HERITABILITY ESTIMATION WITH UNKNOWN PEDIGREE BASED ON THE JOINT DISTRIBUTION OF MARKER AND PHENOTYPIC DATA USING MARKOV CHAIN MONTE CARLO TECHNIQUES

by

Jing Zhang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
July 2021

Contents

List of Tables

List of Figures

Abstract

The heritability of a quantitative trait is a very important parameter to quantify the genetic variation present in a population. Although traditional techniques for estimating heritability require accurate information of the genetic relationship among individuals, pedigree structure is generally lacking in natural population. Nowadays, the development of DNA markers is making possible to reconstruct pedigree accurately with sufficient markers. These reconstructed pedigrees have then been used with restricted maximum likelihood under a general linear mixed model to estimate heritability. In this thesis, we use markers and phenotypic observations jointly to estimate the pedigree and heritability simultaneously. We develop a MCMC sampling method of moving through the sibship configuration space and space of parameters of quantitative trait, and finding the configuration and optimal parameter values that maximizes the full joint likelihood or posterior distribution of proposed family structure and genetic variance components. Using this method, we estimate of heritabilities of 318 abalone at different time points separately and independently. Both MLE and Bayes estimate are superior to two-step method using insufficient markers (two microsatellite markers). We also give the discussion about the choices of prior distributions of parameters in the model. At the end, we extend our method to incorporate with observations at multiple time points, but we don't obtain any significant improvements.

Acknowledgements

The completion of this work was a great challenge for me. I owe many remarkable people for helping and supporting me accomplish my goal. First and formost I am extremly grateful to my supervisor, Dr. Bruce Smith for his invaluable advice, continuous support, and patience during my PhD study. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. It is a great honour and privilege to be your student.

I would also like to thank Dr. Christophe Herbinger, and Dr. Ammar Sarhan for their treasured suggestions, and I am grateful to them for accepting to serve on my dissertation committee. My stay at Dalhousie was enriched by the teaching of many dedicated professors. Dr. George Gober gave me the first statistics lecture, and I still remember his passions about Bayesian inference. Dr. Michael Dowd, Dr. Toby Kenney, Dr. David Hamilton, Dr. Hong Gu, Dr. Joanna Mills Flemming, Dr. Edward Susko, Dr. Keith Thompson, thank you so much for communicating your knowledge to me. I have learned a lot from you, not only the statistics also the way to deliver the courses. All of you inspired me to be a better teacher.

I came to Canada straight from the warmth of the home of my wonderful parents. Their love and support has always been with me and despite the tears of separation,

they encouraged and supported me to pursue my dreams in Canada. I love you so much, Mom and Dad.

Chapter 1

Introduction

Abalone is one of the most valuable mollusks in the international market, but it has a low growth rate. Aquaculture of abalone has become very popular throughout the world due to their high commercial value and over-exploitation of most wild stocks (Gordon and Cook (2004)). In Chile, abalones are not an indigenous species. However, significant quantities of farmed abalone, the red abalone especially, are produced today since they had been introduced for culture purposes from other countries 20 years ago (Flores-Aguilar et al. (2007)). Abalone only reach commercial size in around 4 years, and their slow growth is a main concern for abalone growers. Thus, it is important to continue increasing their growth rates through selective breeding and genetic improvement (Viana (2002)).

The existence of additive genetic variability for the growth trait in the population determines the success of a selective breeding program (Falconer and Mackay (1996)). Response to selection or genetic gain depends on heritability, the proportion of total phenotypic variance in a population that is attributable to the additive genetic effect (Lynch and Walsh (1998)). Additionally, heritability is not always constant for a population, as a consequence of changes in the magnitude of the genetic variance due

to modifications in allele frequencies or appearance of new variants in the population (Visscher et al. (2008)). Therefore, estimating the variance components and heritability for growth traits of abalone throughout the productive cycle is necessary in order to optimise the selection process. The changes in heritability of growth traits for red abalone, measured during 3 years from juvenile stage (27 months) to the harvesting adult stage (51 months) have been estimated by Brokordt et al. (2015). In their study, variance components for growth traits were estimated using the animal model with a restricted maximum likelihood (REML) approach (Johnson and Thompson (1995)). They fitted the animal model with fixed effects (covariates), and random effects including additive genetic effect and other random effect (i.e., a confounded maternal effect, common environmental effects and non-additive genetic effects).

One of the greatest technical limitations of animal model with REML for estimating genetic variance components is their requirement for pairs of individuals of known relatedness or full pedigree structure. In natural populations, detailed knowledge of pedigree is absent in all but the most carefully studied populations, and even then may be subject to error. Therefore, it is important and necessary to develop methods for estimating variance components and heritability with unknownn pedigree structure. In this thesis, we estimate the heritability of a growth trait with a red abalone data set used in Brokordt et al. (2015), but we assume the family structure is unknown.

This introductory chapter presents some basic concepts of quantitative-genetic

theory. In particular, we give the definition of a number of genetic terms, such as heritability (broad-sense and narrow-sense), additive and dominance effects, breeding value, etc. We also explain the genetic covariance between different types of relatives. We then introduce our hybrid data set and give a preliminary study of phenotypic observations. At the end of this chapter, we provide an outline of the remainder of the thesis.

1.1 Introduction of Quantitative-Genetic Theory and Heritability

The expression of quantitative characters is typically influenced by both genetic and environmental factors. From the perspectives of both evolutionary theory and applied breeding programs, genetic components of variance are important to understand because they determine the rates at which characters respond to selection (Lande (1982); Mousseau et al. (1987)). Environmental variance reduces the efficiency of response to selection by causing the phenotypes of selected individuals to depart from their underlying genotypic values. There are many methods to partition phenotypic variance into its various components. They are based on the principle that the phenotypic resemblance between relatives provides information on the degree of genetic differentiation among individuals. Different components of variance influence the resemblance between relatives to different degrees, and have substantially different influences on the evolutionary process (Fisher (1919); Wright (1922)). The additive component of the genetic variance, which is defined below, is of particular interest because it is

the primary determinant of the degree to which offspring resemble their parents, and governs the rate of response of a character to selection.

The evolutionary response of a trait to selection is a function of the intensity of selection and the fraction of the phenotypic variance attributable to certain genetic effects. The within-generation difference between the mean phenotype μ_S after an episode of selection (but before reproduction) and the mean before selection μ, is called the directional selection differential (Lynch and Walsh (1998)).

$$S = \mu_S - \mu \tag{1.1}$$

The value of S depends on the survivorship and reproductive rates of individuals with difference phenotypes. If all individuals have the same fertility and viability, then $\mu_S = \mu$, and $S = 0$, which indicates the population mean phenotype is not expected to change between generations. The directional selection differential also can be viewed as the covariance of phenotype and relative fitness (Price et al. (1970)).

When the regression of offspring phenotype on the average parent is linear with slope β, the change in the parental mean phenotype leads to an expected change in the mean phenotype across generations equal to

$$\Delta\mu = \mu_o - \mu = \beta S \tag{1.2}$$

where μ_o is the mean phenotype of the offspring. This equation is called the breeders' equation (Lush (1937)) and it combines information on the forces of selection (S) with that on inheritance (β) to produce a predictive equation for evolutionary change across generations.

1.1.1 The Heritability in the Broad Sense

The breeders' equation shows the importance of heritable variation in the evolution of a trait through natural selection. The response to selection across generations is zero if $\beta = 0$, no matter how large S is. Quantification of the correspondence between phenotypic and genotypic values is related to one of the central goals of quantitative genetics - the partitioning of the phenotypic variance into genetic and nongenetic components. The phenotypic value of an individual, Y, can be considered as the sum of the total effects of all loci on the trait, G (the genotypic value), and an environmental effect E (Kempthorne (1957)).

$$Y = G + E \tag{1.3}$$

The covariance between phenotypic and genotypic values can be written as

$$\sigma_{Y,G} = \sigma[(G + E), G] = \sigma_G^2 + \sigma_{G,E} \tag{1.4}$$

If we assume no genotype-environment covariance, i.e., $\sigma_{G,E} = 0$, the squared correlation coefficient can be simplified as

$$\rho^2_{Y,G} = \big(\frac{\sigma_{Y,G}}{\sigma_Y \cdot \sigma_G}\big)^2 = \frac{(\sigma^2_G + \sigma_{G,E})^2}{\sigma^2_Y \sigma^2_G} = \frac{\sigma^2_G}{\sigma^2_Y} \tag{1.5}$$

The squared correlation coefficient is simply the proportion of the total phenotypic variance that is genetic. The quantity $\frac{\sigma^2_G}{\sigma^2_Y}$ is generally referred as broad-sense heritability and noted as H^2 (Plomin (1990)). The broad-sense heritability provides some insight into the partitioning of the phenotypic variance into genetic and residual components without getting stuck in genetic complexities. However, while this leads to methods for estimation, the approach is not particularly informative to the practicing geneticist. The broad-sense heritability estimates the degree to which differences among individuals are genetically based, but not all genetically based differences can be passed from parents to offsprings. The underlying genetic values are essentially unobservable without an extensive breeding program.

1.1.2 The Heritability in the Narrow Sense

Based on the simple fact that related individuals carry copies of many of the same alleles, the resemblance between phenotypes of offspring (Y_o) and their midparents (Y_{mp}) inspires a method to estimate levels of genetic variance of quantitative traits. A midparent value is the average phenotype of a mother (Y_m) and a father (Y_f),

$$Y_{mp} = \frac{Y_m + Y_f}{2} \tag{1.6}$$

We will start with a simple genetic situation - a single locus with purely additive gene effects, diploidy, random mating, and no selection. Let g_m and g_f be the effects of the alleles that the offspring inherits from its mother and father, respectively, and g'_m and g'_f be the effects of the alleles that the parents do not transmit to their offspring. Let E_m, E_f and E_o be the environmental effects on the phenotypes of mother, father and offspring, respectively. The three phenotypes can be written as

$$Y_m = g_m + g'_m + E_m \tag{1.7}$$

$$Y_f = g_f + g'_f + E_f \tag{1.8}$$

$$Y_o = g_m + g_f + E_o \tag{1.9}$$

Therefore, the midparent phenotype can be expressed as

$$Y_{mp} = \frac{g_m + g'_m + E_m + g_f + g'_f + E_f}{2} \tag{1.10}$$

The complete expression for the midparent-offspring covariance, σ_{Y_{mp},Y_o} is quite complex, containing $18 = (6 \times 3)$ terms. However, most of these terms have expected values equal to zero under the following assumptions:

- Under the assumptions of random mating and no selection, there is no covariance between the effects of alleles within individuals. The effects of genes inherited by an offspring have zero covariance with the effects of genes that are not inherited, and the effects of genes in mothers are not correlated with those

in fathers.

- Under the assumption of no genotype-environment covariance, the covariances between genetic effects and environmental effects are all equal to zero.

- Provided the parents do not transmit their environmental effects to their progeny, then the covariance between the environmental effects of parents and environmental effects of offspring are equal to zero.

Assuming all above assumptions are fulfilled, the only potential sources of covariance that exist between midparent and offspring phenotypes are those based on the inherited genes. Therefore,

$$\sigma_{Y_{mp},Y_o} = \sigma\Big[(\frac{Y_m + Y_f}{2}), Y_o\Big] = \sigma\Big[\big(\frac{g_m + g_f}{2}\big), \big(g_m + g_f\big)\Big] = \frac{\sigma^2_{g_m} + \sigma^2_{g_f}}{2} \tag{1.11}$$

Since we assumed the genotypic value to be entirely defined by the additive effects of the two alleles, the total genetic variance in the population is the sum of the variances of maternally and paternally derived genes, $\sigma^2_{g_m} + \sigma^2_{g_f}$. Because the gene effects are purely additive, this quantity is also be referred to as the additive genetic variance, σ^2_A. Thus, we conclude that the covariance between midparent and offspring phenotypes is equal to half of the additive genetic variance in the population, and this relationship holds for any number of loci provided they interact additively.

$$\sigma_{Y_{mp},Y_o} = \frac{\sigma^2_A}{2} \tag{1.12}$$

In order to get the expected least-squares regression of offspring on midparent phenotype, the slope is equal to the covariance divided by the variance of midparent phenotypes. Under random mating, we believe that the phenotypic covariance between parents is zero, $\sigma_{Y_f,Y_m} = 0$. We also believe that the phenotypic variance in the two sexes is the same, and equal to the phenotypic variance in the population, σ^2_Y. Thus,

$$\sigma^2_{Y_{mp}} = \sigma^2\big(\frac{Y_m + Y_f}{2}\big) = \frac{\sigma^2_{Y_m} + \sigma^2_{Y_f}}{4} = \frac{\sigma^2_Y}{2} \tag{1.13}$$

The slope of the least-squares linear regression of offspring phenotype on midparent phenotype is then,

$$\beta_{o,mp} = \frac{\sigma^2_A}{\sigma^2_Y} \tag{1.14}$$

The slope of a midparent-offspring regression provides an estimate of proportion of the phenotypic variance that is attributable to additive factors. The ratio σ^2_A/σ^2_Y is known as the narrow-sense heritability and is denoted as h^2 (Lynch and Walsh (1998)).

Recall the breeders' equation (equation 1.2), where β can be treated as h^2. If S is the change in mean phenotype caused by selection prior to reproduction, then the response to selection across generations can be written as,

$$\Delta\mu = h^2 S \tag{1.15}$$

The narrow-sense heritability can be expressed as the efficiency of the response to selection. Regardless of the strength of selection, when $h^2 = 0$, there is no evolutionary change.

The heritability has a value that lies between 0 and 1. It is important to understand that the heritability of a trait is defined for a given population at a given time. This quantity can vary between populations, and can change from time to time (Visscher et al. (2008)). As allele frequencies change, so does heritability. A population with a high h^2 value may have a heritability drop to zero very quickly, while another population with a much smaller h^2 value may have heritability increase during selection as rare alleles become more frequent. Therefore, heritability is an unreliable predictor for long-term response, although it is generally a good predictor of short-term response.

1.2 Genotypic Value and Decomposition of the Genetic Variance

1.2.1 Allele and Genotype Frequencies

The genetic information encoding for characters resides on extremely long strands of deoxyribonucleic acid (DNA) called chromosomes. DNA sequences that encode for particular products are referred to as gene, and their chromosomal locations are

called loci. Many organisms are said to be diploid since they have two copies of each of several chromosomes, one copy from a female parent, and one from a male parent. As we know, DNA replication is an imperfect process and the two copies of each gene carried by diploid individuals need not be identical to those in the parents when mutations appear. The various forms of gene are called alleles.

When denoting the genotype at a single locus, we refer to the pair of alleles that a diploid individual carries at the locus. Individuals that have two identical alleles are called homozygotes, and those that have different alleles are called heterozygotes. For example, we denote the alleles at a particular diallelic locus as B_1 and B_2. There are three possible genotypes: B_1B_1 and B_2B_2 homozygotes, and B_1B_2 heterozygotes.

Allele frequencies are defined uniquely by genotype frequencies. Suppose we use P_{11}, P_{12}, and P_{22} represent the proportions of the population that are B_1B_1, B_1B_2, and B_2B_2. By definition, $P_{11} + P_{12} + P_{22} = 1$ if these are the only genotypes at the locus. Therefore, the relative frequency of the B_1 allele is

$$p_1 = P_{11} + \frac{1}{2}P_{12} \tag{1.16}$$

Thus, the frequency of an allele can be estimated by the observed frequency of homozygotes plus one-half the observed frequency of all heterozygotes containing that allele.

1.2.2 Genotypic values and Fisher's Decomposition

The phenotype (Y) of an individual can be partitioned to a genotypic value (G) and an environmental deviation (E), where G is the expected phenotype (for a given genotype) resulting from the joint expression of all of the genes underlying the trait. The expression for the genotypic value G can be a complicated function for a multilocus trait. However, we will only consider the simpler situation with direct contribution from a single autosomal locus. For now, we start with the special case in which there are only two alleles. Lynch and Walsh (1998) explained a way to define the genotypic values. The genotypic values of the B_1B_1 and B_2B_2 homozygotes can be set to zero and $2a$ respectively, with $2a$ representing the difference between the mean phenotypes of B_2B_2 and B_1B_1. The genotypic value of B_1B_2 is defined to be $(1+k)a$, where k provides a measure of dominance. Alleles B_1 and B_2 behave in a completely additive fashion when $k = 0$, whereas $k = +1$ implies complete dominance of the B_2 allele, and $k = -1$ implies complete dominance of the B_1 allele. If $k > 1$, the phenotypic expression of the heterozygote exceeds both homozygotes, and the locus is said to exhibit overdominance, whereas $k < -1$ implies underdominance.

The number of copies of a particular allele (B_2) in a genotype ($N_2 = 0$, 1 or 2 for diploid individuals) is referred to as the gene content. It is useful to consider the best linear approximation to the relationship between the gene content and the genotypic value. Genotypic values can be partitioned into their expected values based on additivity ($\widehat{G}$) and deviations from those expectations resulting from dominance

(δ).

The preceding points can be formalized by least-squares regression of genotypic values on the number of B_1 and B_2 alleles in the genotype, N_1 and N_2.

$$G_{ij} = \widehat{G}_{ij} + \delta_{ij} = \mu_G + \alpha_1 N_1 + \alpha_2 N_2 + \delta_{ij} \tag{1.17}$$

The genotypic value of genotype B_iB_j is a function of μ_G, the population mean genotypic value, α_1 and α_2, the slopes of the regression of the predictor variables N_1 and N_2, and δ_{ij}, the residual error. This partitioning of genotypic values into various components is one of several major advances developed in Fisher (1919). For the two alleles case, the model can be reduced to a standard univariate regression by noting that for any individual, $N_1 = 2 - N_2$, then

$$\begin{aligned} G_{ij} &= \mu_G + \alpha_1(2 - N_2) + \alpha_2 N_2 + \delta_{ij} \\ &= \iota + (\alpha_2 - \alpha_1)N_2 + \delta_{ij} \end{aligned} \tag{1.18}$$

where $\iota = \mu_G + 2\alpha_1$ is the intercept. The slope of this regression can be written as

$$\alpha = \alpha_2 - \alpha_1 \tag{1.19}$$

As we know, the slope of a univariate regression is simply the covariance between response and predictor variable divided by the variance of the predictor variable.

Thus, the slope of the regression α is

$$\alpha = \frac{\sigma_{G,N_2}}{\sigma^2_{N_2}} \tag{1.20}$$

The terms σ_{G,N_2} and $\sigma^2(N_2)$ are functions of the gene effects (a and k) and proportions of the B_1 and B_2 alleles (p_1 and p_2). They can be easily derived by using the quantities in Table 1.1. Thus, we obtain

$$\alpha = a[1 + k(p_1 - p_2)] \tag{1.21}$$

Under the assumption of random mating, α is known as the average effect of allelic substitution. It represents the average change in genotypic value that results when a B_2 allele is randomly substituted for a B_1 allele. For the purely additive case where $k = 0$, α is simply equal to a.

Table 1.1: Properties of a single segregating diallelic locus under random mating.

Genotype	Gene Content (N_2)	Genotypic Values (G)	Freq	$G \cdot N_2$	N^2	Regression Value ($\hat{G}$)	Dominance Deviation (δ)
B_1B_1	0	0	p_1^2	0	0	ι	$-\iota$
B_1B_2	1	$(1+k)a$	$2p_1p_2$	$(1+k)a$	1	$\iota + \alpha$	$(1+k)a - \iota - \alpha$
B_2B_2	2	$2a$	p_2^2	$4a$	4	$\iota + 2\alpha$	$2a - \iota - 2\alpha$

1.2.3 Partitioning the Genetic Variance

Once the genotypic values have been partitioned in the above manner, it is a relatively simple step to partition the sources of genetic variation at a locus. Recalling the equation $G = \widehat{G} + \delta$, the total genetic variance can be written as

$$\sigma_G^2 = \sigma_{\widehat{G}+\delta}^2 = \sigma_{\widehat{G}}^2 + 2\sigma_{\widehat{G},\delta} + \sigma_\delta^2 \tag{1.22}$$

Based on the property of least-squares regression, the regression prediction is uncorrelated with the residual error. Thus, the total genetic variance attributable to a locus simplifies to the sum of additive and dominance components. These components are denoted as σ_A^2 and σ_D^2.

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 \tag{1.23}$$

Statistically speaking, σ_A^2 is the amount of variance of G that is explained by the regression on N_2, whereas σ_D^2 is the residual variance for the regression. Biologically speaking, σ_A^2 is the genetic variance associated with the average additive effects of alleles (the additive genetic variance), and σ_D^2 is the genetic variance associated with dominance effects (the dominance genetic variance).

Partitioning the genotypic value into additive and dominance components is very useful. In randomly mating diploid species, a parent donates only one allele per locus to each of its offspring. The transmitted allele exhibits its additive effect when randomly combined with a gene from other parents. The dominance deviation of a

parent, which is a function of the interaction between two parental genes, is eliminated when gametes are produced. Thus, $\widehat{G}$ and δ can be considered as the heritable and nonheritable components of an individual's genotypic value.

1.2.4 Additive Effects and Breeding Values

The additive effects, α_i, can be defined to be the least-squares regression coefficients of genotypic value on gene content (equation 1.17). They are obtained by finding the α_1 and α_2 that minimize the mean-squared residual deviation

$$\begin{aligned} MSE = E(\delta_{ij}^2) &= E[(G_{ij} - \mu_G - \alpha_1 N_1 - \alpha_2 N_2)^2] \\ &= (G_{11} - \widehat{G}_{11})^2 P_{11} + (G_{12} - \widehat{G}_{12})^2 P_{12} + (G_{22} - \widehat{G}_{22})^2 P_{22} \end{aligned}$$

where P_{ij} is the relative frequency of the $ijth$ genotype. For a randomly mating population, setting the partial derivatives of MSE with respect to α_i equal to zero, the solutions are

$$\alpha_2 = p_1 a[1 + k(p_1 - p_2)] = p_1 \alpha \tag{1.24}$$

$$\alpha_1 = -p_2 a[1 + k(p_1 - p_2)] = -p_2 \alpha \tag{1.25}$$

The α_i are often referred to as average effects, but we use additive effects to discriminate them from average effects of higher-order gene actions (such as dominance).

An individual's breeding value, denoted by A, is the sum of the additive effects of its genes. In other words, the breeding value of a B_1B_1 homozygote is simply $2\alpha_1$, that of a heterozygote is $(\alpha_1 + \alpha_2)$, and that of B_2B_2 is $2\alpha_2$. The definitions for additive effects and breeding value present a very useful relationship for a random-mating population. The breeding value of a genotype is equivalent to twice the expected deviation of its offspring mean phenotype from the population mean. Therefore, the breeding value of an individual can be estimated by mating it to many randomly chosen individuals from the population and taking twice the deviation of its offspring mean from the population mean.

1.2.5 Additive Genetic Variance for Multiple Alleles

Although the preceding results were obtained under the assumption of a diallelic locus and random mating, they can be generalized to situations with an arbitrary number of alleles, as well as to nonrandomly mating populations.

As in the diallelic case, with n alleles the additive effects are defined to be the set of α_i that minimizes the mean- squared residual deviations, $E(\delta_{ij}^2)$, obtained from the

least- squares solution for the multiple regression

$$G = \mu_G + \sum_{i=1}^{n} \alpha_i N_i + \delta \tag{1.26}$$

This equation is the n-allele extension of equation 1.17, with N_i being the number of copies of allele i carried by an individual. For example, for the genotype G_{12}, $\sum \alpha_i N_i = \alpha_1 + \alpha_2$, and $\delta_{12} = G_{12} - \mu_G - \alpha_1 - \alpha_2$.

Finally, we consider the general definition of the breeding value (A_{ij}) under random mating. Returning to equation 1.26,

$$\begin{aligned} G_{ij} &= \mu_G + \alpha_i + \alpha_j + \delta_{ij} \\ &= \mu_G + A_{ij} + \delta ij \end{aligned} \tag{1.27}$$

Therefore, the genotypic value can be decomposed into four quantities: the mean of genotypic value for the population, the additive effects of the two genes, and a dominance deviation due to the interaction between the genes. By the properties of least-squares regression, A_{ij} and δ_{ij} are uncorrelated, and

$$\sigma_G^2 = \sigma_{\alpha_i + \alpha_j}^2 + \sigma_{\delta_{ij}}^2 \tag{1.28}$$

This is a completely general equation, applying even to the case of nonrandom mating.

For the special case of random mating, α_i and α_j are uncorrelated, and

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 \tag{1.29}$$

Comparing this with equation 1.23, we find σ_A^2 has a very specific and useful meaning. Under random mating assumption, the additive genetic variance is equivalent to the variance of breeding values of individuals in the population. More detailed explanations can be found in Lynch and Walsh (1998).

1.2.6 A General Least-Squares Model for Genetic Effects for Multilocus Traits

In the previous section, the genetic variance associated with a single locus can be partitioned into additive and dominance components. This approach will be generalized below to account for all of the loci contributing to the expression of a quantitative trait, as well as to allow for variance arising from gene interaction among loci.

As shown in previous section, the dominance effect at a locus was defined to be the deviation of the observed genotypic value from the expectation based on additive effects. Thus, dominance is considered as a measure of nonadditivity of allelic effects within loci. Epistasis describes the nonadditivity of effects between loci. With only two loci, there are actually three ways in which epistatic interactions can arise between loci: additive $\times$ additive ($\alpha\alpha$), additive $\times$ dominance ($\alpha\delta$), and dominance $\times$

dominance ($\delta\delta$). With three loci, there are four additional types of epistasis, ($\alpha\alpha\alpha$), ($\alpha\alpha\delta$), ($\alpha\delta\delta$), and ($\delta\delta\delta$). For example, consider an individual with allele A_i and A_j at one locus and B_k and B_l at another. The genotypic value G_{ijkl}, can be expressed as the sum of the effects within loci and a deviation ϵ due to interaction between loci,

$$G_{ijkl} = \mu_G + (\alpha_i + \alpha_j + \delta ij) + (\alpha_k + \alpha_l + \delta_{kl}) + \epsilon_{ijkl} \tag{1.30}$$

where ϵ_{ijkl} contains all the epistatic interactions.

Lynch and Walsh (1998) provided a detailed statistical procedure to define epistatic effects in two loci linear model and the a general least-squares model for genetic effects. It can be considered as a simple extension of the one-locus linear model introduced previously. With the one-locus model, the additive effect of an allele was defined as the deviation of the phenotype for members of the population with the allele from the population mean phenotype. This definition remains the same with the addition of loci. Letting $G_{i...}$ represent the conditional mean phenotype of individuals with allele i at the first locus without regard to the other allele at the locus or to the genotype at the second locus, then

$$\alpha_i = G_{i...} - \mu_G \tag{1.31}$$

The other three additive effects (α_j, α_k, α_l) can be defined in the same way. Let $G_{ij..}$ represent the conditional mean phenotype of individuals with alleles i and j at the first locus without regard to genotypic state at the second locus. When the mean

genotypic value and the additive effects have been removed, the dominance effect is left as the only unexplained portion of the conditional mean at the locus.

$$\delta_{ij} = G_{ij..} - \mu_G - \alpha_i - \alpha_j \tag{1.32}$$

$$\delta_{kl} = G_{..kl} - \mu_G - \alpha_k - \alpha_l \tag{1.33}$$

Because within each locus the mean value of the additive effects is equal to zero, the mean dominance deviation is equal to zero as well.

The definition of epistatic effects can be expressed in a similar fashion. Letting $G_{i.k.}$ be the mean of phenotype of individuals with gene i at locus 1 and k at locus 2, without regard to the other two genes, the ik'th additive $\times$ additive effect is

$$(\alpha\alpha)_{ik} = G_{i.k.} - \mu_G - \alpha_i - \alpha_k \tag{1.34}$$

It is the deviation of the conditional mean $G_{i.k.}$ from the expectation based on the population mean μ_G and the additive effects α_i and α_k. An additive $\times$ dominance effect measures the interaction between an allele at one locus with a particular genotype at another locus. It is defined as the deviation of the conditional mean $G_{i.kl}$ from the expectation based on all lower-order effects, which include three additive effects, one dominance effect, and two additive $\times$ additive effects involving the constituent genes,

$$(\alpha\delta)_{ikl} = G_{i.kl} - \mu_G - \alpha_i - \alpha_k - \alpha_l - \delta_{kl} - (\alpha\alpha)_{ik} - (\alpha\alpha)_{il} \tag{1.35}$$

Last, for a dominance × dominance effect,

$$\begin{aligned}(\delta\delta)_{ijkl} = G_{ijkl} &- \mu_G - \alpha_i - \alpha_j - \alpha_k - \alpha_l - \delta_{ij} - \delta_{kl} \\ &- (\alpha\alpha)_{ik} - (\alpha\alpha)_{il} - (\alpha\alpha)_{jk} - (\alpha\alpha)_{jl} \\ &- (\alpha\delta)_{ikl} - (\alpha\delta)_{jkl} - (\alpha\delta)_{ijk} - (\alpha\delta)_{ijl}\end{aligned} \tag{1.36}$$

The total genotypic value has been partitioned into a series of effects. First, additive effects of alleles which is the lower-order effects, account for as much of the variance in genotypic values as possible. Then, higher-order effects are defined progressively, each time accounting for as much of the residual variation as possible. In the two-locus case, $(\delta\delta)_{ijkl}$ represents the final part of variation not accounted by additive, dominance, additive × additive, or additive × dominance effects. Therefore, the genotypic value can be expressed as

$$\begin{aligned}G_{ijkl....} = \mu_G &+ [\alpha_i + \alpha_j + \alpha_k + \alpha_l] + [\delta_{ij} + \delta_{kl}] \\ &+ [(\alpha\alpha)_{ik} + (\alpha\alpha)_{il} + (\alpha\alpha)_{jk} + (\alpha\alpha)_{jl}] \\ &+ [(\alpha\delta)_{ikl} + (\alpha\delta)_{jkl} + (\alpha\delta)_{ijk} + (\alpha\delta)_{ijl}] + (\delta\delta)_{ijkl} + ...\end{aligned} \tag{1.37}$$

The open-ended equation only implies that there are more terms when more than two loci are involved in the expression of a trait. The parameters in this model depend upon the genotype frequencies in the population and they change as allele frequencies change. However, the mean value of each effect is always equal to zero.

Provided that mating is random and segregation of loci is independent, there is no statistical relationship between the genes found within or among loci. Therefore, the total genetic variance is simply the sum of the variances of individual effects. Letting $\sigma^2_A = \sigma^2_{\alpha_i} + \sigma^2_{\alpha_j} + \sigma^2_{\alpha_k} + \sigma^2_{\alpha_l}$, $\sigma^2_D = \sigma^2_{\delta_{ij}} + \sigma^2_{\delta_{kl}}$, $\sigma^2_{AA} = \sigma^2_{(\alpha\alpha)_{ik}} + \sigma^2_{(\alpha\alpha)_{il}} + \sigma^2_{(\alpha\alpha)_{jk}} + \sigma^2_{(\alpha\alpha)_{jl}}$, and so on, the total genetic variance can be expressed as

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DD} + ... \tag{1.38}$$

This partitioning of the genetic variance into a series of components was developed independently, but with very different approaches, by both Cockerham (1954) and Kempthorne (1954). Because of the hierarchical way in which genetic effects are defined, the magnitude of genetic variance components might be expected to become progressively smaller at higher stages in the hierarchy.

1.3 Genetic Covariance Between Relatives

The phenotypic variance of a trait can theoretically be partitioned into a number of genetic and environmental components. Because various genetic and environmental sources of variance contribute differentially to the resemblance between different types

of relatives(Fisher (1919); Wright (1922)), the significant practical issue of how these components can be estimated remains.

Assuming no genotype × environmental interaction, let $Y_i = G_i + E_i + e_i$ and $Y_j = G_j + E_j + e_j$ be the phenotypic values of two members of particular relationship, such as parent and offspring or half sibs. G, E and e denote genotypic values, environmental effects and residual deviations, respectively. The phenotypic covariance between relatives i and j can be expressed as

$$\begin{aligned}\sigma_Y(i,j) &= \sigma[(G_i + E_i + e_i), (G_j + E_j + e_j)] \\ &= \sigma_G(i,j) + \sigma_{G,E}(i,j) + \sigma_{G,E}(j,i) + \sigma_E(i,j) \end{aligned} \tag{1.39}$$

When we ignore the issue of genotype-environmental covariance, the equation 1.39 is simplified to

$$\sigma_Y(i,j) = \sigma_G(i,j) + \sigma_E(i,j) \tag{1.40}$$

The phenotypic covariance between relatives has been partitioned into genetic covariance and covariance between environmental effects. The genetic covariance is a natural consequence of relatives inheriting copies of the same genes. Similar to genetic variance, the genetic covariance between relatives can be partitioned into components attributable to additive, dominance, and various epistatic effects. Each term consists of one of the familiar components of genetic variance weighted by a coefficient that

describes the joint distribution of genetic effects in pairs of relatives.

1.3.1 Measures of Relatedness

Many relatedness measures have found their way into the population-genetic and sociobiological literature (Wright (1922); Cotterman (1940); Grafen (1985)). All measures are based upon the concept of identity by descent (IBD). Genes that are identical by descent are direct descendents of a specific gene carried in some ancestral individual.

Lange (2003) provides an easy to follow introduction to the concept of identity by descent, and the derivation of a number of associated probabilities.

Consider a single locus in two diploid individuals. There are 15 possible configurations of identity by descent due to the fact that identity may exist within as well as between individuals (Gillois (1965)). If we ignore the distinction between maternally and paternally derived genes, the 15 possible configurations reduce to 9 identity states. Each of these 9 condensed identity states has an associated probability. The collection of 9 probabilities Δ_1 to Δ_9 associated with the 9 identity states were referred to in Jacquard (1974) as condensed coefficients of identity. These coefficients provide a complete description of the probability distribution of identity by descent

at a single locus in a pair of individuals. The values of the condensed identity coefficients depend on the relationship between individuals.

Suppose that the single genes are drawn randomly from individuals x and y. The probability that these two genes are identical by descent is called coefficient of coancestry, Θ_{xy}. It is also referred to as the coefficient of consanguinity, coefficient of kinship, or coefficient de parenteé. In terms of the condensed coefficient of identity,

$$\Theta_{xy} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8 \tag{1.41}$$

This formula weights each condensed identity coefficient by the conditional probability that a randomly draw gene from x is identical by descent with a randomly selected gene from y at the same locus (Lange (2003)). For example, Δ_8 is the probability that one copy of the gene in individual x is IBD to one copy of the gene in individual y. There are 4 possible ways that this can happen: maternal gene in x is IBD with maternal gene in y; maternal gene in x is IBD with paternal gene in y; paternal gene in x is IBD with maternal gene in y and paternal gene in x is IBD with paternal gene in y. Given this condensed state, if single genes are drawn at random from the two individuals, the probability of drawing the pair which is IBD is $\frac{1}{4}$.

Another useful measure of relatedness is the probability that single-locus genotypes (both genes) of two individuals are identical by descent. The formulation of this measure was called the coefficient of fraternity by Trustrum and Williamson (1961).

It can be denoted as Δ_{xy}.

1.3.2 The Genetic Covariance Between Relatives

Karigl (1981) provides a general recursive procedure for calculating all 9 condensed coefficients of identity for arbitrary pedigrees. The identity coefficients for several common relationship as summarized in the Table 1.2 (Lynch and Walsh (1998)). Under assumption of no inbreeding, Δ_1 to Δ_6 are all equal to zero. Δ_7 is the probability that x and y have two genes IBD, Δ_8 is the probability that x and y have single genes which are IBD, and Δ_9 is the probability that x and y have no genes IBD at the locus of interest. It is fairly straightforward to understand these coefficients. For example, if an offspring arises from the mating of two unrelated parents who share no genes IBD, then a parent and offspring are guaranteed to have 1 and only 1 copy IBD, the copy of the gene which was passed from the parent to the offspring. Hence, $\Delta_7 = 0$, $\Delta_8 = 1$ and $\Theta_{xy} = \frac{1}{4}$.

Before we give the expression of genetic covariance between relatives, we have to make some assumptions: (1) all the genetic variation is attributable to diploid, autosomal loci; (2) mating is random; (3) all loci are unlinked and in gametic phase equilibrium; (4) there is no genetic variation from maternal effects; (5) genotype-environment covariance and interaction are unimportant; (6) there is no sexual dimorphism; (7) selection is not operating on the population. We will assume all of

Table 1.2: Identity coefficients for common relationships.

Relationship	Δ_7	Δ_8	Δ_9	Θ_{xy}	Δ_{xy}
Parent-offspring	0	1	0	1/4	0
Grandparent-grandchild	0	1/2	1/2	1/8	0
Great grandparent-great grandchild	0	1/4	3/4	1/16	0
Half sibs	0	1/2	1/2	1/8	0
Full sibs, dizygotic twins	1/4	1/2	1/4	1/4	1/4
Uncle(aunt)-nephew(niece)	0	1/2	1/2	1/8	0
First cousins	0	1/4	3/4	1/16	0
Double first cousins	1/16	6/16	9/16	1/8	1/16
Second cousins	0	1/16	15/16	1/64	0
Monozygotic twins (clonemates)	1	0	0	1/2	1

these conditions hold. Associated definitions - eg. gametic phase equilibrium, sexual dimorphism, etc can be found in Lynch and Walsh (1998).

Consider a collection of pairs of individuals all of the same type of relationship, and let x and y represent the members of a random pair. From equation 1.38, we can write the genotypic values of the two individuals as follow:

$$\begin{aligned} G_{ijkl....}(x) = \mu_G &+ [\alpha_i^x + \alpha_j^x + \alpha_k^x + \alpha_l^x + ...] + [\delta_{ij}^x + \delta_{kl}^x + ...] \\ &+ [(\alpha\alpha)_{ik}^x + (\alpha\alpha)_{il}^x + (\alpha\alpha)_{jk}^x + (\alpha\alpha)_{jl}^x + ...] \\ &+ [(\alpha\delta)_{ikl}^x + (\alpha\delta)_{jkl}^x + (\alpha\delta)_{ijk}^x + (\alpha\delta)_{ijl}^x + ...] + (\delta\delta)_{ijkl}^x + ... \end{aligned} \tag{1.42}$$

$$G_{ijkl....}(y) = \mu_G + [\alpha_i^y + \alpha_j^y + \alpha_k^y + \alpha_l^y + ...] + [\delta_{ij}^y + \delta_{kl}^y + ...]$$
$$+ [(\alpha\alpha)_{ik}^y + (\alpha\alpha)_{il}^y + (\alpha\alpha)_{jk}^y + (\alpha\alpha)_{jl}^y + ...]$$
$$+ [(\alpha\delta)_{ikl}^y + (\alpha\delta)_{jkl}^y + (\alpha\delta)_{ijk}^y + (\alpha\delta)_{ijl}^y + ...] + (\delta\delta)_{ijkl}^y + ... \quad (1.43)$$

where i, j and k, l represent genes at the first and second loci. Fisher (1919) showed that different types of effects are uncorrelated within individuals and between individuals when all of the preceding assumptions are met. Therefore, the genetic covariance between relatives can be expanded into a series of terms, each describing the covariance between the same kinds of effects in two individuals:

$$\sigma_G(x, y) = \sigma_A(x, y) + \sigma_D(x, y) + \sigma_{AA}(x, y) + \sigma_{AD}(x, y) + \sigma_{DD}(x, y) + ... \quad (1.44)$$

When $x = y$, equation 1.45 reduces to equation 1.39, the usual expression for the genetic variance.

Following Lynch and Walsh (1998), each term in equation 1.45 can be expressed in terms of variance components and coefficients of relationship. The covariance between relatives can be defined as

$$\begin{aligned}\sigma_G(x,y) &= \sum (2\Theta_{xy})^n \Delta_{xy}^m \sigma^2_{A^n D^m} \\ &= 2\Theta_{xy}\sigma^2_A + \Delta_{xy}\sigma^2_D + (2\Theta_{xy})^2\sigma^2_{AA} \\ &+ 2\Theta_{xy}\Delta_{xy}\sigma^2_{AD} + \Delta^2_{xy}\sigma^2_{DD} + (2\Theta_{xy})^3\sigma^2_{AAA} + ... \end{aligned} \tag{1.45}$$

where n and m represent the number of additive effects and the number of dominance effects in a type of gene action. Taking the values of the coefficients Θ_{xy} and Δ_{xy} from Table 1.2, explicit expressions for the genetic covariances between common types of relatives can be calculated, and are given in Table 1.3. Note that, these results are only expanded to include two-locus epistatsis. To obtain the covariance expression for a particular type of relationship, multiply each variance component by its coefficient and sum. For example, the genetic covariance between parent-offspring is $(\sigma^2_A/2) + (\sigma^2_{AA}/4)$. The genetic covariance between full sibs is $(\sigma^2_A/2) + (\sigma^2_D/4) + (\sigma^2_{AA}/4) + (\sigma^2_{AD}/8) + (\sigma^2_{DD}/16)$, but if we assume no dominance and epistasic effects, then the genetic covariance between full sibs is half the additive genetic variance. We will use this result when we estimate heritability after we find the estimation of interclass correlation in section 2.2.

Although these expressions are only expanded to include two-locus epistasis, some features are immediately apparent. First of all, gene action involving dominance only rarely contributes to the covariance between relatives. Second, the coefficient for σ^2_{AA}

Table 1.3: Coefficients for the components of genetic covariance between different types of relatives under the assumptions of random mating, free recombination, and gametic phase equilibrium.

Relationship	σ_A^2	σ_D^2	σ_{AA}^2	σ_{AD}^2	σ_{DD}^2
Parent-offspring	1/2	0	1/4	0	0
Grandparent-grandchild	1/4	0	1/16	0	0
Great grandparent-great grandchild	1/8	0	1/64	0	0
Half sibs	1/4	0	1/16	0	0
Full sibs, dizygotic twins	1/2	1/4	1/4	1/8	1/16
Uncle(aunt)-nephew(niece)	1/4	0	1/16	0	0
First cousins	1/8	0	1/64	0	0
Double first cousins	1/4	1/16	1/16	1/64	1/256
Second cousins	1/32	0	1/1024	0	0
Monozygotic twins (clonemates)	1	1	1	1	1

declines more rapidly with the distance of the relationship than does that for σ_A^2. The expressions in Table 1.3 provides a method of the estimation of the different variance components from linear combinations of different observed genetic covariances between relatives. For example, $2\times$ [($4\times$ half-sib covariance)$-$(parent-offspring covariance)] has an expected value of σ_A^2.

1.4 Hybrid Data Set

The thesis uses a hybrid data set, with genetic marker value and phenotypic observational value for each individual selected from two separate data sets.

Microsatellite markers are both rich and polymorphic in the several fish species (Goff et al. (1992); Colbourne et al. (1996)). Because of the variability of these loci and ease of examination using the polymerase chain reaction (PCR), microsatellites

are popular as genetic markers for a range of applications in fisheries and aquaculture, such as population differentiation (McConnell et al. (1995)), parentage determination in mixed family groups (Herbinger et al. (1995)), and family relatedness in natural populations (Herbinger et al. (1997)). The genetic marker data (microsatellites) that we used in the thesis comes from a study of Atlantic salmon by O'reilly et al. (1998). Four microsatellites from Atlantic salmon were isolated in that study, including three teranucleotide loci and an additional dinucleotide locus. All four loci can be amplified in a single reaction and exhibit nonoverlapping allele size distribution, permitting identification using standard autoradiographic detection methods. The combined probability of match value for all four loci was approximately 3.4×10^{-5} (O'Reilly et al. (1996)). Therefore, given the accuracy of scoring alleles and combined information content, this microsatellite system is ideal for familial identification. This is a large data set consisting of 759 Atlantic salmon from 12 full-sib families, with family size varying from 8 to 140. Each offspring was typed at four microsatellite loci, with 11, 14, 10, and 8 alleles per locus. The empirical allele frequencies of the offspring were used as estimates of the allele frequencies, which are required to calculate the likelihood, as will be seen in the later chapter.

The phenotypic trait observations were chosen from a study on growth of abalone (Brokordt et al. (2015)). Sixty full-sib families were produced using Haliotis *refescens* abalone broodstock randomly obtained from a base population of 600 adults, which

were provided by three different abalone breeding companies (200 abalone per company). The mature broodstock were induced to spawn separately and crossing was conducted following a paternal half-sib nested design (Morse et al. (1977)). Seawater containing gametes of one male was used to fertilize oocytes from three females randomly selected from the base population, for a total of 20 males and 60 females. The entire process of production of families took 3 months approximately. After settling, each full-sib family was cultured separately in 200-L tanks with continuous water flow and constant aeration for 14 months. Individuals from different families were mixed and transferred to baskets placed in a 10000-L raceway-type tank after they reached a size of $\geq$ 20 mm shell length ($\sim$ 14 months). After that, abalone were maintained during 3 years with continuous water flow, with constant aeration, at ambient temperature between 13 ℃ and 20 ℃ during the year. From 27 to 51 months of age, growth traits (the shell length, width and the total and flesh masses) were measured every 4 months in 15-30 individuals per full-sib family for the 60 families. We used the total mass (shell plus soft tissues) as the phenotypic measurements in our hybrid data set, for ease of implementation. Since the flesh mass (only soft tissues) and foot protein were measured in the original study as well, and these measures required sacrifice the animals, there were some missing values after first two measurement time points. The original data set also contained a number of other variables (position, module, tank, and so on) which may be included in the modeling as fixed effect covariate components.

The total of 318 abalones, which were from 12 full-sib families were selected from the original data set, 10 families with size of 30, one family with 10 and the last family with size of 8. The same numbers of salmons were picked up from the Atlantic salmon data set. They also came from 12 full-sib families with the same arrangements of family size as in the abalone data set.

The two separate data sets were merged together, such that each individual has measurements on the phenotypic trait observation plus the genetic marker data at four loci. Given the different sources of the data, it is reasonable to assume that genetic information is statistically independent of phenotypic values given the pedigree structure. More discussion about independence will be covered in chapter 4. Furthermore, we have assumed that individuals in distinct full sib families have no common ancestors.

Some preliminary study of the phenotypic observations has been carried out. Summary statistics for each family are listed in the Table 1.4. The boxplots of weights in figures 1.1 through 1.4 clearly indicate variation among families across all four time points. The differences in distribution between some families are quite large and they are easily identified visually, for example, family 62 and 25. However, the weight distributions of some families are very close, which is likely to make it difficult to distinguish those families successfully, based on phenotype alone.

Table 1.4: Summary statistics of weight measurements for Abalone

		Wgt1			Wgt2			Wgt3			Wgt4		
Family ID	Family Size	Obs	Mean	Std	Obs	Mean	Std	Obs	Mean	Std	Obs	Mean	Std
49	30	30	8.91	3.09	30	13.42	5.47	20	15.97	6.41	17	19.69	7.07
62	30	30	7.92	2.95	30	11.86	4.35	17	14.16	7.44	17	17.15	8.69
68	30	30	14.68	5.34	30	23.67	9.25	20	30.57	11.15	20	36.59	10.89
71	30	30	12.93	5.19	30	20.74	9.67	20	29.05	9.69	19	32.97	11.14
84	30	30	12.81	6.1	30	18.81	10.33	18	24.76	13.4	18	29.07	14.29
96	30	30	11.09	3.51	30	18.36	7.02	21	23.36	8.98	19	27.08	10.61
114	30	30	13.68	5.63	30	20.37	8.58	20	24.09	9.97	19	27.32	10.28
25	30	30	14.67	5.65	30	23.12	9.87	18	32.34	10.88	18	36.77	11.15
57	30	30	10.53	2.85	29	16.05	4.46	20	18.78	4.42	20	23.58	6.36
93	30	30	10.8	5.06	29	16.71	9.68	19	22.41	11.04	19	27.25	11.83
43	10	10	11.00	3.16	10	17.81	6.79	7	24.42	6.84	6	28.6	9.87
65	8	8	11.92	6.38	8	20.77	10.25	8	23.01	11.29	8	25.95	9.87

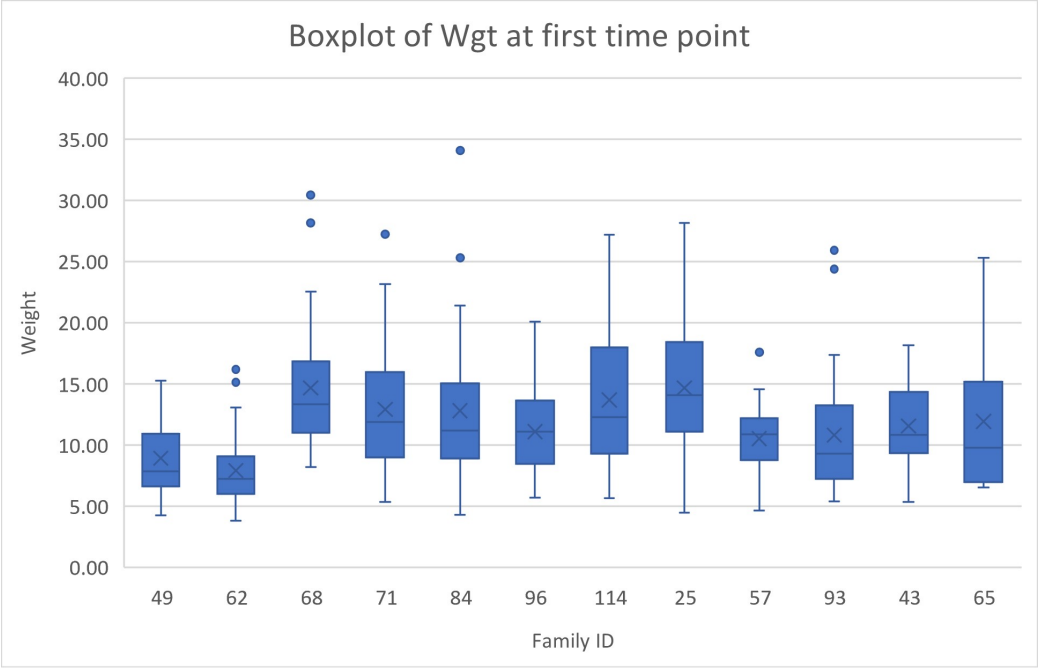

Figure 1.1: Weight values at the first measurement time point

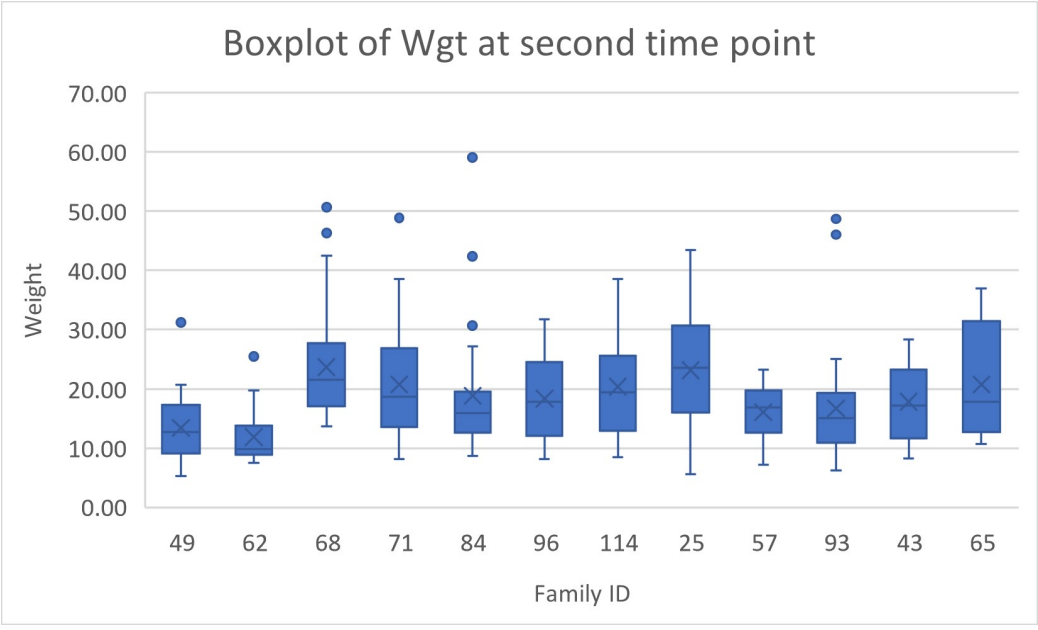

Figure 1.2: Weight values at the second measurement time point

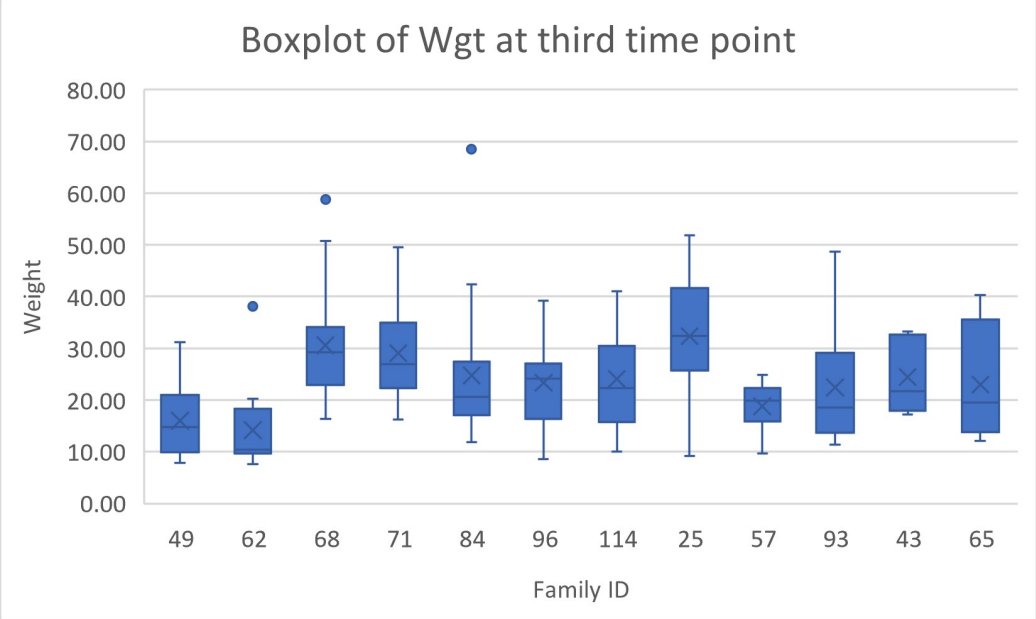

Figure 1.3: Weight values at the third measurement time point

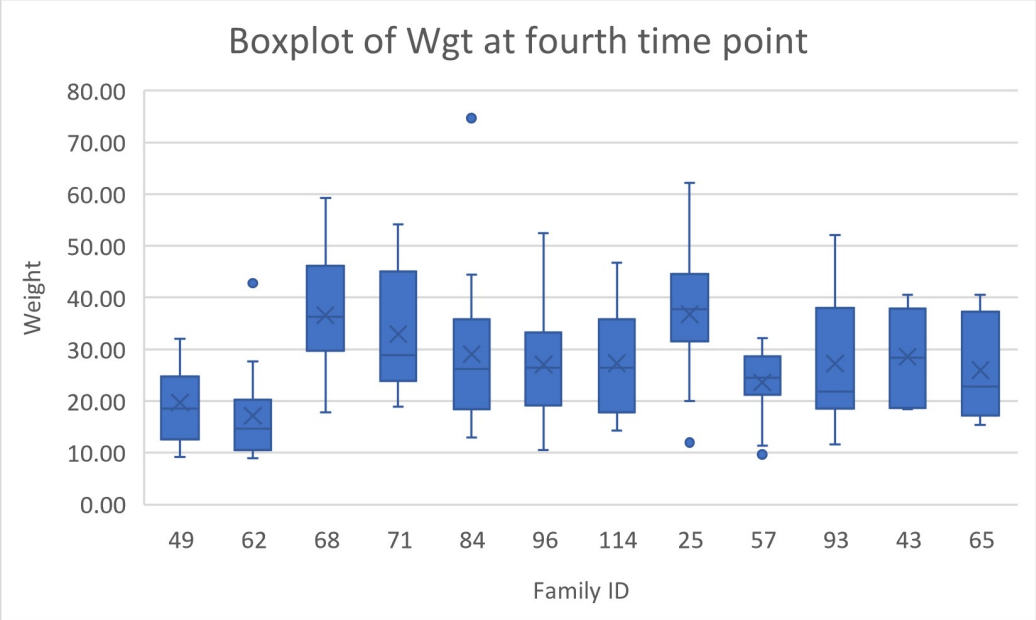

Figure 1.4: Weight values at the fourth measurement time point

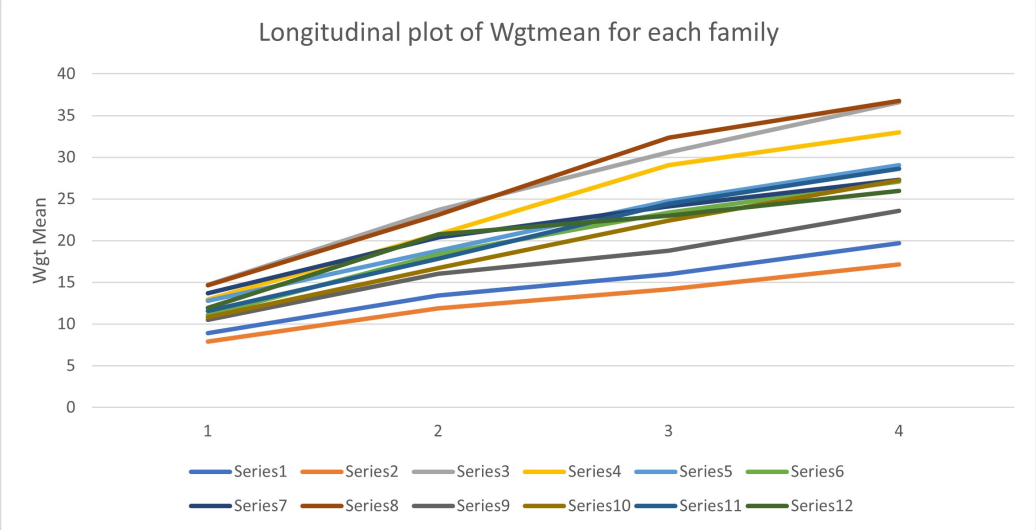

Figure 1.5: The average weight value for each family at four time points

Figure 1.5 shows the average weight in each family at the four measurements, and shows that differences persist over time. For example family 62 (Series2) consistently has the lowest average weight, while family 25 (Series8) has highest or second highest average weight throughout. The ability to detect differences will be largely dependent on the variation of individual measurements about the average curve. In principle one would like to fit a growth curve model to the longitudinal growth measurements, for example a four or five parameter logistic function. However, in the thesis we restrict attention to fitting scalar models at a single time point, or multivariate models for multiple measurements, albeit in the difficult situation of unknown pedigree structure.

In general, the plots and summary statistics give us some confidence that the phenotypic observations should be included in the model of pedigree structure and heritability estimation.

1.5 Outline

This thesis develops a hybrid Markov chain Monto Carlo (MCMC) sampling methods to obtain estimations of heritability using phenotypic observations and genetic marker data simultaneously when the pedigree is unknown. That methodology is then applied to the hybrid data that introduced in section 1.4.

The thesis is organized as follows. In chapter 2 we review three commonly used method to estimate heritability with the knowledge of relatedness or full pedigree structure. Parent-offspring regression has its natural reason to be used since the desire to estimate heritability comes from a specific interest in the resemblance between parent and offspring phenotypes. The statistical inferences are well developed because the compuatations are based on least-squares regression. Sib analysis gives us an alternative when the information from both parent and offspring are not available. The family struture permits one to partition the total phenotypic variance into within- and among-family components, both of which can be interpreted in terms of covariance between relatives. One way random effects analysis of variance (ANOVA) is designed to deal with this kind of data. Fitting a linear mixed model with restricted maximum likelihood method offers more power to deal with unbalanced design and complex but known pedigree structure.

In chapter 3, we present some possible ways to estimate heritability when pedigree is completely unknown. In case that only phenotypic observations are availalbe,

we could use expectation-maximization (EM) algorithm to fit a Gaussian mixture model. We define a heritability like object based on the variance decompositon. As we expect, the estimation results are disappointing since we deal with an extremely challenging situation. With the development of genetic markers, a two-step method has become popular to estimate the variance components of a quantitative trait. The pedigree can be reconstructed based on marker information, and then linear mixed model will provide the estimation of variance components in step 2. With sufficient marker data, this method guarantees an accurate result since the pedigree structure can be accurately reconstructed.

The proposed method is presented in chapter 4. We begin the chapter by defining the joint posterior distribution of pedigree and variance components parameters. We then introduce a moving algorithm and possible prior distributions of model parmeters. In contrast to the two-step method, we believe our method provides a more accurate estimate of heritability when marker information is limited. We also extend our model to incorporate observations at multiple time points. However, when using the first two time points, we don't obtain a significant improvement compared with estimating heritability at each time point separately and independently. Chapter 5 contains a summary of the thesis and some further suggestions.

Chapter 2

Estimation of Heritability with Known Pedigree Structure

There are many methods to estimate the components of variance for quantitative traits. In comparison among the alternatives, two issues should be carefully considered. First, attention has to be given to the kinds of relatives that should be analyzed. Certain kinds of relationships are observed more easily in some species than in others, and some types of phenotypic covariances between relatives are more likely to approximate desired quantities than others. Second, before performing the actual analysis, consideration should be given to the experimental design.

In this chapter, three commonly used methods for estimating narrow-sense heritabilities are introduced, the parent-offspring regression method, sib analysis with the analysis of variance method, and the linear mixed model with maximum likelihood/restricted maximum likelihood estimation. All of these techniques have the requirement of pairs of individuals of known relatedness, or full pedigree structure. At the end of this chapter, we show estimates of heritability for the abalone data from fitting a linear mixed model using restricted maximum likelihood (REML) method. These results can be used as the benchmark to evaluate the performance of our proposed model in chapter 4.

2.1 Parent-Offspring Regression

In some sense, the simplest and most commonly used design for estimating heritability is the regression of offspring phenotypes on those of their parents because the desire to obtain a heritability estimate comes from a specific interest in the resemblance between parent and offspring phenotypes. The regression approach has many advantages. First, the association of parent and offspring is the most easily identified relationship for many species. Second, the statistical inferences are well developed since the computations are based on least-squares regression. Third, the genetic covariance of parent-offspring relationship is not influenced by dominance and epistatic effects. Fourth, parent-offspring regression is the only simple method for heritability estimation that is not biased by the selection of parents (Lynch and Walsh (1998)).

2.1.1 Estimation Procedures with Balanced Data

For ease of presentation, we will consider the rather exceptional situation when all families have the same number of offspring, and to simplify discussion further, we will start with the assumption that only a single offspring and a single parent are observed in each family. The appropriate linear model for such a design is

$$Y_{oi} = \alpha + \beta_{op} Y_{pi} + e_i \tag{2.1}$$

where Y_{oi} and Y_{pi} represent the offspring and parent phenotypes for the ith family, α is the intercept, β_{op} is the regression coefficient, and e_i is the residual deviation from the regression. For this simple linear regression model, in statistical terms, the least-squares regression coefficient, $b_{op} = \mathrm{Cov}(Y_o, Y_p)/\mathrm{Var}(Y_p)$, provides an estimate of β_{op}. If there are no environmental causes of resemblance between parents and offspring, then according to the genetic covariance of (Parent-offspring) relatives that was shown in chapter 1, the expected regression slope b_{op} is

$$\mathrm{E}(b_{op}) = \frac{\sigma(Y_o, Y_p)}{\sigma^2(Y_p)} \simeq \frac{(\sigma^2_A/2) + (\sigma^2_{AA}/4) + (\sigma^2_{AAA}/8) + ...}{\sigma^2_Y} \tag{2.2}$$

For a male parent, it is generally expected that the covariance between parent and offspring environmental values is zero. This is not necessarily the case for a female parent, for whom environmental effects may be shared with the offspring through the non-genetic content of the egg. Thus, single parent-offspring regression usually involves the father, although if the regression slopes for father-offspring and mother-offspring are the same, then shared mother-offspring environmental values can be ruled out. Therefore, under the stated assumptions, a simple (possibly upperwardly biased) estimate of $h^2 = \sigma^2_A/\sigma^2_Y$ is twice the (single) parent-offspring regression, $2b_{op}$. The bias is upward as all terms in equation 2.2 are positive.

When both parents can be measured, we can regress offspring phenotypes on the average phenotypes of their parents (midparent value Y_{mp}) to possibly achieve greater precision. The linear model is slightly changed to

$$Y_{oi} = \alpha + \beta_{o\bar{p}}\Big(\frac{Y_{mi} + Y_{fi}}{2}\Big) + e_i \tag{2.3}$$

where Y_{mi} and Y_{fi} represent the phenotypes of mother and father for the ith family. The least-squares slope of the midparent-offspring regression, $b_{o\bar{p}}$, is a direct estimate of the heritability h^2. To see this, let us assume that the phenotypic variance is the same in both sexes and in both generations. The resemblance between relatives is also be assumed independent of their sex. Then,

$$\begin{aligned} b_{o\bar{p}} &= \frac{\mathrm{Cov}[Y_o, (Y_m + Y_f)/2]}{\mathrm{Var}[(Y_m + Y_f)/2]} \\ &= \frac{[\mathrm{Cov}(Y_o, Y_m) + \mathrm{Cov}(Y_o, Y_f)]/2}{[\mathrm{Var}(Y) + \mathrm{Var}(Y)]/4} \\ &= \frac{2\mathrm{Cov}(Y_o, Y_p)}{\mathrm{Var}(Y)} = 2b_{op} \end{aligned} \tag{2.4}$$

In order to get this result, we have assumed that there is no assortive mating, meaning that individuals with similar phenotypes do not mate preferentially, whereby $\mathrm{Cov}(Y_m, Y_f) = 0$. Therefore, we see that $b_{o\bar{p}} \simeq \sigma_A^2/\sigma_Y^2$, ignoring terms involving epistasis.

What happens when multiple (n) offspring are measured in each family. The expected phenotypic covariance of a parent i and the average of its $j = 1, ..., n$ offspring may be written $\sigma[(\sum_{j=1}^{n} Y_{oij}/n), Y_p]$. Since all n of the covariance terms contained in this expression have the same expected value, this reduces to $n\sigma(Y_o, Y_p)/n = \sigma(Y_o, Y_p)$,

which is the same as the expectation for single offspring. Thus, when family sizes are the same, the interpretation of a parent-offspring is the same whether individual offspring data or the progeny means are used in the analysis.

2.1.2 Precision of Estimates

In order to carry out inferences, it is necessary to ascertain the precision of heritability estimates. This is relatively easy to do with parent-offspring analysis. Under the assumption that x and y are bivariate normally distributed, the variance of the estimator is approximately

$$\sigma^2(b) \simeq \frac{\sigma^2(y)(1-\rho^2)}{n\sigma^2(x)} \tag{2.5}$$

where $\rho = \sigma(x,y)/[\sigma(x)\sigma(y)]$ is the correlation coefficient. A detailed derivation using the Delta method is provided in Appendix A. This result is first attributed to Pearson (1895), although there are very few details provided in that very short paper.

Provided the data have been measured or transformed so that the joint distribution of parent and offspring phenotypes is bivariate normal, the sampling variance of single parent-offspring regression estimate of heritability is approximately

$$\mathrm{Var}(b_{op}) \simeq \frac{(1-r_{op}^2)\mathrm{Var}(Y_o)}{N\mathrm{Var}(Y_p)} \tag{2.6}$$

where N is the number of parent-offspring pairs and r_{op} represents the correlation between single offspring and single parents. This expression reduced to $(1-r^2)/N$

when the phenotypic variances in the two generations are equal. This result also applies to regressions involving midparents if $\mathrm{Var}(\bar{Y}_p) = \mathrm{Var}(Y_p)/2$ is substituted for $\mathrm{Var}(Y_p)$ and $r_{o\bar{p}}$ for r_{op}.

Since the joint distribution of offspring and parent phenotypes is approximately bivariate normal, the sampling distribution of a regression coefficient will be approximately normal, and the usual asymptotic confidence interval $b \pm z_{\alpha/2} \times SE(b)$ will be appropriate for sufficiently large N.

For a regression involving single parent, the confidence interval for h^2 is twice that of the interval for the regression coefficient. For a midparent-offspring regression, the confidence interval for the slope is also the confidence interval for h^2.

2.2 Sib Analysis using the Analysis of Variance Method

When we are unable to collect the information from both parent and offspring, the analysis of contemporary relatives, in particular siblings, provides an alternative to parent-offspring regression in estimating quantitative-genetic parameters. There are three types of sib analyses: those employing half-sib families, those employing full-sib families, and those combining both (nested full-sib, half sib families). The family structure permits one to partition the total phenotypic variance into within- and among-family components, both of which can be interpreted in terms of covariance

between relatives. We will review the case of full-sib families, where individuals in different families are unrelated. One way random effects analysis of variance (ANOVA) is designed to deal with this kind of data.

2.2.1 Full-Sib Analysis

We focus on the simplest sib design which is to examine N full-sib families, each with n offspring and show how observed within- and among-family components of variance can be related to the underlying components of variance discussed before (additive effect, dominance effect, ...). We assume throughout that parents have been sampled randomly from the population and randomly mated, so that the simple interpretations of covariances between sibs given in the previous section on genetic covariances can be used here.

Under random mating and free recombination, two times the genetic covariance between full sibs is $\left[\sigma_A^2 + (\sigma_D^2/2) + (\sigma_{AA}^2/2) + (\sigma_{AD}^2/4) + (\sigma_{DD}^2/8)...\right]$. Thus, for the special situation in which dominance and epistasis are of minor importance, and common environmental effects don't contribute to the phenotypic resemblance of full sibs, $2\sigma(FS)$ provides an estimate of σ_A^2, where $\sigma(FS)$ denotes the expected phenotypic covariance between a pair of full sibs.

The traditional approach to analyzing such data is the one-way random effects

model

$$Y_{ij} = \mu + f_i + e_{ij} \tag{2.7}$$

where Y_{ij} is the phenotype of the jth offspring of the ith family, f_i is the effect of i'th family, and the residual error e_{ij} incoroporates all sources of variance from segregation, dominance, and environment. Stated another way, e_{ij} is the deviation of the phenotype of the ij'th individual from the expected value for the ith family. The usual assumptions are that the family effects are i.i.d. $N(0, \sigma_f^2)$, independent of the e_{ij}, which are i.i.d. $N(0, \sigma_e^2)$.

Under these assumptions, the phenotypic variance is partitioned into the between family variance and within family variance components.

$$\sigma_Y^2 = \sigma_f^2 + \sigma_e^2 \tag{2.8}$$

A second consequence of the model is that the phenotypic covariance between members of the same family equals to the between family variance. Noting that full sibs share family effects but have independent residual deviations, it follows that

$$\begin{aligned}\sigma(FS) &= \sigma(Y_{ij}, Y_{ik}) \\ &= \sigma[(\mu + f_i + e_{ij}), (\mu + f_i + e_{ik})] \\ &= \sigma(f_i, f_j) + \sigma(f_i, e_{ik}) + \sigma(e_{ij}, f_i) + \sigma(e_{ij}, e_{ik}) \\ &= \sigma_f^2 \end{aligned} \tag{2.9}$$

Thus, the covariance between full sibs equals the variance among families effect. This is a very useful identity because ANOVA provides a simple way to estimate σ_f^2. Hence, for the ideal case in which a character has no dominance and epistatic variance, the additive variance can be estimated as twice the among-family variance, i.e., $\widehat{\sigma}_A^2 = 2\widehat{\sigma}_f^2$.

In the following section, we will demonstrate the practical utility of this approach by showing how ANOVA generates estimates of the within- and among-family components of variance from phenotypic data.

2.2.2 One-Way Analysis of Variance

Consider the balanced design in which n individuals are sampled from each of N full sib families, so that there are a total $T = Nn$ individuals in the study. The total sum of squares is partitioned into an among- and within-family component in the usual fashion, as

$$SST = n\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 + \sum_{i=1}^{N}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_i\right)^2$$

$$= SS_f + SS_e \tag{2.10}$$

where $\bar{\bar{Y}}$ denotes the grand mean, and $\bar{Y}_i$ represents the observed family mean, $\bar{Y}_i = \sum_{j=1}^{n} Y_{ij}/n$. The within-family sum of squares (SS_e) is simply the sum of the squared deviations of individual measures from their observed family means, while the among-family sum of squares (SS_f) is the sum of the squared deviations of observed family means from the grand mean.

Assuming that the parents are a random sample of the population at large, the sum of squares can be used to obtain unbiased estimates of the within and among-family components of variance in the following way. We note first that the expected within-family sum of squares is

$$\mathrm{E}(SS_e) = \sum_{i=1}^{N}\mathrm{E}\big[\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2\big] = N(n-1)\sigma_e^2 \tag{2.11}$$

This result follows from the fact that $\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2/(n-1)$ is an unbiased estimate of the variance among sibs in the *ith* family and from our assumption that the variance within each family is equal to σ_e^2.

For the among-family sum of squares, similar reasoning leads to

$$\mathrm{E}(SS_f) = n\mathrm{E}\big[\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2\big] = n(N-1)\sigma^2(\bar{Y}_i) \tag{2.12}$$

where $\sigma^2(\bar{Y}_i)$ is the expected variance of the observed family means, here (with a balanced design) assumed to be the same for all families. The variance of observed family means is a function of the variance of the true family means, $(\mu + f_i)$, as well as of their sampling error, $\bar{e}_i = \bar{Y}_i - (\mu + f_i)$. Thus, assuming that the measurement error is independent of the family mean, we get

$$\sigma^2(\bar{Y}_i) = \sigma^2(\mu + f_i) + \sigma^2(\bar{e}_i) \tag{2.13}$$

The first term is the among-family variance since μ is constant. The second is the expected sampling variance of the mean, σ_e^2/n, leading to the final expression for $E(SS_f)$ as

$$\mathrm{E}(SS_f) = (N-1)(\sigma_e^2 + n\sigma_f^2) \tag{2.14}$$

Finally, the variance components can be expressed in terms of the expected sums of squares,

$$\sigma_f^2 = \frac{1}{n}\Big[\frac{\mathrm{E}(SS_f)}{N-1} - \frac{\mathrm{E}(SS_e)}{N(n-1)}\Big] \tag{2.15}$$

$$\sigma_e^2 = \frac{\mathrm{E}(SS_e)}{N(n-1)} \tag{2.16}$$

Writing the mean squares as the sums of squares divided by the corresponding degrees of freedom, $MS_f = SS_f/(N-1)$ and $MS_e = SS_e/(N(n-1))$, and substituting the observed mean squares for their expectations in equations 2.15 and 2.16, leads to the following unbiased method of moment estimators of σ_f^2, σ_e^2 and σ_Y^2,

$$\widehat{\sigma}_f^2 = \frac{MS_f - MS_e}{n} \tag{2.17}$$

$$\widehat{\sigma}_e^2 = MS_e \tag{2.18}$$

$$\widehat{\sigma}_Y^2 = \widehat{\sigma}_f^2 + \widehat{\sigma}_e^2 \tag{2.19}$$

The calculations are organized into the usual one-way ANOVA table (Table 2.1), which also includes the expected mean squares. This general procedure of estimating variance components from observed mean squares is an example of the method of moments, as the unknown variances can be expressed in terms of observable moments

Table 2.1: Summary of a one-way ANOVA involving N independent families, the ith of which contains n_i individuals. The total sample size is $T = \sum_{i=1}^{N} n_i$, and $n_0 = [T - (\sum n_i^2/T)]/(N-1)$, which reduces to n with equal family sizes.

Factor	**df**	**SS**	**MS**	**E(MS)**
Among-families	$N-1$	$SS_f = \sum_{i=1}^{N} n_i(\bar{Y}_i - \bar{\bar{Y}})^2$	$SS_f/(N-1)$	$\sigma_e^2 + n_0\sigma_f^2$
within-families	$T-N$	$SS_e = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$	$SS_e/(T-N)$	σ_e^2
Total	$T-1$	$SST = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\bar{Y}})^2$	$SST/(T-1)$	σ_Y^2

The quantity

$$ICC = \frac{\text{Var}(f)}{\text{Var}(Y)} \tag{2.20}$$

is the intraclass correlation (Fisher (1919); Fisher et al. (1934)). It provides an estimate of the fraction of the phenotypic variance attributable to differences among families. Recalling from above that $\sigma_f^2 = \sigma(FS) \simeq \sigma_A^2/2$, the full-sib ANOVA estimator of the heritability is

$$\widehat{h^2} \simeq 2ICC \tag{2.21}$$

This expression again assumes that contributions from epistatic genetic variance are small and that there is no dominance variance.

A difficulty with method of moments based estimates of variance components is that the estimate of σ_f^2 can be negative due to the sampling error. In such cases the estimates are typically not reported, or the model is deemed to be inappropriate.

When we assume that there are no dominance effects and no common environmental effects, the analysis of variance for full-sib families provides an estimate of the heritability, which is equal to twice the interclass correlation coefficient. However, since the significance of the dominance and environmental components of variance are generally unknown, it is best to avoid the exclusive use of full sibs to estimate

heritability. The analysis of variance for half sibs also can be used to estimate heritability in the same fashion as for full-sibs. The common environmental effects (maternal effects) can be eliminated by using a paternal half-sib mathing design. Analysis of variance for a nested full-sib half-sib mating design is more robust method which provides information on the relative significance of the components of variance associated with dominance and common environmental effects. The detailed explanation can be found in Lynch and Walsh (1998).

2.3 Linear Mixed Model with Restricted Maximum Likelihood Method

ANOVA estimators estimate variance components by equating observed mean squares to expressions describing their expected values, these being functions of the variance components. ANOVA provides the unbiased estimators for the variance components regardless of whether the data are normally distributed, but it has limitations. Firstly, the variety of relatives that could be observed can often not be analyzed jointly with ANOVA. Secondly, a balanced sample size is generally required in the process of estimation using ANOVA. While ANOVA sums of squares have been proposed to account for unbalanced data (Searle et al. (2009)), their sampling properties are poorly understood.

2.3.1 Maximum Likelihood and Restricted Maximum Likelihood Estimates

Maximum likelihood (ML) and restricted maximum likelihood (REML) estimators are ideal for the unbalanced designs that arise in quantitative genetics, since they don't need any special demands on design or balance of data. ML/REML methods provide a powerful approach to estimate variance components in populations with complex but known pedigrees.

Consider a column vector $\mathbf{y}$ containing the phenotypic observations for a trait measured in n individuals. We assume that these observations are described adequately by a linear model with a $p \times 1$ vector of fixed effects ($\boldsymbol{\beta}$) and a $q \times 1$ vector of random effects ($\mathbf{u}$). The elements of the vector $\mathbf{u}$ of random effects can be considered as genetic effects. In the matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{2.22}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are respectively n $timesp$ and $n \times q$ incidence matrices, and $\mathbf{e}$ is the $n \times 1$ column vector of residual deviations assumed to be distributed independently of the random genetic effects. The distributions of the random effects $\mathbf{u}$ and $\mathbf{e}$ are almost always assumed to be independent multivariate normal with mean vectors $\mathbf{0}$ and diagonal covariance matrices. The maximum likelihood method estimates parameters by maximizing the likelihood of the observed data. All the fixed effects are assumed to be known without error in the usual maximum likelihood approach, although this

is rarely true in practice. Therefore, ML estimators tend to yield biased estimates of variance components. The estimates of the residual variance tend to be downwardly biased because the observed deviations of individual phenotypic values from an estimated population mean tend to be smaller than their deviations from the true mean.

Contrary to the behaviour of ML estimators, restricted maximum likelihood (REML) estimators maximize only the portion of the likelihood that doesn't depend on the fixed effects. In this sense, REML is a restricted version of ML. REML doesn't always eliminate all of the bias in parameter estimation, since many methods for obtaining REML estimates can't give negative estimates of a variance component. However, this source of bias also happens with ML, so REML is generally thought to be the better method for analyzing data sets with complex structure. For the completely balanced design, REML method provides identical results to the classical ANOVA.

Foulley (1993) introduced a useful pedagogical connection between ML and REML, beginning from a very simple application, the estimation of the mean and variance of a set of independent observations. We use this example to show how ML and REML procedures can be used to estimate variance components, and to illustrate that the two estimators can differ.

The mixed model can be written as

$$\mathbf{y} = \mathbf{1}\mu + e \tag{2.23}$$

where μ is the population mean (the only fixed effect), $\mathbf{1}$ is an $n \times 1$ column vector of ones (equivalent to the design matrix $\mathbf{X}$), and the covariance matrix of residuals about the mean is assumed to be $\mathbf{R} = \sigma^2\mathbf{I}$.

Assuming the phenotypes are independent of each other and normally distributed, the log-likelihood function of the observed data $\mathbf{y}$ is given by

$$l(\mu, \sigma^2|\mathbf{y}) \propto -\frac{n}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{1}{n\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \tag{2.24}$$

The log-likelihood is proportional to the joint density of the observations given the parameters, although the likelihood is considered as a function of the parameters conditioned on the observed data y_i, $i = 1, 2, \ldots, n$. Maximum likelihood estimators estimate parameters as those values which maximize the log-likelihood. Equivalently, the ML method estimates estimate parameters as those values which maximize the probability (the joint density) of the observed data.

Letting $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, we have

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{\mu})^2 &= \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\bar{y} - \mu)^2 + 2(\bar{y} - \mu)\sum_{i=1}^{n}(y_i - \bar{y}) \\ &= n[V + (\bar{y} - \mu)^2] \end{aligned} \tag{2.25}$$

Substituting this into equation 2.24, the log-likelihood can be expressed as

$$l(\mu, \sigma^2|\mathbf{y}) \propto -\frac{n}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{V + (\bar{y} - \mu)^2}{\sigma^2}\right] \tag{2.26}$$

Setting partial derivatives with respect to μ and σ^2 equal to 0 gives the likelihood equations

$$\frac{\partial l(\mu, \sigma^2|\mathbf{y})}{\partial \mu} = \frac{n(\bar{y} - \mu)}{\sigma^2} \tag{2.27}$$

$$\frac{\partial l(\mu, \sigma^2|\mathbf{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2}\left[1 - \frac{V + (\bar{y} - \mu)^2}{\sigma^2}\right] \tag{2.28}$$

By setting the partial derivative equations equal to zero and solving these equations gives the MLE's for the population mean and variance that maximize the likelihood function given the observed data $\mathbf{y}$. We obtain an estimator for the mean that is completely independent of the variance,

$$\widehat{\mu} = \bar{y} \tag{2.29}$$

which shows that the sample mean is the ML estimate of the parametric value. However, the solution to equation 2.28,

$$\widehat{\sigma}^2 = V + (\bar{y} - \mu)^2 \tag{2.30}$$

is not independent with the estimated mean, $\bar{y}$, unless the estimated mean happens

to coincide perfectly with the true mean μ. The maximum likelihood estimator of σ^2, is obtained by assuming that the mean is estimated without error, resulting in

$$\widehat{\sigma}^2 = V \tag{2.31}$$

This result gives a downwardly biased estimate of the true variance σ^2, the bias being negative since the term ignored in equation 2.30 is necessarily non-negative.

REML removes this bias by accounting for the error in the estimation of μ. From equation 2.30, the expected amount by which $\widehat{\sigma}^2$ underestimates σ^2 is the expected value of $(\bar{y} - \mu)^2$, which is simply the sampling variance of the mean, σ^2/n. Thus, an improved estimator is

$$\widehat{\sigma}^2 = V + \mathrm{E}[(\bar{y} - \mu)^2] = V + \frac{\sigma^2}{n} \tag{2.32}$$

We don't know exactly what this bias is because we don't know true value of σ^2 with certainty. However, the bias is estimable because we have a preliminary estimate of σ^2, the maximum likelihood estimate V. Then, starting with an initial estimate of $\widehat{\sigma}^2(0) = V$, a second improved estimate of variance is

$$\widehat{\sigma}^2(1) = V + \frac{\hat{\sigma}^2(0)}{n} = V + \frac{V}{n}$$

Just as this changes the estimate of the variance, it also changes the estimate of

$(\bar{y} - \mu)^2$. Hence, a third estimate of σ^2 would be

$$\widehat{\sigma}^2(2) = V + \frac{\widehat{\sigma}^2(1)}{n} = V + \frac{V + (V/n)}{n}$$

This sequence suggests an iterative approach to estimate the variance σ^2

$$\widehat{\sigma}^2(t+1) = V + \frac{\widehat{\sigma}^2(t)}{n} \tag{2.33}$$

The final (fixed point) solution to this equation, $\widehat{\sigma}^2$, is obtained by setting $\widehat{\sigma}^2(t+1) = \widehat{\sigma}^2(t)$, yielding

$$\widehat{\sigma}^2 = \frac{n}{n-1}V = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{2.34}$$

which is the unbiased estimator of the variance (sample variance) that we normally use. This fixed point solution assumes a limit, which is guaranteed by noting the Maclauren series for $\frac{1}{1-1/n} = \sum_{k=0}^{\infty} \frac{1}{n^k}$.

With models containing multiple fixed effects, closed solutions are not usually possible to obtain, particularly in complex pedigree analyses involving unbalanced data. However, iterative procedures can still yield solutions that are asymptotically unbiased.

2.3.2 ML Estimates of Variance Components in Linear Mixed Model

Let's reconsider the general linear mixed model, $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e}$, and assume that $\mathbf{u} \sim \text{MVN}(\mathbf{0}, \mathbf{G})$ independent of $\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{R})$. Under this model, $\mathbf{y}$ is also multivariate normal, with mean $\mathbf{X}\beta$ and variance-covariance matrix $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$. The log-likelihood of β and $\mathbf{V}$ given the observed data $(\mathbf{X}, \mathbf{y})$ is

$$l(\beta, \mathbf{V}|\mathbf{X}, \mathbf{y}) \propto -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) \tag{2.35}$$

Now, we consider $\mathbf{u} = \mathbf{a}$ to be the vector of additive genetic values. The variance components that we wish to estimate are embedded with $\mathbf{G}$ and $\mathbf{R}$. We assume that $\mathbf{G} = \sigma_A^2\mathbf{A}$, where $\mathbf{A}$ is the additive genetic relationship matrix, and $\mathbf{R} = \sigma_E^2\mathbf{I}$.

A potential computational difficulty in carrying out maximum likelihood estimation is the need to invert and find the determinant of $\mathbf{V}$. However, $\mathbf{V}$ has a particular patterned form and if $\mathbf{R}$ and $\mathbf{G}$ are of substantially lower dimension than $\mathbf{V}$, then the patterned structure provides a computationally efficient means of calculating the inverse and determinant of $\mathbf{V}$. This special structure will be utilized in chapter 4.

This setup can be extended to estimate additional variance components by using more generalized model

$$\mathbf{y} = \mathbf{X}\beta + \sum_{i=1}^{m} \mathbf{Z}_i\mathbf{u}_i + \mathbf{e} \tag{2.36}$$

where the m vectors of random effects ($\mathbf{u}_i$) are assumed to be uncorrelated, with $\mathbf{u}_i \sim \text{MVN}(0, \sigma_i^2 \mathbf{B}_i)$ and $\mathbf{B}_i$ being a matrix of known constants. This more general model can incorporate estimates of dominance and other nonadditive variances, maternal environmental effects, etc. The log-likelihood is still given by equation 2.35, but now the covariance matrix $\mathbf{V}$ is

$$\mathbf{V} = \sum_{i=1}^{m} \sigma_i^2 \mathbf{Z}_i \mathbf{B}_i \mathbf{Z}_i^T + \sigma_E^2 \mathbf{I} \tag{2.37}$$

We start with the partial derivatives of the log-likelihood with respect to the vector of fixed effects, β. This derivative involves only the last term of equation 2.35. Using a general result for matrix derivatives (Morrison et al. (1976); Graham (2018)), we get

$$\frac{\partial[(\mathbf{y} - \mathbf{X}\beta)^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta)]}{\partial \beta} = -2\mathbf{X}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta) \tag{2.38}$$

which yields

$$\frac{\partial l(\beta, \mathbf{V} | \mathbf{X}, \mathbf{y})}{\partial \beta} = \mathbf{X}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta) \tag{2.39}$$

Now we consider derivatives with respect to the variance components. We first assume the simple case of only two unknown variances, σ_A^2 and σ_E^2. Writing $\mathbf{V}$ in terms of these two components, we have $\mathbf{V} = \sigma_A^2 \mathbf{Z}\mathbf{A}\mathbf{Z}^T + \sigma_E^2 \mathbf{I}$. Using the notation of σ_i^2 to denote the variance component being estimated, we have

$$\frac{\partial \mathbf{V}}{\partial \sigma_i^2} = \mathbf{V}_i = \begin{cases} \mathbf{I} & \text{when } \sigma_i^2 = \sigma_E^2 \\ \mathbf{ZAZ}^T & \text{when } \sigma_i^2 = \sigma_A^2 \end{cases} \tag{2.40}$$

Obtaining the partial derivatives with respect to the variance σ_A^2 and σ_E^2 involves use of two general results from matrix theory (Searle and Khuri (2017)). Specifically, if $\mathbf{M}$ is a square matrix whose elements are functions of a scalar variable x, then

$$\frac{\partial \ln|\mathbf{M}|}{\partial x} = \text{tr}\left(\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial x}\right) \tag{2.41}$$

$$\frac{\partial \mathbf{M}^{-1}}{\partial x} = -\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial x}\mathbf{M}^{-1} \tag{2.42}$$

where *tr*, the trace, denotes the sum of the diagonal elements of a square matrix.

The general partial derivative equation can be expressed as

$$\begin{aligned} \frac{\partial l(\beta, \mathbf{V}|\mathbf{X}, \mathbf{y})}{\partial \sigma_i^2} &= -\frac{1}{2}\text{tr}(\mathbf{V}^{-1}\mathbf{V}_i) + \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T\mathbf{V}^{-1}\mathbf{V}_i\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &\quad + \frac{1}{2}(\hat{\beta} - \beta)^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{V}_i\mathbf{V}^{-1}\mathbf{X}(\hat{\beta} - \beta) \end{aligned} \tag{2.43}$$

where $\mathbf{V}_i$ is given by equation 2.40. Equations 2.39 and 2.43 are directly analogous to equations 2.27 and 2.28 derived before. Note that $\mathbf{V}_i$ is a fixed matrix of known constants, whereas $\mathbf{V} = \sigma_A^2\mathbf{ZAZ}^T + \sigma_E^2\mathbf{I}$ is a function of the variance-component estimates. More generally, with m random effects plus a residual error, equation 2.43

holds for each of the $m+1$ variance components with

$$\frac{\partial \mathbf{V}}{\partial \sigma_i^2} = \mathbf{V}_i = \begin{cases} \mathbf{I} & \text{when } \sigma_i^2 = \sigma_E^2 \\ \mathbf{Z}_i \mathbf{B}_i \mathbf{Z}_i^T & \text{otherwise} \end{cases} \tag{2.44}$$

The maximum likelihood (ML) estimators are obtained by setting equations 2.39 and 2.43 equal to zero and solving. Using equation 2.39, the ML estimate of the vector of fixed effects is

$$\hat{\beta} = (\mathbf{X}^T \widehat{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \widehat{V}^{-1} \mathbf{y} \tag{2.45}$$

This is the best (minimum variance) linear unbiased estimator (BLUE), of β. The ML estimators for the variance components are obtained by setting $\beta = \hat{\beta}$ in equation 2.43, which makes the last term equal to zero. Rearranging, we have

$$\mathrm{tr}(\widehat{\mathbf{V}}^{-1} \mathbf{V}_i) = (\mathbf{y} - \mathbf{X}\hat{\beta})^T \widehat{\mathbf{V}}^{-1} \mathbf{V}_i \widehat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{2.46}$$

We can simplify equation 2.46 by defining a new matrix $\mathbf{P}$ (Lynch and Walsh (1998)) as

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \tag{2.47}$$

It follows that

$$\mathbf{P}\mathbf{y} = \mathbf{V}^{-1}\mathbf{y} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} = \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{2.48}$$

Using this identity, equation 2.46 can be more compactly written as

$$\mathrm{tr}(\widehat{\mathbf{V}}^{-1}\mathbf{V}_i) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{V}_i\widehat{\mathbf{P}}\mathbf{y} \tag{2.49}$$

where we use $\widehat{\mathbf{P}}$ to indicate that $\mathbf{P}$ is a function of $\mathbf{V}$, it depends on the variance components that we are trying to estimate. Even though it may not be immediately apparent, equation 2.49 can be considered as a generalized version of equation 2.31.

In summary, the ML estimates solve the equation 2.45 (for fixed effects) and a set of equations for variance components (equation 2.49). In general, with m random effects plus a residual, the set of $m + 1$ ML equations for the variances of random effects are

$$\mathrm{tr}(\widehat{\mathbf{V}}^{-1}) = \mathbf{y}^T\widehat{\mathbf{P}}\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_E^2 \tag{2.50}$$

$$\mathrm{tr}(\widehat{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_i^2,\ \mathrm{i} = 1 \ldots \mathrm{m} \tag{2.51}$$

where $\widehat{\mathbf{P}}$ uses

$$\widehat{\mathbf{V}} = \sum_{i=1}^{m} \widehat{\sigma}_i^2\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T + \widehat{\sigma}_E^2\mathbf{I} \tag{2.52}$$

These solutions have two troublesome properties. First, the ML vector of fixed

effects $\hat{\beta}$ is a function of the variance-covariance matrix $\widehat{\mathbf{V}}$, which contains the variance components that we wish to estimate. Second, because the solutions involve the inverse of $\widehat{\mathbf{V}}$, they are nonlinear functions of the variance components. Therefore, there is no simple one-step solution. ML estimation of β, σ_A^2 and σ_E^2 requires an iterative procedure.

2.3.3 Restricted Maximum Likelihood (REML)

An extensive discussion of the estimation of random effects, including restricted maximum likelihood, is provided by Robinson et al. (1991). REML is based on a linear transformation of the observation vector $\mathbf{y}$ that removes the fixed effects from the model. We can use a transformation matrix $\mathbf{K}$ associated with the design matrix $\mathbf{X}$ for the model under consideration such that

$$\mathbf{KX} = \mathbf{0} \tag{2.53}$$

Applying this transformation matrix to the mixed model yields

$$\begin{aligned} \mathbf{y}^* = \mathbf{Ky} &= \mathbf{K}(\mathbf{X}\beta + \mathbf{Za} + \mathbf{e}) \\ &= \mathbf{KZa} + \mathbf{Ke} \end{aligned} \tag{2.54}$$

The $\mathbf{y}^*$ can be viewed as the residual deviations from the estimated fixed effect. REML

estimates of variance components are equivalent to ML estimates of the transformed variables. Thus, we can use ML solutions outlined above by making the following substitutions:

$$\mathbf{Ky} \text{ for } \mathbf{y}, \quad \mathbf{KX} = \mathbf{0} \text{ for } \mathbf{X}, \quad \mathbf{KZ} \text{ for } \mathbf{Z}, \quad \mathbf{KVZ}^T \text{ for } \mathbf{V} \tag{2.55}$$

The REML equations can actually be expressed directly in terms of $\mathbf{V}$, $\mathbf{y}$, and $\mathbf{P}$, without finding the matrix $\mathbf{K}$. This result follows from an identity, proven in Searle et al. (2009), that $\mathbf{K}$ satisfies

$$\mathbf{P} = \mathbf{K}^T(\mathbf{KVK}^T)^{-1}\mathbf{K} \tag{2.56}$$

Noting that

$$(\mathbf{y}^*)^T(\mathbf{V}^*)^{-1}\mathbf{y}^* = (\mathbf{y}^T\mathbf{K}^T)(\mathbf{KVK}^T)^{-1}(\mathbf{Ky}) = \mathbf{y}^T\mathbf{Py} \tag{2.57}$$

and substituting the expressions given as 2.55 into equation 2.46, after some rearrangement, the ML equations yield the REML estimators,

$$\text{tr}(\widehat{\mathbf{P}}) = \mathbf{y}^T\widehat{\mathbf{P}}\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_E^2 \tag{2.58}$$

$$\text{tr}(\widehat{\mathbf{P}}\mathbf{ZAZ}^T) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{ZAZ}^T\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_A^2 \tag{2.59}$$

Note that the REML does not give estimates of β, since the fixed effects are removed from the model by setting $\beta^* = \mathbf{0}$.

Equation 2.36 can expend to the general case with m uncorrelated random effects since the transformation $\mathbf{y}^* = \mathbf{Ky}$ satisfying equation 2.53 only depends on the design matrix.

$$\mathbf{y}^* = \sum_{i=1}^{m} \mathbf{KZ}_i\mathbf{u}_i + \mathbf{Ke} \tag{2.60}$$

and the REML equations for the $m+1$ variance components become

$$\mathrm{tr}(\widehat{\mathbf{P}}) = \mathbf{y}^T\widehat{\mathbf{P}}\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_E^2 \tag{2.61}$$

$$\mathrm{tr}(\widehat{\mathbf{P}}\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_i^2,\ \mathrm{i} = 1 \ldots \mathrm{m} \tag{2.62}$$

where $\widehat{\mathbf{P}}$ is now a function of $\widehat{\mathbf{V}} = \sum_{i=1}^{m} \hat{\sigma}_i^2\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T + \hat{\sigma}_E^2\mathbf{I}$.

2.3.4 Solving the ML/REML Equations

The closed analytical solutions for the ML/REML equations are only available in very special cases (e.g., certain completely balanced designs). In principle, the solutions can be derived by performing an exhaustive grid search - computing the log-likelihood of the data at each point on a grid covering the entire range of parameter space, and letting the solution be defined by the point on the grid giving the largest log-likelihood. However, this procedure is impractical, as if β contains more than a few elements,

each element adds to the dimensionality of the search. Under REML, the dimensionality of parameter space can be greatly reduced, but the likelihood function is considerably more complicated to compute.

A variety of iterative techniques for solving ML/REML equations have been developed based on different modifications of two basic approaches: the Newton-Raphson algorithm and the Expectation-maximization algorithm. Both methods start with preliminary estimates of the parameters, and using information on the slope of the likelihood surface, estimates are then moved in a direction that increases the log-likelihood of the data. The updated estimates are subsequently modified in an iterative fashion, until a satisfactory degree of convergence on a final set of estimates has been obtained. The search for ML/REML solutions avoids spending huge amounts of computational time in regions of low likelihood. However, these methods are not guaranteed to reach a global maximum of the likelihood function, but issues with multi-modality can be explored through the use of different starting value. All of the methods are very computationally intensive when large pedigrees are involved, since they usually require the inversion of large matrices at each step. Detailed reviews can be found in Searle et al. (2009).

The lme4 package (Bates et al. (2014)) for R provides functions to fit and analyze linear mixed models, generalized linear mixed models and nonlinear mixed models. The model is described in an *lmer* using formula which in this case includes both

fixed and random effects terms. The formula and data together determine a numerical representation of the model from which the profiled deviance or the profiled REML criterion can be evaluated as a function of some of the model parameters. The *lmer* estimates of variance components for our transformed abalone data (using the logarithm function to get closer to a normal distribution) are presenting in Table 2.2. In calculating these estimates, the true pedigree structure was used and so the estimated heritabilities can be used as the benchmark to evaluate our proposed method in a later chapter. We used the location of the tank in which each full-sib family was held for the first 14 months of life, and the densities in which each families were held during this period, as two fixed effects. The R results indicated that neither of these fixed effect covariates are significant in the model. The heritabilities might be over estimated because we have assumed that the family effect is dominated by additive genetic effect, with nonadditive and epistatic effects having been ignored.

Table 2.2: Estimates of variance components and heritability by using linear mixed model with REML method. Wgt represents the measurment of weight.

Time point	**Variance of family effect**	**Variance of residual**	**Heritability**
Wgt1	272.7080	1413.9953	0.3234
Wgt2	353.7641	1691.4958	0.3459
Wgt3	640.3033	1565.8135	0.5805
Wgt4	535.5469	1427.7274	0.5456

Chapter 3

Estimation of Heritability with Unknown Pedigree Structure

One of the greatest technical limitations of all methods of estimating genetic variance components that were introduced in the previous chapter is their requirement for knowledge of the relationships among the individuals recorded. In natural populations, detailed information of pedigree is absent in all but the most carefully studied populations. Only in special cases can pedigrees be determined from observation of mating activities (Garant et al. (2004)), but this usually requires long-term intensive observation and still may not be entirely reliable.

In this chapter, we provide some possible solutions to estimate genetic variance components when pedigree structure is completely unknown. First, in case that only phenotypic observations are available, finite mixture models shed some light on the problem at hand. This can be considered as a worst-case scenario as it uses only the minimum of information to estimate heritability. Second, when molecular genetic marker data are available, we introduce two commonly used approaches to estimate quantitative genetic characteristics, including the heritability.

3.1 Gaussian Finite Mixture Models

Finite mixture models have been successfully applied in many fields which include agriculture, astronomy, bioinformatics, biology, economics, genetics and so on. This is because finite mixtures of distributions can be used to provide computationally convenient representations for modeling complex distributions of data arising from random phenomena. Finite mixture models made their first recorded appearance in the modern statistical literature in the nineteenth century in a paper by Newcomb (1886). He suggested an iterative reweighting scheme to compute the common mean of a mixture with known proportions from a finite number of univariate normal distributions with known variances. A few years later, Pearson (1894) fitted a mixture of two normal probability density functions with different means μ_1 and μ_2 and different variances σ_1^2 and σ_2^2 to some crab data. Another early reference on mixtures is Holmes (1892), who brought in the concept of mixtures of populations in his suggestion that an average alone was inadequate in consideration of wealth disparity.

The use of maximum likelihood for fitting mixture models received little attention until the 1960s. Dempster et al. (1977) formalized an iterative estimation scheme in a general context through their expectation-maximization (EM) algorithm, which provided theoretical convergence properties of maximum likelihood estimation for the mixture problem. The EM algorithm proved to be a timely catalyst for further research into the applications of finite mixture models.

Assume we observe $Y_1, ..., Y_n$ and that each Y_i is sampled from one of K mixture component distributions. Associated with each random variable Y_i is a label $Z_i \in \{1, ..., K\}$ which indicates which component Y_i belongs to. We don't observe Z_i, which are referred to as latent variables. In the case of discrete random variables, the marginal probability of Y_i can be expressed as

$$P(Y_i = y) = \sum_{k=1}^{K} P(Z_i = k)P(Y_i = y|Z_i = k) = \sum_{k=1}^{K} \pi_k P(Y_i = y|Z_i = k) \tag{3.1}$$

where the π_k are referred to as mixing proportions or mixture weights, and π_k represents the probability that Y_i follows the k'th mixture component distribution $P(Y_i|Z_i = k)$. The mixing proportions are nonnegative and sum to one, $\sum_{k=1}^{K} \pi_k = 1$, and the $P(Y_i|Z_i = k)$ are distinct probability distributions for $k = 1, \ldots, K$. The latent variables Z_i are random variables taking values on the integers $1, 2, \ldots K$, with $P(Z_i = k) = \pi_k$, for $k = 1, 2, \ldots, K$.

3.1.1 A connection between the Gaussian Mixture Model and Heritability

Now assume we have a Gaussian mixture model, where the k'th component distribution is $N(\mu_k, \sigma_k^2)$, with k'th mixting proportion π_k. In this scenario, the conditional distribution of Y_i given that $Z_i = k$ is $N(\mu_k, \sigma_k^2)$, so that the marginal density of Y_i is

$$f_{Y_i}(y) = \sum_{k=1}^{K} P(Z_i = k) f_{Y_i}(y|Z_i = k) = \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k} \phi\Big(\frac{y - \mu_k}{\sigma_k}\Big) \tag{3.2}$$

where $f_{Y_i}(y|Z_i = k)$ denotes the conditional density of Y_i given $Z_i = k$ and $\phi()$ is the standard normal density function.

When the variances of the component distribution are assumed to be same, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2 = \sigma^2$, an estimate of heritability is suggested by the usual decomposition of variance

$$\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y|Z)) + \mathrm{Var}(\mathrm{E}(Y|Z)) \tag{3.3}$$

Under the constant variance assumption, $\mathrm{Var}(Y|Z = k) = \sigma^2$ and $\mathrm{E}(Y|Z = k) = \mu_k$. In the mixture model the means are fixed constants. However, the variance decomposition formula suggests to replace $\mathrm{Var}(\mathrm{E}(Y|Z))$ by the variance of the K means $\mu_1, \mu_2, ..., \mu_K$, using the discrete distribution $P(\mu = \mu_k) = \pi_k$, $k = 1, 2, \ldots, K$.

If estimates of μ_k and σ^2 are available through the EM or some other algorithm, these quantities can be estimated using $\widehat{\mathrm{E}}(\mathrm{Var}(Y|Z)) = \widehat{\sigma}^2$, and with $\widehat{\mathrm{Var}}(\mathrm{E}(Y|Z))$ as the empirical variance of $\widehat{\mu}_1, \widehat{\mu}_2, ...\widehat{\mu}_K$.

Motivated by the variance decomposition formula, a definition for heritability in the mixture model is suggested by consideration of the one way fixed effects ANOVA model, which is the analogue of the finite mixture model when component (family) memberships are known.

In the previous chapter, we discussed how to estimate variance components by using the ANOVA method for the random effect model. The mean squares of among-family and the mean squares of within-family were used to estimate the intraclass correlation, as follows:

$$\widehat{\sigma}_f^2 = \frac{MS_f - MS_e}{n_o} \tag{3.4}$$

$$\widehat{\sigma}_e^2 = MS_e \tag{3.5}$$

$$\widehat{ICC} = \frac{\widehat{\sigma}_f^2}{\widehat{\sigma}_f^2 + \widehat{\sigma}_e^2} \tag{3.6}$$

When the component variances are equal, the Gaussian finite mixture model is a one way fixed effects ANOVA model, but with unknown component membership. The ANOVA method can be used for the fixed effect model as well as for the random effects model. The total sum of squares for the fixed effects model is decomposed into between and within sum of squares as for the random effects model.

The sum squares of between treatment groups is

$$SSTR = \sum_{k=1}^{K} n_k(\bar{Y}_k - \bar{\bar{Y}})^2 = \sum_{k=1}^{K} n_k \bar{Y}_k^2 - n_T \bar{\bar{Y}}^2 \tag{3.7}$$

where n_k and $\bar{Y}_k$ are the sample size and sample mean in k'th group, $\bar{\bar{Y}}$ is the sample mean for all observations in the K groups, and n_T is the total number of observations. Therefore, the expectation of SSTR is:

$$\mathrm{E}(SSTR) = \mathrm{E}\Big[\sum_{k=1}^{K} n_k \bar{Y}_k^2 - n_T \bar{\bar{Y}}^2\Big] = \Big[\sum_{k=1}^{K} n_k \mathrm{E}(\bar{Y}_k^2)\Big] - n_T \mathrm{E}(\bar{\bar{Y}}^2) \tag{3.8}$$

In general, we know $\mathrm{E}(Y^2) = \mathrm{Var}(Y) + (\mathrm{E}(Y))^2$, and also

$$\mathrm{E}(\bar{Y}_k) = \mu_k \tag{3.9}$$

$$\mathrm{Var}(\bar{Y}_k) = \frac{\sigma^2}{n_k} \tag{3.10}$$

$$\mathrm{E}(\bar{\bar{Y}}) = \frac{1}{n_T}\sum_{k=1}^{K}\sum_{j=1}^{n_k} \mathrm{E}(Y_{kj}) = \frac{1}{n_T}\sum_{k=1}^{K}\sum_{j=1}^{n_k} \mu_k = \frac{1}{n_T}\sum_{k=1}^{K} n_k\mu_k = \bar{\mu} \tag{3.11}$$

$$\mathrm{Var}(\bar{\bar{Y}}) = \frac{\sigma^2}{n_T} \tag{3.12}$$

where σ^2 is the population variance, μ_k is the k'th population mean, and $\bar{\mu}$ is the mean of the K population means. Now, we can substitute equations 3.9 - 3.12 into equation 3.8, which simplifies to:

$$\mathrm{E}(SSTR) = \Big[\sum_{k=1}^{K} n_k\Big(\frac{\sigma^2}{n_k} + \mu_k^2\Big)\Big] - n_T\Big[\frac{\sigma^2}{n_T} + \bar{\mu}^2\Big] \tag{3.13}$$

Simplifying, we get:

$$\begin{aligned}\mathrm{E}(SSTR) &= \Big[\sum_{k=1}^{K}\sigma^2\Big] + \Big[\sum_{k=1}^{K} n_k\mu_k^2\Big] - \sigma^2 - n_T\bar{\mu}^2 \\ &= \sigma^2(K-1) + \Big[\sum_{k=1}^{K} n_k(\mu_k - \bar{\mu})^2\Big]\end{aligned} \tag{3.14}$$

Therefore

$$\mathrm{E}(MSTR) = \mathrm{E}\Big[\frac{SSTR}{K-1}\Big] = \sigma^2 + \frac{\Big[\sum_{k=1}^{K} n_k(\mu_k - \bar{\mu})^2\Big]}{K-1} \tag{3.15}$$

With the balanced data, $n_1 = n_2 = ... = n_K = n_o$, the expectation of the mean square due to treatment is:

$$\mathrm{E}(MSTR) = \mathrm{E}\Big[\frac{SSTR}{K-1}\Big] = \sigma^2 + \frac{n_o \sum_{k=1}^{K}(\mu_k - \bar{\mu})^2}{K-1} \tag{3.16}$$

In order to find the expected value of the within treatment group mean square, we recall two theorems regarding χ^2 disributions.

Theorem 1. If $Y_1, \ldots, Y_n$ are independently and identically distributed normal random variables with mean μ and variance of σ^2, then

$$\frac{(n-1)S^2}{\sigma^2} \tag{3.17}$$

follows a χ^2 distribution with $n-1$ degrees of freedom, where $S^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$ represents the sample variance.

Theorem 2. The sum of independently distributed χ^2 random variables is distributed as χ^2, with the degrees of freedom equal to the sum of the degrees of freedom of the individual variables.

Recall the error sum of squares:

$$SSE = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(Y_{kj} - \bar{Y}_k)^2 = \sum_{k=1}^{K}(n_k - 1)S_k^2 \tag{3.18}$$

where S_k is the sample standard deviation of k'th group. Then

$$\frac{SSE}{\sigma^2} = \frac{\sum_{k=1}^{K}(n_k - 1)S_k^2}{\sigma^2} = \sum_{k=1}^{K}\frac{(n_k - 1)S_k^2}{\sigma^2} \tag{3.19}$$

follows a chi-square distribution with $(n_1 - 1) + (n_2 - 1) + ... + (n_K - 1) = n_T - K$ degrees of freedom. As we know, the expected value of a χ^2 random variable is its degrees of freedom, so

$$\mathrm{E}\Big(\frac{SSE}{\sigma^2}\Big) = n_T - K \tag{3.20}$$

Finally,

$$\begin{aligned}\mathrm{E}(MSE) = \mathrm{E}\Big[\frac{SSE}{n_T - K}\Big] &= \mathrm{E}\Big[\frac{\sigma^2}{n_T - K}\cdot\frac{SSE}{\sigma^2}\Big] \\ &= \frac{\sigma^2}{n_T - K}(n_T - K) \\ &= \sigma^2\end{aligned} \tag{3.21}$$

Combining this with Equation (3.16), in the balanced design case, (MSTR - MSE)/n_o provides an estimate of the variance of the population means $\frac{\sum_{k=1}^{K}(\mu_k - \bar{\mu})^2}{K-1}$.

Considering the correspondence between the fixed and random effects models, this suggests the following as an analogue to the ICC for the fixed effects model.

$$\frac{\frac{\sum_{k=1}^{K}(\hat{\mu}_k-\hat{\bar{\mu}})^2}{K-1}}{\frac{\sum_{k=1}^{K}(\hat{\mu}_k-\hat{\bar{\mu}})^2}{K-1}+\hat{\sigma}^2} \tag{3.22}$$

This is akin to the ICC for the random effects model, but with an empirical variance of the means μ_i replacing $\hat{\sigma}_f^2$. An estimate of heritability with the fixed effects model can be computed by multiplying Equation 3.22 by the appropriate coefficients according to the sibship relationship in the family, for example, multiplication by 2 in the full-sib design.

When the component variances assumed to be all same, the Gaussian mixture model can be viewed as fixed effect model. The distinction is that in the one way fixed effects ANOVA model, family memberships are known, whereas in the Gaussian mixture, family (component) memberships are unknown. This means that parameter estimation will be much more difficult in the mixture model setting, but suggests that the same estimate of heritability might be used.

3.1.2 Expectation-Maximization (EM) Algorithm

Suppose we have n observations $Y_1, ..., Y_n$ from Gaussian mixture model distribution with unknown parameters $\theta = \{\pi_1, ..., \pi_k, \mu_1, ..., \mu_k, \sigma_1^2, ..., \sigma_k^2\}]$. The likelihood function is

$$L(\theta|y_1, .., y_n) \propto \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k N(y_i; \mu_k, \sigma_k^2) \tag{3.23}$$

where, for notational convenience, we use $N(y_i; \mu_k, \sigma_k^2)$ to denote the normal density

with mean μ_k and variance σ_k^2, evaluated at y_i. The log-likelihood function is

$$l(\theta) \propto \sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} \pi_k N(y_i; \mu_k, \sigma_k^2) \Big) \tag{3.24}$$

Setting the derivative with respect to μ_k equal to zero, the associated likelihood equation is

$$\sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} \pi_k N(y_i; \mu_k, \sigma_k^2)} \pi_k N(y_i; \mu_k, \sigma_k^2) \frac{(y_i - \mu_k)}{\sigma_k^2} = 0 \tag{3.25}$$

This can't be solved analytically for μ_k. However, if we knew the latent variables Z_i, then we could simply gather all the observations y_i such that $Z_i = k$, and then solve for μ_k.

Intuitively, the latent variables Z_i should help us find the MLEs. First, we compute the posterior distribution of Z_i given the observations:

$$P(Z_i = k|y_i) = \frac{P(y_i|Z_i = k)P(Z_i = k)}{P(y_i)} = \frac{\pi_k N(y_i; \mu_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k N(y_i; \mu_k, \sigma_k^2)} = \gamma_{Z_i}(k) \tag{3.26}$$

Now we rewrite Equation 3.25, the derivative of the log-likelihood with respect to μ_k, as follows

$$\sum_{i=1}^{n} \gamma_{z_i}(k) \frac{(y_i - \mu_k)}{\sigma_k^2} = 0 \tag{3.27}$$

Even though $\gamma_{Z_i}(k)$ depends on μ_k, we pretend for now that it doesn't. We can solve

for μ_k in this equation to get

$$\widehat{\mu}_k = \frac{\sum_{i=1}^{n} \gamma_{Z_i}(k) y_i}{\sum_{i=1}^{n} \gamma_{Z_i}(k)} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{Z_i}(k) y_i \tag{3.28}$$

where we set $N_k = \sum_{i=1}^{n} \gamma_{Z_i}(k)$. N_k can be considered as the effective number of points assigned to component k. Therefore, $\hat{\mu}_k$ is a weighted average of the data with weights $\gamma_{Z_i}(k)$. We can apply the similar method to find $\widehat{\sigma}_k^2$ and $\widehat{\pi}_k$.

$$\widehat{\sigma}_k^2 = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{Z_i}(k)(y_i - \mu_k)^2 \tag{3.29}$$

$$\widehat{\pi}_k = \frac{N_k}{n} \tag{3.30}$$

Again, these equations are not closed-form expressions since $\gamma_{z_i}(k)$ depends on the unknown parameters. If we knew the parameters, we could compute the posterior probabilities $\gamma_{Z_i}(k)$, and if we knew the posteriors, we could easily compute the parameters. This looks like a vicious circle. The EM algorithm was motivated by this situation, and proceeds as follows:

1. Initialize the π_k, μ_k and σ_k^2 and evaluate the log-likelihood with these parameters.

2. E-step: Evaluate the posterior probabilities $\gamma_{Z_i}(k)$ using the current values of parameters in Equation 3.26.

3. M-step: Estimate new parameters $\widehat{\mu}_k$, $\widehat{\sigma}_k^2$ and $\widehat{\pi}_k$ with the current value of $\gamma_{Z_i}(k)$

using Equation 3.28, 3.29 and 3.30.

4. Evaluate the log-likelihood with the new parameter estimates. If the log-likelihood has changed by less than some small value, stop. Otherwise, go back to step 2.

The EM algorithm is sensitive to the initial values of the parameters, so care must be taken in the first step. However, assuming the initial values are valid, one property of the EM algorithm is that the log-likelihood increases at every step.

The EM algorithm can be generally applied to find maximum likelihood estimates for models with latent variables. Let $\boldsymbol{Y}$ be the entire set of observed variables and $\boldsymbol{Z}$ be the entire set of latent variables. The log-likelihood is therefore

$$l(\boldsymbol{\theta}|\boldsymbol{Y}) \propto \log(P(\boldsymbol{Y}|\boldsymbol{\theta})) = \log\Big(\sum_{Z} P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta})\Big) \tag{3.31}$$

where we have marginalized $\boldsymbol{Z}$ out of the joint distribution. We typically don't know $\boldsymbol{Z}$, but the information we do have about $\boldsymbol{Z}$ is contained in the posterior $P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta})$. Since we don't know the complete data log-likelihood, we consider its expectation under the posterior distribution of the latent variables. This corresponds to the E-step above. In the M-step, we maximize the expectation to find a new estimate for the parameters.

In the E-step, we use the current value of the parameter $\boldsymbol{\theta^0}$ to find the posterior

distribution of the latent variables given by $P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta^0})$. This corresponds to $\gamma_{Z_i}(k)$ in equation 3.26. Then, we use this to find the expectation of the complete data log-likelihood, with respect to this posterior, evaluated at an arbitrary $\boldsymbol{\theta^0}$. This expectation is denoted by $Q(\boldsymbol{\theta}, \boldsymbol{\theta^0})$ and it equals

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta^0}) = \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta^0}}[\log P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta})] = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta^0}) \log P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{3.32}$$

In the M-step, we determine the new parameter $\hat{\boldsymbol{\theta}}$ by maximizing Q

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta^0}) \tag{3.33}$$

3.1.3 EM for Gaussian Mixture Models

The complete data likelihood for Gaussian mixture models takes the form

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{Z}) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{\mathbf{I}(Z_i=k)} N(y_i|\mu_k, \sigma_k^2)^{\mathbf{I}(Z_i=k)} \tag{3.34}$$

with complete data log-likelihood

$$l(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{Z}) \propto \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{I}(Z_i = k)(\log(\pi_k) + \log(N(y_i|\mu_k, \sigma_k^2))) \tag{3.35}$$

where $\mathbf{I}$ is the indicator function, $i, e.$ $\mathbf{I}(Z_i = k) = 1$ if $Z_i = k$, and $\mathbf{I}(Z_i = k) = 0$ otherwise.

Note that for the complete data log-likelihood, the logarithm acts directly on the normal density, leading to a simplified solution for the MLE. However, in practice, we don't observe the latent variables, so we consider the expectation of the complete data log-likelihood with respect to the posterior distribution of the latent variables.

$$\begin{aligned} & \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{y}}[\log(P(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}))] \\ & = \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{y}}\left[\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbf{I}(Z_i = k)(\log(\pi_k) + \log(N(y_i|\mu_k, \sigma_k^2)))\right] \\ & = \sum_{i=1}^{n}\sum_{k=1}^{K}\mathrm{E}_{\boldsymbol{Z}|\boldsymbol{y}}\Big[\mathbf{I}(Z_i = k)\Big](\log(\pi_k) + \log(N(y_i|\mu_k, \sigma_k^2))) \end{aligned} \tag{3.36}$$

Since $\mathrm{E}_{\boldsymbol{Z}|\boldsymbol{y}}[\mathbf{I}(Z_i = k)] = P(Z_i = k|\boldsymbol{y})$, we see that this is simply $\gamma_{Z_i}(k)$. Hence, we have

$$\mathrm{E}_{\boldsymbol{Z}|\boldsymbol{y}}[l(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{Z})] = \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{z_i}(k)(\log(\pi_k) + \log(N(y_i|\mu_k, \sigma_k^2))) \tag{3.37}$$

EM proceeds as follows: first choose initial values for $\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}$ and use these in the E-step to evaluate the $\gamma_{Z_i}(k)$. Then, with $\gamma_{Z_i}(k)$ fixed, maximize the expected complete log-likelihood above with respect to μ_k, σ_k^2 and π_k.

3.1.4 Estimation Results with Phenotypic Observations Only

We have investigated the Gaussian mixture model as the worst possible case scenario for estimating heritability, when none of the family memberships are known. In addition, the absence of knowledge of the number of families makes the estimation considerably more difficult (Chen and Khalili (2008); Kasahara and Shimotsu (2015)). In the simpler situation where the number of families is assumed be known, the EM algorithm estimates parameters by maximizing the mixture likelihood and has the benefit that it provides the estimated probabilities of family membership. This is an indication that the phenotypic observations carry some information about pedigree structure, and it motivates us to include them in the model to reconstruct the pedigrees.

In contrast to estimating a heritability like object directly from fitting a Gaussian mixture model, the pedigree reconstruction problem can be considered as the regular clustering analysis problem.

For example, individuals can be grouped into different families by using K-means clusting algorithm, and conditional on the grouping provided by K-means, an estimate of heritability can be obtained by fitting a linear mixed model. This two-step method was applied with our hybrid data (phenotypic observations at the first time point only). First, we select the optimal number of families according to the elbow method, which is one of the most popular methods to determine the optimal K value.

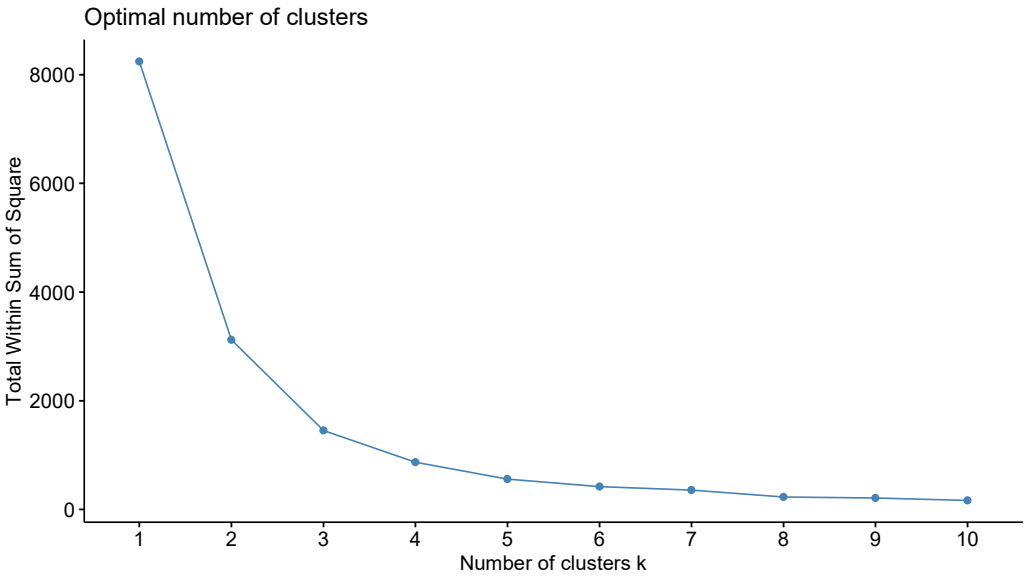

Figure 3.1: The Elbow method to determine the optimal number of families

The eblow method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. From the Figure 3.1, we think 4 or 5 would be most reasonable values for the number of families.

Then, we estimate the pedigree structures with 4 full-sib families and 5 full-sib families using K-means algorithm. For comparison purpose, we estimate the family membership with 12 (true number of families) full-sib families as well. At the end, we use *lmer* function in R to fit the linear mixed model and the estimates of heritabilities are 1.908158, 1.939498 and 1.9800465, respectively.

We also estimate of heritablity like object directly from fitting a Guassian mixture model with 4, 5 and 12 families by using $normalmixEM$ function in "mixtools" package (Benaglia et al. (2009)), which is an R package for analyzing finite mixture models. The estimation results are 1.572833, 1.492110 and 1.815661. It is not surprising that the estimations are disappointing no matter which procedure is used, as there is a lot of observation overlap among the families in our data. Estimating the pedigree structure or heritability by phenotypic observations only will be extremely challenging in such situations.

It is not surprising that the estimations are disappointing no matter which procedure is used, as there is a lot of observation overlap among the families in our data.

Estimating the pedigree structure or heritability by phenotypic observations only will be extremely challenging in such situations.

3.2 Estimation Procedure with Marker Data

In recent years, the extensive development and application of highly polymorphic molecular markers (especially microsatellites) has proven highly useful, particularly in the fields of population and conservation genetics (Frankham et al. (2002)). Different types of markers, particularly codominant microsatellites, have been developed to estimate pairwise coancestries. The estimated coancestry from molecular information then can be used to obtain estimates of heritability (Ritland (1996); Wang (2004)). Another approach involves an explicit reconstruction of groups of a certain coancestry, which can be used as pedigree information in a standard quantitative genetic analysis. This method is performed using Markov chain Monte Carlo (MCMC) procedures to reconstruct sibships within a single generation (Thomas and Hill (2000); Smith et al. (2001)). The reconstructed sibships are then used to estimate variance components and heritability by using the ANOVA method or fitting linear mixed model.

3.2.1 Pairwise techniques with Marker Information

Several methods have been suggested for the estimation of pairwise values of the coefficients of coancestry (Θ_{ij}) and fraternity (Δ_{ij}) from information on shared alleles at codominant marker loci. These methods may be grouped into two categories: method

of moments estimators, which are used to estimate relatedness as a continuous measure, on the basis of shared alleles at marker loci (Lynch and Ritland (1999)) and likelihood techniques, used to determine the likelihood of a pair falling into particular relationship classes given the observed marker information (Mousseau et al. (1998); Thomas and Hill (2000)).

These estimators are not necessarily very efficient unless large numbers of polymorphic loci are assayed, but most of them do provide unbiased estimates. Ritland (1996) made the clever leap of showing how estimates of pairwise relatedness can be combined with estimates of pairwise phenotypic similarity to generate estimates of variance components in natural populations.

Recall that the basic premise underlying all conventional methods for estimating the additive genetic variance of a trait is the fact that, for a character with a purely additive genetic effect, the phenotypic covariance between relatives i and j has expected value $2\Theta_{ij}\sigma_A^2$ (Equation 1.45). The phenotypic similarity of two individuals with phenotypes Y_i and Y_j can be defined as

$$s_{ij} = (Y_i - \bar{Y})(Y_j - \bar{Y}) \tag{3.38}$$

where $\bar{Y}$ is the mean phenotype in the population. Since this expression is in the form of a phenotypic covariance, under the purely additive model, the expected value of s_{ij} is simply $2\Theta_{ij}\sigma_A^2$. Therefore, with a collection of individuals, the observed phenotypic

similarity can be written in the form of a linear model

$$s_{ij} = 2\widehat{\Theta}_{ij}\sigma_A^2 + e_{ij} \tag{3.39}$$

where $\widehat{\Theta}_{ij}$ is the estimated value of Θ_{ij} for the two individuals, and e_{ij} is the residual deviation of the observed similarity from its expectation.

Equation 3.38 suggests that an estimate of the narrow-sense heritability, σ_A^2/σ_Y^2, can be obtained by regressing pairwise measures of phenotypic similarity on estimates of the coefficient of coancestry (with half the slope providing the estimate of σ_A^2, and with the observed phenotypic variance in the population, $\mathrm{Var}(Y)$, providing the estimate of σ_Y^2).

Since the $\widehat{\Theta}_{ij}$ are only estimates, a conventional least-squares analysis would lead to downwardly biased estimates of σ_A^2 as a consequence of the inflated estimate of the variance of relatedness. Ritland (1996) outlines a method that provides an estimate of σ_Θ^2, the actual variance of relatedness, which excludes the sampling variance resulting from the use of a finite number of marker loci. Letting $\mathrm{Var}(\Theta)$ be the estimated actual variance of relatedness and $\mathrm{Cov}(s, \widehat{\Theta})$ be the covariance of phenotypic similarity and estimated relatedness, the heritability can be estimated by

$$\widehat{h}^2 = \frac{\mathrm{Cov}(s, \widehat{\Theta})}{2\mathrm{Var}(\Theta)\mathrm{Var}(Y)} \tag{3.40}$$

under the assumptions of the ideal additive model (assuming random mating and no shared environmental effects).

Pairwise methods have some limitations. First, pairwise techniques may lose valuable information in the form of higher-order relationships. Additionally, the weight placed on information from a single family depends on the number of pairs of individuals that can be chosen from that family. It is dependent only upon family size and not information content. Therefore, pairwise methods do not provide the most efficient estimates for parameters and are prone to larger standard errors than restricted maximum likelihood methods (Thomas and Hill (2000)). Second, obtaining estimates of the allele frequencies at the marker loci is also a problem. Allele frequencies have traditionally been assumed known or have been estimated from the sample. They are subject to further random error, since there are relatives within the sample, which might bias subsequent estimates of pairwise relationships. Finally, there is a question as to how to include other factors such as sex or age in the model. Since pairwise methods operate on a pairwise level, other factors must also be investigated on a pairwise level and as a result, an optimal estimate may not be achieved.

3.2.2 A two-Step estimation procedure with marker information

A two-step procedure is a very popular method to estimate the variance components and heritability. In the first step, family memberships are reconstructed using marker

information. In the second step, the reconstructed pedigree structure is used to form a relationship matrix suitable for use in an animal model, with a linear mixed model used to estimate variance components, as in section 2.3. This approach allows traditional and efficient methods for parameter estimation to be used and hence simplifies the inclusion of additional factors or the use of multivariate analysis if data have been collected from several traits. Of course, parameter estimates are conditional on the pedigree estimated in the first step.

Smith et al. (2001) introduced two Markov chain Monte Carlo algorithms that allow the partitioning of individuals into full-sib groups using single-locus genetic marker data when parental information is not available. They developed a moving algorithm that can search through the sibship configuration space with the aim of locating the configuration that maximizes a criterion - either the full joint likelihood of the proposed family structure, or a score based on the pairwise likelihood ratios of being full-sib or unrelated.

3.2.2.1 The Total Number of Full sibship Configurations

A full sibship configuration of N individuals is a partition of the individuals into different full-sib families. If a particular configuration has K full-sib families, the partition is denoted as $c_1, ..., c_K$, where c_j is the collection of individuals in the j'th family. The space of all possible data configurations consisting of only full-sibs or

unrelated individuals is denoted by $\mathscr{C}$. In combinatorial mathematics, the size of $\mathscr{C}$ is given by the so-called Bell number, which counts the number of possible partitions of a set. These numbers have been studied by mathematicians since the 19th century and are named after Eric Temple Bell, who wrote about them in the 1930s. Starting with $B_0 = B_1 = 1$, the first few Bell numbers are:

1, 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115875, 678570, 4213597, ...

The nth of these numbers, B_n, describes the number of different ways to partition a set that has exactly n elements, or equivalently, the number of equivalence relations on it. For example, $B_3 = 5$ because the 3-element set $\{a, b, c\}$can be partitioned in 5 distinct ways:

$$\Big[a, b, c\Big];\ \Big[a, bc\Big];\ \Big[b, ac\Big];\ \Big[c, ab\Big]\ ;\ \Big[abc\Big]$$

The configuration (the full sib structure) is the parameter to be estimated. The parameter space is non-Euclidean, so usual methods of optimization, such as gradient based methods, cannot be used. The enormous size of the configuration space $\mathscr{C}$ for even a moderate value of N precludes the method of direct enumeration to maximize the probability of the configuration (the family structure) given observed alleles, conditional on the population allelic frequencies. In such cases, methods such as simulated annealing Kirkpatrick et al. (1983) can be used to explore the space of configurations.

3.2.2.2 The Markov Chain Monte Carlo Method and the Metropolis-Hasting Algorithm

Monte Carlo methods are a subset of computational algorithms that use the process of repeated random sampling to make numerical estimations of unknown parameters. The underlying concept is to use randomness to solve the deterministic problems. Monte Carlo methods are commonly used in three problem classes: optimization, numerical integration, and generating samples from probability distributions. In principle, any problem with a probabilistic interpretation can be solved by Monte Carlo methods. For example, the law of large numbers states that the expected value of some random variable can be approximated by taking the empirical mean of independent samples of the variable. Markov chain Monte Carlo (MCMC) samplers are often used when the probability distribution of the variable is parameterised. The main idea is to design a judicious Markov chain with stationary distribution being the distribution of interest. If this can be done then in the limit, the samples generated from the Markov chain will be samples from the desired distribution. The ergodic theorem states that the empirical measure of the random states of the MCMC sampler is an approximation to the stationary distribution, and one consequence of ergodicity is that ensemble averages equal time averages.

Metropolis et al. (1953)'s paper on the statistical mechanics of particles introduced the method of Markov chain Monte Carlo (MCMC) to the world of physics. Hammersley et al. (1965) described the method in a more rigorous statistical framework in term of Markov chains. In 1970, Hastings (1970) provided a generalization of

the original Metropolis algorithm to allow for non-symmetric proposal distributions. Geman and Geman (1984) used the Gibbs sampler on the Bayesian image restoration problem, and Gelfand and Smith (1990) showed how the Gibbs sampler could be applied to help Bayesians solve a much wider class of problems. Robert and Casella (2005) gives a general definition of Markov chain Monte Carlo algorithms: A Markov chain Monte Carlo method for the simulation of a distribution f is any method producing an ergodic Markov chain (X_t) whose stationary distribution is f.

The Metropolis-Hastings algorithm is one of the most important MCMC methods for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult. This sequence of random samples can be used to approximate the distribution of interest. In order to generate a collection of states according to a desired distribution $P(x)$, the Metropolis-Hastings algorithm uses a Markov process which asymptotically reaches a unique stationary distribution $\pi(x)$ such that $\pi(x) = P(x)$.

A Markov process is uniquely defined by its transition probabilities $P(x'|x)$, the probability of moving from any given state x to any other given state x'. When the following two conditions are satisfied, it has a unique stationary distribution $\pi(x)$:

- There must exist a stationary distribution $\pi(x)$. A sufficient but not necessary condition for this is detailed balance, which requires that each transition $x \to x'$ is reversible for each pair of states (x, x'). That is, the probability of being in

state x and transitioning to state x' must be equal to the probability of being state x' and transitioning to state x, $\pi(x)P(x'|x) = \pi(x')P(x|x')$.

- The stationary distribution $\pi(x)$ must be unique. This condition is guaranteed by ergodicity of the Markov process, which requires that every state must be aperiodic and positive recurrent.

The Metropolis-Hasting algorithm involves designing a Markov process that satisfies the two above conditions by constructing the appropriate transition probabilities, such that the process has stationary distribution $\pi(x)$ equal to $P(x)$. The detailed balance condition can be re-written as

$$\frac{P(x'|x)}{P(x|x')} = \frac{P(x')}{P(x)} \tag{3.41}$$

The transition probabilities can be divided into two sub-steps: the proposal step and the acceptance-rejection step. The proposal distribution $q(x'|x)$ is the conditional probability of proposing a state x' given x, and the acceptance ratio $r(x', x)$ is the probability to accept the proposed state x'. The transition probability is the product of the proposal and acceptance probabilities, $P(x'|x) = q(x'|x)r(x', x)$. Now, equation (3.41) can be written as

$$\frac{r(x', x)}{r(x, x')} = \frac{P(x')}{P(x)} \frac{q(x|x')}{q(x'|x)} \tag{3.42}$$

One common choice for an acceptance ratio that fulfills the condition above is the Metropolis choice:

$$r(x', x) = \min\Big(1, \frac{P(x')}{P(x)} \frac{q(x|x')}{q(x'|x)}\Big) \tag{3.43}$$

The Metropolis-Hastings algorithm proceeds in the following manner:

- 1. Initialise
  - 1. Start with an initial state x_0
  - 2. Set $t = 0$
- 2. Iterate
  - 1. Generate a random candidate state x' according to proposal distribution $q(x'|x_t)$.
  - 2. Compute the acceptance probability $r(x', x_t) = \min\left(1, \frac{P(x')}{P(x_t)} \frac{q(x_t|x')}{q(x'|x_t)}\right)$
  - 3. Accept or reject
    - 1. Generate a uniform random number $u \in [0, 1]$.
    - 2. If $u \leq r(x', x_t)$, then accept the candidate state and set $x_{t+1} = x'$.
    - 3. If $u > r(x', x_t)$, then reject the candidate state, and keep the old state forward $x_{t+1} = x_t$
    - 4. Increment, set $t = t + 1$

As shown by Hastings (1970), given that the specified conditions are satisfied, the empirical distribution of the states $x_0, x_1, ..., x_T$ will approach $P(x)$. The number of iterations T required to effectively estimate $P(x)$ depends on number of factors, including the relationship between $P(x)$ and the proposal distribution and the desired accuracy of estimation (Raftery and Lewis (1991)).

3.2.2.3 Pedigree Reconstruction Procedure

The Metropolis-Hastings algorithm is a general tool to sample from a state space, in our case $\mathscr{C}$. Let us define a Markov chain having a stationary distribution on $\mathscr{C}$, denoted by $P(C)$, where $C \in \mathscr{C}$. C_t denotes the tth configuration generated, and the algorithm proceeds by simulating a proposal C' from a proposal probability $q(C'|C_t)$. At the next step, C_{t+1} is randomly assigned to be either the proposal C' with acceptance probability $r(C', C_t)$, or C_t with probability $1 - r(C', C_t)$, where

$$r(C', C_t) = \min\Big(1, \frac{P(C')q(C_t|C')}{P(C_t)q(C'|C_t)}\Big) \tag{3.44}$$

For appropriate choices of q, as described in Hastings (1970), this algorithm is guaranteed to generate samples from the distribution $P(C)$ in the limit as t increases.

The implementation starts by setting the initial configuration C_0 to "all unrelated" in which there are N families each containing one individual. For the distribution $q(C'|C_t)$, we select two individuals I and J independently according to a uniform distribution on $\{1, ..., N\}$. Let c_I and c_J represent the full-sib families to which individuals I and J belong. If $c_I \neq c_J$, then the proposed configuration C' is obtained by moving individual I from group c_I to c_J. If $c_I = c_J$, then individual I is removed from c_I to create a new full-sib family of size one. If c_I has only one member, then the number of families is reduced by 1. In this way, the proposal algorithm can sample configurations with from any number of families between 1 and N. This choice of q satisfies the necessary conditions under which the algorithm will generate samples

from the desired distribution $P(C)$. In Smith et al. (2001), the authors stated that this algorithm also ensures that $q(C'|C_t) = q(C_t|C')$, in which case the Metropolis-Hastings algorithm is the original Metropolis algorithm (Metropolis et al. (1953)). However, after careful checking, I have found that when the cardinalities $\mathbf{card}(c_I)$ and $\mathbf{card}(c_J)$ are not equal, then $q(C'|C_t) \neq q(C_t|C')$, and the acceptance probability $r(C', C_t)$ has been appropriately adjusted.

The distribution used on the space of alleles is the full joint distribution of the observed alleles given the configuration C, conditional on the allele frequencies. This approach was extensively investigated by Painter (1997). The single-locus likelihood for a configuration consisting of K full-sib groups is proportional to the joint probability of the observed alleles, which can be written as

$$\prod_{j=1}^{K} \sum_{g_m(j)} \sum_{g_p(j)} \left[\prod_{i \in c(j)} P(O_i(j))|g_m(j), g_p(j)) \right] P(g_m(j)|p) P(g_p(j)|p) \tag{3.45}$$

where $g_m(j)$ and $g_p(j)$ are the unobserved maternal and paternal genotypes for the j'th full-sib group, $c(j)$ is the group of offspring in the j'th full-sib group, $O_i(j)$ is the observed genotype of the i'th individual in the j'th full-sib group and p denotes the unknown population allele frequencies. In practice the observed allele frequencies are used in place of the population frequencies.

For offspring i in full-sib group j, the probability of the observed genotype $O_i(j)$ at a single locus , conditional on the maternal $g_m(j)$ and paternal $g_p(j)$ genotypes, is

Table 3.1: Probabilities of the segregation events at one locus with two alleles

	Parent's genotypes					
Offspring genotype	xx, xx	xx, xy	xx, yy	xy, xy	xy, yy	yy, yy
xx	1	1/2	0	1/4	0	0
xy	0	1/2	1	1/2	1/2	0
yy	0	0	0	1/4	1/2	1

Table 3.2: Population Genotype Probabilities for Parents, at one locus with two alleles as a function of p_x and p_y, the population allele frequency for allele x and y

Parent genotype	xx	xy	yy
xx	p_x^4	$2p_x^3p_y$	$p_x^2p_y^2$
xy	$2p_x^3p_y$	$4p_x^2p_y^2$	$2p_xp_y^3$
yy	$p_x^2p_y^2$	$2p_xp_y^3$	p_y^4

denoted as $P(O_i(j))|g_m(j), g_p(j))$. These probabilities are determined by the segregation probabilities during the formation of gametes, and are given in Table 3.1. The joint probability of the genotypes at several unlinked loci is just the product, over loci, of the single locus probabilities.

The joint probability of the maternal and paternal genotypes given the population allele frequencies is $P(g_m(j)|p)P(g_p(j)|p)$. Assuming Hardy Weinberg equilibrium and random mating, these probabilities are given in Table 3.2. For example, under the stated assumptions, the probability of drawing a male with genotype xx and a female with genotype xy from the population is $p_x^2 \times p_xp_y = p_x^3p_y$, with the same probability of drawing a female with genotype xx and a male with genotype xy, so the probability of one parent being xx and the the being xy is $2p_x^3p_y$.

The single-locus single full-sib family genotype probabilities can be written as explicit polynomial functions of the allele frequencies under specific genotype configurations, which results in substantial computational saving. The polynomial expressions were derived Painter (1997), and are given in Table 3.3 for each of the 14 possible single-locus genotype configurations of a full-sib family. In the most straightforward case, if a family consists of n individuals of genotype xx and m individuals of genotype yy, then the parental genotypes must have both been xy, and each of the sums in Equation 3.45 contains a single term. Combining the results from Tables 3.2 and 3.1, the probability of the genotypic data (n individuals of genotype xx and m individuals of genotype yy) is

$$(\frac{1}{4})^n \cdot (\frac{1}{4})^m \cdot 4p_x^2 p_y^2 = \frac{4}{4^{n+m}} p_x^2 p_y^2 \tag{3.46}$$

For most of the full-sibship genotypes, the sums in Equation 3.45 will contain multiple terms, and the derivation of the sibship likelihood, while straightforward, can be very tedious.

Using the algorithm described above to propose a new configuration C given the current configuration C_t, it is very common to propose moves for which the new configuration C is infeasible in that the alleles of the individuals in C cannot have arisen by segregation from a pair of parents. For example, suppose a family in the

Table 3.3: Polynomial Equation for the Likelihood of a Single-locus Full-sibship

Full-sibship genotype	Sibship likelihood
xx^n	$\frac{4}{4^n}p_x^2 + (\frac{4}{2^n} - \frac{8}{4^n})p_x^3 + (1 - \frac{4}{2^n} + \frac{4}{4^n})p_x^4$
xy^n	$(2 - \frac{4}{2^n})p_x^2p_y^2 + (\frac{4}{2^n} - \frac{8}{4^n})p_x^2p_y + (\frac{4}{2^n} - \frac{8}{4^n})p_xp_y^2 + \frac{8}{4^n}p_xp_y$
xx^nxy^m	$\frac{8}{4^{n+m}}p_x^2p_y + (\frac{4}{2^{n+m}} - \frac{8}{4^{n+m}})p_x^3p_y + (\frac{4}{4^n2^m} - \frac{8}{4^{n+m}})p_x^2p_y^2$
xx^nyy^m	$\frac{4}{4^{n+m}}p_x^2p_y^2$
xx^nyz^m	$\frac{4}{4^{n+m}}p_x^2p_yp_z$
$xx^nxy^myy^l$	$\frac{4}{2^l4^{n+m}}p_x^2p_y^2$
$xx^nxy^mxz^l$	$\frac{4}{4^{n+m+l}}p_x^2p_yp_z$
$xx^nxy^myz^l$	$\frac{4}{4^{n+m+l}}p_x^2p_yp_z$
$xx^nxy^mxz^lyz^k$	$\frac{4}{4^{n+m+l+k}}p_x^2p_yp_z$
xy^nxz^m	$\frac{4}{2^{n+m}}p_x^2p_yp_z + \frac{8}{4^{n+m}}p_xp_yp_z$
$xy^nxz^myz^l$	$\frac{8}{4^{n+m+l}}(p_x^2p_yp_z + p_xp_y^2p_z + p_xp_yp_z^2)$
$xy^nxz^myw^l$	$\frac{8}{4^{n+m+l}}p_xp_yp_zp_w$
xy^nzw^m	$\frac{16}{4^{n+m}}p_xp_yp_zp_w$
$xy^nxz^myw^lzw^k$	$\frac{8}{4^{n+m+l+w}}p_xp_yp_zp_w$

current configuration C_t has 4 genotypes represented and the proposal is to move an individual with a 5'th genotype into that family. As only 4 genotypes can arise from segregation of alleles from a pair of parents, this proposal is infeasible. There are other cases where a proposed family has less than 4 genotypes, but is infeasible. For example, a proposed family with the three genotypes AB, CD, AE at a single locus is infeasible as the parents have at most 4 distinct alleles between them.

In simulations which will be described later, it was commonly found that 70-80% of proposed configurations are infeasible, and this happens particularly often when a current configuration has several large families. When a proposed configuration is infeasible, no likelihood calculation is carried out, and the proposal is not counted. That is, we remain at (t, C_t), and generate another proposal.

A considerable proportion of the computational time in sampling the genotypic likelihood using the MH algorithms is involved in checking for the feasibility of proposed configurations. We had originally thought to expand the proposal algorithm to include block moves, whereby whole families would be merged, but because the feasibility assessment is done after adding a single individual to an existing feasible group, it was decided that there would be no computational gain to moving several individuals at a time, and substantial program revision required to implement this, so block moves have not been considered.

Wang (2004) introduced another likelihood method for pedigree reconstruction with simple and robust models of typing error incorporated into it. This new method makes improvement by using more efficient MCMC algorithms for calculating the full likelihood function and searching for the maximum likelihood configration with block moves. It can deal with very complex pedigree structure and provides an very accurate result. Jones and Wang (2010) developed colony, a computer program implementing Wang (2004)'s method to simultaneously infer sibship and parentage among individuals using multiloucs genotype data. Colony can be used for both diploid and haplodiploid species; it can use dominant and codominant markers, and can accommodate, and estimate, genotyping error at each locus.

3.2.2.4 Pedigree Reconstruction Error and Heritability Estimation

A statistic that enables measurement of the accuracy of each reconstructed family is useful for the purposes of comparison. We used three different measures to describe the fit between true and predicted configurations. The number of moves represents the minimum number of individuals that need to be relocated from their predicted full-sib groups to their real full-sib groups to get a perfectly matched configuration. It is also equal to the number of individuals that must be removed from the true and predicted configurations to make them identical.

Other statistics which can be used to measure the accuracy are based on the number of correctly and incorrectly classified pairs. For example, the number or proportion of full-sib pairs incorrectly classified as unrelated pairs (E_1), and the proportion of unrelated pairs incorrectly classified as full-sib pairs (E_2). Let us consider an example of two full-sib groups with 10 individuals each:

True configuration: (A A A A A A A A A A), (B B B B B B B B B B)

Prediction 1: (A A A A A A A A A B), (B B B B B B B A), (A), (B)

Prediction 2: (A A A A A A A A A A), (B B B B B B), (B B B B)

Four moves are needed to get to the correct configuration for both predictions. There are a total of 190 pairs of individuals for two families with 10 individuals in each, 90 are full-sib pairs and 100 are unrelated pairs. In prediction 1, 34 full-sib pairs are incorrectly classified as unrelated and 16 unrelated pairs are incorrectly classified

as full-sibs. Therefore, the proportion of $E_1 = 34/90 = 0.378$ and $E_2 = 16/100 = 0.16$ for prediction 1. In prediction 2, the proportion of $E_1 = 24/90 = 0.276$ and $E_2 = 0$ since none of the unrelated pairs are incorrectly classified. Even though the number of moves criterion is the same for the two predicted configurations, prediction 2 would seem to provide the more accurate result, assuming some weight is given to unrelated incorrectly classified as full-sibs. The simulations below will show that misclassifying unrelated individuals as full-sibs is the more serious error in terms of biasing the estimate of heritability.

In the two-step estimation procedure, the reconstructed pedigree is used to estimate variance components for a quantitative trait. Errors of genuinely unrelated individuals who are classified as related lead to a large bias in estimates, particularly within family variance and heritability, as compared to the errors of genuinely related individuals who are classed as unrelated. In the previous example, prediction 2 should give more accurate estimation of variance components. The results from a small simulation study confirm this conclusion.

We generate phenotypic observations with different heritability values (0.2, 0.4, 0.6) for 200 individuals belonging to 10 full-sib families. The heritabilities were estimated by fitting a linear mixed model with the REML method. First of all, we fit the model with the true family structure. Then, we incrementally increased the error rates of full-sib pairs incorrectly classified as unrelated (E1), and unrelated pairs incorrectly

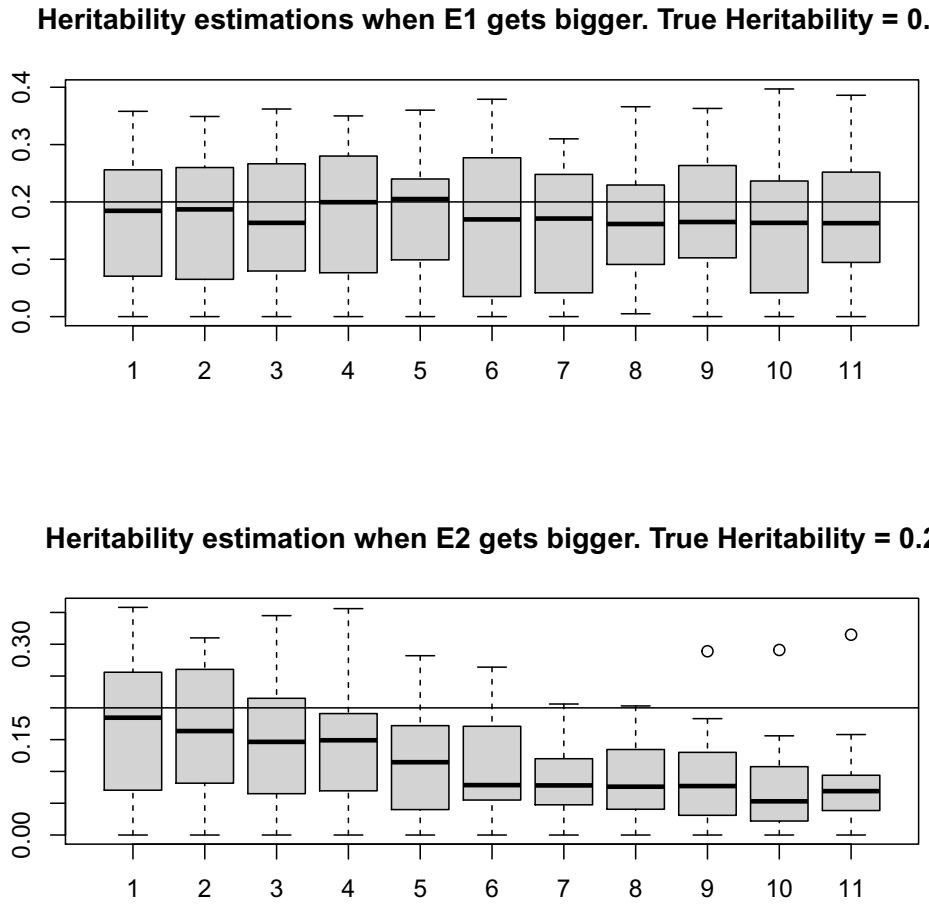

Figure 3.2: Estimated heritability vs misclassification error rates. True heritability = 0.2

classified as full-sib pairs (E2). The boxplots of estimated heritability (Figures 3.2 - 3.4) clearly indicate that when E2 increased, the estimate of heritability becomes poorer. The variable on the horizontal axis represents the average number of individual misclassified in each family plus 1, for example, if 2 individuals are misclassified in each family on average, the value on X axis would be 3.

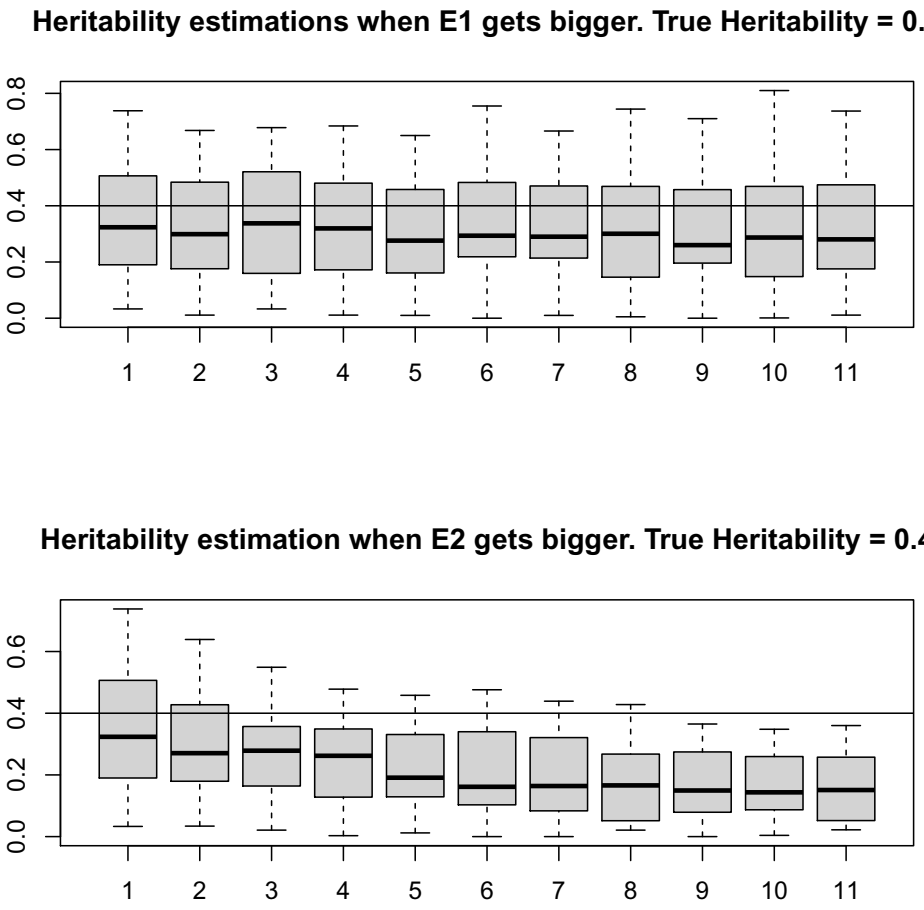

Figure 3.3: Estimated heritability vs misclassification error rates. True heritability = 0.4

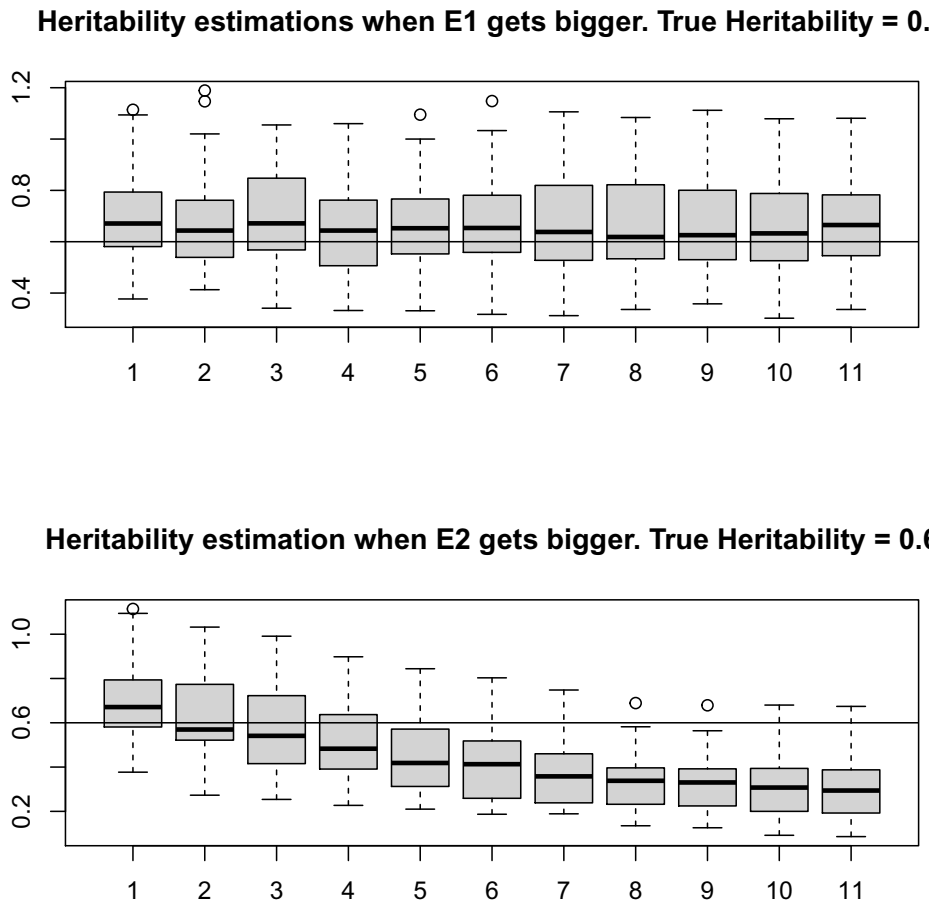

Figure 3.4: Estimated heritability vs misclassification error rates. True heritability = 0.6

Table 3.4: Two-step estimation results for hybrid data set (after log-transformation)

	Accuracy of Pedigree Reconstruction			Heritability Estimation			
Number of Loci	Number of Moves	E1	E2	$\widehat{h_1^2}$	$\widehat{h_2^2}$	$\widehat{h_3^2}$	$\widehat{h_4^2}$
4	7	203	5	0.3167	0.3249	0.5990	0.5583
2	203	3213	1997	0.1705	0.1891	0.2941	0.3326

3.2.2.5 Two-step Estimation Result of Hybrid Data

We used the two-step estimation method for our hybrid data (after log-transformation). Using MCMC method that was introduced by Smith et al. (2001), the pedigree was reconstructed with different levels of genetic marker information (2 loci, and 4 loci). The sample size was 318, and the number of alleles per locus were 11, 14, 10 and 8. The loci with 11 and 14 alleles were used for reconstructions with 2 loci.

The accuracy of the pedigree reconstructions using the marker data, and the associated two step estimates of heritability at each time point are included in Table 3.4. It is clear that with the more accurately estimated pedigree structure, the estimated heritabilities are closer to the results in Table 2.2, where the true pedigree is used. With an inaccurately estimated pedigree (when 2 loci were used for reconstruction), the estimated heritabilities are about half of the estimates with true pedigree. As we discussed in section 3.2.2.4, E2 (unrelated pairs incorrectly classified as full-sib pairs) gives more damage on estimation results.

Chapter 4

Heritability Estimation Using MCMC Procedure with Marker Data and Phenotypic Observation

As we discussed in chapter 3, fitting a Gaussian mixture model or carrying out another clustering analysis indicates that there is information on the family membership contained in the phenotypic observations, while the two-step method fitting a linear mixed model conditional on an estimated pedigree shows some promises if marker data is sufficiently informative to provide a reasonably precise pedigree reconstruction. In this chapter, we will introduce a hybrid Markov chain Monte Carlo approach, where marker information (X) and phenotypic observation (Y) can be used jointly and simultaneously to estimate both pedigree structure and heritability of the quantitative trait. In the case with insufficient marker information, our proposed method is able to provides more accurate results compare with two-step method, regardless the estimation of pedigree or parameter of quantitative trait.

4.1 The General Procedure

The full joint distribution of the observed phenotypic values and the marker alleles given the quantitative trait parameters $\boldsymbol{\Theta}$ and pedigree configuration Π is denoted by $P(\mathbf{Y}, \mathbf{X}|\boldsymbol{\Theta}, \Pi)$. In general the pedigree of a group of individuals of unknown relatedness will include different relationships, such as parent-offspring, full-sibs, half-sibs, cousins, etc. In this thesis, Π is assumed to belong to the space of all possible data configurations consisting only of full-sibs or unrelated individuals.

The loci of microsatellite marker data used are assumed to be from non-coding regions of the genome, or to be otherwise unlinked with the genes associated with the phenotypic trait. In this case $P(\mathbf{Y}, \mathbf{X}|\boldsymbol{\Theta}, \Pi) = P(\mathbf{Y}|\boldsymbol{\Theta}, \Pi)P(\mathbf{X}|\boldsymbol{\Theta}, \Pi)$. Furthermore, the distribution of the marker data is assumed to not depend on the parameters of the phenotypic trait distribution, so $P(\mathbf{X}|\boldsymbol{\Theta}, \Pi) = P(\mathbf{X}|\Pi)$. The distribution of alleles $P(\mathbf{X}|\Pi)$ will depend on the population allele frequencies $\boldsymbol{\Psi}$. This notation has been suppressed throughout for convenience, and in addition, we have followed the usual practice of replacing $\boldsymbol{\Psi}$ by an empirical estimate.

In the Bayesian context, when a joint prior distribution $P(\boldsymbol{\Theta}, \Pi)$ is specified, the full joint posterior is

$$P(\boldsymbol{\Theta}, \Pi|\mathbf{X}, \mathbf{Y}) \propto P(\mathbf{Y}, \mathbf{X}|\boldsymbol{\Theta}, \Pi)P(\boldsymbol{\Theta}, \Pi)$$

Under the stated assumptions, and additionally assuming independent prior distributions for $\boldsymbol{\Theta}$ and Π, the joint posterior is

$$P(\boldsymbol{\Theta}, \Pi|\mathbf{X}, \mathbf{Y}) \propto P(\mathbf{Y}|\boldsymbol{\Theta}, \Pi)P(\mathbf{X}|\Pi)P(\boldsymbol{\Theta})P(\Pi) \tag{4.1}$$

Restricting to full-sib or unrelated relationships, the conditional distribution of the phenotypic data $\mathbf{Y}$ given the pedigree, $P(\mathbf{Y}|\boldsymbol{\Theta}, \Pi)$ follows by specification of a specific model, such as the random effects model or general mixed model considered in chapter 2. It will be discussed in detail in following sections. The polynomial equations developed by Painter (1997), and set down in Table 3.3, provide the details for the calculation of $P(\mathbf{X}|\Pi)$. Prior distributions for $\boldsymbol{\Theta}$ and Π are discussed in the section 4.1.2.

As a side note, the general form of the posterior distribution (Equation 4.1) includes various models considered in previous chapters, depending on the forms of information available. For example, with known pedigree, the terms $P(\mathbf{X}|\boldsymbol{\Pi})$ and $P(\boldsymbol{\Pi})$ drop out and we estimate the parameter of quantitative trait using phenotypic observations, possibly also including prior distribution, so that $P(\boldsymbol{\Theta}|\mathbf{Y}) \propto P(\mathbf{Y}|\boldsymbol{\Theta})P(\boldsymbol{\Theta})$. This is the case that was covered in Chapter 2. With unknown pedigree and without the marker data, $P(\Theta, \boldsymbol{\Pi}|Y) \propto P(Y|\Theta, \boldsymbol{\Pi})P(\boldsymbol{\Pi})$. Summing over Π this gives a finite mixture model similar to that considered in Equation 3.1. When the estimated pedigree structure can be obtained from marker data, we replace the unknown $\boldsymbol{\Pi}$ by $\widehat{\boldsymbol{\Pi}}$ which maximizes $P(\mathbf{X}|\boldsymbol{\Pi})P(\boldsymbol{\Pi})$ (step 1) and then estimate quantitative genetic

parameters by maximizing $P(\mathbf{Y}|\mathbf{\Theta}, \widehat{\mathbf{\Pi}})P(\mathbf{\Theta})$ (step 2). This is the basis of logical structure for two-step estimation method introduced in section 3.2.2.

The original objective of the thesis was to estimate heritability. In the more general context discussed here, the goal of inference is to evaluate the full joint posterior distribution $P(\mathbf{\Theta}, \Pi|\mathbf{X}, \mathbf{Y})$ or the joint likelihood $P(\mathbf{Y}, \mathbf{X}|\mathbf{\Theta}, \Pi)$, depending on the philosophical viewpoint taken. These will be evaluated by sampling, using the Metropolis-Hastings algorithm.

Samples from the marginal distribution $P(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y})$ will then be obtained by marginalizing, and the associated sampled values of $\mathbf{\Theta}$ will be used to generate indirect sample of heritability. For example, if $\mathbf{\Theta} = (\mu, \sigma_f^2, \sigma_e^2)$, and the t'th sampled value is $\mathbf{\Theta}^{(t)} = (\mu_t, \sigma_{f,t}^2, \sigma_{e,t}^2)$. Then for the oneway random effects model, the t'th indirectly sampled value of the intraclass correlation is $\frac{\sigma_{f,t}^2}{\sigma_{f,t}^2 + \sigma_{e,t}^2}$, which when multiplied by 2 gives the t'th sampled value of heritability, h^2 for the full-sib model.

4.1.1 A Hybrid Proposal Algorithm

Let $(\mathbf{\Theta}, \Pi)^{(t)}$ be the state at iteration t. We propose the candidate state $(\mathbf{\Theta}, \Pi)'$ from the proposal distribution $q((\mathbf{\Theta}, \Pi)'|(\mathbf{\Theta}, \Pi)^{(t)})$. We assume that the parameters of the trait are independent of the pedigree configuration and choose independent proposals, using $q((\mathbf{\Theta}, \Pi)'|(\mathbf{\Theta}, \Pi)^{(t)}) = q(\mathbf{\Theta}'|\mathbf{\Theta}^{(t)})q(\Pi'|\Pi^{(t)})$. At the acceptance/rejection step,

$(\boldsymbol{\Theta}, \Pi)^{(t+1)}$ is assigned to be either $(\boldsymbol{\Theta}, \Pi)'$ with acceptance probability ρ, or $(\boldsymbol{\Theta}, \Pi)^{(t)}$ with probability $1 - \rho$, where the acceptance probability ρ is given by

$$\rho = \min\Big(1, \frac{P(\boldsymbol{\Theta}, \Pi|\mathbf{X}, \mathbf{Y})q((\boldsymbol{\Theta}, \Pi)'|(\boldsymbol{\Theta}, \Pi)^{(t)})}{P((\boldsymbol{\Theta}, \Pi)^{(t)}|X, Y)q((\boldsymbol{\Theta}, \Pi)^{(t)}|(\boldsymbol{\Theta}, \Pi)')}\Big) \tag{4.2}$$

The Metropolis-Hastings algorithm guarantees that as $t \to \infty$, the marginal distribution of $(\boldsymbol{\Theta}, \Pi)^{(t)}$ will converge to the stationary distribution $P(\boldsymbol{\Theta}, \Pi|\mathbf{X}, \mathbf{Y})$.

We observed that when using informative marker distributions, after running the chain a moderately large number of iterations, whereby the estimate of Π is close to the true configuration, most proposals Π differing from $\Pi^{(t)}$ will be rejected because small changes to the configuration can lead to large differences in the genetic and/or phenotypic contributions to the likelihood or posterior. After noting this we modified the proposal distribution such that instead of updating both $\boldsymbol{\Theta}$ and Π at each iteration, a hybrid moving algorithm is used with three different options to propose the candidate state $(\boldsymbol{\Theta}, \Pi)'$. In one case we update the parameters of trait $\boldsymbol{\Theta}$ only, in a second case we update the pedigree configuration Π only, and in the third case we update both $\boldsymbol{\Theta}$ and Π. At each iteration we choose from among these three options with preset probabilities p_1, p_2 and p_3, where $\sum_{i=1}^{3} p_i = 1$. This moving algorithm gives greater efficiency in searching the parameter space by increasing the acceptance probability ρ. A few selected simulation results are shown below to illustrate the difference between the two proposal strategies.

4.1.2 Choice of Prior Distribution

Specification of an appropriate prior distribution is the most substantial aspect of a Bayesian analysis which differentiates it from a classical analysis. In Equation 4.1, $P(\Pi)$ represents the prior distribution on the pedigree configuration. As was mentioned in chapter 3, we don't know the total number of elements in the configuration space apart from this being given by a finite Bell number. Without knowing the support, it is not clear what family of distributions on the configuration space might provide a reasonable prior. We do know one valid distribution - the uniform distribution with all configurations being equally likely. Even in that case, because the cardinality of the configuration space is unknown, we can only specify the uniform distribution up to an unknown constant of proportionality. However, looking at Equation 4.1 and Equation 4.2, it is clear that the unknown constant of proportionality disappears from both numerator and denominator when we calculate the acceptance probability ρ. This lack of requirement to know the normalizing constant is responsible for much of the applicability of the Metropolis-Hastings algorithm in general, and in Bayesian statistics in particular, as was a motivating factor in the original work of Metropolis et al. (1953).

The choice of prior distribution $P(\boldsymbol{\Theta})$ for parameters of the phenotypic distribution $P(\mathbf{Y}, \mathbf{X}|\boldsymbol{\Theta}, \Pi)$ will depend on the parameters present in the model that we try

to fit. In the next section we consider the choice of priors when using the one way random effects model with parameters μ, σ_f^2, σ_e^2 .

4.2 Implementation of Linear Mixed Model for Single Observation

Recall the one-way random effects model that we used in section 2.2 for the analysis of the full-sib model with known family memberships,

$$Y_{ij} = \mu + f_i + e_{ij}$$

where Y_{ij} is the phenotype of the j'th offspring of the i'th family. We assume the family effects are i.i.d $N(0, \sigma_f^2)$, independent of the residual errors e_{ij} which are assumed to be i.i.d. $N(0, \sigma_e^2)$.

4.2.1 Conditional Log-Likelihood Function

Let us start with a single family. Where n_i is the number of offspring in the i'th family, the phenotypic observations for that family are $\mathbf{Y}_i^T = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$. Under the assumptions on the model, $\mathbf{Y}_i^T$ has multivariate normal distribution with mean vector $\mu\mathbf{1}$ and covariance matrix $\mathbf{\Sigma} = \sigma_e^2\mathbf{I} + \sigma_f^2\mathbf{J}$, where $\mathbf{J} = \mathbf{1}\mathbf{1}^T$. $\mathbf{\Sigma}$, and $\mathbf{I}$ and $\mathbf{J}$ are all $n_i \times n_i$ matrices.

The patterned structure of $\mathbf{\Sigma}$ allows its inverse and determinant to be written explicitly as

$$\mathbf{\Sigma^{-1}} = \frac{1}{\sigma_e^2}\Big(\mathbf{I} - \frac{\sigma_f^2}{\sigma_e^2 + n_i\sigma_f^2}\mathbf{J}\Big) \tag{4.3}$$

$$|\mathbf{\Sigma}| = (\sigma_e^2)^{n_i - 1}(\sigma_e^2 + n_i\sigma_f^2) \tag{4.4}$$

This result is quite well known to applied statisticians. Details are provided, for example, in Searle et al. (1966). The determinant is found by using elementary row operations to reduce $\mathbf{\Sigma}$ to row echelon form, and then taking the product of the diagonal entries. The form of the inverse can be verified by showing that $\mathbf{\Sigma\Sigma^{-1}} = \mathbf{I}$, which is a straightforward calculation. Given these results, the contribution to the overall log-likelihood $l(\mathbf{\Theta}|\mathbf{Y})$ from the data $\mathbf{Y_i}$ of the i'th family, conditional on the configuration, is

$$\begin{aligned} l_i(\mathbf{\Theta}|\mathbf{Y}_i) = \ln P(\mathbf{y}_i|\mu, \sigma_f^2, \sigma_e^2) &\propto -\frac{1}{2}\ln(|\mathbf{\Sigma}|) - \frac{1}{2}(\mathbf{y}_i - \mu\mathbf{1})^T\mathbf{\Sigma^{-1}}(\mathbf{y}_i - \mu\mathbf{1}) \\ &= -\frac{1}{2}\Big((n_i - 1)\ln(\sigma_e^2) + \ln(\sigma_e^2 + n_i\sigma_f^2)\Big) - \frac{1}{2}\Big(\frac{\sum_{j=1}^{ni}(y_{ij} - \mu)^2}{\sigma_e^2} \\ &- \frac{\sigma_f^2 n_i^2}{\sigma_e^2(\sigma_e^2 + n_i\sigma_f^2)}(\bar{y}_i - \mu)^2\Big) \end{aligned} \tag{4.5}$$

where $\mathbf{\Theta} = (\mu, \sigma_f^2, \sigma_e^2)$ and $\bar{y}_i$ is the sample mean of the observations from family i. The assumptions of the random effects model imply that, conditional on the configuration, observations from different families are independent, whereby the overall log-likelihood, conditional on the family configuration, is

$$l(\Theta|\mathbf{Y}) = \sum_{i=1}^{I} l_i(\mathbf{\Theta}_i|\mathbf{Y}_i)$$

Fixed effects are easily accommodated. Conditional on the family membership, a mixed model for family i can be written as:

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + f_i\mathbf{1} + \mathbf{e_i}$$

where $\mathbf{X}_i$ is a matrix of covariates for family i, and $\boldsymbol{\beta}$ is a fixed parameter to be estimated, and $\mathbf{1}$ is a vector of 1's of length n_i. Assumptions on the random effects f_i and $\mathbf{e_i}$ are as before.

All that is required to incorporate the fixed effects terms is to replace $(\mathbf{y_i} - \mu\mathbf{1})$ by $(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ in the first line of Equation 4.5, with the associated adjustments to $y_{ij} - \mu$ and $\bar{y}_i - \mu$ in the second and third lines of that equation.

The model can be extended to the general linear mixed model and expressed in matrix form as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ where $\boldsymbol{\beta}$ represents a vector of fixed effects and $\mathbf{u}$ represents a vector of random effects. The first element of the vector $\boldsymbol{\beta}$ is typically the population mean with the first column of $\mathbf{X}$ being a column of 1's. We assume that $\mathbf{u} \sim \text{MVN}(\mathbf{0}, \boldsymbol{G})$ and $\mathbf{e} \sim \text{MVN}(\mathbf{0}, \boldsymbol{R})$. Under this model, $\mathbf{Y}$ is also multivariate normal, with mean X$\boldsymbol{\beta}$ and variance-covariance matrix $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$. In this case, the log likelihood for the data from family i was given by Equation 2.35 in section 2.3.2 as

$$l(\boldsymbol{\beta}, \mathbf{V}|\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \propto \ -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta})^T\mathbf{V^{-1}}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta})$$

For this more general model, the variance-covariance matrix $\mathbf{V}$ again has a patterned structure which aids in efficient computation of the inverse and determinant of $\mathbf{V}$.(Searle et al. (1966))). However, this entails evaluating the inverse and determinants of $\mathbf{G}$ and $\mathbf{R}$, and so unless these matrices are of dimension 3 or less, the underlying calculations need to be done using Gaussian elimination. This precludes using any but the simplest one way random effects model in the MCMC scheme which is used below, that is, fitting a full-sib model.

4.2.2 Prior Distribution of $\boldsymbol{\Theta}$

In the one way random effects model, the parameters of the phenotypic distribution are $\boldsymbol{\Theta} = (\mu, \sigma_f^2, \sigma_e^2)$. We have assumed that the components of the prior are independent, so that $P(\boldsymbol{\Theta}) = P(\mu)P(\sigma_f^2)P(\sigma_e^2)$.

The prior distribution for μ has been taken to be a normal distribution $N(\boldsymbol{\mu_o}, \boldsymbol{\sigma_o^2})$, which is from the conjugate family. In many situations when choosing the prior for a normal mean, one might take $\boldsymbol{\sigma_o^2}$ large, in which case the prior approaches a non-uniform, albeit improper prior. However, with the one way random effects model without additional covariates, the overall sample mean $\bar{\bar{y}} = \sum_i \sum_j y_{ij} / \sum_i n_i$ will be a consistent estimator of μ, and so an informative normal prior centred at $\bar{\bar{y}}$ and with

small variance seems justified.

Various non-informative prior distributions for variance components have been suggested in the Bayesian literature and software, including an improper uniform density (Gelman et al. (2013)), proper distributions such as an inverse gamma (0.001, 0.001)(Spiegelhalter et al. (1996)), and distributions that depend on the data-level variance (Box and Tiao (1973)). Gelman et al. (2006) explored and made recommendations for prior distributions for variance components. They illustrated that the choice of non-informative prior distribution can have a big effect on inferences, especially for problems where the number of groups is small or the group-level variance σ_f^2 is close to zero. Furthermore, in those works the family configuration structure was assumed known, in which case good estimates of the sample and family variances are available using ANOVA or linear mixed model type estimators as in chapter 2. More generally, the Bernstein-von Mises theorem (Bickel and Doksum (2006)) indicates that a reasonable choice of joint prior is given by a multivariate normal, centred at the MLE, and with covariance equal to the inverse Fisher information.

Without knowledge of the configuration structure, there does not appear to be a good data based informative prior for the variance components, and in this thesis we have used independent non-informative priors for σ_f^2 and σ_e^2, in the form of independent inverse gamma distributions. We have explored several choices of prior parameter, in one case using $\alpha = 2.001$ and $\beta = 2.0$, in which case the mean and

variance of the prior are $\frac{\beta}{\alpha-1} \approx 2$ and $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \approx 4000$. Another choice, recommended by Spiegelhalter et al. (1996) is $\alpha = 0.001$ and $\beta = 0.001$, in which case the inverse gamma distribution is close to the Jeffrey's prior, which is proportional to σ^{-2}. One sees that this is very informative in favour of small variances. In simulations we found that using either of these inverse gamma parameterizations did not work well when the initial value for one of the model variances was close to 0, in which case the sampling algorithm described below often got stuck with a variance in the vicinity of 0.

4.2.3 Moving algorithm for Θ and Π

As we mentioned before, we use independent proposals for the quantitative trait parameter Θ and pedigree configuration Π. We use the algorithm in Smith et al. (2001) to propose a new full-sib family configuration given the current configuration. The method guarantees that each configuration in the total space of all possible configurations is accessible with non-zero probability, and that the next state of Π' only depends on the current state $\Pi^{(t)}$. Independent random walk algorithms are used to propose the new candidate for the trait parameter $\Theta' = \left[\mu', (\sigma_f^2)', (\sigma_e^2)'\right]$, using a Gaussian random walk proposal for μ, and a lognormal proposal for variances, where the logarithm of the variance follows a Gaussian random walk, which ensures that the proposed variances take non-negative values.

4.2.4 Simulation Results

We generated phenotypic observations with different heritabilities (0.2, 0.4 and 0.8) for 759 salmon from 12 full-sib families, with family size varying from 8 to 140, based on the one way random effects model. In most of the simulations, only two loci were used, with 11 and 14 alleles respectively. We created 100 batches with population mean of 1500, variance of family effect of 40000, and residual variances of 360000, 160000, and 60000 according to the true heritability values. We estimated the pedigree structure and parameter of trait (heritability) with the two-step method, and by sampling the likelihood and posterior. The three estimation methods were applied to each of the 100 simulation batches for each choice of parameters. The prior distribution used for mean is normal centered at the sample mean, with a standard deviation of 10, and non-informative prior distribution have been used for variance components. We use more basic moving algorithm to propose the new candidate state where both $\boldsymbol{\Theta}$ and Π are changed at each step.

The box plots (Figure 4.1 - 4.6) show the distribution of estimated heritability and E2, a measure of pedigree accuracy. E2 is the number of unrelated pairs misclassified as full-sibs in the estimated pedigree. As was shown in chapter 3, this type of error is much more important than the number of full-sibs classified as unrelated, in terms of the effect on the estimate of heritability.

The plots show that there is not much difference between the Maximum Likelihood

Estimate (MLE) and Bayes estimates with the choice of prior used, but that both are superior to the two-step approach, regardless the pedigree reconstruction or heritability estimation. The plots indicate that the proposed method for simultaneously estimating pedigree and variance components works quite well, especially when the true heritability is high. In the case with low heritability $h^2 = 0.2$, corresponding to an intraclass correlation of only 0.1, the simultaneous MLE or Bayes estimates show less bias than the two step estimate, but are considerably more variable and prone to some very small estimates. At higher heritabilities, the variabilities of the two step and simultaneous estimates are more similar, but the simultaneous estimates show considerably less bias. In terms of pedigree errors, at least of the important type E2, the simultaneous estimates of pedigree are superior.

Each simulation batch was run for 10^6 Monte-Carlo iterations. For these marker data, pedigree estimates tend to be fairly stable once the pedigree reaches a size of 20-30 families. If on average the number of families at an MCMC iteration is 20, then on average 20 matrix inverses and determinants are required to evaluate the likelihood for each of the 10^6 proposals for each simulation batch. For a full-sib pedigree, there is no need for numerical inversion or determinant calculation, and the calculations are straight forward. For any other pedigree structure, extensive simulation would be impractical.

Some simulations were also carried out using all four available marker loci, to see

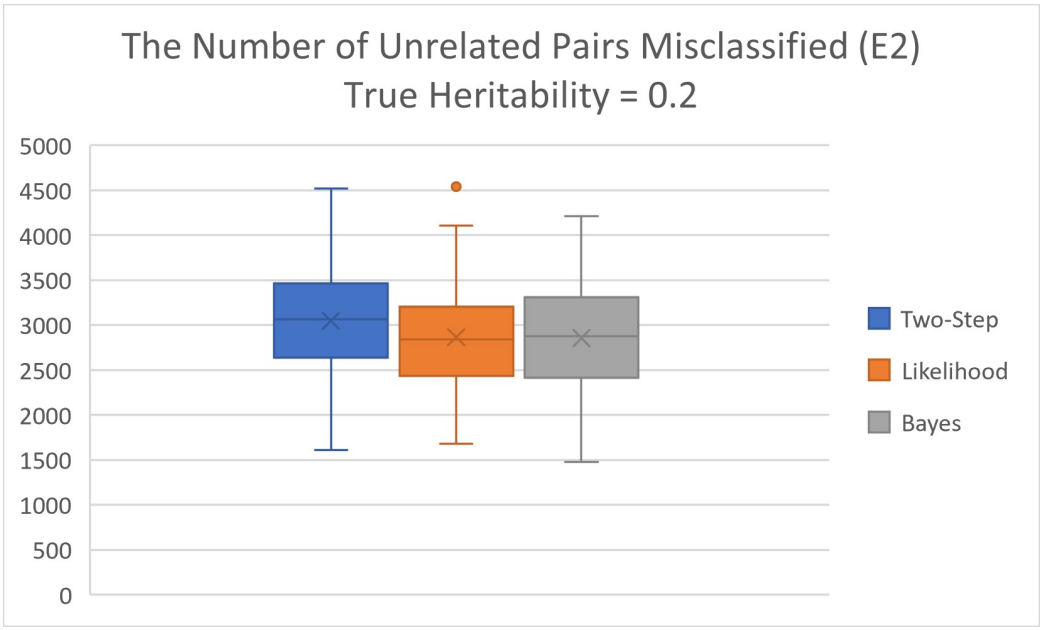

Figure 4.1: Performance of pedigree reconstruction with 2 loci when h^2=0.2

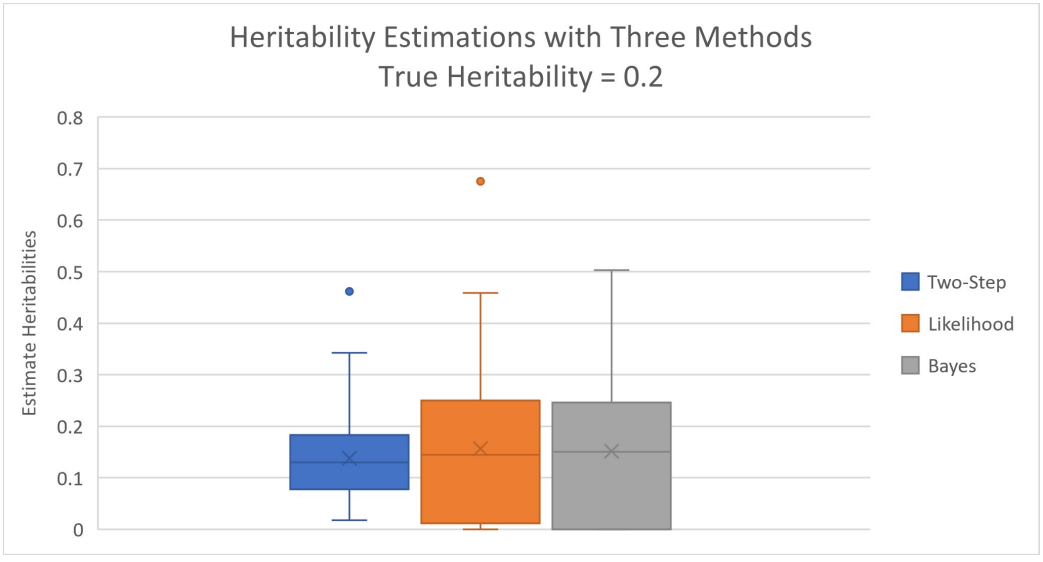

Figure 4.2: Heritability estimation with 2 loci when h^2=0.2. The variations of MLE and Bayes methods are larger than two-step, but the mean and median of estimated values are closer to the true heritability.

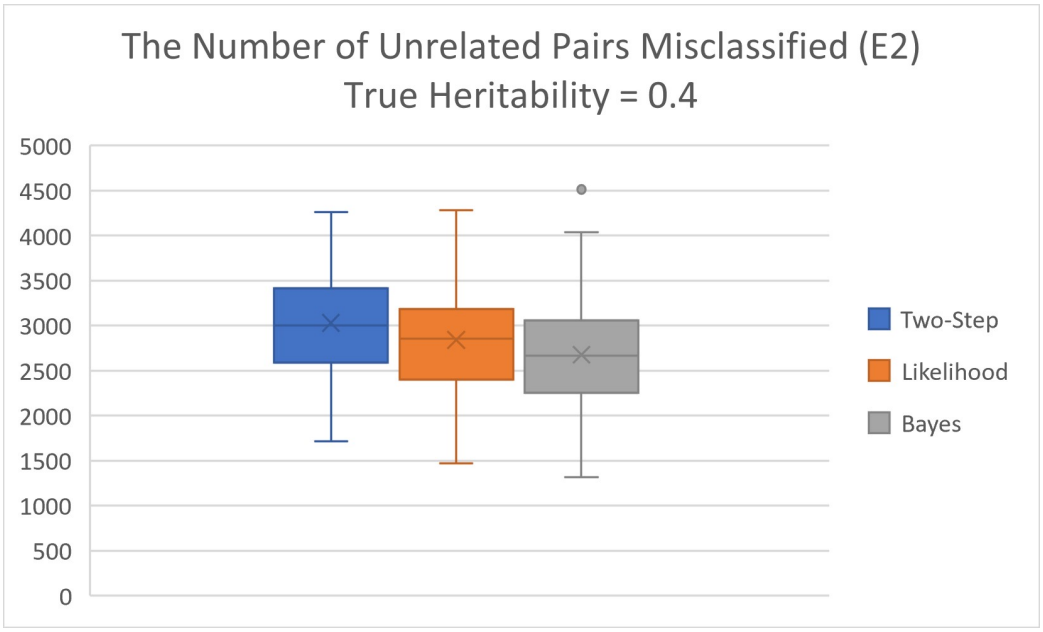

Figure 4.3: Performance of pedigree reconstruction with 2 loci when h^2=0.4

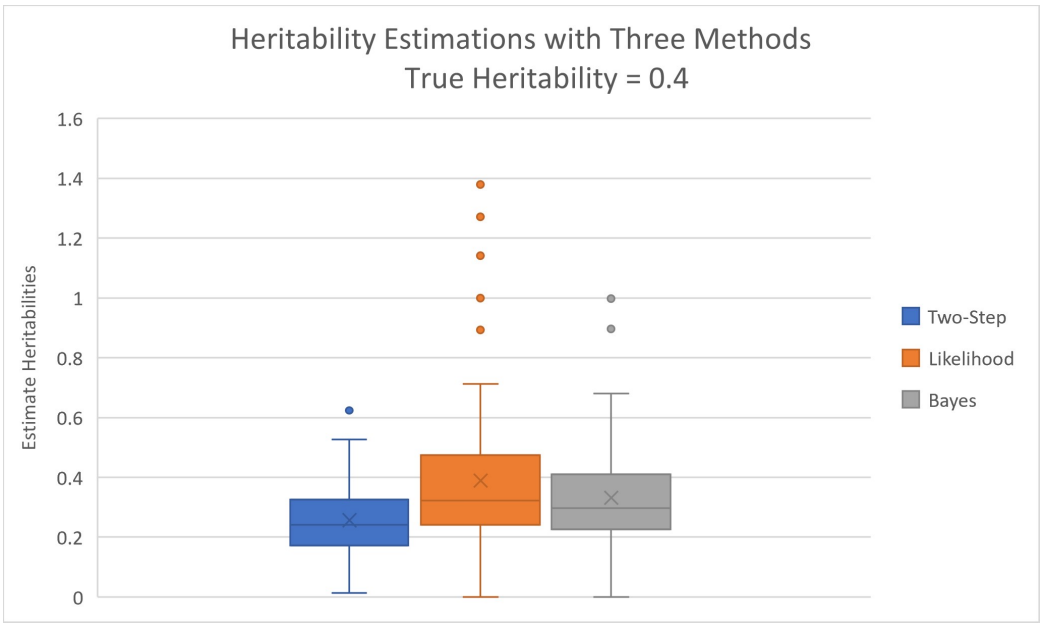

Figure 4.4: Heritability estimation with 2 loci when h^2=0.4

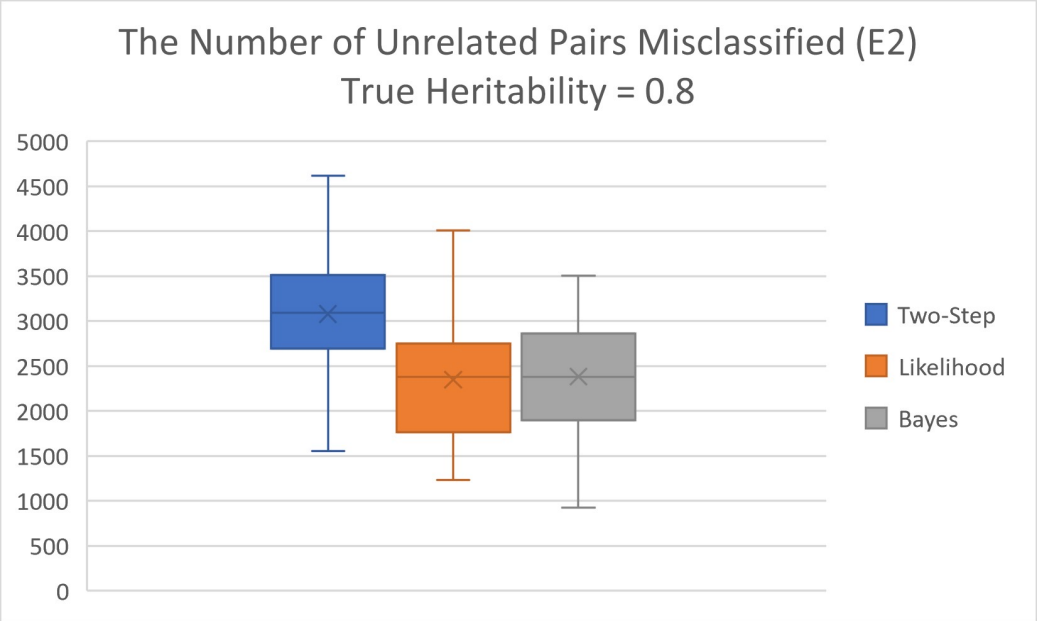

Figure 4.5: Performance of pedigree reconstruction with 2 loci when h^2=0.8. A significant improvement has been made by using MLE or Bayes method when the true heritability is large.The E2 values are significantly reduced. The phenotypic variances among families provides lot of information on pedigree reconstruction.

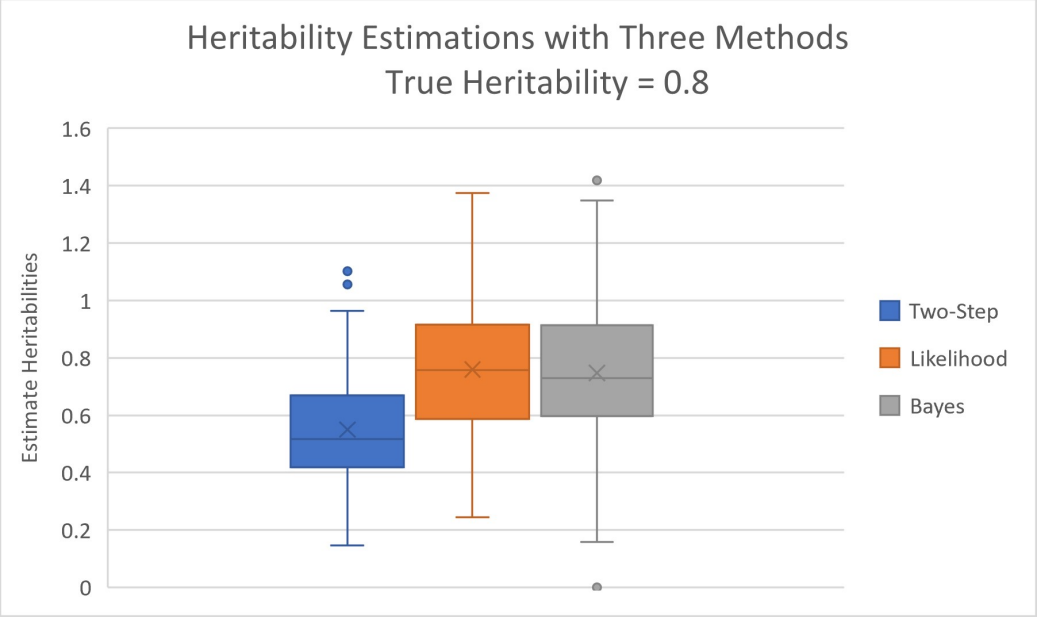

Figure 4.6: Heritability estimation with 2 loci when h^2=0.8. The estimation results are better when we use MLE or Bayes method. The median or mean of estimated values are much closer to the true value.

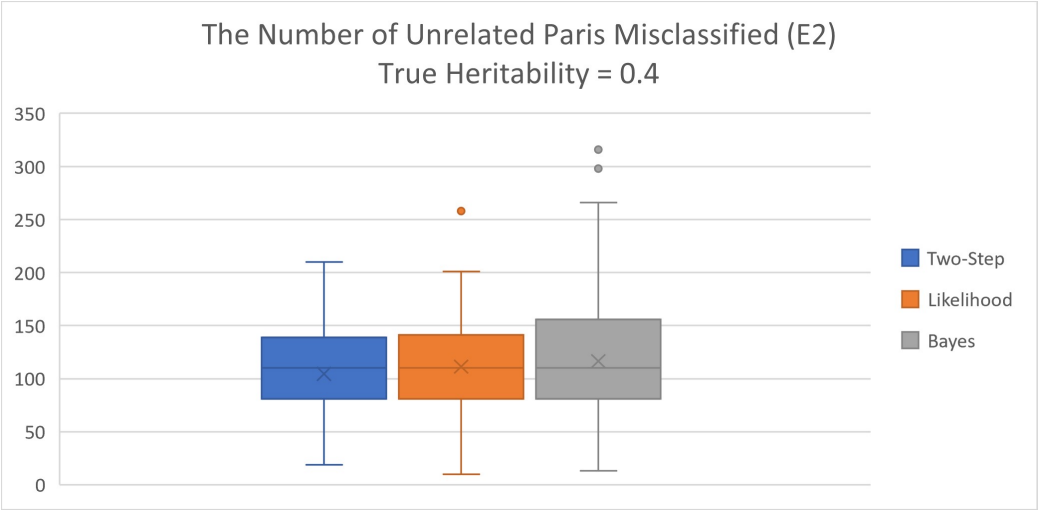

Figure 4.7: Performance of pedigree reconstruction with 4 loci when h^2=0.4.

the effect of having a very good estimated pedigree on the estimate of heritability. Figure 4.7- 4.8 show the results with true heritability equal to 0.4, and is representative of results with other heritability values. With the 4 informative markers used, in repeated runs, the estimated pedigree using genotypic data only, has only 7 or 8 out of 756 individuals incorrectly classified. This means that when carrying out the two-step procedure, the pedigree used is very close to the true pedigree. The two step procedure estimates the phenotypic parameters using a method which should be close to optimal - the one way random effects model with REML - when the pedigree is known. It is remarkable that the estimates from the simultaneous methods which randomly explore the parameter space should be so close to those from the two-step model. The two-step procedure gives just slightly less variable estimates than the sampled likelihood, with the mean and median of the estimates being essentially identical. The simultaneous estimators maximizing the posterior are a bit more variable.

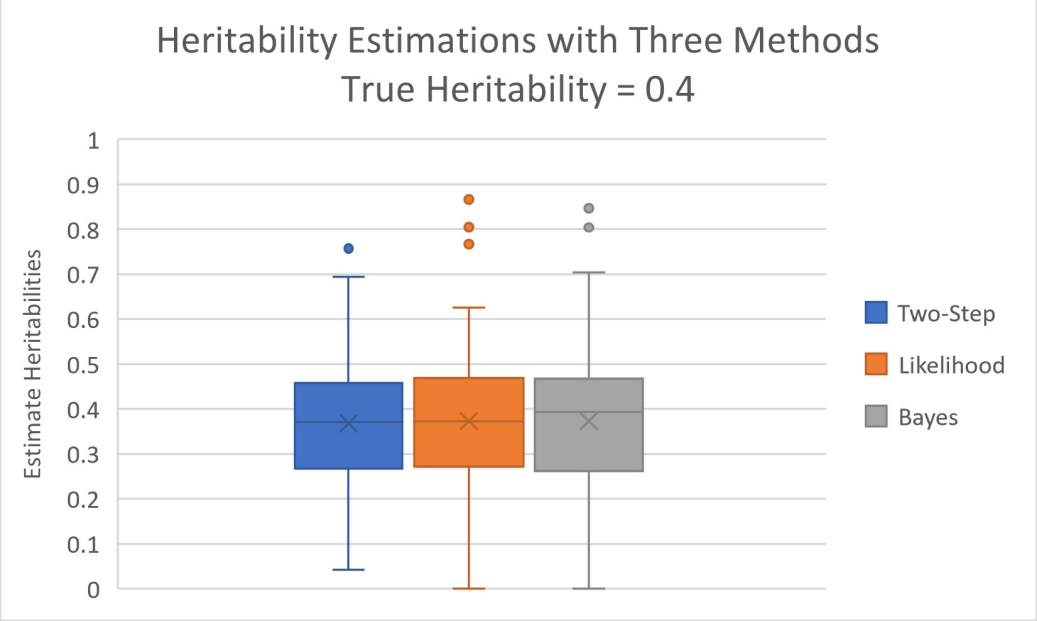

Figure 4.8: Heritability estimation with 4 loci when h^2=0.4. The estimation results are quite similar with all three methods when 4 loci used.

4.2.5 Real Data Analysis Results

We ran the proposed Markov chain Monte Carlo (MCMC) method for 3×10^6 iterations with different levels of genetic marker information (two loci, and four loci). Each run began at the all-unrelated configuration having 318 full-sib groups of size 1. We used the sample mean and the half of the sample variance as the starting points of the sub-parameters $(\mu, \sigma_f^2, \sigma_e^2)$ for the quantitative trait. The proposal distributions for the three parameters were as in the simulations described in section 4.2.3. We estimated the pedigree structure and heritability by sampling the likelihood and posterior. The prior distribution used for μ is normal centered at the sample mean, with a variance of 100. This informative prior is justified by the law of large numbers which says that the overall sample mean will converge to the common population

mean μ. We explored several choices for the prior distributions of the variance components (σ_f^2, and σ_e^2). We used independent priors for the variance components in all cases. We explored the use of improper priors, inverse gamma with both parameters equal to 0.001, which was recommended by Spiegelhalter et al. (1996) and for which neither the mean nor the variance is finite, and inverse gamma with $\alpha = 2.00001$ and $\beta = 2.0$, for which the mean and variance are finite. Then, we obtained the sampled intraclass correlation values from sampled variance components, and generated the sampled heritabilities by multiplying ICC by 2, as required for a full-sib model.

We note that when independent inverse gamma priors with identical parameters are used for the variance components, then a transformation of variables shows that the prior for the ICC has a beta distribution with mode at 1/2, which is an informative prior. Although the IG(0.001, 0.001) has been proposed by Spiegelhalter et al. (1996) as non-informative for variance components, in reality, that distribution favours small variance components.

A few results are summarized in Table 4.1. This shows the maximum likelihood estimates (MLE) and maximum *a posteriori* estimates (MAP), which are the heritability estimates where the likelihood or posterior for variance components are maximized over the 3×10^6 iterations. We also include the estimated heritability after fitting a linear mixed model with known pedigree, and the estimate from the two-step procedure, where pedigree is first estimated using marker data only, after

Table 4.1: Estimation of heritability from different methods

Time Point	Mixed Model With Known Pedigree	With 4 Loci Proposed Method			With 2 Loci Proposed Method		
		Two-Step	MLE	MAP	Two-Step	MLE	MAP
1	0.3234	0.3167	0.4620	0.3454	0.1705	0.3460	0.2126
2	0.3459	0.3249	0.3320	0.3317	0.1891	0.3459	0.3008
3	0.5805	0.5990	0.5116	0.5602	0.2941	0.5650	0.3750
4	0.5456	0.5583	0.5996	0.4640	0.3326	0.5250	0.5222

which a mixed model is fit conditional on the estimated pedigree. These two methods are included for comparative purposes.

The two-step method provides a reliable estimation of heritability when we have enough marker information to accurately estimate the pedigree, as it estimates the variance components conditional on the reconstructed pedigree, which is close to the true pedigree with four informative markers. Our proposed MCMC method is designed to sample from the joint likelihood or posterior distribution using both marker data and phenotypic observations, and estimate the pedigree and variance components simultaneously. Even when using just two loci, when the genetic marker data is insufficient to provide a very accurate pedigree estimate, the sampling method is still able to provide a reasonable estimate of heritability, which is in line with the simulation results of section 4.2.4. The Bayesian estimates in Table 4.1 used IG(0.001, 0.001) priors for the variance components.

As opposed to the MAP which is presented in Table 4.1, and which is the analogue of the MLE but with prior included, most Bayesian analyses focus instead on

the posterior mean or median, with the posterior mean being the Bayes estimator under squared error loss, and a Bayesian analysis often also reports a credible interval. Additionally, unlike the MLE, the MAP estimate is not invariant under reparameterization - the MAP estimate of heritability can't be computed from the MAP estimate of the variance components. We are not able to find the MAP estimate, posterior mean or median of heritability directly, since we don't generate a Markov chain from the posterior distribution of heritability. In this situation, we think that the plugged-in MAP value provides a reasonable estimation of heritability.

In an attempt to construct a sample of independent values of heritability, every 1000'th value of the variance components were selected from 2×10^6 iterates, after removing an initial transient of 10^6 iterates, and the heritability was calculated using the sampled variance components. The mean and median of posterior distributions are listed in the Table 4.2, again using IG (0.001, 0.001) priors for variance components.

In general, the proposed method with simultaneous estimation of variance components and pedigree provides a reasonably accurate estimate of heritability even when only two loci are used, regardless of whether the MLE or MAP are reported.

Table 4.2: Mean and median of posterior distribution of heritability

	With 4 Loci		With 2 Loci	
Time point	Mean	Median	Mean	Median
1	0.2876	0.2676	0.1607	0.1454
2	0.2748	0.2558	0.2499	0.2307
3	0.4102	0.3894	0.2689	0.2478
4	0.4338	0.4142	0.3379	0.3164

4.2.6 Assessing Convergence When Using a Single Chain or Multiple Chains

Establishing convergence of Markov chain Monte Carlo (MCMC) is one of the most important steps of the Bayesian analysis. In the thesis we have generally made calculations based on single long samples from a likelihood or posterior distribution. We used standard convergence diagnostics - trace plots and sample autocorrelation function - to assess the convergence of the sampler, and only used runs for which the diagnostics suggest convergence and independence. Multiple chains (with dispersed starting points) can be used for estimation as well. However, we didn't find a significant difference between these two methods.

When we use multiple chains, we use an ANOVA type procedure to assess convergence, which is similar to the Gelman-Rubin diagnostic (Gelman and Rubin (1992)), but we modify the procedure slightly in order to select a set of chains on which to carry out inference.

The following is a representative analysis using phenotypic observations at time

point 4, and genotypes at 4 loci. The goal is to draw independent samples from the posterior distribution of ICC, which is the half of heritability in the full-sibs family structure. An informative prior $N(\bar{X}, 10)$ was used for μ, and IG(0.001, 0.001) priors for each of the variance components.

To assess stability with respect to the starting point, 10 sets of starting iterates for the variance components were generated by drawing two variance components independently from an uniform distribution on $(0, V)$, where V was the sample phenotypic variance. For each of the 10 replicates, the MCMC posterior sampling algorithm was run for 3×10^6 iterations, and these were subsampled to get 200 points, after dropping an initial transient. Figure 4.9 shows the resulting 2000 sampled values of ICC, together with the sample partial autocorrelation function, and a histogram.

It is clear that sampler was stuck in two of the 10 replicates, and no formal assessment is needed to conclude that these combined 10×200 points do not represent a sample of 2000 points from a stationary distribution. However, it is instructive to consider a one way anova with run number (1 through 10) as a factor. The logarithm of ICC was used as the outcome in the ANOVA, after looking at normal QQ plots of residuals for untransformed ICC, and log, logit and square root transforms.

The ANOVA output in Table 4.3 suggests that after removing runs 3 and 6, the

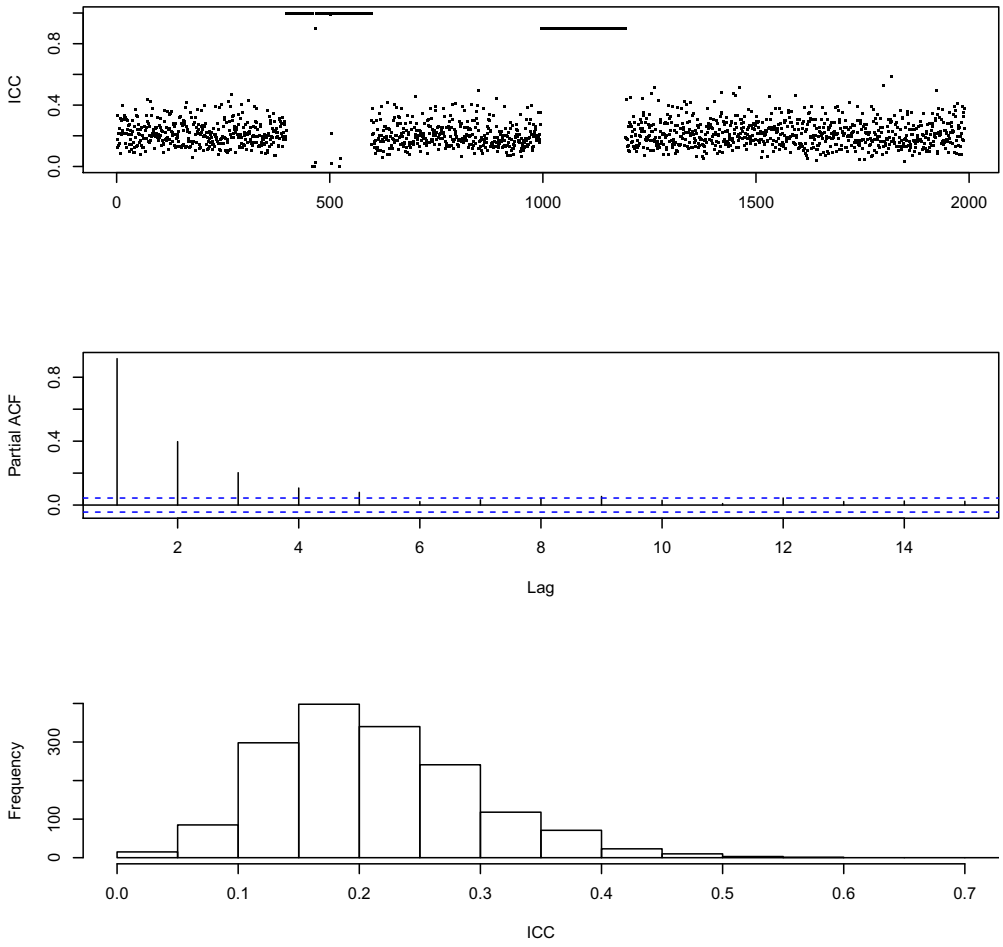

Figure 4.9: ICC with 10 replicates started at random points. top: sampled ICC; middle: estimated partial autocorrelation function; bottom: estimated posterior distribution

	Estimate	Std. Error	t	p-value
Intercept	-1.58	0.06	-25.0	<.001
run 2	-0.04	0.08	-0.4	0.63
run 3	1.08	0.08	12.1	<.001
run 4	-0.03	0.08	-0.4	0.67
run 5	-0.11	0.08	-1.2	0.21
run 6	1.47	0.08	16.5	<.001
run 7	-0.04	0.08	-0.5	0.61
run 8	0.03	0.08	0.4	0.66
run 9	-0.08	0.08	-0.9	0.34
run 10	-0.05	0.08	-0.6	0.55

Table 4.3: ANOVA of log(ICC) on run number 1-10

remaining ICC's will be from a stationary distribution. However, the variance is inflated by runs 3 and 6, and a second anova (not shown) after removing those two runs indicates that samples from run 5 are significantly different from the remaining 7 runs. A third analysis, after removing runs 3, 5 and 6 is summarized in Figure 4.10 and Table 4.4. There is no discernible pattern in the 1400 sampled values of ICC. The partial autocorrelation function supports a conclusion of independent observations, and the ANOVA of log(ICC) on run shows no evidence of run to run variation, with an overall F-test p-value equal to .066. The median and mean of these 1400 sampled ICCs are .205 and .214, respectively.

As we mentioned, all the results presented in previous section are calculated based on single long samples from a posterior distribution or likelihood. Figure 4.11 is instructive of the case where a single long sample is used, and shows the trace, PACF and histogram using 30,000 observations sub-sampled from run 10 alone.

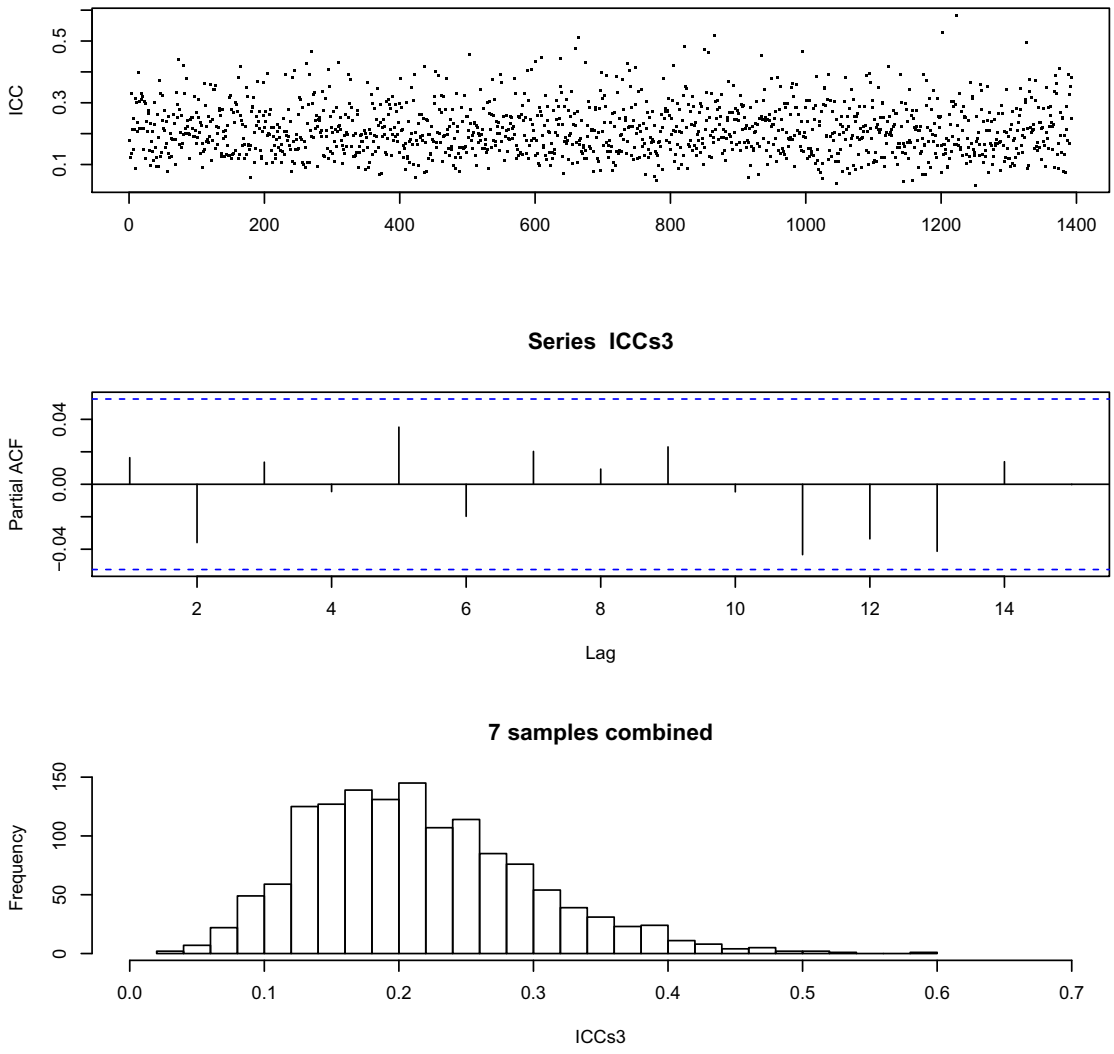

Figure 4.10: ICC for 7 replicates started at randomly sampled values of the variance components. Replicates 2,5 and 6 have been dropped due to evidence of non-stationarity.

	Estimate	Std. Error	t	p-value
Intercept	-1.58	0.03	-55.36	< 0.001
run 2	-0.04	0.04	-1.05	0.29
run 4	-0.03	0.04	-0.93	0.35
run 7	-0.04	0.04	-1.12	0.26
run 8	0.03	0.04	0.96	0.33
run 9	-0.08	0.04	-2.11	0.03
run 10	-0.05	0.04	-1.32	0.18

Table 4.4: ANOVA of log(ICC) on run number 1, 2, 4, 7, 8, 9, 10

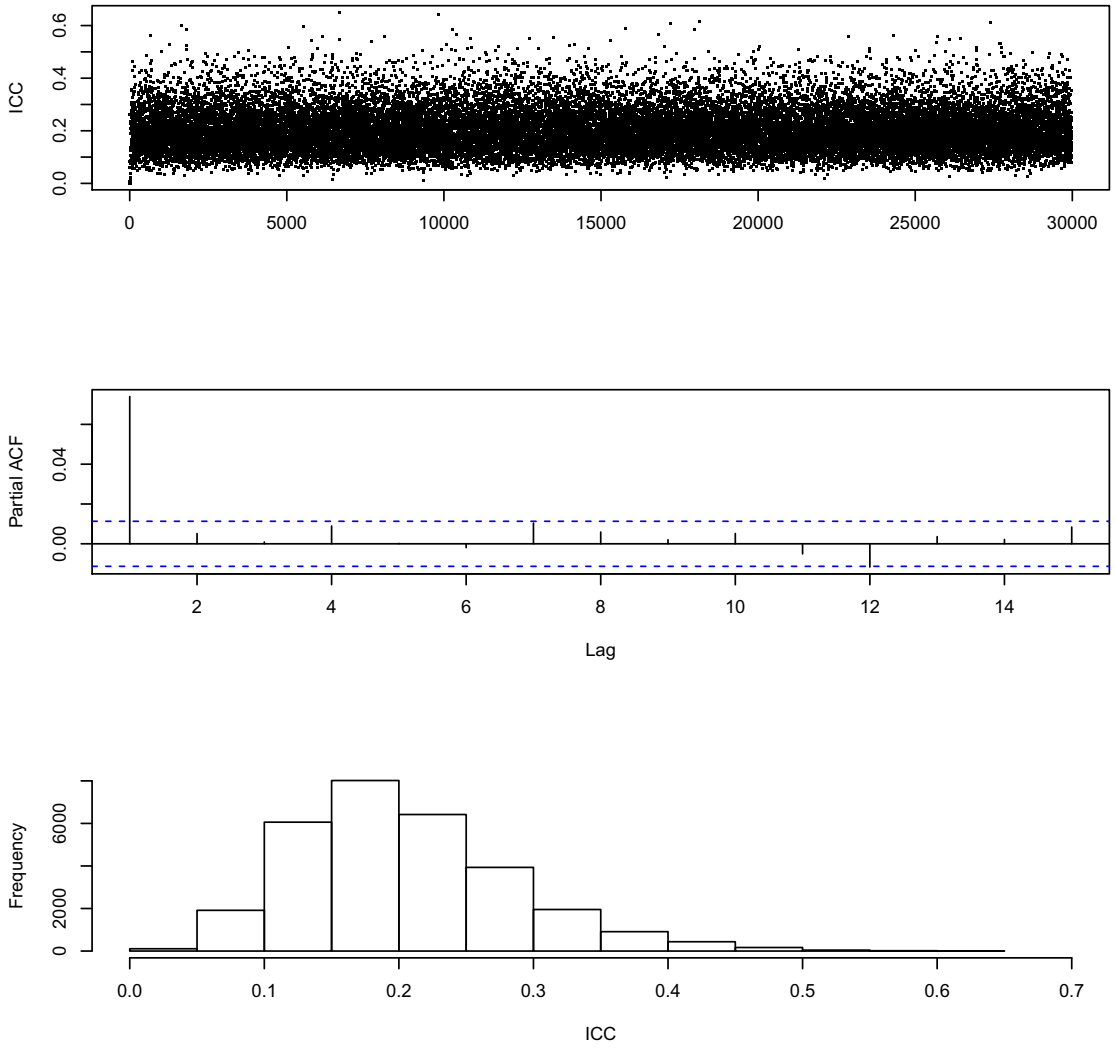

Figure 4.11: 30,000 sampled values of ICC from run 10.

The trace plot visually shows no evidence of non-stationarity. However, the estimated PACF shows a lag 1 value substantially greater than 0, indicating non-independence, and compatible with an AR(1) process. After further sub-sampling by taking every 10'th of the last 20,000 observations, we arrive at Figure 4.12, which is compatible with a random sample of 2000 points from a stationary distribution. The median and mean of these ICC values are .193 and .205, respectively.

In this case, as seen in 4.13, there are minimal differences among the posterior distributions regardless of sampling scheme - multiple samples, one long sample showing some dependence, or a single sample after further sub-sampling. The median estimates of ICC, for example, are .205 using multiple chains, and .193 for a single long chain with or without sub-sampling to ensure independence. The additional sub-sampling is an attempt to ensure the nominal level of credible interval. In this case the empirical 95% credible intervals for ICC are similar for all methods - (.079,.393) for the single chain without sub-sampling, (.079, .404) for single chain with sub-sampling, and (.083, .397) using multiple chains.

4.3 Extension to Multiple Observations

The model developed for a single time point can be extended to include observations at multiple time points. For computational purposes, it is essential to maintain a covariance structure for which the covariance matrix has explicit inverse and determinant. One approach is to define a simple model that assumes the observations at different times are conditionally independent given the pedigree.

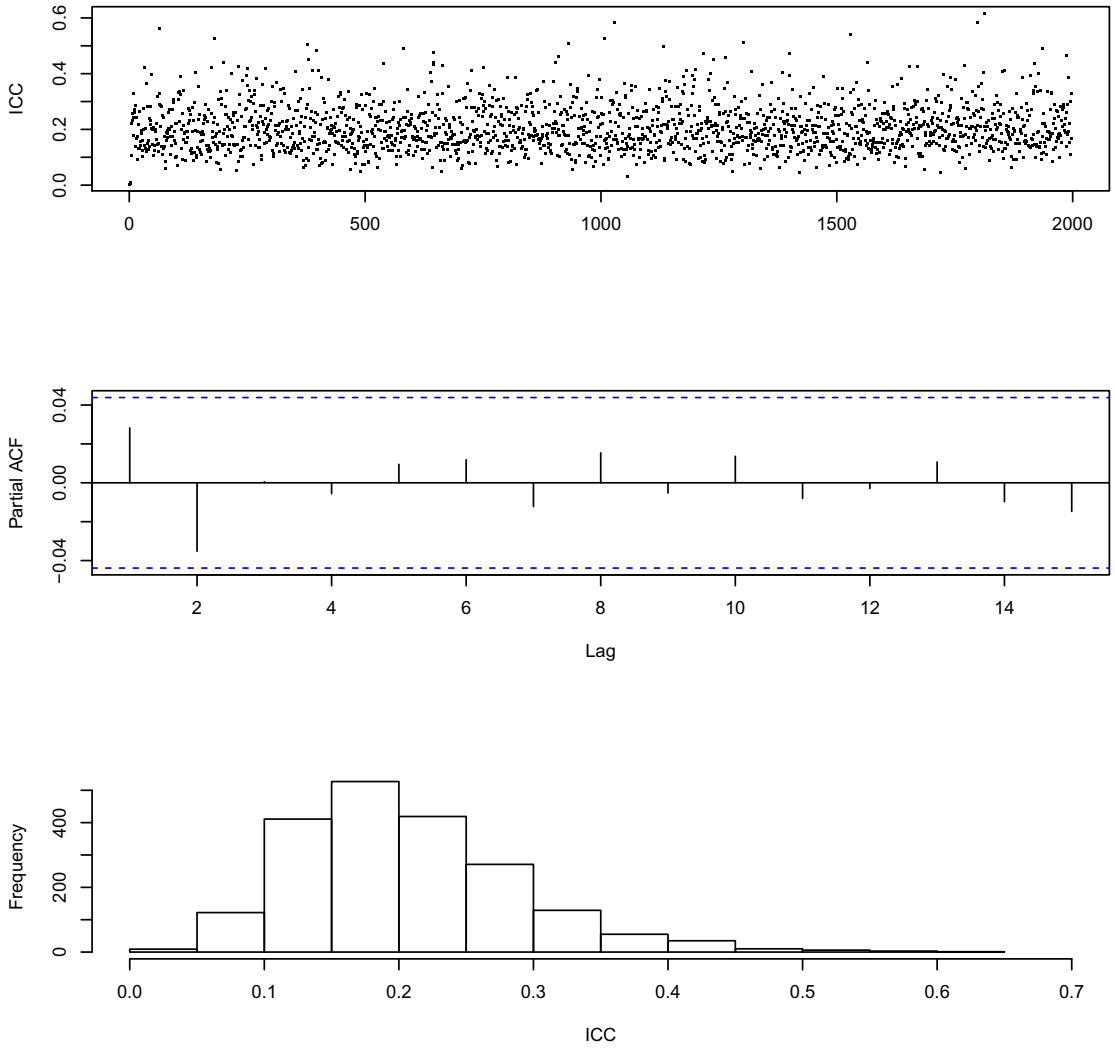

Figure 4.12: 2,000 sub-sampled values of ICC from run 10.

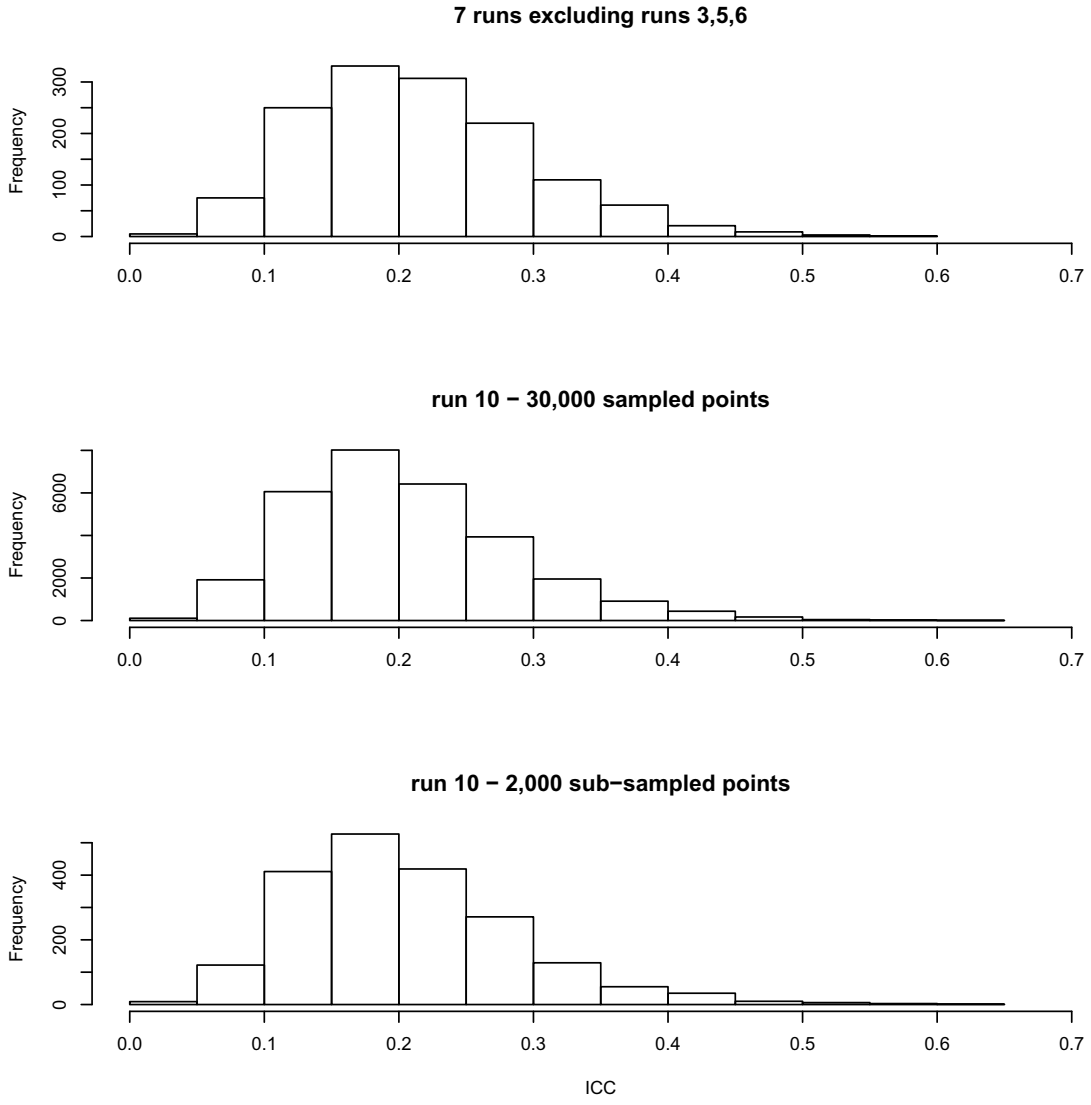

Figure 4.13: Histogram estimates of posterior distribution of ICC. top: using sampled values from 7 chains started at random points; middle: using 30,000 observations from a single chain; bottom: using 2,000 sub-sampled points from the single chain.

Take the example of observations at two time points. Let Y_{ijk} be the observation on subject j in family i at time $k \in (1, 2)$, the model can be expressed as

$$Y_{ij1} = \mu_1 + f_{i1} + e_{ij1} \tag{4.6}$$

$$Y_{ij2} = \mu_2 + f_{i2} + e_{ij2} \tag{4.7}$$

We also assume that all random terms at times 1 and 2 are independent, and that the family effects and additive error terms are independent at each time point.

Fixed effect terms are easily accommodated. The population means μ_1 and μ_2 can be replaced by terms such as $\mathbf{X}\boldsymbol{\beta}_1$ and $\mathbf{X}\boldsymbol{\beta}_2$. We allow for different model parameters at times 1 and 2, and the model can be generalized to several time points.

The observations at times 1 and 2 are conditionally independent given the pedigree. However, they are dependent when integrated over the pedigree. This means that in the Metropolis-Hastings (MH) sampling algorithm, when a configuration is proposed, the joint conditional likelihood, conditioned on the proposed pedigree, is calculated using observations at times 1 and 2. This uses the same proposed configuration at each time point, rather than running two independent runs of the MH algorithm, one for the time 1 observations, and one for the time 2 observations.

This model was fit to the observations at times 1 and 2 of our merged data, using four marker loci, and sampling from the likelihood. The MLE of heritabilities for the joint likelihood at times 1 and 2 are 0.3076 and 0.3901. The traces and ACF's for both times didn't show any patterns to suggest lack of convergence, or dependence of the sampled observations. However, in comparing with the estimates from running independent chain at each time point (Table 4.1), the joint estimation doesn't offer any improvements.

An alternative approach to maintain the covariance matrix in a special structure admitting explicit calculations of inverse and determinant is to use a one way mixed effect model conditioning an observation at a time point on observations for the same individual at one or more previous times. We still use the first two time points for illustration.

Let Y_{ijk} be the observation on subject j in family i at time $k \in (1, 2)$. A model with independent increments, but with measurement at the second time point being measurement at time 1 plus an increment is as follows:

$$Y_{ij1} = \mu_1 + f_{i1} + e_{ij1} \tag{4.8}$$

$$Y_{ij2} = \beta Y_{ij1} + \mu_2 + f_{i2} + e_{ij2} \tag{4.9}$$

We assume that all random terms at times 1 and 2 are independent, and that the family effects and additive error terms are independent at each time point. At the second time point, conditional on the observation at the first time point, that is $Y_{ij1} = y_{ij1}$, the model at time 2 has the same block diagonal covariance structure as the model at time 1. This model uses a regression approach, regressing the observation at time 2 on time 1.

Using this model and sampling with likelihood, the MLE of heritability at first times 1 and 2 are 0.2773 and 0.2093, respectively, where the estimate of heritability at time 2 can be expressed as

$$\widehat{h_2^2} = 2 \times \widehat{ICC} = 2 \times \left(\frac{\hat{\beta}^2 \widehat{\sigma_{f_1}^2} + \widehat{\sigma_{f_2}^2}}{\hat{\beta}^2 \widehat{\sigma_{f_1}^2} + \widehat{\sigma_{f_2}^2} + \hat{\beta}^2 \widehat{\sigma_{e_1}^2} + \widehat{\sigma_{e_2}^2}} \right) \tag{4.10}$$

The smaller estimate at the second point reflects that the time two increment has estimated family variance $\hat{\sigma}_{f_2}^2$ which is essentially 0. This is not a bad estimation of $\sigma_{f_2}^2$. The box plots (Figure 4.14) of residuals from the regression of time 2 on time 1 observations, by family, show that the residuals do not retain much variation between the different families.

The estimated regression coefficient ($\hat{\beta}$) from fitting the model with regression of Y_{ij2} on Y_{ij1} is 1.5096. This is the estimated value of β where the overall likelihood, now a function of 4 parameters (β, μ_2, $\sigma_{f_2}^2$ and $\sigma_{e_2}^2$) plus the configuration, is maximized. Therefore, the estimated population mean at the second time point can be computed

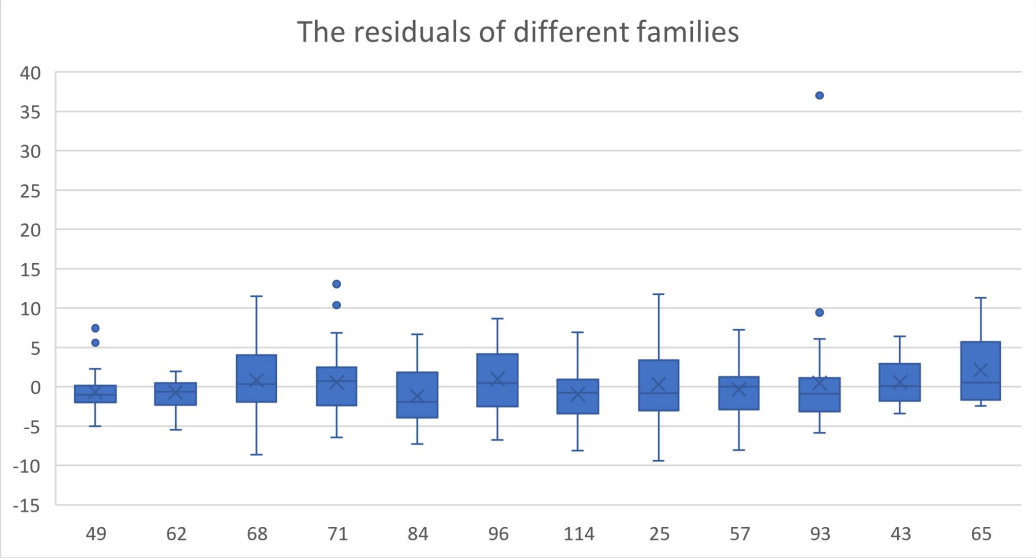

Figure 4.14: The box plots of residuals for each family

as

$$\widehat{\mathrm{E}}(\mathbf{Y}_2) = 1.5096 \times \widehat{\mathrm{E}}(\mathbf{Y}_1) + \widehat{\mathrm{E}}(\mu_2) \tag{4.11}$$

Using this equation, the estimated population mean at the second time point is 1828.72. This turns out to be a reasonably good estimate, as the sample mean of phenotypic observations at time 2 is 1837.242. Our estimate from the regression model is within 10 of the sample mean of Y_2.

Chapter 5

Conclusions and Further Suggestions

5.1 Conclusions

Estimates of the genetic variance of quantitative traits in a population are important because they inform us about its potential ability to evolve in response to novel selective challenges. This corresponds to one of Brookfield's definitions of the evolvability of a population:"a description of its current standing crop of genetic variability, and the consequence of the extent and nature of this variation for the population's ability to respond to current selective pressures" (Brookfield (2009)). The analysis of a series of relationships between relatives provides the basis for partitioning the phenotypic variance into its elementary components. In practice, we are often confronted with difficulties. Some of the variance is essentially beyond reach in a statistical sense, such as the variance caused by higher-order epistatic interactions. Nevertheless, with appropriate experimental designs, most of the fundamental sources of variance (additive and dominance genetic variance, and environmental variance due to common maternal environments) can be estimated to a good degree. Most practical applications of quantitative genetics have been concerned with only the additive genetic component of the phenotypic variance, other components are treated as noise. The

ratio σ_A^2/σ_Y^2 has come to be known as the narrow-sense heritability of a trait.

The parent-offspring regression method is commonly used for estimating heritability since the desire to obtain a heritability estimate stems from a particular interest in the resemblance between parent and offspring phenotypes, so it is natural that this resemblance should be measured directly. However, in the situation where the information of both parent and offspring generations are not available, the analysis of sibs provides an alternative to estimate quantitative genetic parameters. The analysis of variance (ANOVA) is designed to deal with these kinds of data. The total phenotypic variance can be partitioned into within family and among family components, both of which can be interpreted in terms of genetic covariances between relatives.

In section 2.2, we introduced a procedure for the estimation of heritability with a full-sib family design which is appropriate for our our hybrid data set that includes only full-sib or unrelated relationships. When the natural population presents the investigator with highly unbalanced family sizes and fragmentary data from numerous kinds of relationships, maximum likelihood and restricted maximum likelihood estimators with a general mixed model are preferred, since they don't place any special demands on the design or balance of data and they can be obtained readily for any arbitrary pedigree of individuals. We illustrated how ML and REML procedures can be used to estimate variance components and how these estimates differ, using a simple example in section 2.3. We then introduced the ML and REML equations for variance component estimation under the general mixed model. At the end of

chapter 2, we presented the estimates of heritability for measurements at each time point in the hybrid data set. These can be considered as the best estimation results, and they can be used as the bench mark for evaluating other estimation procedures.

All traditional techniques for estimating variance components that were introduced in chapter 2 require knowledge of the relationships among the individuals recorded. Except in humans, some animals in zoological parks, and some domesticated species, relationships of free-ranging individuals are generally unknown, and even in the best situations, paternity is often uncertain. In chapter 3, we showed how to estimate variance components in the the absence of direct observations on relationships. Motivated by the variance decomposition formula, we defined a heritability like object in the mixture model by consideration of the one way fixed effects ANOVA model, which is the analogue of the Gaussian mixture model when component memberships are known. The expectation-maximization algorithm can be used to estimate the parameters in the Gaussian mixture model. However, when the number of families is completely unknown and the distributions of phenotypic observation of each family are indistinguishable, this method doesn't provide a good estimation of heritability. We also used the classic clustering method K-means to group the individuals first, and then fitted random effects model to estimate the variance components conditional on the estimated family memberships. The associated grouping into full-sib families was not accurate, leading to poor estimates of variance components. Estimating heritability with unknown pedigree is very challenging, but

we still believe that phenotypic observations carry some information on the pedigree structure, and that they should be incorporated in the estimation procedure.

With the development of highly polymorphic molecular markers (especially microsatellies), molecular-based tools for inferring genetic relationships have become popular. In section 3.2, we explained a detailed two-step procedure for estimating variance components: first, families of sibs are reconstructed using a Markov chain Monte Carlo (MCMC) based method to maximize the configuration likelihood using marker data, and second, the reconstructed sibships are used to estimate variance components by fitting the mixed model. Using the MCMC method that was developed by Smith et al. (2001), the pedigree was reconstructed with different levels of genetic marker information (two loci, and four loci). The estimation results clearly indicate that with a more accurately estimated pedigree, the two step estimated heritabilities are closer to the best estimation results using a mixed model with known pedigree. We noted that unrelated pairs incorrectly classified as full-sib pairs are more detrimental to the two-step estimation procedure than are full-sib pairs incorrectly classified as unrelated. In particular, a large true full-sib group split into two moderate sized but different full-sib groups, has very little impact on the estimate of heritability.

The performance of two-step method heavily depends on the reconstructed pedigree. With insufficient marker information, the accuracy of estimation is not guaranteed. Therefore, we developed a hybrid Markov chain Monto Carlo procedure, where

marker information ($\mathbf{X}$) and phenotypic observations ($\mathbf{Y}$) can be used jointly and simultaneously to estimate both the pedigree and heritability. We started with development of the joint posterior distribution $P(\mathbf{\Theta}, \Pi|\mathbf{X}, \mathbf{Y}) \propto P(\mathbf{Y}|\mathbf{\Theta}, \Pi)P(\mathbf{X}|\Pi)P(\mathbf{\Theta})P(\Pi)$ and used the Metroplis-Hastings algorithm to generate values of $(\mathbf{\Theta}, \Pi)$, where $\mathbf{\Theta}$ represents the parameter of the quantitative trait and Π is the pedigree configuration. The associated sampled values of $\mathbf{\Theta}$ from the marginal distribution $P(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y})$ were used to create indirect samples of heritability.

Restricting to full-sib or unrelated pedigrees, the random effects model and general linear mixed model were considered, providing the conditional distribution of $P(\mathbf{Y}|\mathbf{\Theta}, \Pi)$. In section 4.2, we discussed the details of this implementation. The patterned structure of covariance matrix of $\mathbf{Y}$ allows us to find the inverse and determinant explicitly. Given this result, we were able to write the conditional log-likelihood function of $l(\mathbf{\Theta}|\mathbf{Y})$. We could develop a robust method to deal with non-Gaussian family effects or errors, especially long tailed distributions. We could substitute a multivariate t-distribution for the multivariate normal, and this accommodates the same pattern in the covariance matrix. Therefore, we are still able to compute the likelihood. For a Bayesian analysis of the random effects model we applied independent prior distributions for each of the sub-parameters (population mean, variance of family effects and variance of errors) in the $\mathbf{\Theta}$. An informative normal distribution was selected as the prior distribution of μ, centered at overall sample mean of the observations with small variance. Various prior distributions for variance components

(σ_f^2, and σ_e^2) were discussed, and we explored the use of non-informative/improper uniform and inverse gamma (0.001, 0.001) priors to fit the model.

Box plots for the simulation study showed that the MLE or Bayes estimates were superior to the two-step approach in the situation where insufficient marker information was used. The proposed simultaneous estimation procedure worked quite well, especially when the true heritability was high. From the analysis for the real hybrid data, when only two marker loci were used, compared with the two-step method, the MLE and MAP with the simultaneous estimation procedure were closer to the estimate of heritability from fitting a mixed model with true pedigree. Since the MAP estimate is not invariant under reparameterization, the MAP of heritability can't be obtained directly from MAP of variance components, and so we examined the mean and median of samples from the posterior distribution. When four loci were used and independent inverse gamma (0.001, 0.001) prior distributions were chosen for variance components, the sampling algorithm behaved well in terms of the convergence diagnostics, and there were no substantial differences from estimates using non-informative priors.

We have extended our method for multiple observations. First, we assumed the observations at different times are conditionally independent given the pedigree, and this setup promoses the covariance matrix has explicit inverse and determinant. We used the first two time points as the example and estimated the heritabilities at times

1 and 2 simultaneously by maximized the joint likelihood of both observations. Another approach to maintain the covariance matrix has explicit inverse and determinant was to regress the observation at time 2 to time 1 with an independent increment. However, the results of joint estimation method or regression approach with multiple observations didn't indicate any significant improvements compared with estimation independently for each time point.

5.2 Further Suggestions

There are a number of areas for further exploration. The proposal used had independent random walks on the log scale for variance components. The speed at which the parameter space is explored is dependent on the variance of the innovation in the random walk. A more extensive investigation of the influence of the variance is needed. There were related issues which arose when a variance component came close to zero, after which the sampler was not able to escape from such a local point. We plan to explore the use of so-called independence chains for proposals, to see if as sampler can be developed which is not so prone to being stuck near local optima.

In the Bayesian context there are a number of questions related to the choice of prior. If there is a supposition as to the value of the underlying heritability, say in the form of an informative beta prior, what are the informative joint priors on variance

components which are compatible with the prior on heritability. Also, we used independent priors on the pedigree. Suppose that there is information on the pedigree - for example a fixed number of sires and dams being used in a mating experiment, but with parentage unknown, such as might happen in a fisheries context. More generally, suppose that there is a prior on the number of families. How does a prior distribution on family memberships follow, keeping in mind that feasible configurations must satisfy genetic constraints on alleles?

The concept of heritability was developed in the context of a scalar quantity. Are there related concepts which are appropriate to multivariate traits? In the one-way random effects model, the ICC, which is the ratio of the family variance to the total variance, could be generalized by replacing variance by the generalized variance - the determinant of the covariance matrix. In this way there is an analogue of the wide sense heritability in the multivariate case, at least in a statistical sense, but does it have biological relevance.

Often when serial measurements on a characteristic are made, a growth curve model is fit, such as a four or five parameter logistic model. Can such models be developed which allow for between and within family variation? This might be accomplished through some form of hierarchical model whereby the parameters of the growth curve represent a level in the hierarchy that incorporates family effects. Conditional on the parameters, time series of observations for individuals could then be

based on a growth curve model. With unknown pedigree, sampling from a likelihood or posterior would be very challenging computationally.

Appendix A

Expectiations, Variances and Covariances of Compound Variables

A.1 The Delta Method

Consider an arbitrary expression f, which is a function of x. Performing a Taylor series expansion around an arbitrary constant c,

$$f = f(c) + (x-c)\frac{\partial f(c)}{\partial x} + (x-c)^2\frac{\partial^2 f(c)}{2\partial x^2} + (x-c)^3\frac{\partial^3 f(c)}{3\cdot 2\partial x^3} + ... \tag{A.1}$$

where $f(c)$ refers to the function evaluated at $x = c$, and the partial derivatives are first evaluated with respect to x, after which c is substituted for x.

Consider the case where x is a random variable and we wish to determine the expected value of the function f averaged over all x. Generally speaking, the mean value of a function is only equal to the function evaluated at the mean of x in the special cases in which the function is linear in x or x is a constant. Hence, we cannot just directly substitute the sample mean when trying to evaluate the mean of some

function of the data. However, we can get around this problem by expanding f about the mean of x, using Equation 1 with $c = \mu_x$, and then take the expectation,

$$\begin{aligned} E(f) &= E\big[f(\mu_x) + (x-\mu_x)\frac{\partial f(\mu_x)}{\partial x} + (x-\mu_x)^2\frac{\partial^2 f(\mu_x)}{2\partial x^2} + ...\big] \\ &= f(\mu_x) + E(x-\mu_x)\frac{\partial f(\mu_x)}{\partial x} + E(x-\mu_x)^2\frac{\partial^2 f(\mu_x)}{2\partial x^2} + ... \end{aligned} \tag{A.2}$$

By the definition of a mean, $E(x-\mu_x) = 0$, and $E(x-\mu_x)^2$ is the expected variance of x, $\sigma^2(x)$. Thus, ignoring third and higher-order terms,

$$E(f) \simeq f(\mu_x) + \sigma^2(x)\frac{\partial^2 f(\mu_x)}{2\partial x^2} \tag{A.3}$$

A.1.1 Expectation and Variance of Complex Variables

The same approach can be used to derive expressions for the expectations of functions that depend on more than a single variable. In this case, f must be expanded around the means of each of component variables. With two componenet variables, for example, an expansion around μ_x and μ_y leads to

$$E(f) = f(\mu_x, \mu_y) + \sigma^2(x)\frac{\partial^2 f(\mu_x, \mu_y)}{2\partial x^2} + \sigma(x, y)\frac{\partial^2 f(\mu_x, \mu_y)}{\partial x \partial y} + \sigma^2(y)\frac{\partial^2 f(\mu_x, \mu_y)}{2\partial y^2} + ... \quad \text{(A.4)}$$

The similar approach can be used to obtain an expression for the variance of a function. Again expanding around $c = \mu_x$, and substituting for f from Equation 7.1

$$\begin{aligned}
\sigma_f^2 &= E\{[f - E(f)]^2\} \\
&= E\left\{\left[\left(f(\mu_x) + (x - \mu_x)\frac{\partial f(\mu_x)}{\partial x} + ...\right) - \left(f(\mu_x) + \sigma^2(x)\frac{\partial^2 f(\mu_x)}{2\partial x^2} + ...\right)\right]^2\right\} \\
&= E\left\{\left[(x - \mu_x)\frac{\partial f(\mu_x)}{\partial x} + [(x - \mu_x)^2 - \sigma^2(x)]\frac{\partial^2 f(\mu_x)}{2\partial x^2} + ...\right]^2\right\} \quad \text{(A.5)}
\end{aligned}$$

Ignoring all but the two lowest-order terms, and noting that $E(x - \mu_x) = 0$,

$$\begin{aligned}\sigma_f^2 \simeq\ & E[(x-\mu_x)^2]\Big[\frac{\partial f(\mu_x)}{\partial x}\Big]^2 + 2E\big[(x-\mu_x)^3\big]\Big[\frac{\partial f(\mu_x)}{\partial x}\Big]\Big[\frac{\partial^2 f(\mu_x)}{2\partial x^2}\Big] \\ & - 2E(x-\mu_x)\sigma^2(x)\Big[\frac{\partial f(\mu_x)}{\partial x}\Big]\Big[\frac{\partial^2 f(\mu_x)}{2\partial x^2}\Big] + E\big[(x-\mu_x)^4\big]\Big[\frac{\partial^2 f(\mu_x)}{2\partial x^2}\Big]^2 \\ & - 2E\big[(x-\mu_x)^2\big]\sigma^2(x)\Big[\frac{\partial^2 f(\mu_x)}{2\partial x^2}\Big]^2 + \sigma^4(x)\Big[\frac{\partial^2 f(\mu_x)}{2\partial x^2}\Big]^2 \\ & = \sigma^2(x)\Big[\frac{\partial f(\mu_x)}{\partial x}\Big]^2 + 2\mu_{3x}\Big[\frac{\partial f(\mu_x)}{\partial x}\Big]\Big[\frac{\partial^2 f(\mu_x)}{2\partial x^2}\Big] \\ & + \big[\mu_{4x} - \sigma^4(x)\big]\Big[\frac{\partial^2 f(\mu_x)}{2\partial x^2}\Big]^2 \end{aligned} \tag{A.6}$$

where $\mu_{3x} = E[(x-\mu_x)^3]$ and $\mu_{4x} = E[(x-\mu_x)^4]$ are the third and fourth moments about the mean of x.

When f is a function of two variables, an approximation often used in place of Equation 7.6 is obtained by ignoring all but the first-order terms. Then, if f is a function of n variables,

$$\sigma_f^2 \simeq \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma(x_i,x_j)\big(\frac{\partial f}{\partial x_i}\big)\big(\frac{\partial f}{\partial x_j}\big) \tag{A.7}$$

where $\sigma(x_i,x_j)$ is a variance when $i=j$ and a covariance otherwise, and the partial derivatives are evaluated at the expectations for all underlying variables.

A.1.2 Expectations and Variances of Ratios

Letting $f = u/v$, and $\partial f/\partial u = v^{-1}$, $\partial f/\partial v = -u/v^2$, giving $\partial^2 f/\partial u^2 = 0$, $\partial^2 f/\partial v^2 = 2u/v^3$, and $\partial^2 f/\partial u \partial v = \partial^2 f/\partial v \partial u = -1/v^2$. Evaluating these second-order partials at μ_u and μ_v (the expected values of u and v), Euqation 7.4 gives

$$E(\frac{u}{v}) \simeq \frac{\mu_u}{\mu_v}\Big(1 + \frac{\sigma^2(v)}{\mu_v^2} - \frac{\sigma(u,v)}{\mu_u \mu_v}\Big) \tag{A.8}$$

Likewise, from Equation 7.7,

$$\begin{aligned}\sigma^2(u/v) &\simeq \sigma^2(u)\Big(\frac{1}{\mu_v}\Big)^2 + \sigma^2(v)\Big(\frac{\mu_u}{\mu_v^2}\Big)^2 - 2\sigma(u,v)\Big(\frac{1}{\mu_v}\Big)\Big(-\frac{\mu_u}{\mu_v^2}\Big) \\ &= \Big(\frac{\mu_u}{\mu_v}\Big)^2\Big(\frac{\sigma^2(u)}{\mu_u^2} - \frac{2\sigma(u,v)}{\mu_u\mu_v} + \frac{\sigma^2(v)}{\mu_v^2}\Big)\end{aligned} \tag{A.9}$$

Both Equation 7.8 and 7.9 are approximations since $\partial^3 f/\partial v^3 \neq 0$.

A.2 Sample Mean and Variances of Regression Coefficients

The least-squares regression coefficient is given by $b = u/v$, where $u = \mathrm{Cov}(x, y)$ and $v = \mathrm{Var}(x)$. Since $\mathrm{Var}(x)$ and $\mathrm{Cov}(x, y)$ are unbiased estimators of the variance and covariance, $\mu_u = \sigma(x, y)$ and $\mu_v = \sigma^2(x)$. Under the assumption that x and y are bivariate normally distributed, we can also obtain the variances and covariance of u and v:

$$\sigma^2(u) = \frac{\sigma^2(x)\sigma^2(y) + [\sigma(x,y)]^2}{n} \tag{A.10}$$

$$\sigma^2(v) = \frac{2\sigma^4(x)}{n} \tag{A.11}$$

$$\sigma(u,v) = \frac{2\sigma^2(x)\sigma(x,y)}{n} \tag{A.12}$$

$$\tag{A.13}$$

Expression for the variances and covariances of moments about the origin and expression for the variances and covariances of other bivariate moment can be found in Stuart et al. (1963).

Subsituting these equations into Equation 7.8, we get

$$\begin{aligned} E(b) = E(\frac{u}{v}) &\simeq \frac{\mu_u}{\mu_v}\Big(1 + \frac{\sigma^2(v)}{\mu_v^2} - \frac{\sigma(u,v)}{\mu_u\mu_v}\Big) \\ &= \frac{\sigma(x,y)}{\sigma^2(x)}\Big[1 + \frac{2\sigma^4(x)/n}{(\sigma^2(x))^2} - \frac{2\sigma^2(x)\sigma(x,y)/n}{\sigma(x,y)\sigma^2(x)}\Big] \\ &= \frac{\sigma(x,y)}{\sigma^2(x)} \end{aligned} \tag{A.14}$$

In the similar way, substituting these equations into Equation 7.9, we obtain,

$$\begin{aligned} \sigma^2(b) = \sigma^2(\frac{u}{v}) &\simeq \Big(\frac{\mu_u}{\mu_v}\Big)^2\Big(\frac{\sigma^2(u)}{\mu_u^2} - \frac{2\sigma(u,v)}{\mu_u\mu_v} + \frac{\sigma^2(v)}{\mu_v^2}\Big) \\ &= \frac{\sigma^2(y)(1-\rho^2)}{n\sigma^2(x)} \end{aligned} \tag{A.15}$$

where $\rho = \sigma(x, y)/[\sigma(x)\sigma(y)]$ is the correlation coefficient.

Bibliography

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*, 2014.

Tatiana Benaglia, Didier Chauveau, David Hunter, and Derek Young. mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software*, 32 (6):1–29, 2009.

Peter J Bickel and Kjell Doksum. Mathematical statistics: Basic concepts and selected ideas, vol. i, 2006.

George EP Box and George C Tiao. Bayesian inference in statistical analysis. Technical report, Wisconsin University Madison Dept of Statistics, 1973.

Katherina B Brokordt, Federico M Winkler, William J Farías, Roxana C González, Fabio Castaño, Philippe Fullsack, and Christophe M Herbinger. Changes of heritability and genetic correlations in production traits over time in red abalone (h aliotis rufescens) under culture. *Aquaculture Research*, 46(9):2248–2259, 2015.

John FY Brookfield. Evolution and evolvability: celebrating darwin 200. *Biology Letters*, 5(1):44–46, 2009.

Jiahua Chen and Abbas Khalili. Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103(484): 1674–1683, 2008.

C Clark Cockerham. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39 (6):859, 1954.

John Kenneth Colbourne, Bryan D Neff, Jonathan M Wright, and Mart R Gross. Dna fingerprinting of bluegill sunfish (lepomis macrochirus) using (gt) n microsatellites and its potential for assessment of mating success. *Canadian Journal of Fisheries and Aquatic Sciences*, 53(2):342–349, 1996.

Charles William Cotterman. *A calculus for statistico-genetics*. PhD thesis, The Ohio State University, 1940.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

DS Falconer and TF Mackay. C. 1996. *Introduction to Quantitative Genetics*, 4, 1996.

Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

Ronald Aylmer Fisher et al. Statistical methods for research workers. *Statistical Methods for Research Workers*, (5th Ed), 1934.

Roberto A Flores-Aguilar, Alfonso Gutierrez, Andres Ellwanger, and Ricardo Searcy-Bernal. Development and current status of abalone aquaculture in chile. *Journal of Shellfish Research*, 26(3):705–711, 2007.

Jean Louis Foulley. A simple argument showing how to derive restricted maximum likelihood. *Journal of Dairy Science*, 76(8):2320–2324, 1993.

Richard Frankham, Scientist Emeritus Jonathan D Ballou, David A Briscoe, and Jonathan D Ballou. *Introduction to conservation genetics*. Cambridge university press, 2002.

Dany Garant, Loeske EB Kruuk, Robin H McCleery, and Ben C Sheldon. Evolution in a changing environment: a case study with great tit fledging mass. *The American Naturalist*, 164(5):E115–E129, 2004.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

Michel Gillois. Relation d'identité en génétique i.-postulats et axiomes mendéliens. In *Annales de l'IHP Probabilités et statistiques*, volume 2, pages 1–94, 1965.

Deborah J Goff, Katherine Galvin, Hillary Katz, Monte Westerfield, Eric S Lander, and Clifford J Tabin. Identification of polymorphic simple sequence repeats in the genome of the zebrafish. *Genomics*, 14(1):200–202, 1992.

H Roy Gordon and Peter A Cook. World abalone fisheries and aquaculture update: supply and market dynamics. *Journal of Shellfish Research*, 23(4):935–940, 2004.

Alan Grafen. A geometric view of relatedness. *Oxford Surveys in Evolutionary Biology*, 2(2):28–89, 1985.

Alexander Graham. *Kronecker products and matrix calculus with applications*. Courier Dover Publications, 2018.

JM Hammersley, DC Handscomb, and George Weiss. Monte carlo methods. *Physics Today*, 18(2):55, 1965.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Christophe M Herbinger, Roger W Doyle, Elizabeth R Pitman, Danielle Paquet, Kate A Mesa, Dianne B Morris, Jonathan M Wright, and Douglas Cook. Dna fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. *Aquaculture*, 137(1-4):245–256, 1995.

CM Herbinger, RW Doyle, CT Taggart, SE Lochmann, Al L Brooker, Jonathan M Wright, and D Cook. Family relationships and effective population size in a natural cohort of atlantic cod (gadus morhua) larvae. *Canadian Journal of Fisheries and Aquatic Sciences*, 54(S1):11–18, 1997.

George K Holmes. Measures of distribution. *Publications of the American Statistical Association*, 3(18-19):141–157, 1892.

Albert Jacquard. The genetic structure of populations. 1974.

DL Johnson and Robin Thompson. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of Dairy Science*, 78(2):449–456, 1995.

Owen R Jones and Jinliang Wang. Colony: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10(3):551–555, 2010.

G Karigl. A recursive algorithm for the calculation of identity coefficients. *Annals of Human Genetics*, 45(3):299–305, 1981.

Hiroyuki Kasahara and Katsumi Shimotsu. Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110(512):1632–1645, 2015.

Oscar Kempthorne. The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London. Series B-Biological Sciences*, 143(910): 103–113, 1954.

Oscar Kempthorne. An introduction to genetic statistics. 1957.

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

Russell Lande. A quantitative genetic theory of life history evolution. *Ecology*, 63(3): 607–615, 1982.

Kenneth Lange. *Mathematical and statistical methods for genetic analysis*. Springer Science & Business Media, 2003.

JL Lush. Animal breeding plans. *Animal breeding plans.*, 1937.

Michael Lynch and Kermit Ritland. Estimation of pairwise relatedness with molecular markers. *Genetics*, 152(4):1753–1766, 1999.

Michael Lynch and Bruce Walsh. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

Stewart K McConnell, Patrick O'Reilly, Lorraine Hamilton, Jonathan M Wright, and Paul Bentzen. Polymorphic microsatellite loci from atlantic salmon (salmo salar): genetic differentiation of north american and european populations. *Canadian Journal of Fisheries and Aquatic Sciences*, 52(9):1863–1872, 1995.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Donald F Morrison, Lauriston C Marshall, and Harry L Sahlin. Multivariate statistical methods. 1976.

Daniel E Morse, Helen Duncan, Neal Hooker, and Aileen Morse. Hydrogen peroxide induces spawning in mollusks, with activation of prostaglandin endoperoxide synthetase. *Science*, 196(4287):298–300, 1977.

Timothy A Mousseau, Derek A Roff, et al. Natural selection and the heritability of fitness components. *Heredity*, 59(Pt 2):181–197, 1987.

Timothy A Mousseau, Kermit Ritland, and Daniel D Heath. A novel method for estimating heritability using molecular markers. *Heredity*, 80(2):218–224, 1998.

Simon Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, pages 343–366, 1886.

Patrick T O'Reilly, Lorraine C Hamilton, Stewart K McConnell, and Jonathan M Wright. Rapid analysis of genetic variation in atlantic salmon (salmo salar) by pcr multiplexing of dinucleotide and tetranucleotide microsatellites. *Canadian Journal of Fisheries and Aquatic Sciences*, 53(10):2292–2298, 1996.

PT O'reilly, C Herbinger, and Jonathan M Wright. Analysis of parentage determination in atlantic salmon (salmo salar) using microsatellites. *Animal Genetics*, 29 (5):363–370, 1998.

Ian Painter. Sibship reconstruction without parental information. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 212–229, 1997.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Karl Pearson. Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. *Proceedings of the Royal Society of London*, 59:69–71, 1895.

Robert Plomin. *Nature and nurture: An introduction to human behavioral genetics.* Thomson Brooks/Cole Publishing Co, 1990.

George R Price et al. Selection and covariance. *Nature*, 227:520–521, 1970.

Adrian E Raftery and Steven Lewis. How many iterations in the gibbs sampler? Technical report, WASHINGTON UNIV SEATTLE DEPT OF STATISTICS, 1991.

Kermit Ritland. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, 67(02):175–185, 1996.

Christian P Robert and George Casella. Monte carlo statistical methods (springer texts in statistics), 2005.

George K Robinson et al. That blup is a good thing: the estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

Shayle R Searle and Andre I Khuri. *Matrix algebra useful for statistics*. John Wiley & Sons, 2017.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

SR Searle et al. Matrix algebra for the biological sciences (including applications in statistics). *Matrix Algebra for the Biological Sciences (including applications in statistics).*, 1966.

Bruce R Smith, Christophe M Herbinger, and Heather R Merry. Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, 158(3):1329–1338, 2001.

David Spiegelhalter, Andrew Thomas, Nicky Best, and Wally Gilks. Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*, pages 1–59, 1996.

Alan Stuart, Maurice G Kendall, et al. *The advanced theory of statistics*. Griffin, 1963.

Stuart C Thomas and William G Hill. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, 155(4):1961–1972, 2000.

GB Trustrum and JH Williamson. The correlations between relatives in a random mating diploid population. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 57, pages 315–320. Cambridge University Press, 1961.

MT Viana. Abalone aquaculture, an overview. *WORLD AQUACULTURE-BATON ROUGE-*, 33(1):34–39, 2002.

Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 2008.

Jinliang Wang. Sibship reconstruction from genetic data with typing errors. *Genetics*, 166(4):1963–1979, 2004.

Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338, 1922.