MAXIMUM LIKELIHOOD METHODS FOR GENETIC ANALYSIS OF MULTIVARIATE PEDIGREE DATA

by

George Ebow Bonney

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1408

July 1982

MAXIMUM LIKELIHOOD METHODS FOR GENETIC ANALYSIS OF MULTIVARIATE PEDIGREE DATA

by

GEORGE EBOW BONNEY

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

Gary G. Koch

Reader

K K Namboodiri

Reader

TABLE OF CONTENTS

LIST OF FIGURES

Figure | Page

LIST OF TABLES

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

No doubt the starting point of all genetical thought was the common observation of resemblance between relatives. Yet the degree of this resemblance depends remarkably on the trait observed. Thus a child may resemble a parent with respect to one trait but more strikingly resemble a more distant relative with respect to some other trait. The genetic basis of the covariation of different traits is naturally of general biological interest. However we shall concern ourselves here with our own species.

In human genetics this interest arises quite naturally when the traits under study relate to a disease such as hypertension; the genetic basis of covariation of the many symptoms and predisposing factors may fruitfully be investigated.

There are two basic questions that can be asked in such an analysis.

1. Is there a single disorder or several? e.g., does one gene control obesity, another subcutaneous fat thickness, etc., or is there a single major fundamental genetic defect that causes the whole spectrum of characteristics?
2. What may be used as a measure of the "innate" trait? Such an index can be used for further genetic analysis (e.g. to determine the position of the underlying genes relative to

some known genes), and also to identify individuals at risk.

Our aim in this dissertation is to develop statistical procedures based on the genetic components of covariance, for answering these questions. We are motivated by the following considerations:

Denote by $\underset{\sim}{x}$ the measurement of a p-variate trait on an individual, and suppose that in the population the covariance matrix of $\underset{\sim}{x}$ may be partitioned into genetic and environmental components as follows

$$V = V_{mg} + V_{pg} + V_e$$

where $V_{mg} + V_{pg} = V_g$ is the contribution to the covariance matrix by genetic factors, broken down into V_{mg} representing the contribution of a major gene to the genetic covariance, and V_{pg} representing the contribution of many separately indistinguishable genes. Now consider the linear index

$$I(\underset{\sim}{x}) = \underset{\sim}{a}'\underset{\sim}{x} ;$$

its variance may be partitioned as follows

$$\underset{\sim}{a}'V\underset{\sim}{a} = \underset{\sim}{a}'V_g\underset{\sim}{a} + \underset{\sim}{a}'V_e\underset{\sim}{a} = \underset{\sim}{a}'V_{mg}\underset{\sim}{a} + \underset{\sim}{a}'V_{pg}\underset{\sim}{a} + \underset{\sim}{a}'V_e\underset{\sim}{a}.$$

$I(\underset{\sim}{x})$ may be used as a measure of the innate trait if it has a high genetic variance relative to its total variance. If there is a single genetic defect the major gene component of the genetic variance will be relatively large. Thus the two basic questions may be simply answered by examining the eigen structures of $V^{-1}V_{pg}$ and $V^{-1}V_{mg}$.

The problem is that V_{mg}, V_{pg} and V_e are not known. Hence we shall consider first the statistical estimation of the genetic components of covariance. It is also of interest to know whether or not all of these components are statistically important. We shall describe maximum likelihood methods for estimation and hypothesis testing.

We are dealing here with the human species. Experimentation is therefore generally precluded. The data will typically comprise observations on related individuals--families and extended families, i.e., pedigrees. Because of heterogeneity a single large pedigree is better than a large number of nuclear families analyzed together (Elston and Rao, 1978). The rich correlational architecture of pedigrees unfortunately creates problems in modeling and analysis. Lack of independence between observations removes pedigree analysis from the main stream of statistical inference methodology. It also creates computational problems.

Lange, Boehnke and Spence (1981) have described a model for arbitrary pedigree structure with multivariate traits, but it only allows for polygenic inheritance and is thus not general. In particular it cannot be used for the single gene analysis required to answer the questions raised above. Elston and Stewart (1971) presented a general model that allows a wide variety of genetic mechanisms to be specified. However,pedigrees of arbitrary structure can only be analyzed on the assumption of single or few genetic loci. There are also computational problems for some genetic mechanisms even for simple pedigrees; the accurate calculation of the likelihood under the mode of inheritance involving both major and polygenic loci, the mixed model, is not feasible even with our present day computing facilities. Moreover, the model is essentially meant for univariate traits, although multivariate generalizations for some genetic models are trivial and are implied in Elston and Stewart's (1971) paper and used explicitly by Beaty (1978) and Simpson (1981).

We shall develop a model for multivariate traits that has all the features of the Elston and Stewart model, and includes the essentials of the Lange, Boehnke and Spence model. The computational problem is studied analytically in considerable depth, yielding recurrence formulas for the exact calculation of the likelihood for even the mixed model, and for arbitrary pedigree structure.

The rest of this chapter is devoted to a review of the literature. The objective is to present a simplified expository description of the basic statistical models in genetics, the main emphasis being on those notions which are essential to a proper understanding of the theory and methodology developed later. In particular we shall fix the definitions and explain concisely the origin and use of the genetic components of covariance that enter into our model for quantitative traits. Secondly, we shall describe the Lange, Boehnke and Spence, and Elston and Stewart models, and briefly review the methods for testing common genetic control (pleitropy) in multivariate traits.

1.2 The Simple Mendelian Model and Fisher's Decomposition

Genetic traits are due ultimately to extremely small 'particles' or genes, which are present in every nucleate cell of the body. The scientific understanding of the transmission of genetic traits began with the two laws of Gregor Mendel. The law of segregation is the first. It states that traits are controlled by pairs of genes which segregate or separate during the formation of the reproductive cells, thus passing into different gametes. The pairs are restored when fertilization occurs, and this leads to the production of different types of offspring in certain definite proportions. The term Mendelian inheritance is used to refer to this phenomenon. Segregation analysis is the set of procedures that enable Mendelian inheritance to be inferred.

The thrust of Mendel's First Law lies in its use to make statistical predictions of the consequences of different kinds of mating. Consider a gene with two alternative forms (alleles): A and a. Since an individual receives either A or a from each parent, his genotype is AA, Aa, or aa. With respect to this gene then, there are six mating types in the population: AA x AA, AA x Aa, AA x aa, Aa x Aa, Aa x aa, aa x aa. For example, if both parents are heterzygotes (Aa), then the mating type is Aa x Aa and an offspring will be AA, Aa or aa with probabilities ¼, ½, ¼ respectively. Now suppose we have a large population with q of A and 1-q of a. If the proportions of the three genotypes AA, Aa and aa are q^2, $2q(1-q)$ and $(1-q)^2$ respectively, then if the population is large enough so that probabilities correspond to frequencies,we can easily verify, using the elementary probability laws and taking account of the six mating types, that on random mating the genetic structure of the population remains unchanged, i.e., the offspring distribution is also q^2, $2q(1-q)$ and $(1-q)^2$ for the three genotypes. This is the Hardy-Weinberg Law.

The genotype of an individual is distributed over the population as a random variable. Denote it by the letter t, and the population distribution by ψ_t. For a quantitative p-vector trait $\underset{\sim}{z}$, let the contribution of the genotype t to its variablity be $\underset{\sim}{\mu}_t$. Then for the simple two-allele case above, the genetic contribution to the mean of $\underset{\sim}{z}$ is

$$\begin{aligned} \underset{\sim}{\mu}_g &= \sum_t \psi_t \underset{\sim}{\mu}_t \\ &= q^2 \underset{\sim}{\mu}_{AA} + 2q(1-q) \underset{\sim}{\mu}_{Aa} + (1-q)^2 \underset{\sim}{\mu}_{aa}, \end{aligned} \quad (1.1)$$

and the covariance matrix is

$$\begin{aligned} V_g &= \sum_t \psi_t \underset{\sim}{\mu}_t \underset{\sim}{\mu}'_t - \underset{\sim}{\mu}_g \underset{\sim}{\mu}'_g \\ &= q^2 \underset{\sim}{\mu}_{AA} \underset{\sim}{\mu}'_{AA} + 2q(1-q) \underset{\sim}{\mu}_{Aa} \underset{\sim}{\mu}'_{Aa} + (1-q)^2 \underset{\sim}{\mu}_{aa} \underset{\sim}{\mu}'_{aa} \qquad (1.2) \\ &\quad - \underset{\sim}{\mu}_g \underset{\sim}{\mu}'_g . \end{aligned}$$

It should be noted that in this formulation the random variable t is the underlying factor for the p-vector effect $\underset{\sim}{\mu}_t$, so the genetic covariance between two components of the vector trait $\underset{\sim}{z}$ is the population mean of the cross-products of effects due to the same genotypes, corrected for the product of the means.

For quantitative traits it is helpful to think in terms of the effects of the genes (alleles) that constitute the genotype. Denote the paternal allele of an individual by v and the maternal allele w. Then we can write

$$\underset{\sim}{\mu}_{vw} = \underset{\sim}{\alpha}_v + \underset{\sim}{\alpha}_w + \underset{\sim}{\delta}_{vw} \qquad (1.3)$$

where $\underset{\sim}{\alpha}_v$ and $\underset{\sim}{\alpha}_w$ respectively denote the paternal and maternal gene effects, and $\underset{\sim}{\delta}_{vw}$ represents the interaction effect between v and w. For a quantitative trait, the genotype of an individual is generally unknown so the usual 'linear model' definitions of $\underset{\sim}{\alpha}_v$, $\underset{\sim}{\alpha}_w$ and δ_{vw} cannot be used. We choose $\underset{\sim}{\alpha}_v$ and $\underset{\sim}{\alpha}_w$ to minimize

$$(\underset{\sim}{\mu}_{vw} - \underset{\sim}{\alpha}_v - \underset{\sim}{\alpha}_w)'(\underset{\sim}{\mu}_{vw} - \underset{\sim}{\alpha}_v - \underset{\sim}{\alpha}_w) \qquad (1.4)$$

summed over the whole population, and set

$$\underset{\sim}{\delta}_{vw} = \underset{\sim}{\mu}_{vw} - \underset{\sim}{\alpha}_v - \underset{\sim}{\alpha}_w$$

which implies $\sum_v q_v \underset{\sim}{\alpha}_v = \underset{\sim}{0}$, $\sum q_w \underset{\sim}{\alpha}_w = \underset{\sim}{0}$,

assuming without loss of generality that $\sum_v \sum_w q_v q_w \underset{\sim}{\mu}_{vw} = \underset{\sim}{0}$, where q_v and q_w denote the frequencies of v and w, so that

$$\underset{\sim}{\alpha}_v = \sum_q q_w \underset{\sim}{\mu}_{vw} , \quad \underset{\sim}{\alpha}_w = \sum_v q_v \underset{\sim}{\mu}_{vw}$$

and $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (1.5)

$$\underset{\sim}{\delta}_{vw} = \underset{\sim}{\mu}_{vw} - \sum_w q_w \underset{\sim}{\mu}_{vw} - \sum_v q_v \underset{\sim}{\mu}_{vw} .$$

This is the multivariate generalization of Fisher's (1918) decomposition for a single genetic locus (Lange et al., 1981). Thus Fisher defined the additive effect of a gene as "equal to the mean of the values of the genotypes which contain the gene, weighted by the frequencies of the other genes in those genotypes." When there is no dominance the additive effects are independent of gene frequencies.

We note that Fisher's decomposition is arbitrary. It is based on the assumption that the dominance effects can be defined such that the sum of squares of their values is minimized. The values $\underset{\sim}{\alpha}_v$ are clearly not related simply to the physiological effects of the genes, since they depend on the frequencies of other genes. The $\underset{\sim}{\alpha}_v$ can therefore be different in different populations, if the populations have different gene frequencies even if the genotypic values $\underset{\sim}{\mu}_{vw}$ are the same. However Fisher's decomposition has the advantage of defining additive and dominance effects which have zero covariance so that the genetic component of the variance of $\underset{\sim}{z}$ may be partitioned into

$$V_g = V_a + V_d$$

where

$$\begin{aligned} V_a &= \sum_{v,w} q_v q_w (\underset{\sim}{\alpha}_v + \underset{\sim}{\alpha}_w)(\underset{\sim}{\alpha}_v + \underset{\sim}{\alpha}_w)' \\ &= 2 \sum_v q_v \underset{\sim}{\alpha}_v \underset{\sim}{\alpha}'_v \end{aligned} \tag{1.7}$$

and

$$V_d = 4 \sum_{v>w} q_v^2 q_w^2 \, \underset{\sim}{d}_{vw} \underset{\sim}{d}'_{vw} \qquad (1.8)$$

where

$$\underset{\sim}{d}_{vw} = \underset{\sim}{\mu}_{vw} - \tfrac{1}{2}(\underset{\sim}{\mu}_{vv} + \underset{\sim}{\mu}_{ww}). \qquad (1.9)$$

This simple exposition does not take into account differential survival of the genotypes, or viabilities of the gametes, which lead to the phenomenon of selection. We have also assumed that there is no mutation, i.e., genes do not change their forms, and that for qualitative traits individuals are classified without error according to genotypes. The modifications needed when these assumptions do not hold form the bulk of Mathematical Population Genetics, which we shall not go into.

If the trait is controlled by two or more loci epistatic interactions (interactions of genes at different loci) must also be considered. The basic model (1.3) becomes

$$\underset{\sim}{\mu} = \underset{\sim}{a} + \underset{\sim}{\delta} + \underset{\sim}{\gamma} \qquad (1.10)$$

where $\underset{\sim}{a}$, $\underset{\sim}{\delta}$ and $\underset{\sim}{\gamma}$ denote respectively the additive, dominance and epistatic interaction effects. The genetic component of variance may therefore be partitioned accordingly into

$$V_g = V_a + V_d + V_\gamma + 2\,\mathrm{cov}(\underset{\sim}{a},\underset{\sim}{\delta}) + 2\,\mathrm{cov}(\underset{\sim}{\delta},\underset{\sim}{\gamma}) + 2\,\mathrm{cov}(\underset{\sim}{a},\underset{\sim}{\gamma}) \qquad (1.11)$$

where, by suitable definition of $\underset{\sim}{\alpha}$, $\underset{\sim}{\delta}$ and $\underset{\sim}{\gamma}$ the covariances are made zero. Cockerham (1954), using a system of orthogonal coordinates, further partitioned the epistatic variance into four components. As far as the author is aware Cockerham's decomposition of epistatic variance has not been used for the analysis of human data, and so will not be described here.

1.3 Identity and Kinship Coefficients

Fisher (1918) used his decomposition of genetic variance for a quantitative trait to calculate the genetic correlation between relatives. It is remarkable that the genetic covariances between relatives turn out to be linear functions of the components of genetic variance. General formulas for the genetic covariance of any two individuals are available in terms of one of the following measures of relatedness: kinship coefficient (coefficient de parenté, Malécot 1948), coefficients of identity by descent (Gillois 1964), or the condensed coefficients of identity (Jacquard 1974). We shall now briefly discuss these measures and show how they have been used to partition the genetic covariance between relatives.

However complex the relationship between two individuals i and j, the genetic implication of this relationship is simply that there is a positive probability that a gene in one and a corresponding gene in the other may both be copies of a given gene in one of their common ancestors. These measures of relatedness therefore depend on the notion of identity of genes at the same locus. Two genes g_1 and g_2 are identical by descent (IBD) if one is a physical copy of the other or they are both physical copies of the same ancestral gene. The kinship coefficient Φ_{ij} of two individuals i and j is the probability that a gene selected randomly from i and a gene selected randomly from the same locus of j are IBD. If i and j are the same person, we are more interested in the inbreeding coefficient f_i, defined as the probability that the two genes he possesses at a given locus are IBD. The inbreeding coefficient of an individual is equal to the kinship coefficient of his parents, i.e., $f_i = \Phi_{k\ell}$, where k and ℓ are the parents of i. If the parents are not related $f_i = 0$. If $f_i > 0$, i is said to be inbred. The kinship coefficient of the individual i and his inbreeding coefficient are related: $\Phi_{ii} = \frac{1}{2}(1 + f_i)$.

At a given autosomal locus, the two individuals i and j have a total of four genes. Suppose u_i is the gene transmitted to i by his father, and v_i that by his mother. Correspondingly for j we have u_j and v_j. Depending on the pedigrees of i and j, these genes can be identical with one another or not. The relation of identity is transitive: two genes which are each identical to a third gene are identical to each other. Taking this identity into account Gillois (1964) has enumerated 15 possible cases ('identity modes') with respect to the identity of the four genes of i and j. By analysis of the pedigree which contains i and j, we can attach probabilities $\delta_1, \delta_2, \ldots, \delta_{15}$ to these states. These probabilities are called "coefficients of identity by descent" of i and j. They contain all the information about the relationship of i and j that we require for genetic purposes.

Thinking in terms of the genotypes of individuals without regard to which genes came from which of his parents, Jacquard (1974) reduced the fifteen states to nine: $\Sigma_1, \Sigma_2, \ldots, \Sigma_9$ and called the associated probabilities $\Delta_1, \Delta_2, \ldots, \Delta_9$ "condensed coefficients of identity." Writing "$u_i \sim v_j$" for "u_i and v_j are IBD," and "iff" for "if and only if," the condensed states are:

Σ_1 iff $u_i \sim v_i \sim u_j \sim v_j$

Σ_2 iff $u_i \sim v_i$ and $u_j \sim v_j$

Σ_3 iff $u_i \sim v_i \sim u_j$ or $u_i \sim v_i \sim v_j$

Σ_4 iff $u_i \sim v_i$

Σ_5 iff $u_i \sim u_j \sim v_j$ or $v_i \sim u_j \sim v_j$

Σ_6 iff $u_j \sim v_j$

Σ_7 iff $u_i \sim u_j$ and $v_i \sim v_j$, or $u_i \sim v_j$ and $v_i \sim u_j$

Σ_8 iff $u_i \sim u_j$ or $u_i \sim v_j$ or $v_i \sim u_j$ or $v_i \sim v_j$

Σ_9 iff none of the 4 genes are i.b.d.

It follows that

$$\Phi_{ij} = \Delta_{1ij} + \tfrac{1}{2}(\Delta_{3ij} + \Delta_{5ij} + \Delta_{7ij}) + \tfrac{1}{4}\Delta_{8ij}, \qquad (1.12)$$

$$f_i = \Delta_{1ij} + \Delta_{2ij} + \Delta_{3ij} + \Delta_{4ij}, \qquad (1.13)$$

$$f_j = \Delta_{1ij} + \Delta_{2ij} + \Delta_{5ij} + \Delta_{6ij}, \qquad (1.14)$$

and if neither i nor j is inbred,

$$\Phi_{ij} = \tfrac{1}{2}\Delta_{7ij} + \tfrac{1}{4}\Delta_{8ij} \, . \qquad (1.15)$$

Lange et al. (1976) discuss algorithms for computing the symmetric matrices $\Phi = (\Phi_{ij})$ and $\Delta_7 = (\Delta_{7ij})$. Lange et al. (1981) have derived the genetic covariance of a trait X_i measured on the individual i and a trait Y_j measured on the individual j. We shall briefly present their result. Suppose X_i and X_j are determined by the same autosomal locus whose r-th allele has frequency q_r. Denote the decomposition of X_i into additive and dominance components by $\alpha_r + \alpha_s + \delta_{rs}$ and that of Y_j by $\beta_t + \beta_u + \beta_{tu}$. Then assuming $E(X_i) = E(Y_i) = 0$ the genetic component of covariance is

$$\begin{aligned}\mathrm{cov}_G(X_i, Y_j) = {} & \Delta_{7ij} \sum_{rs} (\alpha_r + \alpha_s + \delta_{rs})(\beta_r + \beta_s + \gamma_{rs}) q_r q_s \\ & + \Delta_{8ij} \sum_{rst} (\alpha_r + \alpha_s + \delta_{rs})(\beta_r + \beta_t + \delta_{rt}) q_r q_s q_t \\ & + \Delta_{9ij} \sum_{rstu} (\alpha_r + \alpha_s + \delta_{rs})(\beta_t + \beta_u + \delta_{tu}) q_r q_s q_t q_u\end{aligned}$$

$$= 2\Phi_{ij}(2 \sum_{r} \alpha_r\beta_r q_r) + \Delta_{7ij} \sum_{rs} \delta_{rs}\gamma_{rs}q_r q_s$$

$$= 2\Phi_{ij}\sigma_{axy} + \Delta_{7ij}\ \sigma_{dxy} \qquad (1.16)$$

where $\sigma_{axy} = \mathrm{cov}_a(X,Y)$ and $\sigma_{dxy} = \mathrm{cov}_d(X,Y)$ (Lange et al., 1981). The partition of genetic covariance in (1.16) is valid only if neither i nor j is inbred. If either i or j is inbred but there is no dominance, i.e., $\sigma_{dxy} = 0$, then

$$\mathrm{cov}_G(X_i,Y_j) = 2\Phi_{ij}\sigma_{axy} \qquad (1.17)$$

is still correct, for the following reason. Gene frequencies in the offspring population differ from that of the parent population when there is inbreeding, but in the absence of dominance, the additive genetic effects, and consequently the additive genetic variance, do not depend on gene frequency.

One important use of the kinship coefficient is in the derivation of the distribution of the genetic effects on quantitative traits, of a large number of loci which are separately indistinguishable--the so called polygenic loci--over biologically related individuals. The work we shall now describe was originated by Fisher (1918), but has received more formal treatment only recently by Lange (1978), who has presented central limit theorems for polygenic traits over pedigrees of arbitrary structure. Here we shall state only one special case of Lange's result, for later use. Let ξ_r be the effect of the r-th locus on a univariate quantitative trait, with corresponding additive and dominance variances $\sigma^2_{a(r)}$ and $\sigma^2_{d(r)}$ respectively. Assume $E(\xi_r) = 0$ and further that the locus effects are additive. Now let

$$S_n = \sum_{r=1}^{n} \xi_{(r)} \text{ and } s^2_n = \sum_{r=1}^{n} (\sigma^2_{a(r)} + \sigma^2_{d(r)}) \qquad (1.18)$$

and let $\underset{\sim}{S}_n$ and $\underset{\sim}{s}_n$ be Nx1 vectors of quantities S_n and s_n for each of N individuals in a pedigree.

Then we may state the following special case of Lange's theorem.

Theorem 1.1 (Lange 1978)

$$\frac{\underset{\sim}{S}_n}{\underset{\sim}{s}_n} \xrightarrow{p} N_N(\underset{\sim}{0}, 2\Phi\sigma_a^2 + \Delta\sigma_d^2) \qquad (1.19)$$

where $\Phi = (\Phi_{ij})$, $\Delta = (\Delta_{ij})$, i,j are individuals in the pedigree,provided

(i) the sequence $\xi_{(1)}, \xi_{(2)}, \ldots,$ is m-dependent, m finite;

(ii) for some M and $\delta > 0$,

$$E\{|\xi_{(r)}|^{2+\delta}\} \leq M,$$

(iii) Both

$$\lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} \sigma^2_{a(r)} \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} \sigma^2_{d(r)} \qquad (1.20)$$

exist and at least one of them is positive.

It is important to note that the limits here refer to the number of genetic loci, not the number of individuals in the pedigree. Thus the result holds for single individuals as well. The three sufficient conditions for asymptotic normality are all plausible for any group of individuals. The first is the requirement that the sequence $\xi_1, \xi_2, \ldots$ be m-dependent, i.e., ξ_k and ξ_ℓ are independent whenever $|k-\ell| > m$. This condition prevents clustering of loci, and is automatically satisfied if loci k and ℓ are on different chromosomes. The second requirement is the boundedness condition, which is always satisfied since the effect of any locus on the trait is necessarily finite. The last condition is merely a requirement that the variances stabilize as the number of loci increases. Extensions of the results to multivariate traits are given by Lange et al. (1981). Necessary conditions for asymptotic normality are as yet unknown. The rate of convergence is also not known in general terms, although Elston (1980) has given a demonstration that approximate normality could hold for as few as three equal and additive loci.

1.4 Modeling a Quantitative Phenotype

Quantitative traits like height and weight show continuous variation in the total range of the phenotype. The theory we have outlined above for the genetic basis of such traits is based "on the supposition of Mendelian inheritance," Fisher (1918), i.e., that the genes governing the traits are transmitted in Mendelian fashion. But on to their effects must be superimposed effects of the environment. To a particular genotype, therefore, there corresponds not a single measure, but a set of measures of the trait. The genotype therefore defines a frequency distribution of phenotype values, and consequently the quantitative trait may be regarded as a random variable in a statistical sense, dependent on the genotype and environment. The decomposition of observed statistics (in particular variances and covariances) into meaningful genetic and environmental components is basic to the study of these traits. A linear additive model is commonly assumed. The usual p-variate model is

$$\underset{\sim}{z} = \underset{\sim}{g} + \underset{\sim}{c} + \underset{\sim}{e} \qquad (1.21)$$

where $\underset{\sim}{z}$ is the vector phenotypic value ,

$\underset{\sim}{g}$ is the vector genotypic value,

$\underset{\sim}{c}$ is the common or family environmental effect,

and $\underset{\sim}{e}$ is the random environmental effect,

with $\text{var}(\underset{\sim}{z}) = V$, $\text{var}(\underset{\sim}{g}) = V_g$, $\text{var}(\underset{\sim}{c}) = V_c$,

$\text{var}(\underset{\sim}{e}) = V_e$, $\text{cov}(\underset{\sim}{g},\underset{\sim}{c}) = V_{gc}$, and

$$\mathrm{cov}(\underset{\sim}{g},\underset{\sim}{e}) = \mathrm{cov}(\underset{\sim}{c},\underset{\sim}{e}) = \underset{\sim}{0}.$$

Thus, the total phenotypic covariance matrix is given by

$$V = V_g + V_c + 2V_{gc} + V_e. \qquad (1.22)$$

The terms in (1.22) are the commonly estimated components of covariance, when interest centres mainly on resolving the phenotypic covariance into genetic and environmental components. The more interesting case occurs when the investigation includes the actual genetic mechanism. Then $\underset{\sim}{g}$, and consequently V_g, must be partitioned into genetically meaningful components. The "additive" and "dominance" components we defined in 1.3 are the simplest and most commonly used genetic components. Elston and Rao (1978) listed the following assumptions for the application of this model (1.21) to family data, although they considered only the univariate case:

(i) a linear model exists for the quantitative trait which assumes no genotype-environment interactions;

(ii) genotype-environment covariance matrix is in equilibrium;

(iii) if twins occur in the data, their phenotypic similarity due to common prenatal and postnatal environment, irrespective of zygosity, is the same as for ordinary siblings;

(iv) adoptions, if they occur in the data, are random with regard to genetic or environmental variables, and true parents are assumed to exert no influence on the children either prior to or after the adoption.

1.5 Estimation of the Genetic Components of Covariance

Theoretically, the statistical problem of estimating the genetic components of covariance falls into the general category of covariance components estimation for random or mixed designs with unbalanced data. Searle (1971) presents a comprehensive review of the huge literature on the statistical estimation of components of variance, with obvious extensions to components of covariance. Here we shall give only a brief summary of the general methods. In the next section we shall review the methods for the particular case of pedigree data--the type of data structure to be studied in this dissertation.

The estimation procedures may be classified into three basic categories: Analysis of Variance (ANOVA) estimators, Symmetric Sums of Products (SSP) estimators and Maximum Likelihood (ML) estimators. When the design is balanced these methods lead essentially to the same estimator. In human genetics the ANOVA estimators were until recently the best known. Henderson (1953) presented the following methods for obtaining ANOVA estimators of components of variance or covariance:

Method 1: Equate simple sums of products to their expectations under the assumption of a random effects model.

Method 2: For mixed models, adjust the data for the fixed effects by the method of least squares, then apply Method 1 to the adjusted data.

Method 3: Use some conventional method to compute mean products for the non-orthogonal data, and then equate these mean squares to their expectations and solve for the estimates.

We often have more equations than parameters, in which case weighted least squares methods are used to solve the equations. Henderson indicated that Method 1 leads to biased estimates if the assumptions of a mixed model are appropriate or if certain elements of the model are correlated. Method 2 adjusts for the mixed model, but the estimates are still biased if there is interaction between fixed and random effects. Method 3 gives unbiased estimates,but may require a great deal of computation. Rohde and Tallis (1969) have derived the exact first- and second-order moments of estimates of components of covariance, under normality assumptions.

The SSP approach was developed principally by Koch (1967,1968) for random effects models, and extended to mixed models by Forthofer and Koch (1974). The method uses the fact that expected values of products of observations are linear functions of the variance components. Sums of these products, and hence means of them, therefore provide unbiased estimators of the components.

In connection with the ANOVA and SSP estimators Searle (1971) notes the following problems.

(i) There are infinitely many quadratic forms that can be used, but the procedures give no criteria for selecting the quadratic forms to be used.

(ii) The only known property of the estimators is that they are unbiased for random models and, with the SSP and Henderson Method 3, unbiased for mixed models as well. Searle doubts the usefulness of the unbiasedness property for unbalanced designs.

(iii) Variances of estimators are somewhat tractable only if normality is assumed. Even then, the variances are functions of the unknown components.

(iv) The ANOVA methods can yield negative components of variance.

(v) These methods only estimate components of covariance. They give no guidance on the problem of estimating the fixed effects of the model, which are usually also unknown.

Maximum Likelihood methods overcome most of these problems. However ML equations for estimating components of covariance from unbalanced data cannot be solved explicitly. Even if solutions could be found, the problem of using these to derive non-negative estimates of variances must be considered. Thompson (1962) suggests a restricted maximum likelihood procedure, confined to just that portion of the set of sufficient statistics which is location invariant. But Searle in his review reaches the conclusion that "Explicit maximum likelihood estimators must be despaired of." Hartley and Rao (1967) have developed a general set of equations from which specific estimates may be obtained by iteration. They showed further that these estimators were consistent and asymptotically efficient. However, to our knowledge their method has not been specifically extended to include components of covariance, so we shall not describe them here. Maximum likelihood methods have been developed specifically for pedigree data. We shall describe these in the next section.

1.6 Maximum Likelihood Methods for Estimating the Genetic Components of Covariance from Pedigrees

For detailed analyses of the mode of inheritance of the trait, we must use family data or, preferably, their extended form: pedigree data. As Elston and Rao (1978) pointed out, genetic heterogeneity from family to family may obscure the mode of inheritance, but a large single pedigree is more likely to be homogeneous. We shall generally consider pedigrees.

The term "pedigree" has been given a technical definition in graph-theoretic terms by Lange and Elston (1975). In ordinary language it refers simply to a family tree including spouses. Lange and Elston (1975) define a pedigree as "simple" if there are no consanguineous marriages, and, except for the original parents, the members of each mating pair in the pedigree is as follows: one is related to someone in the previous generation, the other is an unrelated person "marrying into the pedigree." Otherwise it is said to be a "complex" pedigree.

Two methods have been developed to handle simultaneously the lack of balance and the biological dependencies in pedigree data. We shall describe these in turn.

The method of Lange et al. (1976) extended to multivariate traits by Lange et al. (1981) assumes the phenotypes of an arbitrary pedigree follow a multivariate normal distribution, whose covariance matrix is expressed as a function of additive genetic covariance, a dominance covariance and an environmental covariance. Estimates are then derived by the scoring algorithm (Rao, 1973).

For a pedigree of n members, arrange the p-variate data as

$$\underset{\sim}{w} = (x_{11}, \ldots, x_{1n}, \ldots, x_{p1}, \ldots, x_{pn})' .$$

Then using the decomposition of genetic covariance given by formula (1.16), the phenotypic covariance matrix of $\underset{\sim}{w}$ may be partitioned as

$$\mathrm{var}(\underset{\sim}{w}) = \sum_{k} \sigma_k \Omega_k = \Omega \qquad (1.23)$$

where σ_k are scalars and Ω_k are np x np symmetric matrices. The log likelihood function is then, apart from a constant

$$\ell = -\tfrac{1}{2}\, \ell n|\Omega| \; -\tfrac{1}{2}(\underset{\sim}{w} - A\underset{\sim}{\mu})'\, \Omega^{-1}\, (\underset{\sim}{w} - A\underset{\sim}{\mu}) \qquad (1.24)$$

where $A\underset{\sim}{\mu} = \mathcal{E}(\underset{\sim}{w})$. The score vector is

$$\underset{\sim}{s} = \begin{pmatrix} \dfrac{\partial \ell}{\partial \underset{\sim}{\mu}} \\ \\ \dfrac{\partial \ell}{\partial \underset{\sim}{\sigma}} \end{pmatrix} = \begin{pmatrix} \underset{\sim}{s}_{\underset{\sim}{\mu}} \\ \\ \underset{\sim}{s}_{\underset{\sim}{\sigma}} \end{pmatrix} \qquad (1.25)$$

where $\dfrac{\partial \ell}{\partial \mu_k} = \dfrac{\partial \underset{\sim}{\mu}'}{\partial \mu_k} \cdot A'\Omega^{-1}(\underset{\sim}{w} - A\underset{\sim}{\mu})$

and

$$\frac{\partial \ell}{\partial \sigma_k} = -\tfrac{1}{2}\mathrm{tr}(\Omega^{-1}\Omega_k) + \tfrac{1}{2}(\underset{\sim}{w} - A\underset{\sim}{\mu})'\Omega^{-1}\Omega_k\Omega^{-1}(\underset{\sim}{w} - A\underset{\sim}{\mu}).$$

The information matrix is

$$\underset{\sim}{I} = \begin{pmatrix} (-\mathcal{E}\, \dfrac{\partial^2 L}{\partial \mu_k \partial \mu_\ell}) & \underset{\sim}{0} \\ \\ \underset{\sim}{0} & (-\mathcal{E}\, \dfrac{\partial^2 \ell}{\partial \sigma_k \partial \sigma_\ell}) \end{pmatrix} = \begin{pmatrix} I_{\underset{\sim}{\mu}} & \underset{\sim}{0} \\ \\ \underset{\sim}{0} & I_{\underset{\sim}{\sigma}} \end{pmatrix} \qquad (1.26)$$

where

$$I_{\mu} = (-\mathcal{E}\, \frac{\partial^2 \ell}{\partial \mu_k \partial \mu_\ell}) = A'\Omega^{-1}A$$

$$(1.27)$$

$$\underset{\sim}{I}_{\sigma} = (-\mathcal{E}\, \frac{\partial^2 \ell}{\partial \sigma_k \partial \sigma_\ell}) = (\tfrac{1}{2}\mathrm{tr}(\Omega^{-1}\Omega_k\Omega^{-1}\Omega_\ell))$$

and

$$-\mathcal{E}\, \frac{\partial^2 \ell}{\partial \mu_k \partial \sigma_k} = 0.$$

If there are m independent pedigrees and ℓ_r is the log likelihood for r-th pedigree, then the scoring algorithm updates $\underset{\sim}{\mu}$ and $\underset{\sim}{\sigma}$ by adding the increments

$$\Delta\underset{\sim}{\mu} = \left(\sum_{r=1}^{m} I_{\underset{\sim}{\mu} r} \right)^{-1} \sum_{r=1}^{m} s_{\underset{\sim}{\mu} r}$$
$$\Delta\underset{\sim}{\sigma} = \left(\sum_{r=1}^{m} I_{\underset{\sim}{\sigma} r} \right)^{-1} \sum_{r=1}^{m} s_{\underset{\sim}{\sigma} r} \ . \tag{1.28}$$

Lange et al. (1981) noted the major disadvantage of the method: the matrix inversion Ω^{-1} and matrix multiplication $\Omega^{-1}\Omega_k$ must be done for each pedigree at each iteration. Their method implicitly assumes a polygenic model, and is therefore not general.

The second method is due is Elston and Stewart (1971) whose algorithm may be used to compute the unconditional likelihood of a simple pedigree under any genetic model. Before we describe their algorithm let us explain their notation. The measurement on an individual who is related to a member of the pedigree in a previous generation is denoted by x; the measurement on his spouse is denoted y. For the original parents, the measurement on one is arbitrarily labeled x, and the other y. The measurement is then subscripted hierarchically to reflect the individual's pedigree (we may have more than one pedigree), generation within his pedigree and birth order within his generation. Thus the original parents (generation zero) have one subscript i_0 to denote pedigree number. Their children (generation 1) have two subscripts $i_0 i_1$ - i_0 is the same as their parents' and i_1 indexes their birth order. Grandchildren (generation 2) have three subscripts $i_0 i_1 i_2$ - $i_0 i_1$ are the same as their parents' and i_2 their birth order within family, and so on (see Figure 1.1). In general the measurement on an individual in the j-th generation of the pedigree, counting the original parents as generation 0, will be of the form $x_{i_0 i_1 i_2 \ldots i_j}$: this being for the i_j-th child of the i_{j-1}-th child .... of the i_j-th child of the i_0-th original parents. The measurement on his spouse is $y_{i_0 i_1 i_2} \ldots i_j$.

We shall first construct the likelihood for a random pedigree for an oligogenic model.

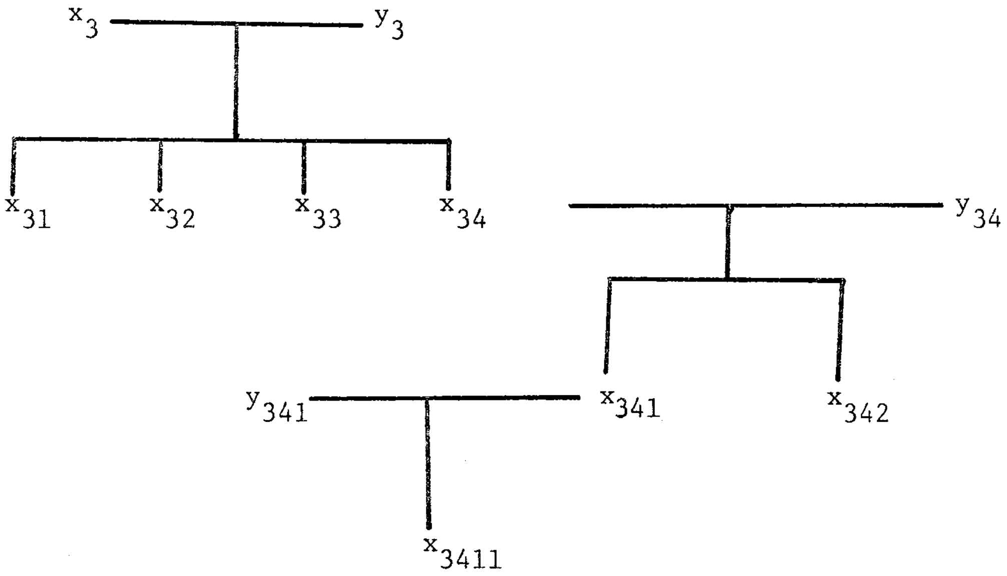


Fig. 1.1 Illustration of the notation for the phenotypes of the members of a pedigree: this is the third pedigree ($i_0 = 3$) in a set of pedigrees.

Let the number of possible genotypes be k, assumed arranged in order and indexed u = 1, 2, ..., k. Let $g_u(x)$ denote the conditional probability distribution of the phenotype x given the u-th genotype, and p_{stu} the probability that an individual has genotype u given his parents have genotypes s and t (s,t, = 1, 2, ..., k). Then for a sibship of size r the likelihood is

$$L = \prod_{i=1}^{r} \sum_{u=1}^{k} p_{stu} g_u(x_i) \qquad (1.29)$$

Let ψ_v be the probability that a person selected randomly from the population has genotype v, then the likelihood of observing a particular phenotype on an original parent or someone who marries into the pedigree is

$$L = \sum_{v=1}^{k} \psi_v g_v(y). \qquad (1.30)$$

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

$$L = \prod_{i=1}^{r} \sum_{u=1}^{k} p_{stu} g_u(x_i) \sum_{v=1}^{k} \psi_v g_v(y_i) \ . \qquad (1.31)$$

This is a function of the parental genotypes s and t; but s and t of this generation are the u and v of the previous generation. Thus the likelihood of the j-th generation (including spouses), given the parents' parents' genotypes s_{j-1} and t_{j-1} can be written as

$$\begin{cases} \Gamma_0 = \sum_{s_0=1}^{k} \psi_{s_0} g_{s_0}(x_{i_0}) \sum_{t_0=1}^{k} \psi_{t_0} g_{t_0}(y_{i_0}) \\ \Gamma_j = \prod_{i_j} \sum_{s_j=1}^{k} p_{s_{j-1} t_{j-1} s_j} g_{s_j}(s_{i_0 i_1} \ldots i_j) \sum_{t_j=1}^{k} \psi_{t_j} g_{t_j}(y_{i_0 i_1 \ldots i_j}); j \geq 1 \end{cases}$$

(1.32)

By starting at the most recent generation and successively moving up the pedigree, we can write the unconditional likelihood of observing the entire pedigree as

$$L = \Gamma_0(\Gamma_1(\Gamma_2(\Gamma_3 \ldots))) \qquad (1.33)$$

in which Γ is regarded as an operator, not a mathematical function.

Under different models only p_{stu} (the transition probabilities) and $g_u(x)$ (the conditional phenotypic probabilities given the genotype) will vary.

In the polygenic model the p_{stu} are normal densities, and the summations are replaced by integrals. The rationale they gave was this: the number of loci in a genotype that is heterozygous tends to a constant value; hence the variance of the population of gametes transmitted by any genotype also tends to a constant value, say σ^2, equal to half the additive genetic variance, i.e., $\sigma^2 = \frac{1}{2}\sigma_a^2$. Under random mating uniting gametes are uncorrelated and so the variance of genotypes within a sibship is always the same. If the parents' genotypes are s and t, the gametic distributions are $N(\frac{s}{2},\sigma^2)$ and $N(\frac{t}{2},\sigma^2)$. The genotype distribution within the sibship is thus $N(\frac{s+t}{2},\sigma^2)$. Setting $m = \frac{1}{2}(s+t)$ the analogue of the transmission probability is

$$\phi(u-m,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(\frac{u-m}{\sigma})^2} \qquad (1.34)$$

i.e., the ordinate at 0 of the distribution $N(u-m,\sigma^2)$, $g_u(x)$ is replaced by the conditional density of x given u, $g(x|u) \sim N(u+\mu,\sigma_e^2)$ where μ is the mean over all genotypes. The genotypic variance is equal to the additive genetic variance under this model and is therefore $\sigma_a^2 = 2\sigma^2$. Thus ψ_v, the genotype distribution in the population is replaced by $\phi(v,\sigma_a^2)$. With these changes the conditional likelihood for the j-th generation is

$$\begin{cases} \Gamma_0 = \int_{s_0} \phi(s_0,\sigma_a^2)\ \phi(s_0+\mu-x_{i_0},\sigma_e^2) \int_{t_0} \phi(t_0,\sigma_a^2)\ \phi(t_0+\mu-y_{i_0},\sigma_e^2) \\ \Gamma_j = \prod_{i_j} \int_{s_j} \phi(s_j-\frac{1}{2}(s_{j-1}+t_{j-1}),\frac{1}{2}\sigma_a^2)\ \phi(s_j+\mu-x_{i_0 i_1 \ldots i_j},\sigma_e^2) \\ \qquad \int_{t_j} \phi(t_j,\sigma_a^2)\ \phi(t_j+\mu-y_{i_0 i_1 \ldots i_j},\sigma_e^2); \quad j \geq 1 \end{cases} \qquad (1.35)$$

where $\int_s$ denotes integration with respect to $s \varepsilon (-\infty,\infty)$.

For the mixed model, in which the genotype is made up of a few major loci and a polygenic loci, the likelihood is obtained simply by combining (1.32) and (1.35) as follows

$$\begin{cases} \Gamma_0 = \sum\limits_{s_0} \psi_{s_0} \int_{a_0} \phi(a_0,\sigma_a^2)\ \phi(a_0+\mu_{s_0}-x_{i_0},\sigma_e^2) \\ \qquad \sum\limits_{t_0} \psi_{t_0} \int_{b_0} \phi(b_0,\sigma_a^2)\ \phi(b_0+\mu_{t_0}-y_{i_0},\sigma_e^2) \\ \Gamma_j = \prod\limits_{i_j} \sum\limits_{s_j} p_{s_{j-1}t_{j-1}s_j} \int_{a_j} \phi(a_j-\tfrac{1}{2}(a_{j-1}+b_{j-1}),\tfrac{1}{2}\sigma^2)\phi(a_j+\mu_{s_j}-x_{i_0i_1\ldots i_j},\sigma_e^2) \\ \qquad \sum\limits_{t_j} \psi_{t_j} \int_{b_j} \phi(b_j,\sigma_a^2)\ \phi(b_j+\mu_{t_j}-y_{i_0i_1\ldots i_j},\sigma_e^2); \quad j \geq 1. \end{cases} \tag{1.36}$$

The integrals that appear in (1.35) and (1.36) are easily expressed as functions of normal densities. Thus, the Elston and Stewart algorithm, indicated by Elston and Stewart (1971) for (1.35) and considered in detail later in this dissertation for (1.36), is a highly efficient means of computing the likelihood of a pedigree.

For the purpose of estimating variance components it is used in conjunction with algorithms that maximize functions without taking derivatives, i.e., by searching the likelihood surface. ML estimators have the advantage that their asymptotic variances are easily derived, even though the estimators themselves cannot be written down explicitly. In fact they are found simply by inverting the information matrix given in (1.26).

For oligogenic models Lange and Elston (1975) have extended the algorithm to complex pedigrees. Cannings et al. (1976,1978) used a graph theoretic approach to generalize this procedure so that the complete likelihood can be calculated from any arbitrary subset of the original pedigree. Their methods allow for going down from the original parents when calculating the likelihood.

Elston (1973) extended the model to allow for variable age of onset.

For the special case of nuclear families Morton and MacLean (1974) developed conditional likelihoods. Elston and Rao (1978) have noted that the conditional likelihoods may be obtained by dividing the unconditional likelihoods by the corresponding likelihoods of the two parents. Go et al. (1978) have demonstrated in a simulation study that the unconditional likelihood is appreciably more efficient for parameter estimation than the conditional. However, for detection of major genes, their relative efficiencies depended on the precise hypotheses tested (MacLean et al.,1975, Go et al.,1978). On the other hand Elston and Rao (1978) have remarked that for pedigrees, there is nothing comparable to the likelihood conditional on the parental phenotypes since a large pedigree can contain individuals who are at the same time parents and offspring of other individuals in the pedigree.

Further developments have included the calculation of the likelihood for the case of linked traits (Ott,1974) and the approximation of the likelihood under a mixed model by a mixture of normal distributions (Graepel, 1981).

The Elston and Stewart algorithm has not been extended to multivariate traits and so cannot be used to compute genetic components of covariance, and hence to test associated hypotheses. This is a major goal of the present study.

1.7 Testing the Simplest Genetic Hypothesis of Covariation

We now return to the question of the genetic basis for the covariation of two or more traits: is it due to the same gene(s) or to different genes for each trait? Two main approaches have been suggested in the literature for answering this question. One method, Eaves and Gale (1973), rescales the additive effects to a common value employing a minimum chi-square criterion to estimate the scales. They do not say, but since they assume a purely polygenic model, the hypothesis they are really testing is common polygenic control, and that may not be very helpful.

The second method, Elston et al. (1975), essentially reduces the problem to a univariate one by determining a linear function of the traits that best fits the phenotypic distribution assumed to be a mixture of two or three normal distributions. Segregation and linkage analyses are then performed on the linear function. A variation of the method, Goldin et al. (1980), determines the linear function that best fits the phenotypic distribution under an assumed genetic model. The method, together with the slight variation noted, has been applied with some success to study the segregation and linkage to known markers of the hyperlipoproteinemias (Elston et al., 1975), genetic covariation of cholesterol and triglycerides (Namboodiri et al., 1975), hypertriglyceridemias (Namboodiri et al., 1977), and von Willebrand's disease (Goldin et al., 1980). However the procedure does not provide a test of the hypothesis of genetic covariation, and it is not known whether or not the linear index so determined is optimal in terms of having the maximum realizable heritability, or major gene component of genetic variance. If the index is not optimum in one or both senses, then it is not the best measure of the "innate" trait underlying the multivariate phenotype, although it could provide valuable insight into the phenomenon under study if the index so determined is interpretable. Besides, the genetic components of covariance are also important as descriptive statistics in the analysis of the genetic basis of covariation, and hence estimates should be obtained if at all possible.

1.8 Objectives and Outline of Remaining Chapters

Our aim in this dissertation is to develop maximum likelihood methods for analyzing the genetic basis of covariation. To this end we shall develop a model for multivariate traits that has all the features of the Elston and Stewart model, and includes the essentials of Lange's model. This will allow us to study polygenic inheritance and combined major gene and polygenic inheritance in pedigrees with arbitrary structure; this cannot be done accurately even for univariate traits at present.

The components of the model are specified in Chapter II. In Chapter III we discuss the calculation of the likelihood of a simple pedigree under oligogenic and polygenic models of inheritance. Formulas for the mixed model are presented in Chapter IV. Since the feasibility of computations is by far the greatest drawback to the use of the complete Elston-Stewart model, even for univariate traits, we study the problem analytically in depth yielding recurrence formulas that overcome the computational problems.

The formulas derived in Chapters III and IV are extended to complex pedigrees in Chapter V. Estimation and hypothesis testing are considered in Chapter VI. In Chapter VII we discuss our conclusions and mention problems for further research.

CHAPTER II

COMPONENTS OF THE MATHEMATICAL MODEL

2.1 Introduction

Genetic analysis has as its basic aim the determination of the mode of inheritance of a particular trait, especially with a view to establishing single-gene effects. Statistical models that provide the framework for genetic analysis abound and have had a long history. However they have been developed mainly for univariate traits as outlined in the previous chapter. We define in this chapter a setting in which genetically meaningful statistical analyses may be done on multivariate data gathered on human families and pedigrees. We do this by developing a model that has all the features of the general model presented by Elston and Stewart (1971) but includes the essentials of the model presented by Lange et al., (1981) for polygenic traits.

In general the mathematical model has four major components. The first component specifies the joint distribution of mating individuals in the population, the second describes the relationship between phenotype and genotype, the third describes how genetic variability is passed on from one generation to the next, while the last component deals with sampling from the population. Elston and Stewart (1971) and Elston and Yelverton (1975) have described these components in detail for univariate traits.

2.2 Joint Genotypic Distribution of Mating Individuals

We shall assume mating is random with respect to the loci under consideration. Then the joint genotypic distribution of mating types is simply the product of the population genotypic frequencies or densities of the genotypic value. Following Elston (1980) we shall write ψ_t for the population frequency or density of the value of genotype t. The distribution of the mating type s x t is therefore $\psi_s \cdot \psi_t$.

To generalize the specification to include inbreeding we can write the equilibrium distribution for say the single autosomal locus, two allele case, as

Genotype	AA	Aa	aa
Frequency	$\psi_{AA} = q^2+fq(1-q)$	$\psi_{Aa} = 2q(1-q)(1-f)$	$\psi_{aa} = (1-q)^2+fq(1-q)$

where f is the inbreeding coefficient--the correlation between uniting gametes. Since f is related to the parental genetic correlation, the specification above can also be used for assortative mating if mates are selected according to genotype. When f is used as a measure of inbreeding, it may be defined as the kinship coefficient of his parents, which can be calculated from the pedigree structure.

If we consider a polygenotype G, the population density of genotypic value of an individual i is given by Lange's result (our Theorem 1.1) as $\psi_G = \phi(\underset{\sim}{g},2\Phi_{ii}V_g)$, where we have assumed no dominance so that V_g is equal to the additive covariance matrix. The density of the mating type G x H (of individuals i and j respectively) is then given by

$$\psi_G \cdot \psi_H = \phi(\underset{\sim}{g},2\Phi_{ii}V_g)\cdot\phi(\underset{\sim}{h},2\Phi_{jj}V_g).$$

For a mixed genotype, containing oligogenic and polygenic components, the distribution of the mating type sG x tH is

$$\psi_{sG}\psi_{tH} = \psi_s\phi(\underset{\sim}{g},2\Phi_{ii}V_g) \cdot \psi_t\phi(\underset{\sim}{h},2\Phi_{jj}V_g).$$

2.3 Genotype-Phenotype Relationship

Denote the distribution of vector phenotype $\underset{\sim}{z}$ given the oligogenotype u by $g_u(\underset{\sim}{z})$, given the polygenotype H by $g_h(\underset{\sim}{z})$, and given the mixed genotype uH by $g_{uh}(\underset{\sim}{z})$. These may generally be referred to as g-functions, or penetrances.

2.3.1 Qualitative phenotype

When all components of the vector trait are qualitative we need to specify penetrances for all combinations of the vector values. As an example, suppose in a family study of breast cancer there is some interest in whether left (z_1), right (z_2) or both breasts are affected (1), or not (0). Then for two alleles at an autosomal locus we need to specify the penetrances given in Table 2.1. Each genotype involves four g-functions. However only three of them need to be specified; the fourth is then automatically given by the condition that the row probabilities in the table must sum to unity. Dominance will reduce the table to only two rows.

The risk function approach to specifying the phenotypic distribution (Falconer (1965), Curnow (1972) and others) may also be easily generalized for multivariate traits. The simplest case occurs with a dichotomy where the condition is expressed if $\underset{\sim}{z} \geq \underset{\sim}{0}$ (meaning $z_i \geq \theta_i$, $i = 1,2,\ldots,p$), where $\underset{\sim}{\theta}$ is a p-vector of threshold constants. There is no problem letting some components of $\underset{\sim}{z}$ reverse their inequalities. This approach effectively makes the phenotypic

TABLE 2.1

PENETRANCE FOR AN AUTOSOMAL MONOGENIC TWO-ALLELE INHERITANCE WITH QUALITATIVE BIVARIATE PHENOTYPE

Genotype	Penetrances				Total
AA	$g_{AA}(0,0)$	$g_{AA}(1,0)$	$g_{AA}(0,1)$	$g_{AA}(1,1)$	1
Aa	$g_{Aa}(0,0)$	$g_{Aa}(1,0)$	$g_{Aa}(0,1)$	$g_{Aa}(1,1)$	1
aa	$g_{aa}(0,0)$	$g_{aa}(1,0)$	$g_{aa}(0,1)$	$g_{aa}(1,1)$	1

distribution discrete, and without loss of generality we may let $x = 0$ if $\underset{\sim}{z} < \underset{\sim}{\theta}$, and $x = 1$ if $\underset{\sim}{z} \geq \underset{\sim}{\theta}$. Then assuming the density of $\underset{\sim}{z}$ conditional on t is $\phi(\underset{\sim}{\mu}_t - \underset{\sim}{z}, V_g)$,

$$g_t(1) = \int_{\underset{\sim}{z} \geq \underset{\sim}{\theta}} \phi(\underset{\sim}{\mu}_t - \underset{\sim}{z}, V_g)$$

and

$$g_t(0) = 1 - g_t(1).$$

The extension to a polychotomy is obvious.

2.3.2 Quantitative phenotype

When each component of the phenotypic vector is quantitative it may be reasonable to assume $\underset{\sim}{z}$ is multinormal. We can then set $g_t(\underset{\sim}{z}) = \phi(\underset{\sim}{\mu}_t - \underset{\sim}{z}, V_e)$ for oligogenic models where $\underset{\sim}{\mu}_t$ is the vector mean of the oligogenotype t, V_e is the environmental covariance matrix assumed equal for all oligogenotypes, and $\phi(\underset{\sim}{\mu}_t - \underset{\sim}{z}, V_e)$ is the ordinate at $\underset{\sim}{\mu}_t$ of the multinormal distribution $N(\underset{\sim}{z}, V_e)$. It should be noted though that whereas a continuous univariate distribution can virtually always be transformed to a normal distribution, this is not true of a multivariate distribution.

The situation where some components of $\underset{\sim}{z}$ are qualitative and others are quantitative would seem difficult to deal with in general terms. Where it is reasonable to 'group' each quantitative component into a few categories the likelihood can be computed as if each component of $\underset{\sim}{z}$ was qualitative (section 2.3.1). In general however it would seem reasonable to specify the penetrances of the quantitative components conditional on particular combinations of the qualitative components.

2.4 The Mode of Inheritance

The third component of the model is the mode of inheritance, i.e., the genotypic distribution of the offspring conditional on the two parental genotypes. For oligogenic inheritance, the p_{stu}, the probability that an individual has genotype u given that his parents' genotypes are s and t, is exactly the same as for univariate traits.

Some modifications are however necessary for polygenic inheritance. We shall use Lange's (1978) result described in section 1.3 to derive the desired p_{stu} which is a probability density function in this case. This leads to a formulation that includes the Elston and Stewart (1971) formulation as a special case. This more general formulation allows for consanguineous matings in the pedigree.

Let F, G, H be the respective polygenotypes of two parents P_1 and P_2 and their child C. Denote the corresponding p-variate values of the polygenotypes by $\underset{\sim}{s}_{P_1}$, $\underset{\sim}{s}_{P_2}$ and $\underset{\sim}{s}_{c}$. Assume $E(\underset{\sim}{s}_{P_1}) = E(\underset{\sim}{s}_{P_2}) = E(\underset{\sim}{s}_{c}) = \underset{\sim}{0}$. Then we can state the following result.

Theorem 2.1

In a population that has been undergoing random mating but in which chance consanguineous matings occur, assuming no dominance, the distribution of offspring genotypic value $\underset{\sim}{s}_c$ conditional on parental genotypic values $\underset{\sim}{s}_{p_1}$ and $\underset{\sim}{s}_{p_2}$ is given by:

$$p_{FGH} = \phi(\underset{\sim}{s}_c - [\alpha_1 \underset{\sim}{s}_{p_1} + \alpha_2 \underset{\sim}{s}_{p_2}],\ \alpha_3 V_g)$$

where

$$\alpha_1 = \frac{\Phi_{p_1c}\Phi_{p_2p_2} - \Phi_{p_2c}\Phi_{p_1p_2}}{\Phi_{p_1p_1}\Phi_{p_2p_2} - \Phi^2_{p_1p_2}},$$

$$\alpha_2 = \frac{\Phi_{p_2c}\Phi_{p_1p_1} - \Phi_{p_1c}\Phi_{p_1p_2}}{\Phi_{p_1p_1}\Phi_{p_2p_2} - \Phi^2_{p_1p_2}},$$

$$\alpha_3 = 2\{\Phi_{cc} - \frac{\Phi_{p_1p_1}\Phi^2_{p_2c} + \Phi_{p_2p_2}\Phi^2_{p_1c} - 2\Phi_{p_1p_2}\Phi_{p_1c}\Phi_{p_2c}}{\Phi_{p_1p_1}\Phi_{p_2p_2} - \Phi^2_{p_1p_2}}\},$$

and

Φ_{AB} is the kinship coefficient between A and B.

Proof:

Let $x = \underset{\sim}{a}'\underset{\sim}{s}_{p_1}$, $y = \underset{\sim}{a}'\underset{\sim}{s}_{p_2}$, $z = \underset{\sim}{a}'\underset{\sim}{s}_c$ and $\sigma^2_A = \underset{\sim}{a}'V_g\underset{\sim}{a}$. Then we only need to show that for all $\underset{\sim}{a} \neq \underset{\sim}{0}$, the density of z given x and y is normal with mean $\alpha_1 x + \alpha_2 y$ and variance $\alpha_3 \underset{\sim}{a}'V_g\underset{\sim}{a}$. It follows from Lange's theorem presented in Theorem 1.1 that

$$(x,y,z)' \sim N_3(\underset{\sim}{0}, 2\Phi\sigma^2_A)$$

where

$$\Phi = \left(\begin{array}{cc|c} \Phi_{p_1p_1} & \Phi_{p_1p_2} & \Phi_{p_1c} \\ \Phi_{p_2p_1} & \Phi_{p_2p_2} & \Phi_{p_2c} \\ \hline \Phi_{cp_1} & \Phi_{cp_2} & \Phi_{cc} \end{array}\right)$$

is the matrix of kinship coefficients, partitioned in an obvious manner. The density of z given x and y is then clearly normal (see for example Anderson, 1958, p. 28) with mean

$$\mu = \sigma^2_A[\Phi_{p_1c} \quad \Phi_{p_2c}] \cdot \frac{1}{\sigma^2_A} \begin{pmatrix} \Phi_{p_1p_1} & \Phi_{p_1p_2} \\ \Phi_{p_2p_1} & \Phi_{p_2p_2} \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$= \alpha_1 x + \alpha_2 y,$$

and variance

$$\sigma^2 = 2\sigma^2_A \left\{ \Phi_{cc} - [\Phi_{p_1c} \quad \Phi_{p_2c}] \begin{pmatrix} \Phi_{p_1p_1} & \Phi_{p_1p_2} \\ \Phi_{p_2p_1} & \Phi_{p_2p_2} \end{pmatrix}^{-1} \begin{pmatrix} \Phi_{p_1c} \\ \Phi_{p_2c} \end{pmatrix} \right\}$$

$$= \alpha_3 \sigma^2_A ,$$

after simplification. That concludes the proof.

<u>Corollary 2.1</u>

If all parents are unrelated, then under random mating the distribution of offspring genotypic value $\underset{\sim}{s}_c$ conditional on parental genotypic values $\underset{\sim}{s}_{p_1}$ and $\underset{\sim}{s}_{p_2}$ is given by

$$p_{FGH} = \phi(\underset{\sim}{s}_c - \tfrac{1}{2}[\underset{\sim}{s}_{p_1} + \underset{\sim}{s}_{p_2}], \tfrac{1}{2}V_g)$$

Proof:

If all parents are unrelated

$$\Phi_{p_1p_1} = \Phi_{p_2p_2} = \Phi_{cc} = \tfrac{1}{2},$$

$$\Phi_{p_1p_2} = 0, \; \Phi_{p_1c} = \Phi_{p_2c} = \tfrac{1}{4},$$

giving

$$\alpha_1 = \alpha_2 = \alpha_3 = \tfrac{1}{2}.$$

We shall call the quantities α_1,α_2 and α_3 "relatedness coefficients" They describe how the relationship between mates affects the offspring genotypic value. When all parents are unrelated the three coefficients are all equal to ½ (corollary 2.1) which is what Elston and Stewart used for univariate traits. By using the relatedness coefficients in the definition of p_{FGH} we therefore achieve an integration of the Lange, Boehnke and Spence and the Elston and Stewart models. In Chapter V the relatedness coefficients are used to define "generalized status indices" which summarize in numerical codes, the relevant information on every individual in the pedigree. This device will be shown to be the key to the computational problem in maximum likelihood analysis of pedigree data.

The relevant conditional "probabilities" for mixed inheritance where each genotype contains both oligogenic and polygenic components is simply the product of the two corresponding conditional "probabilities," i.e.,

$$p_{sFtGuH} = p_{stu} \cdot p_{FGH} .$$

The assumption here is that major genes and polygenes are transmitted independently.

2.5 Sampling Considerations

The three components described above enable us to write down the likelihood of a pedigree provided the pedigree can be regarded as randomly selected from the population under study. The likelihood function clearly has the same form as for the univariate case (section 1.6). However, in human genetics many traits studied are rare diseases for which random sampling is inefficient; most pedigrees in such a situation will contain only unaffected individuals, yielding no information at all about the kind of genetic segregation that underlies the disease. To avoid this, pedigrees are selected for study or ascertained, via probands--affected individuals who bring their pedigrees to the notice of the investigator. Thus every pedigree in the sample contains at least one proband, implying that only a subset of the population is included in the sampling frame, and so the likelihood needs modification.

For a pedigree of size n and phenotypes $\underset{\sim}{z}_1, \underset{\sim}{z}_2, \ldots, \underset{\sim}{z}_n$, let $L(\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n)$ be the likelihood assuming random sampling, and $\pi(\underset{\sim}{z}_i)$ the probability that an affected individual brings his pedigree into the (possibly conceptual) sampling frame from which a random sample of pedigrees is drawn for study. Then assuming that individuals are independently ascertained the likelihood conditional on the sampling procedure is

$$L_s = \frac{P(\text{at least one proband} \mid \underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n) \cdot L(\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n)}{P(\text{at least one proband})}$$

$$= \frac{1 - \prod_i [1 - \pi(\underset{\sim}{z}_i)]}{1 - \sum_{\underset{\sim}{z}_1} \cdots \sum_{\underset{\sim}{z}_n} L(\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n)\{\prod_i [1-\pi(\underset{\sim}{z}_i)]\}} \cdot L(\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n)$$

$$= (\text{ascertainment correction}) \times L(\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n);$$

summations change to integrals for continuous $\underset{\sim}{z}$'s, and $\pi(\underset{\sim}{z})$ is generally unknown.

Once $\pi(\underset{\sim}{z}_i)$ is specified we can calculate the likelihood of a pedigree with the necessary correction for ascertainment. The formulation above is the multivariate generalization of the Elston and Yelverton (1975) approach. They discussed various choices for $\pi(z)$ and the computational problems that arise. Elston and Sobel (1979) generalized the formulation to allow for the fact that some relatives may be outside the proband sampling frame. In this dissertation we shall concentrate on problems relating to the specification and computation of $L(\underset{\sim}{z}_1,\ldots,\underset{\sim}{z}_n)$, the unconditional likelihood, leaving aside ascertainment corrections for multivariate traits (see Dawson, 1981). Elston (1980) has pointed out that the family or pedigree is effectively random if the probands can be regarded as random and the type and number of relatives included are independent of a proband's phenotype. When we assume random sampling as indicated, there are two possible sources of bias. The existence of certain types of relatives in the population may depend on the random proband's phenotype. Then also the willingness of individuals to be sampled, whether probands or relatives, may depend on their phenotypes (Elston (1980)). These potential sources of bias should therefore be borne in mind when interpreting the results of the kind of analyses proposed in this dissertation.

CHAPTER III

OLIGOGENIC AND POLYGENIC INHERITANCE IN SIMPLE PEDIGREES

3.1 Introduction

We have discussed in Chapter II the mathematical specification of the various components of the likelihood function for a multivariate trait. We now turn to the actual calculation of the likelihood of a pedigree when the components of the mathematical model have been specified. In this chapter we shall confine our attention to oligogenic and polygenic modes of inheritance. Oligogenic inheritance involves a few and separately distinguishable genetic loci, whereas polygenic inheritance involves an indefinitely large number of separately indistinguishable genetic loci acting cumulatively on the trait of interest. The mode of inheritance which involves both oligogenic and polygenic inheritance will be considered in the next chapter.

In our development we first present the results in terms of the operators introduced by Elston and Stewart (1971). We then dispense with them as Ott (1974) did, and think in terms of the total configuration of the pedigree. This leads, in the case of polygenic inheritance with multinormal phenotypes, to a simple recurrence formula which is the multivariate generalization of an algorithm developed and programmed in 1972 by Dr. Philip Green, III (unpublished).

To facilitate the description of the likelihood function we start with our own definition and classification of pedigrees.

3.2 Definition and Classification of Pedigrees

Graph theoretic definitions and classification of pedigrees were first given by Lange and Elston (1975). A different classification system, also formulated in graph theoretic terms has been provided by Cannings, Thompson and Skolnick (1976,1978). The graph theoretic approach confers sophistication and some semblance of mathematical rigor to the developments presented by the authors, but it throws the subject into obscurity for the intended users, and, in my view, has not as yet led to really new insights which are otherwise difficult to see. We shall present simple definitions which are virtually equivalent to those of Lange and Elston (1975).

Definition 3.1

A pedigree is a pictorial representation of a set of related individuals in which lines of descent are indicated; either both parents or no parent must be indicated for each member of the pedigree.

Definition 3.2

A spouse-pair is the set of two individuals who have had a biological union resulting in offspring.

Definition 3.3

Two spouse-pairs are connected if they have one member in common, i.e., there are only three distinct individuals involved. If a spouse-pair is not connected to any other spouse-pair it is isolated.

Definition 3.4

A spouse-connected set is the set of all individuals in spouse-pairs that can be arranged so that adjacent spouse-pairs are connected.

Let us note that however complex the pedigree structure, there are only two types of individuals in a pedigree: those whose parental lines are indicated and those whose parental lines are not indicated. We shall use this fact to classify pedigrees.

Definition 3.5

A pedigree is simple if the parental lines are indicated for only one or no member of every isolated spouse-pair and spouse-connected set.

If for some spouse-pairs and spouse-connected sets parental lines are indicated for more than one member the pedigree is complex.

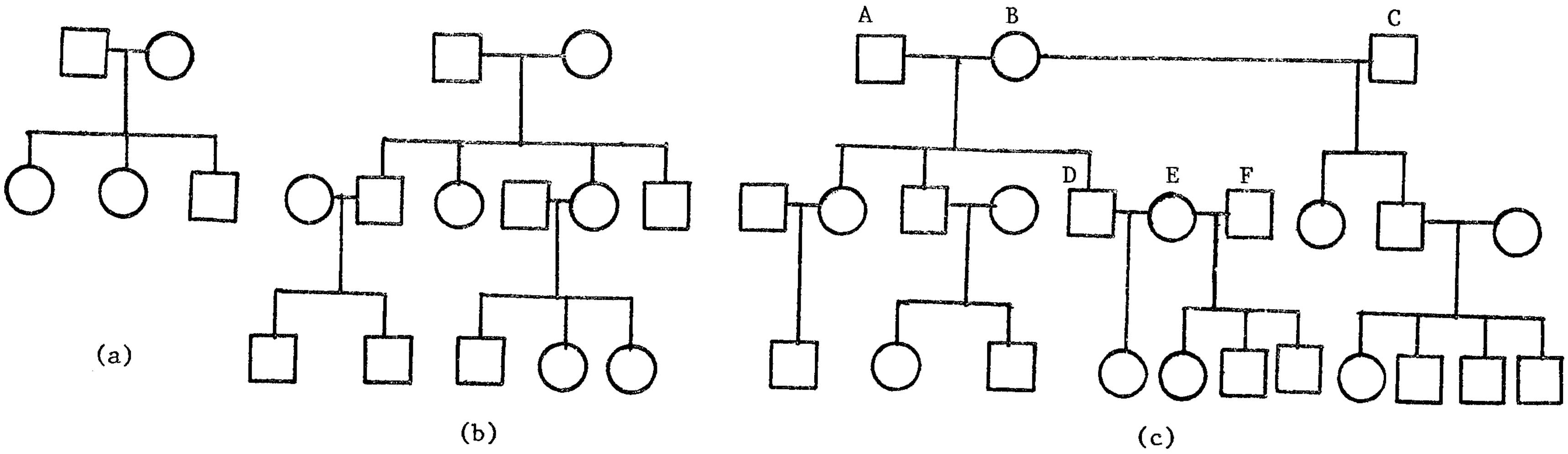
Under this definition a nuclear family with at least one child (Fig. 3.1a) is the simplest pedigree.

We can distinguish between two types of complex pedigrees: those in which the lines of descent form loops as in Fig. 3.3 and those in which they do not (Fig. 3.2). Loops may be due to consanguineous matings (Fig. 3.3a), or not (Fig. 3.3b).

Only simple pedigrees will be considered in this chapter. Complex pedigrees will be dealt with in Chapter V.

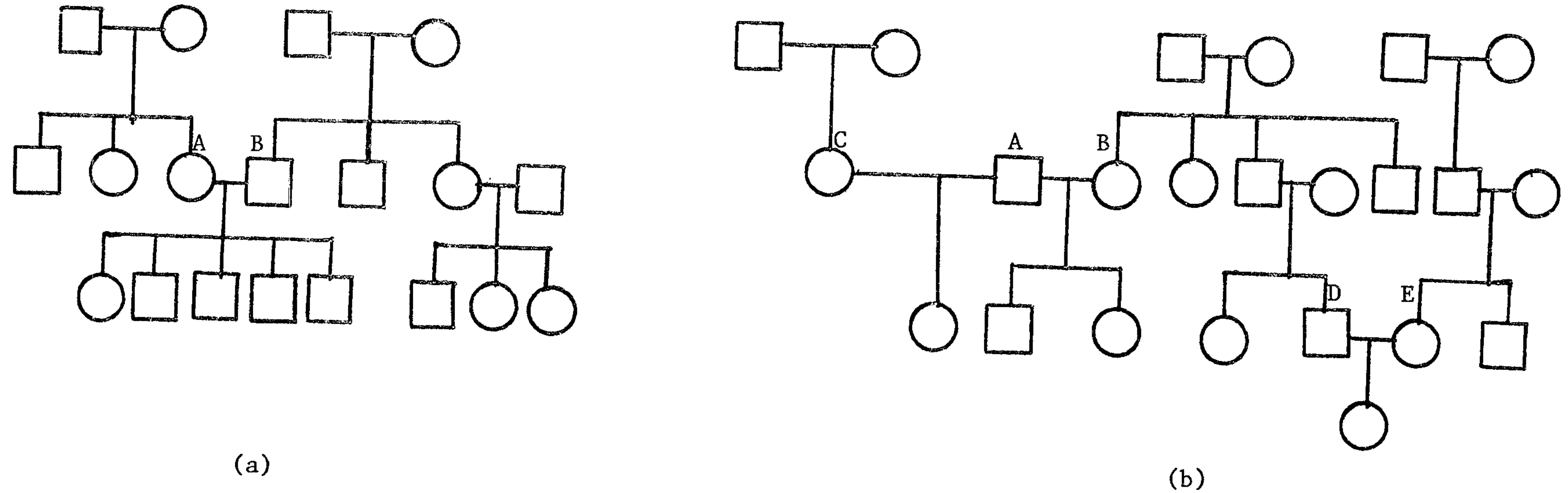
3.3 Oligogenic Inheritance in Simple Pedigrees

Consider a p-variate trait $\underset{\sim}{z}$. The likelihood of a pedigree having the observed values of $\underset{\sim}{z}$ depends on the genetic mechanism. We start with oligogenic inheritance, the case with only a few major loci. Let k be the number of genotypes underlying the observable or phenotypic variation in $\underset{\sim}{z}$, arranged in some specified order so that it is meaningful to talk of the u-th genotype, $u = 1,2,\ldots,k$, and $g_u(\underset{\sim}{z})$ be the conditional probability, given the u-th genotype, of observing $\underset{\sim}{z}$. Similarly we may let p_{stu} be the probability that an individual has genotype u, given that his parents' genotypes are s and t ($s,t,u = 1,2,\ldots,k$). In this setting the likelihood of observing a



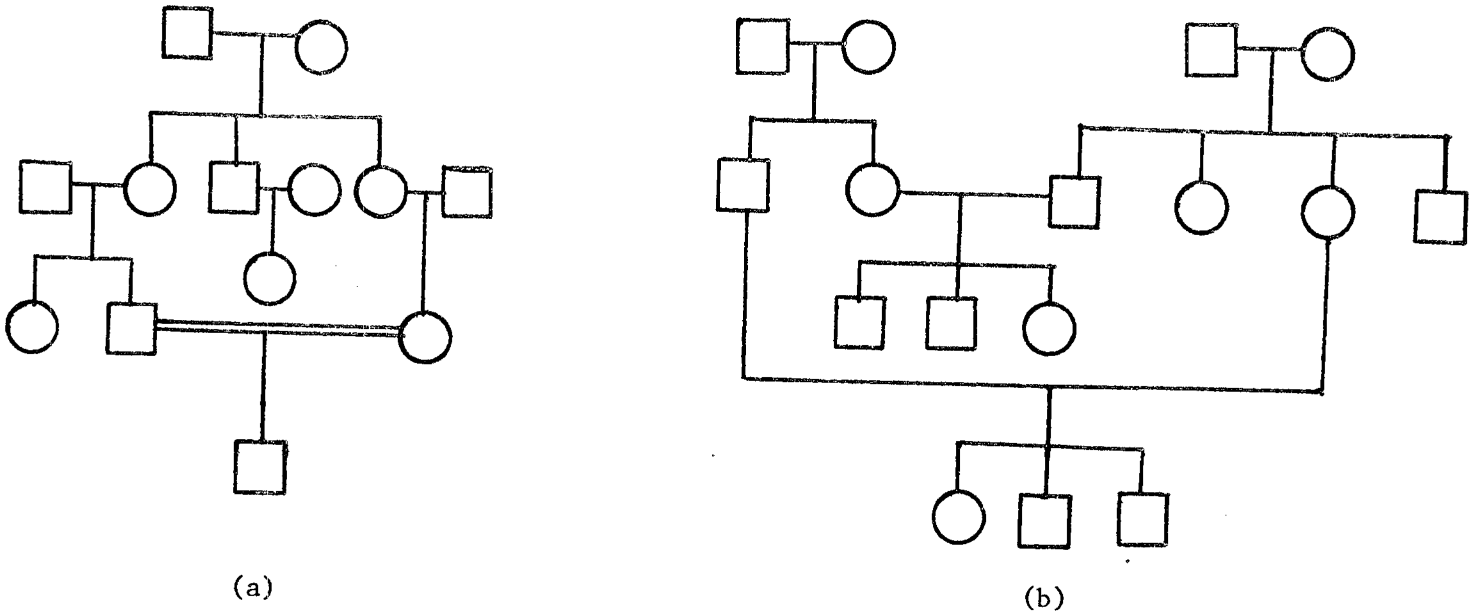
(a) Nuclear family;
(b) Simple extended family with no half-sibs;
(c) Extended families including half-sibs. Note that (A,B) and (B,C) are connected spouse-pairs so that {A,B,C} is a spouse-connected set. Similarly, {D,E,F} is a spouse-connected set.

Fig. 3.1 Examples of simple pedigrees



(a) Ancestors of both members of spouse-pair (A,B) are indicated.
(b) Ancestors of more than one member of spouse-connected set {A,B,C} and spouse-pair (D,E) are indicated.

Fig. 3.2 Examples of complex pedigrees without loops



(a) Consanguineous loop;
(b) Non-consanguineous loop.

Fig. 3.3 Examples of complex pedigrees with loops

sibship and their spouses in the j-th generation, conditional on the sibs' parents being of genotypes s_{j-1}, t_{j-1} respectively follows in an obvious manner from the univariate case (1.32), i.e.,

$$\begin{cases} \Gamma_0 = \sum_{s_0=1}^{k} \psi_{s_0} g_{s_0} (\underset{\sim}{x}_{i_0}) \sum_{t_0=1}^{k} \psi_{t_0} g_{t_0} (\underset{\sim}{y}_{i_0}) \\ \Gamma_j = \prod_{i_j} \sum_{s_j=1}^{k} p_{s_{j-1}t_{j-1}s_j} g_{s_j} (\underset{\sim}{x}_{i_0 i_1, \ldots i_j}) \sum_{t_j=1}^{k} \psi_{t_j} g_{t_j} (\underset{\sim}{y}_{i_0 i_1 \ldots i_j}), \end{cases} \tag{3.1}$$

$$j \geq 1$$

where $\underset{\sim}{x}_{i_0 i_1 \ldots i_j}$ is the phenotype value of $\underset{\sim}{z}$ of the j-th sib in the i-th generation and $\underset{\sim}{y}_{i_0 i_1 \ldots i_j}$ is the corresponding phenotype of the spouse, and ψ_{t_j} denotes the population frequency of the t_j-th genotype. Consequently the likelihood of observing the entire pedigree is, from (1.33),

$$L_0 = \Gamma_0(\Gamma_1(\Gamma_2(\Gamma_3 \ldots))) \tag{3.2}$$

where L_0 is the likelihood under oligogenic inheritance. As for the case of univariate traits, we assume implicitly that mating is random with respect to the k genotypes, and pedigrees are random. We further assume that all p variates are measured on every available member of the pedigree. For missing values (unavailable individuals) the corresponding $g_u(\underset{\sim}{z})$ may be simply set to 1 wherever they occur, as for the univariate case.

If only one locus is involved, p_{stu} will take on one of the values of 0, ¼, ½, or 1; if w independent loci are involved it must be the product of w factors each of which is 0, ¼, ½, or 1. We should note that the p_{stu} are identical to those for univariate traits and so the genetic transmission matrices presented by Elston and Stewart (1971) for the cases of one autosomal locus, one X-linked locus, multiple unlinked loci and two linked autosomal loci, may be used.

The obvious advantage of the Elston and Stewart representation ((3.1) and (3.2)) is that the likelihood of an individual can be computed first and the result attached as a factor to the appropriate term in the summation for his parents. The individual is then no longer needed in further computations. This process of elimination is repeated for each individual. Thus the highest order of the matrices that enter into the calculations is p x p, for a p-variate trait, regardless of the size of the pedigree.

However in terms of the Γ operators the nature of the likelihood function is obscure. As Ott (1974) did, we can write down the expanded form of (3.2) without the operators. Assume the n individuals in the pedigree (assumed simple) are ordered so that children have higher numbers than their parents. Let I denote the set of individuals in the pedigree whose parental lines are indicated, M the set of those whose parental lines are not indicated, and T the total set of individuals indicated in the pedigree, i.e., T = I + M. Note that we are using the letter I for individuals 'in' the pedigree and M for 'marry ins'. These are the terms used originally by Elston and Stewart (1971). With these conventions, the likelihood function (3.2) of the entire pedigree is simply

$$L_0 = \prod_{i=1}^{n} \sum_{u_i=1}^{k} p_i(u_i)\, g_{u_i}(\underset{\sim}{z}_i) \qquad (3.3)$$

where

$$p_i(u_i) = \begin{cases} P_{stu} & \text{if} \quad i \in I \\ \psi_{u_i} & \text{if} \quad i \in M \end{cases} \qquad (3.4)$$

An alternative form for (3.3) is easily seen to be

$$L_0 = \sum_{\substack{\text{all n-tuples}\\(u_1,u_2,\ldots,u_n)}} \prod_{i=1}^{n} p_i(u_i) g_{u_i}(\underset{\sim}{z}_i). \tag{3.5}$$

Since for every i, u_i can take k different values, there are k^n different n-tuples, i.e., the summation in (3.5) is over k^n expressions, each of which is the product of n factors. This n-dimensional summation can also be written

$$L_0 = \sum_{u_1} \sum_{u_2} \ldots \sum_{u_n} \prod_{i=1}^{n} p_i(u_i) \prod_{i=1}^{n} g_{u_i}(\underset{\sim}{z}_i), \tag{3.6}$$

a representation that we shall refer to in the sequel.

3.4 Polygenic Inheritance in Simple Pedigrees

3.4.1 The Elston-Stewart formulation

Consider now that each component of $\underset{\sim}{z}$ is under polygenic control, i.e., $\underset{\sim}{z}$ is under the control of an infinite number of genetic loci which are separately indistinguishable, but additive, and in Hardy-Weinberg equilibrium. We have shown in section 2.2 that ψ_{t_j}, the population distribution becomes $\phi(\underset{\sim}{t}_j,V_g)$; the distribution of an offspring conditional on his parents' genotypes, $p_{s-1,t-1,s_j}$, becomes the density $\phi(\underset{\sim}{s}_j - [\alpha_{1j}\underset{\sim}{s}_{j-1} + \alpha_{2j}\underset{\sim}{t}_{j-1}], \alpha_{3j}V_g)$. The relatedness factors α_{1j}, α_{2j} and α_{3j} are each equal to ½ while we are dealing with simple pedigrees with no consanguineous matings. Substituting in (3.1) and changing all summations to integrals we obtain

$$\begin{cases} \Gamma_0 = \int_{\underset{\sim}{s}_0} \phi(\underset{\sim}{s}_0,V_g) g_{\underset{\sim}{s}_0}(\underset{\sim}{x}_{i_0}) \int_{\underset{\sim}{t}_0} \phi(\underset{\sim}{t}_0,V_g) g_{\underset{\sim}{t}_0}(\underset{\sim}{y}_{i_0}) \\ \Gamma_j = \prod_{i_j} \int_{\underset{\sim}{s}_j} \phi(\underset{\sim}{s}_j - \tfrac{1}{2}[\underset{\sim}{s}_{j-1} + \underset{\sim}{t}_{j-1}], \tfrac{1}{2}V_g) \cdot g_{\underset{\sim}{s}_j}(\underset{\sim}{x}_{i_0\, i_1 \,\ldots\, i_j}) \\ \qquad\qquad x \int_{\underset{\sim}{t}_j} \phi(\underset{\sim}{t}_j,V_g) g_{\underset{\sim}{t}_j}(\underset{\sim}{y}_{i_0\, i_1 \ldots i_j}) \qquad j \geq 1 \end{cases} \tag{3.7}$$

where $g_{\underset{\sim}{s}_j}(x_{i_0 i_1 \ldots i_j})$ and $g_{\underset{\sim}{t}_j}(y_{i_0 i_1 \ldots i_j})$ are the phenotypic distributions conditional on the genotypic vectors $\underset{\sim}{s}_j$ and $\underset{\sim}{t}_j$ respectively. In the case of multinormal phenotypes,

$$g_{\underset{\sim}{s}_j}(\underset{\sim}{x}_{i_0 i_1 \ldots i_j}) = \phi(\underset{\sim}{s}_j + \underset{\sim}{\mu} - \underset{\sim}{x}_{i_0 i_1 \ldots i_j}, V_e) \tag{3.8}$$

and

$$g_{\underset{\sim}{t}_j}(\underset{\sim}{y}_{i_0 i_1 \ldots i_j}) = \phi(\underset{\sim}{t}_j + \underset{\sim}{\mu} - \underset{\sim}{y}_{i_0 i_1 \ldots i_j}, V_e) \tag{3.9}$$

For multinormal phenotypes (3.7) therefore becomes

$$\left\{ \begin{aligned} \Gamma_0 &= \int_{\underset{\sim}{s}_0} \phi(\underset{\sim}{s}_0, V)\ \phi(\underset{\sim}{s}_0 + \underset{\sim}{\mu} - \underset{\sim}{x}_{i_0}, V_e) \int_{\underset{\sim}{t}_0} \phi(\underset{\sim}{t}_0, V_g)\ \phi(\underset{\sim}{t}_0 + \underset{\sim}{\mu} - \underset{\sim}{y}_{i_0}, V_e) \\ \Gamma_j &= \prod_{i_j} \int_{\underset{\sim}{s}_j} \phi(\underset{\sim}{s}_j - \tfrac{1}{2}[\underset{\sim}{s}_{j-1} + \underset{\sim}{t}_{j-1}], \tfrac{1}{2}V_g)\ \phi(\underset{\sim}{s}_j + \underset{\sim}{\mu} - \underset{\sim}{x}_{i_0 i_1 \ldots i_j}, V_e) \\ &\qquad \times \int_{\underset{\sim}{t}_j} \phi(\underset{\sim}{t}_j, V_g)\ \phi(\underset{\sim}{t}_j + \underset{\sim}{\mu} - \underset{\sim}{y}_{i_0 i_1 \ldots i_j}, V_e), \quad j \geq 1. \end{aligned} \right. \tag{3.10}$$

The likelihood involves integrals of products of multinormal densities. These can be evaluated analytically. Before deriving the necessary formula let us first state the following result which is really a multivariate generalization of the method of 'completing the squares.' The result must be well know. For example, it seems implicit in results on quadratic forms (e.g., Searle, 1971, pp. 68 following). We state it here because we use it repeatedly in what follows.

Lemma 3.1

For p-vectors $\underset{\sim}{x}$, $\underset{\sim}{y}$, a pxp symmetric, nonsingular matrix Q, and a scalar c,

$$\underset{\sim}{x}'\underset{\sim}{Q}\underset{\sim}{x} + 2\underset{\sim}{x}'\underset{\sim}{y} + c = (\underset{\sim}{x}+\underset{\sim}{z})'\ Q(\underset{\sim}{x}+\underset{\sim}{z}) + c - \underset{\sim}{z}'Q\underset{\sim}{z} \tag{3.11}$$

where

$$\underset{\sim}{z} = \underset{\sim}{Q}^{-1}\underset{\sim}{y}. \tag{3.12}$$

Proof

By multiplying out we see that

$$(\underset{\sim}{x}+\underset{\sim}{z})'\,\underset{\sim}{Q}(\underset{\sim}{x}+\underset{\sim}{z}) = \underset{\sim}{x}'\underset{\sim}{Q}\underset{\sim}{x} + 2\underset{\sim}{x}'\underset{\sim}{Q}\underset{\sim}{z} + \underset{\sim}{z}'\underset{\sim}{Q}\underset{\sim}{z} \qquad (3.13)$$

and the result immediately follows if we set

$$\underset{\sim}{Q}\underset{\sim}{z} = \underset{\sim}{y}. \qquad (3.14)$$

We shall now state the main result.

Theorem 3.1

Let $A_i (i = 1,2,...,m)$ be pxp matrices not all zero, C_i pxp symmetric, nonsingular matrices such that $\sum_i C_i^{-1}$ is nonsingular, and $\underset{\sim}{b}_i$ be p-vectors. Then

$$\int_{\underset{\sim}{s}} \prod_{i=1}^{m} \phi(\underset{\sim}{s}+A_i\underset{\sim}{t}+\underset{\sim}{b}_i, C_i) = \alpha\phi(\underset{\sim}{t}+\underset{\sim}{\nu}, H) \qquad (3.15)$$

where $$H = [\sum_i A_i'C_i^{-1}A_i - (\sum_i C_i^{-1}A_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1}A_i)]^{-1}, \qquad (3.16)$$

$$\underset{\sim}{\nu} = H[\sum_i A_i'C_i^{-1}\underset{\sim}{b}_i - (\sum_i C_i^{-1}A_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1}\underset{\sim}{b}_i)], \qquad (3.17)$$

and

$$\alpha = (2\pi)^{(1-\frac{m}{2})p} \prod_i |C_i|^{-\frac{1}{2}} |\sum_i C_i^{-1}|^{-\frac{1}{2}} \; |H|^{\frac{1}{2}} \qquad (3.18)$$

$$\times\; e^{-\frac{1}{2}\{\sum_i \underset{\sim}{b}'_i C_i^{-1}\underset{\sim}{b} \;-\; (\sum_i C_i^{-1}\underset{\sim}{b}_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1}\underset{\sim}{b}_i) \;-\; \underset{\sim}{\nu}'H^{-1}\underset{\sim}{\nu}\}}$$

Proof

$$\prod_i \phi(\underset{\sim}{s}+A_i\underset{\sim}{t}+\underset{\sim}{b}_i, C_i) = ae^{-\frac{1}{2}d}$$

where $$a = (2\pi)^{-mp/2} \prod_i |C_i|^{-\frac{1}{2}} \qquad (3.19)$$

and

$$\begin{aligned} d &= \sum_i (\underset{\sim}{s}+A_i\underset{\sim}{t}+\underset{\sim}{b}_i)'\ C_i^{-1}(\underset{\sim}{s}+A_i\underset{\sim}{t}+\underset{\sim}{b}_i) \\ &= \underset{\sim}{s}'(\sum_i C_i^{-1})\underset{\sim}{s} + 2\underset{\sim}{s}' \sum_i C_i^{-1}(A_i\underset{\sim}{t}+\underset{\sim}{b}_i) + \sum_i (A_i\underset{\sim}{t}+\underset{\sim}{b}_i)'\ C_i^{-1}(A_i\underset{\sim}{t}+\underset{\sim}{b}_i) \\ &= (\underset{\sim}{s}+\underset{\sim}{g})'\ (\sum_i C_i^{-1})\quad (\underset{\sim}{s}+\underset{\sim}{g}) + h - \underset{\sim}{g}'(\sum_i C_i^{-1})\underset{\sim}{g} \end{aligned}$$

by Lemma 3.1, where

$$\begin{aligned} \underset{\sim}{g} &= (\sum_i C_i^{-1})^{-1} \sum_i C_i^{-1}(A_i\underset{\sim}{t}+\underset{\sim}{b}_i) \\ &= (\sum_i C_i^{-1})^{-1}\ [(\sum_i C_i^{-1}A_i)\underset{\sim}{t} + \sum_i C_i^{-1}\underset{\sim}{b}_i], \end{aligned}$$

and

$$\begin{aligned} h &= \sum_i (A_i\underset{\sim}{t}+\underset{\sim}{b}_i)'\ C_i^{-1}(A_i\underset{\sim}{t}+\underset{\sim}{b}_i) \\ &= \underset{\sim}{t}'(\sum_i A_i'C_i^{-1}A_i)\underset{\sim}{t} + 2\underset{\sim}{t}'(\sum_i A_i'C_i^{-1}\underset{\sim}{b}_i) + \sum_i \underset{\sim}{b}_i'C_i^{-1}\underset{\sim}{b}_i. \end{aligned}$$

We therefore have

$$\begin{aligned} \int_{\underset{\sim}{s}} ae^{-\frac{1}{2}d} &= ae^{-\frac{1}{2}\{h-\underset{\sim}{g}'(\sum_i C_i^{-1})\underset{\sim}{g}\}} \int_{\underset{\sim}{s}} e^{-\frac{1}{2}\{(\underset{\sim}{s}+\underset{\sim}{g})'(\sum_i C_i^{-1})(\underset{\sim}{s}+\underset{\sim}{g})\}} \\ &= ae^{-\frac{1}{2}\{h-\underset{\sim}{g}'(\sum_i C_i^{-1})\underset{\sim}{g}\}} \cdot (2\pi)^{p/2}\left|\sum_i C_i^{-1}\right|^{-\frac{1}{2}}, \end{aligned}$$

using Aiken's integral (e.g., Searle, 1971, p. 24).

Now

$$\begin{aligned} \underset{\sim}{g}'(\sum_i C_i^{-1})\underset{\sim}{g} &= [(\sum_i C_i^{-1}A_i)\underset{\sim}{t} + \sum_i C_i^{-1}\underset{\sim}{b}_i]'(\sum_i C_i^{-1})^{-1}[(\sum_i C_i^{-1}A_i)\underset{\sim}{t} + \sum_i C_i^{-1}\underset{\sim}{b}_i] \\ &= \underset{\sim}{t}'(\sum_i C_i^{-1}A_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1}A_i)\underset{\sim}{t} \\ &\quad + 2\underset{\sim}{t}'(\sum_i C_i^{-1}A_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1}\underset{\sim}{b}_i) \\ &\quad + (\sum_i C_i^{-1}\underset{\sim}{b}_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1}\underset{\sim}{b}_i), \end{aligned}$$

so that

$$
\begin{aligned}
\underset{\sim}{h}-\underset{\sim}{g}'(\sum_i C_i^{-1})\underset{\sim}{g} &= \underset{\sim}{t}'[\sum_i A_i' C_i^{-1} A_i - (\sum_i C_i^{-1} A_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1} A_i)]\underset{\sim}{t} \\
&\quad + 2\underset{\sim}{t}'[\sum_i A_i' C_i^{-1} \underset{\sim}{b}_i - (\sum_i C_i^{-1} A_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1} \underset{\sim}{b}_i)] \\
&\quad + \sum_i \underset{\sim}{b}_i' C_i^{-1} \underset{\sim}{b}_i - (\sum_i C_i^{-1} \underset{\sim}{b}_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1} \underset{\sim}{b}_i) \\
&= \underset{\sim}{t}'H^{-1}\underset{\sim}{t} + 2\underset{\sim}{t}'H^{-1}\underset{\sim}{\nu} + w
\end{aligned}
$$

where H and $\underset{\sim}{\nu}$ are defined by (3.16) and (3.17) respectively, and

$$
w = \sum_i \underset{\sim}{b}_i' C_i^{-1} \underset{\sim}{b}_i - (\sum_i C_i^{-1} \underset{\sim}{b}_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1} \underset{\sim}{b}_i)
$$

$$
\underset{\sim}{h}-\underset{\sim}{g}'(\sum_i C_i^{-1})\underset{\sim}{g} = (\underset{\sim}{t}+\underset{\sim}{\nu})'H^{-1}(\underset{\sim}{t}+\underset{\sim}{\nu}) + w - \underset{\sim}{\nu}'H^{-1}\underset{\sim}{\nu},
$$

by Lemma 3.1.

Thus

$$
\begin{aligned}
\int a e^{-\frac{1}{2}d} &= a(2\pi)^{p/2}|\sum_i C_i^{-1}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\{\underset{\sim}{h}-\underset{\sim}{g}'(\sum_i C_i^{-1})\underset{\sim}{g}\}} \\
&= a(2\pi)^{p/2}|\sum_i C_i^{-1}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\{w-\underset{\sim}{\nu}'H^{-1}\underset{\sim}{\nu}\}} \cdot e^{-\frac{1}{2}(\underset{\sim}{t}+\underset{\sim}{\nu})'H^{-1}(\underset{\sim}{t}+\underset{\sim}{\nu})} \\
&= \alpha\ \phi(\underset{\sim}{t}+\underset{\sim}{\nu}, H)
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha &= a(2\pi)^{p/2}|\sum_i C_i^{-1}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\{w-\underset{\sim}{\nu}'H^{-1}\underset{\sim}{\nu}\}} \cdot (2\pi)^{p/2}|H|^{\frac{1}{2}} \\
&= (2\pi)^{-mp/2}\prod_i |C_i|^{-\frac{1}{2}} \cdot (2\pi)^{p/2}|\sum_i C_i^{-1}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\{w-\underset{\sim}{\nu}'H^{-1}\underset{\sim}{\nu}\}} \cdot (2\pi)^{p/2}|H|^{\frac{1}{2}} \\
&= (2\pi)^{(1-m/2)p}\prod_i |C_i|^{-\frac{1}{2}}\ |\sum_i C_i^{-1}|^{-\frac{1}{2}}\ |H|^{\frac{1}{2}} \\
&\qquad \times e^{-\frac{1}{2}\{\sum_i \underset{\sim}{b}_i' C_i^{-1} \underset{\sim}{b}_i - (\sum_i C_i^{-1}\underset{\sim}{b}_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1}\underset{\sim}{b}_i) - \underset{\sim}{\nu}'H^{-1}\underset{\sim}{\nu}\}}.
\end{aligned}
$$

That concludes the proof.

The formula (3.15) can be used recursively to evaluate the Γ's in (3.10) and hence the likelihood (3.2) of polygenic inheritance in a simple pedigree.

It is evident from the form of H in (3.16) that complicated matrix expressions can result from the use of Theorem 3.1. In certain important cases we can simplify H using the following result.

Lemma 3.2

For symmetric nonsingular matrices A and B such that $A^{-1} + B^{-1}$ is also nonsingular,

$$A + B = \{A^{-1}-A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}\}^{-1} = \{B^{-1}-B^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}\}^{-1} \tag{3.22}$$

Proof

$$\begin{aligned}\{A^{-1}-A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}\}^{-1} &= \{I-(A^{-1}+B^{-1})^{-1}A^{-1}\}^{-1}A \\ &= \{I-(I+AB^{-1})^{-1}\}^{-1}A \\ &= \{(I+AB^{-1}-I)(I+AB^{-1})^{-1}\}^{-1}A \\ &= (I+AB^{-1})BA^{-1}A \\ &= A+B.\end{aligned}$$

The second part follows by symmetry.

To illuatrate the use of these formulas, let us first calculate the likelihood of an individual picked at random from the population. This is of course the same as the population distribution of a polygenic trait $\underset{\sim}{z}$, which is given by

$$L_R = \int_{\underset{\sim}{s}} \phi(\underset{\sim}{s},V_g)\ \phi(\underset{\sim}{s}+\underset{\sim}{\mu}-\underset{\sim}{z},V_e). \tag{3.23}$$

In Theorem 3.1 set

$$\underset{\sim}{t} = \underset{\sim}{\mu} - \underset{\sim}{z}$$

$$A_1 = \underset{\sim}{0},\ A_2 = I,$$

$$\underset{\sim}{b}_1 = \underset{\sim}{b}_2 = \underset{\sim}{0},$$

$$C_1 = V_g,\ C_2 = V_e$$

Then

$$\begin{aligned} H &= [\sum_i A_i' C_i^{-1} A_i - (\sum_i C_i^{-1} A_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1} A_i)]^{-1} \\ &= [V_e^{-1} - V_e^{-1}(V_g^{-1}+V_e^{-1})^{-1}V_e^{-1}]^{-1} \\ &= V_g + V_e \text{ by Lemma 3.2.} \end{aligned}$$

$$\underset{\sim}{\nu} = H[\sum_i A_i' C_i^{-1} \underset{\sim}{b}_i - (\sum_i C_i^{-1} A_i)'(\sum_i C_i^{-1})^{-1}(\sum_i C_i^{-1} \underset{\sim}{b}_i)] = \underset{\sim}{0}$$

$$\alpha = (2\bar{\Lambda})^{(1-2/2)p}|V_g|^{-\frac{1}{2}}|V_e|^{-\frac{1}{2}}|V_g^{-1}+V_e^{-1}|^{-\frac{1}{2}}|H|^{\frac{1}{2}}\, e^{-\frac{1}{2}\cdot 0} = 1,$$

since $$|V_g|^{-\frac{1}{2}}|V_e|^{-\frac{1}{2}}|V_g^{-1}+V_e^{-1}|^{-\frac{1}{2}}|H|^{\frac{1}{2}} = 1;$$

for $$|V_g|^{-\frac{1}{2}}|V_e|^{-\frac{1}{2}}|V_g^{-1}+V_e^{-1}|^{-\frac{1}{2}} = |V_g+V_e|^{-\frac{1}{2}} = |H|^{-\frac{1}{2}}.$$

Substituing in Theorem 3.1 we obtain

$$L_R = \int_{\underset{\sim}{s}} \phi(\underset{\sim}{s},V_g)\ \phi(\underset{\sim}{s}+\underset{\sim}{\mu}-\underset{\sim}{z},V_e) = \phi(\underset{\sim}{\mu}-\underset{\sim}{z},V_g+V_e), \qquad (3.24)$$

i.e., the distribution of a polygenic trait in the population is multinormal with mean $\underset{\sim}{\mu}$ and covariance matrix $V_g + V_e$.

As another important application we shall compute the likelihood of a sib given the parental genotypic effects are $\underset{\sim}{s}_{P_1}$ and $\underset{\sim}{s}_{P_2}$. This is also the distribution among sibs conditional on the parents' genotypes. It is given by

$$L_{s|p} = \int_{\underset{\sim}{s}} \phi(\underset{\sim}{s}-\underset{\sim}{m},\tfrac{1}{2}V_g)\ \phi(\underset{\sim}{s}+\underset{\sim}{\mu}-\underset{\sim}{z},V_e) \qquad (3.25)$$

where $\underset{\sim}{m} = \frac{1}{2}[\underset{\sim}{s}_{P_1}+\underset{\sim}{s}_{P_2}]$. The likelihood can easily be evaluated if we note that the integral can be rewritten as

$$\int_{\underset{\sim}{s}-\underset{\sim}{m}} \phi(\underset{\sim}{s}-\underset{\sim}{m},\tfrac{1}{2}V_g)\ \phi(\underset{\sim}{s}-\underset{\sim}{m}+\underset{\sim}{m}+\underset{\sim}{\mu}-\underset{\sim}{z},V_e). \qquad (3.26)$$

Substituting in Theorem 3.1,

$$\underset{\sim}{t} = \underset{\sim}{m} + \underset{\sim}{\mu} - \underset{\sim}{z}$$

$$A_1 = \underset{\sim}{0},\ A_2 = I$$

$$\underset{\sim}{b}_1 = \underset{\sim}{b}_2 = \underset{\sim}{0}$$

$$C_1 = \tfrac{1}{2}V_g,\ C_2 = V_e,$$

we obtain

$$H = [V_e^{-1}-V_e^{-1}(2V_g^{-1}+V_e^{-1})^{-1}V_e^{-1}]^{-1}$$

$$= \tfrac{1}{2}V_g + V_e \text{ by Lemma 3.2,}$$

$$\underset{\sim}{\nu} = \underset{\sim}{0}, \text{ and } \alpha = 1.$$

Thus

$$L_{s|p} = \phi(\tfrac{1}{2}[\underset{\sim}{s}_{p_1}+\underset{\sim}{s}_{p_2}] + \underset{\sim}{\mu} - \underset{\sim}{z},\ \tfrac{1}{2}V_g+V_e) \qquad (3.27)$$

So our model implies that the distribution of the trait among sibs given the parents' genotypes is multinormal with mean $\underset{\sim}{\mu} + \tfrac{1}{2}[\underset{\sim}{s}_{p_1}+\underset{\sim}{s}_{p_2}]$ and covariance matrix $\tfrac{1}{2}V_g + V_e$. Without loss of generality $\underset{\sim}{\mu}$ may be set to $\underset{\sim}{0}$. The results in (3.24) and (3.27) then reduce to the mathematical model for polygenic inheritance according to our formulation in section 2.4.

3.4.2 Multivariate generalization of Green's Recurrence formula

We have indicated above that Theorem 3.1 can be used recursively to compute the likelihood of the entire pedigree. The essence of the procedure is to integrate out the polygenic effect in the integral for each individual and attach the result as a factor to the remaining integrals. It is then possible by a slightly different representation of the likelihood function to develop a simple recurrence relation that defines the factors that result from such an elimination process, and enables us to write down an explicit expression for the likelihood of the entire pedigree. We first define a set of status indices that gives the relevant information on an individual in a simple pedigree.

Definition 3.6: Status indices

The status indices of the i-th individual (each set of monozygotic (MZ) sibs being considered a single individual for this purpose) in a simple pedigree are the following scalar quantities.

$$a_i = \begin{cases} 0 & \text{if } i \text{ is not observed} \\ 1 & \text{if } i \text{ is an observed singleton} \\ v & \text{if } i \text{ is a set of } v \text{ observed MZ sibs} \end{cases}$$

$$b_i = \begin{cases} 1 & \text{if } i \in M \\ 2 & \text{if } i \in I \end{cases}$$

$$c_i = \tfrac{1}{2} \times \{\text{number of children of } i \text{ observed}\},$$

$$c_{ij} = \tfrac{1}{2} \times \{\text{number of children } i \text{ and } j \text{ have in common that are observed}\},$$

and

$$d_{ij} = \begin{cases} 1 & \text{if } i,j \text{ are a parent-offspring pair, i.e., } i = p_1(j) \text{ or } p_2(j), \text{ or } j = p_1(i) \text{ or } p_2(i). \\ 0 & \text{otherwise .} \end{cases}$$

We can now state the main result.

Theorem 3.2

Under polygenic inheritance the likelihood of a simple pedigree with n individuals, each one observed, is

$$L_p = 2^{n_I}(2\pi)^{-\frac{1}{2}np}|V_g|^{-n/2}|V_e|^{-n/2}\{\prod_{i=1}^{n}|D_{i,i}^{(i)}|^{-\frac{1}{2}}\}e^{-\frac{1}{2}D_{0,0}^{(0)}} \qquad (3.28)$$

where n_I = the number of individuals in I, and the D's satisfy the recurrence relations

$$D^{(m-1)}_{i,j} = D^{(m)}_{i,j} - D^{(m)'}_{m,i} D^{(m)-1}_{m,m} D^{(m)}_{m,j} \tag{3.29}$$

$$(m = 1,2,\ldots,n;\ i,j = 0,1,2,\ldots,n)$$

with initial conditions defined by

$$D^{(n)}_{i,i} = a_i V_e^{-1} + (b_i+c_i)V_g^{-1} \tag{3.30}$$

$$D^{(n)}_{i,j} = (c_{ij}-d_{ij})V_g^{-1}, \qquad i \neq j \tag{3.31}$$

$$D^{(n)}_{i,0} = -V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i),\ \text{a pxl vector} \tag{3.32}$$

$$D^{(n)}_{0,0} = \sum_{i=1}^{n} (\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)'V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i),\ \text{a scaler.} \tag{3.33}$$

Proof

The polygenic version of the representation (3.6) is

$$L_p = \int_{\underset{\sim}{s}_1}\cdots\int_{\underset{\sim}{s}_n} \prod_{i\varepsilon I} \phi(\underset{\sim}{s}_i-\tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)}+\underset{\sim}{s}_{p_2(i)}],\tfrac{1}{2}V_g) \prod_{i\varepsilon M} \phi(\underset{\sim}{s}_i,V_g) \times \prod_{i\varepsilon T} \phi(\underset{\sim}{s}_i+\underset{\sim}{\mu}_i-\underset{\sim}{z}_i,V_e) \tag{3.34}$$

$$= \int_{\underset{\sim}{s}_1}\cdots\int_{\underset{\sim}{s}_n} \alpha\, e^{-\frac{1}{2}\beta^{(n)}}$$

where L_p is the likelihood under polygenic inheritance,

$$\alpha = \prod_{i\varepsilon I} (2\pi)^{-\frac{1}{2}p}|\tfrac{1}{2}V_g|^{-\frac{1}{2}} \prod_{i\varepsilon M} (2\pi)^{-\frac{1}{2}p}|V_g|^{-\frac{1}{2}} \prod_{i\varepsilon T} (2\pi)^{-\frac{1}{2}p}|V_e|^{-\frac{1}{2}}$$

$$= 2^{n_I}(2\pi)^{-np}|V_g|^{-n/2}|V_e|^{-n/2}, \tag{3.35}$$

and

$$\beta^{(n)} = 2\sum_{i\varepsilon I} (\underset{\sim}{s}_i-\tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)}+\underset{\sim}{s}_{p_2(i)}])'V_g^{-1}(\underset{\sim}{s}_i-\tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)}+\underset{\sim}{s}_{p_2(i)}]) + \sum_{i\varepsilon M} \underset{\sim}{s}_i'V_g^{-1}\underset{\sim}{s}_i + \sum_{i\varepsilon T} (\underset{\sim}{z}_i-\underset{\sim}{\mu}_i-\underset{\sim}{s}_i)'V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i-\underset{\sim}{s}_i) \tag{3.36}$$

Expanding and rearranging so that quadratic terms in $\underset{\sim}{s}_i$ come first, followed by cross-product terms, linear terms and constants, we obtain

$$\begin{aligned}\beta^{(n)} = {} & \sum_{i\varepsilon T} \underset{\sim}{s}' V_e^{-1} \underset{\sim}{s}_i + 2 \sum_{i\varepsilon I} \underset{\sim}{s}_i' V_g^{-1} \underset{\sim}{s}_i + \sum_{i\varepsilon M} \underset{\sim}{s}_i' V_g^{-1} \underset{\sim}{s}_i \\ & + \tfrac{1}{2} \sum_{i\varepsilon I} \underset{\sim}{s}_{p_1(i)}' V_g^{-1} \underset{\sim}{s}_{p_1(i)} + \tfrac{1}{2} \sum_{i\varepsilon I} \underset{\sim}{s}_{p_2(i)}' V_g^{-1} \underset{\sim}{s}_{p_2(i)} \\ & - 2 \sum_{i\varepsilon I} \underset{\sim}{s}_{p_1(i)}' V_g^{-1} \underset{\sim}{s}_i - 2 \sum_{i\varepsilon I} \underset{\sim}{s}_{p_2(i)}' V_g^{-1} \underset{\sim}{s}_i + \sum_{i\varepsilon I} \underset{\sim}{s}_{p_1(i)}' V_g^{-1} \underset{\sim}{s}_{p_2(i)} \\ & - 2 \sum_{i\varepsilon T} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i)' V_e^{-1} \underset{\sim}{s}_i + \sum_{i\varepsilon T} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i)' V_e^{-1} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i)\end{aligned} \tag{3.37}$$

$$\begin{aligned} = {} & \sum_{i=1}^{n} \underset{\sim}{s}_i' [a_i V_e^{-1} + b_i V_g^{-1} + c_i V_g^{-1}] \underset{\sim}{s}_i \\ & + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \underset{\sim}{s}_i' [-d_{ij} V_g^{-1} + c_{ij} V_g^{-1}] \underset{\sim}{s}_j \\ & - 2 \sum_{i=1}^{n} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i)' V_e^{-1} \underset{\sim}{s}_i + \sum_{i=1}^{n} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i)' V_e^{-1} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i)\end{aligned} \tag{3.38}$$

$$= \sum_{i=1}^{n} \underset{\sim}{s}_i' D_{i,i}^{(n)} \underset{\sim}{s}_i + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \underset{\sim}{s}_i' D_{i,j}^{(n)} \underset{\sim}{s}_j + 2 \sum_{i=1}^{n} D_{i,0}^{(n)\prime} \underset{\sim}{s}_i + D_{0,0}^{(n)} \tag{3.39}$$

The D's are obtained by the following argument. First consider $D_{i,i}^{(n)}$ appearing in the quadratic forms in $\underset{\sim}{s}_i$. If individual i is observed we have one quadratic form involving V_e^{-1} (from the first summation in (3.37)). Hence the term $a_i V_e^{-1}$. This individual belongs to either the set I or M, never both. If he belongs to I we have, from the second summation in (3.37), one quadratic form in $\underset{\sim}{s}_i$ involving $2V_g^{-1}$. On the other hand, if he belongs to the set M we have one term from the third term in (3.37) involving V_g^{-1}. Hence the second term in $D_{i,i}^{(n)}$ is $b_i V_g^{-1}$. The third term in $D_{i,i}^{(n)}$ is $c_i V_g^{-1}$. This comes from the fact that if the individual i is a parent, the fourth and fifth summations in (3.37) imply every child of i contributes to the quadratic form in $\underset{\sim}{s}_i$ a term involving $\tfrac{1}{2}V_g^{-1}$.

The derivation of $D_{i,j}^{(n)}$ follows similarly by inspection of the cross-products (bilinear forms) in $\underset{\sim}{s}_i$ and $\underset{\sim}{s}_j$. Also the origin of $D_{i,0}^{(n)}$ and $D_{0,0}^{(n)}$ is clear.

We start the elimination process with the n-th individual in the pedigree, so first pick out the term in $\underset{\sim}{s}_n$ from $\beta^{(n)}$, and group together as $\gamma^{(n)}$ all terms in $\underset{\sim}{s}_{n-1}$, $\underset{\sim}{s}_{n-2}$, etc. Then (3.39) becomes

$$\beta^{(n)} = \underset{\sim}{s}_n' D_{n,n}^{(n)} \underset{\sim}{s}_n + 2[\sum_{j=1}^{n-1} D_{n,j}^{(n)} \underset{\sim}{s}_j + D_{i,0}^{(n)}]' \underset{\sim}{s}_n + \gamma^{(n)} \tag{3.40}$$

where $\gamma^{(n)} = \sum_{i=1}^{n-1} \underset{\sim}{s}_i' D_{i,i}^{(n)} \underset{\sim}{s}_i + 2 \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \underset{\sim}{s}_i' D_{i,j}^{(n)} \underset{\sim}{s}_j + 2 \sum_{i=1}^{n-1} D_{i,0}^{(n)'} \underset{\sim}{s}_i + D_{0,0}^{(n)}$.

Applying Lemma (3.1) we can rewrite (3.40) as

$$\beta^{(n)} = (\underset{\sim}{s}_n + \underset{\sim}{\tau}^{(n)})' D_{n,n}^{(n)} (\underset{\sim}{s}_n + \underset{\sim}{\tau}^{(n)}) + \gamma^{(n)} - \delta^{(n)} \tag{3.41}$$

where $\underset{\sim}{\tau}^{(n)} = D_{n,n}^{(n)^{-1}} [\sum_{j=1}^{n-1} D_{n,j}^{(n)'} + D_{n,0}^{(n)}]$

and $\delta^{(n)} = \underset{\sim}{\tau}^{(n)'} D_{n,n}^{(n)} \underset{\sim}{\tau}^{(n)}$.

Integrating out $\underset{\sim}{s}_n$ we obtain

$$\alpha \int_{\underset{\sim}{s}_1} \cdots \int_{\underset{\sim}{s}_n} e^{-\frac{1}{2}\beta(n)} = \alpha \int_{\underset{\sim}{s}_n} e^{-\frac{1}{2}(\underset{\sim}{s}_n + \underset{\sim}{\tau}^{(n)})' D_{n,n}^{(n)} (\underset{\sim}{s}_n + \underset{\sim}{\tau}^{(n)})} \times \int_{\underset{\sim}{s}_1} \cdots \int_{\underset{\sim}{s}_{n-1}} e^{-\frac{1}{2}(\gamma^{(n)} - \delta^{(n)})}$$

$$= \alpha (2\pi)^{p/2} |D_{n,n}^{(n)}|^{-\frac{1}{2}} \int_{\underset{\sim}{s}_1} \cdots \int_{\underset{\sim}{s}_{n-1}} e^{-\frac{1}{2}\beta^{(n-1)}} \tag{3.42}$$

where $\beta^{(n-1)} = \gamma^{(n)} - \delta^{(n)}$. (3.43)

Repeat the process, i.e., define matrices $D_{i,j}^{(n-1)}$ and integrate out $\underset{\sim}{s}_{n-1}$. But note that

$$\gamma^{(n)} = \sum_{i=1}^{n-1} \underset{\sim}{s}'_i D_{i,i}^{(n)} \underset{\sim}{s}_i + 2 \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \underset{\sim}{s}_i' D_{i,j}^{(n)} \underset{\sim}{s}_j + 2 \sum_{i=1}^{n-1} D_{i,0}^{(n)'} \underset{\sim}{s}_i + D_{0,0}^{(n)} \tag{3.44}$$

and

$$\delta^{(n)} = [\sum_{j=1}^{n-1} D_{n,j}^{(n)'} \underset{\sim}{s}_j + D_{n,0}^{(n)}]' D_{n,n}^{(n)^{-1}} [\sum_{j=1}^{n-1} D_{n,j}^{(n)'} \underset{\sim}{s}_j + D_{n,0}^{(n)}]$$

$$= \sum_{i=1}^{n-1} \underset{\sim}{s}_i' D_{n,i}^{(n)} D_{n,n}^{(n)^{-1}} D_{n,i}^{(n)'} \underset{\sim}{s}_i + 2 \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \underset{\sim}{s}_i' D_{n,i}^{(n)} D_{n,n}^{(n)^{-1}} D_{n,j}^{(n)'} \underset{\sim}{s}_j \qquad (3.45)$$

$$+ 2 \sum_{i=1}^{n-1} [D_{n,i}^{(n)} D_{n,n}^{(n)^{-1}} D_{n,0}^{(n)}]' \underset{\sim}{s}_i + D_{n,0}^{(n)'} D_{n,n}^{(n)^{-1}} D_{n,0}^{(n)},$$

so that

$$\beta^{(n-1)} = \gamma^{(n)} - \delta^{(n)}$$

$$= \sum_{i=1}^{n-1} \underset{\sim}{s}_i' [D_{i,i}^{(n)} - D_{n,i}^{(n)} D_{n,n}^{(n)^{-1}} D_{n,i}^{(n)'}] \underset{\sim}{s}_i$$

$$+ 2 \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \underset{\sim}{s}_i' [D_{i,j}^{(n)} - D_{n,i}^{(n)} D_{n,n}^{(n)^{-1}} D_{n,j}^{(n)'}] \underset{\sim}{s}_j \qquad (3.46)$$

$$+ 2 \sum_{i=1}^{n-1} [D_{i,0}^{(n)} - D_{n,i}^{(n)'} D_{n,n}^{(n)-1} D_{n,0}^{(n)}]' \underset{\sim}{s}_i + D_{0,0}^{(n)} - D_{n,0}^{(n)'} D_{n,n}^{(n)^{-1}} D_{n,0}^{(n)}$$

$$= \sum_{i=1}^{n-1} \underset{\sim}{s}_i' D_{i,i}^{(n-1)} \underset{\sim}{s}_i + 2 \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \underset{\sim}{s}_i' D_{i,j}^{(n-1)^{-1}} \underset{\sim}{s}_j + \sum_{i=1}^{n-1} D_{i,0}^{(n-1)'} \underset{\sim}{s}_i + D_{0,0}^{(n-1)} \qquad (3.47)$$

where

$$D_{i,j}^{(n-1)} = D_{i,j}^{(n)} - D_{n,i}^{(n)'} D_{n,n}^{(n)^{-1}} D_{n,j}^{(n)} \qquad (3.48)$$

$$(i,j = 0,1,2,\ldots,n-1)$$

Hence the integral with respect to $\underset{\sim}{s}_{n-1}$ in (3.42) is of the same form as the integral with respect to $\underset{\sim}{s}_n$, so that by using the recurrence relation (3.48) repeatedly we find

$$\int_{\underset{\sim}{s}_1} \cdots \int_{\underset{\sim}{s}_n} \alpha\, e^{-\frac{1}{2}\beta^{(n)}} = \alpha (2\pi)^{np/2} \prod_{i=1}^{n} |D_{i,i}^{(i)}|^{-\frac{1}{2}} e^{-\frac{1}{2} D_{0,0}^{(0)}}. \qquad (3.49)$$

Substitution of α in (3.35) concludes the proof.

With regard to the use of the above result, it should be noted that the n p-vector integrations can be carried out in any order. However it is computationally advantageous to choose the order so that at each step as many as possible of the $D_{i,j}$ are zero. The best order for a simple pedigree seems to be: unmarried individuals first, followed by spouse-pairs, within each spouse-pair eliminating the marry-in first. This order implies $D_{i,j} = \underset{\sim}{0}$ unless i,j refer to parent-child, or spouses, or i = j, or i or j = 0. The computations may then be simplified as follows. Each unmarried individual has the same $D^{(n)}_{i,i}$ given by

$$D^{(n)}_{i,i} = 2V_g^{-1} + V_e^{-1} \tag{3.50}$$

and since the elimination of such an individual does not affect $D_{i,i}$ for any other unmarried, all such individuals together contribute to the likelihood the factor

$$|2V_g^{-1} + V_e^{-1}|^{-\frac{1}{2}n_0} \tag{3.51}$$

where n_0 is the number of observed unmarrieds in the pedigree. Elimination of unmarrieds results in the following;

$$D^{(n)}_{0,0} = \sum_{i=1}^{n} (\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)$$

$$D^{(n-1)}_{0,0} = D^{(n)}_{0,0} - D^{(n)\,'}_{n,0}D^{(n)^{-1}}_{n,n}D^{(n)}_{n,0}$$

$$= D^{(n)}_{0,0} - (\underset{\sim}{z}_n-\underset{\sim}{\mu}_n)'V_e^{-1}(2V_g^{-1}+V_e^{-1})^{-1}V_e^{-1}(\underset{\sim}{z}_n-\underset{\sim}{\mu}_n)$$

$$D^{(n-2)}_{0,0} = D^{(n-1)}_{0,0} - D^{(n-1)\,'}_{n-1,0}D^{(n-1)^{-1}}_{n-1,n-1}D^{(n-1)}_{n-1,0}$$

$$= D^{(n)}_{0,0} - (\underset{\sim}{z}_n-\underset{\sim}{\mu}_n)'V_e^{-1}(2V_g^{-1}+V_e^{-1})^{-1}V_e^{-1}(\underset{\sim}{z}_n-\underset{\sim}{\mu}_n)$$

$$- (\underset{\sim}{z}_{n-1}-\underset{\sim}{\mu}_{n-1})'V_e^{-1}(2V_g^{-1}+V_e^{-1})^{-1}V_e^{-1}(\underset{\sim}{z}_{n-1}-\underset{\sim}{\mu}_{n-1})$$

- - - - - - - -

Thus, if there are n_0 unmarrieds,

$$\begin{aligned} D^{(n-n_0)}_{0,0} &= D^{(n)}_{0,0} - \sum_{i=n-n_0+1}^{n} (\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)'V_e^{-1}(2V_g^{-1}+V_e^{-1})^{-1}V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i) \\ &= \sum_{i=1}^{n} (\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)'V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i) - \sum_{i=n-n_0+1}^{n} (\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)'V_e^{-1}(2V_g^{-1}+V_e^{-1})^{-1}V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i) \end{aligned} \tag{3.52}$$

Within a given spouse-pair let the individual whose parental line is indicated be i and the marry-in be j = i+1. Then to eliminate j we note that

$$\begin{aligned} D^{(j)}_{j,j} &= D^{(j+1)}_{j,j} - D^{(j+1)'}_{j,j+1}D^{(j+1)^{-1}}_{j+1,j+1}D^{(j+1)}_{j,j+1} \\ &= D^{(j+2)}_{j,j} - D^{(j+2)'}_{j,j+2}D^{(j+2)^{-1}}_{j+2,j+2}D^{(j+2)}_{j,j+2} - D^{(j+1)'}_{j,j+1}D^{(j+1)^{-1}}_{j+1,j+1}D^{(j+1)}_{j,j+1} \\ &\text{- - - - - - - -} \\ &= D^{(n)}_{j,j} - \sum_{r=1}^{n-j} D^{(j+r)'}_{j,j+r}D^{(j+r)^{-1}}_{j+r,j+r}D^{(j+r)}_{j,j+r} \\ &= a_jV_e^{-1} + (1+c_j)V_g^{-1} - \sum_{r=1}^{n-j} D^{(j+r)'}_{j,j+r}D^{(j+r)^{-1}}_{j+r,j+r}D^{(j+r)}_{j,j+r} \quad . \end{aligned}$$

From the definition of the D's we see that $D^{(j+r)}_{j,j+r} = \underset{\sim}{0}$ unless j+r is a child of j. Thus

$$D^{(j)}_{j,j} = a_jV_e^{-1} + (1+c_j)V_g^{-1} - \sum_{\ell} V_g^{-1}D^{(\ell)^{-1}}_{\ell,\ell}V_g^{-1} \tag{3.53}$$

where the summation is over all children of i and j including those eliminated in the first step, i.e., among the unmarrieds.

Similarly,

$$\begin{aligned} D^{(j)}_{j,0} &= D^{(n)}_{j,0} - \sum_{r=1}^{n-j} D^{(j+r)'}_{j,j+r}D^{(j+r)^{-1}}_{j+r,j+r}D^{(j+r)}_{j+r,0} \\ &= -V_e^{-1}(\underset{\sim}{z}_j-\underset{\sim}{\mu}_j) - \sum_{\ell} (-V_g^{-1})D^{(\ell)^{-1}}_{\ell,\ell}D^{(\ell)}_{\ell,0} \end{aligned} \tag{3.54}$$

After elimination of j = i+1, the marry-in, we now consider the elimination of i, the one "in" the pedigree. The relevant D's are computed as follows:

$$\begin{aligned} D^{(i)}_{i,i} &= D^{(i+1)}_{i,i} - D^{(i+1)'}_{i,i+1} D^{(i+1)^{-1}}_{i+1,i+1} D^{(i+1)}_{i,i+1} \\ &= D^{(j)}_{i,i} - D^{(j)'}_{i,j} D^{(j)^{-1}}_{j,j} D^{(j)}_{i,j}, \text{ (since } j = i+1) \\ &= a_i V_e^{-1} + (1+c_i) V_g^{-1} - \sum_{\ell} V_g^{-1} D^{(\ell)^{-1}}_{\ell,\ell} V_g^{-1} - D^{(j)'}_{i,j} D^{(j)^{-1}}_{j,j} D^{(j)}_{i,j} \end{aligned} \tag{3.55}$$

using (3.53).

Similarly,

$$D^{(i)}_{i,0} = -V_e^{-1}(\underset{\sim}{z}_i - \underset{\sim}{\mu}_i) - \sum_{\ell} (-V_g^{-1}) D^{(\ell)^{-1}}_{\ell,\ell} D^{(\ell)}_{\ell,0} - D^{(j)'}_{i,j} D^{(j)^{-1}}_{j,j} D^{(j)}_{i,0} \tag{3.56}$$

and

$$D^{(i)}_{0,0} = D^{(j)}_{0,0} - D^{(j)'}_{j,0} D^{(j)^{-1}}_{j,j} D^{(j)}_{j,0} . \tag{3.57}$$

After elimination of i we have

$$\begin{aligned} D^{(i-1)}_{0,0} &= D^{(i)}_{0,0} - D^{(i)'}_{i,0} D^{(i)^{-1}}_{i,i} D^{(i)}_{i,0} \\ &= D^{(j)}_{0,0} - D^{(j)'}_{j,0} D^{(j)^{-1}}_{j,j} D^{(j)}_{j,0} - D^{(i)'}_{i,0} D^{(i)^{-1}}_{i,i} D^{(i)}_{i,0} \end{aligned} \tag{3.58}$$

It therefore follows that for each spouse-pair we only need to keep track of $\sum_{\ell} V_g^{-1} D^{(\ell)^{-1}}_{\ell,\ell} V_g^{-1}$ and $\sum_{\ell} (-V_g^{-1}) D^{(\ell)^{-1}}_{\ell,\ell} D^{(\ell)}_{\ell,0}$ over all the children ℓ of i and j.

The univariate version of the remarkable results derived in this section were first worked out in 1972 in slightly different form by Dr. Philip Green, III (unpublished), who used them to develop the computer program PLYGEN for calculating the likelihood of a simple pedigree under polygenic inheritance. However, the explanation given here for the $D_{i,j}^{(n)}$ is my own. It is regrettable that he did not publish the results. In recognition of his original work I have chosen to call formula (3.29) Green's recurrence relation.

CHAPTER IV

COMBINED OLIGOGENIC AND POLYGENIC INHERITANCE IN SIMPLE PEDIGREES

4.1 Introduction

The model for combined oligogenic and polygenic inheritance incorporates a few major (distinguishable) loci and an infinite number of separately indistinguishable loci. For this reason this mode of inheritance is often called the mixed model. Conceptually we are regarding the entire genome as contributing to the variability in the trait under study although only the effects of a few loci may be distinguishable. This makes better sense biologically than either the oligogenic model or the purely polygenic model, both of which are easily seen to be special limiting cases of the mixed model.

The mixed model was originally proposed by Elston and Stewart (1971). For nuclear families, Morton and Maclean (1974) described a version that included a common within sibship environmental effect. They calculated the likelihood conditional on the parent's phenotypes. Ott (1979) extended their formulation to pedigrees, but used an unconditional likelihood. Boyle and Elston (1979) have indicated how allowance could be made for other random effects including other environmental effects and effects due to assortative mating.

In spite of the superiority of the model and the extensions mentioned above, its use is seriously limited by an unwieldy computational problem. Present computing facilities can only handle 10 or fewer individuals and even that at a cost of \$300 (Ott, 1979). For this reason approximations to the likelihood function have been considered by Lalouel et al. (1981) and Graepel (1981). These are very recent and how well they approximate the likelihood has not been fully demonstrated. Furthermore we wish to deal with the multivariate version.

The computational problem is not insurmountable. We shall show that the 'problem' is largely the result of the particular representation of the likelihood function used. We shall then use the results on polygenic inheritance, presented in the last chapter, to derive recurrence formulas for calculating the exact likelihood. This circumvents the 'peculiar' computational problem, and altogether obviates the use of approximations to the likelihood function.

4.2 The Likelihood Function

To write down the likelihood of multivariate data on a simple pedigree we first recall the specification of the components of the mixed model defined in Chapter II. On the assumption that the major genotype and polygenotype components of the mixed genotype are independent in the population and are transmitted independently, the population distribution of the mixed genotype is

$$\psi_{tG} = \psi_t \phi(\underset{\sim}{s}_G, V_g) \qquad (4.1)$$

and the distribution of offspring mixed genotype given that of the parents is

$$p_{uFvGwH} = p_{uvw} p_{FGH} = p_{uvw} \phi(\underset{\sim}{s}_H - \tfrac{1}{2}[\underset{\sim}{s}_F + \underset{\sim}{s}_G], \tfrac{1}{2}V_g) \qquad (4.2)$$

assuming no consanguineous mating.

It then follows from (3.1) and (3.7) that

$$\left\{\begin{aligned} \Gamma_0 &= \sum_{u_0} \psi_{u_0} \int_{\mathbf{s}_0} \phi(\mathbf{s}_0, V_g) g_{u_0\mathbf{s}_0}(\mathbf{x}_{i_0}) \sum_{v_0} \psi_{v_0} \int_{\mathbf{t}_0} \phi(\mathbf{t}_0, V_g) g_{v_0\mathbf{t}_0}(\mathbf{y}_{i_0}) \\ \Gamma_j &= \prod_{i_j} \sum_{u_j} p_{u_{j-1}v_{j-1}u_j} \int_{\mathbf{s}_j} \phi(\mathbf{s}_j - \tfrac{1}{2}[\mathbf{s}_{j-1} + \mathbf{t}_{j-1}], \tfrac{1}{2}V_g) g_{u_j\mathbf{s}_j}(\mathbf{x}_{i_0 i_1 \ldots i_j}) \\ &\qquad \times \sum_{v_j} \psi_{v_j} \int_{\mathbf{t}_j} \phi(\mathbf{t}_j, V_g) g_{v_j\mathbf{t}_j}(\mathbf{y}_{i_0 i_1 \ldots i_j}); \quad j \geq 1 \end{aligned}\right. \tag{4.3}$$

where $g_{u_j\mathbf{s}_j}(\mathbf{x}_{i_0 i_1 \ldots i_j})$ is the probability distribution or the density function of $\mathbf{x}_{i_0 i_1 \ldots i_j}$ conditional on u_j-th oligogenotype and the value $\mathbf{s}_j$ of the random variable representing the polygenotype. The likelihood of the entire pedigree is then

$$L_M = \Gamma_0(\Gamma_1(\Gamma_2 \ldots)) \tag{4.4}$$

as before.

When it is reasonable to assume multinormal phenotype we can set

$$g_{u_j\mathbf{s}_j}(\mathbf{z}) = \phi(\mathbf{s}_j + \boldsymbol{\mu}_{u_j} - \mathbf{z}, V_e) \tag{4.5}$$

where $\boldsymbol{\mu}_{u_j}$ is the vector mean of u_j-th oligogenotype.

The distribution of the trait in the population is then found from (3.24) to be

$$\sum_u \psi_u \int_{\mathbf{s}} \phi(\mathbf{s}, V_g)\ \phi(\mathbf{s} + \boldsymbol{\mu}_u - \mathbf{z}, V_e) = \sum_u \psi_u \phi(\boldsymbol{\mu}_u - \mathbf{z}, V_g + V_e) \tag{4.6}$$

which is a mixture of multinormals where the covariance matrix within each distribution is $V_g + V_e$.

4.3 The 'Peculiar' Computational Problem

We now discuss the problem of computing the likelihood function for multinormal phenotypes. If we assume individuals in the simple pedigree are numbered so that children have higher numbers than their parents, then as in (3.3)

$$L_M = \sum_{i=1}^{n} \prod_{u_i} p(u_i)\int_{\underset{\sim}{s}_i} f(\underset{\sim}{s}_i)\ \phi(\underset{\sim}{s}_i+\underset{\sim}{\mu}_i-\underset{\sim}{z}_i,V_e) \qquad (4.7)$$

where

$$p(u_i) = \begin{cases} \psi_{u_i} & \text{if } i \varepsilon M \\ p_{stu} & \text{if } i \varepsilon I \end{cases}$$

and

$$f(\underset{\sim}{s}_i) = \begin{cases} \phi(s_i,V_g), & \text{if } i \varepsilon M \\ \phi(\underset{\sim}{s}_i-\tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)}+\underset{\sim}{s}_{p_2(i)}],\ \tfrac{1}{2}V_g), & \text{if } i \varepsilon I \end{cases}$$

$$= \sum_{\substack{\text{all n-tuples}\\(u_1,u_2,\ldots,u_n)}} \prod_{i=1}^{n} p(u_i)\int_{\underset{\sim}{s}_1}\int_{\underset{\sim}{s}_2}\cdots\int_{\underset{\sim}{s}_n} \prod_{i=1} f(\underset{\sim}{s}_i)\phi(\underset{\sim}{s}_i+\underset{\sim}{\mu}_{u_i}-\underset{\sim}{z}_i,V_e) \qquad (4.8)$$

Since for all i, $u_i = 1,2,\ldots,k$, there are k^n different n-tuples. The n-tuples here correspond to Ott's (1979) genotype vectors. The integrals appearing in (4.8) represent the likelihood L_p under a polygenic model (3.34) except that the mean $\underset{\sim}{\mu}_i$ is now replaced by $\underset{\sim}{\mu}_{u_i}$, the mean of the oligogenotype u_i. Hence (4.8) can be rewritten as

$$L_M = \sum_{\substack{\text{all n-tuples}\\(u_1,u_2,\ldots u_n)}} \prod_{i=1}^{n} p(u_i)L_p((\underset{\sim}{\mu}_{u_1},\underset{\sim}{\mu}_{u_2},\ldots,\underset{\sim}{\mu}_{u_n}),V_g,V_e) \qquad (4.9)$$

Ott (1979) used the Lange (1978) formulation for a polygenic model and so avoided the integrals in (4.8). But the result was an L_p involving the inverse of an nxn matrix Ω. For a p-variate trait the dimension of Ω is np x np! Of course, Ott did not consider multivariate traits, but in his algorithm all the k^n possible quantities to be summed in (4.9) are stored in computer memory. For a single major locus with two alleles, there are three major genotypes, i.e., $k = 3$, and we need to compute and store the 3^n terms, each involving the inverse of an np x np matrix. Tuples or genotype vectors with zero probabilities of course need not be considered, but even for nuclear families the elimination of these leaves $4 + 2^n + 3^{n-2}$ tuples with positive probability. This can still be too many for large families. We note however that the problem of having to compute a sum of k^n terms is not peculiar to the mixed model. The oligogenic models share the same problem. The likelihood of say a multinormal trait under an oligogenic model, from (3.6) is clearly

$$L_0 = \sum_{\substack{\text{all n-tuples} \\ (u_1,u_2,\ldots,u_n)}} \prod_{i=1}^{n} p_i(u_i)\ \phi(\underset{\sim}{\mu}_{u_i} - \underset{\sim}{z}_i, V_e). \tag{4.10}$$

This can be computed easily and quickly,even for multivariate traits, in large pedigrees by the program GENPED (Kaplan, Unpublished), which uses the Elston and Stewart (1971) recursive formulation or equivalently (3.3), i.e.,

$$L_0 = \prod_{i=1}^{n} \sum_{u_i=1}^{k} p(u_i)\ \phi(\underset{\sim}{\mu}_{u_i} - \underset{\sim}{z}_i, V_e)\ . \tag{4.11}$$

Hence the problem is not really with the number of terms to be summed per se, but the fact that the k^n possible n-tuples each involving the inverse of a large matrix are stored in computer memory, and we soon exceed our storage capabilities.

4.4 An Exact Recursive Algorithm

We now derive formulas that can be used to calculate the exact likelihood recursively. The essence of the method is to use the formula for the polygenic model, Theorem 3.2, to evaluate analytically the integrals occurring in (4.8), and then to rewrite the resulting function in the form of (4.11). The result is summarized in the following theorem.

Theorem 4.1

The likelihood of a simple pedigree under the mixed model with multinormal phenotype is

$$L_M(\underset{\sim}{\mu}_1,\ldots,\underset{\sim}{\mu}_k,V_g,V_e) = 2^{n_I}(2\pi)^{-\frac{1}{2}np}|V_g|^{-n/2}|V_e|^{-n/2} \times \prod_{i=1}^{n}\{|D^{(i)}_{i,i}|^{-\frac{1}{2}}\sum_{u_i=1}^{k} p(u_i)e^{-\frac{1}{2}w(u_i)}\} \qquad (4.12)$$

where

$$w(u_i) = (\underset{\sim}{z}_i-\underset{\sim}{\mu}_{u_i})'V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_{u_i}) - D^{(i)}_{i,0}D^{(i)^{-1}}_{i,i}D^{(i)}_{i,0}\,, \qquad (4.13)$$

$p(u_i)$ are defined by (4.7) and the D's are defined by Theorem 3.2 with $\underset{\sim}{\mu}_{u_i}$ replacing $\underset{\sim}{\mu}_i$.

Proof

Using Theorem 3.2, (4.8) becomes

$$L_M = \sum_{\substack{\text{all n-tuples}\\(u_1,u_2,\ldots,u_n)}} \prod_{i=1}^{n} p(u_i)\int_{\underset{\sim}{s}_1}\int_{\underset{\sim}{s}_2}\cdots\int_{\underset{\sim}{s}_n} \prod_{i\varepsilon I}\phi(\underset{\sim}{s}_i-\tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)}+\underset{\sim}{s}_{p_2(i)}],\tfrac{1}{2}V_g)$$

$$\times \prod_{i\varepsilon M}\phi(\underset{\sim}{s}_i,V_g)\prod_{i\varepsilon T}\phi(\underset{\sim}{s}_i+\underset{\sim}{\mu}_{u_i}-\underset{\sim}{z}_i,V_e) \qquad (4.14)$$

$$= \sum_{\substack{\text{all n-tuples}\\(u_1,u_2,\ldots,u_n)}} \prod_{i=1}^{n} p(u_i)\cdot 2^{n_I}(2\pi)^{-np/2}|V_g|^{-n/2}|V_e|^{-n/2}\prod_{i=1}^{n}|D^{(i)}_{i,i}|^{-\frac{1}{2}}e^{-\frac{1}{2}D^{(0)}_{0,0}}$$

(provided $\underset{\sim}{\mu}_{u_i}$ replaces $\underset{\sim}{\mu}_i$ in the $D_{i,0}$'s and $D_{0,0}$'s) (4.15)

$$= 2^{n_I}(2\pi)^{-np/2}|V_g|^{-n/2}|V_e|^{-n/2}\prod_{i=1}^{n}|D^{(i)}_{i,i}|^{-\frac{1}{2}} \sum_{\substack{\text{all n-tuples}\\(u_1,u_2,\ldots,u_n)}} \prod_{i=1}^{n} p(u_i)e^{-\frac{1}{2}D^{(0)}_{0,0}}. \qquad (4.16)$$

Note that the recurrence relations (3.29) imply

$$\begin{aligned}
D^{(0)}_{0,0} &= D^{(1)}_{0,0} - D^{(1)'}_{1,0} D^{(1)^{-1}}_{1,1} D^{(1)}_{1,0} \\
&= D^{(2)}_{0,0} - D^{(2)'}_{2,0} D^{(2)-1}_{2,2} D^{(2)}_{2,0} - D^{(1)'}_{1,0} D^{(1)^{-1}}_{1,1} D^{(1)}_{1,0} \\
&\quad ---- \qquad\qquad (4.17) \\
&= D^{(n)}_{0,0} - \sum_{i=1}^{n} D^{(i)'}_{i,0} D^{(i)^{-1}}_{i,i} D^{(i)}_{i,0} \\
&= \sum_{i=1}^{n} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_{u_i})' V_e^{-1} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_{u_i}) - \sum_{i=1}^{n} D^{(i)'}_{i,0} D^{(i)^{-1}}_{i,i} D^{(i)}_{i,0} \\
&= \sum_{i=1}^{n} w(u_i),
\end{aligned}$$

so that (4.16) becomes

$$2^{n_I}(2\pi)^{-np/2}|V_g|^{-n/2}|V_e|^{-n/2}\prod_{i=1}^{n}|D^{(i)}_{i,i}|^{-\frac{1}{2}} \sum_{\substack{\text{all n-tuples}\\(u_1,u_2,\ldots,u_n)}} \prod_{i=1}^{n} p(u_i)e^{-\frac{1}{2}w(u_i)} \qquad (4.18)$$

$$= 2^{n_I}(2\pi)^{-np/2}|V_g|^{-n/2}|V_e|^{-n/2}\prod_{i=1}^{n}|D^{(i)}_{i,i}|^{-\frac{1}{2}} \prod_{i=1}^{n} \sum_{u_i=1}^{k} p(u_i)e^{-\frac{1}{2}w(u_i)}$$

and the proof is complete.

4.5 Some Remarks

In the formula presented in Theorem 4.1 the form of $w(u_i)$ is interesting. The first term is obviously related to the oligogenic component. The second term is therefore the difference that the polygenic component of the genotype makes to the likelihood of an individual. It is also easy to see, by comparing (4.12) and (4.11), that to compute the likelihood of the mixed model, we only need to replace the g-function for individual i, in the oligogenic model, by

$$|D_{i,i}|^{-\frac{1}{2}} e^{-\frac{1}{2}w(u_i)} \tag{4.19}$$

and use

$$|V_g|^{-n/2} |V_e|^{-n/2} \tag{4.20}$$

as a scaling factor to the whole likelihood. Furthermore, by assuming the order of integration in (4.14) follows the "best" order for polygenic models (see page 16 and following), we can use (3.50), (3.53), (3.54), (3.55) and (3.57) to calculate the D's.

CHAPTER V

COMPLEX PEDIGREES

5.1 Introduction

Lange and Elston (1975) have presented algorithms for calculating the likelihood of a set of univariate data over pedigrees of arbitrary structure. The underlying genetic models were, however, restricted to oligogenic inheritance, i.e., a finite set of loci. In this chapter we shall generalize their algorithms to polygenic and mixed models of inheritance. By analytically evaluating the integrals, we obtain simple recurrence formulas. Our results are presented for a general p-variate trait $\underset{\sim}{z}$.

5.2 Complex Pedigree With No Loops

According to our classification of pedigrees in section 3.2, complex pedigrees have at least one spouse-pair or spouse-connected set for which two or more parental lines are indicated. There may or may not be loops in the lines of descent in a complex pedigree. And loops, when they occur, may be the result of consanguineous matings or not. We shall first consider complex pedigrees with no loops (Fig. 3.2).

The method of Lange and Elston (1975) involves breaking up the complex pedigree into simple ones. This means splitting one member of every spouse-pair or spouse-connected set for which two or more parental lines are indicated, i.e., in place of one member of every such spouse-pair or spouse-connected set, create two separate but phenotypically and genotypically identical individuals. Designate one of the resulting simple pedigrees as the "root" pedigree. Then if n is the size of the root pedigree, the likelihood of the root pedigree and the adjacent "bits" of pedigree, under oligogenic inheritance, is given by Lange and Elston (1975), using our notation, by

$$L_0 = \prod_{i=1}^{n} \{\sum_{u_i} p(u_i) g_{u_i}(z_i) \prod_j \frac{L_j(u_i)}{\psi_{u_i} g_{u_i}(z_i)}\} , \qquad (5.1)$$

where $L_j(u_i)$ is the likelihood of j-th broken bit of pedigree given all individuals identical to i have genotype u_i, and

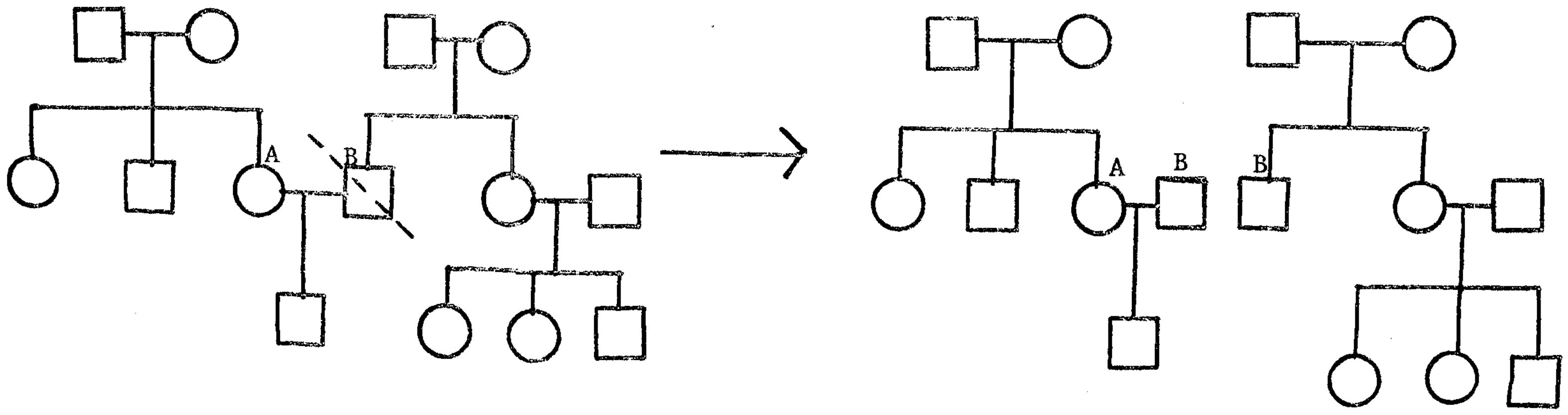
$$p\ (u_i) = \begin{cases} p_{stu} & \text{if} \quad i \varepsilon I \\ \psi_{u_i} & \text{if} \quad i \varepsilon M \end{cases} .$$

If we compare (5.1) with the likelihood of a simple pedigree (from (3.3)),

$$L_0 = \prod_{i=1}^{n} \sum_{u_i} p(u_i) g_{u_i}(z_i)$$

we see immediately that the process is one of calculating $\prod_j \frac{L_j(u_i)}{\psi_{u_i} g_{u_i}(z)}$ and attaching the result as a factor to the term for the individual i in the root pedigree.

For concreteness, let us consider the pedigree in Fig. 5.1. The pedigree may be broken at A or B. Let us break it at B, i.e., in place of B create two separate but genotypically and phenotypically identical individuals. We can take pedigree 1 as the root, and calculate the likelihood of pedigree 2 given B has a particular genotype u. We thus obtain $L_2(u)$, which we attach to the term for B in the root pedigree. Note here that for this case (5.1) is equivalent to



Complex pedigree

Simple pedigree 1

Simple pedigree 2

The complex pedigree is broken up at B into two simple pedigrees.

Fig. 5.1 Breaking up a complex pedigree

$$L_0 = \sum_u \frac{L_1(u)L_2(u)}{\psi_u g_u(x)} \tag{5.2}$$

where x is the phenotype of B. We extend this to a multivariate polygenic model simply by writing

$$L_p = \int_{\underset{\sim}{s}} \frac{L_1(\underset{\sim}{s})\cdot L_2(\underset{\sim}{s})}{\phi(\underset{\sim}{s},V_g)g_{\underset{\sim}{s}}(\underset{\sim}{x})} \tag{5.3}$$

where $\underset{\sim}{s}$ is the vector effect of the polygenotype of B and $\underset{\sim}{x}$ is the corresponding vector phenotype. We shall now evaluate (5.3) for the case where it is reasonable to assume multinormal phenotype, i.e., $g_{\underset{\sim}{s}}(\underset{\sim}{x})=\phi(\underset{\sim}{s}+\underset{\sim}{\mu}-\underset{\sim}{x},V_e)$. We first calculate $L_1(\underset{\sim}{s})$ and $L_2(\underset{\sim}{s})$ by the elimination procedure outlined for polygenic inheritance over simple pedigrees in section 3.4.2. This is easily done if we "eliminate" all individuals except B, separately for each of the two simple pedigrees. Then

$$\begin{aligned} L_1(\underset{\sim}{s}) &= K_1|V_g|^{-\frac{n_1}{2}}|V_e|^{-\frac{n_1}{2}}\prod_{i=1}^{n_1-1}|D^{(i)}_{1,i,i}|^{-\frac{1}{2}}e^{-\frac{1}{2}\beta_1^{(1)}(\underset{\sim}{s})} \\ L_2(\underset{\sim}{s}) &= K_2|V_g|^{-\frac{n_2}{2}}|V_e|^{-\frac{n_2}{2}}\prod_{i=1}^{n_2-1}|D^{(i)}_{2,i,i}|^{-\frac{1}{2}}e^{-\frac{1}{2}\beta_2^{(1)}(\underset{\sim}{s})} \end{aligned} \tag{5.4}$$

where the first subscript (1 or 2) denotes simple pedigree 1 or 2, K_1 and K_2 are constants independent of parameters,

$$\begin{aligned} \beta_1^{(1)}(\underset{\sim}{s}) &= \underset{\sim}{s}'D^{(1)}_{1,1,1}\underset{\sim}{s} + 2D^{(1)'}_{1,1,0}\underset{\sim}{s} + D^{(1)}_{1,0,0} \\ \beta_2^{(1)}(\underset{\sim}{s}) &= \underset{\sim}{s}'D^{(1)}_{2,1,1}\underset{\sim}{s} + 2D^{(1)'}_{2,1,0}\underset{\sim}{s} + D^{(1)}_{2,0,0} \end{aligned} \quad . \tag{5.5}$$

Substituting in (5.3) we obtain

$$\begin{aligned} L_p = K(|V_e||V_g|)^{-\frac{1}{2}(n_1+n_2)}\cdot|V_g|^{\frac{1}{2}}|V_e|^{\frac{1}{2}}\prod_{i=1}^{n_1-1}|D^{(i)}_{1,i,i}|^{-\frac{1}{2}}\prod_{i=1}^{n_2-1}|D^{(i)}_{2,i,i}|^{-\frac{1}{2}} \\ \times \int_{\underset{\sim}{s}} e^{-\frac{1}{2}\{\beta_1^{(1)}(\underset{\sim}{s})+\beta_2^{(1)}(\underset{\sim}{s})-\underset{\sim}{s}'V_g^{-1}\underset{\sim}{s}-(\underset{\sim}{s}+\underset{\sim}{\mu}-\underset{\sim}{x})'V_e^{-1}(\underset{\sim}{s}+\underset{\sim}{\mu}-\underset{\sim}{x})\}} \end{aligned}$$

The exponent in the integrand equals

$$\{\ \} = \underset{\sim}{s}'[D^{(1)}_{1,1,1}+D^{(1)}_{2,1,1}-V_g^{-1}-V_e^{-1}]\underset{\sim}{s} + 2[D^{(1)}_{1,1,0}+D^{(1)}_{2,1,0}+V_e^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})]'\underset{\sim}{s}$$

$$+ D^{(1)}_{1,0,0} + D^{(1)}_{2,0,0} - (\underset{\sim}{x}-\underset{\sim}{\mu})'V_e^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})$$

$$= \underset{\sim}{s}'D^{(1)}_{1,1}\underset{\sim}{s} + 2D^{(1)'}_{1,0}\underset{\sim}{s} + D^{(1)}_{0,0}, \text{ say.}$$

Using the results of section 3.4.2 we obtain after integration

$$L_p = K(|V_g||V_e|)^{-\frac{N}{2}}|D^{(1)}_{1,1}|^{-\frac{1}{2}} \prod_{i=1}^{n_1-1} |D^{(i)}_{1,i,i}|^{-\frac{1}{2}} \prod_{i=1}^{n_2-1} |D^{(i)}_{2,i,i}|^{-\frac{1}{2}} e^{-\frac{1}{2}D^{(0)}_{0,0}} \tag{5.6}$$

where K is a constant independent of parameters,

$N = n_1 + n_2 - 1$ is the total size of complex pedigree;

$$\begin{aligned} D^{(1)}_{1,1} &= D^{(1)}_{1,1,1} + D^{(1)}_{2,1,1} - V_g^{-1} - V_e^{-1} \\ D^{(1)}_{1,0} &= D^{(1)}_{1,1,0} + D^{(1)}_{2,1,0} + V_e^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}) \\ D^{(1)}_{0,0} &= D^{(1)}_{1,0,0} + D^{(1)}_{2,0,0} - (\underset{\sim}{x}-\underset{\sim}{\mu})'V_e^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}) \\ D^{(0)}_{0,0} &= D^{(1)}_{0,0} - D^{(1)'}_{1,0} D^{(1)^{-1}}_{1,1} D^{(1)}_{1,0} \quad . \end{aligned} \tag{5.7}$$

As a first step toward generalization, suppose B had m_B spouses whose parental lines were indicated, then breaking the pedigree at B results in m_B simple pedigrees each containing B as a marry-in and one simple pedigree in which B is an offspring. Each pedigree containing B as a marry-in contributes to the likelihood a factor similar to that of pedigree 1 above. In each pedigree 'eliminate' all individuals except the one identical to B.

Then analogous to the case above

$$D^{(1)}_{1,1} = \sum_j D^{(1)}_{j,1,1} - m_B(V_g^{-1}+V_e^{-1})$$

$$D^{(1)}_{1,0} = \sum_j D^{(1)}_{j,1,0} + m_B V_e^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}) \tag{5.8}$$

$$D^{(1)}_{0,0} = \sum_j D^{(1)}_{j,0,0} - m_B(\underset{\sim}{x}-\underset{\sim}{\mu})'V_e^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})$$

where the summation $\sum_j$ is over all simple pedigrees containing B.

Generalization of the above procedure to more complex pedigrees without loops is not difficult. Break up the pedigree as above, select one of the resulting simple pedigrees as root. For an individual in the root pedigree which was broken compute the likelihood given the polygenotype of this individual as above and attach the likelihood as a factor, the procedure being repeated until only the root pedigree remains. The equivalent of (5.1) for polygenic models is then

$$L_p = \prod_{i=1}^{n} \{\int_{\underset{\sim}{s}_i} f(\underset{\sim}{s}_i) g_{\underset{\sim}{s}_i}(\underset{\sim}{z}_i) \prod_{n=1}^{m_i} \frac{L_j(\underset{\sim}{s}_i)}{\phi(\underset{\sim}{s}_i,V_g) g_{\underset{\sim}{s}_i}(\underset{\sim}{z}_i)}\} \tag{5.9}$$

where
$$f(s_i) = \begin{cases} \phi(\underset{\sim}{s}_i,V_g) & \text{if } i \in M \\ \phi(\underset{\sim}{s}_i-\tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)}+\underset{\sim}{s}_{p_2(i)}],\tfrac{1}{2}V_g) & \text{if } i \in I \end{cases}$$

The root pedigree is simple, so the likelihood of the whole complex pedigree can be computed from the recurrence formulas in Theorem 3.2 provided the D's are redefined as follows

$$\begin{aligned}
D^{(n)}_{i,i} &= a_i V_e^{-1} + (b_i+c_i)V_g^{-1} + \sum_{r\varepsilon B_i} D^{(1)}_{r,1,1} - m_i(V_g^{-1}+V_e^{-1}) \\
D^{(n)}_{i,j} &= (c_{ij}-d_{ij})V_g^{-1} \\
D^{(n)}_{i,0} &= -V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i) + \sum_{r\varepsilon B_i} D^{(1)}_{r,i,0} + m_i V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i) \qquad (5.10) \\
D^{(n)}_{0,0} &= \sum_{i\varepsilon T} (\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)'V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i) + \sum_{r\varepsilon B_i} D^{(1)}_{r,0,0} \\
&\quad - \sum_{i\varepsilon B} m_i(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)'V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i)
\end{aligned}$$

where B_i denotes simple pedigrees, other than the root, that result from breaking the complex pedigree at individual i in the root pedigree, m_i denotes the number of such pedigrees, and B denotes all individuals broken in the root pedigree. The modifications to the D's are obvious extensions of (5.8). Note the $D^{(n)}_{i,j}$ remains the same as in Theorem (3.2) since it is the matrix of the bilinear form in $\underset{\sim}{s}_i$ and $\underset{\sim}{s}_j$, and i and j refer only to individuals in the root pedigree. With these modifications, we find that the likelihood becomes

$$L_p = K|V_g|^{-\frac{N}{2}}|V_e|^{-\frac{N}{2}} \prod_{i=1}^{n} \left(|D^{(i)}_{i,i}|^{-\frac{1}{2}} \prod_{r\varepsilon B_i} \prod_{\ell} |D^{(\ell)}_{r,\ell,\ell}|^{-\frac{1}{2}}\right) e^{-\frac{1}{2}D^{(0)}_{0,0}} \qquad (5.11)$$

where N is the number of all individuals in the complex pedigree, and $D^{(\ell)}_{r,\ell,\ell}$ results from the elimination of the ℓ-th individual in the r-th simple pedigree connected to the individual i in the root pedigree.

Formulas (5.1) and (5.9) are strictly speaking meant for the case where a complex pedigree without loops can be broken as indicated into a simple "root" pedigree and simple "adjacent" pedigrees. However they can be used for the case where some of the adjacent pedigrees are complex pedigrees without loops; we only need to calculate the $L_j(u_i)$ and $L_j(\underset{\sim}{s}_i)$ for such adjacent pedigrees as for complex pedigrees.

5.3 Combined Oligogenic and Polygenic Inheritance in Complex Pedigrees With No Loops

By combining (5.1) and (5.9), the likelihood of a complex pedigree with no loops, under multifactorial inheritance is easily seen to be

$$L_M = \prod_{i=1}^{n} \{\sum_{u_i} p(u_i) \int_{\underset{\sim}{s}_i} f(\underset{\sim}{s}_i) g_{u_i \underset{\sim}{s}_i}(\underset{\sim}{z}_i) \prod_{j=1}^{m_i} \frac{L_j(u_i,\underset{\sim}{s}_i)}{\psi_{u_i} \phi(\underset{\sim}{s}_i, V_g) g_{u_i \underset{\sim}{s}_i}(\underset{\sim}{z}_i)}\} \qquad (5.12)$$

where $L_j(u_i,\underset{\sim}{s}_i)$ is the likelihood of the j-th broken bit of pedigree given all individuals identical to i have major genotype u_i and polygenotype value $\underset{\sim}{s}_i$,

$$p(u_i) = \begin{cases} p_{stu} & \text{if } i \in I, \\ \psi_{u_i} & \text{if } i \in M, \end{cases}$$

$$f(\underset{\sim}{s}_i) = \begin{cases} \phi(\underset{\sim}{s}_i - \tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)} + \underset{\sim}{s}_{p_2(i)}], \tfrac{1}{2}V_g), & i \in I, \\ \phi(\underset{\sim}{s}_i, V_g), & i \in M, \end{cases}$$

and $g_{u\underset{\sim}{s}_i}(\underset{\sim}{z}_i) = \phi(\underset{\sim}{s}_i + \underset{\sim}{\mu}_{u_i} - \underset{\sim}{z}_i, V_e)$ for multinormal phenotypes. As expected, the integrals have the same form as for the polygenic model (5.9) and can therefore be evaluated by (5.11), with $\underset{\sim}{\mu}_{u_i}$ replacing $\underset{\sim}{\mu}_i$. After evaluating the integrals, the likelihood function reduces to the form for oligogenic inheritance (5.1), with the recurrence relations already developed. Thus (5.12) becomes

$$L_M = K(|V_g||V_e|)^{-\frac{N}{2}} \cdot \prod_{i=1}^{n} \{|D_{i,i}^{(i)}|^{-\frac{1}{2}} \prod_{r \in B_i} \prod_{\ell} |D_{r,\ell,\ell}^{(\ell)}|^{-\frac{1}{2}}$$

$$\times \sum_{u_i=1}^{k} p(u_i) e^{-\frac{1}{2}w(u_i)} \prod_{j=1}^{m_i} \frac{L_j(u_i)}{\psi_{u_i}}\} \qquad (5.13)$$

where $w(u_i) = (\underset{\sim}{z}_i - \underset{\sim}{\mu}_{u_i})' V_e^{-1} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_{u_i}) - D_{i,0}^{(i)'} D_{i,i}^{(i)^{-1}} D_{i,0}^{(i)}$, and in the calculation of $L_j(u_i)$, the g-functions are replaced by the corresponding $e^{-\frac{1}{2}w}$.

5.4 Non-Consanguineous Loops

When there are loops in the pedigree, Lange and Elston (1975) suggest breaking them as above, i.e., replacing one member of every spouse pair with both parental lines observed by two separate but phenotypically and genotypically identical individuals. Breaking all loops this way results in either a simple pedigree or a complex pedigree without loops. Suppose there are b breaks and at the j-th break n_j identical individuals are created, and let $L(v_1,v_2,\ldots,v_b)$ be the likelihood of the pedigree without loops when all n_j identical individuals resulting from the j-th break have their genotypes fixed at v_j (= 1,2,...,b). Then Lange and Elston calculate the likelihood under oligogenic inheritance by

$$L_0 = \sum_{\substack{\text{all } b \text{ tuples}\\ (v_1,v_2,\ldots,v_b)}} \frac{L(v_1,v_2,\ldots,v_b)}{\prod_j (\psi_{v_j} g_{v_j}(z))^{n_j-1}}, \qquad (5.14)$$

in our notation. We can extend this to polygenic inheritance and for multivariate traits with multinormal phenotypes simply by writing

$$L_p = \int_{\underset{\sim}{s}_1^0}\int_{\underset{\sim}{s}_2^0}\cdots\int_{\underset{\sim}{s}_b^0} \frac{L(\underset{\sim}{s}_1^0,\underset{\sim}{s}_2^0,\ldots,\underset{\sim}{s}_b^0)}{\prod_j[\phi(\underset{\sim}{s}_j,V_g)\ \phi(\underset{\sim}{s}_j+\underset{\sim}{\mu}_j-\underset{\sim}{z}_j,V_e)]^{n_j-1}} \qquad (5.15)$$

where $L(\underset{\sim}{s}_1^0,\underset{\sim}{s}_2^0,\ldots,\underset{\sim}{s}_b^0)$ is the likelihood of the broken pedigree given that the genotypic values of the b individuals broken are fixed at $\underset{\sim}{s}_1^0,\underset{\sim}{s}_2^0,\ldots,\underset{\sim}{s}_b^0$. The likelihood function (5.15) looks complicated, but it is equal to the likelihood function without loops (5.11). We can most easily explain this if we consider the case where breaking a looped pedigree of size n results in a simple pedigree of size $n' \; (= n + \sum_j n_j - b)$. From section 3.4.2 and in particular (3.36), (5.15) becomes

$$L_p = \alpha \prod_{j=1}^{b} [(2\pi)^{p/2}|V_g|^{\frac{1}{2}}|V_e|^{\frac{1}{2}}]^{n_j-1} \int_{\underset{\sim}{s}_1^0}\int_{\underset{\sim}{s}_2^0}\cdots\int_{\underset{\sim}{s}_b^0}\int_{\underset{\sim}{s}_1}\int_{\underset{\sim}{s}_2}\cdots\int_{\underset{\sim}{s}_{n''}} e^{-\frac{1}{2}\{\beta^{(n')} - \sum_{j=1}^{b} [\underset{\sim}{s}_j^{0'} V_g^{-1} \underset{\sim}{s}_j^{0}]^{n_j-1} - \sum_{j=1}^{b} [(\underset{\sim}{z}_j-\underset{\sim}{\mu}_j-\underset{\sim}{s}_j^{0})'V_e^{-1}(\underset{\sim}{z}_j-\underset{\sim}{\mu}_j-\underset{\sim}{s}_j^{0})]^{n_j-1}\}} \tag{5.16}$$

where n'' is the number of those not fixed (n''+b=n),

$$\{\underset{\sim}{s}_1,\underset{\sim}{s}_2,\ldots \underset{\sim}{s}_{n''}\} \text{ excludes } \{\underset{\sim}{s}_1^0,\underset{\sim}{s}_2^0,\ldots,\underset{\sim}{s}_b^0\},$$

$$\beta^{(n')} = 2\sum_{i\varepsilon I} (\underset{\sim}{s}_i-\tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)}+\underset{\sim}{s}_{p_2(i)}])'V_g^{-1}(\underset{\sim}{s}_i-\tfrac{1}{2}[\underset{\sim}{s}_{p_1(i)}+\underset{\sim}{s}_{p_2(i)}]) + \sum_{i\varepsilon M} \underset{\sim}{s}_i' V_g^{-1} \underset{\sim}{s}_i + \sum_{i\varepsilon T} (\underset{\sim}{x}_i-\underset{\sim}{\mu}_i-\underset{\sim}{s}_i)'V_e^{-1}(\underset{\sim}{x}_i-\underset{\sim}{\mu}_i-\underset{\sim}{s}_i) \tag{5.17}$$

and

$$\alpha = \text{constant} \times (|V_g||V_e|)^{-\frac{n'}{2}} . \tag{5.18}$$

Now note that when we break the complex pedigree at the j-th position resulting in n_j identical individuals, n_j-1 of these are artificially created marry-ins. These contribute n_j-1 terms to the second summation in (5.17) and n_j-1 terms to the third summation. But the terms contributed by the artificially created marry-ins are subtracted from $\beta^{(n')}$ in the exponent term of (5.16). Hence the exponent in (5.16) equals $\beta^{(n)}$ as originally defined in (3.36). We also note that

$$\alpha \prod_{j=1}^{b} [(2\pi)^{p/2} |V_g|^{\frac{1}{2}} |V_e|^{\frac{1}{2}}]^{n_j-1} = \text{constant} \times (|V_g||V_e|)^{-n/2} \tag{5.19}$$

There is no trouble changing the order of integrations in (5.16) to correspond to $\underset{\sim}{s}_1, \underset{\sim}{s}_2, \ldots, \underset{\sim}{s}_n$. We find therefore that the likelihood of the complex pedigree is the same as the likelihood of the simple pedigree resulting from breaking the loops, without the artificially created marry-ins. Thus we need not break up the loops at all, conceptually. We can start from any computationally convenient individual and "eliminate" until we have covered all individuals in the pedigree, using the recursive formula given in Theorem 3.2.

The observation we have made above is true for the more general case where breaking all loops results in a complex pedigree without loops. Thus formula (5.11) can be used for complex pedigrees with loops as well.

Let us now briefly consider the combined oligogenic and polygenic inheritance in complex pedigrees with non-consanguineous loops. By combining (5.14) and (5.15), we can write down the desired likelihood as

$$L_M = \sum_{\substack{\text{all b-tuples}\\(v_1,v_2,\ldots,v_b)}} \int_{\underset{\sim}{s}_1^0}\int_{\underset{\sim}{s}_2^0}\cdots\int_{s_b^0} \frac{L(\underset{\sim}{s}_1^0 v_1, \underset{\sim}{s}_1^0 v_2, \ldots, \underset{\sim}{s}_b^0 v_b)}{\prod_j [\psi_{v_j}\,\phi(\underset{\sim}{s}_j^0, V_g)\,\phi(\underset{\sim}{s}_j^0+\underset{\sim}{\mu}_{v_j}-\underset{\sim}{z}_j, V_e)]^{n_j-1}} \qquad (5.20)$$

Using (5.11) to evaluate the integrals, we have

$$L_M = K(|V_g||V_e|)^{-N/2}\prod_{i=1}^{n}\left(|D_{i,i}^{(i)}|^{-\frac{1}{2}}\prod_{r\varepsilon B_i}\prod_{\ell}|D_{r,\ell,\ell}^{(\ell)}|^{-\frac{1}{2}}\right)$$

$$\times \sum_{\substack{\text{all b-tuples}\\(v_1,v_2,\ldots v_b)}} \frac{L(v_1,v_2,\ldots,v_b)}{\prod_j(\psi_{v_j})^{n_j-1}} \qquad (5.21)$$

where in the calculation of $L(v_1,v_2,\ldots v_b)$ the g-function for individual i is replaced by $e^{-\frac{1}{2}w(u_i)}$ as in (5.13).

5.5 Consanguineous Loops

The complication introduced by consanguineous matings is that the relatedness factors α_1,α_2, and α_3 defined for polygenic models in section 2.4 are no longer all equal to ½. These are the only quantities that change, which means that in our formulas only the D's need to be modified; the form of the likelihood functions remains the same as for non-consanguineous pedigrees.

Recalling from section 2.4 that for polygenic inheritance the transmission density for individual i is

$$\phi(\underset{\sim}{s}_i - [\alpha_{1i}\underset{\sim}{s}_{p_1(i)} + \alpha_{2i}\underset{\sim}{s}_{p_2(i)}],\ \alpha_{3i}V_g),$$

the $\beta^{(n)}$ that enters into the derivations in section 3.4.2 (specifically expression (3.36) and following) becomes

$$\beta^{(n)} = \sum_{i\varepsilon I}\frac{1}{\alpha_{3i}}(\underset{\sim}{s}_i-[\alpha_{1i}\underset{\sim}{s}_{p_1(i)}+\alpha_{2i}\underset{\sim}{s}_{p_2(i)}])'V_g^{-1}(\underset{\sim}{s}_i-[\alpha_{1i}\underset{\sim}{s}_{p_1(i)}+\alpha_{2i}\underset{\sim}{s}_{p_2(i)}])$$

$$+\sum_{i\varepsilon M}\underset{\sim}{s}_i'V_g^{-1}\underset{\sim}{s}_i + \sum_{i\varepsilon T}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i-\underset{\sim}{s}_i)'V_e^{-1}(\underset{\sim}{z}_i-\underset{\sim}{\mu}_i-\underset{\sim}{s}_i)\ . \tag{5.22}$$

Expanding and rearranging so that quadratic terms in $\underset{\sim}{s}_i$, come first, followed by bilinear forms in $\underset{\sim}{s}_i$ and $\underset{\sim}{s}_j$, linear terms in $\underset{\sim}{s}_i$ and constants, we have

$$\begin{aligned}\beta^{(n)} = {} & \sum_{i\varepsilon T} \underset{\sim}{s}_i' V_e^{-1} \underset{\sim}{s}_i + \sum_{i\varepsilon I} \frac{1}{\alpha_{3i}} \underset{\sim}{s}_i' V_g^{-1} \underset{\sim}{s}_i + \sum_{i\varepsilon M} \underset{\sim}{s}_i' V_g^{-1} \underset{\sim}{s}_i \\ & + \sum_{i\varepsilon I} \frac{\alpha_{1i}^2}{\alpha_{3i}} \underset{\sim}{s}'_{p_1(i)} V_g^{-1} \underset{\sim}{s}_{p_1(i)} + \sum_{i\varepsilon I} \frac{\alpha_{2i}^2}{\alpha_{3i}} \underset{\sim}{s}'_{p_2(i)} V_g^{-1} \underset{\sim}{s}_{p_2(i)} \\ & - 2 \sum_{i\varepsilon I} \frac{\alpha_{1i}}{\alpha_{3i}} \underset{\sim}{s}'_{p_1(i)} V_g^{-1} \underset{\sim}{s}_i - 2 \sum_{i\varepsilon I} \frac{\alpha_{2i}}{\alpha_{3i}} \underset{\sim}{s}'_{p_2(i)} V_g^{-1} \underset{\sim}{s}_i \\ & + 2 \sum_{i\varepsilon I} \frac{\alpha_{1i}\alpha_{2i}}{\alpha_{3i}} \underset{\sim}{s}'_{p_1(i)} V_g^{-1} \underset{\sim}{s}_{p_2(i)} - 2 \sum_{i\varepsilon T} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i)' V_e^{-1} \underset{\sim}{s}_i \\ & + \sum_{i\varepsilon T} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i)' V_e^{-1} (\underset{\sim}{z}_i - \underset{\sim}{\mu}_i) \qquad (5.23)\end{aligned}$$

By comparing (5.23) with (3.37), we see that the status indices (definition 3.6) must be generalized as follows.

$$a_i = \begin{cases} 0 & \text{if } i \text{ is not observed} \\ 1 & \text{if } i \text{ is observed} \\ v & \text{if } i \text{ represents } v \text{ monozygotic sibs} \end{cases}$$

$$b_i = \begin{cases} 1 & \text{if } i \ \varepsilon\ M \\ \alpha_{3i}^{-1} & \text{if } i \ \varepsilon\ I \end{cases}$$

$$c_i = \begin{cases} \alpha_{1i}^2 \alpha_{3i}^{-1} \times \text{the number of children of } i \text{ if } i \text{ is male} \\ \alpha_{2i}^2 \alpha_{3i}^{-1} \times \text{the number of children of } i \text{ if } i \text{ is female} \end{cases}$$

$$c_{ij} = \alpha_{1i}\alpha_{2i}\alpha_{3i}^{-1} \times \text{the number of children } i \text{ and } j \text{ have in common}$$

$$d_{ij} = \begin{cases} \alpha_{i1}\alpha_{3i}^{-1} & \text{if } j \text{ is the father of } i \\ \alpha_{i2}\alpha_{3i}^{-1} & \text{if } j \text{ is the mother of } i \\ 0 & \text{otherwise} \end{cases} \qquad (5.24)$$

The D's then have the same definition as in Theorem 3.2. With these changes the formulas for non-consanguineous loops can be used for consanguineous loops as well.

CHAPTER VI

MAXIMUM LIKELIHOOD ESTIMATION AND HYPOTHESIS TESTING

6.1 Introduction

For univariate traits it has been said that the single most important question concerning inheritance is whether or not a single locus can account for most of the genetic variation (Elston and Stewart, 1971; Elston and Rao, 1978). With regard to multivariate traits, the most important question is possibly this: Are the component traits under common genetic control? A positive answer would mean the pleiotropic expression of at least one locus.

Our initial motivation for considering the problem of estimating the covariance component matrices was that, by examining their eigen structures, we could answer the above question. The eigen structures also provide a linear index that could be used as a measure of the innate trait and for identifying individuals at risk. But since we have advocated the use of the maximum likelihood method of estimation we also have a natural statistic, namely the likelihood ratio criterion, for answering the question. We shall discuss both methods in this chapter.

The estimation and tests proposed will be based largely on the multivariate mixed model developed in Chapter IV. The other modes of inheritance are special cases of this, and furthermore tests related to them are trivial extensions of univariate tests, so for brevity we shall not discuss them here.

6.2 Estimation of Genetic Components of Covariance

Theoretically the statistical problem of estimating the genetic components of covariance, from pedigree data, falls into the general category of covariance components estimation for random or mixed designs with unbalanced data. The desirability of, and problems associated with, maximum likelihood estimation in this general setting have already been discussed. Closed form solutions to the likelihood equations cannot be found in general, Lange et al. (1981). In fact for the models we have described, it does not as yet appear feasible to derive likelihood equations which could be solved iteratively, though Ott (1979) made a start in this direction. We therefore resort to direct search methods, and it is for this reason that the formulas we have developed in Chapters III, IV and V are useful. We can use them in conjunction with computer routines, such as MAXLIK (Kaplan and Elston, 1972) to determine maximum likelihood estimates of the genetic components of covariance by searching the likelihood surface. MAXLIK also uses numerical methods to determine the sample information matrix, which is inverted to obtain the covariance matrix of the estimates.

As for univariate traits the likelihood surface may have several local maxima, and there is no general way of knowing how many such maxima exist. The advice of Elston (1980) is to search for more than one local maximum whenever a particular set of parameters are estimated.

6.3 The Likelihood Ratio Test for the Major Gene Hypothesis

The test developed here is a multivariate generalization of one of the tests for major gene proposed by Elston and Stewart (1971). Consider a mixed model with one two-allele locus, plus a polygenic component. The three genotypes may be denoted AA, Aa and aa. For the univariate trait z, we may assume the conditional phenotypic distribution

$$g_t(z) = \phi(\mu_t - z, \sigma_e^2), \quad t = AA, Aa, aa,$$

and also that the genotype frequencies ψ_{AA}, ψ_{Aa} and ψ_{aa} depend on a single parameter q, the gene frequency. The model then depends on seven parameters

$$q, \mu_{AA}, \mu_{Aa}, \mu_{aa}, \sigma_g^2, \sigma_e^2.$$

The test for major gene effect is then equivalent to testing the hypothesis

$$H_0\text{:} \quad q = 0, \mu_{AA} = \mu_{Aa} = \mu_{aa}.$$

For a p-variate trait $\underset{\sim}{z}$, the μ's become p×1 vectors and the σ^2's become p×p matrices. The parameters of the mixed model are

$$q, \underset{\sim}{\mu}_{AA}, \underset{\sim}{\mu}_{Aa}, \underset{\sim}{\mu}_{aa}, V_g, V_e$$

so that the null hypothesis for major gene becomes

$$H_0\text{:} \quad q = 0, \underset{\sim}{\mu}_{AA} = \underset{\sim}{\mu}_{Aa} = \underset{\sim}{\mu}_{aa}. \tag{6.1}$$

Using the likelihood under the mixed model discussed in Chapter IV, we may compute the likelihood ratio as

$$\lambda = \frac{\max\limits_{\underset{\sim}{\theta}} L_M(\underset{\sim}{\theta}, \underset{\sim}{z} | H_0)}{\max\limits_{\underset{\sim}{\theta}} L_M(\underset{\sim}{\theta}, \underset{\sim}{z})} \tag{6.2}$$

where $\max\limits_{\underset{\sim}{\theta}} L_M(\underset{\sim}{\theta}, \underset{\sim}{z} | H_0)$ is the ML computed with the restrictions imposed by H_0 (6.1), and $\max L_M(\underset{\sim}{\theta}, \underset{\sim}{z})$ is the unrestricted maximum likelihood. Usually the statistic $-2 \log_e \lambda$ has asumptotic χ^2 distribution with

2p+1 degrees of freedom. It should however be noted that under H_0, q is indeterminate when $\underset{\sim}{\mu}_{AA} = \underset{\sim}{\mu}_{Aa} = \underset{\sim}{\mu}_{aa}$ and the latter are not (distinctly) determinate when q=0. These boundaries can cause problems. Wolfe (1971) suggests from a simulation study of the sampling distribution of the likelihood ratio for independent multinormal mixtures, that for hypotheses involving boundary values of parameters, doubling the degrees of freedom gives a better fit to the sampling distribution. With the dependencies in pedigree data, Wolfe's suggestion needs verifying here--an area of future research.

Other hypotheses can also be tested. For example, the hypothesis of Hardy-Weinberg equilibrium proportions for the major genotype can easily be tested, by noting that the genotypic distribution is given by $\psi_{AA} = q^2$, $\psi_{Aa} = 2q(1-q)$ and $\psi_{aa} = (1-q)^2$ corresponds to the restriction $\psi_{Aa} = 2(\psi_{AA} \cdot \psi_{aa})^{\frac{1}{2}}$. This is of course one of the tests for univariate traits that carry over to the multivariate situation. Another is the test of the presence of simple Mendelian segregating proportions via transmission probabilities (Elston and Stewart, 1971). Other tests are the hypotheses of dominance: $\underset{\sim}{\mu}_{AA} = \underset{\sim}{\mu}_{Aa}$ or $\underset{\sim}{\mu}_{Aa} = \underset{\sim}{\mu}_{aa}$, and the absence of polygenic inheritance which corresponds to the restriction $V_g = \underset{\sim}{0}$. The degrees of freedom for $-2 \log_e \lambda$ in each of these tests equals the number of restrictions imposed by the null hypothesis.

6.4 The Ratio of Generalized Genetic Variances

The generalized variance, defined as the determinant of the covariance matrix is accepted as an overall measure of variation in multivariate populations or samples. From the mixed model we can obtain estimates for V_{mg} and V_{pg}, the covariance matrices for the major gene locus and polygenic loci respectively. The proportion of genetic variation in the multivariate trait due to the major gene is then

$$\Lambda = \frac{|V_{mg}|}{|V_{mg} + V_{pg}|} \tag{6.3}$$

with sample estimate

$$\hat{\Lambda} = \frac{|\hat{V}_{mg}|}{|\hat{V}_{mg} + \hat{V}_{pg}|} . \tag{6.4}$$

$\hat{\Lambda}$ formally resembles the Wilk's Λ statistic whose distribution is well known (See for example Anderson, 1958). However Wilk's Λ requires $\hat{V}_{mg}$ and $\hat{V}_{pg}$ to have independent Wishart distributions, but in our case their functional relationships with $\underset{\sim}{z}$ cannot be determined explicitly. We cannot therefore demonstrate that $\hat{V}_{mg}$ and $\hat{V}_{pg}$ are Wishart matrices. We shall use the Taylor series method to derive a consistent estimate of the sampling variance from which standard errors can be computed. Now let

$$V_{mg} = (\sigma_{mg_{ij}}) \text{ and } V_{pg} = (\sigma_{pg_{ij}}).$$

Since these are covariance matrices

$$\sigma_{mg_{ij}} = \sigma_{mg_{ji}} \text{ and } \sigma_{pg_{ij}} = \sigma_{pg_{ji}} .$$

Let $\underset{\sim}{w}_{mg}$ be the $\frac{1}{2}p(p+1) \times 1$ vector of the distinct elements in V_{mg} and $\underset{\sim}{w}_{pg}$ the corresponding vector for distinct elements in V_{pg}; $\underset{\sim}{w}_{mg}$ is formed simply by rolling out by rows (or columns) the lower half of V_{mg}, i.e.,

$$\underset{\sim}{w}_{mg} = [\sigma_{mg_{11}}, \sigma_{mg_{21}}, \sigma_{mg_{22}}, \sigma_{mg_{31}}, \ldots, \sigma_{mg_{33}}, \ldots, \sigma_{mg_{p1}}, \sigma_{mg_{p2}}, \ldots, \sigma_{mg_{pp}}]' .$$

Similarly

$$\underset{\sim}{w}_{pg} = [\sigma_{pg_{11}}, \sigma_{pg_{21}}, \sigma_{pg_{22}}, \sigma_{pg_{31}}, \ldots, \sigma_{pg_{33}}, \ldots, \sigma_{pg_{p1}}, \sigma_{pg_{p2}}, \ldots, \sigma_{pg_{pp}}]' .$$

Now let $\underset{\sim}{w} = \begin{pmatrix} \underset{\sim}{w}_{mg} \\ \underset{\sim}{w}_{pg} \end{pmatrix}$. Then $\underset{\sim}{w}$ is the $p(p+1) \times 1$ vector of all distinct major gene and polygenic components of variances and covariances. Denote the $p(p+1) \times 1$ vector of partial derivatives of Λ with respect to the elements of $\underset{\sim}{w}$ by $\underset{\sim}{d}$. Let $\hat{\underset{\sim}{w}}$ be the sample estimate of $\underset{\sim}{w}$, with estimated covariance matrix $\hat{V} = (\text{cov}(\hat{w}_i, \hat{w}_j))$, and $\underset{\sim}{d}(\hat{\underset{\sim}{w}})$ the values of the derivatives evaluated at the sample values $\hat{\underset{\sim}{w}}$, then the Taylor series approximation to the variance of $\hat{\Lambda}$ is given by

$$\text{var}(\hat{\Lambda}) \doteq \underset{\sim}{d}'(\hat{\underset{\sim}{w}})\ V\ \underset{\sim}{d}(\hat{\underset{\sim}{w}}) \tag{6.5}$$

See for example Rao (1973).

To use the formula we need to calculate the derivatives $\underset{\sim}{d}$, or equivalently

$$\frac{\partial\Lambda}{\partial\sigma_{mg_{ij}}} \text{ and } \frac{\partial\Lambda}{\partial\sigma_{pg_{ij}}}, \qquad i \leq j = 1,2,\ldots,p.$$

It is more convenient to evaluate the matrices

$$\frac{\partial\Lambda}{\partial V_{mg}} = \left(\frac{\partial\Lambda}{\partial\sigma_{mg_{ij}}}\right) \text{ and } \frac{\partial\Lambda}{\partial V_{pg}} = \left(\frac{\partial\Lambda}{\partial\sigma_{pg_{ij}}}\right).$$

We need two results on derivatives of determinants. The first is standard (see for example Rao (1973, p. 72). For a symmetric nonsingular matrix X

$$\frac{\partial}{\partial X}|X| = |X|\{2X^{-1} - \text{diag}(X^{-1})\}.$$

The second result is an easy extension of this. Let X and Y be symmetric matrices such that X+Y is symmetric and non-singular. Then

$$\frac{\partial}{\partial X}|X+Y| = \frac{\partial}{\partial(X+Y)}|X+Y| \cdot \frac{\partial}{\partial X}(X+Y)$$

$$= |X+Y|\{2(X+Y)^{-1} - \text{diag}(X+Y)^{-1}\}\cdot \underset{\sim}{I}$$

$$= |X+Y|\{2(X+Y)^{-1} - \text{diag}(X+Y)^{-1}\}.$$

Using these two results we find

$$\frac{\partial \Lambda}{\partial V_{mg}} = \{|V_{mg}+V_{pg}| \cdot \frac{\partial}{\partial V_{mg}}|V_{mg}| - |V_{mg}| \cdot \frac{\partial}{\partial V_{mg}}|V_{mg}+V_{pg}|\}/|V_{mg}+V_{pg}|^2$$

$$= \frac{|V_{mg}|}{|V_{mg}+V_{pg}|}\{(2V_{mg}^{-1}-\text{diag }V_{mg}^{-1}) - (2(V_{mg}+V_{pg})^{-1} - \text{diag}(V_{mg}+V_{pg})^{-1}\}$$

$$= \Lambda\{2V_{mg}^{-1}-\text{diag }V_{mg}^{-1}) - (2(V_{mg}+V_{pg})^{-1} - \text{diag}(V_{mg}+V_{pg})^{-1})\} \qquad (6.6)$$

Similarly,

$$\frac{\partial \Lambda}{\partial V_{pg}} = -\frac{|V_{mg}|}{|V_{mg}+V_{pg}|^2} \cdot \frac{\partial}{\partial V_{pg}}|V_{mg}+V_{pg}|$$

$$= -\Lambda\{2(V_{mg}+V_{pg})^{-1} - \text{diag}(V_{mg}+V_{pg})^{-1}\}\ . \qquad (6.7)$$

The vector $\underset{\sim}{d}$ is formed by rolling out $\frac{\partial \Lambda}{\partial V_{mg}}$ and $\frac{\partial \Lambda}{\partial V_{pg}}$ as for $\underset{\sim}{w}$ above.

The standard error of $\hat{\Lambda}$ is approximately estimated by

$$\text{se}(\hat{\Lambda}) = \sqrt{\text{var}(\hat{\Lambda})}\ . \qquad (6.8)$$

It follows that the statistic

$$t = \frac{\hat{\Lambda}}{\text{se}(\hat{\Lambda})} = \frac{\hat{\Lambda}}{\sqrt{\text{var}(\hat{\Lambda})}} \qquad (6.9)$$

is asymptotically $N(0,1)$. The estimate of the standard error determined above is very approximate, since the derivatives $\underset{\sim}{d}(\hat{\underset{\sim}{w}})$ are evaluated at the sample values $[\hat{V}_{mg},\hat{V}_{pg}]$. The true values $[V_{mg},V_{pg}]$ are unknown and the sample estimates are likely to be biased.

6.5 Index Procedures

We now consider problems related to the genetic analysis of a linear function of the components of $\underset{\sim}{z}$. We shall refer to this function as an index and denote it by

$$I(\underset{\sim}{z}) = \underset{\sim}{a}'\underset{\sim}{z} \tag{6.10}$$

Two issues of particular interest are:

(i) the choice of $\underset{\sim}{a}$ to maximize the heritability of $I(\underset{\sim}{z})$, and

(ii) the statistical significance of the major gene component to the additive genetic variance of $I(\underset{\sim}{z})$.

It is hoped that the index $I(\underset{\sim}{z})$ would be a better measure of the "innate" trait than each component of $\underset{\sim}{z}$ separately.

The heritability of the index is

$$h_I^2 = \frac{\underset{\sim}{a}'V_g\underset{\sim}{a}}{\underset{\sim}{a}'(V_g+V_e)\underset{\sim}{a}} \tag{6.11}$$

with sample estimate

$$\hat{h}_I^2 = \frac{\underset{\sim}{a}'\hat{V}_g\underset{\sim}{a}}{\underset{\sim}{a}'(\hat{V}_g+\hat{V}_e)\underset{\sim}{a}} \tag{6.12}$$

where $\hat{V}_g$ and $\hat{V}_e$ are ML estimates. The problem of maximizing h_I^2 is the familiar problem of maximizing the ratio of two quadratic forms. The solution is given by the largest root (λ_1) of the determinental equation

$$|V_g - \lambda(V_g+V_e)| = 0 \tag{6.13}$$

with sample equivalent

$$|\hat{V}_g - \hat{\lambda}(\hat{V}_g+\hat{V}_e)| = 0; \tag{6.14}$$

$\hat{\underset{\sim}{a}}$ is the characteristic vector corresponding to $\hat{\lambda}_1$. The statistical significance of $\hat{h}_I$ may be determined from the distribution of Roy's largest root criterion. However, this test will not be exact since $\hat{V}_g$ and $\hat{V}_e$ are correlated and may not be Wishart matrices.

We next ask whether an index so determined has a significant major gene component in its genetic variance. This can be done using univariate methods. Notice however that the genetic variance of the index may be partitioned into major gene and polygenic components as follows.

$$\mathrm{Var}_g[I(\underset{\sim}{z})] = \underset{\sim}{a}'V_g\underset{\sim}{a} = \underset{\sim}{a}'V_{mg}\underset{\sim}{a} + \underset{\sim}{a}'V_{pg}\underset{\sim}{a} \qquad (6.15)$$

The sample estimate of the proportion of genetic variance due to the major gene is

$$\hat{\nu} = \frac{\underset{\sim}{a}'\hat{V}_{mg}\underset{\sim}{a}}{\underset{\sim}{a}'(\hat{V}_{mg}+\hat{V}_{pg})\underset{\sim}{a}} \quad . \qquad (6.16)$$

We shall obtain approximate large sample estimates of the standard error of $\hat{\nu}$ below.

A different question that may be asked in the spirit of the one above is whether we can find a linear function that has maximum major gene variance relative to the total genetic variance. This proportion is

$$\tau = \max_{\underset{\sim}{b}} \frac{\underset{\sim}{b}'V_{mg}\underset{\sim}{b}}{\underset{\sim}{b}'(V_{mg}+V_{pg})\underset{\sim}{b}} \qquad (6.17)$$

with sample estimate

$$\hat{\tau} = \max_{\underset{\sim}{b}} \frac{\underset{\sim}{b}'\hat{V}_{mg}\underset{\sim}{b}}{\underset{\sim}{b}'(\hat{V}_{mg}+\hat{V}_{pg})\underset{\sim}{b}}$$

$$= \lambda_1,$$

the largest characteristic root of $\hat{V}_{mg}(\hat{V}_{mg}+\hat{V}_{pg})^{-1}$, and $\underset{\sim}{b}_{max}$ is the associated characteristic vector.

We shall now derive approximate large sample variances for the statistics $\hat{h}_I^2$, $\hat{\nu}$ and $\hat{\tau}$. Consider the ratio

$$f(V_1,V_2) = \frac{\underset{\sim}{a}'V_1\underset{\sim}{a}}{\underset{\sim}{a}'(V_1+V_2)\underset{\sim}{a}} \tag{6.18}$$

At the sample values we have

$$\hat{f} = f(\hat{V}_1,\hat{V}_2) = \frac{\underset{\sim}{a}'\hat{V}_1\underset{\sim}{a}}{\underset{\sim}{a}'(\hat{V}_1+\hat{V}_2)\underset{\sim}{a}} \tag{6.19}$$

$$\frac{\partial f}{\partial V_1} = (\underset{\sim}{a}'(V_1+V_2)\underset{\sim}{a} \cdot \underset{\sim}{a}\underset{\sim}{a}' - \underset{\sim}{a}'V_1\underset{\sim}{a} \cdot \underset{\sim}{a}\underset{\sim}{a}')/[\underset{\sim}{a}'(V_1+V_2)\underset{\sim}{a}]^2 \ .$$

Evaluating at the sample values, we have

$$\left.\frac{\partial f}{\partial V_1}\right|_{(\hat{V}_1,\hat{V}_2)} = \frac{\underset{\sim}{a}'\hat{V}_2\underset{\sim}{a}}{[\underset{\sim}{a}'(\hat{V}_1+\hat{V}_2)\underset{\sim}{a}]^2} \cdot \underset{\sim}{a}\underset{\sim}{a}' \ . \tag{6.20}$$

Similarly,

$$\left.\frac{\partial f}{\partial V_2}\right|_{(\hat{V}_1,\hat{V}_2)} = \frac{\underset{\sim}{a}'\hat{V}_1\underset{\sim}{a}}{[\underset{\sim}{a}'(\hat{V}_1+\hat{V}_2)\underset{\sim}{a}]^2} \ \underset{\sim}{a}\underset{\sim}{a}' \ . \tag{6.21}$$

Now let $\hat{V}$ be the estimate of the covariance matrix of parameter estimates, and $\underset{\sim}{d}(\hat{\underset{\sim}{w}})$ be the p(p+1) × 1 vector formed from the distinct elements of

$$\left(\left.\frac{\partial f}{\partial V_1}\right|_{(V_1,V_2)} , \left.\frac{\partial f}{\partial V_2}\right|_{(V_1,V_2)}\right) \quad \text{, as in section 6.4.}$$

Then the large sample Taylor series approximation to the variance of $\hat{f}$ is obtained by substituting $\hat{V}$ and $\underset{\sim}{d}(\hat{\underset{\sim}{w}})$ in (6.5).

6.6 Some Remarks

We have discussed maximum likelihood methods for estimating parameters of the multivariate versions of the genetic models, and derived large sample tests for the single gene hypothesis. The methods presented would be adequate for segregation analysis of multivariate traits with human pedigree data.

CHAPTER VII

SUMMARY AND CONCLUSIONS

We have concerned ourselves in this thesis with the problem of analyzing the genetic basis of covariation of multivariate traits in families and extended families, or pedigrees, using maximum likelihood methods. Specifically we have developed a model that is general with regard to the genetic mechanism of transmission as well as pedigree structure and consanguinity. The problem of computing the resulting likelihood function has been studied analytically in some depth, yielding recursive formulas that bring the calculations envisaged within the capabilities of present computing facilities, thus removing what could be the most serious objection to using the model.

It has been possible to bring the many facets of the problem of genetic analysis,as well as the complexities of pedigree structure, into one computational scheme by integrating the Elston and Stewart (1971) model and the Lange, Boehnke and Spence (1981) model. Ours is the second attempt at some kind of integration of the two models; Ott's (1979) was the first. What is new in our approach is that we start with the Elston and Stewart basic formulation, but replace their conditional density function, p_{FGH}, for the polygenic component by a more general one derived from the model by Lange et al. The new p_{FGH} depends on quantities α_1, α_2 and α_3 which we call relatedness coefficients; the original Elston and Stewart model is the special case:

$\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{2}$. This formulation allows us to handle consanguinity in pedigrees when the traits have multinormal phenotypes. The α's are determined by the pedigree structure and are assumed known. It is interesting to note that if we multiply α_1 and α_2 by $\frac{1}{1+\rho}$ and α_3 by $1+\rho$, where ρ is the correlation between polygenotypes of mates, we allow for assortative mating. Furthermore if α_1, α_2 and α_3 are assumed to be arbitrary parameters to be estimated, we allow for cultural inheritance that is completely confounded with polygenic inheritance, Rice et al. (1978), as quoted in Boyle and Elston (1979).

The oligogenic component of our model is exactly the same as in the original Elston and Stewart model. It therefore follows that reparametrization in terms of transmission probabilities can be done for linkage studies and tests of mode of transmission.

The ideas used by Lange and Elston (1975) to calculate the likelihood of oligogenic inheritance in complex pedigrees have been extended to polygenic inheritance and combined oligogenic and polygenic inheritance: the mixed model. However complex the pedigree structure,it can be broken (at least conceptually) into simple pedigrees and every individual described by his status indices, together with pointers or other indices that indicate who the individual's spouse and children are.

An important feature of our model and computational methods is that the highest order of matrices that enter into our calculations is $p \times p$ for a p-variate trait, instead of $np \times np$ (where n is the size of the pedigree) as in Lange et al. This, as noted in Chapter IV,is one reason it is not feasible to compute the likelihood as Ott formulated it. The other reason is that in Ott's model it is necessary to store all possible k^n genotype combinations--a problem completely circumvented by the use of our recurrence formulas.

From the definition of terms in the recurrence formula (3.29), it is evident that the key to the calculation of the likelihood function, and hence the development of an efficient computer software for pedigree analysis,is the status indices which are functions of the relatedness coefficients.

Our analyses have further shown that the exact calculation of the likelihood under the mixed model of inheritance, involving both oligogenic and polygenic components, is possible. In their review of statistical methods for genetic analysis of quantitative univariate traits in families and pedigrees, Boyle and Elston (1979) stated the need for a computing algorithm for the likelihood under the mixed model of inheritance. Now we have found it. The algorithm only requires modification of the g-functions in the corresponding oligogenic model by the recurrence formulas we have developed, a fact that will obviously facilitate the development of general computing software for pedigree analysis.

The actual estimation and hypothesis testing procedures are discussed in Chapter VI. Estimates are found by direct search of the likelihood surface; hence the need for fast and accurate algorithms for calculating the likelihood. Tests are based on the likelihood ratio criterion and an examination of the eigen structure of the genetic components of covariances. Standard errors of estimates are derived by the Taylor series approximation method. With the dependencies in pedigree data, and the fact that most of the tests involve boundary values of some of the parameters, the statistical procedures outlined are crude at best. Numerical studies could lead to refinements as in Wolfe (1971).

The work we have presented here is probably the first attempt at developing a general model and computing algorithm for the genetic analysis of multivariate traits. We have no doubt concentrated on the barest essentials. As with the original Elston and Stewart model, it is easy to foresee a whole generation of modelists springing up and embellishing the model with such additions as ascertainment corrections, age of onset functions, common family or sibship environments, general environmental covariates, and the like. However these would be at the cost of increasing the number of parameters to be estimated, which is already formidable even for a small number of traits. Yet given the realities of the non-experimental data we have to work with, such efforts are desiderata, and in theory at least, easy to do. It is also possible that closer scrutiny, practical considerations in software development and numerical studies would suggest modifications and refinements to the formulas and algorithms we have derived.

In conclusion we remark that the genetic basis of covariation is of such practical importance that any new advances in analytic and computing tools will be welcome.

REFERENCES

Anderson, T. W. (1958). An introduction to multivariate statistical analysis. New York: Wiley and Sons, Inc.

Beaty, T. L. H. (1978) Likelihood analysis of multivariate phenotypes in pedigrees. Ph.D. Thesis, University of Michigan.

Boyle, C. R. and Elston, R. C. (1979). Multifactorial genetic models for quantitative traits in humans. Biometrics 35, 55-68.

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.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39, 859-882.

Dawson, D. V. (1981). Problems of ascertainment in pedigree analysis. Ph.D. Thesis, University of North Carolina at Chapel Hill.

Eaves, L. J. and Gale, J. S. (1974). A method for analyzing the genetic basis of covariation. Behaviour Genetics 4, 253-267.

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

Elston, R. C. and Boklage, C. E. (1978). International Twin Conference, 2nd, 1977, Washington, D. C. New York: Alan Liss.

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 (Lond.) 39, 67-87.

Elston, R. C. and Rao, D. C. (1978). Statistical modeling and analysis in human genetics. Ann. Rev. Biophys. Bioeng. 7, 253-286.

Elston, R. C. and Sobel, E. (1979). Sampling considerations in gathering and analysis of pedigree data. American Journal of Human Genetics 31, 62-69.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. Human Heredity 21: 523-542.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinb. 52, 399-433.

Forthofer, R. N. and Koch, G. G. (1974). An extension of the symmetric sum approach to the estimation of variance components. Biom. Z. Bd. 16. 1974. Heft 1.S. 3-14.

Gillois, M. (1966). Le concept d'identité et son importance en génétique. Annales de Génétique 9, 58-65.

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.

Goldin, L. R. (1978). Genetic analysis of Von Willebrand's disease in two large pedigrees: a multivariate approach. Ph.D. Thesis, University of North Carolina at Chapel Hill.

Graepel, G. J. (1981). Multifactorial models and likelihoods for the segregation analysis of quantitative traits. Ph.D. Thesis, University of North Carolina at Chapel Hill.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. Biometrika 54, 93-108.

Henderson, C. R. (1953). Estimation of variance and covariance components. Biometrics 9, 226-252.

Jacquard, A. (1974). The genetic structure of populations. New York: Springer Verlag.

Kempthorne, O. (1957). An introduction to genetic statistics. New York: John Wiley & Sons, Inc.

Koch, G. G. (1967). A general approach to the estimation of variance components. Technometrics 9, 93-118.

Koch, G. G. (1968). Some further remarks on "A general approach to the estimation of variance components." Technometrics 10, 551-558.

Lalouel, J. M. and Morton, N. E. (1981). Complex segregation analysis with pointers. To appear in Human Heredity.

Lange, K. (1978). Central limit theorems for pedigrees. Journal of Math. Biology 6, 59-66.

Lange, K., Boehnke, M. and Spence, M. A. (1981). Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. Submitted for publication.

Lange, K. and Elston, R. C. (1975). Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigree. Human Heredity 25, 95-105.

Lange, K., Westlake, J., and Spence, M. A. (1976). Extensions to pedigree analysis. III. Variance components by the scoring method. Annals Human Genetics 39, 485-491.

MacLean, C. J., Morton, N. E., and Lew, R. (1975). Analysis of family resemblance. IV.Operational characteristics of segregation analysis. American Journal of Human Genetics 27, 365-384.

Malecot, G. (1948). Les Mathématiques d'l'hérédité . Paris, Masson et Cie.

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.

Namboodiri, K. K., Elston, R. C., Glueck, C. J., Fallat, R., Buncher, C. R., and Tsang, R. (1975). Bivariate analysis of cholesterol and triglyceride levels in families in which probands have IIb lipoprotein phenotype. American Journal of Human Genetics 27, 454-471.

Namboodiri, K. K., Elston, R. C., and Hames, C. (1977). Segregation and linkage analyses of a large pedigree with hypertriglyceridemia. American Journal of Medical Genetics 1, 151-171.

Ott, J. (1974). Computer simulation in human linkage analysis. American Journal of Human Genetics 26, 64A.

Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. American Journal of Human Genetics 31, 161-175.

Rao, C. R. (1973). Linear statistical inference and its applications, 2nd ed. New York: Wiley.

Rohde, C. A. and Tallis, G. M. (1969). Exact first- and second-moments of estimates of components of covariance. Biometrika 57, 517-526.

Searle, S. R. (1971). Linear models. New York: Wiley.

Simpson, J. M. (1981). The genetic analysis of quantitative traits in humans. Ph.D. Thesis, University of New South Wales.

Thompson, W. A., Jr. (1962). The problem of negative estimates of variance components. Annals of Math. Stat. 33, 273-289.

Wolfe, J. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixture of multinormal distributions. _Naval Personnel and Training Research Laboratory, Technical Bulletin STB_ _72_, 2.

ABSTRACT

GEORGE EBOW BONNEY. Maximum Likelihood Methods for Genetic Analysis
of Multivariate Pedigree Data. (Under the direction of
Robert C. Elston.)

The problem of analyzing the genetic basis of covariation of multivariate traits in families and extended families, or pedigrees, is studied using maximum likelihood methods. A general model is presented that includes all the features of the Elston and Stewart (1971) model and the essentials of the Lange, Boehnke and Spence (1981) model. The integration of the two models, achieved through the introduction of quantities we call "relatedness coefficients," makes it possible to bring the many facets of the problem of genetic analysis as well as the complexities of pedigree structure into one computational scheme.

Simple definitions and classifications of pedigrees are introduced and a device first used by Dr. Philip Green, III, is extended and used to define "status indices" that summarize in numerical codes the relevant information on every individual in the pedigree. The status indices turn out to be simple functions of the relatedness coefficients, and they enable us to formulate likelihood functions for continuous traits in complex pedigrees.

The computational problem is studied analytically in depth yielding recurrence formulas for the exact calculation of the likelihood function even for the case of combined oligogenic and polygenic inheritance in simple as well as complex pedigrees, thus removing what has been by far the most serious objection to the use of such models.

ACKNOWLEDGMENTS

I should first thank my thesis advisor, Dr. Robert C. Elston, for his guidance and enthusiastic support. I would also thank the other members of my thesis committee, Drs. Kadambari K. Namboodiri, Michael J. Symons, Gary G. Koch, and Kerry L. Lee, for their encouragement and helpful suggestions. I also had fruitful discussions with Ellen E. Kaplan on computer algorithms, and with Dr. Philip Green, III, whose unpublished algorithm I have extended in parts of this thesis.

In addition I wish to thank the Population Dynamics Programme at the University of Ghana, and the Ghana Government for their financial support.

I must also thank my wife, Efua Esaaba Bonney for her remarkable understanding, encouragement and prayer support, without which this thesis would not have been completed. Two are really better than one. And, my two children Aba and Kofi, who, by being themselves throughout the period of writing, have taught me practical lessons in "nature-nurture."

Mrs. Gay Hinnant did this excellent typing.

Finally, I must remember my roots and in particular single out my uncle John Obu-Simpson (alias Kweku Semi) who at great personal expense supervised and provided for my early childhood education through the prestigious Mfantsipim School, Cape Coast, Ghana, where my love for science was born and nurtured and a good foundation for mathematics laid. To him I gladly dedicate this thesis.

TO

JOHN OBU-SIMPSON

Alias

KWEKU SEMI