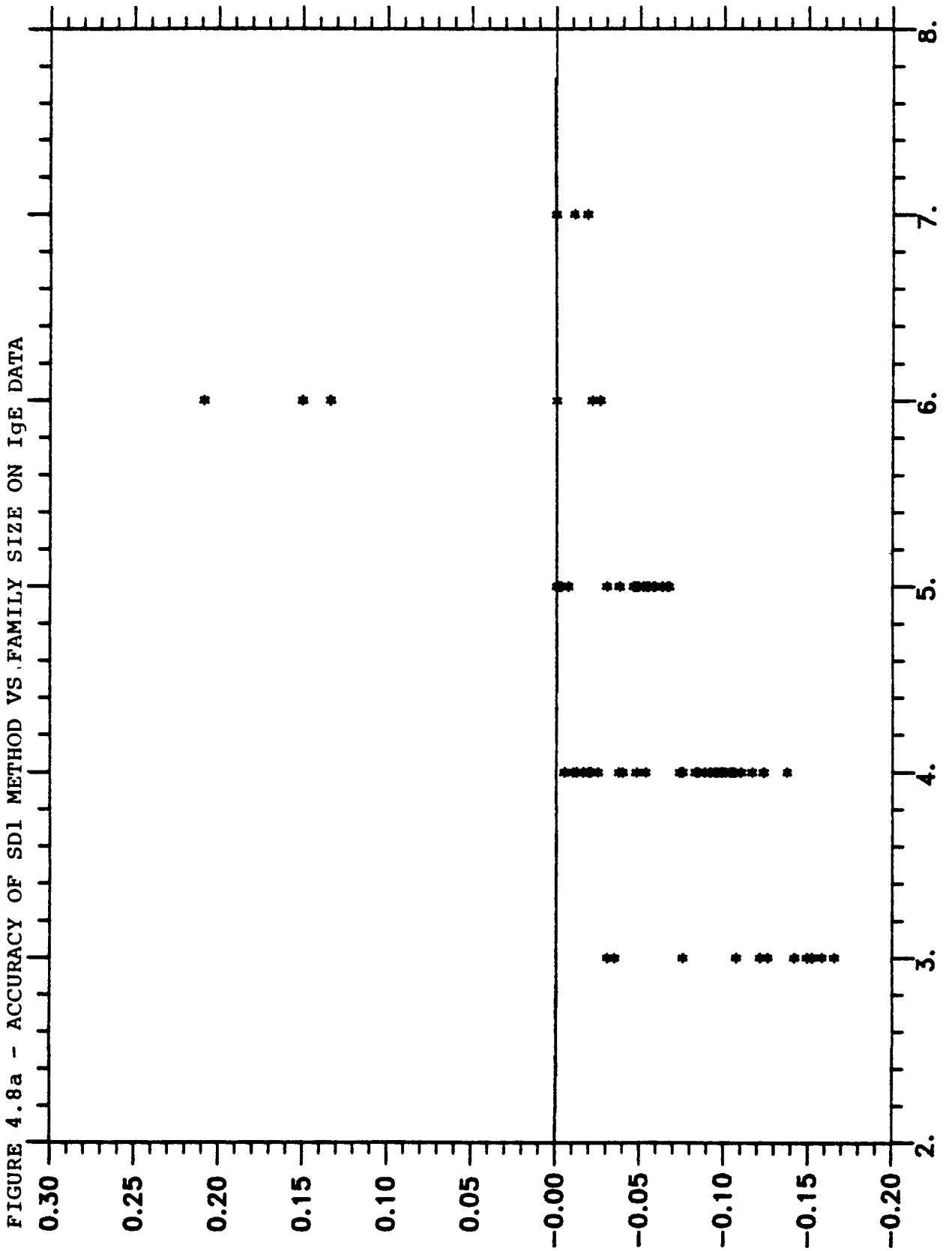


FIGURE 4.8a - ACCURACY OF SD1 METHOD VS FAMILY SIZE ON IgE DATA
DIFFERENCE
0.30
0.25
0.20
0.15
0.10
0.05
-0.00
-0.05
-0.10
-0.15
-0.20
2.
3.
4.
5.
6.
7.
8.
FAMILY SIZE

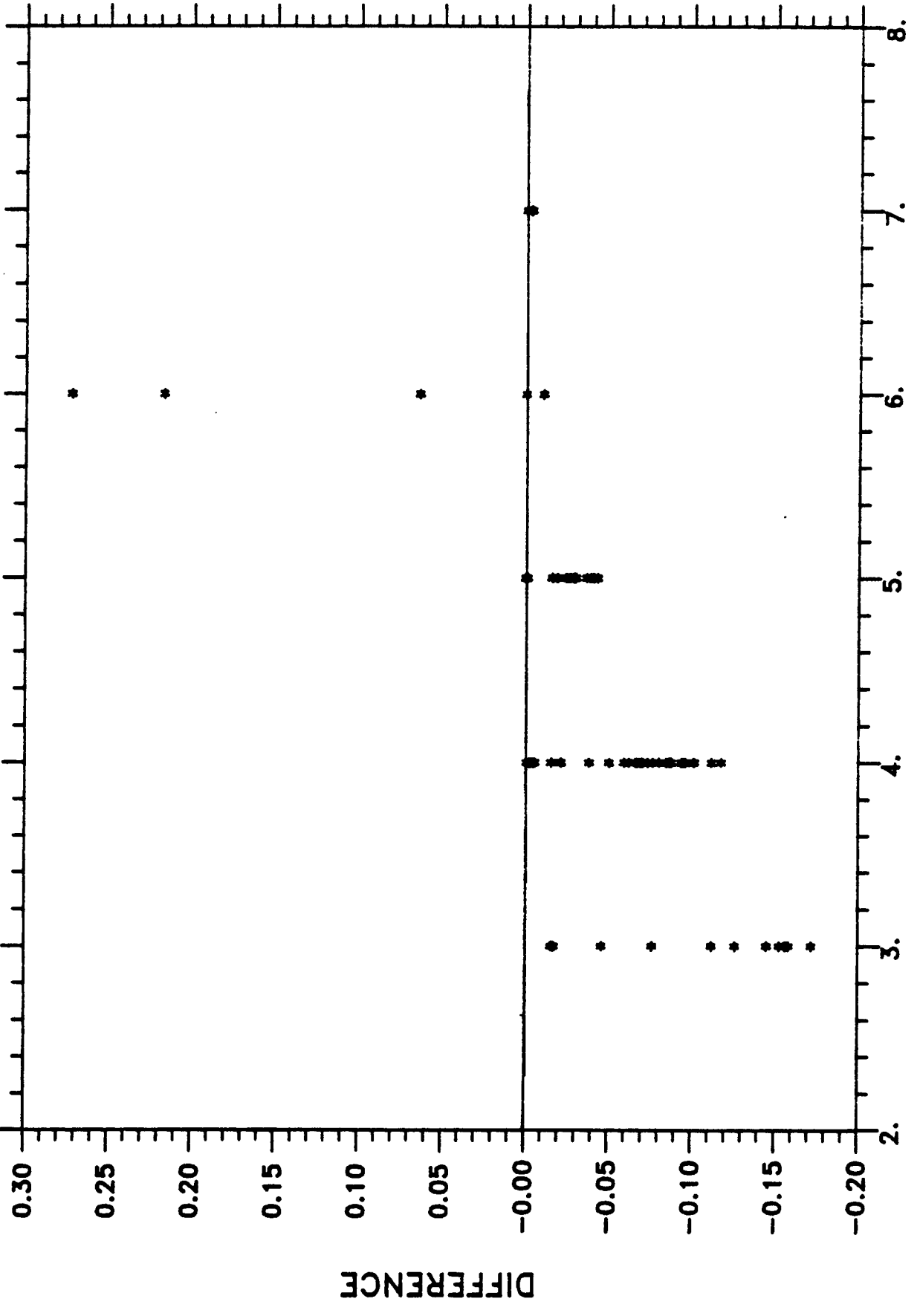
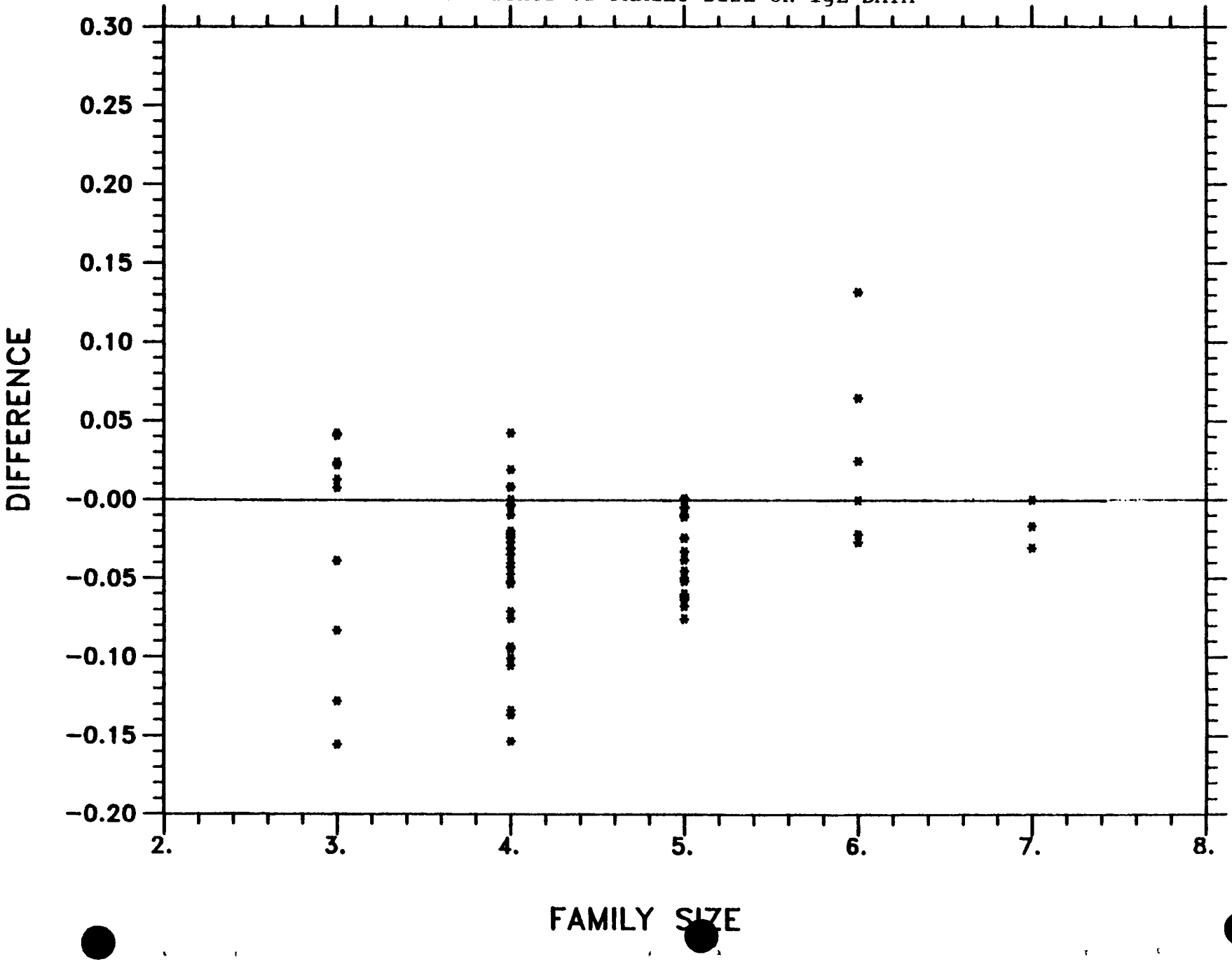


FIGURE 4.8b - ACCURACY OF SD2 METHOD VS FAMILY SIZE ON IgE DATA

FIGURE 4.8c - ACCURACY OF ML2 METHOD VS FAMILY SIZE ON IgE DATA



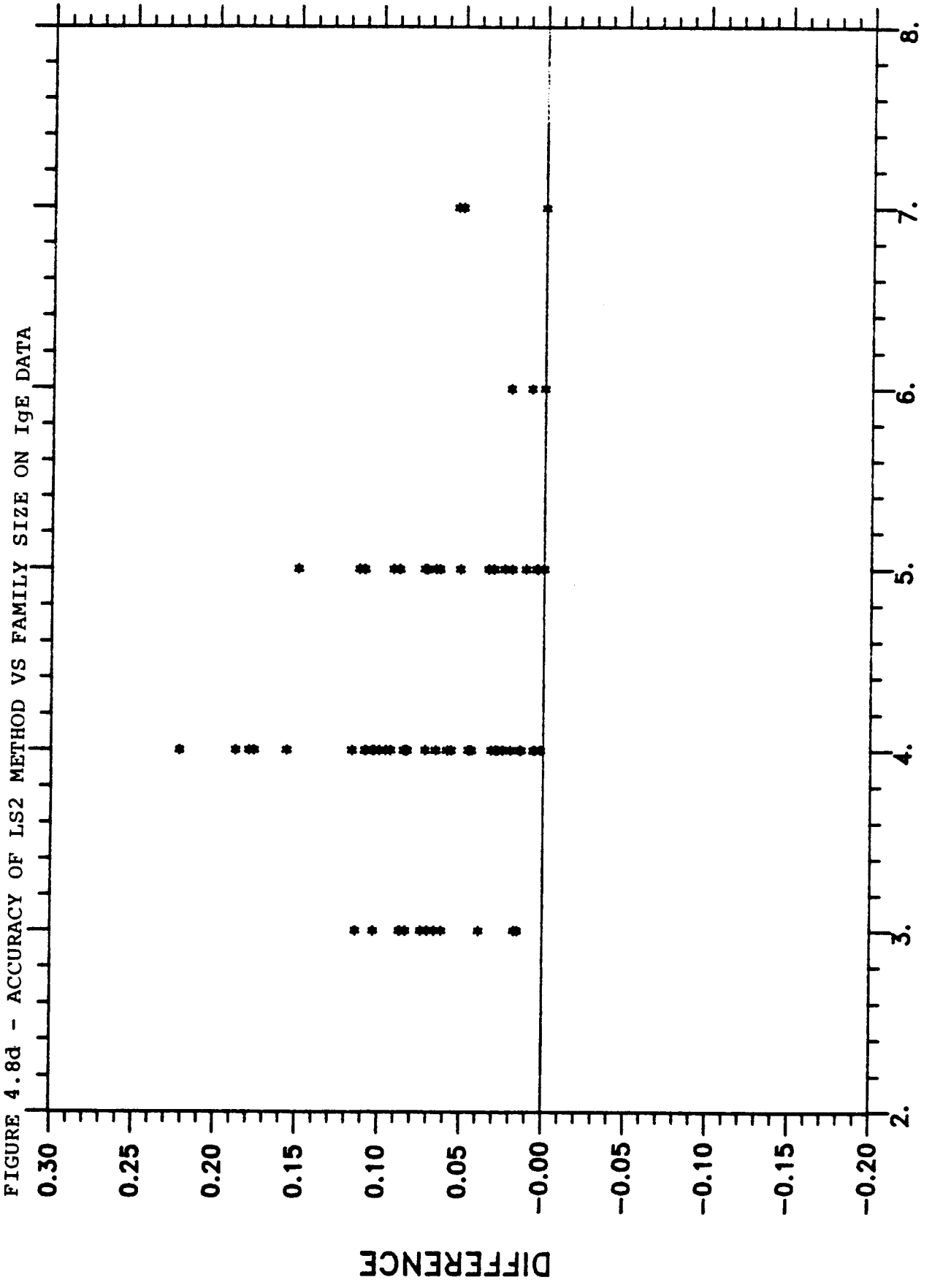


FIGURE 4.8d - ACCURACY OF LS2 METHOD VS FAMILY SIZE ON IgE DATA

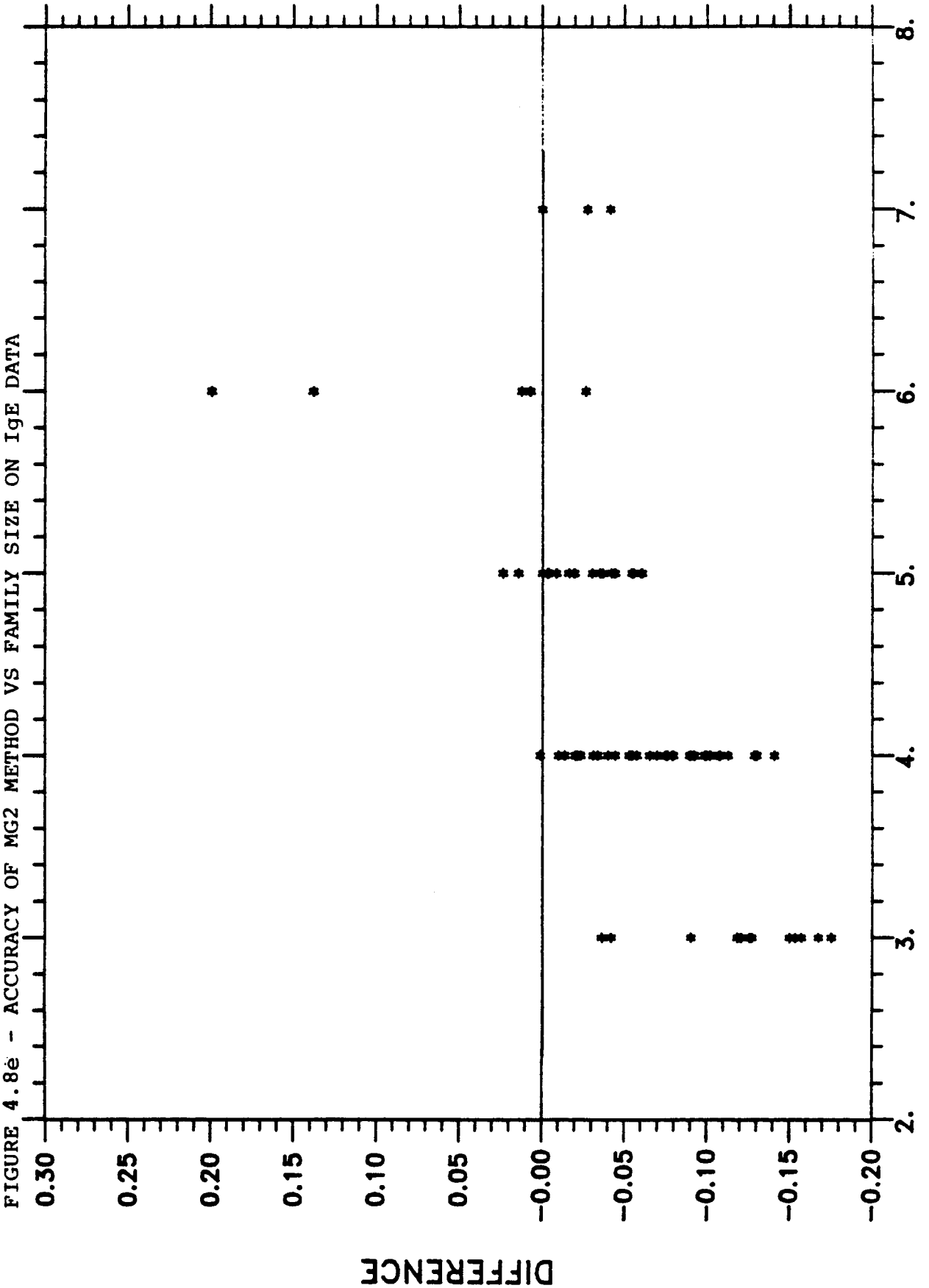


FIGURE 4.8e - ACCURACY OF MG2 METHOD VS FAMILY SIZE ON IgE DATA

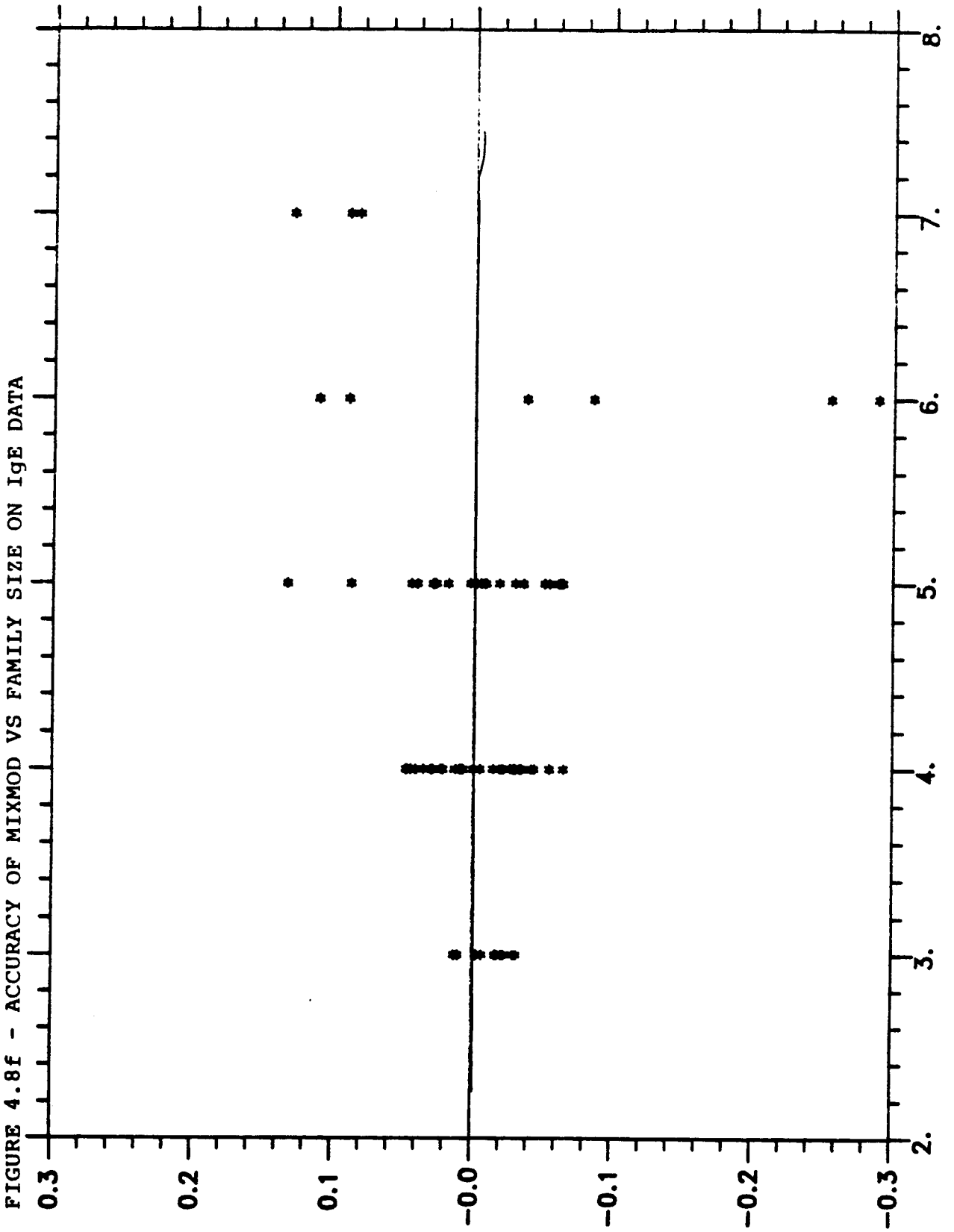


FIGURE 4.8f - ACCURACY OF MIXMOD VS FAMILY SIZE ON IgE DATA

CHAPTER V

THE SEGREGATION ANALYSIS OF IgE

Immunoglobulin E (IgE) is a serum protein associated with immunologic responses. Increasing levels of IgE have been found to be associated with allergic atopic disease (Johannson, 1967). There have been numerous studies done to determine whether the level of IgE is under genetic control (Grundbacher, 1967, Gerrard, et. al., 1978 and Ott, 1979). The analyses of Gerrard et. al. and Ott were done on the same dataset collected by Gerrard. The two groups have published conflicting results from the same data. Rao et. al. (1980) published results which attempt to resolve this conflict. The following discussion is a further attempt at clarification.

5.1. Description of the Data

The data were collected on 173 white Canadian two-generation families, with a total of 781 individuals. The selection of families was done randomly with respect to atopic disease. The size of the families ranged from three to twelve members. All analyses were done on age- and sex-adjusted data.

5.2. Summary of Previous Analyses

The data were originally analyzed by Gerrard et. al. (1978) using the model of Morton and MacLean which is discussed in section 2.1. It can be recalled that their method uses the conditional likelihood of the sibship phenotypes given the parents' phenotypes for estimation and hypothesis testing. They carried out segregation analysis on two transformations of the data, using the p-transformations discussed in 2.1.3. A mixture of two distributions was found to fit the data significantly better ($\chi^2_2 = 10.25$) than a single distribution. Using the more conservative transformed values (i.e. those using the p estimated from a single distribution) they found evidence for a major gene segregating in the data ($\chi^2_3 = 12.25$). As expected, using the less conservative transformation the conclusion of the presence of a major gene was confirmed. After publication of these results, Ott (1979) reanalyzed a subset of the data. He removed one family of size twelve from the data as his program could not handle such a large family. Using the joint likelihood of the parent and offspring phenotypes, and the transformed data which best fit the single distribution, he found no evidence for a single gene ($\chi^2 = 3.36$). There appears to be at least three possible reasons for the discrepancy between the investigators. They are the following:

1) Either the joint or the conditional likelihood is not appropriate for these data.

2) The difference in sample size, 769 vs. 781, causes the difference.

3) The accuracy of the numerical calculations may be questionable.

The importance of these proposed reasons is discussed in the following sections.

5.3. Analysis

To help resolve the differences between the investigators the program MIXMOD (Spence, et. al., 1979) was used. It computes both joint and conditional likelihoods for nuclear families, in a numerical method similar to that of Morton and MacLean, and with the same parameterization. Unless otherwise noted the computations in this section were done by MIXMOD.

In Table 5.1 are given the maximum log likelihood values for the unrestricted (q=t=d=0) models for relevant combinations of transformations and type of likelihood (conditional or joint) for the IgE dataset, with and without the single large family. The p=-.639 transformation is the transformation used by both Ott and Gerrard, et. al., and is given by

$$Y = \frac{r}{p}\left[\left(\frac{X}{r} + 1\right)^{p} - 1\right]$$

with p=-.639, and r = 6 as discussed in Chapter II. The value of p was obtained assuming the data were from a single normal

distribution. The N(0,1) transformation was obtained as follows:

Let Y = transformed value,

X = original value,

F(X) = sample cumulative distribution function,

$\Phi^{-1}(t)$ = inverse of cumulative normal distribution function,

then $Y = \Phi^{-1}(F(X))$,

The N(0,1) transformation was applied to the whole dataset as a single sample. The 2N(0,1) transformation was the same transformation as the N(0,1) transformation, but applied separately to each generation. To be conservative, the χ^2 values derived from the log likelihoods are assumed to have 3 degrees of freedom.

5.3.1. Sample Size

The joint likelihood was calculated for all three transformations, with and without the large family included. This large family accounted for one and one half percent (12/781) of the sample. With all three transformations the contribution it made to the χ^2 value was larger than expected given its relation to the total sample size. For the three transformations p=-.639,N(0,1),2N(0,1) the single family was responsible for 16%, 4%, and 29%, respectively of the difference between the log likelihoods of the unrestricted and restricted models. For the 2N(0,1) transformation the effect of this family is to change the conclusion of the test;

(see Table 5.1) with the large family included there is evidence for a major gene, without the family this is no longer true. The other two transformations yield nonsignificant conclusions for both the censored and full samples. The large differences in log likelihood caused by the censorship of the data demonstrate the importance of large families when studying genetic models. The differences suggest a great deal of caution be used if data are censored.

5.3.2. Numerical accuracy

It is difficult to ascertain the relative accuracy of the different programs. This is a result of the different parameter schemes for each of the programs and the fact that that parameter estimates are published to only two or three decimal place accuracy. Although not significantly statistically different, they are significantly numerically different. The problems with accuracy are illustrated in Table 5.1 where it can be seen that for the same dataset, using the p=-.639 transformation and the conditional likelihood, χ^2 values of 14.74 and 12.24 were obtained for the same hypothesis using two different programs. For the censored dataset, using the joint likelihood Ott has obtained a χ^2 value of 3.36 while MIXMOD obtained a value of 4.60. Although the different values of the statistics do not change the conclusions about the fit of the model, it should be recognized that numerical accuracy may present problems in marginally

TABLE 5.1

LOG LIKELIHOOD AND RESULTING χ^2 FOR HYPOTHESIS q=t=d=0

Transformation	Type of Likelihood	Sample	Unrestricted Log Likelihood	Restricted Log Likelihood	χ^2
p = -.639*	Conditional	All	558.50	564.62	12.24
p = -.639**	Joint	Censored	—	—	3.36
p = -.639	Conditional	All	570.17	577.54	14.74
p = -.639	Joint	All	1031.94	1034.68	5.48
p = -.639	Joint	Censored	1021.27	1023.57	4.60
N(0,1)	Conditional	All	590.08	595.78	11.40
N(0,1)	Joint	All	1059.03	1061.84	5.62
N(0,1)	Joint	Censored	1047.43	1050.12	5.38
2N(0,1)	Conditional	All	592.54	595.73	6.38
2N(0,1)	Joint	All	1055.51	1059.69	8.36
2N(0,1)	Joint	Censored	1043.99	1046.91	5.90

*Gerrard, et al. (1978) Published Results
**Ott (1979) Published Results

fitting models.

5.3.3. Joint vs Conditional Likelihood

From Table 5.1 it can be seen that the use of the conditional or joint likelihood is an important factor in determining whether the major gene model fits the data. For the p=-.639 transformation the χ^2 based on the conditional likelihood indicates strong evidence for a major gene while the value based on the joint likelihood does not. The N(0,1) transformation, which is more conservative than the p transformation since all the non-normality in the data which might simulate a major gene effect is removed, still demonstrates this discrepancy between the χ^2 based on different types of likelihood. The 2N(0,1) transformation is used to determine if heterogeneity between the two generations may have simulated a major gene effect. Here the test based on the conditional likelihood indicates the presence of a major gene. The test based on the joint likelihood gives conflicting answers. With censored sample there is a nonsignificant major gene effect while the complete sample indicates significance. Going by the rule that one should use all the data is making decisions, one can conclude that heterogeneity between generations did not simulate the major gene effect.

The results here concur with the results of Rao, et. al. (1980), who found that the choice of joint or conditional

likelihood can change the conclusion of the analysis. Apparently the transformations p=-.639 and N(0,1) eliminate the skewness caused by the major gene and result in heterogeneity between generations. This heterogeneity confounds the effect of a major gene when the joint likelihood is studied, but has no effect on the conditional likelihood. When this heterogeneity is eliminated in the 2N(0,1) transformation, the effect of the major gene is no longer confounded and is significant with either the joint or conditional likelihood.

CHAPTER VI

SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

In the previous discussion models for studying whether quantitative traits in man are under genetic control have been presented. Chapter II discussed these various models in detail. One of these models, the Morton-MacLean model, is limited in that it can only work with two-generation data and it is not robust against environmentally caused skewness. This limitation of size is significant as it is believed that large families with at least two generations are most informative when studying genetic hypotheses. The model using the approach outlined by Elston-Stewart does not suffer from this limitation. It is also parameterized in such a way that environmentally caused skewness should not cause a problem. The model is also parameterized so that the effects of the various factors which influence a particular trait can be estimated. These factors include a major gene effect, a polygenic effect, and effect due to common nuclear family, an effect due to common sibship and an unexplained random effect. In segregation analysis we are primarily interested in characterizing the major gene effect, the other effects

are essentially confounders. We wish to determine if this major gene effect is significant and, if so, the nature of its behavior, i.e. does it display dominant or codominant behavior.

As well as having presented the models, Chapter II presented the likelihoods for the models, making it possible to estimate the parameters and test hypotheses by applying the theory of maximum likelihood. This likelihood computation can be quite extensive, particularly for more than two-generation data, as it is essentially the sum of 3^N normal density functions where N is the number of individuals in the pedigree of interest. It becomes so extensive that we either need faster computers or a way of approximation to calculate the function.

Chapter III presented five methods of approximating the function. They all involve estimating the parameters of either a single or mixture of normal distributions. The general methods outlined are suitable for the model with all the effects mentioned earlier, as well as for more than two-generation data. The methods were investigated using only two-generation data, and with only the major gene and polygenic effect present. There is no theoretical problem with generalizing the procedure to more generations or introducing all the parameters; the time and patience of a programmer are all that are required.

The proposed methods of approximation were studied on up to twelve families of size 8, generated from three different genetic models. This size was chosen because this is the largest size family for which exact calculations are practical. For this size family and for this set of generated data none of the proposed methods could be said to be failures. However two methods seem to offer advantages over the others. The SD1 method was extremely quick, reasonably accurate, at least as accurate as all but the ML methods and unlikely to bias the results towards finding a genetic effect when none is actually present. The ML3 method was the slowest of all the methods proposed but, at least for the data generated here, it was the most accurate in approximating the true function. It seems that any future investigation should use a combination of these two methods. The SD1 method is probably suitable when initially investigating the likelihood surface. Because of its apparent superior accuracy the ML3 method is probably best used when final estimates are made.

A small subset of the data discussed in Chapter V was used to study the robustness of the methods of approximation on small families. The results indicate they are not robust for small families, but they do indicate that accuracy improves with larger families. The improvement in accuracy is dramatic and indicates that if an algorithm is drawn up such that approximations are made when at least six individuals are incorporated accurate approximations are possible.

Chapter V was a discussion of the segregation analysis of a large dataset of nuclear families. The trait of interest was the IgE level in blood serum samples. The discussion revolved around the importance of choosing a conditional or joint likelihood, the effect of censoring a sample and the problems of numerical accuracy. The results indicated that it is most likely that there is a major gene segregating in this dataset.

Probably the most important extension of the work presented here is to incorporate the approximation algorithm into a program able to handle more than two-generation data. As was mentioned earlier, this is not theoretically difficult, just time consuming. When this is accomplished it would be appropriate to evaluate the likelihood from a larger pedigree than was used here using more than one of the approximation methods. It is inpractical to compare these methods to the exact likelihood for a larger pedigree, but the comparison among the approximations would shed light on their suitability.

It also seems important in the future to add the various environmental correlations which were described previously but were not incorporated in the present program. Again there is no theoretical difficulty with this; the only requirement is a great deal of patience on the part of the programmer. Instead of evaluating the function $f(a_j + b_j, \underline{\theta})$ defined in section 3.2 the function $f(a_j + b_j - c_{j-1,j}, \underline{\theta})$ would be evaluated.

BIBLIOGRAPHY

Bartlett, M.S. and MacDonald, P.D.M. (1968). Least Squares Estimation of Distribution Mixtures, *Nature, 217,* 195-196.

Bhattacharya, C.G. (1967). A Simple Method of Resolution of a Distribution into Gaussian Components, *Biometrics, 23,* 115-137.

Boyle, C.R. and Elston, R.C. (1979). Multifactorial Genetic Models for Quantitative Traits in Humans. *Biometrics, 35,* 55-68.

Bryant, P. (1978). Comment on 'Estimating Mixtures of Normals Distributions', *Journal of the American Statistical Association, 73,* 748-749.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1976). The recursive derivation of likelihoods on complex pedigrees, *Advances in Applied Probability, 8,* 622-625.

Cannings, C., Thompson, E.A., and Skolnick, M.H. (1978). Probability functions on complex pedigrees, *Advances in Applied Probability, 10,* 26-61.

Clark, M.W. (1977). GETHEN: A Computer Program for the Decomposition of Mixtures of Two Normal Distributions by the Method of Moments, *Computers and Geosciences, 3,* 257-267.

Clarke, B.R. and Heathcote, C.F. (1978). Comment on 'Estimating Mixtures of Normal Distributions', *Journal of the American Statistical Association, 73,* 749-750.

Cohen, A.C. (1967). Estimation in Mixtures of Two Normal Distributions, *Technometrics, 9,* 15-28.

Day, N. E. (1969). Estimating the Components of a Mixture of Normal Distributions, *Biometrika, 56,* 463-474.

Dick, N. P. and Bowden, D.C. (1973). Maximum Likelihood Estimation for Mixtures of Two Normal Distributions, *Biometrics, 29,* 781-790.

Elandt-Johnson, R.C. (1971). Probability Models and Statistical Methods in Genetics, John Wiley and Sons, New York.

Elston, R.C. (1973). Ascertainment and age of onset in pedigree analysis, *Human Heredity, 23,* 105-112.

Elston, R.C. (1980). Segregation Analysis, In *Current Developments in Anthropological Genetics: Theory and Methods.* Eds. J.H. Mielke and M.H. Crawford, Vol. I, 325-352.

Elston, R.C., Namboodiri, K.K., Glueck, C.J., Fallat, R., Tsang, R. and Leuba, V. (1975). Study of the genetic transmission of hypercholesterolemia and hypertriglyceridemia in a 195 member kindred, *Annals of Human Genetics, 39,* 67-87.

Elston, R.C. and Stewart, J. (1971). A General Model for the Genetic Analysis of Pedigree Data, *Human Heredity, 21,* 523-542.

Elston, R.C. and Yelverton, K.C. (1975). General Models for Segregation Analysis, *American Journal of Human Genetics, 27,* 31-45.

Fryer, J.G. and Robertson, C.A. (1972). A Comparison of Some Methods for Estimating Mixed Normal Distributions, *Biometrika, 59,* 639-649.

Gerrard, J.W., Rao, D.C. and Morton, N.E. (1978). A Genetic Study of Immunoglobulin E, *The American Journal of Human Genetics, 30,* 46-58.

Go, R.C.P., Elston, R.C. and Kaplan, E.B. (1978). Efficiency and Robustness of Pedigree Segregation Analysis. *American Journal of Human Genetics, 30,* 28-37.

Gregor, J. (1969). An Algorithm for the Decomposition of a Distribution into Gaussian Components, *Biometrics, 25,* 79-93.

Grundbacher, F.J. (1975). Causes of Variation in Serum IgE Levels in Normal Population, *Journal of Allergy and Clinical Immunology, 56,* 104-111.

Hasselblad, V. (1966). Estimation of Parameters for a Mixture of Normal Distributions, *Technometrics, 8,* 431-444.

Hosmer, D.W. (1973a). On MLE of the Parameters of a Mixture of Two Normal Distributions When the Sample Size is Small, *Communications in Statistics, 1,* 217-227.

Hosmer, D.W. (1973b). A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of A Mixture of Two Normal Distributions Under Three Different Types of Sample, *Biometrics, 29,* 761-770.

Heathcote, C.R. (1977). The Integrated Squared Error Estimation of Parameters, *Biometrika, 64,* 255-264.

Johannson, S.G.O. (1967). Raised Levels of a New Immunoglobulin Class (IgND) in Asthma, *The Lancet, II,* 951-953.

Kaplan, E.B. and Elston, R.C. (1978). A Subroutine for Maximum Likelihood Estimation (MAXLIK). Institute of Statistics Mimeo Series No. 823, University of North Carolina.

Lange, K. and Elston, R.C. (1975). Extensions to Pedigree Analysis. I. Likelihood Calculations for Simple and Complex Pedigrees, *Human Heredity, 25,* 95-105.

Lee, K.K. (1978). A Genetic Analysis of Serum Cholesterol and Blood Presure Levels in a Large Pedigree, Institute of Statistics Mimeo Series No. 1174, University of North Carolina at Chapel Hill.

MacLean, C.J., Morton, N.E., Lew, R. (1975). Analysis of Family Resemblance. IV. Operational Characteristics of Segregation Analysis, *American Journal of Human Genetics, 27,* 365-384.

Morton, N.E. and MacLean, C.J. (1974). Analysis of Family Resemblance. III. Complex Segregation of Quantitative traits, *American Journal of Human Genetics, 26,* 489-503.

Murphy, E.A. and Bolling, D.R. (1967). Testing of Single Locus Hypothesis Where There is Incomplete Separation of the Phenotypes, *American Journal of Human Genetics, 19,* 322-334.

Ott, J. (1979). Maximum Likelihood Estimation by Counting Methods Under Polygenic and Mixed Models in Human Pedigrees, *American Journal of Human Genetics, 31,* 161-75.

Paulson, A.S., Holcomb, E.W., and Leitch, R.A., (1975). The Estimation of the Parameters of the Stable Laws, *Biometrika, 62,* 163-170.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution, *Philosophical Transactions of the Royal Society of London, 185,* 71-110.

Preston, E.J. (1953). A Graphical Method for Analysis of Statistical Distributions into Two Normal Components, *Biometrika, 40,* 460-464.

Quandt, R.E. and Ramsey, J.B. (1978). Estimating Mixtures of Normal Distributions and Switching Regressions, *Journal of the American Statistical Association, 73,* 730-737.

Rao, C.R. (1948). Advanced Statistical Methods in Biometric Research, John Wiley and Sons, New York.

Rao, D.C., Lalouel, J.M., Morton, N.E., and Gerrard, J.W. (1980). Immunoglobulin E Revisited, *American Journal of Human Genetics*, 32, 620-625.

Robertson, C.A. and Fryer, J.G. (1970). The Bias and Accuracy of Moment Estimators, *Biometrika*, 57, 57-65.

Spence, M.A., Westlake, J. and Lange, K. (1979). MIXMOD: A Mixed Model Segregation Analysis Package (Unpublished Program Documentation).

Tan, W.Y. and Chang, W.C. (1972). Comparisons of Method of Moments and Methods of Maximum Likelihood in Estimating Parameters of a Mixture of Two Normal Densities, *Journal of the American Statistical Association*, 67, 702-709.