International Journal of Statistics and Probability; Vol. 3, No. 2; 2014
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Estimating Statistical Measures of Pleiotropic and Epistatic Effects in the Genomic Era

Charles J. Mode[1]

[1] Professor Emeritus, Department of Mathematics, Drexel University, Philadelphia, PA, USA

Correspondence: Charles J. Mode, Professor Emeritus, Department of Mathematics, Drexel University, Philadelphia, PA 19104, USA. E-mail: cjmode@comcast.net

Received: January 24, 2014 Accepted: March 28, 2014 Online Published: April 1, 2014

doi:10.5539/ijsp.v3n2p81 URL: http://dx.doi.org/10.5539/ijsp.v3n2p81

Abstract

Recent developments in the technology for sequencing the genomes of various species has had a profound effect of the working paradigms of various fields of genetics. Included among these fields is the classical field of quantitative genetics, which is a subfield of statistical genetics, that is devoted to traits that can be quantified on some continuous scale and are often influenced by alleles at many loci. In recent years, many investigators have conducted genome wide sweeps and have used a variety of statistical criteria to judge whether identified regions of the human genome have a significant influence on the expression of some quantitative trait such as measurements on patients with Alzheimer's disease. From the point of view of quantitative genetics, the regions of a genome that have some influence on a quantitative trait may be viewed as loci, and variations among these loci at the *DNA* level, such as nucleotide substitutions or other markers, may be used as working definitions of alleles, and, therefore, can be used to determine whether an individual carries a particular allele at some locus. Given such data, an investigator can identify the genotype of each individual in a study, with respect to the loci under consideration as well as the two alleles present at each locus in a diploid species such as man. This ability to use these working definitions to identify the genotype of each individual in a sample results in a significant change in the working paradigm of sub-field of quantitative genetics, called variance and covariance analysis, because effects and components of variance and covariance may be estimated directly in a sense that will be described in detail in the paper.

Keywords: loci, alleles, genotypic distribution, phenotypic, genetic and environmental covariance matrices, positive definite matrices, direct estimation of effects, variance and covariance components, multivariate analysis, set theory, class of all subsets of a set, corresponding effect for each subset

1. Introduction

Following the sequencing of the human genome in 2000 as well as that of other species, sequencing technology has advanced to the point at which it is becoming financially feasible, see Church (2006), to sequence the genomes of individuals in samples under study by an investigator or teams of investigators. This technological advance has led investigators to conduct genome wide sweeps to search for regions of the human genome as well as those of other species such that there is evidence to suggest that these regions are implicated in the expression of complex quantitative traits. In an interesting paper by Stranger et al. (2010), the impact of genome wide association studies on the genetics of complex traits is discussed in depth. Among these complex traits are Alzheimer's disease (*AD*) and immune-mediated diseases such as rheumatoid arthritis. For the case of *AD*, in a recent paper Raj et al. (2012) have reported that 11 regions of the human genome are involved in susceptibility to this disease, and, moreover, there is evidence that four of these regions form a protein-protein interaction network that is under natural selection. Similarly, in a paper by Rossin et al. (2011), it has been found that proteins encoded in genomic regions associated with immune-mediated disease physically interact and this interaction may also suggest some biological mechanisms underlying such diseases. In some samples of patients with *AD*, the genome of each individual in the sample has been sequenced so that, in principle, the genotype of each individual may be determined with respect to 11 loci, regions of the human genome, or in combinations of these loci, whenever there are working definitions of at least two alleles at each locus.

This ability to identify the genotype of each individual in a sample with respect to some quantitative trait provides a concrete basis for extending some of the techniques of classical quantitative genetics into the era of sequenced

genomes. In classical quantitative genetics, however, the loci under consideration as well as the alleles at each locus were unknown to an investigator so that these notions were treated abstractly and the parameters of a model could only be estimated indirectly. Briefly, a variety of analysis of variances and covariance procedures were used to estimate parameters of interest indirectly. A concrete example of such procedures may be found in the material accompanying Table 1 in Mode and Robinson (1959). But, as will be shown in subsequent sections, whenever the genotype of each individual in a sample may be identified with respect to some loci with at least two alleles at each locus, the parameters of the model be estimated in a straight forward manner based on elementary methods of statistical estimation.

As is recognized by many who at sometime during their careers have worked in the field of quantitative genetics, the subject known as components of variance analysis began with the publication of a paper on correlations among relatives on the supposition of Mendelian inheritance by Fisher (1918). In this paper, Fisher attempted to reconcile existing biometrical theories with Mendelian genetics that led to describing genetic variation in terms of components of variance. Evidently at that time, the use of the word "variance" was relatively new, because in the beginning of the paper Fisher emphasized the word by representing it in upper case letters. Subsequently, the ideas of Fisher were extended to include epistasis and other interactions among alleles in the seminal papers of Cockerham (1954) and Kempthorne (1954). The techniques introduced in these papers have been applied in the current genomic era. Examples of the ideas introduced by Cockerham have used in the paper Kao and Zeng (2002), and those of Kempthorne have been used and extended in the paper Mao et al. (2006). The ideas of Kempthorne were also used and extended in the paper of Mode and Robinson (1959) as well as in unpublished lecture notes by the author written and presented during the period 1960 to 1966. Furthermore, the roots of the ideas presented in this paper are extensions of the some of the unpublished material in the lecture notes complied by the author during the period 1960 to 1966. Many of the themes of statistical genetics as they existed during the 1950s were summarized and extended in Kempthorne's well known book (Kempthorne, 1957).

During the years following Fisher's seminal work and those cited above by Cockerham and Kempthorne, an extensive literature on quantitative genetics has evolved. It is beyond the scope of this paper to review this literature and in what follows a few books on the subject will be cited. A book that has been very popular with quantitative geneticists is that of Falconer and MacKay (1996) as well as earlier editions. Another book of interest on quantitative genetics is that of Bulmer (1980). Both of these books contain extensive lists of references on quantitative genetics. A more recent book on genetics and analysis of quantitative traits is that of Lynch and Walsh (1998). This is an influential tome, consists of over 900 pages and contains what seems to be the most extensive treatment of the subject of quantitative genetics published in the 20th century. The principal focus of this book is a biological and evolutionary point of view along with an extensive use of applied statistical methods. There is also an extensive list of papers on quantitative genetics that a reader, who is interested in quantitative genetics, may wish to peruse. The book by Liu (1998) on statistical genetics focuses on statistical genetics along with linkage, mapping and quantitative trait linkage (QTL) analysis. Two recent books on statistical genetics are those of Laird and Lange (2011) and Wu, Ma and Casella (2010). It is suggested that if a reader is interested in an overview of the material that is being taught in courses in quantitative and statistical genetics or simply an introduction to these subjects, that the books cited above be consulted.

2. Pleiotropism and the Phenotypic, Genetic and Environmental Covariance Matrices for the Case of One Autosomal Locus

Pleiotropism is a term used by geneticists to describe cases in which several traits, discrete or quantitative, seem to be governed by alleles at a single autosomal locus or a locus on chromosomes that govern the sex of an individual. Because quantitative traits are important in agriculture and medicine, the focus of attention in this paper will be quantitative traits, whose genetics seems to be governed by alleles at a single autosomal locus. Let $\mathbb{A}$ denote the set of alleles at some autosomal locus and let the symbols x and y denote alleles in the set $\mathbb{A}$. In what follows, the number of alleles in this set will be assumed to be finite. The genotype of an individual will be denoted by the symbol (x, y), where x and y are alleles in $\mathbb{A}$ contributed by the maternal and paternal parents, respectively. Let $\mathbb{G}$ denote the set of genotypes under consideration so that for every genotype (x, y), it follows that $(x, y) \in \mathbb{G}$. To take into account pleiotropic effects for $k \geq 2$ quantitative traits, let the $W_1, W_2, \ldots W_k$ be k random variables characterizing the phenotypic variation in these k traits among individuals in a population or sample under consideration. It will be assumed that these random variables take values in the sets $\mathfrak{R}_1, \mathfrak{R}_2, \ldots \mathfrak{R}_k$ of quantitative measurements expressed in terms of real numbers. As a first step in developing a succinct notation that will be used extensively in what follows, let $\boldsymbol{W}$ denote a $k \times 1$ column vector whose components are the

random variables $W_1, W_2, \ldots W_k$. In what follows, the symbol $\mathbb{R}_k$ will denote the set $k \geq 2$ dimensionally set real numbers, which is the set of all possible values that may be realized by the vector $\boldsymbol{W}$, and let the symbol T denote the transpose of a vector or matrix.

As the technology for sequencing genomes of individuals continues to develop, it will become financially feasible to sequence the genomes of all individuals in a sample. Furthermore, in some cases, it has been possible to find evidence that some region of a genome has been implicated in the phenotypic expression of the $k \geq 2$ quantitative traits under consideration. Moreover, it may also be possible to differentiate the alleles in this genomic region so that the genotype (x, y) of each individual in a population or sample may be identified in terms of the bases or other characteristics of the *DNA* in a genomic region by identifying each of the alleles x and y. It will at the discretion of an investigator as to whether the maternal and paternal alleles are identified if such information is indeed available. Actually in what follows all that is necessary is that two alleles carried by an individual at a particular locus be identifiable. In any event, whenever it is possible to identify the genotype of each individual in a population or sample, it will be feasible to develop methods of statistical estimation that can take advantage of this information and provide more direct methods for estimating and drawing inferences about the parameters in a quantitative genetic model accommodating pleiotropic effects.

From the statistical point of view, the model under consideration will be a mixture of a discrete and a continuous multivariate distributions. Let

$$\mathbb{D}_{Geno} = (p(x, y) \mid (x, y \in G)) \tag{2.1}$$

denote the genotypic distribution of the population under consideration, where $p(x, y) \geq 0$ for all $(x, y) \in G$ and

$$\sum\nolimits_{(x,y)} p(x, y) = 1. \tag{2.2}$$

By way of interpreting the distribution $\mathbb{D}_{Geno}$, let the random pair (X, Y) denote a genotype of an individual chosen at random from a population. Then,

$$P[(X, Y) = (x, y)] = p(x, y).$$

It should be mentioned that if a sample of individuals resulted from matings of parents whose genotypes were known, then, for the case of one locus, the theoretical genotypic distribution could be predicted by well known Mendelian theory. But, in general, in most samples of sequenced individuals, the genotypes of their parents are not known so that there would not be a theoretical basis for predicting the form of the genotypic distribution.

Given that $(X, Y) = (x, y)$, let $f(\boldsymbol{w} \mid (x, y))$ denote the conditional density of the random vector $\boldsymbol{W}$. It will be assumed that this distribution has the expectation vector

$$E[\boldsymbol{W} \mid (x, y)] = \boldsymbol{\mu}(x, y) = \int_{\mathbb{R}_k} f(\boldsymbol{w} \mid (x, y))\, d\boldsymbol{w} \tag{2.3}$$

and covariance matrix

$$E\left[(\boldsymbol{W} - \boldsymbol{\mu}(x, y))(\boldsymbol{W} - \boldsymbol{\mu}(x, y))^T \mid (x, y)\right] = \boldsymbol{\Psi}(x, y) \tag{2.4}$$

with finite elements for all genotypes $(x, y) \in \mathbb{G}$. From these definitions, it follows that the unconditional $k \times 1$ expectation vector of the population is

$$\boldsymbol{\mu} = \sum\nolimits_{(x,y)} p(x, y)\, \boldsymbol{\mu}(x, y), \tag{2.5}$$

and the $k \times k$ unconditional covariance matrix of the population is

$$\boldsymbol{\Psi} = \sum\nolimits_{(x,y)} p(x, y)\, \boldsymbol{\Psi}(x, y). \tag{2.6}$$

By definition, the marginal or gene frequency of the maternal allele x in the population is

$$p(x) = \sum\nolimits_{y} p(x, y) \tag{2.7}$$

for all $x \in \mathbb{A}$. The marginal frequency $p(y)$ of the paternal allele y is defined similarly. A population is said to be in Hardy-Weinberg equilibrium if

$$p(x, y) = p(x)\, p(y) \tag{2.8}$$

for all genotypes $(x, y) \in \mathbb{G}$.

In the formulation under consideration, vectors that are measures of first and second order effects of alleles x and y on the k quantitative traits under consideration will be defined in terms conditional expectations of the mean vector $\boldsymbol{\mu}(x,y)$ for each genotypic $(x,y) \in \mathbb{G}$ with respect to the genotypic distribution. In general, one would not expect that a population would be in a Hardy-Weinberg equilibrium so that for any particular sample an investigator may wish to test the hypothesis that a population was in this type of equilibrium. To accommodate the case in which a population is not in a Hardy-Weinberg equilibrium, vectors of first order effects will be defined in terms of conditional distributions. By definition, for $p(x) \neq 0$, the conditional distribution of the allele $y \in \mathbb{A}$, given x, is

$$p(y \mid x) = \frac{p(x,y)}{p(x)}. \tag{2.9}$$

Thus, the conditional expectation of the conditional mean vector $\boldsymbol{\mu}(x,y)$, given x, is

$$\boldsymbol{\mu}(x) = \sum\nolimits_y p(y \mid x)\, \boldsymbol{\mu}(x,y). \tag{2.10}$$

It is interesting to note that if the population is in a Hardy-Weinberg equilibrium, then

$$\boldsymbol{\mu}(x) = \sum\nolimits_y \frac{p(x)\,p(y)}{p(x)} \boldsymbol{\mu}(x,y) = \sum\nolimits_y p(y)\, \boldsymbol{\mu}(x,y). \tag{2.11}$$

The conditional expectation of $\boldsymbol{\mu}(x,y)$, given y such that $p(y) \neq 0$, is defined similarly.

The first order effect of the allele x is defined by the vector equation

$$\boldsymbol{\alpha}(x) = \boldsymbol{\mu}(x) - \boldsymbol{\mu} \tag{2.12}$$

for all $x \in \mathbb{A}$. Similarly, the first order effect of allele y is defined by

$$\boldsymbol{\alpha}(y) = \boldsymbol{\mu}(y) - \boldsymbol{\mu} \tag{2.13}$$

for all $y \in \mathbb{A}$. The second order effect of alleles x and y due to their interacting is defined by the vector equation

$$\boldsymbol{\alpha}(x,y) = \boldsymbol{\mu}(x,y) - \boldsymbol{\mu} - \boldsymbol{\alpha}(x) - \boldsymbol{\alpha}(y) \tag{2.14}$$

for all genotypes $(x,y) \in \mathbb{G}$. Equivalently,

$$\boldsymbol{\mu}(x,y) = \boldsymbol{\mu} + \boldsymbol{\alpha}(x) + \boldsymbol{\alpha}(y) + \boldsymbol{\alpha}(x,y) \tag{2.15}$$

for all genotypes $(x,y) \in \mathbb{G}$. This equation reminds one of an analysis of variance model of a multivariate experimental design with two factors. If $\boldsymbol{\alpha}(x,y) = \mathbf{0}$, the zero vector, for all genotypes $(x,y) \in \mathbb{G}$, then the alleles x and y act additively with respect to the k quantitative measurements. But, if $\boldsymbol{\alpha}(x,y) \neq \mathbf{0}$ for some genotype (x,y), then there is interaction among the alleles.

In a similar fashion, with respect to the random vector $\boldsymbol{W}$ with k phenotypic measurements for genotypes (x,y), it can be seen that the vector equation

$$\boldsymbol{W} = \boldsymbol{\mu} + (\boldsymbol{\mu}(x,y) - \boldsymbol{\mu}) + (\boldsymbol{W} - \boldsymbol{\mu}(x,y)) \tag{2.16}$$

is valid all genotypes $(x,y) \in \mathbb{G}$. By definition, the total phenotypic covariance matrix with respect to k quantitative traits in the population is

$$\boldsymbol{cov}_P[\boldsymbol{W}] = \sum\nolimits_{(x,y)} p(x,y)\, E\left[(\boldsymbol{W} - \boldsymbol{\mu})(\boldsymbol{W} - \boldsymbol{\mu})^T \mid (x,y)\right], \tag{2.17}$$

where the conditional expectation is taken with respect to the multivariate conditional density $f(\boldsymbol{w} \mid (x,y))$ for each genotype (x,y). The covariance matrix

$$\boldsymbol{cov}_G[\boldsymbol{W}] = \sum\nolimits_{(x,y)} p(x,y)\, (\boldsymbol{\mu}(x,y) - \boldsymbol{\mu})(\boldsymbol{\mu}(x,y) - \boldsymbol{\mu})^T, \tag{2.18}$$

measuring the covariation of the mean vectors $\boldsymbol{\mu}(x,y)$ for genotypes governing $k \geq 2$ traits around the mean vector $\boldsymbol{\mu}$ for the population, is called the genetic covariance matrix. Finally, the covariance matrix

$$cov_E[\boldsymbol{W}] = \boldsymbol{\Psi} = \sum\nolimits_{(x,y)} p(x,y)\, E\left[(\boldsymbol{W} - \boldsymbol{\mu}(x,y))(\boldsymbol{W} - \boldsymbol{\mu}(x,y))^T \mid (x,y)\right] \tag{2.19}$$

is called the environmental covariance matrix in the context of Equation (2.16). Observe that it is the same matrix as that defined in Equation (2.6) .

From Equation (2.16) it follows that

$$\begin{aligned} & E\left[(\boldsymbol{W}-\boldsymbol{\mu})(\boldsymbol{W}-\boldsymbol{\mu})^T \mid (x,y)\right] \\ = \; & E\left[(\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})(\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})^T \mid (x,y)\right] + E\left[(\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})(\boldsymbol{W}-\boldsymbol{\mu}(x,y))^T \mid (x,y)\right] \\ & + E\left[(\boldsymbol{W}-\boldsymbol{\mu}(x,y))(\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})^T \mid (x,y)\right] + E\left[(\boldsymbol{W}-\boldsymbol{\mu}(x,y))(\boldsymbol{W}-\boldsymbol{\mu}(x,y))^T \mid (x,y)\right]. \end{aligned} \tag{2.20}$$

But,

$$\begin{aligned} & E\left[(\boldsymbol{W}-\boldsymbol{\mu}(x,y))(\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})^T \mid (x,y)\right] \\ = \; & E\left[(\boldsymbol{W}-\boldsymbol{\mu}(x,y)) \mid (x,y)\right](\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})^T \\ = \; & \mathbf{0}_{k\times 1}(\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})^T = \mathbf{0}_{k\times k}, \end{aligned} \tag{2.21}$$

where, as indicated, $\mathbf{0}_{k\times 1}$ is a $k\times 1$ of zeros and $\mathbf{0}_{k\times k}$ is a $k\times k$ matrix of zeros. By way of validating this last step in (2.21), let X and Y one dimensional random variables, taking values in the set of real numbers. Then, it is well known that $E[XY \mid X] = XE[Y \mid X]$. Moreover, it can be shown that if X and Y are matrices such that the product XY is well defined, then the equation $E[XY \mid X] = XE[Y \mid X]$ also is valid for matrices. Furthermore, it can be shown that $E[XY \mid Y] = E[X \mid Y]Y$ is also valid for matrices. It was this well known property that was used to justify the second expression in Equation (2.21). These properties were also be used to show that the third term on the right in Equation (2.20) is equal to $\mathbf{0}_{k\times k}$, a $k\times k$ matrix of zeros. If $\boldsymbol{f}(\boldsymbol{X})$ is a matrix valued function of a matrix $\boldsymbol{X}$, then it can also be shown that $E[\boldsymbol{f}(\boldsymbol{X}) \mid \boldsymbol{X}] = \boldsymbol{f}(\boldsymbol{X})$

Therefore, Equation (2.20) reduces to

$$\begin{aligned} & E\left[(\boldsymbol{W}(x,y)-\boldsymbol{\mu})(\boldsymbol{W}(x,y)-\boldsymbol{\mu})^T \mid (x,y)\right] \\ = \; & (\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})(\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})^T + E\left[(\boldsymbol{W}-\boldsymbol{\mu}(x,y))(\boldsymbol{W}-\boldsymbol{\mu}(x,y))^T \mid (x,y)\right]. \end{aligned} \tag{2.22}$$

By multiplying this equation by $p(x,y)$ and summing over all genotypes $(x,y) \in \mathbb{G}$, it follows that

$$\boldsymbol{cov}_P[\boldsymbol{W}] = \boldsymbol{cov}_G[\boldsymbol{W}] + \boldsymbol{cov}_E[\boldsymbol{W}]. \tag{2.23}$$

It is interesting to note that this equation was derived by using properties of conditional expectations and is free as to any assumptions that may be made about the distributions of the terms of the right in Equation (2.16). For example, in classical quantitative genetics, for either one trait or many with pleiotropic effects, it was often assumed that the genetic effects $(\boldsymbol{\mu}(x,y)-\boldsymbol{\mu})$ and environmental effects $(\boldsymbol{W}-\boldsymbol{\mu}(x,y))$ were distributed independently. But, as can be seen from the derivation of (2.23) just described, such assumptions are not necessary. It is also important to note that, even though equation was derived under the assumption that only one autosomal locus was under consideration, Equation (2.23) would also be valid under the assumption that two or more autosomal loci were under consideration, but the formal details of a proof this statement will not be given here.

As is well known, the principal diagonal elements of any covariance matrix are the variances of the elements of a random vector $\boldsymbol{W}$ under consideration. In particular, let $var_P[W_\nu]$ denote the phenotypic variance for trait ν in the random the random vector $\boldsymbol{W}$. Similarly, let $var_G[W_\nu]$ and $var_E[W_\nu]$ denote, respectively, the genetic and environmental variances of trait $\nu = 1,2,\cdots,k$. Then, from (2.23), it follows that

$$var_P[W_\nu] = var_G[W_\nu] + var_E[W_\nu] \tag{2.24}$$

for every trait ν. A measure of heritability of a trait that has used extensively by many investigators is the ratio

$$H_\nu = \frac{var_G[W_\nu]}{var_G[W_\nu] + var_E[W_\nu]} \tag{2.25}$$

for for every trait $\nu = 1,2,\cdots,k$. In an actual application involving the analysis of data based in the structure presented in this section, an investigator may wish to estimate the fraction H_ν for every trait $\nu = 1,2,\cdots,k$. Even though many investigators have used a ratio of form (2.25) to estimate heritability of a quantitative trait, the details in the calculations may vary among investigators. It is suggested that if a reader is interesting in pursuing

these details, the books cited in the introduction be consulted. Three recent papers containing applications of the concept of heritability are those of Yang et al. (2010), Zaitlen et al. (2013) and Price et al. (2011). It should also be mentioned that, even though only one autosomal locus has been under consideration is this section, Equations (2.23) and (2.24) are also be valid if two or more autosomal loci were under consideration. Hence, the methods outlined in this section are also valid for any combination of two or more autosomal loci.

3. Partitioning the Genetic Covariance Matrix Into Component Covariance Matrices for the Case of One Autosomal Locus

From Equation (2.15), it can be seen from Equation (2.16) that the $k \times 1$ vector expression $\mu(x,y) - \boldsymbol{\mu}$ may be expressed as the $k \times 1$ vector equation

$$\mu(x,y) - \boldsymbol{\mu} = \boldsymbol{\alpha}(x) + \boldsymbol{\alpha}(y) + \boldsymbol{\alpha}(x,y) \tag{3.1}$$

for every genotype $(x,y) \in \mathbb{G}$. From the definition of the vector $\boldsymbol{\alpha}(x)$ in Equation (2.12), it can be seen that its expectation with respect to the genotypic distribution $\mathbb{D}_{Geno}$ is

$$E_{\mathbb{D}_{Geno}}[\boldsymbol{\alpha}(x)] = \sum\nolimits_{(x,y)} p(x,y)\,\boldsymbol{\alpha}(x) = \sum\nolimits_{x} p(x)\,\boldsymbol{\alpha}(x) = \mathbf{0}, \tag{3.2}$$

a $k \times 1$ zero vector. Similarly, it can be shown that

$$E_{\mathbb{D}_{Geno}}[\boldsymbol{\alpha}(y)] = \mathbf{0} \tag{3.3}$$

and

$$E_{\mathbb{D}_{Geno}}[\boldsymbol{\alpha}(x,y)] = \mathbf{0}, \tag{3.4}$$

where $\mathbf{0}$ is $k \times 1$ vector of zeros.

By definition, the additive genetic covariance matrix is

$$\boldsymbol{cov}_A[\boldsymbol{W}] = E_{\mathbb{D}_{Geno}}\left[\left(\boldsymbol{\alpha}(x)(\boldsymbol{\alpha}(x))^T\right) + \left(\boldsymbol{\alpha}(y)(\boldsymbol{\alpha}(y))^T\right)\right] \tag{3.5}$$

and the intra-allelic interaction covariance matrix is defined as

$$\boldsymbol{cov}_{IAI}[\boldsymbol{W}] = E_{\mathbb{D}_{Geno}}\left[\left(\boldsymbol{\alpha}(x,y)(\boldsymbol{\alpha}(x,y))^T\right)\right]. \tag{3.6}$$

Observe that the matrices in (3.4) and (3.5) are of order $k \times k$.

If a population is not in a Hardy-Weinberg equilibrium, then it is necessary to define cross-covariance matrices. For example, the covariance matrix of the vectors $\boldsymbol{\alpha}(x)$ and $\boldsymbol{\alpha}(y)$ is defined by the matrix expression

$$E_{\mathbb{D}_{Geno}}\left[(\boldsymbol{\alpha}(x))(\boldsymbol{\alpha}(y))^T\right]. \tag{3.7}$$

The covariance matrices of the pairs of vectors $\boldsymbol{\alpha}(x)$ and $\boldsymbol{\alpha}(x,y)$ as well as $\boldsymbol{\alpha}(y)$ and $\boldsymbol{\alpha}(x,y)$ are defined similarly. If the population is in a Hardy-Weinberg equilibrium, then it can be shown that all cross covariance matrices are zero matrices, and it follows that the genetic covariance matrix may be partitioned into and additive and intra-allelic interaction covariance matrices. Thus, when a population is in a Hardy-Weinberg equilibrium, the $k \times k$ matrix equation

$$\boldsymbol{cov}_G[\boldsymbol{W}] = \boldsymbol{cov}_A[\boldsymbol{W}] + \boldsymbol{cov}_{IAI}[\boldsymbol{W}] \tag{3.8}$$

is valid.

If, however, the population is not in a Hardy-Weinberg equilibrium, then the analogue of Equation (3.8) is more complicated. In this case, it will be helpful to express Equation (3.8) in a more general form. To that end, consider the $3k \times 1$ column vector

$$\boldsymbol{\Phi}(x,y) = \begin{pmatrix} \boldsymbol{\alpha}(x) \\ \boldsymbol{\alpha}(y) \\ \boldsymbol{\alpha}(x,y) \end{pmatrix} \tag{3.9}$$

and let

$$\boldsymbol{\Psi}(x,y) = \boldsymbol{\Phi}(x,y)\,\boldsymbol{\Phi}^T(x,y) \tag{3.10}$$

denote a $3k \times 3k$ matrix of $k \times k$ sub-matrices on the principal diagonal and $k \times k$ cross-products matrices off the principal diagonal. The expectation $\boldsymbol{\Psi}_G$ of this matrix with respect to the genotypic distribution is defined by

$$\boldsymbol{\Psi}_G = \sum\nolimits_{(x,y)} p(x,y)\, \boldsymbol{\Psi}(x,y). \tag{3.11}$$

In general, when the population is not in a Hardy-Weinberg equilibrium, it will be helpful to represent the matrix $\boldsymbol{\Psi}_G$ in the succinct partitioned form

$$\boldsymbol{\Psi}_G = \begin{pmatrix} \boldsymbol{\Psi}_G(1,1) & \boldsymbol{\Psi}_G(1,2) & \boldsymbol{\Psi}_G(1,3) \\ \boldsymbol{\Psi}_G(2,1) & \boldsymbol{\Psi}_G(2,2) & \boldsymbol{\Psi}_G(2,3) \\ \boldsymbol{\Psi}_G(3,1) & \boldsymbol{\Psi}_G(3,2) & \boldsymbol{\Psi}_G(3,3) \end{pmatrix}, \tag{3.12}$$

where every sub-matrix is of order $k \times k$. In terms of this matrix, it follows from the vector in (3.9) that the additive covariance matrix $\boldsymbol{cov}_A[\boldsymbol{W}]$ in (3.8) is

$$\boldsymbol{cov}_A[\boldsymbol{W}] = \boldsymbol{\Psi}_G(1,1) + \boldsymbol{\Psi}_G(2,2), \tag{3.13}$$

and the intra-allelic interaction covariance matrix in $(3,9)$ is

$$\boldsymbol{cov}_{IAI}[\boldsymbol{W}] = \boldsymbol{\Psi}_G(3,3). \tag{3.14}$$

Let $k \times k$ matrix $\boldsymbol{R}_G[\boldsymbol{W}]$ denote the sum

$$\boldsymbol{R}_G[\boldsymbol{W}] = \sum\nolimits_{i \neq j} \boldsymbol{\Psi}_G(i,j). \tag{3.15}$$

Given these definitions, the desired extension of Equation (3.9) to the case a population is not in a Hardy-Weinberg equilibrium has the form

$$\boldsymbol{cov}_G[\boldsymbol{W}] = \boldsymbol{cov}_A[\boldsymbol{W}] + \boldsymbol{cov}_{IAI}[\boldsymbol{W}] + \boldsymbol{R}_G[\boldsymbol{W}]. \tag{3.16}$$

Let $\boldsymbol{X}$ denote any $k \times 1$ random vector with a covariance matrix $\boldsymbol{cov}[\boldsymbol{X}] = \left(cov_{ij}[\boldsymbol{X}]\right)$ such that all its elements are finite. If $i = j$, then $cov_{ii}[\boldsymbol{X}] = var[X_i]$, where X_i is the i-th component of the vector $\boldsymbol{X}$. In terms of this notation, the ν-th component on the principal diagonal is matrix Equation (3.16) has the form

$$var_G[W_\nu] = var_A[W_\nu] + var_{IAI}[W_\nu] + cov_G[W_\nu], \tag{3.17}$$

for $\nu = 1, 2, \cdots, k$, where $cov_G[W_\nu]$ is the element of the matrix $\boldsymbol{R}_G[\boldsymbol{W}]$ in row corresponding to position $\nu\nu$. For each trait $\nu = 1, 2, \cdots, k$, an investigator may wish to compute that ratio

$$r_A(\nu) = \frac{var_A[W_\nu]}{var_G[W_\nu]} \tag{3.18}$$

as a measure the contribution of the additive genetic variance $var_A[W_\nu]$ to the total genetic $var_G[W_\nu]$. Moreover, it is interesting to note that the ratio

$$r_{IAI}(\nu) = \frac{var_{IAI}[W_\nu]}{var_G[W_\nu]} \tag{3.19}$$

and may be interpreted as a measure of the contribution of the intra-allelic interaction variance to the total genetic variance for trait ν. An investigator may also be interested in computing the ratio

$$1 - r_A(\nu) - r_{IAI}(\nu) = \frac{cov_G[W_\nu]}{var_G[W_\nu]}$$

as a measure of departure from a Hardy-Weinberg equilibrium. Observe that if $cov_G[W_\nu]$ is close to zero, then a Hardy-Weinberg disequilibrium would have a minimal effect on estimating $r_A(\nu)$ or $r_{IAI}(\nu)$. But, if $\mid cov_G[W_\nu] \mid$ is relatively large, then the effect of a disequilibrium could be significant in estimating either of these ratios.

Correlation matrices may also be of interest to some investigators. For example, for the case of the genetic covariance matrix $\boldsymbol{cov}_G[\boldsymbol{W}]$ in (3.16), an investigator may wish to look at the correlation matrix derived from this covariance matrix. Let $\sigma^2_G(\nu) = var_G[W_\nu]$ denote the variance of trait ν and let $\sigma_G(\nu, \nu') = cov_G[W_\nu, W_{\nu'}]$ denote

the covariance of traits ν and ν' such that $\nu \neq \nu'$. Then, the genetic correlation coefficient of traits ν and ν' has the form

$$\rho_G(\nu, \nu') = \frac{cov_G[W_\nu, W_{\nu'}]}{\sigma_G(\nu)\sigma_G(\nu')} \tag{3.20}$$

If $\nu = \nu'$, then

$$\rho_G(\nu, \nu) = \frac{\sigma_G^2(\nu)}{\sigma_G^2(\nu)} = 1. \tag{3.21}$$

Then, the $k \times k$ genetic correlation has the form

$$\boldsymbol{corr}_G[\boldsymbol{W}] = (\rho_G(\nu, \nu')). \tag{3.22}$$

Observe that all the components of the principal diagonal of this matrix are 1. If an investigator were also interested in the correlation matrices corresponding to the covariance matrices $\boldsymbol{cov}_A[\boldsymbol{W}]$ and $\boldsymbol{cov}_{IAI}[\boldsymbol{W}]$ in (3.16), then the correlation matrices for these covariance matrices could also be computed. For all the matrices just mentioned, as well as those that may arise in subsequent sections of this paper, the value of correlation coefficients may be interpreted as measure of pleiotropic effects.

4. Estimation of Mean Genetic Vector and Covariance Matrices From Data for the Case of One Locus

It should be stated at the outset that one could assume that the multivariate $k \times 1$ random vectors $\boldsymbol{W}(x, y)$ under consideration for each genotype were distributed independently with multivariate normal distributions with expectation vector $\boldsymbol{\mu}(x, y)$ and covariance matrix $\boldsymbol{\Psi}(x, y)$ for all genotypes $(x, y) \in \mathbb{G}$. But, even though such assumptions may be a valuable in coming to grips with some of the problems that will be encountered in subsequent sections, particularly those involving multiple comparisons, a decision was made to minimize the formalism presented in this paper with the hope that a less involved notation would attract more readers with interests in statistical genetics and related disciplines. Thus, for the most part, the methods of estimation to be described in this section are extension of the method of moments, which is among the most primitive methods used is statistical estimation, but is, nevertheless, still effective in many situations in which problems of estimation of parameters arise. If, however, a reader is interested in pursuing theoretical treatments of multivariate statistics, it is suggested that the books by Anderson (1984) and Muirhead (1982) be consulted.

For each genotype $(x, y) \in \mathbb{G}$, suppose there are $n(x, y) \geq 2$ individuals of genotype (x, y) that are measured with respect to the k quantitative traits under consideration, and let the random $k \times 1$ vectors $\boldsymbol{W}_\nu(x, y)$ for $\nu = 1, 2, \cdots, n(x, y)$ denote the observed data for the $n(x, y)$ individuals of genotype (x, y). It will be assumed that these random vectors are distributed independently and that each vector has the same distribution as the phenotypic vector $\boldsymbol{W}(x, y)$ for individuals of genotype (x, y) defined in section 2. Then, the random vector

$$\widehat{\boldsymbol{\mu}}(x, y) = \frac{1}{n(x, y)} \sum\nolimits_{\nu=1}^{n(x,y)} \boldsymbol{W}_\nu(x, y) \tag{4.1}$$

is an estimator of the $k \times 1$ expectation vector $\boldsymbol{\mu}(x, y)$ defined in (2.3) for all genotypes $(x, y) \in \mathbb{G}$. Given the estimator $\widehat{\boldsymbol{\mu}}(x, y)$, it can be shown that the random matrix

$$\widehat{\boldsymbol{\Psi}}(x, y) = \frac{1}{n(x, y) - 1} \sum\nolimits_{\nu=1}^{n(x,y)} (\boldsymbol{W}_\nu(x, y) - \widehat{\boldsymbol{\mu}}(x, y))(\boldsymbol{W}_\nu(x, y) - \widehat{\boldsymbol{\mu}}(x, y))^T \tag{4.2}$$

is an estimator of the covariance matrix $\boldsymbol{\Psi}(x, y)$ defined in (2.4) for all genotypes $(x, y) \in \mathbb{G}$. It also is well known that both these estimators are unbiased in the sense that $E[\widehat{\boldsymbol{\mu}}(x, y) \mid (x, y)] = \boldsymbol{\mu}(x, y)$ and $E\left[\widehat{\boldsymbol{\Psi}}(x, y) \mid (x, y)\right] = \boldsymbol{\Psi}(x, y)$ for all genotypes $(x, y) \in \mathbb{G}$.

Let

$$n = \sum\nolimits_{(x,y)} n(x, y) \tag{4.3}$$

denote that total number of observed individuals. Then,

$$\widehat{p}(x, y) = \frac{n(x, y)}{n} \tag{4.4}$$

is an estimator of the frequency of genotype (x, y) in the population. If it is assumed that the observations $\{n(x, y) \mid (x, y) \in \mathbb{G})\}$ are a sample from a multinomial distribution with the genotypic distribution $\mathbb{D}_{Geno}$ in (2.1) as

its probabilities with sample size n, then

$$E\left[\widehat{p}\,(x,y)\right] = \frac{1}{n}E\left[n\,(x,y)\right] = \frac{1}{n}np\,(x,y) = p\,(x,y) \tag{4.5}$$

for all genotypes $(x,y) \in \mathbb{G}$ so that $\widehat{p}\,(x,y)$ is an unbiased estimator of $p\,(x,y)$ under these assumptions. From (2.5) it follows that the random variable

$$\widehat{\boldsymbol{\mu}} = \sum\nolimits_{(x,y)} \widehat{p}\,(x,y)\,\widehat{\boldsymbol{\mu}}\,(x,y) \tag{4.6}$$

is an estimator of the unconditional expectation vector $\boldsymbol{\mu}$ defined in (2.5) . Similarly, the random matrix

$$\widehat{\boldsymbol{\Psi}} = \sum\nolimits_{(x,y)} \widehat{p}\,(x,y)\,\widehat{\boldsymbol{\Psi}}\,(x,y) \tag{4.7}$$

is an estimator of the unconditional covariance matrix $\boldsymbol{\Psi}$ defined in (2.6).

Given the estimators described above, it follows that the covariance matrices on the right side of Equation (2.23) can be estimated. For example, from the definition of the environmental matrix in Equations (2.22) and (2.23) , it follows that

$$\widehat{\boldsymbol{cov}}_E\left[\boldsymbol{W}\right] = \widehat{\boldsymbol{\Psi}}_E = \widehat{\boldsymbol{\Psi}}, \tag{4.8}$$

see (4.7). Similarly, the random matrix

$$\widehat{\boldsymbol{cov}}_G\left[\boldsymbol{W}\right] = \sum\nolimits_{(x,y)} \widehat{p}\,(x,y)\left(\widehat{\boldsymbol{\mu}}\,(x,y) - \widehat{\boldsymbol{\mu}}\right)\left(\widehat{\boldsymbol{\mu}}\,(x,y) - \widehat{\boldsymbol{\mu}}\right)^T \tag{4.9}$$

is an estimator of the genetic matrix in (2.23) . From these results, it follows from Equation (2.23) that the random matrix

$$\widehat{\boldsymbol{cov}}_P\left[\boldsymbol{W}\right] = \widehat{\boldsymbol{cov}}_G\left[\boldsymbol{W}\right] + \widehat{\boldsymbol{cov}}_E\left[\boldsymbol{W}\right] \tag{4.10}$$

is an estimator of the phenotypic covariance matrix in (2.23). Given the estimates of covariance matrices on the right in (4.10) , an investigator may wish to estimate the measure of heritability of each trait, using the formula in Equation (2.25) .

Although the formal details will not be given here, it is easy to see that the effects defined in section 2 as well as the variance components defined in section 3 could be estimated directly without recourse to any analysis of variance procedure. This estimation procedure could be carried out either under the assumption that the population was in a Hardy-Weinberg equilibrium or was not in such an equilibrium. In either case, all the covariance matrices defined in section 3 could be estimated.

It is easy to see from (4.9) that the genetic covariance is symmetric, because

$$\left(\left(\widehat{\boldsymbol{\mu}}\,(x,y) - \widehat{\boldsymbol{\mu}}\right)\left(\widehat{\boldsymbol{\mu}}\,(x,y) - \widehat{\boldsymbol{\mu}}\right)^T\right)^T = \left(\widehat{\boldsymbol{\mu}}\,(x,y) - \widehat{\boldsymbol{\mu}}\right)\left(\widehat{\boldsymbol{\mu}}\,(x,y) - \widehat{\boldsymbol{\mu}}\right)^T \tag{4.11}$$

for every genotype $(x,y) \in \mathbb{G}$. Therefore, since the transpose of a sum of matrices is the sum of the transpose of each matrix in (4.9) in the sum, it follows that

$$\left(\widehat{\boldsymbol{cov}}_G\left[\boldsymbol{W}\right]\right)^T = \widehat{\boldsymbol{cov}}_G\left[\boldsymbol{W}\right] \tag{4.12}$$

so that $\widehat{\boldsymbol{cov}}_G\left[\boldsymbol{W}\right]$ is a symmetric matrix, and will thus have real eigenvalues. As suggested in section 3, suppose an investigator wished to estimate the genetic correlation matrix $\widehat{\boldsymbol{corr}}_G\left[\boldsymbol{W}\right]$ defined in (3.22) . As will be shown subsequently, if an estimate of genetic covariance matrix is positive definite, then the corresponding estimates of the genetic correlation matrix will be such that all correlation coefficients $\widehat{\rho}$ will have values in the open interval $(-1,1)$.

Recall that for $k \geq 2$, a $k \times k$ symmetric real matrix $\boldsymbol{A}$ is positive definite if, and only if, $\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w} > 0$ for all $k \times 1$ vectors $\boldsymbol{w} \in \mathbb{R}_k$ such that $\boldsymbol{w} \neq \boldsymbol{0}$, where $\mathbb{R}_k$ is the set of $k \times 1$ of vectors real numbers and $\boldsymbol{0}$ is the $k \times 1$ vector of zeroes. If for some vector $\boldsymbol{w} \in \mathbb{R}_k$, such that $\boldsymbol{w} \neq \boldsymbol{0}$ and $\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w} = 0$, then the matrix $\boldsymbol{A}$ is said to be positive semi-definite. From this definition, it also follows that any square sub-matrix $\boldsymbol{B}_\nu$ of $\boldsymbol{A}$ consisting of ν rows and ν columns is also positive definite for $\nu = 1, 2, \cdots, k-1$.

A positive definite matrix may be characterized in a number of ways. For example, it is well known that a matrix $\boldsymbol{A}$ is positive definite if, and only, its eigenvalues are all positive. If a reader is interested in finding a proof of this

statement as well as a treatment of positive semi-definite matrices, it is suggested that the key phrase, "positive definite matrices" be typed into an internet search engine, where a wealth of material on linear algebra may be found in brief but reliable accounts. Among the many examples on the internet is Wikipedia article with this title. It is interesting to observe that it follows from its construction that the covariance matrix $\widehat{\boldsymbol{corr}}_G[\boldsymbol{W}]$ is positive semi-definite. For example, for any $\boldsymbol{w} \in \mathbb{R}_k$ it can be seen from (4.11) that $\boldsymbol{w}^T\left(\widehat{\boldsymbol{\mu}}(x,y) - \widehat{\boldsymbol{\mu}}\right) = u(x,y) = \left(\widehat{\boldsymbol{\mu}}(x,y) - \widehat{\boldsymbol{\mu}}\right)^T \boldsymbol{w}$, where $u(x,y) \in \mathbb{R}_1$ for all genotypes $(x,y) \in \mathbb{G}$. Therefore,

$$\begin{aligned} \boldsymbol{w}^T \widehat{\boldsymbol{cov}}_G[\boldsymbol{W}]\,\boldsymbol{w} &= \sum_{(x,y)} \widehat{p}(x,y)\,\boldsymbol{w}^T\left(\widehat{\boldsymbol{\mu}}(x,y) - \widehat{\boldsymbol{\mu}}\right)\left(\widehat{\boldsymbol{\mu}}(x,y) - \widehat{\boldsymbol{\mu}}\right)^T \boldsymbol{w} \\ &= \sum_{(x,y)} \widehat{p}(x,y)\,u^2(x,y) \geq 0 \end{aligned} \tag{4.13}$$

for all $\boldsymbol{w} \in \mathbb{R}_k$.

At this juncture, it seems fitting to point out that a necessary property of any covariance matrix $\boldsymbol{A}$ is it must be positive semi-definite. To see that this statement is indeed true, let $\boldsymbol{W}$ be a $k \times 1$ column vector of random variables with values in $\mathbb{R}_k$ with covariance matrix $\boldsymbol{A}$ and let $\boldsymbol{a} \in \mathbb{R}_k$. Then, let $Z = \boldsymbol{a}^T\boldsymbol{W}$ denote a random variable taking values in $\mathbb{R}_1$. Then, from the definition of the variance of a random variable, it follows that $var[Z] \geq 0$ for all $\boldsymbol{a} \in \mathbb{R}_k$. But, it is well known that $var[Z] = \boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a} \geq 0$ so that the matrix $\boldsymbol{A}$ must be positive semi-definite. Because the estimate of the genetic covariance matrix will always be positive semi-definite when the estimation procedure outlined in this section is used, it has a definite advantage when compared with that used by Mode and Robinson (1959), which was carried out within a framework of an analysis of covariance table, because when using such procedures one could not, in general, guarantee that an estimated genetic covariance matrix was indeed always be positive semi-definite so on some occasions an investigator may get negative estimates of a variance component. If a reader is interested in the method of estimation just mentioned, it is suggested that the formal procedures accompanied Table 1 of the paper by Mode and Robinson just cited be consulted. It should also be mentioned that at time this work was done in the 1950s, the presumed set of loci as well as the alleles at each locus governing the quantitative genetics of the traits under consideration were treated abstractly and without any knowledge of their locations in the genome of a species under study in an experiment. Consequently, the estimation procedure suggested in this section was not even conceivable the during the 1950s.

But, as mentioned above, it would be desirable to know whether the estimated genetic covariance matrix is indeed positive definite so that all estimated correlation coefficients would lie in the open interval $(-1, 1)$. It is recommended, therefore, that if an investigator is interested in estimating the genetic correlation matrix and using it to draw genetic inferences about a population, a first step would be that of computing the eigenvalues of the estimated genetic covariance matrix to determine whether all is eigenvalues are positive. Many software packages contain programs for calculating the eigenvalues of a square symmetric matrix so that a software developer would have little difficulty in finding a program to compute the eigenvalues of a covariance matrix. It should also be mentioned that if one or more of these eigenvalues are near zero, then the genetic covariance matrix would be nearly singular.

Among the several ways of characterizing a real positive definite matrix is a procedure known as the Sylvester's criterion. For any symmetric $k \times k$ matrix $\boldsymbol{A} = \left(a_{ij}\right)$, define k determinants as follows:

$$D_1 = a_{11}, D_2 = \det\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \cdots, D_k = \det \boldsymbol{A}. \tag{4.14}$$

Then, the matrix $\boldsymbol{A}$ is positive definite if, and only if, $D_\nu > 0$ for all $\nu = 1, 2, \cdots, k$. From the computational point of view, the Sylvester criterion could also be used to check whether a square matrix is positive definite, but to carry out such a procedure a software developer would need to have access to a program based on vary efficient algorithms for finding the numerical value of the determinants in (4.13). As will be shown below, this criterion also has interesting implications for testing whether all estimated correlation coefficients $\widehat{\rho}$ belong to the open interval $(-1, 1)$.

For suppose, $\boldsymbol{A}$ is the estimate of the genetic covariance $\widehat{\boldsymbol{cov}}_G[\boldsymbol{W}]$ matrix in (4.12) and suppose it has been determined that the matrix is positive definite. To simplify the notation, represent the element of this matrix by $\widehat{cov}_G[\boldsymbol{W}] = \left(\sigma_{ij}\right)$. To further simplify the notation, the symbol $\widehat{\circ}$, indicating estimates of parameters are under consideration, has been omitted for the elements of the estimated genetic covariance matrix. In this matrix $\sigma_{ii} = \sigma_i^2$, the estimated variance of trait i for $i = 1, 2, \cdots, k$, and the covariance $\sigma_{ij} = \rho_{ij}\sigma_i\rho_j$, where $i \neq j$, ρ_{ij} is an estimate of the genetic correlation between traits i and j, and σ_i and σ_j are, respectively, the estimated genetic standard deviations for traits i and j. Because of symmetry, $\rho_{ij} = \rho_{ji}$ for all $i \neq j$. Because the estimated genetic covariance

has, by assumption, been shown to be positive definite, it follows the Sylvester's criterion will be satisfied. In this notation, the first two determinates in Sylvester's criterion are $D_1 = \sigma_1^2 > 0$ and

$$D_2 = \det\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix} = \sigma_1^2\sigma_2^2 - \rho_{12}^2\sigma_1^2\sigma_2^2 = \sigma_1^2\sigma_2^2\left(1-\rho_{12}^2\right) > 0. \tag{4.15}$$

But, because $\sigma_1^2\sigma_2^2 > 0$, this equation, implies that $1 - \rho_{12}^2 > 0$. Thus, $\rho_{12}^2 < 1$, which implies $-1 < \rho_{12} < 1$. By preceding systematically in this way by choosing 2×2 sub-matrices corresponding to two rows and two columns, it can be shown the all estimated correlation coefficients $\widehat{\rho}_{ij}$ for all $i \neq j$ satisfy the condition $-1 < \widehat{\rho}_{ij} < 1$. For example, the 2×2 sub-matrix corresponding to rows and columns 1 and 3 of a $k \times k$ positive definite matrix has the form

$$\begin{bmatrix} \sigma_1^2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{31}\sigma_3\sigma_1 & \sigma_3^2 \end{bmatrix}, \tag{4.16}$$

which has the same form as the matrix in (4.14), which implies that

$$-1 < \widehat{\rho}_{13} < 1. \tag{4.17}$$

5. Measures of Pleiotropism and Epistasis for the Case of Two Autosomal Loci

Let $\mathbb{A}_1$ denote the set of alleles at locus 1 and let $\mathbb{A}_2$ denote the set of alleles at locus 2. It will be assumed that each of these sets contains at least two alleles but the total number of alleles at each locus is finite. Any genotype will be denoted by a symbol of the form (x_1, y_1, x_2, y_2), where the subscript, 1 or 2, denotes and locus, the symbols x and y denote alleles contributed the maternal and paternal parent, respectively. In order to simplify the notation, a genotype will often be denoted by the single letter symbol $z = (x_1, y_1, x_2, y_2)$. Let $\mathbb{G}$ denote the set of genotypes under consideration. The number of genotypes in the set $\mathbb{G}$ may be quite large, particularly if there are several alleles at each locus. Because the set of genotypes contains all possible combinations of alleles at each locus, the set $\mathbb{G}$ may be represented as the product set

$$\mathbb{G} = \mathbb{A}_1 \times \mathbb{A}_1 \times \mathbb{A}_2 \times \mathbb{A}_2. \tag{5.1}$$

For every genotype $z \in \mathbb{G}$, let $p(z)$ denote the frequency of genotype z in the population. Then, $p(z) \geq 0$ for all $z \in \mathbb{G}$ and

$$\sum\nolimits_{z\in\mathbb{G}} p(z) = 1. \tag{5.2}$$

Just as in the case on one locus, the genotypic distribution will be denoted by

$$\mathbb{D}_{Geno} = \{p(z) \mid z \in \mathbb{G}\}. \tag{5.3}$$

As in the foregoing sections, to accommodate pleiotropism it will also be supposed that $k \geq 2$ traits are under consideration and that the random $k \times 1$ phenotypic vector $\boldsymbol{W}$ along with its distribution characterize the observed variation with respect to k traits in a population or sample that is under consideration. Just as in section 2, the $k \times 1$ conditional expectation or genetic vector

$$\boldsymbol{\mu}(z) = E[\boldsymbol{W} \mid z] \tag{5.4}$$

along with the conditional covariance matrix

$$\boldsymbol{\Psi}(z) = E\left[(\boldsymbol{W} - \boldsymbol{\mu}(z))(\boldsymbol{W} - \boldsymbol{\mu}(z))^T \mid z\right], \tag{5.5}$$

which are defined for all genotypes $z \in \mathbb{G}$, will play essential roles in this section.

In section 3, where only one autosomal locus was under consideration, an effect was introduced that was a measure of interactions of alleles at one locus. When two of more autosomal loci are under consideration, however, interactions among alleles at different loci will need to accommodated in a model that will form the basis for a statistical analysis of data gathered in an experiment devoted to quantitative genetics. In classical genetics, interactions among alleles at different loci are referred to as epistasis. Briefly, the primary focus of attention in this section is to extend the results in section 3 by defining effects, which are functions of the genetic expectations in (5.4), that are measures of epistasis as well as other interactions among sets of alleles. Just as in the foregoing sections, the unconditional genetic expectation vector

$$\boldsymbol{\mu} = \sum\nolimits_{z\in\mathbb{G}} p(z)\boldsymbol{\mu}(z) \tag{5.6}$$

and unconditional covariance matrix

$$\mathbf{\Psi} = \sum\nolimits_{z \in \mathbb{G}} p(z)\, \mathbf{\Psi}(z) \tag{5.7}$$

will play roles in what follows.

As a first step in defining effects for the case of two autosomal loci, it will be necessary to represent the genotypic distribution in a more explicit form. For every genotype $z = (x_1, y_1, x_2, y_2) \in \mathbb{G}$, let $p(x_1, y_1, x_2, y_2)$ denote its frequency in the population. Then, by definition,

$$p(x_1) = \sum\nolimits_{(y_1, x_2, y_2)} p(x_1, y_1, x_2, y_2) \tag{5.8}$$

is the marginal frequency distribution of alleles $x_1 \in \mathbb{A}_1$ in the population. The marginal frequencies of alleles y_1, x_2 and y_2 are defined defined similarly. To lighten the notation, no subscripts, such as $p_1(x_1)$, will be attached to marginal distributions, because the subscript on each allele will denote the locus under consideration. Let $p(y_1)$, $p(x_2)$ and $p(y_2)$ denote, respectively, the marginal frequencies of alleles y_1, x_2 and y_2. For the case of two autosomal loci, a population is said to be in linkage equilibrium if

$$p(x_1, y_1, x_2, y_2) = p(x_1)\, p(y_1)\, p(x_2)\, p(y_2) \tag{5.9}$$

for all genotypes $z = (x_1, y_1, x_2, y_2) \in \mathbb{G}$.

One would not, in general, expect that a population would be in linkage equilibrium for the case under consideration so that it becomes necessary to define $k \times 1$ vectors of effects using conditional distributions. For example, the conditional distribution of the alleles (y_1, x_2, y_2), given x_1, is

$$p(y_1, x_2, y_2 \mid x_1) = \frac{p(x_1, y_1, x_2, y_2)}{p(x_1)} \tag{5.10}$$

for $p(x_1) \neq 0$. It is easy to see that if (5.9) is satisfied, it follows that

$$p(y_1, x_2, y_2 \mid x_1) = p(y_1)\, p(x_2)\, p(y_2) \tag{5.11}$$

for all triples of alleles $(y_1, x_2, y_2) \in \mathbb{A}_1 \times \mathbb{A}_2 \times \mathbb{A}_2$. Let the $k \times 1$ vector $\boldsymbol{\mu}(x_1)$ denote conditional expectation

$$\boldsymbol{\mu}(x_1) = \sum\nolimits_{(y_1, x_2, y_2)} \boldsymbol{\mu}(x_1, y_1, x_2, y_2)\, p(y_1, x_2, y_2 \mid x_1) \tag{5.12}$$

for all $x_1 \in \mathbb{A}_1$. Then, just as in the one autosomal locus case, the vector $\boldsymbol{\alpha}(x_1)$ of effects for allele x_1 in the population is defined by

$$\boldsymbol{\alpha}(x_1) = \boldsymbol{\mu}(x_1) - \boldsymbol{\mu} \tag{5.13}$$

for all $x_1 \in \mathbb{A}_1$. Observe that

$$E_{\mathbb{D}_{Geno}}[\boldsymbol{\alpha}(x_1)] = \sum\nolimits_{x_1} p(x_1)\, \boldsymbol{\alpha}(x_1) = \mathbf{0}, \tag{5.14}$$

a $k \times 1$ vector of zeroes. The first order effects just defined can easily be extended to define the $k \times 1$ vectors $\boldsymbol{\alpha}(y_1)$, $\boldsymbol{\alpha}(x_2)$ and $\boldsymbol{\alpha}(y_2)$, for all alleles $y_1 \in \mathbb{A}_1$ and all pairs $(x_2, y_2) \in \mathbb{A}_2 \times \mathbb{A}_2$.

For the case of two autosomal loci, it is possible to define many more effects than for the case of one autosomal locus. To provide a framework for defining and classifying these effects, it will be helpful to consider a set $\mathfrak{S} = (1, 2, 3, 4)$ of four positions that are occupied by alleles at the two loci under consideration. For example, for any genotype, positions 1 and 2 are occupied by alleles at locus 1, and positions 3 and 4 are occupied by alleles at locus 2. To get a grasp of how many effects that can be defined for the case under consideration, it is helpful to think of the class $\mathfrak{T}$ of all subsets of the set $\mathfrak{S}$. Among the sets in $\mathfrak{T}$ is φ, the empty set. To provide a means for describing effects that are measures of interactions of alleles at one locus or effects that are measures of epistatic interaction among alleles at two or more loci, it is useful to enumerate classes of subsets the the class $\mathfrak{T}$. For $\nu = 0, 1, 2, 3, 4$, let $\mathfrak{T}_\nu$ denote the class of subset containing ν positions.

Thus, the class $\mathfrak{T}_0 = \{\varphi\}$ contains only the empty set, but the class $\mathfrak{T}_1$ consists of the singletons

$$\mathfrak{T}_1 = \{(1), (2), (3), (4)\}, \tag{5.15}$$

which were used to define the first order effects listed above. From elementary combinatorics, it is easy to see that the number of sets in the class $\mathfrak{T}_2$ is

$$\binom{4}{2} = 6.$$

In particular, the subsets in the class $\mathfrak{T}_2$ are

$$\mathfrak{T}_2 = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\} \tag{5.16}$$

Observe that the subclass of sets of positions

$$\mathfrak{T}_{2IAI} = \{(1,2),(3,4)\} \tag{5.17}$$

will form a basis for defining second order effects that are measures of intra-allelic interactions at loci 1 and 2, respectively. But the subclass of sets

$$\mathfrak{T}_{2EPI} = \{(1,3),(1,4),(2,3),(2,4)\} \tag{5.18}$$

provide a basis for defining second order epistatic effects among the two loci under consideration. There are

$$\binom{4}{3} = 4$$

subsets in the class $\mathfrak{T}_3$, which consists of the subsets

$$\mathfrak{T}_3 = \{(1,2,3),(1,2,4),(1,3,4),(2,3,4)\}. \tag{5.19}$$

As will be shown subsequently, the third order effects corresponding to the set in this class provide a basis for defining various types of effects that are measures of epistatic interactions. Finally, the class of sets

$$\mathfrak{T}_4 = \{\mathfrak{S} = (1,2,3,4)\} \tag{5.20}$$

contains all positions and provides a basis for defining a fourth order effects that are measures of epistatic and intralocus interactions among any set of four alleles at the two loci under consideration.

From the foregoing discussion, it can be seen that a total of 15 effects may be defined for the case of two autosomal loci, and to derive a formula for computing each effect, it would be necessary to consider 15 conditional probability distributions. Rather than deriving a formula for each of these 15 subsets of $\mathfrak{S}$, it will be helpful to set down general formulas for marginal and conditional distributions. Let

$$\mathfrak{E} = \bigcup_{\nu=1}^{3} \mathfrak{T}_\nu \tag{5.21}$$

denote the union of the classes of subsets of $\mathfrak{S}$ corresponding to effects under consideration. For every set $A \in \mathfrak{E}$, let A^c denote its complement with respect to $\mathfrak{S}$, and let $z(A)$ and $z(A^c)$ denotes the sets of alleles in genotype z corresponding, respectively, to the positions in the sets A and A^c. For any genotype $z \in \mathbb{G}$, $z = z((A), z(A^c))$, where, by definition, the positions in the sets A are fixed in the operations that follow. For example, the formula denotes

$$p(z(A)) = \sum\nolimits_{z(A^c)} p(z(A), p(A^c)) \tag{5.22}$$

the marginal distribution of the alleles in the position of the set A, where the sum runs over all alleles in the positions of the set A^c. Thus, in this succinct notation

$$p(z(A^c) \mid z(A)) = \frac{p(z(A), p(A^c))}{p(z(A))} \tag{5.23}$$

is the conditional distribution of the alleles in the positions in set A^c, given the fixed alleles in the positions of the set A such that $p(z(A)) \neq 0$. Let $\boldsymbol{\mu}(z(A))$ denote the conditional expectation of the vector $\boldsymbol{\mu}(z)$, given the fixed alleles in the positions corresponding to the set A. Then, for every $A \in \mathfrak{E}$ let

$$\boldsymbol{\mu}(z(A)) = \sum\nolimits_{z(A^c)} p(z(A^c) \mid z(A)) \boldsymbol{\mu}(z((A), z(A^c))) \tag{5.24}$$

where the sum runs over all the alleles in the positions of the set A^c.

To illustrate using formula (5.24), consider those sets $A \in \mathfrak{T}_1$. If $A = \{1\}$, then $\boldsymbol{\mu}(z(A)) = \boldsymbol{\mu}(x_1)$ so that $\boldsymbol{\alpha}(x_1) = \boldsymbol{\mu}(x_1) - \boldsymbol{\mu}$ could be estimated for for all alleles $x_1 \in \mathbb{A}_1$. The remaining first order effects, $\boldsymbol{\alpha}(y_1)$, $\boldsymbol{\alpha}(x_2)$ and $\boldsymbol{\alpha}(y_2)$ could also be derived and estimated in a similar way, using formula (5.24). The number of sets in the class $\mathfrak{T}_2$ is 6 so that there are 6 conditional expectations of the form $\boldsymbol{\mu}(A)$ for $A \in \mathfrak{T}_2$. Suppose, for example, $A = \{x_1, y_1\}$. Then $\boldsymbol{\mu}(A) = \boldsymbol{\mu}(x_1, y_1)$, and, just as in the case of one autosomal locus, the effect $\boldsymbol{\alpha}(x_1, y_1)$ would be defined as

$$\boldsymbol{\alpha}(x_1, y_1) = \boldsymbol{\mu}(x_1, y_1) - \boldsymbol{\mu} - \boldsymbol{\alpha}(x_1) - \boldsymbol{\alpha}(y_1) \tag{5.25}$$

for all pairs of alleles $(x_1, y_1) \in \mathbb{A}_1 \times \mathbb{A}_1$. Observe this effect is a measure of intralocus interactions of alleles. If the alleles at locus 1 acted in a purely additive manner, one would expect that $\boldsymbol{\alpha}(x_1, y_1)$ would be small, particularly if $x_1 = y_1$ were the same allele. But, if $x_1 \neq y_1$, then there may be intra-allelic interaction between the two alleles so that this effect may be greater than that for homozygous genotypes. If $A = \{x_1, x_2\}$, then the effect

$$\boldsymbol{\alpha}(x_1, x_2) = \boldsymbol{\mu}(x_1, x_2) - \boldsymbol{\mu} - \boldsymbol{\alpha}(x_1) - \boldsymbol{\alpha}(x_2) \tag{5.26}$$

is a measure of interloci or epistatic interactions for all pairs of alleles $(x_1, x_2) \in \mathbb{A}_1 \times \mathbb{A}_2$. It is of interest to note that among the four remaining effects for sets $A \in \mathfrak{T}_2$, three would be measures on interloci interactions and one would be an effect similar to that in (5.25) corresponding to positions $\{3, 4\}$.

For every set $A \in \mathfrak{T}_3$, there corresponds and effect $\boldsymbol{\alpha}(z(A))$ and, as can be seen from (5.19), there are four subsets in $\mathfrak{T}_3$. To illustrate the procedure used to define each of these effects, suppose $A = \{1, 2, 3\}$. Then, $\boldsymbol{\mu}(x_1, y_1, x_2)$ is the conditional expectation that needs to be derived, using formula (5.24). Briefly, for each subset B of A, such that $A \neq B$, there will be an effect that is used in defining the effect $\boldsymbol{\alpha}(z(A))$. In particular,

$$\begin{aligned} \boldsymbol{\alpha}(x_1, y_1, x_2) &= \boldsymbol{\mu}(x_1, y_1, x_2) - \boldsymbol{\alpha}(x_1) - \boldsymbol{\alpha}(y_1) - \boldsymbol{\alpha}(x_2) \\ &\quad -\boldsymbol{\alpha}(x_1, y_1) - \boldsymbol{\alpha}(x_1, x_2) - \boldsymbol{\alpha}(y_1, x_2) \end{aligned} \tag{5.27}$$

for all triples $(x_1, y_1, x_2) \in \mathbb{A}_1 \times \mathbb{A}_1 \times \mathbb{A}_2$. The procedure illustrated in (5.27) may also be used to derive an expression for $\boldsymbol{\alpha}(z(A))$ for any set $A \in \mathfrak{T}_3$ such that $A \neq \{1, 2, 3\}$.

The last effect that needs to be defined is that for the class

$$\mathfrak{T}_4 = \{\mathfrak{S} = (1, 2, 3, 4)\}.$$

Let $\alpha(z)$, where $z = (x_1, y_1, x_2, y_2)$, denote this effect. Then, $\alpha(z)$ is determined by solving the equation

$$\boldsymbol{\mu}(z) = \boldsymbol{\mu} + \sum\nolimits_{A \in \mathfrak{T}_1} \boldsymbol{\alpha}(z(A)) + \sum\nolimits_{A \in \mathfrak{T}_2} \boldsymbol{\alpha}(z(A)) + \sum\nolimits_{A \in \mathfrak{T}_3} \boldsymbol{\alpha}(z(A)) + \boldsymbol{\alpha}(z) \tag{5.28}$$

for $\alpha(z)$ for all genotypes $z \in \mathbb{G}$. The number of solutions to this equation depends on the number of alleles at each locus. For example, if there are two alleles at each locus, then there are 16 possible values of $\boldsymbol{\alpha}(z) = \boldsymbol{\alpha}((x_1, y_1, x_2, y_2))$. Moreover, each solution is a $k \times 1$ column vector, corresponding to the $k \geq 2$ traits under consideration. Let $\alpha_\nu(x_1, y_1, x_2, y_2)$ denote the effect for trait $\nu = 1, 2, \cdots, k$ in the $k \times 1$ vector $\boldsymbol{\alpha}(z)$. An investigator may be interesting in estimating the component of variance

$$var[W_\nu] = \sum\nolimits_{z \in \mathbb{G}} p(z) \alpha_\nu^2(z) \tag{5.29}$$

for each $\nu = 1, 2, \cdots, k$. But, because this is a summary statistic, it may mask the most interesting effects, i.e. those with the largest values, making up this variance component. Furthermore, because each effect in the sum (5.29) has been estimated, it would be possible to inspect the numerical values of each term in the sum of Equation (5.29) for indications of usually large values that would be indicative of significant interactions among the alleles at the two loci under consideration. In the next section, a procedure for inspecting the numerical values of all the estimated effects will be suggested.

Before proceeding to the next section, it is interesting to note that Hemani et al. (2013) have provided an evolutionary perspective on epistasis and the missing heritability and have also suggested that genome-wide association studies would be improved by searching directly for epistatic effects. It seems plausible, therefore, that the measures of epistatic effects defined in this section as well those for multi-loci effects that will be introduced in a

subsequent section may provide a means for searching for epistatic effects not only in quantitative genetics but for also in measuring these effects in genome-wide association studies.

6. Searching for Unusual Effects and Interactions Among the Alleles at Two Autosomal Loci

The strategy for searching unusual effects and interactions at the two autosomal loci will be that of inspecting the squares of each effects and then picking out the largest of them as indicators of unusual effects of alleles or interactions among the alleles at the two loci. It will be assumed that $k \geq 2$ traits are under consideration, but to simplify the notation each set of effects will not be distinguished by the subscript ν, indicating the trait, but it will be tacitly be assumed that any search procedure would be carried out for each trait. Let

$$\mathfrak{E}_1 = \left\{ \alpha^2 (z(A)) \mid A \in \mathfrak{T}_1 \right\} \tag{6.1}$$

denote the set of squared first order effects. For the case there are two alleles at each locus, any position may be occupied by one of two alleles, and because there are four positions under consideration, there 8 effects in the set $\mathfrak{E}_1$. Consequently, in this case, an investigator could easily find the effect with the largest value by inspection, but in cases with a large number of effects, it would be necessary to write a computer program or use existing software to find largest of them.

The set of second order effects may be partitioned into two subsets, see (5.17) and (5.18). The set of squared effects for intra-allelic interactions is

$$\mathfrak{E}_{2IAI} = \left\{ \alpha^2 (z(A)) \mid A \in \mathfrak{T}_{2IAI} \right\}. \tag{6.2}$$

In this case each of the two sets in $A \in \mathfrak{T}_{2IAI}$ contains two positions and each position may be occupied by two alleles. Hence, the number of effects in the set $\mathfrak{E}_{2IAI}$ is 8. The set of squared effects for second order epistatic interactions is

$$\mathfrak{E}_{2EPI} = \left\{ \alpha^2 (z(A)) \mid A \in \mathfrak{T}_{2EPI} \right\}. \tag{6.3}$$

For this case, there are 4 sets of two positions in the set $\mathfrak{T}_{2EPI}$, and each position occupied by two alleles. Consequently, the number of effects in the set $\mathfrak{E}_{2EPI}$ is 16.

The set of squared effects for third order epistatic interactions is

$$\mathfrak{E}_{3EPI} = \left\{ \alpha^2 (z(A)) \mid A \in \mathfrak{T}_3 \right\}. \tag{6.4}$$

Each of the four sets in $\mathfrak{T}_3$ consists of 3 positions and each position may be occupied with one of two alleles. Consequently, for each set of three positions there will correspond $2^3 = 8$ effects and, because there of 4 sets in $\mathfrak{T}_3$, there will be a total of 32 effects in the set $\mathfrak{E}_{3EPI}$. Rather than relying on a visual inspection of 32 effects, in this case, as well as in cases in which 3 or more autosomal loci are under consideration, an investigator may prefer to write a computer program or use existing software to find the largest squared effect. On the other hand, an investigator may not want to focus on the largest effect in each case just enumerated, but have a computer order the squared effects from the smallest to the largest so that, for example, attention could be focused on the three largest squared effects. For the case of two autosomal loci, the set of squared fourth order effects is

$$\mathfrak{E}_{4EPI} = \left\{ \alpha^2 (z(A)) \mid A \in \mathfrak{T}_4 = \{1, 2, 3, 4\} \right\}. \tag{6.5}$$

As mentioned in section 5, the number of squared effects in this set is 16 so that in this case an investigator may wish to select the largest one in a search for unusual epistatic interactions of fourth order.

If an investigator were interested in estimating a components of the total genetic variance for any trait, it would be possible to do so for any of the sets of squared effects listed above. For example, suppose attention was focused on sum of the variance components corresponding to the squared effects in the set $\mathfrak{E}_{3EPI}$. For every $A \in \mathfrak{T}_3$, let $p(z(A))$ denote the corresponding marginal probability. Then, for the trait $\nu = 1, 2, \ldots k$ under consideration, the component of the genetic variance corresponding to the set $\mathfrak{E}_{3EPI}$ is defined by

$$var_{\mathfrak{E}_{3EPI}} [W_\nu] = \sum\nolimits_{A \in \mathfrak{T}_3} p(z(A)) \alpha^2 (z(A)), \tag{6.6}$$

and the variance components corresponding to the other sets listed above could be estimated similarly. On the other hand, one could proceed as in the one locus case discussed in section and define a $15k \times 1$ column vector $\mathbf{\Phi}(x, y)$ as in (3.9) and a $15k \times 15k$ covariance matrix $\mathbf{\Psi}_G$ as in (3.11), but the formal details of this derivation will

be left to the reader as an exercise. This covariance matrix may be particularly interesting when the sample or population is not in linkage equilibrium at the two loci under consideration.

At this point in the discussion, it is appropriate to mention the pioneering paper by Cockerham (1954), who was among the first to consider variance components for epistatic interactions in quantitative genetics and introduced a distinctive nomenclature for such interactions. In his terminology but in a different notation, the genetic variance component

$$var_A[W_\nu] = \sum\nolimits_{B\in\mathfrak{T}_1} p(z(B))\,\alpha^2(z(A)), \tag{6.7}$$

corresponding to the set $\mathfrak{T}_1$, would be called the additive component with subscript (A) for trait $\nu = 1, 2, \cdots, k$. The component of variance $var_D[W_\nu]$, corresponding to the set $\mathfrak{E}_{2IAI}$ would be called the dominance component with subscript (D), and the variance component $var_{AA}[W_\nu]$ corresponding to the set $\mathfrak{E}_{2EPI}$ would be called the additive by additive component with subscript (AA). Similarly, the variance components corresponding to the set $\mathfrak{E}_{3EPI}$ would be designated as the additive by dominant component, with subscript (AD), Finally, that corresponding to the set $\mathfrak{E}_{4EPI}$ would be labeled the dominant by dominant, with subscript (DD), component of the genetic variance for any trait $\nu = 1, 2, \cdots, k$. It should also be mentioned that Equation (3.16) , representing a partition of the genetic covariance matrix into component matrices for the case of one autosomal locus, could also generalized to the case to two autosomal loci using the set-based classification of effects outlined in this section, but this matrix partition will also be left to the reader as an exercise.

As will be discussed in the next section, when many loci are under consideration, the potential number of effects that may be considered can be very large. In such cases, to find unusually large effects and interactions even for the relatively simple case of only two alleles per locus may entail searches of hundreds of thousands or even millions of squared effects along with testing for their statistical significance. As has been widely recognized, when large numbers of statistical tests are considered, there is a risk of false discovery rates. It is beyond the scope of this paper to discuss the technicalities underlying false discovery rates, but it is suggested that an interested reader consult the papers Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001, 2005).

7. Overview of Cases for $l > 2$ Autosomal Loci

Let $l > 2$ denote the number of autosomal loci under consideration in a some diploid species such as man. Then, the number of positions that may be occupied by alleles at each locus is $N = 2l$. Let $\mathfrak{S}$ the set of N positions numbered $1, 2, \cdots, N$, let $\mathfrak{T}$ denote the class of all subsets of $\mathfrak{S}$ and let $\mathfrak{T}_\nu$ denote that class of subsets containing ν positions for $\nu = 0, 1, 2, \cdots, N$. Then, as is well known, the number of subsets in the class $\mathfrak{T}$ is 2^N and the number of subsets in the class $\mathfrak{T}_\nu$ is

$$\binom{N}{\nu} \tag{7.1}$$

for $\nu = 0, 1, 2, \cdots, N$. From the point of view of judging the feasibility of defining effects based on $\nu \geq 1$ positions, this number is essential, because it gives us the number of effects that may be defined. In what follows, it will also be helpful to recall the well known equation from combinatorics,

$$\sum\nolimits_{\nu=0}^{N} \binom{N}{\nu} = 2^N, \tag{7.2}$$

that is sometimes listed in text books on introductory probability theory. Because, no effect is associated with the empty set φ, which is the only member of the class $\mathfrak{T}_0$ and $\binom{N}{0} = 1$, it follows that for the case of l autosomal loci, the number of effects that may be defined is $2^N - 1$.

By way of an illustrative example, suppose the number of autosomal loci under consideration is $l = 4$ so that $N = 8$. Thus, in this case it would be possible to define

$$2^8 - 1 = 255 \tag{7.3}$$

effects. It is also interesting to note that if there were 2 alleles per locus, then the number of possible genotypes would be $2^8 = 256$. If an investigator had access to a large sample of individuals whose genomes had be sequenced, it may be feasible to consider this many possible genotypes. But the study of 256 genotypes in small samples would be problematic, because even in a sample of size $n = 256$, all possible genotypes may be not represented and in rare samples there may only one observation for each genotype. If a sample were sufficiently large, an investigator

may undertake a four autosomal loci study, but as a first approximation, a decision may be made of to define and estimate only first, second and third order effects rather dealing all the 255 possible effects.

It is clear that the number of sets in the class $\mathfrak{T}_1$ is 8 so that 8 first order effects could be defined and estimated, using the principles outlined in section 5 for the case of two loci. The number of second order effects would be

$$\binom{8}{2} = 28 \tag{7.4}$$

and the number of third order effects would be

$$\binom{8}{3} = 56. \tag{7.5}$$

If an investigator were skilled in writing computer code, it would be possible to write programs to enumerate the sets in the classes $\mathfrak{T}_2$ and $\mathfrak{T}_3$ as well as to compute the effects associated with each class. Ideally, it would be close to optimal if a team of investigators were working on the project. At a minimum, a team should consists of at least two individuals: one individual should have expertise in genetics and managing data bases consisting of individuals whose genomes have been sequenced and a second individual would expertise in writing software. It is interesting to note that with respect to the class $\mathfrak{T}_2$, there would be 4 sets of two positions such that effects that were measures of intra-allelic interactions could be estimated, but the remaining 24 sets would from a basis for defining and estimating effects that are measures of epistatic interactions among the alleles at four loci taken two at a time. All 56 effects corresponding to subsets in the class $\mathfrak{T}_3$ could be interpreted as measures of epistatic interactions among the alleles at the four loci under consideration. In order to include the phenomenon of pleiotropism in the formulation, it will be assumed, as in previous sections, that all effects are $k \times 1$ column vectors, where $k \geq 2$.

To provide a succinct overview of considering only first, second and third order effects in a variance component model, it will again to helpful to let z denote a genotype in the set $\mathbb{G}$ of all possible genotypes, and let $z(A)$ denote a set of alleles corresponding to the positions in any set $A \in \mathfrak{T}$. Next suppose that all the effects in the classes

$$\{\boldsymbol{\alpha}(z(A)) \mid A \in \mathfrak{T}_\nu\} \tag{7.6}$$

have been defined and estimated for $\nu = 1, 2, 3$, using the principles outlined in section 5. Then, the linear model expressing a genetic value $\boldsymbol{\mu}(z)$ as a function of effects would have the form

$$\boldsymbol{\mu}(z) = \boldsymbol{\mu} + \sum\nolimits_{A \in \mathfrak{T}_1} \boldsymbol{\alpha}(z(A)) + \sum\nolimits_{A \in \mathfrak{T}_2} \boldsymbol{\alpha}(z(A)) + \sum\nolimits_{A \in \mathfrak{T}_3} \boldsymbol{\alpha}(z(A)) + \boldsymbol{\alpha}_R(z), \tag{7.7}$$

where the remainder effect is determined by solving Equation (7.7) for $\boldsymbol{\alpha}_R(z)$ for every genotype $z \in \mathbb{G}$. For any trait ν in the vector valued terms in this equation, an investigator could search for usually large squared effects following the search procedures suggested in section 6 with a goal of finding epistatic interactions that would be of interesting for interpreting the data. In this connection, and difference in these effects among the k traits under consideration would attributable to pleiotropism.

To provide a measure of the adequacy of the approximation in (7.7) based on only first, second and third order effects is would be of interest to estimate the variances for each trait in vector of remainder effects $\boldsymbol{\alpha}_R(z)$. For example, let $\alpha_R^{(\nu)}(z)$ the element in this vector for trait ν for $\nu = 1, 2, \cdots, k$ and let

$$var_R[W_\nu] = \sum\nolimits_{z \in \mathbb{G}} p(z) \left(\alpha_R^{(\nu)}(z)\right)^2 \tag{7.8}$$

denote the estimated variance component corresponding to the remainder term in (7.7). Then, the ratio

$$\frac{var_R[W_\nu]}{var_G[W_\nu]}, \tag{7.9}$$

where $var_G[W_\nu] = \boldsymbol{cov}_{\nu\nu}[W_\nu]$ is element $\nu\nu$ on the principal diagonal of the genetic covariance matrix $\boldsymbol{cov}_G[\boldsymbol{W}]$ in (2, 23), is a measure of the contribution of the variance component in (7.8) to the total genetic variance for trait $\nu = 1, 2, \cdots, k$. In relative terms, small values of this fraction for each trait ν would be indicators of the goodness approximation in (7.7), using only first, second and third order effects.

At this point in the discussion of variance component models that are used in quantitative genetics, it is appropriate to mention that the number of genotypes under consideration may be significantly reduced if only three genotypes

are identified at any locus for the case of two alleles per locus. To demonstrate this idea, it is helpful to resort to methods for representing genotypes used in classical Mendelian genetics. Suppose at some locus there are to alleles B and b. Then, there are four genotypes BB, Bb, bB and aa. If, however, the heterozygotes Bb and bB are lumped into one class, then only three genotypes would be distinguishable at each locus. Thus, if this idea were used, the number of distinguishable genotypes with respect to four autosomal loci would be

$$3^4 = 81. \tag{7.10}$$

This number is significantly less than 256, which was the number of genotypes in which 8 positions were used in defining genotypes. If this classification of genotypes were used, the procedures for estimating effects would need to be modified by taking into account the lumping of heterozygotes into one class.

During the past five to ten years, a rather large number of papers have been published by members of the genetic community in which genome wide sweeps have been made of the human genome with goals of finding genomic regions that are implicated in such neurological conditions as Alzheimer's and Parkinson's diseases. In an interesting paper, Raj et al. (2012) considered 11 regions (loci) in the human genome that have been implicated with Alzheimer's disease and present evidence that four of these loci are involved in protein interaction network that has been maintained in the population by positive natural selection. It is also stated in this paper that 12 loci have been implicated in Parkinson's disease. It is of interest, therefore, to consider the extent the methodology presented in this paper may be applied to traits whose genetics are governed by 11 or 12 loci with two alleles per locus. If three genotypes were distinguished at each locus, then, for the case of 11 loci the number of genotypes that could be distinguished for 11 loci is

$$3^{11} = 177,147, \tag{7.11}$$

and for the case of 12 loci, this number is $531,441$. When it is required that the genomes of all individuals in a sample have been sequenced, it is doubtful at the present time whether of samples sizes of $n > 177,147$ or $n > 531,441$ would be available to an investigator or a team of investigators who are interested in applying the ideas presented in this paper.

Even though sample sizes this magnitude may not be available to an investigator, their research, however, could proceed by using the available data to detect epistatic and pleiotropic interactions among the alleles represented in the sample. For example, for the case of 11 loci, it is suggested that an investigator perform a preliminary survey of the data to estimate the number $n(\boldsymbol{x}, \boldsymbol{y})$ of individuals of genotype $(\boldsymbol{x}, \boldsymbol{y})$ are available, where the alleles in $\boldsymbol{x}$ and $\boldsymbol{y}$ have been ascertained with respect to 11 or fewer loci. If some number or numbers $n(\boldsymbol{x}, \boldsymbol{y})$ are too small to yield reliable statistical information, then an investigator may make a decision to restrict attention to a subset of the 11 loci such that the sample sizes for each genotype are judged sufficiently large to draw statistically reliable inferences from the data as to the presence of epistatic and pleiotropic interactions.

Acknowledgements

A word of thanks is due Dr. Towfique Raj, Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, who called the author's attention to recent papers devoted to the estimation of heritability of various quantitative traits in humans, which have been cited in this paper. A cooperative research effort involving the author and Dr. Raj's group is also in progress, with a goal of writing software to implement some of the ideas set forth in this paper, along with results not included in this paper, and applying them in a quantitative genetic analysis of data from samples of patients whose genomes have been sequenced.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed.). New York, Chichester, Brisbane, and Singapore: John Wiley and Sons Inc.

Benjamini, Y., & Hochberg, Y. (1995). Controlling false discovery rate: A practical & powerful approach to multiple testing. *J. Roy. Soc. Ser. B, 57*, 289-300.

Benjamini, Y. & Yekutieli, D. (2001). The control of false discovery rate under dependency. *Ann. Statist., 29*, 1165-1188. http://dx.doi.org/10.1214/aos/1013699998

Benjamini, Y. & Yekutieli, D. (2005). False discovery rate adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc., 100*, 71-93. http://dx.doi.org/10.1198/016214504000001907

Bulmer, M. G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford: Clarendon Press.

Church, G. M. (2006). Genomes For All. *Scientific American, 294*, 46-54. http://dx.doi.org/10.1038/scientificamerican0106-46

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

Falconer, D., & MacKay, T. F. C. (1996). *Introduction to Quantitative Genetics*. New York: Longman.

Fisher, R. A. (1918). The correlation among relatives on the assumption of Mendelian inheritance. *Trans. Royal Soc., Edinburgh, 52*, 399-433. http://dx.doi.org/10.1017/S0080456800012163

Hemani, G., Knott, S., & Haley, C. (2013). An Evolutionary Perspective on Epistasis & the Missing Heritability. *PLoS Genet, 9*(2), e1003295. http://dx.doi.org/10.1371/journal.pgen.1003295

Kao, C.-H., & Zeng, Z.-B. (2002). Modeling Epistasis of Quantitative Trait Loci Using Cockerham Model. *Genetics, 160*, 1243-1261.

Kempthorne, O. (1954). The Correlations Between Relatives in a Random Mating Population. *Proc. Royal Soc. London, B 143*, 103-113. http://dx.doi.org/10.1098/rspb.1954.0056

Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. New York: John Wiley & Sons.

Laird, N. M., & Lange, C. (2011). *The Fundamentals of Modern Statistical Genetics*. New York, Dordrecht, Heidelberg, London: Springer. http://dx.doi.org/10.1007/978-1-4419-7338-2

Liu, B. H. (1998). *Statistical Genomics-Linkage, Mapping and QTL Analysis*. Boca Raton, London, New York and Washington, D. C.: CRC Press.

Lynch, M., & Walsh, B. (1998). *Genetics and the Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates, Inc.

Mao, Y., Nicole, R., Ma, L., Dvorkin, D., & Da, Y. (2006). Detection of SNP epistasis effects of quantitative traits using extended Kempthorne model. *Physiol Genomics, 28*, 46-52. http://dx.doi.org/10.1152/physiolgenomics.00096.2006

Mode, C. J., & Robinson, H. F. (1959). Pleiotropism and the genetic variance and covariance. *Biometrics, 15*, 518-537. http://dx.doi.org/10.2307/2527650

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York, Chichester, Brisbane, and Singapore: John Wiley & Sons Inc.

Price, A. L., Helgason, A., Thorleifsson, G., McCarroll, S. A., Kong, A., & Stefansson, K. (2011). Single Tissue and Cross-Tissue Heritability of Gene Expression via Identity-by-Descent in Related and Unrelated Individuals. *PloS Genet, 7*(2), e1001317. http://dx.doi.org/10.1371/journal.pgen.1001317

Raj, T., Shulman, J. M., Keenan, B. T., Lori, B., Chibnik, L. B., Evans, D. A., ... De Jager, P. L. (2012). Alzheimer Disease Susceptibility Loci: Evidence for a Protein Network under Natural Selection. 2012 The American Society of Human Genetics. http://dx.doi.org/10.1016/j.ajhg.2012.02.022

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., ... Daly, M. J. (2011). Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genet, 7*(1), e1001273. http://dx.doi.org/10.1371/journal.pgen.1001273

Stranger, B. E., Eli, A., Stahl, E. A., & Raj, T. (2010). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics, 187*(2), 367-383. http://dx.doi.org/10.1534/genetics.110.120907

Wu, R. L., Ma, C.-X., & Casella, G. (2010). Statistical Genetics of Quantitative Traits-Linkage, Maps AND QTL. ISBN978-1-4419-1912-0, Springer Science.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., ... Visscher, P. M. (2010). Common SNPs explain a large proportion of heritability for human height. *Nat. Genet., 42*(7), 565-569. http://dx.doi.org/10.1038/ng.608

Zaitlen, N, Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., & Price, A. L. (2013). Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genet, 9*(5), e1003520. http://dx.doi.org/10.1371/journal.pgen.1003520