Functional Quantitative Genetics and the Missing Heritability Problem

Serge Sverdlov

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Elizabeth A. Thompson, Chair

Joseph Felsenstein

Bruce S. Weir

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Functional Quantitative Genetics
and the Missing Heritability Problem

Serge Sverdlov

Chair of the Supervisory Committee:
Professor Elizabeth A. Thompson
Department of Statistics

In classical quantitative genetics, the correlation between the phenotypes of individuals with unknown genotypes and a known pedigree relationship is expressed in terms of probabilities of IBD states. In existing models of the inverse problem where genotypes are observed but pedigree relationships are not, probabilities and correlations have either a Bayesian or a hybrid interpretation.

We introduce the IBF (Identity by Function) model based on the classic infinite allele mutation process. Describing genetic resemblance in terms of functional states defines a genetic architecture for a trait without reference to specific alleles or a population, treating a gene-scale functional region, rather than a SNP, as a QTL, and emphasizing locus weights and patterns of dominance over multiple alleles. This allows the reconciliation of bottom-up (genome sequence based) and (pedigree/population) calculations of heritability, as well as phenotype and gene effect prediction. We perform these calculations with simulated, pig, and human traits. For related computational problems, we describe an algorithm for the estimation in large scale variance components problems by matrix decomposition methods techniques related to the Sparse Bayesian Learning/Relevance Vector Machine framework. Additionally, we describe a combinatorial framework for decomposing nonlinear genetic effects due to dominance and epistasis, and a method for adjusting the Genomic Relationship Matrix for linkage disequilibrium by SNP selection and weighting.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ARD: Automatic Relevance Determination

BLUE: Best Linear Unbiased Estimator (or Estimate)

BLUP: Best Linear Unbiased Predictor (or Prediction)

DAG: Directed Acyclic Graph

DCS: Directional Consistency of Substitution

EM: Expectation/Maximization [algorithm]

FGRM: Functional Genomic Relationship Matrix

GWAS: Genome Wide Association Study

GRM: Genomic Relationship Matrix

HUMVEE: Hierarchical Unbiased Minimum Variance Ensemble Estimator

HWE: Hardy-Weinberg Equilibrium

IBD: Identity by Descent

IBF: Identity by Function

IBS: Identity by State

LD: Linkage Disequilibrium

LDGRM: Linkage Disequilibrium [adjusted] Genomic Relationship Matrix

MAF: Minor Allele Frequency

ML: Maximum Likelihood

QP: Quadratic Program (or Programming)

QTL: Quantitative Trait Locus

REML: Residual Maximum Likelihood

RVM: Relevance Vector Machine

SBL: Sparse Bayesian Learning

SKAT: SNP-set (Sequence) Kernel Association Test

SNP: Single Nucleotide Polymorphism

SVD: Singular Value Decomposition

ACKNOWLEDGMENTS

I am grateful for the astounding effort and time Elizabeth Thompson devoted to guiding me in turning a tangle of ideas into a work of science. I'm also grateful for all the help and guidance of committee members Joseph Felsenstein, Vladimir Minin, and Bruce Weir, which is reflected in all chapters of this thesis.

Much of the variance components chapter benefited from the guidance of Marina Meila as part of the Computational Learning project class and follow-on work. The linkage disequilibrium chapter benefitted from Hoyt Koepke's ideas on optimization methods.

Access to, and help with interpreting the formats for, human genotype and phenotype data, as well as computing resources, were provided by Univerity of Washington's members of the GENEVA project, Genetics Coordinating Center, and the Biostatistics IT team. In particular I am grateful for the time spent by David Levine and Robert Moulton helping me with this project.

DEDICATION

To my parents, who are not as annoying as I say they are.

Chapter 1

INTRODUCTION

1.1 The Missing Heritability Problem

The field of statistical quantitative genetics comprises two overlapping bodies of methodology and theory:

1. Top-down approaches, deriving genetic resemblances between individuals based on pedigree and population relationships

2. Bottom-up approaches, based on identifying and combining the effects of individual genetic variants

The convergence between these two bodies of knowledge, and the need to reconcile them, is a key theoretical challenge of the present generation of quantitative genetic methodologies. The *Missing Heritability* problem, as highlighted by Maher (2008), is a primary manifestation of this challenge. For some traits, such as human height, top-down methods point to a high level of heritability, without identifying the individually responsible genetic variants. Bottom up methods do identify the individual genetic variants associated with the traits; but together such variants account for a small fraction of the heritability.

A number of explanations for the gap have been proposed since the problem has been formulated (e.g. Manolio et al., 2009). We split these explanations into three categories:

1. Consistent explanations within the paradigm of classical quantitative genetics.

   Under these explanations, the gap will be closed given unlimited sample size, or under ideal experimental conditions. A trait may have a high heritability architecture composed of a large number of variants, each with a small individual contribution to the heritability, and correspondingly hard to individually detect. This can be either

due to small effect size, or low frequency alleles, possibly singletons (i.e. infrequent enough to be encountered exactly once within the study population), with large effect size. Methods for extrapolating the heritability of variants that cannot be individually identified at a given sample size have been proposed (Park et al., 2010). Indirect approaches to the estimation of heritability without identifying individual SNP effects (Yang et al., 2010) can also be interpreted as following this approach.

2. Known approximations within the classical paradigm.

   Unlike Mendelian inheritance, which follows probabilistic laws that are as exact as any in biology, population phenomena are subject to stochastic modeling based on approximate assumptions. The exact treatment of multiple linked genes, and higher order epistatic interactions between genes, is often infeasible due to a combinatorial increase in the number of parameters. Systematic discrepancies due to acknowledged approximations to known biological phenomenona made in the classical framework include effects of gene-environment interactions, population structure, linkage and linkage disequilibrium, and epistasis, including epistasis in the presence of linkage and linkage disequilibrium. A prominent critique based on higher order epistatic components is that of Zuk et al. (2012).

3. Explanations outside the paradigm.

   These involve biological mechanisms that are not modeled in the classical framework, epigenetic sources of resemblance between relatives (Slatkin, 2009; Tal et al., 2010), copy number variations, and a role for mutation on a within-pedigree time scale. Resemblance due to alternative biological phenomena may still be captured in classical models. A closer pair of relatives may tend to be phenotypically similar for a variety of reasons, properly accounted for as due to environment, population structure, or epigenetics, a confounding effect encountered in multiple branches of genetic analysis. An additive genetic model fit to such a trait will, to some extent, capture the relationship between genotypic similarity and phenotypic similarity, and allocate the effect to individual genes. Though the allocation to individual genes will have no biological

interpretation, the model may be effective at predicting phenotype prediction from genotype, regardless of the underlying mechanism.

1.2 Heritability

1.2.1 Problem Context

Our primary concern is with complex polygenic quantitative traits. For our purposes, a trait is complex if it takes on multiple states depending on multiple genetic and environmental influences. Polygenic, as opposed to Mendelian or oligogenic, refers to the assumption that the trait architecture is made up of many genes with small individual influence, not excluding the possibility that some *major genes* have individually meaningful influence. Under this assumption there are sufficiently many genes that the genetic contribution to the trait takes on a continuum of values, and we should not expect to identify the individual contributions of genes at the smallest end of the effect size spectrum The word *gene*, or *genetic variant*, is used here loosely, and we will devote significant attention later on to the question of what constitutes the appropriate indivisible unit of genetic analysis, or *quantitative trait locus (QTL)*. A trait is quantitative if it is measured as a real value; as opposed to a discrete trait such as the presence or absence of a disease. For mathematical convenience, whenever necessary we will assume a scale such that any positive or negative trait value is realizable for any individual, though values outside some typical range will be deemed improbable by a reasonable model.

1.2.2 What is Heritability?

The textbook definitions of heritability (e.g. Lynch and Walsh, 1998; Falconer and MacKay, 1996) of trait or phenotype P distinguish between narrow sense heritability

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \tag{1.1}$$

and broad sense heritability

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} \tag{1.2}$$

where σ^2_P, the variance of the phenotype in a chosen reference population, is experimentally measurable. The variance explainable by genetic factors or causes, σ^2_G, can be defined in an experimentally measurable way in presence of monozygotic twins or clones as the (reference population) covariance between a pair of genetically identical individuals. The definition of σ^2_A, the variance explainable by additive genetic factors, is more subtle. There are two paths to defining σ^2_A:

1. Top-down definition, as a scaled coefficient from a parent-offspring regression, or some other pedigree design

2. Bottom-up definition, by contributions from individual loci, as derived in appendix A

For example Zuk et al. (2012) propose splitting the definition of narrow-sense heritability h^2 into 'true' heritability h^2_{all}, composed by adding heritability contributions of individual loci, and 'apparent' heritability h^2_{pop}, the appropriately scaled parent-offspring correlation. Correspondingly Zuk et al. (2012) make a distinction between *missing* and *phantom* heritability. Specifically they emphasize a point made by Falconer and MacKay (1996), that applications of classical theory to narrow sense heritability estimation commonly ignore epistatic terms, assuming that $\sigma^2_A \gg \sigma^2_{AA}$. Under the decomposition of the genetic variance into higher order variance components (Cockerham, 1954; Kempthorne, 1954), the correlation between (non-fraternal) relatives includes rapidly vanishing higher order additive epistatic terms $\sigma^2_{AA}, \sigma^2_{AAA}, \ldots$. However, if the epistatic terms are collectively large, the heritability, as estimated by a model without epistatic terms, will vary by different levels of pedigree relationships. Thus, a method based on detecting identity by descent among distant relatives will yield a dramatically different (and lower) heritability estimate, compared with traditional pedigree methods necessarily limited to a manageable number of generations. Population, or apparent, heritability, h^2_{pop}, is not a true heritability, according to this argument, as it necessarily depends on the pedigree structure of the underlying measurement. It is defined above in terms of the parent-offspring correlation, but it could as legitimately have been defined in terms of, e.g., correlation between second cousins, and would then refer to a biologically different quantity.

From our perspective, the 'true' heritability h^2_{all} is at least as problematic. To compute the heritability contributions of individual loci we must unambiguously define the loci, and allele frequencies at each locus, as discussed in appendix A. Not being independent of these model choices, the 'true' heritability defined in this way does not refer to a quantity measurable in an ideal experiment. The only definitions of heritability that meet this standard are the pedigree-based measurements, which are defined up to the choice of the pedigree. Thus we may choose twins, parent-offspring pairs, first cousins, etc. as a canonical experiment for the definition of heritability; that heritability measured by means of other pedigree structures is the same value is a falsifiable, model dependent hypothesis.

Makowsky et al. (2011) discusses the distinction between predictive R^2 and heritability. Since narrow sense heritability is the fraction of the variance attributable to additive contributions from genes, it can be interpreted as the maximum R^2 of any phenotype predictor linear in allele dosages. Likewise, broad sense heritability is the maximum R^2 of any predictor. The trick with applying such a definition of heritability is the choice of reference population. If, as in some of the measurements performed by Makowsky et al. (2011), the sample on which R^2 is evaluated contains pedigree relatives, or population structure, the prediction R^2 need not relate to the heritability of a randomly mating population. For example, consider the R^2 of prediction in a structured population with two highly phenotypically distinct families. A high R^2 predictor must only distinguish between the families on the basis of genotype, and need not make accurate inferences about relationships within each family. As the families need not have been randomly selected, the high R^2 reveals little about the heritability of the underlying trait.

1.2.3 Heritability of Human Height

Human height is one of the most widely studied quantitative traits (Visscher et al., 2010). Galton (1886) is credited with the origin of the term *regression*, in the context of regression toward the mean, in the context of the study of the relationship of between the heights of children and parents. Height is thought to be highly heritable, influenced by a large number of genes, and thus the canonical example of a complex polygenic trait (e.g. Lango Allen et al.,

2010).

Three ways of computing heritability for height do not appear to reconcile:

1. Bottom up computation, based on effect sizes of individually significant SNPs (Lango Allen et al. 2010), $h^2 \approx 10\%$.

2. Distant relatives using inferred relationship matrices (Yang et al. 2010), $h^2 \approx 45\%$.

3. Twins or close relatives (Visscher et al. 2007), $h^2 \approx 80\%$.

We can exhaustively classify the potential explanations:

1. Non-Answers (i.e. the experiments were wrong)

   Such explanations would include: Insufficient sample size, random sampling error, differences in study populations or environmental conditions, differences in the trait being measured or procedures for adjusting for age or sex.

2. Approximations made by the experimental designs relative to the Classical Model

   This includes the neglect of the higher order additive epistatic terms, neglect of alleles of insufficient effect size to pass the significance threshold (which could in principle could be overwhelmed by an impractically large sample size), lack of complete linkage disequilibrium between markers and QTLs, or markers within QTLs; multiple alleles, unaccounted for by the bottom up computation's biallelic model; or the presence of environmental effects.

3. Assumptions made by the Classical Model

   The twin concerns of linkage (non-independent assortment) and linkage disequilibrium, and violations of Hardy-Weinberg (and more broadly, gametic phase) equilibrium beyond the inbreeding adjustment, are both assumed away for tractability within the classical framework. Likewise, several population phenomena enter the classical framework as statistical adjustments and not fundamental concepts; primarily, excess relatedness, population structure, and admixture.

4. Ambiguities within the Classical Model

   We will deal with the key question of what constitutes a locus and allele at multiple points within this dissertation.

5. Well-known biological phenomena outside Classical Model

   Non-Autosomal inheritance is not fully represented in the classical model, especially when comparing sex-adjusted height in a mixed sex population.

6. Not-fully-understood biological phenomena outside Classical Model

   Here we classify epigenetic inheritance, copy number variants, and maternal effects for sibling studies.

1.3 Outline

This dissertation is an effort to construct a model compatible with both the top-down and bottom-up approaches. This requires a careful re-examination of the sources of probabilistic uncertainty underlying existing models. The classical framework, built at a time when individual genes could not be observed, deals with uncertainty about the identity of genes as they are propagated through pedigrees by Mendelian segregation. Modern data, with the rapidly approaching availability of full DNA sequences for massive numbers of study subjects, remove uncertainty about individuals' genotypes, but not about the effects of specific alleles. Classical methods cannot be directly applied to fully sequenced genotypes; our contribution, in part, is to clarify the probabilistic interpretations of existing models that can be so applied, and to construct a model with a probabilistic interpretation consistent with both the classical framework and the availability of complete genotypes.

As we attempt to formulate a unified framework, we will identify the sources of uncertainty in the estimation of heritability, and argue that properly accounted for, this uncertainty is large enough to explain the missing heritability gap.

The framework we develop explicitly accounts for the phenomenon of dominance, and we develop extensions to examine the impact of, and possibility of adjusting for, dominance,

epistasis, and linkage disequilibrium.

In Chapter 2 (*Theory*), largely based on previously published work (Sverdlov and Thompson, 2013), we develop the Identity by Function framework. Here, we contrast the probabilistic interpretations of the classical and modern approaches to quantiatitive genetics, and develop our own approach that is a reconciliation between them.

The use of this framework to make inferences with genetic data is covered in Chapters 3 (*Methodology*) and 4 (*Variance Component Inference*). In classical quantitative genetics a distinction is made between fixed effect estimation and random effect prediction given variance components, and variance component inference. Characteristically, a prominent textbook separates its treatment of general matrix-based methods into *Estimation of Breeding Values* (Lynch and Walsh, 1998, Chapter 26) and *Variance-component Estimation with Complex Pedigrees* (Lynch and Walsh, 1998, Chapter 27). Likewise, we split the applications into methodology for using IBF parameters for individual and gene level inferences, and methodology for inferring IBF parameters themselves.

We present results of analyses of simulated and real data in Chapter 5. The simulation illustrates the methods presented in Chapter 3 (*Methodology*). The methods of Chapters 3 (*Methodology*) and 4 (*Variance Component Inference*) are applied to a pig dataset with multiple traits and relatively sparse SNP genotypes, and a human height dataset with imputed dense SNP data to which we apply genome partitioning.

We then present two extensions: a combinatorial approach to dominance, epistasis, and a framework for adjusting existing genomic relationship matrix methods for linkage disequilibrium. The conclusion chapter enumerates the contributions of the thesis, and discusses potential future directions for the IBF framework and related subjects in quantitative genetics.

Chapter 2

THEORY: IDENTITY BY FUNCTION

2.1 Introduction

2.1.1 Correlation Between Pedigree and Genotype Relatives

Quantitative genetic models can be described as explaining phenotypic resemblance between individuals on the basis of genetic resemblance, that is, the expected or actual sharing of genes. As we transition from pedigree information to full-sequence genetic information, we change the way we pose a typical question in quantitative genetics, inviting answers with conceptually incompatible probabilistic interpretations:

1. What is the correlation between the heights of an uncle and a nephew?

2. What is the correlation between the heights of individual A_1 with genotype $\mathbf{g}_1$ and individual A_2 with genotype $\mathbf{g}_2$?

The first question has a well-studied interpretation. We can sample n nominally unrelated uncle-nephew pairs from the population and compute their pairwise correlation. The second question only becomes meaningful in the context of models where the set of genotypic effects, or the function mapping genotype to expected phenotype, is treated as random. The meaning of the correlation is not reducible to an ideal sample from the population, and is model dependent. The two major model classes have two distinct interpretations: the purely Bayesian, reflected in the genomic selection (Meuwissen et al., 2001) and reproducing kernel Hilbert spaces (Gianola and van Kaam, 2008) frameworks, and an interpretation based on the inference of the parameters of classical theory from genomic data (Visscher, 2009).

2.1.2 The Three Model Classes

We can thus identify three distinct classes of quantitative genetic models:

1. Classical theory answers the first question. It predicts the expected fraction of genome shared between two individuals under the principle of Identity by Descent (IBD), based on their pedigree relationship, and maps this measure of genetic similarity to a phenotypic correlation.

2. Several classes of Bayesian, or random effect, models answer the second question on the principle of Identity by State (IBS). The correlation is a function of the similarity between the actual genotypes $\mathbf{g}_1$ and $\mathbf{g}_2$, without reference to pedigrees and descent.

3. Hybrid models answer the second question by inferring relatedness parameters between A_1 and A_2 from genotype data, and substituting these into the equations of classical theory. This admits neither a pure classical, nor a pure Bayesian interpretation.

2.1.3 Our Alternative Model

Our aim is to construct a model within which the second question can be meaningfully asked with a probabilistic interpretation in the non-Bayesian context of a population genetic process based on the classical infinite allele model. The model is generative in the sense of describing a joint distribution of genotypes and effects, given parameters describing a population process and trait architecture.

We adapt the genetic architecture from Fisher (1918), additive across loci but not within a locus, and treat alleles symmetrically, from a neutralist perspective, as draws from an infinite allele process. It is the need to reconcile this symmetry with diploidy and dominance that makes the model mathematically nontrivial. The result is a definition of correlation between two individuals' phenotypes given their genotypes (question 2) with respect to a population process. The thought experiment with respect to which we define our correlation is a replay of evolution. Classically, correlation is defined with respect to a repetition of the same pedigree relationship. For our model, it is the draw of functionally different alleles from the infinite allele process at the same mutation events that gave rise to our observed genotypes.

The model allows for infinitely many alleles at each locus. This is useful for full sequence

genomic data, where we routinely encounter novel alleles. The model's genetic architecture parameters specify the relative importance of loci and the pattern of dominance within a locus. These parameters are features of the mutation process that creates alleles, not of individual alleles. They are specific to a trait, but not to a particular population.

2.1.4 Quantitative Trait Locus Scale and Linkage Disequilibrium

Our model treats a functional region, such as a gene, as a quantitative trait locus (QTL). Each (non-synonymous) sequence in that region found in the population is an allele. This contrasts with treating each minimal length unit of genomic variation, such as a SNP or indel, as a QTL, as is the implicit practice in the GWAS or whole-exome (e.g. Kiezun et al., 2012) literature, and has analogies to region-based association models such as SKAT (Wu et al., 2011). In earlier QTL mapping methods based on sparse markers and linkage analysis (e.g. Lynch and Walsh, 1998, chapters 14–16), the QTL was treated as a point along the chromosome and did not need to be defined in sequence terms. Likewise, in the literature related to the genomic relatedness matrix (e.g. Yang et al., 2010), SNPs used to compute kinship are described as markers in linkage disequilibrium with point causal variants. For complete genome or dense SNP genotype data, for the human genome 1 million SNPs or approaching 50 SNPs per gene, the point abstraction runs into difficulties. Marker SNPs may be inside exons and causal, and multiple markers may occur within the same causal gene. The question of whether these multiple markers should be treated as separate QTLs cannot be avoided. In the GWAS context, it is typical for multiple SNPs in or near a gene to be associated with a trait. Methods extending GWAS approaches to polygenic traits (e.g. Guan and Stephens, 2011) treat this as a variable selection problem, as though only a single SNP per functional region is causal, and the other SNPs in LD with the causal SNP are a nuisance and a source of collinearity. This solution is practical, but restricts the biological effects that can be modeled at gene scale.

We leave outside the scope of this thesis the question of defining the boundaries of a biological functional region, and the complication of overlapping genes; we will use the terms gene and functional region interchangeably. From a gene-as-QTL perspective, every new

mutation is likely to create a novel gene-scale haplotype, implying an infinite allele population model. From a SNP-as-QTL perspective, every new mutation is likely to target a new SNP at a different base pair location, implying a diallelic, infinite sites model. In SNP-as-QTL, multiple nearby SNPs that occur together in non-random patterns to form functional sequences must be accounted for as linkage disequilibrium (LD), and their action should be expected to exhibit epistasis. The simplest consequence of switching from SNP-as-QTL to gene-as-QTL is reducing the importance of short-range (within-gene) LD and epistasis, introducing instead greater scope for dominance phenomena through the interaction of many alleles. Consider an extreme example, two SNPs within one codon: $[C/A]G[A/T]$. The ancestral sequence CGA codes for Arginine, as does either individual SNP mutation, CGT or AGA. However, the two SNPs together, AGT, code for Serine. If we treat the two SNPs as separate QTLs, we would expect both LD and epistasis in any trait affected by this sequence. If, alternatively, we treat the whole codon as a single QTL, there are simply four possible alleles with different frequencies and effects.

The ideal conditions of the classical (Cockerham, 1954; Kempthorne, 1954) decomposition of genetic covariance into variance component terms include an unlinked set of quantitative trait loci, linkage equilibrium, and the absence of mutation and selection. To reconcile quantitative genetic models with genomic data, we must define the QTL on a scale which best fits the classical approximation. Unlike the gene-as-QTL approach, the SNP-as-QTL approach conflicts with the ideal conditions by requiring either the violation of linkage equilibrium or the introduction of mutation or selection. When two SNP-scale loci are part of the same functional region, it is biologically plausible that they are in linkage disequilibrium. Linkage disequilibrium between linked, nearby loci decays under random mating due to recombination; but recombination *within* a functional region creates new functional haplotypes (mutation) with varying phenotypic effects (potential selection). In the gene-as-QTL approach, there may be linkage equilibrium or disequilibrium on a between-gene scale. Linkage disequilibrium on a within-gene scale, however, is not conceptually a part of the model. Instead, the diversity of gene haplotypes is captured by treating each haplotype as an allele.

2.2 Existing Quantitative Genetic Models

For all of the three existing model classes we consider, we can write the genotypic value as $G = \mathbf{g} \cdot \mathbf{f}$ for a suitably encoded genotype vector $\mathbf{g}$ and effect vector $\mathbf{f}$. For the commonly used additive model over diallelic SNPs, the genotype vector is a vector of minor allele dosages (0, 1, or 2) for each SNP locus, and the effect the additive contribution to the trait per allele. This simplest model is more restrictive than the Fisher (1918) model, since it excludes dominance and the possibility of more than two alleles. For a more general model, the genotype vector may contain a 0 or 1 as an indicator for the genotype containing an arbitrarily complex combination of alleles, with the most general possible model including a separate indicator column for each possible genotype.

2.2.1 Type 1: Fisher's Classical Polygenic Model and Identity by Descent

In classical quantitative genetics as originated by Fisher (1918) and presented in current form e.g. by Lynch and Walsh (1998), the non-environmental component of randomness is due to the unknown genotypes of the two individuals. Thus genotypes $\mathbf{g}$ are random, and the effects $\mathbf{f}$ are fixed.

The genotypes are modeled as being sampled from the population, with covariance between relatives due to the sharing of genes, as expressed by the probabilities of the identity by descent (IBD) states. If we were given genotypes of the two individuals in this setting, no source of randomness would remain, and the conditional covariance would be zero. This does not imply that given complete genotype information we know the genotypic values of the two individuals, but rather that we can express them in terms of (fixed) model parameters with no random variables; the number of such parameters in a general polygenic model may well be too large to even consider estimating.

2.2.2 Type 2: Bayesian, Random Effect, and Identity by State Models

In several classes of models, genotypes $\mathbf{g}$ are fixed, and the effects $\mathbf{f}$ are random.

The reproducing kernel Hilbert space framework described by Gianola and van Kaam (2008) is a general form for a broad class of such models, which may include dominance

and epistatic interaction terms. In this framework, the covariance between trait values is explicitly a function of the similarity between genotypes, expressed e.g. as a function of the norm of the difference between the SNP dosage vectors. The description given by de los Campos et al. (2009) argues for judging such covariance structures in terms of their predictive power.

This class of similarity functions is broad, but inherently applies to the measurement of similarities between diploid genomes. A separate class of models of genetic similarity exists for haploid genomes, motivated by specific applications. These include models of the divergence between genome sequences on an evolutionary time scale (e.g. Felsenstein, 2004, Chapter 11), and local distance measures used in sequence alignment applications.

In the class of genomic selection models (Meuwissen et al., 2001), a prior on each component of $\mathbf{f}$ is imposed. The prior may place a finite weight on 0, may be heavy tailed, and may be hierarchical, with hyperparameters shared by the components of $\mathbf{f}$. The family of analogous models with different priors and computational methods, recently reviewed by Kärkkäinen and Sillanpää (2012), has become known as the Bayesian alphabet (Gianola et al., 2009).

In such a model, the phenotypes of two individuals given their genotypes have a joint distribution, and therefore a covariance. The interpretation of this distribution is a Bayesian prior, our uncertainty about the effects.

2.2.3 Type 3: Hybrid and Realized Relationship Models

In these models, genotype data are used to infer relatedness parameters. Then, for the purposes of variance component and heritability estimation or prediction, we treat these as the true parameter values, i.e. probabilities of IBD gene sharing. Such models naturally emerged from a historical context in light of classical theory. "Population and quantitative genetics theory is built with parameters that describe relatedness, and the estimation of these parameters from genetic markers enables progress in fields as disparate as plant breeding, human disease gene mapping and forensic science." (Weir et al., 2006) With the availability of limited genetic marker data, it became possible to infer relatedness param-

eters from genotype, when complete pedigrees were unavailable (Thompson, 1975). As we have moved from sparse to dense marker maps, it was noticed that inferred relatedness parameters could outperform pedigree based ones in practice (Hayes et al., 2009), by capturing the actual Mendelian variability within classes of relatives (Hill et al., 2011).

The estimated fraction of genome shared by descent is inferred from marker data, either by identifying individual segments passed down without recombination (Browning and Browning, 2012) or from the realized relationship matrix (Hayes et al., 2009). This is a probabilistic inference based, crucially, not only on genotype data, but on population data, such as allele frequencies and assumptions about linkage disequilibrium.

A type 3 model can be reinterpreted either as a type 1 or type 2 model:

1. In a type 3 model, the covariance between two individuals' phenotypes given their genotypes is the classical covariance, given the inferred relatedness. Neglecting the difference between true and inferred relatedness parameters, or conditioning on the value of the relatedness parameters to be equal to the inferred values, a type 3 model can be interpreted as a a type 1 model.

2. The type 1 covariance is a function of the relatedness parameters, and the inferred relatedness parameters are, in turn, a function of genotype vectors. Substituting in the inferred relatedness parameters, we obtain covariances as a function of genotype vectors. This is, formally, a type 2 model.

2.2.4 Population Neutrality

A reference population is needed by classical (type 1) models to specify the probabilities of the random founder genotypes. In a model where genotypes are given and not random, a reference population is not needed. A property of an individual, or a biological property of a single allele in isolation, is *population neutral*. We define population neutrality as the formal invariance of a estimator to the properties of a population, such as allele frequencies and models of linkage disequilibrium. This characterizes the estimator, not its distribution. Its distribution in a population will vary with the distribution of alleles in that population.

The effect of substitution of specific allele A for specific allele B, in a model without epistasis, is population neutral. The average effect of substituting allele A for a random allele (drawn from a population) is not population neutral. The prediction of phenotype from genotype is population neutral; the prediction of a breeding value is not. Under the assumption of the additive model, classical additive effects (and therefore breeding values) do not algebraically depend on the allele frequencies, leading to breeding value estimation methods that do not explicitly refer to the reference population. In the general case, however, the breeding value may be defined as the best predictor of the offspring trait value when mated with a randomly chosen individual from the reference population; such a definition makes the population dependence explicit. Likewise, the heritability of a trait is always dependent on the reference population, and not population neutral.

In the Bayesian (type 2) context, it is difficult to motivate the idea that the prior for the biological effect of an allele would depend on population frequency of that allele. This is clear from a neutralist perspective. Even from a selectionist perspective, such a dependence would only arise from an explicit model of selective forces.

Type 2 models do not have a reference population, but type 3 models do. The type 3 methods explicitly model the fraction of genome shared IBD between the two individuals. Several definitions of IBD may be used:

1. Classically, IBD is defined with respect to a particular pedigree, or with respect to a founder generation (Thompson, 2013). By choosing extreme pedigrees, the complete population history of the species, or the pedigree containing no relationships, we find IBD fractions of 1 or 0, respectively, for any pair of chromosomes or individuals.

2. The inferred kinship is an estimate of the realized or actual fraction of IBD genome (Cockerham and Weir, 1983; Hill et al., 2011) whereas the classical kinship coefficient is the expectation of this fraction. Such identity must be defined with respect to pedigree founders, or an ad-hoc mechanism.

3. An alternative definition of kinship based on IBD described by Powell et al. (2010) uses the concept of the current reference population, and defines kinship coefficients

that may be negative, and thus cannot be interpreted classically as IBD probabilities.

Whichever form of IBD is assumed, we explicitly or implicily refer to a model of a reference population, and depend on at least such parameters as allele frequencies.

2.3 Functional Effect Distribution Model

2.3.1 Overview

By analogy with genome shared identical by descent (IBD) or by state (IBS) we introduce the concept of *Identity by Function (IBF)* and the *Functional Identity States*, S_0 through S_4. At a single locus, two individuals are functionally identical if they share a homozygous or heterozygous genotype, AA:AA (S_4) or AB:AB (S_3). They have partial functional identity if they share one identical allele, AA:AB (S_2) or AB:AC (S_1). In all other states (S_0), no alleles are shared between the two individuals. Intuitively we would expect more phenotypic resemblance, the more similarity there is between genotypes, i.e. the more loci are functionally identical.

In practice, we must define locus and identity in genomic sequence terms. A working definition is that a locus is the region surrounding an annotated gene including some fixed upstream and downstream flanking sequences. Identity, given phased full sequence genotypes, is sequence identity excluding known synonymous variation.

As illustrated in Figure 2.1, the five functional identity states are a different grouping of the 15 partitions of 4 chromosomes. In IBD theory, the 15 IBD states (Thompson, 1974) are partitioned into the 9 Jacquard coefficient (Jacquard, 1972) classes, of which two pairs are mirror images, making 7 classes. Of these, 3 classes do not involve allele sharing between the individuals, and are grouped into one covariance class for our purposes.

2.3.2 Single Locus, Infinite Allele Model

Consider the standard neutral infinite allele model (e.g., Kimura and Crow, 1964). At each mutation event, we draw a new allele from $\mathcal{A}$, the set of all possible alleles at a locus. Mutations are assumed to be irreversible. When we observe the allele states in the current

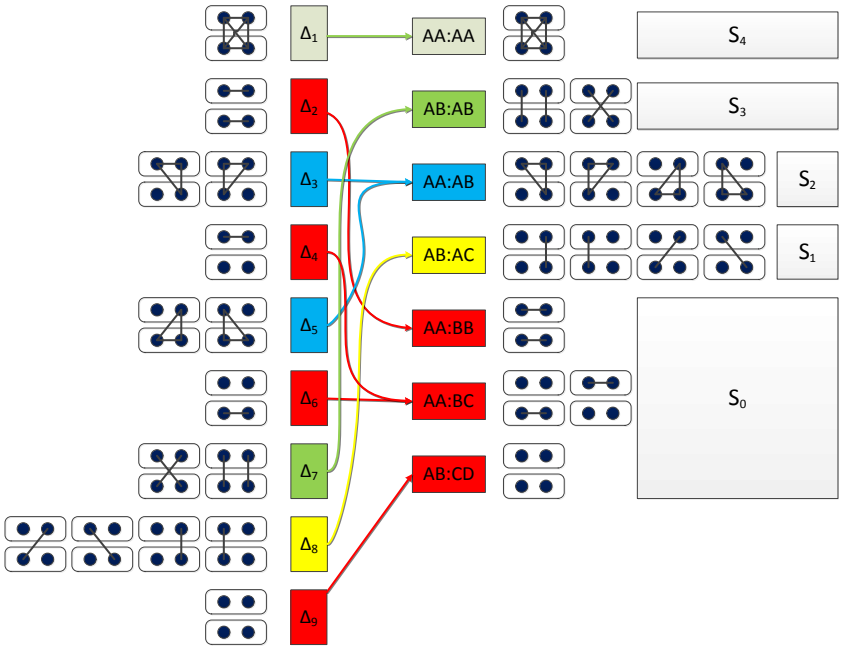


Figure 2.1: Functional Identity States. The 15 partitions of four chromosomes in two diploid individuals correspond to the 9 Jacquard coefficients (left) and map to the 5 Functional Identity States (right).

generation, we can distinguish identical and distinct alleles; but we do not know what mutation event produced a given allele, nor the relationship between allele and phenotype.

Each creation of a novel allele by a new mutation (or, more restrictively, by a non-synonymous mutation) represents an independent draw from the potential allele distribution. This model explicitly rejects the use of information about sequence similarity or mutation distance. There is no sense in which two alleles separated by a single point mutation are more similar than two alleles that, after many mutations, no longer resemble one another in sequence; two alleles are either identical in function, or different. Alleles are abstract objects, members of set $\mathcal{A}$, and do not correspond directly to real-valued quantities associated with phenotypes. Instead, we introduce a genotypic value mapping function, which turns the allele into a quantitative effect associated with a particular trait.

For a haploid organism with a trait controlled by a single locus, the genotypic trait value depends on a single allele. The genotypic value is then a function of the allele. If alleles are sampled independently from an arbitrary distribution, so are the corresponding genotypic values. We define the haploid genotypic value function $G(\cdot)$ as a mapping $\mathcal{A} \to \mathfrak{R}$. Then the real-valued random variables $G_i = G(A_i)$ are iid. Assuming all required moments exist, $\mathrm{E}[G_i] = \mu_G$, $\mathrm{Var}[G_i] = \sigma_G^2$, and $\mathrm{Cov}[G_i, G_j] = 0$. In this simple situation we can think of a genotypic value associated with each allele, and independently generated by each mutation event. The assumption that the moments exists excludes some models of interest, such as models where some effects are drawn from the Cauchy distribution.

2.3.3 Diploid Model

For a diploid organism with a trait controlled by a single locus, the genotypic trait value depends on two alleles, in interchangeable order. Define the diploid genotypic value function $G(\cdot, \cdot)$ as a mapping $(\mathcal{A}, \mathcal{A}) \to \mathfrak{R}$, subject to the symmetry constraint $G(A, B) = G(B, A)$. Likewise, define real-valued random variables $G_{ij} = G(A_i, A_j)$. The G_{ij} are no longer all independent; they are only independent in the case $G_{ij} \perp G_{kl}$ if $\{i, j\} \bigcap \{k, l\} = \emptyset$. Otherwise there is a number of patterns of dependence we can enumerate. By symmetry, the expectations and variances of all homozygotes and all heterozygotes are separately equal,

$\mathrm{E}[G_{ii}] = \mu_{AA}$ and $\mathrm{E}[G_{ij}] = \mu_{AB}$, and $\mathrm{Var}[G_{ii}] = \sigma^2_{AA}$ and $\mathrm{Var}[G_{ij}] = \sigma^2_{AB}$. We introduce two new correlation parameters, ρ_2 and ρ_3:

$$\mathrm{Cov}[G_{ii}, G_{ij}] = \rho_3 \sigma_{AA} \sigma_{AB} \tag{2.1}$$

$$\mathrm{Cov}[G_{ij}, G_{ik}] = \rho_2 \sigma^2_{AB} \tag{2.2}$$

$$\mathrm{Cov}[G_{ii}, G_{jk}] = \mathrm{Cov}[G_{ii}, G_{jj}] = \mathrm{Cov}[G_{ij}, G_{kl}] = 0 \tag{2.3}$$

The five covariances correspond to the five functional identity states, using the compact notation $AB = G(A, B)$, etc.:

S_0**:** No shared alleles:

$$C_0 = \mathrm{Cov}(AB, CD) = \mathrm{Cov}(AB, CC) = \mathrm{Cov}(AA, BB) = 0$$

S_1**:** Sharing 2 alleles: $C_1 = \mathrm{Cov}(AB, AC) = \rho_2 \sigma^2_{AB}$

S_2**:** Sharing 3 alleles: $C_2 = \mathrm{Cov}(AA, AB) = \rho_3 \sigma_{AB} \sigma_{AA}$

S_3**:** Full heterozygous functional identity: $C_3 = \mathrm{Cov}(AB, AB) = \sigma^2_{AB}$

S_4**:** Full homozygous functional identity: $C_4 = \mathrm{Cov}(AA, AA) = \sigma^2_{AA}$

Not every choice of parameters $\rho_3, \rho_2, \sigma_{AA}, \sigma_{AB}, \mu_{AA}, \mu_{AB}$ corresponds to a valid joint distribution for the effect vector $[G_{ii}; G_{ij}]$. Assuming second moments exist, the requirement is that the covariance matrix for the effects be positive semidefinite. We derive the positive semidefinite constraint

$$\rho_3^2 \leq \rho_2 \leq \frac{1}{2} \tag{2.4}$$

in appendix B. The constraint is necessary for the existence of an infinite allele process; that is, for the existence of a joint distribution for the effect vector for an arbitrarily large number k of alleles. It is also sufficient because the proof in appendix B gives a covariance matrix from which we can construct a multivariate normal distribution for the effect vector for k alleles.

A special case of the diploid model is the *additive model*. We may associate an additive effect with each allele, $C(\cdot) : \mathcal{A} \to \mathfrak{R}$, and define the genotypic value function $G(A_i, A_j) = C(A_i) + C(A_j)$. Then, by the derivation in appendix C,

$$\mu_{AA} = \mu_{AB} = 2\mu_C \tag{2.5}$$

$$\rho_2 = 1/2; \quad \rho_3 = 1/\sqrt{2} \tag{2.6}$$

$$\sigma_{AA} = 2\,\sigma_C; \quad \sigma_{AB} = \sqrt{2}\,\sigma_C; \quad \sigma_{AB}/\sigma_{AA} = 1/\sqrt{2} \tag{2.7}$$

2.3.4 Generative Models for Genotypic Values

Multivariate Normal For any choice of μ_{AA}, μ_{AB}, σ^2_{AA}, σ^2_{AB}, ρ_2, and ρ_3 satisfying the positive semidefinite constraint (2.4), there exists a multivariate normal distribution for the genotypic values. We summarize some of the properties of such distributions.

Using the covariance matrix expression in appendix B, we can derive the conditional distribution of a heterozygote's genotypic value, conditional on its two homozygotes. The conditional expectation can be decomposed into the mid-homozygote value, a heterozygosity bias, and an effect specific to ρ_3:

$$\begin{aligned}\mathrm{E}\left[G_{AB}|G_{AA} = g_{AA}, G_{BB} = g_{BB}\right] &= \frac{g_{AA} + g_{BB}}{2} + (\mu_{AB} - \mu_{AA}) \\ &\quad + \left(\frac{\sigma_{AB}}{\sigma_{AA}}\rho_3 - \frac{1}{2}\right)(g_{AA} - \mu_{AA} + g_{BB} - \mu_{AA}) \\ &= \mu_{AB} + 2\rho_3\frac{\sigma_{AB}}{\sigma_{AA}}\left(\frac{g_{AA} + g_{BB}}{2} - \mu_{AA}\right)\end{aligned} \tag{2.8}$$

$$\mathrm{Var}\left[G_{AB}|G_{AA} = g_{AA}, G_{BB} = g_{BB}\right] = \sigma^2_{AB}\left(1 - 2\rho_3^2\right) \tag{2.9}$$

For the purely additive model $AB = (AA + BB)/2$, when $\mu_{AB} = \mu_{AA}$ and when $\sigma_{AB}/\sigma_{AA} = \rho_3 = 1/\sqrt{2}$, the conditional variance for the heterozygote is 0, and the conditional mean is the simple midpoint between the two homozygotes.

Through the derivation in appendix D, assuming $\mu_{AB} = \mu_{AA}$ we can compute the probability of overdominance, that is, the probability that for two distinct alleles A and B, some ordering of genotypic values other than $AA < AB < BB$ or $AA > AB > BB$ prevails:

$$P_{OD} = \frac{2}{\pi}\tan^{-1}\sqrt{1 - 4\rho_3\frac{\sigma_{AB}}{\sigma_{AA}} + 2\frac{\sigma_{AB}^2}{\sigma_{AA}^2}} \tag{2.10}$$

$P_{OD} = 0$ in the additive case, and positive otherwise.

Effect and Affinity Model Multivariate normality implies the possibility of overdominance, and requires a marginal normal distribution. A second class of generative models compatible with this framework associates with each allele 1) its homozygous effect, and 2) random values unobservable in the homozygote, that govern how heterozygous effects are formed by merging two homozygous effects. This allows arbitrary marginal distributions to be imposed directly for homozygotes, and indirectly for heterozygotes, and the explicit construction of complete or partial dominance orders.

For each allele A_i, we construct two random variables, the homozygote value G_{ii} and the strength or affinity S_i. These are independent for different i, but G_{ii} and S_i need not be independent. We define the heterozygote value as a weighted blend of the homozygote values,

$$G_{ij} = G_{ii}\ w\left(S_i, S_j\right) + G_{jj}\ w\left(S_j, S_i\right) \tag{2.11}$$

There is no possibility of overdominance in this model unless w is allowed to exceed 1. A partial dominance model can be constructed by a "soft cutoff" weight function,

$$w_\sigma\left(A, B\right) = \frac{1}{1 + e^{\frac{B-A}{\sigma}}} \quad \text{or} \quad w_0\left(A, B\right) = \begin{cases} 1, & A > B \\ 1/2, & A = B \\ 0, & B < A \end{cases} \tag{2.12}$$

In the limit as $\sigma \to 0$, this becomes a hard cutoff corresponding to strict dominance.

The simplest model of dominance is dominance with respect to the trait itself; that is, $G_{ii} = S_i$, in which case a heterozygote has the genotypic value of the higher of its homozygotes. Another choice is the random dominance order, where S_i is a continuous random variable independent of G_{ii}. An intermediate model would let S_i be correlated with G_{ii} but not be deterministically related.

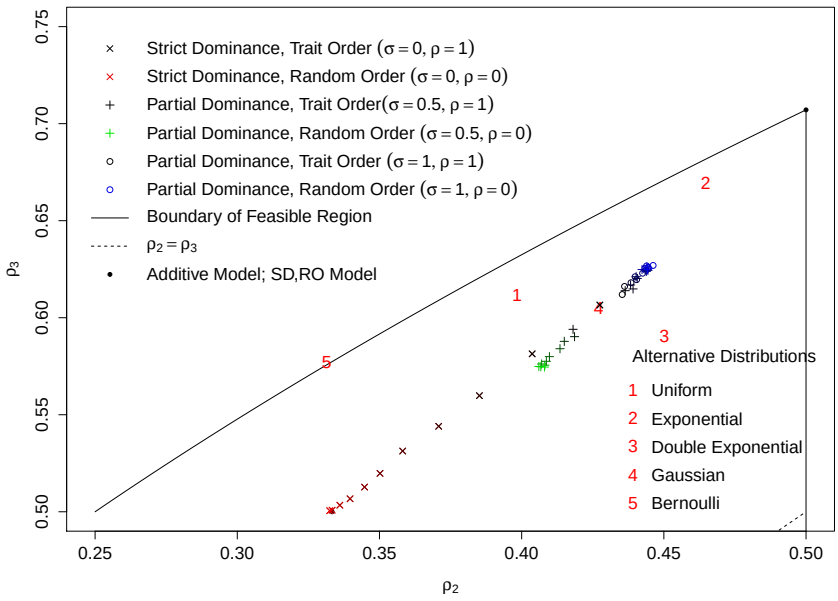


Figure 2.2: Feasible region $\rho_3^2 \leq \rho_2 \leq \frac{1}{2}$ and (ρ_2, ρ_3) for various simulated distributions and dominance models. Numbers correspond to the Strict Dominance, Trait Order model under different distributional assumptions, as described in the text. Symbols correspond to the Partial Dominance model for different values of σ. For each symbol, full color is Random Order ($\rho = 0$), and black is Trait Order ($\rho = 1$).

The random dominance order case has a closed form solution regardless of distribution. By the derivation in appendix E,

$$\mu_{AA} = \mu_{AB}; \quad \sigma_{AB} = \sigma_{AA} \tag{2.13}$$

$$\rho_2 = 1/3; \quad \rho_3 = 1/2 \tag{2.14}$$

Other cases do not yield closed form solutions, but we can compute ρ_2 and ρ_3 by simulation. In Figure 2.2, we simulate the strict dominance model with trait dominance order for several distributions. Location and scale are irrelevant for the calculation of these correlations; the distributions shown are the uniform, normal, exponential, the double exponential (Laplace), and the Bernoulli ($p = 0.5$). For the normal distribution, we simulate different values of the partial dominance σ generate S_i and G_{ii} as jointly normal with different choices of correlation ρ. The choice of $\sigma = 0$ and $\rho = 0$ is the random strict dominance order model, and increasing the parameter $\sigma \to \infty$ from strict to partial dominance attracts toward the additive model. Other parameter choices lie approximately on the line between those two models.

2.4 Multiple Loci

2.4.1 Multiple loci and covariance between two individuals

Suppose the genome consists of L loci. The trait's genotypic value is a sum of contributions from the L loci, each following an infinite allele process with its own μ_{AA}, μ_{AB}, σ^2_{AA}, σ^2_{AB}, ρ_2, and ρ_3, designated $\mu_{AA(l)}$ etc. This is a genetic architecture additive across loci, but not within a locus.

We introduce the strong assumption of independence of effects across loci. This is not an assumption about linkage equilibrium, but about the mutation processes that generate alleles at each locus. For the sake of model completeness, we may assume unlinked Mendelian segregation. Extending the model to a linked pattern of inheritance, so long as inheritance is independent of trait values, does not affect the problem, since in the end we are conditioning on final allele state.

At a given locus, two individuals are in one of the identity states defined above, S_0 through S_4. We can express the relationship between the two individuals through a vector $S(l)$, where $S(l) \in \{0, 1, 2, 3, 4\}$ gives the identity state at locus l. The covariance between phenotypes P_A and P_B of the two individuals can be written in terms of $S(l)$, without reference to specific alleles:

$$\mathrm{Cov}\left(P_A, P_B\right) = \sum_{l=1}^{L} C_{S(l)} \tag{2.15}$$

Two individuals that share an identical genotype, whether it is homozygote or heterozygote (S_4 or S_3) experience the same contribution from that locus to the trait, and a correlation of 1. Individuals sharing no alleles (S_0) have independent contributions to the trait from the locus, and a correlation of 0. It is also possible to share two (S_1) or three (S_2) alleles, with corresponding effect correlations of ρ_2 and ρ_3 respectively.

2.4.2 Variance Component Representation

For a single locus, assuming an environmental error term with variance σ^2_E, we can write the covariance of a vector of individuals' trait values,

$$\mathrm{Cov}(\mathbf{P}) = C_4\mathbf{I}_4 \ + C_3\mathbf{I}_3 \ + C_2\mathbf{I}_2 \ + C_1\mathbf{I}_1 \ + \sigma^2_E\mathbf{I} \tag{2.16}$$

where $C_4 = \sigma^2_{AA}$, $C_3 = \sigma^2_{AB}$, $C_2 = \rho_3\sigma_{AB}\sigma_{AA}$, $C_1 = \rho_2\sigma^2_{AB}$; $\mathbf{I}_k$ is a matrix whose entry ij is an indicator, 1 when individuals i and j are in identity state k and 0 otherwise. We can write the positive semidefinite constraint (2.4) in terms of the variance components,

$$\rho_2 = \frac{C_1}{C_3}; \quad \rho_3 = \frac{C_2}{\sqrt{C_4 C_3}} \tag{2.17}$$

$$2C_1 \leq C_3; \quad 2C_2^2 \leq 2C_4C_1 \tag{2.18}$$

Both the additive model (2.5) and the strict dominance random ordering model (2.13) constrain the relationship between the four variance components down to one parameter. In the additive model we can solve

$$\frac{C_1}{C_3} = \rho_2 = \frac{1}{2}; \quad \frac{C_2}{\sqrt{C_4 C_3}} = \rho_3 = \frac{1}{\sqrt{2}}; \quad \frac{C_3}{C_4} = \frac{\sigma^2_{AB}}{\sigma^2_{AA}} = \frac{1}{2} \tag{2.19}$$

$$[C_1, C_2, C_3, C_4] \propto \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{2}, 1\right] \tag{2.20}$$

and equivalently, in the strict dominance random ordering model:

$$\frac{C_1}{C_3} = \rho_2 = \frac{1}{3}; \quad \frac{C_2}{\sqrt{C_4 C_3}} = \rho_3 = \frac{1}{2}; \quad \frac{C_3}{C_4} = \frac{\sigma^2_{AB}}{\sigma^2_{AA}} = 1 \tag{2.21}$$

$$[C_1, C_2, C_3, C_4] \propto \left[\frac{1}{3}, \frac{1}{\sqrt{2}}, 1, 1\right] \tag{2.22}$$

If the distribution of the effects G_{ij} is multivariate normal, so is that of $\mathbf{P}$. The posterior distributions of G_{ij} conditional on observed $\mathbf{P}$ is then also normal and expressible in matrix terms. Any such inferences are, of course, also conditional on known IBF parameters μ_{AA}, μ_{AB}, σ^2_{AA}, σ^2_{AB}, ρ_2, and ρ_3.

In the multi-locus case, the number of variance component parameters required to specify the model is very large:

$$\mathrm{Cov}(\mathbf{P}) = \sigma^2_E \mathbf{I} + \sum_{l=1}^{L} C_{4(l)}\mathbf{I}_{4(l)} \; + C_{3(l)}\mathbf{I}_{3(l)} \; + C_{2(l)}\mathbf{I}_{2(l)} \; + C_{1(l)}\mathbf{I}_{1(l)} \tag{2.23}$$

at each locus, the variances $C_{4(l)}$ and $C_{3(l)}$ must be nonnegative, and the positive semidefinite constraint must be satisfied. Estimation of these parameters from data poses special challenges. Under the restrictive assumption of equal IBF parameters for each locus, we can aggregate the identity state matrices over loci, and thus estimate only four variance component (plus σ^2_E). In this case the inference problem is a conventional mixed model, though with a nonlinear positive semidefinite constraint (2.4).

The most general inference problem is that in which IBF parameters are allowed to vary by locus. This is a mixed model with a number of variance components proportional to the number of loci, and quite possibly larger than the number of observations in $\mathbf{P}$. In the genomic selection literature surveyed by Kärkkäinen and Sillanpää (2012), analogous sparse regression problems are addressed by imposing single or hierarchical priors on weight parameters and finding point estimates via MAP (maximum a posteriori) methods or posterior

distribution estimates via MCMC methods. In the machine learning literature, the work of Tipping (2001) and Faul and Tipping (2001) on what in statistical terminology would be called random coefficient regression proposes an algorithm for large scale local optimization of models with large numbers of variance components.

2.4.3 Locus Weight and Identity State Contribution

Our method expresses the covariance between two individuals, given the IBF parameters, as sum of contributions by identity state, weighted over loci. Some, but not all, inferred relationship IBD (type 3) and IBS (type 2) covariance structures can also be expressed in this way. For a simple diallelic SNP model to which all three approaches are applicable, we can express all three models in terms that allow direct comparison. Consider two individuals X and Y. Each pair of SNP dosages at (X_l, Y_l), is an identity state for the pair of individuals at locus l. Each model defines a covariance, relatedness, or similarity index r_{XY}. We decompose r_{XY} into a locus weight and an identity state contribution $C_l(X_l, Y_l)$ where we fix $C_l(2,2) = 1$ and $C_l(2,0) = C_l(0,2) = 0$. Extracting a constant k, we can write

$$r_{XY} = k + \sum_{l=1}^{L} w_l C_l(X_l, Y_l) \tag{2.24}$$

Representative IBD method The realized relationship coefficient is justified in IBD terms by Powell et al. (2010) and here considered representative of a type 3, hybrid, model. It is defined for biallelic genotypes, and to express it we will arbitrarily designate one of the alleles at each locus as the A allele. It is computed for genotypes of individuals X and Y in terms of p_l, the population frequency of the A allele at locus l, and X_l and Y_l, the respective dosages of the A alleles at locus l in the two individuals:

$$r_{XY} = \frac{1}{2L} \sum_{l=1}^{L} \frac{(X_l - 2p_l)(Y_l - 2p_l)}{p_l(1-p_l)} \tag{2.25}$$

The classical phenotypic covariance, in the additive model, is proportional to this coefficient. For this parametrization, $w_l = 2/Lp_l$, and the normalized contributions $C_l(X_l, Y_l)$ are given in Table 2.1.

Representative IBS method An early estimator of relatedness on the basis of IBS genome sharing (Nejati-Javaremi et al., 1997) counts the number of the pairs of alleles IBS between the two individuals at each locus (0, 2 or 4) and averages this count over polymorphic loci. The arbitrariness of this definition is in the choice of polymorphic loci to be equally weighted; for example, adding loci presumed polymorphic, but actually not polymorphic, would skew the score by adding 4's into the average. This is equivalent to the similarity index:

> "For a given single locus, a similarity index S_{xy} between 2 individuals x and y is calculated, where $S_{xy} = 1$ when genotype $x = kk$ (i.e., both alleles at locus l are identical) and genotype $y = kk$. $S_{xy} = 0.5$ when $x = kk$ and $y = kl$, or vice versa, or when $x = kl$ and $y = kl$, $S_{xy} = 0.25$ when $x = kl$ and $y = km$, and $S_{xy} = 0$ when the 2 individuals have no alleles in common at the locus (Eding and Meuwissen, 2001). The similarity index was averaged over loci." (Hayes and Goddard, 2008)

We adapt the similarity index directly as our identity state contribution $C_l(X_l, Y_l) = S_{xy}$, and choose weights $w_l = 1/L$ for the L arbitrarily chosen loci.

The IBF method Following the IBF definition of the covariance between individuals (Eq. 2.15) we choose the identity state contribution $C_l(X_l, Y_l) = C_{S(l)}/\sigma^2_{AA(l)}$. We scale by choosing weights $w_l = \sigma^2_{AA(l)}$ and introducing the parameter $v = \sigma_{AB(l)}/\sigma_{AA(l)}$.

Comparison To allow comparison with the IBD method, we consider only diallelic combinations, excluding e.g. $AB : AC$. The results are demonstrated in Table 2.1. We can observe that the IBD method treats the two alleles asymmetrically, unless the allele frequency of $p = 0.5$ imposes symmetry. This offsets the asymmetry of the weighting scheme, $w_l = 2/Lp_l$; of course when $p_l = 0.5$, $w_l = 1/L$.

The key observation is that the same identity state contribution function is derived in the simplest and most symmetric case from each approach. The IBS similarity index is constructed from arbitrary symmetry considerations. The IBD-based realized relationship

	IBD	IBD with $p = 1/2$	IBS	IBF	IBF with $v = \rho_3 = \frac{\sqrt{2}}{2}$
AA:AA	1	1	1	1	1
AA:AB	1/2	1/2	1/2	$\rho_3 v$	1/2
AA:BB	0	0	0	0	0
AB:AB	$\frac{1}{4}\left(1 + \frac{p}{1-p}\right)$	1/2	1/2	v^2	1/2
AB:BB	$\frac{1}{2}\frac{p}{1-p}$	1/2	1/2	$\rho_3 v$	1/2
BB:BB	$\frac{p}{1-p}$	1	1	1	1

Table 2.1: Comparison of identity state contributions under IBD, IBS, and IBF methods.

method gives the same function when the allele frequency $p = 0.5$. The IBF parametrization $v = \rho_3 = \frac{\sqrt{2}}{2}$ corresponds to the additive model.

We can interpret the difference thus: the IBF framework should deviate from the IBS method when there is dominance in the genetic architecture of the specific trait. The IBD method should deviate when major and minor allele frequencies differ.

2.5 Applications

For a set of known IBF parameters, and given data in the form of genotype-phenotype pairs, there exists a conditional distribution of the effects G_{ij}, from which other inferences can be made about the effects themselves, classical variance components and heritabilities, and phenotypes and breeding values.

Under the assumption of normality and known IBF parameters, most operations involved in these applications are reducible to linear algebra. Section 2.4.2 above described the separate, and greater, computational challenge of estimating the IBF parameters from data.

2.5.1 Association

The generative model produces genotypic values G_{ij} for each allele pair at each locus. By classical methods (Lynch and Walsh, 1998, Chapter 4), these can be mapped to additive effects α_i, dominance deviations δ_{ij}, and the additive variance σ^2_A, given allele frequencies.

For fixed IBF parameters the distribution of effects G_{ij}, if assumed multivariate normal as in Section 2.3.4, is jointly multivariate normal with the phenotype distribution in Section 2.4.2. Then the effect distribution conditional on observed phenotype is also normal and has a matrix form.

Formally, this implies the ability to perform a GWAS style test for the significance of individual loci and alleles. The power to do so is severely limited by the large number of parameters, one for each allele pair at each locus. Such testing is practical when IBF parameters are either assumed known or inferred from outside the sample. Significance for an individual allele is easiest to achieve when IBF parameters induce sparsity or concentration on particular loci.

2.5.2 Heritability

An important application of type 3 models (Yang et al., 2010) has been the estimation of heritability explainable by common variants. It has been pointed out (Visscher, 2010; Makowsky et al., 2011) that such estimates of heritability are not the same as predictive R^2. With reasonable sample size, heritability can be estimated with limited error, but the large number of individual effects needed to build an accurate predictive model cannot. The narrow-sense heritability estimation problem, ignoring complications, is reducible to the estimation of the additive variance, σ^2_A, which in a type 3 context (e.g. Yang et al., 2010) is a natural variance component parameter of the model. In the setting where σ^2_A is not a model parameter, we construct it, bottom-up, from additive effects and dominance deviations. The additive variance, σ^2_A, is a quadratic form in the G_{ij}, and a random variable under our model, for fixed allele frequencies. There are two ways to calculate an estimate of σ^2_A, and therefore heritability, conditional on observed data:

1. Substitute the expectations of individual G_{ij} into the quadratic form.

2. Compute the expectation of the quadratic form over the uncertainty about the G_{ij}.

By convexity the second, expected, σ^2_A is greater than the first σ^2_A. We interpret the second as an estimate of the true additive variance; and the first as the estimate of the additive variance explained by the model as inferred given finite data.

In Chapter 5, we illustrate heritability estimation in simulated and real data. In Section 2.6 we develop an approach to decomposing heritability estimates into Explained and Unexplained components associated with the convexity effect described above.

2.5.3 Prediction

Given the IBF parameters, a covariance between each pair of individuals, and therefore a covariance matrix among n individuals, can be computed. The IBF covariance can be interpreted as another kernel in the framework of Gianola and van Kaam (2008), and applied to the prediction of phenotype for an individual with known genotype, given a set of genotype-phenotype pairs. Such prediction has many analogies, and is algebraically equivalent to prediction in Henderson's BLUP, Kriging, or Gaussian Processes (Robinson, 1991). A prediction variance is computed together with the prediction, and is generally lower when data on "neighbors" of the individual to be predicted is available. A neighbor conducive to such prediction is an individual that shares genotype at important loci; the importance weight of a locus roughly corresponds to high variance in its IBF parameters.

2.5.4 Prediction, Populations, and Population Neutrality

Population neutral applications of the IBF model, in particular phenotype and individual genotypic effect prediction, do not require us to specify a reference population in the form of a set of allele frequencies. Unlike phenotype prediction, breeding value prediction is not population neutral in a model with dominance; indeed the breeding value can be defined through regression of the offspring trait onto a parent crossed with a mate chosen at random from a given population. Classically, we can predict the breeding value by inferring, and summing, the additive effects α_i at each locus.

The bottom-up reconstruction of α_i, breeding value, and heritability thus take a set of allele frequencies as an input. These may be be directly estimated from the same population as the effects; alternatively, we may transpose our effect estimates onto a different population, or use an adjusted allele frequency estimation method (e.g. Sillanpää, 2011).

2.6 Explained and Unexplained Heritability Decomposition

Classically, the contributions of each locus to the variance components σ^2_A and σ^2_D are functions of the allele frequencies p_i and additive effects α_i and dominance residuals δ_{ij} The α_i and δ_{ij}, in turn, coefficients and residuals of a weighted regression of the genotypic effects G_{ij} onto an allele dosage matrix, with the weighting a function of the p_i. Generally, σ^2_A and σ^2_D, and for our non-epistatic model $\sigma^2_G = \sigma^2_A + \sigma^2_D$, are quadratic in G_{ij}; that is, each has the form $\mathbf{G}'\mathbf{Q}\mathbf{G}$ where $\mathbf{G}$ is the vector of the G_{ij} and $\mathbf{Q}$ depends on p_i.

In our model, both the G_{ij} and the p_i are random variables; thus σ^2_A and σ^2_D are random. Our model has variance components, $C_1 ... C_4$, which determine the IBF parameters at each locus. These do not translate deterministically to the classical variance components σ^2_A and σ^2_D, but to *expected* σ^2_A and σ^2_D, either under given allele frequencies, or under a distribution of allele frequencies.

Consider a fixed population with known allele frequencies p_i, and a single locus trait with known IBF parameters and environmental variance σ^2_E. If we have complete knowledge of the G_{ij}, the variance components σ^2_A and σ^2_D are given deterministically by the classical calculations. If we have no knowledge of G_{ij}, we can determine the expectations over possible realizations of genotypic values, $\mathrm{E}\left[\sigma^2_A\right]$ and $\mathrm{E}\left[\sigma^2_D\right]$. From a Bayesian perspective, the former case corresponds to a degenerate complete-data posterior, and the latter to a prior, on genotypic effects. If we observe some pairs of genotypes and phenotypes, and condition on this information, we obtain a posterior over genotypic effects, and the Bayes estimator of the variance components is the expectation under that posterior.

Under the normal model, taking for simplicity $\mu_{AA} = \mu_{AB} = 0$,

1. The prior is the IBF normal genotypic value model and has the form $\mathbf{G} \sim N(\mathbf{0}, \mathbf{\Sigma}_1)$

2. The posterior is obtained by conditioning the prior on observed phenotypes, and has

the form $\mathbf{G} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$

3. Complete knowledge of $\mathbf{G}$ is equivalent to $\mathbf{G} \sim N(\boldsymbol{\mu}_3, \mathbf{0})$

The expectation of a quadratic in a random vector is a sum of mean and covariance terms. Under the assumption of normality, so is the variance (e.g. Mathai and Provost, 1992):

$$\mathrm{E}\left[\mathbf{G}'\mathbf{Q}\mathbf{G}\right] = \boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu} + \mathrm{tr}\left[\mathbf{Q}\boldsymbol{\Sigma}\right] \tag{2.26}$$

$$\mathrm{Var}\left[\mathbf{G}'\mathbf{Q}\mathbf{G}\right] = 4\boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\mu} + 2\mathrm{tr}\left[\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\Sigma}\right] \tag{2.27}$$

In the methodology of Yang et al. (2010) addressing the "Missing Heritability" problem, variance component estimates imply a heritability that may be significantly higher than the R^2 achievable by prediction via the corresponding covariance model or attributable to individual markers. Here, we decompose the *Total* heritability or variance component estimate into an *Explained* and an *Unexplained* component. The mean term is the variance component, or equivalently heritability, *Explained* by inferred effects; it is the variance component we would get by substituting the predicted (i.e. expected, conditional on observed data) genotypic values into the classical expression for the variance component. The covariance term then corrects for the bias due to the genotypic values being predicted from limited data. These inferences are illustrated by the simulation in Section 5.2.4. In the extreme cases, when we observe no phenotype data, we have only the covariance term (unless $\mu_{AA} \neq \mu_{AB}$). Given infinite phenotype data, the mean term converges to the true variance component value; both the covariance term and the error term (SD) vanish. More generally, the heritability in the *Explained* category increases with sample size, and the estimated adjustment in the *Unexplained* category decreases with sample size. In the simulation, we obtain Table 2.2 by conditioning on 1000 genotype-phenotype pairs.

	σ^2_A	σ^2_D	$\sigma^2_G = \sigma^2_A + \sigma^2_D$
Explained	0.41	0.05	0.46
Unexplained	0.01	0.02	0.04
Total	0.43	0.07	0.50
SD	0.04	0.01	0.04

Table 2.2: Decomposition of the expected variance components into Explained and Unexplained components for simulated case.

Chapter 3

METHODOLOGY: IDENTITY BY FUNCTION

3.1 Introduction

In Chapter 2 (Theory) we have described the Identity by Function model of the genotype-phenotype relationship and the IBF states. In this chapter we will describe procedures for mapping genetic data to the model, inference of the model's parameters from data, and for manipulating the fitted model to make predictions about phenotypes, breeding values, and allele effects.

First we will describe the mapping between genetic data and our model of identity states. The *state identity array* is a general representation of identity states incorporating ambiguity, which can be used to model missing data.

We define identity states relative to a locus. As discussed in the Theory chapter, there are two approaches to defining a locus or Quantitative Trait Locus (QTL). We can treat the minimal unit of genetic variability, such as a SNP, as a locus, or we can aggregate genetic variation across a functional gene-scale region as a QTL into distinct functional alleles. For the purposes of this chapter we assume the genotype data available to us are in the form of bialleic SNPs, possibly phased. For biallelic SNP locus data, we will deal with the issues of missing data imputation and ambiguous genotypes. For gene-scale QTL, given phased SNP data, we discuss the use of bioinformatic databases for classification of SNPs into gene-related synonymous allele groups.

We also describe a dictionary based method for mapping limited resolution SNP data to a reference panel of complete sequences. This has applications usable for phasing, imputation and for estimation of probabilies of identity states involving unknown alleles.

Having constructed the state identity array from genotype information, we will treat the IBF parameters σ^2_{AA}, σ^2_{AB}, ρ_2, and ρ_3 at each locus as known, together with the global environmental variance, σ^2_E. Together, these allow us to construct the phenotype covariance

matrix and and the covariances between phenotypes and effects. Given this, we will form inferences about genotypic values and additive and dominance effects. This in turn enables phenotype and breeding value prediction. From inferred effects, and estimated uncertainties of inferred effects, we can construct estimates of classical additive and dominance variances. By appropriate scaling and simplification, the IBF covariance matrix becomes the functional Genomic Relationship Matrix (*fGRM*), substitutable like the GRM (e.g. Hayes et al., 2009; VanRaden, 2008) for the numerator matrix in heritability estimation and prediction.

This chapter considers the problems of inference when IBF parameters are known. In Chapter 4, we will consider the inference of IBF parameters themselves from genotype and phenotype information. The problem is analogous to the classical ML/REML variance component problem, and thus readily solvable under the assumption of equal or fixed locus weights. Without this assumption, the problem corresponds to a search in a high-dimensional variance component space.

3.2 From Genotype to State Identity Array

3.2.1 Representations of State Identity

We have defined the identity state $S_0 \dots S_4$ for a pair of individuals at a locus. We will now describe a representation for the identity state of multiple individuals across multiple loci, which allows us to incorporate ambiguity about the genotype. At the raw sequence level, such ambiguity may come from missing genotype data or instrument uncertainty about calling individual base pairs. At an intermediate sequence interpretation level, ambiguity may be due to incomplete coverage of genetic variants, or uncertainty about local phasing. At the level of bioinformatic sequence interpretation, uncertainty about the functional identity of distinct sequences may also be represented as identity state ambiguity. Any of these phenomena may potentially be represented by considering the relationship between the pair of individuals at the locus as a probability mixture of the five states. We assign such identity state probabilities by marginalizing away the joint distribution of the genotypes, which need not be independent for different individuals at the same locus.

Whatever the model of state uncertainty, we can define a *state identity array* T_{ijkl} of

probabilities that individuals i and j are in state S_k at locus l with the following properties:

1. l (for locus) is the label of the locus, ranging from 1 to L
2. i and j (for individuals) are labels of individuals, ranging from 1 to N
3. The array is symmetric with respect to i and j; that is, $T_{ijkl} = T_{jikl}$
4. k (for klass) is the label of the identity state, from 0 to 4
5. T_{ijkl} are the $T_{ijkl} \geq 0$ for all i, j, k, l
6. $\sum_{k=0}^{4} T_{ijkl} = 1$ for all i, j, l
7. Every individual is in identity class S_3 or S_4 with themselves (see page 20 for state descriptions) at a given locus: $T_{ii3l} + T_{ii4l} = 1$

Not every possible state identity array is internally consistent, in the sense of representing some mixture of genotypes for each individual. For example, only homozygotes can be in state S_4, and only heterozygotes can be in state S_3; therefore is individual A is in S_4 with B with probability 1, and individual B is in S_4 with C with probability 1, then individual C must have zero probability of being in S_3 with A. Violations of internal consistency may occur if we manipulate the array arbitrarily, for example for imputation, but not if we construct the array directly from mixtures of genotypes. Inconsistent identity states can manifest themselves as violations of positive semidefiniteness assumptions when constructing covariance matrices from a given identity state array.

3.2.2 State Identity from SNP data

A common representation for biallelic SNP array genomes is the dosage encoding. At each locus, one allele is designated as ancestral and one as minor. An ancestral homozygote is encoded as a 0, a minor allele homozygote as a 2, and a heterozygote as a 1. Table 3.1 demonstrates the mapping from this encoding to the IBF states.

	0	1	2
0	S_4	S_2	S_0
1	S_2	S_3	S_2
2	S_0	S_2	S_4

Table 3.1: Mapping of SNP combinations to IBF states.

When some SNP data is missing, as commonly indicated in databases by the sentinel dosage value -1, we can impute identity states by treating the missing state as a mixture of the genotypes weighted according to the population genotype frequencies. We estimate the genotype frequencies p_0, p_1, and p_2 from the available non-missing SNPs at the same locus. Alternatively, we can estimate (possibly out of sample) the single overall minor allele frequency p, and use the Hardy-Weinberg genotype frequencies $p_0 = (1-p)^2$, $p_1 = 2p(1-p)$, and $p_2 = p^2$.

Table 3.2 shows the mapping from the allele frequencies to identity state probabilities. The coefficients associated with the state S_k become the nonzero entries in the identity state array T_{ijkl} for the corresponding individuals i and j at locus l.

Another SNP encoding incorporates dosage uncertainty by allowing dosage d to take on intermediate values $0 \leq d \leq 2$. This can be interpreted as a mixture between the two adjacent integer dosages, with the real-valued dosage an expectation of the integer value. A way to transform dosage expectations to genotype probabilities suitable for programming by vectorized operations is as follows:

$$p_0 = \max(0, 1 - |0 - d|) \tag{3.1}$$

$$p_1 = \max(0, 1 - |1 - d|) \tag{3.2}$$

$$p_2 = \max(0, 1 - |2 - d|) \tag{3.3}$$

Given genotypes in the form of a dosage matrix d_{il}, this approach generates three matrices, p_{0il}, p_{1il}, and p_{2il}, such that

$$0p_{0il} + 1p_{1il} + 2p_{2il} = d_{il} \tag{3.4}$$

	-1	0	1	2
-1	$2p_0p_2S_0$ $+2\left(p_0p_1+p_1p_2\right)S_2$ $+p_1p_1S_3$ $+\left(p_0p_0+p_2p_2\right)S_4$	p_0S_4 $+p_1S_2$ $+p_2S_0$	p_0S_2 $+p_1S_3$ $+p_2S_2$	p_0S_0 $+p_1S_2$ $+p_2S_4$
0	$p_0S_4+p_1S_2+p_2S_0$	S_4	S_2	S_0
1	$p_0S_2+p_1C_3+p_2S_2$	S_2	S_3	S_2
2	$p_0S_0+p_1S_2+p_2S_4$	S_0	S_2	S_4

Table 3.2: Mapping of SNP combinations to IBF states with missing data.

From this representation we can directly compute the identity state array:

$$T_{ij4l} = p_{0il} * p_{0jl} + p_{2il} * p_{2jl} \tag{3.5}$$

$$T_{ij3l} = p_{1il} * p_{1jl} \tag{3.6}$$

$$T_{ij2l} = p_{0il} * p_{1jl} + p_{1il} * p_{0jl} + p_{2il} * p_{1jl} + p_{1il} * p_{2jl} \tag{3.7}$$

$$T_{ij1l} = 0 \tag{3.8}$$

$$T_{ij0l} = p_{0il} * p_{2jl} + p_{2il} * p_{0jl} \tag{3.9}$$

3.2.3 Multiallelic Data

When more than two alleles are possible at a locus, the dosage representation is not applicable. Genotype data are encoded in the form of paired vectors of allele labels when there is no ambiguity. It is still helpful for programming purposes to vectorize the operation of forming the identity state array.

Let $\mathbf{a}_1$ and $\mathbf{a}_2$ be the two allele encoding vectors for individual i, and likewise $\mathbf{b}_1$ and $\mathbf{b}_2$ for individual j. The vectors are across loci (that is, they have L entries).

Let $\mathbf{z}_a$ and $\mathbf{z}_b$ be the indicator vectors for homozygosity of the two individuals:

$$\mathbf{z}_{al} = I\left[\mathbf{a}_{1l}{=}\mathbf{a}_{2l}\right] \tag{3.10}$$

Then we form the identity count vectors $\mathbf{c}_0, \mathbf{c}_1, \ldots \mathbf{c}_4$, corresponding to the possible numbers of times the alleles are in identity states:

$$\mathbf{c}_{ml} = I\left[m = I\left[\mathbf{a}_{1l}{=}\mathbf{b}_{1l}\right] + I\left[\mathbf{a}_{1l}{=}\mathbf{b}_{2l}\right] + I\left[\mathbf{a}_{2l}{=}\mathbf{b}_{1l}\right] + I\left[\mathbf{a}_{2l}{=}\mathbf{b}_{2l}\right]\right] \tag{3.11}$$

From this we can write down the rules for forming the state identity array (assuming memberwise multiplication throughout):

$$T_{ij4l} = \mathbf{c}_{4l} \tag{3.12}$$

$$T_{ij3l} = \left(1 - \mathbf{z}_{bl}\right)\left(1 - \mathbf{z}_{al}\right)\mathbf{c}_{2l} \tag{3.13}$$

$$T_{ij2l} = \left[\mathbf{z}_{al}\left(1 - \mathbf{z}_{bl}\right) + \mathbf{z}_{bl}\left(1 - \mathbf{z}_{al}\right)\right]\mathbf{c}_{2l} \tag{3.14}$$

$$T_{ij1l} = \mathbf{c}_{1l} \tag{3.15}$$

$$T_{ij0l} = \mathbf{c}_{0l} \tag{3.16}$$

3.2.4 Phased Sequence Data and Genome Partitioning

For relatively low resolution, SNP based datasets, the choice of methodologies is restricted by the available data. Given a complete genome sequence, or a high resolution approximation, we are faced with a number of discretionary choices about the mapping between genetic data and model inputs. In particular, our method relies on partitioning the genome into functional regions, and identifying functionally distinct alleles within the region.

We processed human genotype-phenotype datasets as described in Chapter 5. Each of the datasets available was originally sequenced using a 0.5 to 1 million SNP array. These data were preprocessed using the BEAGLE model (Browning and Browning, 2007). This resulted in genotypes phased and imputed to the resolution of the 1000 Genomes project (1000 Genomes Project Consortium, 2012). This approach has significant limitations. We do not introduce additional treatment for any uncertainty contributed by the probabilistic phasing and imputation, and, of course, statistical imputation cannot identify rare variants. It does, however, serve as a proxy for the challenge that would be faced if, or once, complete phased genome sequences became available.

The genotype data are available to us in the form of tuples of individual identifier (e.g. *ID001*), SNP-scale variant identifier (e.g. *RS001*), and allele (e.g. *T*). The task

is to map this to functional regions and functional alleles; that is, to tuples of individual identifier (e.g. *ID001*), gene-scale functional region identifier (e.g. *BRCA2*), and allele (e.g. *FunctionalAllele5*).

We used the ENSEMBL database (Flicek et al., 2014), with release 72 of the Homo Sapiens genome and genome variation database, referencing the GRCh37 consensus assembly of the human genome.

First, we selected the set of variants with protein altering consequences. We used a narrow definition, selecting only variants which unambiguously alter some catalogued (computationally predicted or experimentally observed) protein transcript. Thus, synonymous variants within an exon were excluded, but frame shift variants included. The MISO sequence onthology browser (Eilbeck et al., 2005) was helpful in classifying the entries in the transcript variation, which may be categorized in slightly different ways in different transcript record sources. We constructed a 'magic number' bit field query to match transcript variants. The transcript database stores the set of possible variant consequences for each variant-transcript pair as a bit field, and set intersection by a bitwise and operation is an efficient way of performing this query. Table 3.3 enumerates the sequence ontology choices and the construction of the bit field for the query.

Also available in the transcript database were variant impact scores, including PolyPhen scores (Adzhubei et al., 2010). While this is a popular and powerful approach for classifying variants into potentially damaging and non-damaging classes, we chose not to incorporate such information. Incorporating such scores would be contrary to the logic of the IBF approach, and, in particular, its symmetric, infinite allele assumptions. The method aims to distinguish alleles with non-identical function, and then infer from the data their relative strengths. Additionally, we are interested in the effect of alleles which differ at multiple SNP loci within the same gene-scale region, and it would require a substantial separate research effort to extend variant impact score methodology to multiple mutations.

The set of SNPs implicated in protein altering variation is then intersected with the (1000 genome based) set of SNPs in our imputed dataset. As the nomenclature for SNPs can vary between databases, ENSEMBL's variant synonym database is used to match identifiers. Each SNP is associated with a transcript for which it has a protein altering consequence,

Ensemble Database Bit Identifier	Sequence Ontology Code	Included	Bit Contribution
0	splice _acceptor_variant	No	0
1	splice_donor_variant	No	0
2	stop_lost	Yes	4
3	coding_sequence_variant	No	0
4	missense_variant	Yes	16
5	stop_gained	Yes	32
6	synonymous_variant	No	0
7	frameshift_variant	Yes	128
8	nc_transcript_variant	No	0
9	non_coding_exon_variant	No	0
10	mature_miRNA_variant	No	0
11	NMD_transcript_variant	No	0
12	5_prime_UTR_variant	No	0
13	3_prime_UTR_variant	No	0
14	incomplete_terminal_codon_variant	Yes	16384
15	intron_variant	No	0
16	splice_region_variant	No	0
17	downstream_gene_variant	No	0
18	upstream_gene_variant	No	0
19	initiator_codon_variant	Yes	524288
20	stop_retained_variant	No	0
21	inframe_insertion	Yes	2097152
22	inframe_deletion	Yes	4194304
23	transcript_ablation	No	0
24	transcript_fusion	No	0
25	transcript_amplification	No	0
26	transcript_translocation	No	0
27	TFBS_ablation	No	0
28	TFBS_fusion	No	0
29	TFBS_amplification	No	0
30	TFBS_translocation	No	0
31	regulatory_region_ablation	No	0
32	regulatory_region_fusion	No	0
33	regulatory_region_amplification	No	0
34	regulatory_region_translocation	No	0
35	feature_elongation	No	0
36	feature_truncation	No	0
Total			6832308

Table 3.3: Choices of transcript variant classifications for ENSEMBL database query.

and the transcript is in turn associated with a gene. A single SNP may be associated with more than one gene. Most SNPs are not associated with any gene. Not every gene has a SNP with a protein-altering consequence matched with the set of SNPs in our dataset.

For each gene with at least one such SNP, the name of each functional allele is constructed by concatenating the names of the protein altering variants in an appropriately matched order. Thus it resembles a DNA sequence, but is not one, because the base pairs contained are not consecutive and need not be on the same strand.

Within the Identity State Array, we now treat each gene as a locus, and form deterministic identity states by matching functional allele names.

3.2.5 The Dictionary Method

The dictionary method applies when we are given limited resolution genotype data for individuals in our study. Suppose we also have a dictionary of alleles at that locus, a random sample of completely sequenced, phased haplotypes. This dictionary contains n haplotypes from $n/2$ individuals, which are made up of some number of common variants, at most n, each repeated between 1 and n times. For the purposes of the dictionary method, we will use the term *common variant* synonymously with variant found in the dictionary, and *rare variant* with one not found in the dictionary. An individual observed allele in the study is either a common variant or a a rare variant. Two or more alleles may be in an identity state by having either common or rare alleles in common.

We represent the process for generating novel alleles with the Polya urn model, which yields the Ewens Sampling Formula with parameter θ. Under this model if we observe a new allele, with no genotype information, having previously observed n alleles which we have recorded into the dictionary,

1. The new allele has probability $\frac{1}{n+\theta}$ of being identical with any of the n already observed alleles in the dictionary

2. It has probability $P = \frac{\theta}{n+\theta}$ of being some unobserved ('rare') allele

The Ewens Sampling Formula,

$$\Pr\left(a_1,\ \ldots,\ a_n;;\theta\right) = \frac{n!}{\theta\left(\theta+1\right)\ldots\left(\theta+n-1\right)}\prod_{j=1}^{n}\frac{\theta^{a_j}}{j^{a_j}a_j!} \tag{3.17}$$

for $a_1+2a_2+3a_3+\cdots+na_n=n$, gives the probability that among the n alleles, the distinct allele counts are distributed according to the pattern a_i. For example $a_n=1$ corresponds to the case where all n alleles are identical (i.e. there is a single allele class with count n), and $a_1=n$ to the case where all n alleles are distinct (i.e. there are n allele classes, each with count 1).

Standard results related to the estimattion of θ can be found in, e.g., Ewens (2004). Given observed allele counts a_i, the MLE of θ can be derived as

$$\ln\Pr\left(a_1,\ \ldots,\ a_n;;\theta\right) = C-\ln\left[\theta\left(\theta+1\right)\ldots\left(\theta+n-1\right)\right] + \left(\sum_j a_j\right)\ln\theta \tag{3.18}$$

$$\frac{\theta}{\theta}+\frac{\theta}{\theta+1}+\ldots\ \frac{\theta}{\theta+n-1}=\sum_j a_j \tag{3.19}$$

We can interpret $K=\sum_j a_j$ as the number of distinct alleles; if $K\to 1$, $\widehat{\theta}\to 0$, and if $K\to n$, $\widehat{\theta}\to\infty$. $E\left[\mathbf{K}\right]$ is the same as this MLE; that is, the method of moments estimator is the also MLE, as K is a natural parameter of an exponential family.

$$\sum_{i=0}^{n-1}\frac{\theta}{\theta+i}=1+\sum_{i=1}^{n-1}\frac{nP}{nP+i\left(1-P\right)}=K \tag{3.20}$$

The following expression can be stably solved for P in the $[0,1]$ interval:

$$\sum_{i=1}^{n-1}\frac{P}{nP+i\left(1-P\right)}=\frac{K-1}{n} \tag{3.21}$$

Asymptotically:

$$\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{P}{P+\frac{i}{n}\left(1-P\right)}\approx\int_0^1\frac{P}{P+x\left(1-P\right)}dx=\frac{P}{1-P}\ln P \tag{3.22}$$

So for large n,

$$\frac{K-1}{n-1}=\frac{P}{P-1}\ln P \tag{3.23}$$

Thus, even though P is a function of both n and θ, and the estimator (either method of moments or MLE) of θ is a function of k and n involving an infinite sum, for large $n>100$,

P	$\frac{K-1}{n-1}$		P	$\frac{K-1}{n-1}$
0	0.00000		0.00000	0
0.001	0.00691		0.00011	0.001
0.01	0.04652		0.00154	0.01
0.1	0.25584		0.02692	0.1
0.25	0.46210		0.09665	0.25
0.5	0.69315		0.28467	0.5
0.75	0.86305		0.57683	0.75
0.9	0.94824		0.81290	0.9
0.99	0.99498		0.98013	0.99
0.999	0.99950		0.99800	0.999

Table 3.4: Asymptotic probability of encountering an allele not previously recorded in dictionary.

we have an estimator of P as an inverse of a closed form involving only $\frac{K-1}{n-1}$, bypassing the need to separately estimate θ. Table 3.4 shows solutions of this expression for representative values of P; that is, the estimated probabilities that the next allele encountered will be a rare allele, given the diversity in a large dictionary.

The probabilities given above are for a randomly drawn new allele with no genotype information. In practice, suppose we observe a limited set of SNPs for the new allele, smaller than the set of SNPs in the dictionary. Some of the dictionary alleles are incompatible at the observed SNPs. Those that are compatible at all observed SNPs may or may not be the same allele; and the new allele may be a rare allele. The simplest model for the type of the new allele is constructed by filtering away the incompatibility. The new allele has probability $\frac{1}{n_c+\theta}$ of being identical with any of the n_c already observed alleles in the dictionary which are consistent, and probability $P_c = \frac{\theta}{n_c+\theta}$ of being some unobserved ('rare') allele.

With this model we can perform several types of inferences: phasing, expected state decomposition, and computation of identity state probabilities. For example, suppose we know

unphased SNP genotypes, at a specific locus (e.g. 0120211111). Generalized phasing means not only assigning probabilities to each assignment of SNP to one of two complementary haplotypes, but assigning a probability to each possible combination of the form:

1. Common Variant 5/Common Variant 3

2. Common Variant 5/Rare Variant A

3. Rare Variant A/Rare Variant A (homozygote only)

4. Rare Variant A/Rare Variant B

We now have the unconditional probabilities of each common variant, and the unconditional probability of encountering a rare variant. Applying the Ewens Sampling Formula with $\theta = nP/(1-P)$, we can compute the conditional probabilities to encountering distinct or identical rare variants, given that we observe two rare variants. We compute the phased haplotype pair probabilities by naively conditioning on compatibility; that is, by dividing the unconditional probability of observing each compatible variant pair by the sum of unconditional probabilities of such pairs.

We can decompose the total 200% allele dosage into expected contributions from dictionary allele, comprising a total dosage $2(1-P_c)$ broken out into K distinct alleles, and $2P_c$ expected dosage from rare alleles. Such a decomposition follows directly from the marginal probabilities of the dictionary alleles in the haplotype pair distribution in the phasing step. The expected dosages can be used, in a GWAS context, as haplotype dosages; that is, a regression of a trait Y onto dummy variables X_j corresponding to dictionary alleles

$$Y_i = \mu + d_{1i}X_{1i} + d_{2i}X_{2i} + \cdots + \epsilon_i \tag{3.24}$$

can be carried out by substituting the expected dosages for true values of X_j $(0, 1, 2)$.

To find IBF identity state probabilities, we consider 4 alleles at a time. For up to 4 rare variants, given a value of $\theta = nP/(1-P)$, the Ewens Sampling Formula gives the probabilities of sampling some combination of identical rare variants. The probabilities of

the identity states $S_0 \ldots S_4$ for a pair of individuals at a locus are computed by summing identity combinations due to:

1. Matching common variants, with match probability due to alleles being assigned to the same compatible dictionary allele class

2. Matching rare variants, with match probability due to Ewens Sampling Formula; when both individuals have at least one rare allele, we can evaluate the probability that these alleles are identical

3. Combinations of the two types of matches, namely, matches on one common allele and one rare allele; these can only be of type S_3 $(AB : AB)$

Potential extensions to the model include:

1. Weighting the reference population used to build the dictionary to account for population structure;

2. Population models with more than the single θ parameter of the Ewens Sampling Formula;

3. Use of biologically motivated SNP probability models, possibly including SNPs outside the reference region, to weigh the probabilities of matching the proposed allele with the dictionary alleles.

3.3 Locus Weights, Identity State Array, and Covariance

At each locus l from 1 to L, we have locus-specific IBF parameters σ^2_{AAl}, σ^2_{ABl}, ρ_{2l}, and ρ_{3l}. Previously, for a single locus, assuming an environmental error term with variance σ^2_E, we have written the covariance of a vector of individuals' trait values,

$$\mathrm{Cov}(\mathbf{P}) = C_4\mathbf{I}_4 \ + C_3\mathbf{I}_3 \ + C_2\mathbf{I}_2 \ + C_1\mathbf{I}_1 \ + \sigma^2_E\mathbf{I} \tag{3.25}$$

where $C_4 = \sigma^2_{AA}$, $C_3 = \sigma^2_{AB}$, $C_2 = \rho_3\sigma_{AB}\sigma_{AA}$, $C_1 = \rho_2\sigma^2_{AB}$; $\mathbf{I}_k$ is a matrix whose entry ij is an indicator, 1 when individuals i and j are in identity state k and 0 otherwise. In the multi-locus case, we attach locus labels to the variance components,

$$\mathrm{Cov}(\mathbf{P}) = \sigma^2_E\mathbf{I} + \sum_{l=1}^{L} C_{4l}\mathbf{I}_{4l} \; + C_{3l}\mathbf{I}_{3l} \; + C_{2l}\mathbf{I}_{2l} \; + C_{1l}\mathbf{I}_{1l} \tag{3.26}$$

Rewriting this in terms of the identity state array,

$$\mathrm{Cov}(\mathbf{P}_i, \mathbf{P}_j) = \sigma^2_E\delta_{ij} + \sum_{k=1}^{4}\sum_{l=1}^{L} C_{kl}T_{ijkl} \tag{3.27}$$

Recall that in the additive model,

$$\frac{C_1}{C_3} = \rho_2 = \frac{1}{2}; \quad \frac{C_2}{\sqrt{C_4C_3}} = \rho_3 = \frac{1}{\sqrt{2}}; \quad \frac{C_3}{C_4} = \frac{\sigma^2_{AB}}{\sigma^2_{AA}} = \frac{1}{2} \tag{3.28}$$

$$[C_1, C_2, C_3, C_4] \propto \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{2}, 1\right] \tag{3.29}$$

and equivalently, in the strict dominance random ordering model:

$$\frac{C_1}{C_3} = \rho_2 = \frac{1}{3}; \quad \frac{C_2}{\sqrt{C_4C_3}} = \rho_3 = \frac{1}{2}; \quad \frac{C_3}{C_4} = \frac{\sigma^2_{AB}}{\sigma^2_{AA}} = 1 \tag{3.30}$$

$$[C_1, C_2, C_3, C_4] \propto \left[\frac{1}{3}, \frac{1}{\sqrt{2}}, 1, 1\right] \tag{3.31}$$

We will define locally scaled covariances c_{kl} such that $c_{4l} = 1$ and

$$C_{kl} = \sigma^2_U w_l c_{kl} = \sigma^2_l c_{kl} \tag{3.32}$$

thus either scaling the variance terms with weights w_l and a single universal variance σ^2_U or with locus variance scales σ^2_l such that

$$\sigma^2_l = C_{4l} \tag{3.33}$$

3.4 Inference Given IBF Parameters

3.4.1 Genotype Distribution Model

Consider the IBF model at a single locus. With four alleles, labelled A, B, C, D, the model defines genotypic values, which can be represented as a matrix $\mathbf{G}_M$ or a genotypic value

vector $\mathbf{G}$, eliminating diploid redundancy:

$$\mathbf{G}_M = \begin{bmatrix} G_{AA} & G_{AB} & G_{AC} & G_{AD} \\ G_{BA} & G_{BB} & G_{BC} & G_{BD} \\ G_{CA} & G_{CB} & G_{CC} & G_{CD} \\ G_{DA} & G_{DB} & G_{DC} & G_{DD} \end{bmatrix}; \quad \mathbf{G} = \begin{bmatrix} G_{AA} \\ G_{BB} \\ G_{CC} \\ G_{DD} \\ G_{AB} \\ G_{AC} \\ G_{AD} \\ G_{BC} \\ G_{BD} \\ G_{CD} \end{bmatrix} \tag{3.34}$$

We have previously shown that the joint covariance matrix for $\mathbf{G}$ has the form

$$\begin{aligned} &\mathrm{Cov}\left(([G_{AA}\ G_{BB}\ G_{CC}\ G_{DD}]\,;[G_{AB}\ G_{AC}\ G_{AD}\ G_{BC}\ G_{BD}\ G_{CD}])'\right) \\ &= \begin{bmatrix} \begin{bmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{bmatrix}\sigma^2_{AA} & \begin{bmatrix} 1&1&1&0&0&0 \\ 1&0&0&1&1&0 \\ 0&1&0&1&0&1 \\ 0&0&1&0&1&1 \end{bmatrix}\rho_3\sigma_{AB}\sigma_{AA} \\ \begin{bmatrix} 1&1&0&0 \\ 1&0&1&0 \\ 1&0&0&1 \\ 0&1&1&0 \\ 0&1&0&1 \\ 0&0&1&1 \end{bmatrix}\rho_3\sigma_{AB}\sigma_{AA} & \begin{bmatrix} 1&0&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&0&1&0&0&0 \\ 0&0&0&1&0&0 \\ 0&0&0&0&1&0 \\ 0&0&0&0&0&1 \end{bmatrix}\sigma^2_{AB} + \begin{bmatrix} 0&1&1&1&1&0 \\ 1&0&1&1&0&1 \\ 1&1&0&0&1&1 \\ 1&1&0&0&1&1 \\ 1&0&1&1&0&1 \\ 0&1&1&1&1&0 \end{bmatrix}\rho_2\sigma^2_{AB} \end{bmatrix} \end{aligned} \tag{3.35}$$

We make the simplifying assumption $E[\mathbf{G}] = \mathbf{0}$, introduce the assumption of normality, and treat the IBF model as a prior distribution

$$\mathbf{G} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_0) \tag{3.36}$$

Observing phenotypes $\mathbf{P} = \mathbf{p}$ and the IBF covariance matrix $\mathbf{V}$ of the corresponding individuals, inferred elsewhere, we can write the joint distribution

$$\begin{bmatrix} \mathbf{G} \\ \mathbf{P} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_0 & \mathbf{C} \\ \mathbf{C}' & \mathbf{V} \end{bmatrix}\right) \tag{3.37}$$

and the conditional distribution

$$\mathbf{G}|\mathbf{P} = \mathbf{p} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.38}$$

$$\boldsymbol{\mu} = \mathbf{C}\mathbf{V}^{-1}\mathbf{p} \tag{3.39}$$

$$\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0 - \mathbf{C}\mathbf{V}^{-1}\mathbf{C}' \tag{3.40}$$

We've just introduced $\mathbf{C}$, the matrix of covariances between genotypic values and phenotypes. The covariances are due only to phenotyped individual's genotypes at the same locus. The covariance of a genotypic value can most generally be written, in terms of the identity state array, as

$$\mathbf{C}_{i(g)} = \sigma_l^2 \sum_{k=0}^{4} T_{ijkl} c_{kl} \tag{3.41}$$

where i is the phenotyped individual, and j is a hypothetical individual with genotype g at locus l.

3.4.2 Phenotype Prediction

The marginal distribution of the phenotype of an individual with known genotype has the form of the prior,

$$\mathbf{P}_{new} \sim N\left(0, \Sigma_0\right) \text{ where } \Sigma_0 = \mathrm{Cov}(\mathbf{P}_{new}, \mathbf{P}_{new}) = \sigma_E^2 + \sum_{k=3}^{4}\sum_{l=1}^{L} C_{kl} T_{ijkl} \tag{3.42}$$

The joint distribution of the "new" individual's unknown phenotypic value $\mathbf{P}_{new}$ and the known phenotypes $\mathbf{P}$ of the training set of individuals is

$$\begin{bmatrix} \mathbf{P}_{new} \\ \mathbf{P} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_0 & \mathbf{C} \\ \mathbf{C}' & \mathbf{V} \end{bmatrix} \right) \tag{3.43}$$

where the covariances in $\mathbf{C}$ and $\mathbf{V}$ are constructed using equation (3.27). The predicted value of $\mathbf{P}_{new}$ has the conditional distribution

$$\mathbf{P}_{new} | \mathbf{P} = \mathbf{p} \sim N\left(\mu, \Sigma\right) \tag{3.44}$$

$$\mu = \mathbf{C}\mathbf{V}^{-1}\mathbf{p} \tag{3.45}$$

$$\Sigma=\Sigma_0 - \mathbf{C}\mathbf{V}^{-1}\mathbf{C}' \tag{3.46}$$

Like other inferences in this chapter, this conditional distribution is also conditional on fixed values of the IBF parameters.

For the evaluation of the performance of the prediction model, we perform leave-one-out cross validation. That is, given n individuals with known phenotypes $\mathbf{P}$ and genotypes, we

treat each phenotype as unknown in turn, and predict its conditional mean and standard deviation given the other $n-1$ phenotypes. Collecting these means and standard deviations into the vectors $\mathbf{P}^*$ and $\boldsymbol{\sigma}^* = [\sigma_i^*]$, we evaluate the predictive performance via the ordinary least squares regression

$$\mathbf{P} = \alpha\ \mathbf{1} + \beta \mathbf{P}^* + \boldsymbol{\epsilon} \tag{3.47}$$

We expect the prediction $\mathbf{P}^*$ to be an unbiased estimator of $\mathbf{P}$, implying $\alpha = 0$ and $\beta = 1$. The R^2 of this regression is a measure of prediction quality. Makowsky et al. (2011) discuss the relationship between predictive R^2 and heritability, and we refer to this topic elsewhere, e.g. in Section 2.5.2.

3.4.3 Effect and Breeding Value Inference

From equation (3.38) we can directly write the conditional means and variances of the individual genotypic values at a locus. In order to reconstruct the additive effects α_i and dominance deviations δ_{ij}, we need to introduce some of the classical machinery of population genetics, in particular allele frequencies.

For a fixed number of alleles (four for illustration purposes) the gene incidence/design matrix X can be constructed. The columns correspond to alleles, and the rows correspond to diploid genotypes. The entries are the counts (0, 1, or 2) of the allele in the corresponding genotype:

$$\mathbf{X} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{3.48}$$

Our subsequent calculations will be with respect to allele and genotype frequencies. For illustration we will take allele frequencies f_A, f_B, f_C, f_D and assume Hardy-Weinberg equilibrium, but, elsewhere, equilibrium is not a model assumption unless explicitly stated.

The frequencies enter the calculation through the diploid weight matrix, with the diagonal weight terms summing to 1:

$$\mathbf{W} = \begin{bmatrix} f_A f_A & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & f_B f_B & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & f_C f_C & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & f_D f_D & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2f_A f_B & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2f_A f_C & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2f_A f_D & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2f_B f_C & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2f_B f_D & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2f_C f_D \end{bmatrix} \tag{3.49}$$

The matrix representation of weighted regression allows a compact formulation of the classical Fisher model. The regression coefficient matrix maps genotypic values to regression coefficients,

$$reg_{\mathbf{W}}(\mathbf{X}) = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W} \tag{3.50}$$

The hat matrix projects to fitted values:

$$hat_{\mathbf{W}}(\mathbf{X}) = \mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W} \tag{3.51}$$

Residual matrix is the complement of the hat matrix:

$$res_{\mathbf{W}}(\mathbf{X}) = \mathbf{I} - hat_{\mathbf{W}}(\mathbf{X}) \tag{3.52}$$

The $\boldsymbol{\alpha}$ values are found by removing the grand mean, and then performing the regression onto the X design matrix:

$$\boldsymbol{\alpha} = reg_{\mathbf{W}}(\mathbf{X})\, res_{\mathbf{W}}(\mathbf{1})\, \mathbf{G} \tag{3.53}$$

We can decompose the genotypic values into the grand mean M, the additive component A, and the dominance deviation D:

$$\mathbf{G} = \mathbf{M} + \mathbf{A} + \mathbf{D} \tag{3.54}$$

$$\mathbf{M} = hat_{\mathbf{W}}(\mathbf{1})\, \mathbf{G} \tag{3.55}$$

$$\mathbf{A} = hat_{\mathbf{W}}(\mathbf{X})\, res_{\mathbf{W}}(\mathbf{1})\, \mathbf{G} \tag{3.56}$$

$$\mathbf{D} = res_{\mathbf{W}}(\mathbf{X})\, res_{\mathbf{W}}(\mathbf{1})\, \mathbf{G} \tag{3.57}$$

Thus $\boldsymbol{\alpha}$ and $\mathbf{D}$ correspond to the vectors of additive effects and dominance deviations, respectively. Both have the form $\mathbf{v} = \mathbf{MG}$, and since $\mathbf{G} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can compute the means and variances of each effect:

$$\mathbf{v} \sim N(\mathbf{M}\boldsymbol{\mu}, \mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}) \tag{3.58}$$

Breeding value prediction can be performed by summing the α estimates across the genome.

3.4.4 Heritability Reconstruction

We have written the decomposition of the genotypic value into grand mean, additive, and dominance terms. The classical additive and dominance variances can be written as the population variances of the corresponding terms, with the appropriate population weights. The variances are:

$$\sigma_A^2 = \mathbf{A}'\mathbf{W}\mathbf{A} = \mathbf{G}' \left[res'_{\mathbf{W}}(\mathbf{1})\, hat'_{\mathbf{W}}(\mathbf{X}) \;\; \mathbf{W} \; hat_{\mathbf{W}}(\mathbf{X})\, res_{\mathbf{W}}(\mathbf{1}) \right] \mathbf{G} = \mathbf{G}'\mathbf{S}_A\mathbf{G} \tag{3.59}$$

$$\sigma_D^2 = \mathbf{D}'\mathbf{W}\mathbf{D} = \mathbf{G}' \left[res'_{\mathbf{W}}(\mathbf{1})\, res'_{\mathbf{W}}(\mathbf{X}) \;\; \mathbf{W} \; res_{\mathbf{W}}(\mathbf{X})\, res_{\mathbf{W}}(\mathbf{1}) \right] \mathbf{G} = \mathbf{G}'\mathbf{S}_D\mathbf{G} \tag{3.60}$$

Both are quadratic forms in G, and positive semidefinite by symmetry.

For the simplest biallelic case, under Hardy-Weinberg equilibrium:

$$\mathbf{X} = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \end{bmatrix} \tag{3.61}$$

$$\mathbf{W} = \begin{bmatrix} p^2 & 0 & 0 \\ 0 & 2\,p\,(1-p) & 0 \\ 0 & 0 & (1-p)^2 \end{bmatrix} \tag{3.62}$$

We can derive the regression matrices, using a symbolic algebra program,

$$hat_{\mathbf{W}}(\mathbf{1}) = \begin{bmatrix} p^2 & -2\,p\,(p-1) & (p-1)^2 \\ p^2 & -2\,p\,(p-1) & (p-1)^2 \\ p^2 & -2\,p\,(p-1) & (p-1)^2 \end{bmatrix} \tag{3.63}$$

$$res_{\mathbf{W}}(\mathbf{1}) = \begin{bmatrix} -p^2+1 & 2\,p\,(p-1) & -(p-1)^2 \\ -p^2 & 2\,p^2-2\,p+1 & -(p-1)^2 \\ -p^2 & 2\,p\,(p-1) & 2\,p-p^2 \end{bmatrix} \tag{3.64}$$

$$hat_{\mathbf{W}}(\mathbf{X}) = \begin{bmatrix} -(p-2)\,p & 2\,(p-1)^2 & -(p-1)^2 \\ -p\,(p-1) & 2\,p^2-2\,p+1 & -p\,(p-1) \\ -p^2 & 2\,p^2 & -p^2+1 \end{bmatrix} \tag{3.65}$$

$$res_{\mathbf{W}}(\mathbf{X}) = \begin{bmatrix} 1-2\,p+p^2 & -2\,(p-1)^2 & (p-1)^2 \\ p\,(p-1) & -2\,p\,(p-1) & p\,(p-1) \\ p^2 & -2\,p^2 & p^2 \end{bmatrix} \tag{3.66}$$

From these, we can derive the quadratic forms $\mathbf{S}_A$ and $\mathbf{S}_D$ corresponding to the additive and dominance variances:

$$\mathbf{S}_A = 2p\,(1-p) \begin{bmatrix} p^2 & -p\,(2\,p-1) & p\,(p-1) \\ -p\,(2\,p-1) & (2\,p-1)^2 & -(p-1)\,(2\,p-1) \\ p\,(p-1) & -(p-1)\,(2\,p-1) & (p-1)^2 \end{bmatrix} \tag{3.67}$$

$$\mathbf{S}_D = p^2\,(1-p)^2 \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} \tag{3.68}$$

By equations 2.26 and 2.27, the expectation of a quadratic form in a random vector is a sum of mean and covariance terms, and under the assumption of normality, so is the variance:

$$\mathrm{E}\left[\mathbf{G}'\mathbf{Q}\mathbf{G}\right] = \boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu} + \mathrm{tr}\left[\mathbf{Q}\boldsymbol{\Sigma}\right]$$

$$\mathrm{Var}\left[\mathbf{G}'\mathbf{Q}\mathbf{G}\right] = 4\boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\mu} + 2\mathrm{tr}\left[\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\Sigma}\right]$$

Since

$$\boldsymbol{\mu} = \mathbf{C}\mathbf{V}^{-1}\mathbf{p} \tag{3.69}$$

$$\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0 - \mathbf{C}\mathbf{V}^{-1}\mathbf{C}' \tag{3.70}$$

we can write

$$\mathrm{E}\left[\mathbf{G}'\mathbf{Q}\mathbf{G}\right] = \mathbf{p}'\mathbf{V}^{-1}\mathbf{C}'\mathbf{Q}\mathbf{C}\mathbf{V}^{-1}\mathbf{p} + \mathrm{tr}\left[\mathbf{Q}\boldsymbol{\Sigma}_0\right] - \mathrm{tr}\left[\mathbf{Q}\mathbf{C}\mathbf{V}^{-1}\mathbf{C}'\right] \tag{3.71}$$

Only the first term depends directly on phenotype values. In the Explained and Unexplained variance terminology of Section 2.6, the first term is the Explained variance, and the other two correspond to the Unexplained. The $\mathrm{tr}\left[\mathbf{Q}\boldsymbol{\Sigma}_0\right]$ can be interpreted as the expected variance under the prior, when no phenotype data is observed. As more individuals are observed, the Unexplained component decreases due to the third term, and the Explained component increases. With complete information, $\boldsymbol{\Sigma}$ vanishes and only the Explained component remains.

3.4.5 The Functional Genomic Relationship Matrix

When the amount of observed phenotype data is small, the prior dominates,

$$\mathrm{E}\left[\mathbf{G}'\mathbf{Q}\mathbf{G}\right] \approx \mathrm{tr}\left[\mathbf{Q}\boldsymbol{\Sigma}_0\right] \tag{3.72}$$

This expression can be evaluated at a locus, requiring only allele frequencies, if we assume the additive model. Let $\mathbf{Q}_l$ be the quadratic form for additive variance evaluated using allele frequencies at locus l. Let $\boldsymbol{\Sigma}^*_{0l}$ be the unscaled additive form of the IBF prior at locus l; that is, $\boldsymbol{\Sigma}_{0l} = \sigma^2_l\boldsymbol{\Sigma}^*_{0l}$ Then for a fixed set of weights, such as $w_l = 1$, the overall additive variance, summing over loci, is

$$\sigma^2_A = \sigma^2_U \sum_{l=0}^{L} w_l \mathrm{tr}\left[\mathbf{Q}_l\boldsymbol{\Sigma}^*_{0l}\right] \tag{3.73}$$

We can use this relationship to find σ^2_U given σ^2_A. Given σ^2_U, the weights w_l, and the assumption of the additive model, we can construct the complete IBF phenotype matrix. The functional Genomic Relationship Matrix (fGRM) is defined by analogy to the classical relationship matrix, for which

$$\mathrm{Cov}(\mathbf{P}) = \sigma^2_A\mathbf{A} + \sigma^2_E\mathbf{I} \tag{3.74}$$

To create the appropriately scaled $\mathbf{A}$ from the IBF covariance matrix, we define

$$\mathbf{A}_{ij} = \frac{\sum_{k=1}^{4} \sum_{l=1}^{L} C_{kl} T_{ijkl}}{\sum_{l=0}^{L} w_l \text{ tr} \left[\mathbf{Q}_l \mathbf{\Sigma}_{0l}^* \right]} \tag{3.75}$$

For the biallelic SNP, Hardy-Weinberg equilibrium case, we can derive a compact closed form for these expressions. Equation 3.67 gives the expression for $\mathbf{Q}_l$ using SNP frequencies.

Under the additive model,

$$\mathbf{\Sigma}_{0l}^* = \begin{bmatrix} 1 & 1/2 & 0 \\ 1/2 & 1/2 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix} \sigma_{AA}^2 \tag{3.76}$$

Using symbolic algebra software,

$$\text{tr} \left[\mathbf{Q}_l \mathbf{\Sigma}_{0l}^* \right] = \sigma_{AA}^2 \; p \, (1 - p) \tag{3.77}$$

Under the random order strict dominance model,

$$\mathbf{\Sigma}_{0l}^* = \begin{bmatrix} 1 & 1/2\sqrt{2} & 0 \\ 1/2\sqrt{2} & 1 & 1/2\sqrt{2} \\ 0 & 1/2\sqrt{2} & 1 \end{bmatrix} \sigma_{AA}^2 \tag{3.78}$$

Again using symbolic algebra software,

$$\text{tr} \left[\mathbf{Q}_l \mathbf{\Sigma}_{0l}^* \right] = \sigma_{AA}^2 \left(6 - 4\sqrt{2} \right) \left(2\,p^2 - 2\,p + \sqrt{2} + 2 \right) p \, (1 - p) \tag{3.79}$$

The two expressions are numerically very similar, as demonstrated in Figure 3.1. The scaling factors have a range from 0 to 0.25. If p is distributed $Uniform(0, 1)$ or $Uniform(0, 1/2)$ over loci, the average scaling factor is $1/6 \approx 0.167$.

3.5 Summary

In this chapter, we introduced the state identity array, a method for representing genetic similarity regardless of the specific genotype, as well as several types of genotype ambiguity. We applied the method to several missing or ambiguous genotype situations. Within the

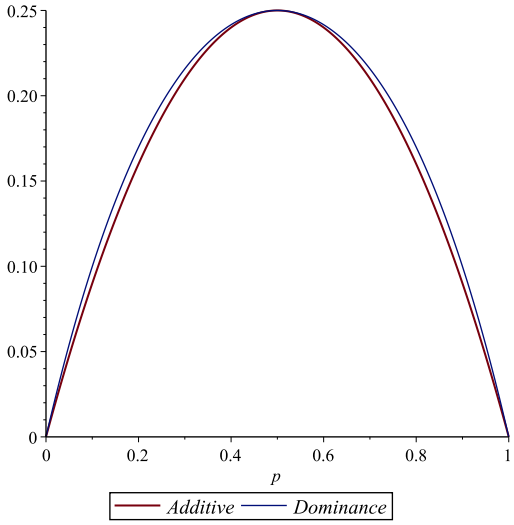


Figure 3.1: fGRM scaling factor under the Additive and Random Order Strict Dominance models.

context of such ambiguity, the dictionary method is an approach to imputation and phasing for inferring IBF genotypes based on a population model. We described our approach to classifying functional genome regions and alleles. We develop tools for phenotype and breeding value prediction and effect estimation, analogous to their classical counterparts. The method for bottom-up heritability construction is an implementation of the classical definition of heritability; but in our context, the genotypic effects are random rather than fixed, leading to the Explained/Unexplained decomposition.

Chapter 4

VARIANCE COMPONENT ESTIMATION

4.1 Introduction

In classical quantitative genetics a distinction is made between fixed effect estimation and random effect prediction given variance components, and variance component inference. Characteristically, a prominent textbook separates its treatment of general matrix-based methods into *Estimation of Breeding Values* (Lynch and Walsh, 1998, Chapter 26) and *Variance-component Estimation with Complex Pedigrees* (Lynch and Walsh, 1998, Chapter 27).

The former class of methods involves inference about individual (e.g. animal, gamete) effects, and is often referred to by the optimality properties of its estimators, BLUP (Best Linear Unbiased Predictor) and BLUE (Best Linear Unbiased Estimator). The variance components, such as additive and environmental variance, or their functions, such as heritability, typically enter the estimator equations in a nonlinear fashion. Therefore, if we are simultaneously estimating the variance components and using them to estimate individual effects, the estimator is no longer linear, and neither necessarily unbiased or best. Variance components must be estimated separately, by methods of the latter class, typically by fitting an ML or REML likelihood.

Our method is distinguished from the classical framework by the interpretation of the probabilities, as well as by the large number of variance components. The number of variance components scales with the number of genes, typically creating an underdetermined ($p \gg n$) system in which the variance components themselves are not necessarily identifiable. This is in contrast to the more common case, encountered e.g. in Yang et al. (2010), where a single variance component serves as a hyperparameter governing the regularization of a number of random effects.

Two approaches to fitting such a large number of variance components are feasible. In a

fully Bayesian treatment, we would impose a prior on the variance components themselves. Alternatively, we can find a single representative set of variance components. The fully Bayesian treatment is superior in providing an interpretable measure for the uncertainty about variance, but is not computationally feasible when prediction with a single set of variance components is computationally expensive.

We will first describe the Tipping-Faul approach to mass variance component estimation, which is the regression equivalent to our problem. Then, we use the Tipping-Faul tools to show that the simplest, univariate case of their regression has a closed form. We will use this closed form to initialize the search for a high likelihood solution. Returning from the special case of regressions to the general variance component problem, we derive the incremental update procedure, an optimization method based on matrix decomposition. Then we describe how these procedures can be combined to generalize the Tipping-Faul method to arbitrary variance components, and apply it to our specific problem.

4.2 Automatic Relevance Determination and the Tipping-Faul Approach

The most analogous well-studied underdetermined problem with a large number of variance components is the Relevance Vector Machine (Tipping, 2001; Faul and Tipping, 2001) relying on the Automatic Relevance Determination prior and methodology (Neal, 1995; MacKay, 1996).

The core of the method is a local optimization algorithm for a likelihood expressed in the variance components, typically providing a highly sparse solution with most variance components set to zero. Recent work on regressions using this approach emphasizes the multiplicity of local optima, except in highly specialized cases (Wipf and Rao, 2004; Wipf and Nagarajan, 2007).

4.2.1 Historical Context

The Relevance Vector Machine was described by Tipping (2001) as a technique competing with Support Vector Machines in the classification and regression settings, producing sparse linear kernel-based classifiers and predictors. The methodology developed for this purpose is directly applicable in the non-kernel sparse regression and classification settings.

The literature sometimes informally interchanges the terms Relevance Vector Machine (RVM), Automatic Relevance Determination (ARD), and Sparse Bayesian Learning (SBL). **Sparse Bayesian Learning** is the broadest of the terms. Generally we do not think of the Bayesian setting as conducive to identifying sparse solutions; if we have a prior, over e.g. regression coefficients, that allows sparse solutions, the posterior would typically assign some probability to each possible sparse solution, as well as to possible non-sparse solutions. At some point, optimization is needed to choose which parameters to cut to induce sparsity.

A hierarchical Bayesian model for data $\mathbf{x}$ has the likelihood

$$\Pr(\mathbf{x}) = \Pr(\mathbf{x}|\theta)\ \Pr(\theta|h)\ \Pr(h) \tag{4.1}$$

with a prior on θ determined by hyperparameters h. Rather than compute a posterior $\Pr(h|\mathbf{x})$ it is possible to choose a single h^* to maximize $\Pr(h|\mathbf{x})$. From a Bayesian perspective this is an approximation valid when the posterior distribution of h is close to degenerate; from a frequentist perspective the model is converted to the maximum likelihood setting parametrized by h. In the statistics literature this is Empirical Bayes or Objective Bayes, and in the machine learning literature Type 2 Maximum Likelihood or Evidence Approximation.

Given fixed hyperparameters h, we still have a Bayesian model of the form

$$\Pr(\mathbf{x}) = \Pr(\mathbf{x}|\theta)\ \Pr(\theta|h) \tag{4.2}$$

When the θ are regression coefficients or weights, sparsity is generated when the choice of h creates restrictions on the possible values of θ, fixing some weights to zero with probability 1. Optimizing over h can thus create a Bayesian model constrained over a limited, sparse set of coefficients.

Automatic Relevance Determination (ARD) is the specific class of prior and models to which the algorithms we will describe apply. The term was used for the improper prior used in a neural network setting by MacKay (1994), and to describe the sparse regression framework based on that prior by Wipf and Nagarajan (2007), common to all RVM variants.

The ARD prior for a regression coefficient $w_i \sim N(0, \alpha_i^{-1})$ is an improper flat hyperprior over $\ln \alpha_i$. Both Tipping (2001) and Bishop (2003) derive the improper logarithmically uniform prior as a limit of the Gamma precision prior, and justify it on the basis on invariance

to scale. The ARD method, as distinct from the ARD prior, is the application of the ARD prior to a set of regression coefficients, in the context of a regression or classification problem, and the use of Sparse Bayesian Learning to estimate point values of hyperparameters in the context of that prior. Put together, these generate sparse regression solutions by pushing some $\alpha_i \to \infty$ and therefore imposing the constraint $w_i = 0$.

The **Relevance Vector Machine**, as introduced by Tipping (2001), is a particular application of the Automatic Relevance Determination procedure. The important algorithms for ARD/SBL have been developed in the RVM context, but strictly speaking RVM is a special case of ARD. ARD can be interpreted as a regression onto an arbitrary set of features, and the feature vectors are composed of kernel functions evaluated between pairs of data vectors in the training set.

The original RVM papers have also introduced a classification version of the ARD regression, by applying a Laplace approximation of a logistic likelihood to construct a proxy for the Gaussian likelihood to which the ARD algorithm applies.

4.2.2 Regression Model

In the regression setting, we choose regression coefficients or weights $\mathbf{w}$ to fit the equation

$$\mathbf{y}=\boldsymbol{\phi}\mathbf{w}+\boldsymbol{\epsilon} \tag{4.3}$$

The dimensions of this equation are $N \times 1 = (N \times M)(M \times 1) + (N \times 1)$. There are N observations, and M basis vectors denoted as $\boldsymbol{\phi}_i(x_j)$ which we can think of as M functions extracting the i-th feature from the j-th input vector x_j, and M weights w_i, many of which can be zero in a sparse solution. In scalar notation,

$$y_j = \sum_{i=0}^{M} \boldsymbol{\phi}_i(x_j) w_i + \epsilon_j \tag{4.4}$$

Typically, but not necessarily, $\boldsymbol{\phi}_0(x_j) = 1$ and w_0 is referred to as the bias.

The weights and the error terms are given independent prior distributions with hyperparameters,

$$w_i \sim N\left(0, \alpha_i^{-1}\right) \tag{4.5}$$

$$\epsilon_j \sim N\left(0, \sigma^2\right) \tag{4.6}$$

An individual α_i is allowed to be notionally $+\infty$, implying $w_i \sim N\left(0, 0\right)$ so $w_i = 0$ with probability 1; this is the case that imposes a sparsity constraint. For fixed hyperparameters α_i and σ^2, we can write the joint distribution of the data $\mathbf{y}$ using $\mathbf{A}$= diag $(\alpha_1 \ldots \alpha_M)$:

$$\mathbf{y} \sim N\left(\mathbf{0}, \boldsymbol{\phi}\mathbf{A}^{-1}\boldsymbol{\phi}' + \sigma^2\mathbf{I}\right) \tag{4.7}$$

Given observed data $\mathbf{y}$=$\mathbf{t}$, for $\mathbf{C} = \boldsymbol{\phi}\mathbf{A}^{-1}\boldsymbol{\phi}' + \sigma^2\mathbf{I}$ this gives the multivariate normal log likelihood function, up to a constant c,

$$L\left(\alpha_1 \ldots \alpha_M, \sigma^2\right) = c - \frac{1}{2}\left[\ln \det \mathbf{C} + \mathbf{t}'\mathbf{C}^{-1}\mathbf{t}\right] \tag{4.8}$$

The learning procedure is to find the hyperparameters α_i and σ^2 that maximize L.

4.2.3 Overall Optimization Strategy

Given a likelihood in any of the above forms, the goal of the optimization is to iteratively update $\alpha_1 \ldots \alpha_M, \sigma^2$, improving the likelihood until convergence. There are three possibilities: the regression case where we optimize over $\alpha_1 \ldots \alpha_M$ given a fixed σ^2, the regression case where we optimize over both $\alpha_1 \ldots \alpha_M$ and σ^2, and the classification case, where we optimize over $\alpha_1 \ldots \alpha_M$ and do not have an explicit σ^2, but the Laplace approximation to the likelihood function changes.

We will repeatedly refer to feature vectors being added to or removed from the basis. A vector is in the basis if its corresponding $\alpha_i < \infty$, and therefore its w_i estimate is not forced to zero. Each update step changes one of the α_i, possibly changing an α_i to or from ∞ and thereby adding or removing a feature vector from the basis.

4.2.4 EM and MacKay Updates

We briefly describe the two original ARD optimization algorithms first demonstrated in Tipping (2001), superseded by the algorithm described by Tipping and Faul (2003).

EM Algorithm The EM procedure is discussed in Tipping (2001) and dismissed as slow. The w_i and σ^2 are treated as latent variables, and at each iteration, are set to their expected values given the observed data and previous versions of the parameter estimates. The α_i are then maximized using these latent variables.

McKay Fixed Point Updates The fixed point procedure used by Tipping (2001) follows ideas developed in a similar context by MacKay (1994). Differentiating the likelihood gives equations for the optimal values of the α_i and σ^2 which are of course not solvable analytically. The equations can be rearranged, expressing the α_i and σ^2 as a function of α_i and σ^2; the optimal values are a fixed point of this function. The function can be iterated from some initial values, and if the rearrangement is chosen in a way that happens to converge to the fixed point, the optimal values will be found.

Convergence Considerations Wipf and Nagarajan (2007) point out that both of these algorithms have fixed points at any $\alpha_i = +\infty$, whether or not it is a local optimum. The implementation of both algorithms by Tipping (2001) uses the heuristic of detecting an α_i that tends to $+\infty$ as it exceeds an arbitrary numerical threshold, and permanently removing the corresponding $\boldsymbol{\phi}_i$ from the basis. Thus, once removed from the basis, a feature cannot return.

No global optimality guarantees are available for any of these algorithms. The EM algorithm has the property of converging to a local maximum due to the classical KL divergence argument. The MacKay fixed point procedure, though described as faster in practice, is not proven to converge to a local maximum.

4.2.5 Tipping/Faul Updates

This procedure is described by Faul and Tipping (2001) and Tipping and Faul (2003), and sometimes referred to as the "Fast Tipping Update."

Given $\alpha_1 \dots \alpha_M, \sigma^2$, this algorithm chooses an i (by a round-robin, or random, or some other process) and finds the optimal α_i holding all other α_j and σ^2 fixed. This guarantees improvement of the likelihood at each step. Faul and Tipping (2001) establish that at any

joint maximum over all α, the Hessian with respect to the α is negative semi-definite. This ensures that the algorithm converges to a local maximum with respect to $\alpha_1 \ldots \alpha_M$, so long as σ^2 is fixed. However, if the σ^2 is reestimated, at each iteration or less frequently, there is no explicit convergence guarantee.

The algorithm proceeds by the decomposition of the likelihood's primary covariance matrix:

$$\mathbf{C} = \sigma^2\mathbf{I} + \sum_m \alpha_m^{-1}\boldsymbol{\phi}_m\boldsymbol{\phi}_m' = \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i' \tag{4.9}$$

where

$$\mathbf{C}_{-i} = \sigma^2\mathbf{I} + \sum_{m\neq i} \alpha_m^{-1}\boldsymbol{\phi}_m\boldsymbol{\phi}_m' \tag{4.10}$$

Matrix identities (rank 1 updates to the determinant and inverse) are used to show:

$$\det \mathbf{C} = \left|1 + \alpha_i^{-1}\boldsymbol{\phi}_i'\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\right| \det \mathbf{C}_{-i} \tag{4.11}$$

$$\ln\det \mathbf{C} = \ln\det \mathbf{C}_{-i} + \ln\left|1 + \alpha_i^{-1}\boldsymbol{\phi}_i'\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\right| \tag{4.12}$$

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i'\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\phi}_i'\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i} \tag{4.13}$$

$$\mathbf{t}'\mathbf{C}^{-1}\mathbf{t} = \mathbf{t}'\mathbf{C}_{-i}^{-1}\mathbf{t} - \frac{\mathbf{t}'\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i'\mathbf{C}_{-i}^{-1}\mathbf{t}}{\alpha_i + \boldsymbol{\phi}_i'\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i} \tag{4.14}$$

Defining scalars

$$s_i = \boldsymbol{\phi}_i'\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i \tag{4.15}$$

$$q_i = \boldsymbol{\phi}_i'\mathbf{C}_{-i}^{-1}\mathbf{t} \tag{4.16}$$

we obtain the likelihood decomposition

$$L\left(\alpha_1 \ldots \alpha_M, \sigma^2\right) = c - \frac{1}{2}\left[\ln\det \mathbf{C} + \mathbf{t}'\mathbf{C}^{-1}\mathbf{t}\right] \tag{4.17}$$

$$= c - \frac{1}{2}\left[\ln\det \mathbf{C}_{-i} + \mathbf{t}'\mathbf{C}_{-i}^{-1}\mathbf{t} + \ln\left(1 + s_i/\alpha_i\right) - \frac{q_i^2}{\alpha_i + s_i}\right]$$

Thus to maximize with respect to α_i holding the other α_j and the σ^2 constant, we need only maximize the component

$$l\left(\alpha_i\right) = \frac{1}{2}\left[\ln\left(\frac{\alpha_i}{\alpha_i + s_i}\right) + \frac{q_i^2}{\alpha_i + s_i}\right] \tag{4.18}$$

This expression is optimized at $\alpha_i = +\infty$ when $s_i > q_i^2$, and otherwise at

$$\alpha_i = \frac{s_i^2}{q_i^2 - s_i} \tag{4.19}$$

The essential insight is that, holding all other α_j and the σ^2 constant, we can optimize with respect to α_i and thus ensure improvement of the likelihood. The optimization can set α_i to $+\infty$, removing $\boldsymbol{\phi}_i$ from the active basis, change an α_i that was previously $+\infty$ to a real number, adding $\boldsymbol{\phi}_i$ into the active basis, or it can merely change α_i from one real number to another.

Since $s_i = \boldsymbol{\phi}_i' \mathbf{C}_{-i}^{-1} \boldsymbol{\phi}_i$ and $q_i = \boldsymbol{\phi}_i' \mathbf{C}_{-i}^{-1} \mathbf{t}$ are not convenient to compute directly, we instead compute $S_i = \boldsymbol{\phi}_i' \mathbf{C}^{-1} \boldsymbol{\phi}_i$ and $Q_i = \boldsymbol{\phi}_i' \mathbf{C}^{-1} \mathbf{t}$ including the current value of α_i, then back out $s_i = \alpha_i S_i / (\alpha_i - S_i)$ and $q_i = \alpha_i Q_i / (\alpha_i - S_i)$.

Further methods for quickly computing q_i and s_i without full matrix operations are derived by Tipping and Faul (2003).

The estimation of σ^2 proceeds using the MacKay update

$$\sigma^2 = \frac{|\mathbf{t} - \boldsymbol{\phi}\boldsymbol{\mu}|^2}{N - \sum_{i=1}^{M^+} 1 + \alpha_i \boldsymbol{\Sigma}_{ii}} \tag{4.20}$$

The summation is over the M^+ elements of the basis for which $\alpha_i < \infty$. Note that the calculation of $\boldsymbol{\Sigma}$ in the regression case uses the previous value of σ^2. Thus this is justified as a fixed point equation for the optimal σ^2 given all α_i, but does not carry a guarantee of improving the likelihood.

4.2.6 Differences Between the Update Procedures

The Tipping/Faul update adds or removes features by an unambiguous optimization decision, rather than the heuristic of a number exceeding an arbitrary threshold, and can follow a broader range of search paths since it can add, and not only remove, features. The EM and MacKay fixed point approaches must start with all features in the basis, and gradually remove features. Crucially, the Tipping/Faul approach can be initialized with a small basis, adding features one at a time, rather than with a complete basis, removing features one at a time. The most burdensome calculations at each iteration are inversions of matrices scaling

with the size of the basis. If the final basis is sparse, and M is high, the Tipping/Faul approach allows all calculations to be done with small basis-scale matrices rather than $M \times M$ matrices.

4.3 Standalone Estimation of Variance Components

Problem Formulation We extend the Tipping/Faul method to find the variance components for the single regressor case analytically. That is, we want a closed form solution for the simplest random coefficient regression,

$$\mathbf{y} = w\boldsymbol{\phi} + \boldsymbol{\epsilon} \tag{4.21}$$

$$w \sim N\left(0, \alpha^{-1}\right) \text{ or } N\left(0, \sigma_{\boldsymbol{\phi}}^2\right) \tag{4.22}$$

$$\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \tag{4.23}$$

yielding a variance component α^{-1} or $\sigma_{\boldsymbol{\phi}}^2$, together with the residual variance σ^2. As in the Faul/Tipping update process in the previous section, we can form the complete and incomplete covariance matrices:

$$\mathbf{C} = \sigma^2\mathbf{I} + \alpha^{-1}\boldsymbol{\phi}\boldsymbol{\phi}'$$

$$\mathbf{C}_{-i} = \sigma^2\mathbf{I}$$

$$\mathbf{C}_{-i}^{-1} = \sigma^{-2}\mathbf{I}$$

Solving for One Variance Component given Residual Variance The two Faul/Tipping optimization factors are :

$$q = \boldsymbol{\phi}'\mathbf{C}_{-i}^{-1}\mathbf{t} = \left(\sigma^2\right)^{-1}\boldsymbol{\phi}'\mathbf{t}$$

$$s = \boldsymbol{\phi}'\mathbf{C}_{-i}^{-1}\boldsymbol{\phi} = \left(\sigma^2\right)^{-1}\boldsymbol{\phi}'\boldsymbol{\phi}$$

giving the solution

$$\alpha = \frac{s^2}{q^2 - s} = \frac{\left(\sigma^2\right)^{-2}(\boldsymbol{\phi}'\boldsymbol{\phi})^2}{(\sigma^2)^{-2}(\boldsymbol{\phi}'\mathbf{t})^2 - (\sigma^2)^{-1}\boldsymbol{\phi}'\boldsymbol{\phi}} = \frac{(\boldsymbol{\phi}'\boldsymbol{\phi})^2}{(\boldsymbol{\phi}'\mathbf{t})^2 - (\sigma^2)\boldsymbol{\phi}'\boldsymbol{\phi}}$$

when this expression is positive positive, and $\alpha = +\infty$ or $\sigma_{\boldsymbol{\phi}}^2 = 0$ otherwise. Thus the variance component of single effect given σ^2:

$$\sigma_{\boldsymbol{\phi}}^2 = \frac{1}{\alpha} = \frac{(\boldsymbol{\phi}'\mathbf{t})^2}{(\boldsymbol{\phi}'\boldsymbol{\phi})^2} - \frac{\sigma^2}{\boldsymbol{\phi}'\boldsymbol{\phi}} \tag{4.24}$$

Determinant and Inverse Expressions Solving simultaneously for σ_{ϕ}^2 and σ^2 requires a different approach. Applying the Matrix Determinant Lemma to covariance matrix $\mathbf{C}$:

$$\det\left(\mathbf{A}+\mathbf{u}\mathbf{v}'\right)=\left(1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}\right)\det\left(\mathbf{A}\right) \tag{4.25}$$

where

$$\mathbf{C}=\sigma^2\mathbf{I}+\sigma_{\phi}^2\boldsymbol{\phi}\boldsymbol{\phi}'$$

$$\mathbf{A}=\sigma^2\mathbf{I}$$

$$\mathbf{u}\mathbf{v}'=\sigma_{\phi}^2\boldsymbol{\phi}\boldsymbol{\phi}'$$

We obtain the determinant,

$$\det\mathbf{C}=\left(1+\frac{\sigma_{\phi}^2}{\sigma^2}\boldsymbol{\phi}'\boldsymbol{\phi}\right)\det\left(\sigma^2\mathbf{I}\right)\ =\left(1+\frac{\sigma_{\phi}^2}{\sigma^2}\boldsymbol{\phi}'\boldsymbol{\phi}\right)\left(\sigma^2\right)^n \tag{4.26}$$

Similarly, applying Sherman-Morrison formula for matrix inverses,

$$\left(\mathbf{A}+\mathbf{u}\mathbf{v}'\right)^{-1}=\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}$$

where

$$\mathbf{C}=\sigma^2\mathbf{I}+\sigma_{\phi}^2\boldsymbol{\phi}\boldsymbol{\phi}'$$

$$\mathbf{A}=\sigma^2\mathbf{I}$$

$$\mathbf{u}\mathbf{v}'=\sigma_{\phi}^2\boldsymbol{\phi}\boldsymbol{\phi}'$$

$$\mathbf{A}^{-1}=\frac{1}{\sigma^2}\mathbf{I}$$

$$\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}=\frac{\sigma_{\phi}^2}{\sigma^2}\boldsymbol{\phi}'\boldsymbol{\phi}$$

We obtain the inverse,

$$\mathbf{C}^{-1}=\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}=\frac{1}{\sigma^2}\mathbf{I}-\frac{\frac{1}{\sigma^2}\mathbf{I}\sigma_{\phi}^2\boldsymbol{\phi}\boldsymbol{\phi}'\frac{1}{\sigma^2}\mathbf{I}}{1+\frac{\sigma_{\phi}^2}{\sigma^2}\boldsymbol{\phi}'\boldsymbol{\phi}}=\frac{1}{\sigma^2}\left[\mathbf{I}-\frac{\sigma_{\phi}^2}{\sigma^2+\sigma_{\phi}^2\left(\boldsymbol{\phi}'\boldsymbol{\phi}\right)}\boldsymbol{\phi}\boldsymbol{\phi}'\right]$$

Likelihood for both Variance Terms We seek to opimize a likelihood of the form:

$$\ln L \ = -\frac{1}{2}\left[k + \ln\det \mathbf{C} \ + \mathbf{t}'\mathbf{C}^{-1}\mathbf{t}\right]$$

Substituting in the derived expressions for $\det \mathbf{C}$ and $\mathbf{C}^{-1}$,

$$\mathbf{t}'\mathbf{C}^{-1}\mathbf{t} = \frac{1}{\sigma^2}\left[\mathbf{t}'\mathbf{t} - \frac{\sigma_{\boldsymbol{\phi}}^2}{\sigma^2 + \sigma_{\boldsymbol{\phi}}^2\left(\boldsymbol{\phi}'\boldsymbol{\phi}\right)}\left(\mathbf{t}'\boldsymbol{\phi}\right)^2\right]$$

$$\ln L \ = -n\ln\left(\frac{1}{\sigma^2}\right) \ + \ln\left(1 + \frac{\sigma_{\boldsymbol{\phi}}^2}{\sigma^2}\boldsymbol{\phi}'\boldsymbol{\phi}\right) \ + \frac{1}{\sigma^2}\mathbf{t}'\mathbf{t} - \frac{1}{\sigma^2}\frac{\sigma_{\boldsymbol{\phi}}^2}{\sigma^2}\frac{\left(\mathbf{t}'\boldsymbol{\phi}\right)^2}{1 + \frac{\sigma_{\boldsymbol{\phi}}^2}{\sigma^2}\left(\boldsymbol{\phi}'\boldsymbol{\phi}\right)} \tag{4.27}$$

Reparametrizations We will optimize this with the parametrization

$$a = \frac{1}{\sigma^2}$$

$$b = \frac{\sigma_{\boldsymbol{\phi}}^2}{\sigma^2}$$

which allows us to write

$$\ln L \ = -n\ln a \ + \ln\left(1 + b\boldsymbol{\phi}'\boldsymbol{\phi}\right) \ + a\mathbf{t}'\mathbf{t} - ab\frac{\left(\mathbf{t}'\boldsymbol{\phi}\right)^2}{1 + b\left(\boldsymbol{\phi}'\boldsymbol{\phi}\right)}$$

or, using the notation $|\mathbf{t}|^2 = \mathbf{t}'\mathbf{t}$, $|\boldsymbol{\phi}|^2 = \boldsymbol{\phi}'\boldsymbol{\phi}$, and $r = \mathbf{t}'\boldsymbol{\phi}/\left|\mathbf{t}\right|\left|\boldsymbol{\phi}\right|$:

$$\ln L \ = -n\ln a \ + \ln\left(1 + b|\boldsymbol{\phi}|^2\right) \ + a|t|^2 - ab\frac{r^2|\boldsymbol{\phi}|^2|t|^2}{1 + b|\boldsymbol{\phi}|^2}$$

Optimization with respect to Single Parameters Optimizing the likelihood with respect to a:

$$\frac{\partial \ln L}{\partial a} = -\frac{n}{a} + |\mathbf{t}|^2 - b\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1 + b|\boldsymbol{\phi}|^2} = 0$$

$$\frac{1}{a} = \frac{1}{n}|\mathbf{t}|^2 - \frac{1}{n}b\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1 + b|\boldsymbol{\phi}|^2} \tag{4.28}$$

Optimizing the likelihood with respect to b:

$$\frac{\partial \ln L}{\partial b} = \frac{|\boldsymbol{\phi}|^2}{1 + b|\boldsymbol{\phi}|^2} - a\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1 + b|\boldsymbol{\phi}|^2} + ab|\boldsymbol{\phi}|^2\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{\left[1 + b|\boldsymbol{\phi}|^2\right]^2} = 0$$

$$|\boldsymbol{\phi}|^2 - ar^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2 + ab|\boldsymbol{\phi}|^2\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1+b|\boldsymbol{\phi}|^2} = 0$$

$$\frac{1}{a} = r^2|\mathbf{t}|^2 - b\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1+b|\boldsymbol{\phi}|^2} \tag{4.29}$$

Reconstructing the Tipping/Faul Solution from Matrix Approach If a is known:

$$\frac{1}{a} = r^2|\mathbf{t}|^2 - b\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1+b|\boldsymbol{\phi}|^2}$$

$$\frac{1}{ar^2|\mathbf{t}|^2} = 1 - \frac{b}{\frac{1}{|\boldsymbol{\phi}|^2}+b} = \frac{\frac{1}{|\boldsymbol{\phi}|^2}}{\frac{1}{|\boldsymbol{\phi}|^2}+b}$$

$$\frac{1}{|\boldsymbol{\phi}|^2} + b = \frac{ar^2|\mathbf{t}|^2}{|\boldsymbol{\phi}|^2}$$

$$b = \frac{ar^2|\mathbf{t}|^2 - 1}{|\boldsymbol{\phi}|^2}$$

$$\widehat{\sigma^2_{\boldsymbol{\phi}}} = \frac{b}{a} = \frac{r^2|\mathbf{t}|^2}{|\boldsymbol{\phi}|^2} - \frac{\frac{1}{a}}{|\boldsymbol{\phi}|^2} = \frac{r^2|\mathbf{t}|^2 - \sigma^2}{|\boldsymbol{\phi}|^2} \tag{4.30}$$

which is equivalent to equation 4.24.

Optimization with respect to Both Parameters We find the simultaneous optimum by combining equations 4.28 and 4.29. That is, in the two expressions derived from $\frac{\partial \ln L}{\partial a} = 0$ and $\frac{\partial \ln L}{\partial b} = 0$, we equate the terms equal to $1/a$, thus eliminating a. This allows us to solve for b:

$$\frac{1}{n}|\mathbf{t}|^2 - \frac{1}{n}b\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1+b|\boldsymbol{\phi}|^2} = r^2|\mathbf{t}|^2 - b\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1+b|\boldsymbol{\phi}|^2}$$

$$\left[\frac{1}{n} - r^2\right]|\mathbf{t}|^2 = \left(\frac{1}{n} - 1\right)b\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1+b|\boldsymbol{\phi}|^2}$$

$$\left[r^2 n - 1\right] = (n-1)\,b\frac{r^2|\boldsymbol{\phi}|^2}{1+b|\boldsymbol{\phi}|^2}$$

$$\frac{n - \frac{1}{r^2}}{n-1} = \frac{b|\boldsymbol{\phi}|^2}{1+b|\boldsymbol{\phi}|^2}$$

$$\frac{n-1}{n-\frac{1}{r^2}} = 1 + \frac{1}{b|\boldsymbol{\phi}|^2}$$

$$\left(\frac{n-1}{n-\frac{1}{r^2}} - 1\right)|\boldsymbol{\phi}|^2 = \frac{1}{b} = \left(\frac{n-1}{n-\frac{1}{r^2}} - \frac{n-\frac{1}{r^2}}{n-\frac{1}{r^2}}\right)|\boldsymbol{\phi}|^2 = \left(\frac{\frac{1}{r^2}-1}{n-\frac{1}{r^2}}\right)|\boldsymbol{\phi}|^2$$

$$b = \frac{1}{|\boldsymbol{\phi}|^2}\frac{n-\frac{1}{r^2}}{\frac{1}{r^2}-1} = \frac{1}{|\boldsymbol{\phi}|^2}\frac{nr^2-1}{1-r^2}$$

Substituting back, we can solve for $\sigma^2 = 1/a$:

$$\frac{1}{a} = \frac{1}{n}|\mathbf{t}|^2 - \frac{1}{n}b\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1+b|\boldsymbol{\phi}|^2} = \frac{1}{n}|\mathbf{t}|^2 - \frac{1}{n}\frac{1}{|\boldsymbol{\phi}|^2}\frac{nr^2-1}{1-r^2}\frac{r^2|\boldsymbol{\phi}|^2|\mathbf{t}|^2}{1+\frac{1}{|\boldsymbol{\phi}|^2}\frac{nr^2-1}{1-r^2}|\boldsymbol{\phi}|^2}$$

$$= \frac{1}{n}|\mathbf{t}|^2 - \frac{1}{n}\frac{nr^2-1}{1-r^2}\frac{r^2|\mathbf{t}|^2}{1+\frac{nr^2-1}{1-r^2}} = \frac{1}{n}|\mathbf{t}|^2\left[1 - \frac{nr^2-1}{1-r^2}\frac{r^2}{1+\frac{nr^2-1}{1-r^2}}\right]$$

$$= \frac{1}{n}|\mathbf{t}|^2\left[1 - \frac{nr^2-1}{n-1}\right] = |\mathbf{t}|^2\frac{1-r^2}{n-1} \tag{4.31}$$

Results of Joint Optimization Thus the result, plugging in the optimal σ^2:

$$\widehat{\sigma^2} = |\mathbf{t}|^2\frac{1-r^2}{n-1} \tag{4.32}$$

$$\widehat{\sigma^2_{\boldsymbol{\phi}}} = \frac{r^2|\mathbf{t}|^2 - |\mathbf{t}|^2\frac{1-r^2}{n-1}}{|\boldsymbol{\phi}|^2} = \frac{|\mathbf{t}|^2}{|\boldsymbol{\phi}|^2}\frac{nr^2-r^2-1+r^2}{n-1} = \frac{|\mathbf{t}|^2}{|\boldsymbol{\phi}|^2}\frac{nr^2-1}{n-1} \tag{4.33}$$

When $\widehat{\sigma^2_{\boldsymbol{\phi}}}$ is formally negative, that is, $r^2 < 1/n$, the valid solution with the highest likelihood is the sparse solution:

$$\widehat{\sigma^2} = |\mathbf{t}|^2\frac{1}{n} \tag{4.34}$$

$$\widehat{\sigma^2_{\boldsymbol{\phi}}} = 0 \tag{4.35}$$

Compare with the case for known σ^2:

$$\widehat{\sigma^2_{\boldsymbol{\phi}}} = \frac{r^2|\mathbf{t}|^2 - \sigma^2}{|\boldsymbol{\phi}|^2} \tag{4.36}$$

Here, likewise, the optimal choice is $\widehat{\sigma^2_{\boldsymbol{\phi}}} = 0$ if the solution is formally negative.

4.4 Incremental Improvement of the Variance Components

4.4.1 Classification of Optimization Cases

In our context, the general variance component problem is the maximization of the likelihood of the form

$$L = c - \frac{1}{2}\left[\ln\det\mathbf{C} + \mathbf{t}'\mathbf{C}^{-1}\mathbf{t}\right] \tag{4.37}$$

where $\mathbf{t}$ is the phenotype vector and $\mathbf{C}$ is the covariance. $\mathbf{C}$ is a linear combination of positive definite matrices; the problem is to find the coefficients of these matrices subject to some constraints.

We can reduce this problem to a set of optimization operations for restricted cases:

1. Solve for residual variance in isolation

$$\mathbf{C} = a\mathbf{I} \tag{4.38}$$

2. Rescale (1 is a special case)

$$\mathbf{C} = a\mathbf{M} \tag{4.39}$$

3. Solve for residual variance

$$\mathbf{C} = \mathbf{M} + b\mathbf{I} \tag{4.40}$$

4. Rescale including residual variance

$$\mathbf{C} = a\mathbf{M} + b\mathbf{I} \tag{4.41}$$

5. Optimize a component, holding all others fixed

$$\mathbf{C} = a\mathbf{M} + \mathbf{R} \tag{4.42}$$

6. Simultaneously optimize a pair of components (4 is a special case of this)

$$\mathbf{C} = a\mathbf{M} + b\mathbf{R} \tag{4.43}$$

4.4.2 Closed Form Solution

The first two cases can be solved in closed form; with $\mathbf{C} = a\mathbf{M}$, taking n as the dimension of $\mathbf{C}$,

$$L = c - \frac{1}{2}\left[n \ln a \ + \ln \det \mathbf{M} \ + a^{-1} \ \mathbf{t}'\mathbf{M}^{-1}\mathbf{t}\right] \tag{4.44}$$

$$\frac{\partial L}{\partial a} = 0 = -\frac{1}{2}\left[\frac{n}{a} - \frac{\mathbf{t}'\mathbf{M}^{-1}\mathbf{t}}{a^2}\right] \tag{4.45}$$

$$a = \frac{\mathbf{t}'\mathbf{M}^{-1}\mathbf{t}}{n} \tag{4.46}$$

This is a solution for case 2 above. As we would expect, the solution for case 1, $\mathbf{C} = a\mathbf{I}$, a is the sample variance,

$$a = \frac{\mathbf{t}'\mathbf{I}^{-1}\mathbf{t}}{n} = \frac{\mathbf{t}'\mathbf{t}}{n} \tag{4.47}$$

4.4.3 General Solution

For the remaining four cases, we optimize the likelihood using the eigenvalue method used recently Lippert et al. (2011), and originating in Thompson and Shaw (1992) and Thompson and Shaw (1990). We extend the decomposition approach from a the Singular Value Decomposition of Lippert et al. (2011) to a simultaneous diagonalization of two positive definite matrices. The algorithm is as follows:

1. Begin with matrices $\mathbf{A}$ and $\mathbf{B}$

2. Perform the Cholesky Decomposition, computing $\mathbf{Y}$ and $(\mathbf{Y}')^{-1}$:

$$\mathbf{A} = \mathbf{Y}'\mathbf{Y} \text{ and } (\mathbf{Y}')^{-1}\mathbf{A}\mathbf{Y}^{-1} = \mathbf{I} \tag{4.48}$$

3. Let $\mathbf{C} = \mathbf{Y}^{-T}\mathbf{B}\mathbf{Y}^{-1}$

4. Perform the Singluar Value Decomposition, computing $\mathbf{U}$ and $\mathbf{U}'$:

$$\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}' \text{ and } \mathbf{U}'\mathbf{C}\mathbf{U} = \mathbf{D} \tag{4.49}$$

5. Let $\mathbf{T} = \mathbf{Y}^{-1}\mathbf{U}$

In special cases we can simplify the computation further:

1. If $\mathbf{A} = \mathbf{I}$: $\mathbf{Y} = \mathbf{I}$ and $\mathbf{C} = \mathbf{B}$, so we perform SVD on $\mathbf{B}$. This most closely matches the case considered by Lippert et al. (2011).

2. If $\mathbf{B} = \mathbf{I}$: $\mathbf{C} = (\mathbf{Y}')^{-1}\mathbf{Y}^{-1} = \mathbf{A}^{-1}$ so we perform SVD on $\mathbf{A}^{-1}$ and Cholesky on $\mathbf{A}$.

Having obtained the transformed matrices, we observe:

$$\mathbf{T}'\mathbf{A}\mathbf{T} = \mathbf{U}'(\mathbf{Y}')^{-1}\mathbf{A}\mathbf{Y}^{-1}\mathbf{U} = \mathbf{U}'\mathbf{I}\mathbf{U} = \mathbf{U}'\mathbf{U} = \mathbf{I} \tag{4.50}$$

$$\mathbf{T}'\mathbf{B}\mathbf{T} = \mathbf{U}'(\mathbf{Y}')^{-1}\mathbf{B}\mathbf{Y}^{-1}\mathbf{U} = \mathbf{U}'\mathbf{C}\mathbf{U} = \mathbf{D} \tag{4.51}$$

$$\mathbf{T}'\left[a\mathbf{A} + b\mathbf{B}\right]\mathbf{T} = a\mathbf{I} + b\mathbf{D} \tag{4.52}$$

$$a\mathbf{A} + b\mathbf{B} = (\mathbf{T}')^{-1}\left[a\mathbf{I} + b\mathbf{D}\right]\mathbf{T}^{-1} \tag{4.53}$$

Leading to the results:

$$[a\mathbf{A} + b\mathbf{B}]^{-1} = \mathbf{T}[a\mathbf{I} + b\mathbf{D}]^{-1}\mathbf{T}' \tag{4.54}$$

$$\det\left[a\mathbf{A} + b\mathbf{B}\right] = \det(\mathbf{T}')^{-1}\det\left[a\mathbf{I} + b\mathbf{D}\right]\,\det\mathbf{T}^{-1} = \frac{\det\left[a\mathbf{I} + b\mathbf{D}\right]}{\left[\det\mathbf{T}\right]^2} \tag{4.55}$$

Because the matrix inversion and the determinant are now applied to diagonal matrices, the computation of the likelihood is linear in the dimension of the matrix, and we can optimize with respect to a and b with standard nonlinear optimization methods.

4.4.4 Generalizing to Many Variance Components

The method thus described solves the optimization problem for two variance components. The general problem we aim at is the generalized multi-locus case,

$$\mathrm{Cov}(\mathbf{P}) = \sigma_E^2\mathbf{I} + \sum_{l=1}^{L} C_{4l}\mathbf{I}_{4l} \; + C_{3l}\mathbf{I}_{3l} \; + C_{2l}\mathbf{I}_{2l} \; + C_{1l}\mathbf{I}_{1l} \tag{4.56}$$

where, for each locus l, $C_4 = \sigma^2_{AA}$, $C_3 = \sigma^2_{AB}$, $C_2 = \rho_3\sigma_{AB}\sigma_{AA}$, $C_1 = \rho_2\sigma^2_{AB}$; $\mathbf{I}_k$ is a matrix whose entry ij is an indicator, 1 when individuals i and j are in identity state k and 0 otherwise, and the positive definiteness constraint holds:

$$2C_1 \leq C_3; \quad 2C_2^2 \leq 2C_4C_1 \tag{4.57}$$

We optimize over the set of all C_{il} and a single σ^2_E. Starting with an initial vector of variance components, we want a procedure to update one or more components to increase the likelihood until such possibilities are exhausted, with the understanding that both intermediate and optimal solutions may be sparse, with many components temporarily set to zero. The Tipping-Faul method corresponds to case 5 above: we choose one variance component to optimize, holding others fixed. In Tipping-Faul, the form of the constituent matrix $\mathbf{I}_{il}$ is constrained to rank 1, corresponding to a regression. Case 5 is applicable to the general constituent matrix..

Case 6 allows a more general approach to optimization. We choose a subset of matrices with their corresponding variance components $C_{il} \ldots, \sigma^2_E$ to designate as $\mathbf{M}$, and its complement as $\mathbf{R}$ in the Case 6 expression. Finding the optimal coefficients a and b, we rescale the two sets of variance components. The split-optimize cycle is repeated until convergence indicated heuristically indicated by a lack of progress over a sufficient number of cycles.

For our purposes, the variance components have three roles:

1. Overall scaling of the genetic contribution to the phenotype

2. Between loci, relative weighting of loci by relevance to a specific trait

3. Within a locus, covariance structure between alleles, associated with the pattern of dominnance

For a fixed set of IBF parameters up to scale over all loci, such as that given by the Additive or Random Order Strict Dominance models, and a fixed set of weights, the optimization problem is reducible to fitting a single pair of parameters, corresponding to a genetic and an environmental variance. The simplest choice of weights is the equal locus

weight. In a more sophisticated approach, we can apply standalone variance component estimation to each locus, and apply Case 6 to the split between all genetic components and the single environmental component, in effect rescaling the standalone variance. This, in effect, is how we proceed in Chapter 5.

Chapter 5

RESULTS

5.1 Introduction

We will illustrate the methods of Chapters 3 and 4 with three case studies:

1. A simulated single locus trait. With clean data, a large sample size, and a simple error model, we illustrate the application of the range of tools built in Chapter 3 for inference of a number of known parameters, including heritability.

2. Five pig phenotypes of varying heritabilities, with SNP level genotype data. We use both equal and variable weightings, and perform prediction and heritability analysis.

3. A human height phenotype. The population is nominally unrelated, the SNP genotypes are of high density and imputed to a higher density, and genome partitioning procedure is performed to construct regions with multiple alleles from multiple SNPs. We perform prediction and heritability analysis.

5.2 Case Study 1: Simulated Trait

5.2.1 Description

A toolkit for simulation and prediction using the IBF model was built in the R language, as a set of classes using the Reference Class model. We use this simulation to demonstrate the implementation details of the model, and to illustrate our approach to prediction and heritability estimation. The toolkit was published as part of Sverdlov and Thompson (2013). Here we illustrate a trait and population simulated using this toolkit, and a set of inferences following procedures described in Chapter 3.

Simulation	N	θ	Distinct Alleles	Singletons
1	5000	2.0	14	1
2	10000	4.0	27	6

Table 5.1: Summary of infinite allele process population simulations.

5.2.2 Generating Allele Frequencies

Using the *AlleleFrequencyModel* class, the infinite allele model is simulated using the Chinese Restaurant Process with a fixed population size N and Ewens parameter θ. This gives population allele frequencies. The population, simulated with the $AlleleFrequencyModel$ class, is described in Table 5.2 and plotted in Figure 5.1. The top two plots illustrate alternative views of the allele frequency distribution.

Note the two ways of plotting the allele frequency histogram. The typical allele, when sampled by population frequency (i.e. drawn from the population) is a high frequency allele; but when sampling from the distinct allele list, or with equal frequency, the typical allele is a singleton or rare allele.

Using a higher θ parameter implies higher allele diversity (Figure 5.2). The parameter choices and summary counts of distinct alleles and singletons in the finite population are given in Table 5.1.

5.2.3 Calculations with IBF parametrizations

The $IBFParametrization$ class allows the manipulation of sets of IBF parameters. For a single locus, we fix the full set of IBF parameters as a midpoint in covariance space between the Additive and Random Order, Strict Dominance Models with $\mu_{AA} = \mu_{AB} = 0$ and $\sigma_{AA} = 1$. We label this as the *Moderate* parametrization. We will simulate genotypic values using the multivariate normal model with these IBF parameters (Table 5.3).

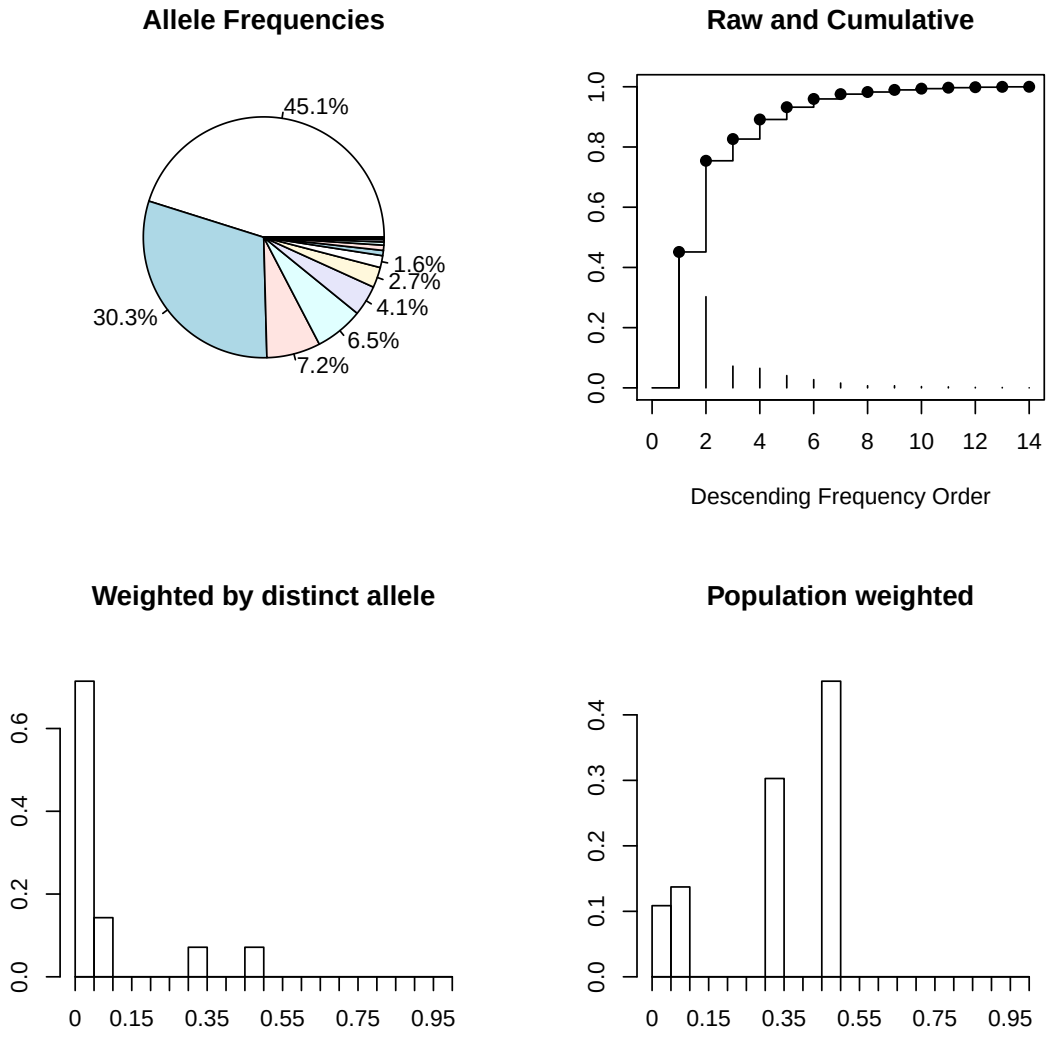


Figure 5.1: Simulated allele frequencies (Simulation 1).

Simulation 1			Simulation 2		
k	N_k	$freq$	k	N_k	$freq$
5	2257	0.4514	1	5914	0.5914
1	1514	0.3028	4	1212	0.1212
3	361	0.0722	10	756	0.0756
4	325	0.065	19	383	0.0383
2	204	0.0408	24	366	0.0366
123	137	0.0274	16	276	0.0276
39	80	0.016	98	210	0.021
14	35	0.007	60	187	0.0187
158	35	0.007	80	172	0.0172
192	21	0.0042	3	157	0.0157
355	16	0.0032	181	126	0.0126
453	8	0.0016	28	89	0.0089
156	6	0.0012	285	50	0.005
971	1	0.0002	6	39	0.0039
			279	22	0.0022
			520	17	0.0017
			608	6	0.0006
			2951	6	0.0006
			2482	2	0.0002
			2637	2	0.0002
			5897	2	0.0002
			7171	1	0.0001
			7594	1	0.0001
			9211	1	0.0001
			9255	1	0.0001
			9563	1	0.0001
			9587	1	0.0001

Table 5.2: Simulated allele counts. k is the order of appearance of a surviving allele in the population history. N_k is the count of the allele in the final population. $freq$ is the allele frequency in the final population.

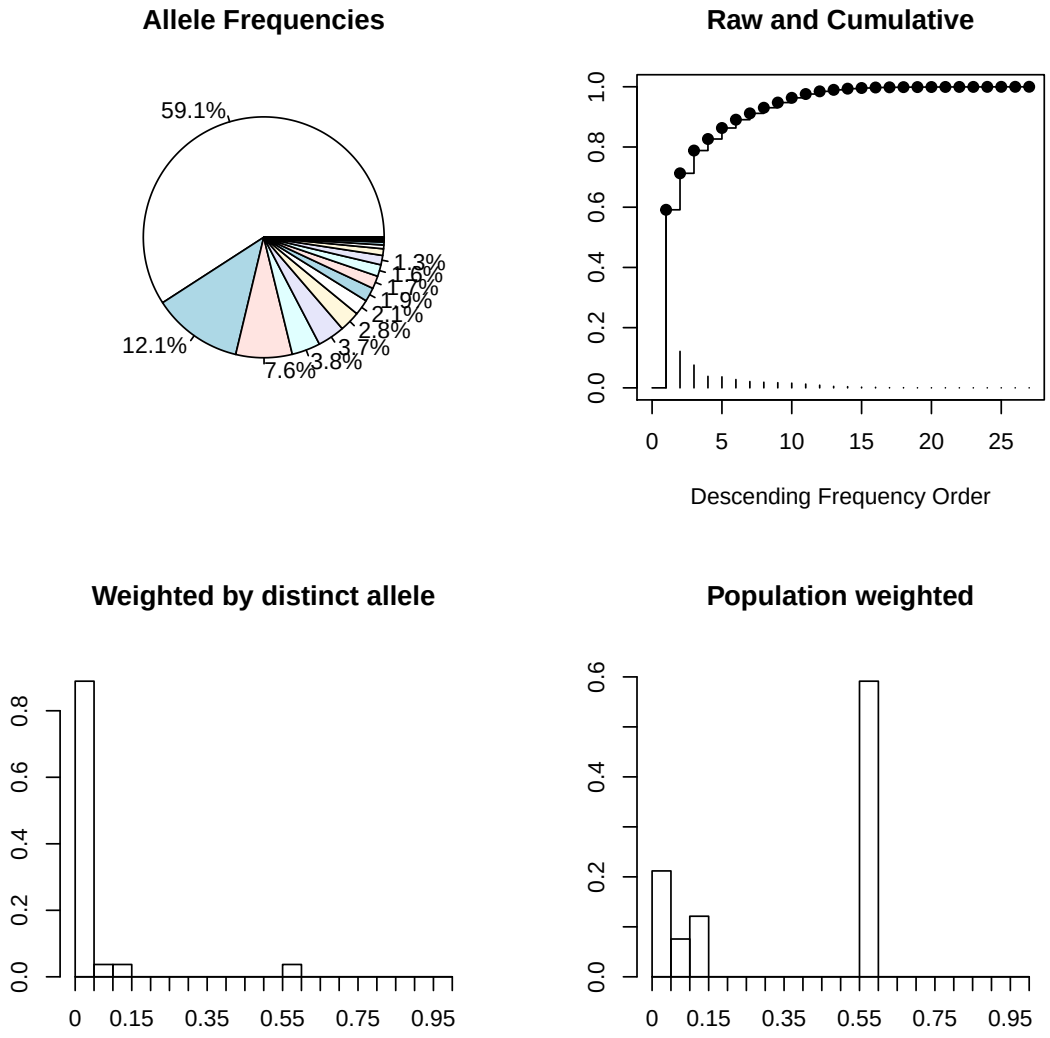


Figure 5.2: Simulated allele frequencies (Simulation 2).

	C_0	C_1	C_2	C_3	C_4
Additive	0.000	0.250	0.500	0.500	1.000
Dominance	0.000	0.333	0.500	1.000	1.000
Moderate	0.000	0.292	0.500	0.750	1.000

	ρ_3	ρ_3^2	ρ_2	σ_{AB}	σ_{AA}
Additive	0.707	0.500	0.500	0.707	1.000
Dominance	0.500	0.250	0.333	1.000	1.000
Moderate	0.577	0.333	0.389	0.866	1.000

	σ^2_{AB}	σ^2_{AA}	μ_{AA}	μ_{AB}
Additive	0.500	1.000	0.000	0.000
Dominance	1.000	1.000	0.000	0.000
Moderate	0.750	1.000	0.000	0.000

Table 5.3: Simulated variance components.

Generating Alleles and IBF states

The $IBFMatrixGenerator$ class manages allele states, IBF state parametrizations, and translates these to covariance matrices given IBF parameters in $IBFParametrization$ objects. In Table 5.4, as an illustration, we use it to generate all allele states for 3 alleles.

Pairwise IBF states are given in Table 5.5 and covariance matrix using "moderate" parametrization are given in Table 5.6.

Simulating Genotypic Values

We will use allele frequencies from simulation 2 in Table 5.1 and the *moderate* parametrization from Table 5.3. This is a single draw from the Gaussian model, henceforth to be treated as the true genotypic values. We plot these genotypic values against allele frequencies as

	allelePairIndex	allele1	allele2
A:A	1	1	1
B:B	2	2	2
C:C	3	3	3
A:B	4	1	2
A:C	5	1	3
B:C	6	2	3

Table 5.4: All generated pairwise allele states for 3 alleles.

	A:A	B:B	C:C	A:B	A:C	B:C
A:A	4	0	0	2	2	0
B:B	0	4	0	2	0	2
C:C	0	0	4	0	2	2
A:B	2	2	0	3	1	1
A:C	2	0	2	1	3	1
B:C	0	2	2	1	1	3

Table 5.5: Pairwise IBF states.

Figure 5.3.

Projection matrix utilities can be used to compute the Fisher α_i, the projections onto additive effects (i.e. breeding values, removing dominance deviations), and the vector of mean-adjusted genotypic values. Since we now have additive effects (α_i), we can also plot them against allele frequencies, as Figure 5.4.

Sampling Genotypes and Genotypic Values

We set up model objects, primarily the *GenotypicValuesRealization* class, which contains and manages a single sample from the Gaussian model, to be treated as the "true" genotypic

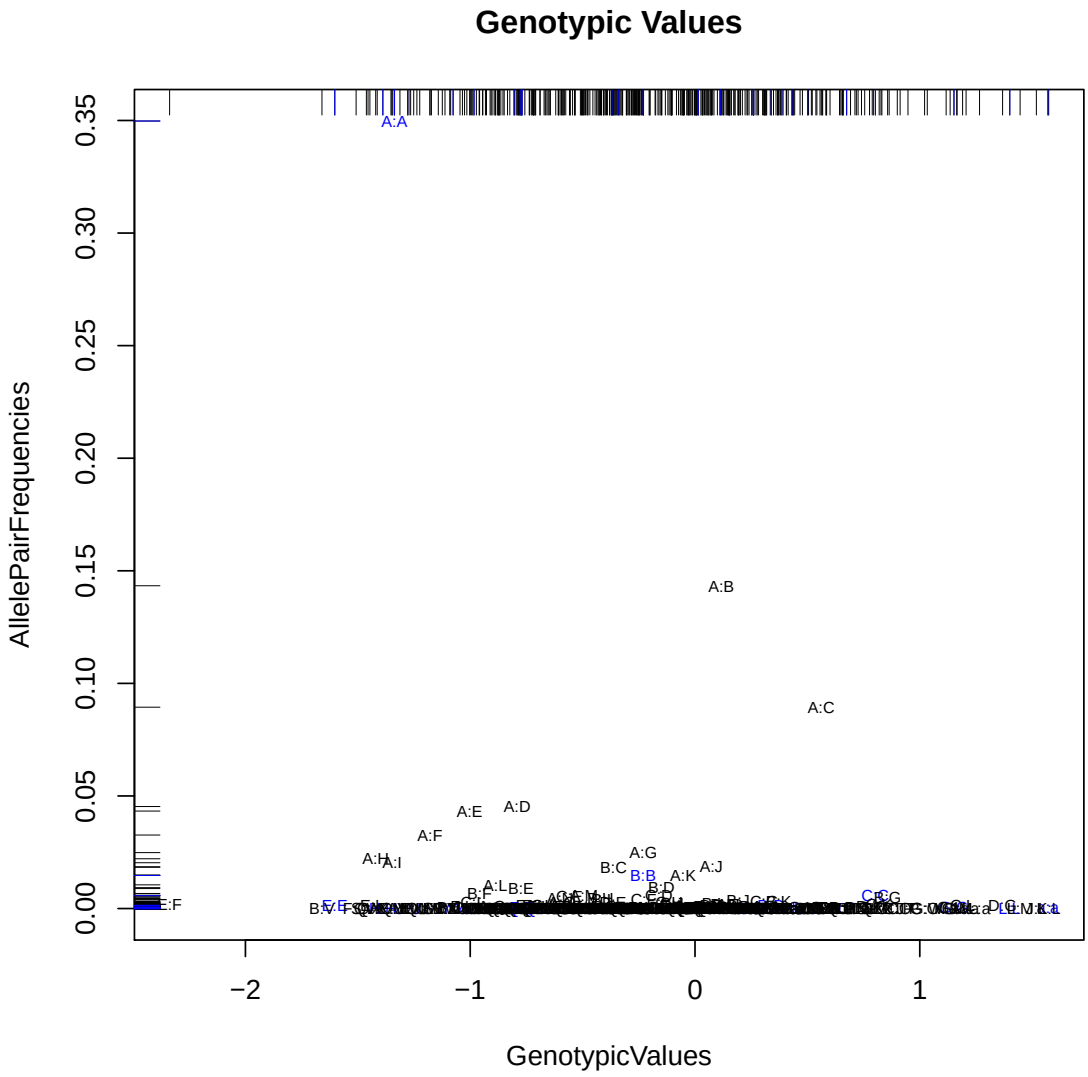


Figure 5.3: Genotypic values plotted aganst allele frequencies. Blue indicated homozygotes.

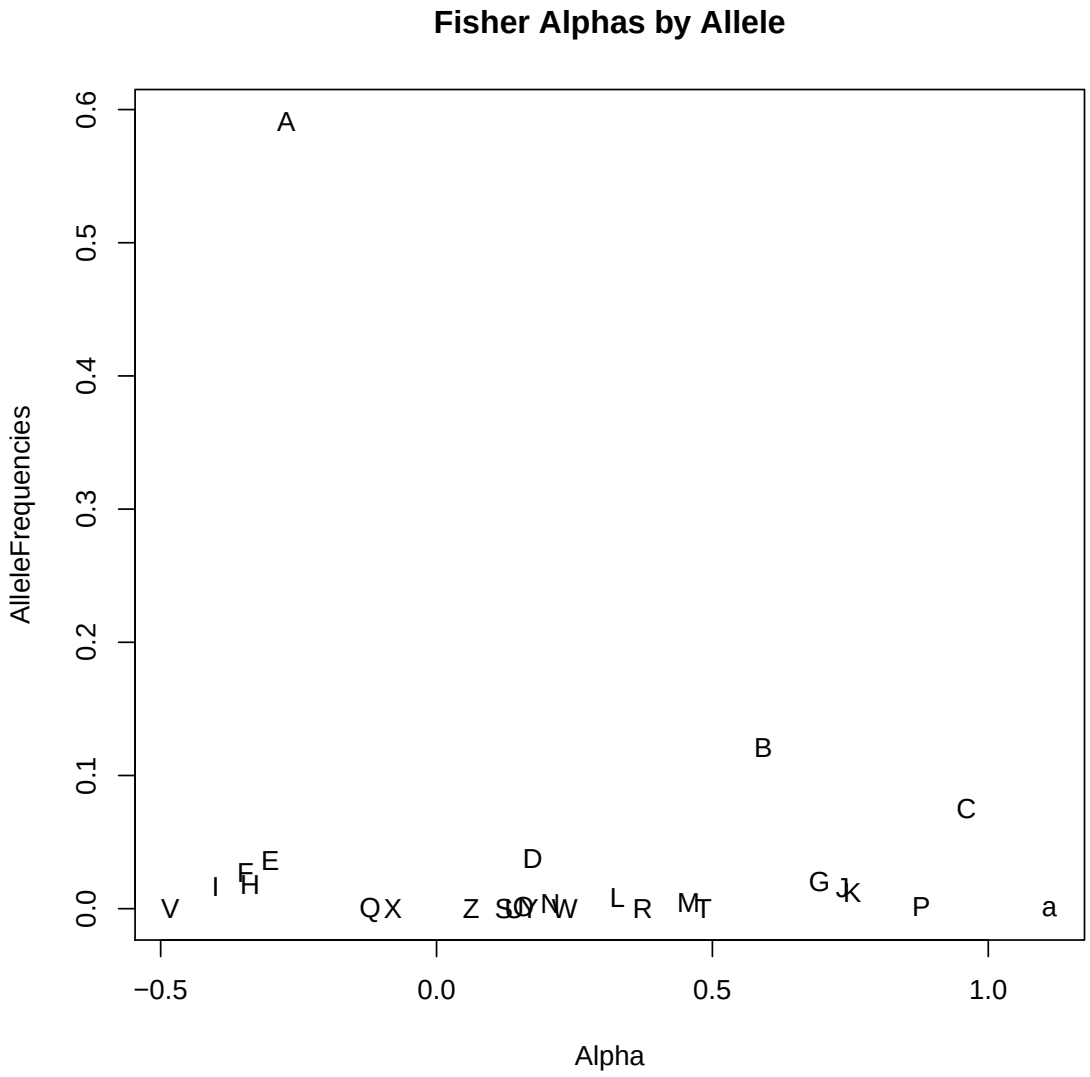


Figure 5.4: Fisher Alphas against allele frequencies.

	A:A	B:B	C:C	A:B	A:C	B:C
A:A	1.00	0.00	0.00	0.50	0.50	0.00
B:B	0.00	1.00	0.00	0.50	0.00	0.50
C:C	0.00	0.00	1.00	0.00	0.50	0.50
A:B	0.50	0.50	0.00	0.75	0.29	0.29
A:C	0.50	0.00	0.50	0.29	0.75	0.29
B:C	0.00	0.50	0.50	0.29	0.29	0.75

Table 5.6: IBF state covariance matrix.

values. We generate a finite sample of genotypes using the allele frequencies, with corresponding "true" genotypic values, and add random environmental contribution with $\sigma_E^2 = 1$. For the observed individuals, we plot the true genotypic values vs. phenotypes including environmental "error" term, observing repeated true genotypes sampled in proportion to population frequency, as Figure 5.5.

Comparing Predicted and True Fisher Alpha

We plot True vs. Predicted Fisher α values, with the area of the square proportional to allele frequency, as Figures 5.6 and 5.7. Note that high frequency alleles are generally well predicted, while rare alleles are sometimes predicted with substantial error. Rare alleles not observed in the finite sample have a common predicted value near zero, consistent with zero global mean.

5.2.4 Variance Component Estimation and Missing Heritability

Prior, Posterior, and Complete Data Posterior

Given the genotype vector $\mathbf{G}$, we can compute the classical variance components involved in the calculation of narrow and broad sense heritability:

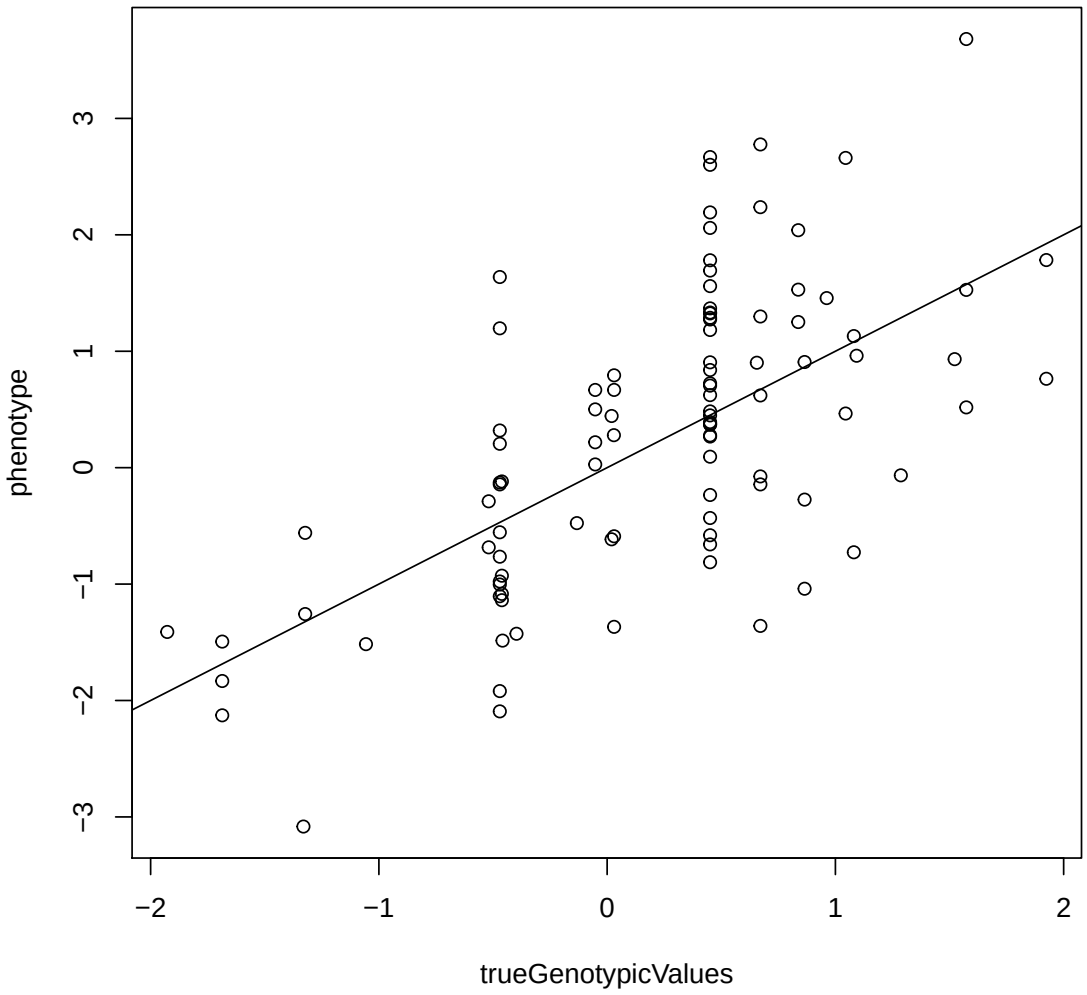


Figure 5.5: True genotypic values vs. phenotypes with error term.

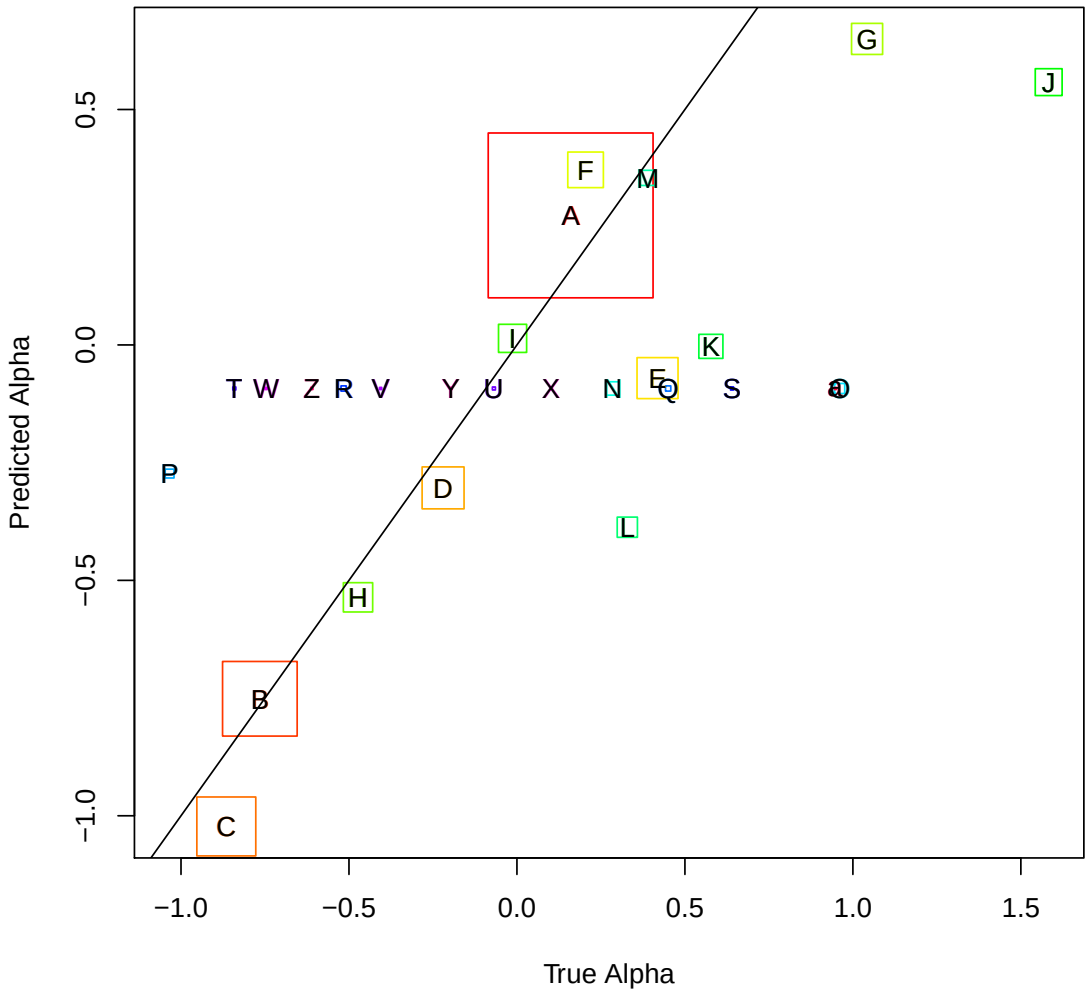


Figure 5.6: True vs. Predicted additive effects (Fisher α). Letters identify alleles. Area of square is proportional to allele frequency.

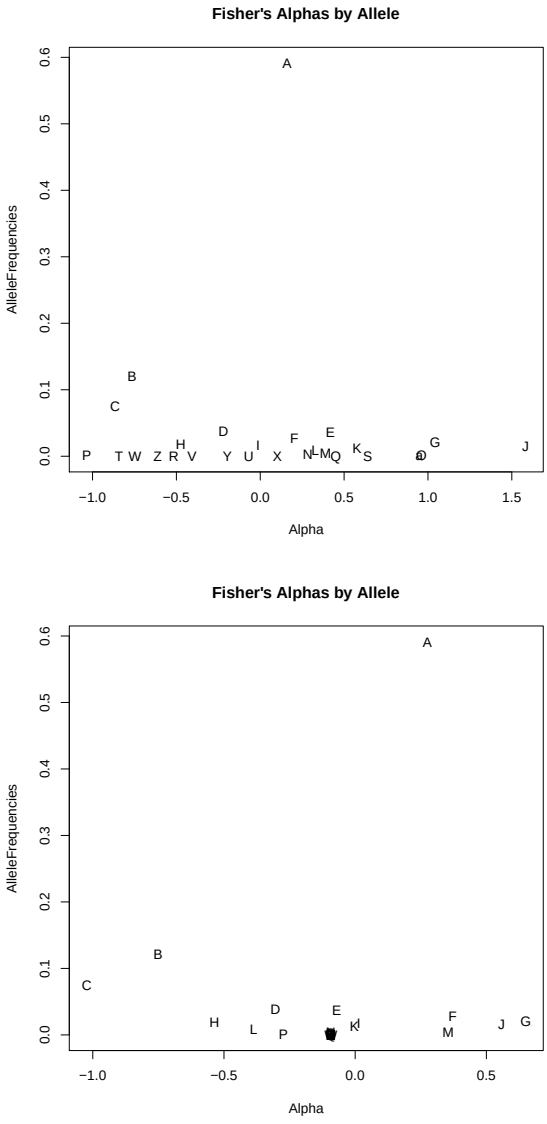


Figure 5.7: Comparison of distribution of true alphas (above) vs. predicted alphas (below) plotted against allele frequency.

SSA	SSD	SSG
σ^2_A	σ^2_D	$\sigma^2_G = \sigma^2_A + \sigma^2_D$

Given the *true* genotype vector $\mathbf{G}$, we can compute the variance components by classical methods, deterministically. There is no model uncertainty about the variance components (i.e. SD=0). All of the variance is attributed to the mean (Explained) effects, and there are no covariance (Unexplained) effects. This is equivalent to the expectation under the complete data posterior; that is, to observing an unlimited number of genotype-phenotype pairs under the true model. These are the actual variance components from the particular genotypic values we generated:

	SSA	SSD	SSG
Explained	0.40	0.14	0.54
Unexplained	0.00	0.00	0.00
Total	0.40	0.14	0.54
SD	0.00	0.00	0.00

Suppose we have no knowledge of the genotypic values, but only fixed IBF parameters. Then we can repeatedly sample vectors of genotypic values, and compute the distribution over these realizations. From the Bayesian perspective, this corresponds to the prior. The expectation is attributed to Unexplained, or covariance, effects, and there is a substantial variability in the variance components as expressed by the standard deviation of the Total term (SD).

	SSA	SSD	SSG
Explained	0.00	0.00	0.00
Unexplained	0.38	0.07	0.45
Total	0.38	0.07	0.45
SD	0.24	0.03	0.25

The average variance components from repeated draws from the Gaussian genotypic vector model, and the corresponding standard deviations, closely match the theoretical expectations given by the matrix formulae in the above expression for the prior:

	SSA	SSD	SSG
Explained	0.37	0.07	0.44
Total	0.37	0.07	0.44
SD	0.22	0.03	0.23

When we actually observe some genotype-phenotype pairs, and condition on that information, the result is intermediate. The table below gives the mean and SD of the posterior distribution of the variance components conditional on the observed data. The (Total) mean is decomposed into Explained and Unexplained parts. The values of the variance components obtained by plugging the inferred genotypic values into the classical formulas are the Explained variance components. But this is an inherently downward biased estimate of the variance components; if there are no genotype-phenotype pairs, it is zero. The Unexplained, or covariance, component, can be interpreted as an adjustment for this bias. With unlimited data under the true model, both the Unexplained component and the SD error term vanish.

	SSA	SSD	SSG
Explained	0.44	0.01	0.45
Unexplained	0.08	0.05	0.13
Total	0.52	0.06	0.58
SD	0.13	0.02	0.13

Trait	N	μ	Raw σ	h^2	σ^s_A	Raw σ^2	σ^2_P	σ^2/σ^2_P	σ/σ_P
T1	2804	-.045	1.21	0.07	0.22	1.46	3.14	0.466	0.683
T2	2715	0.005	1.12	0.16	2.11	1.25	13.19	0.095	0.308
T3	3141	0.706	0.96	0.38	0.66	0.92	1.74	0.531	0.728
T4	3152	-1.073	2.33	0.58	4.93	5.43	8.50	0.639	0.799
T5	3184	37.989	60.45	0.62	3459.09	3654.20	5579.18	0.655	0.809

Table 5.7: PIC dataset phenotype characteristics (Cleveland et al., 2012); additional computed variances and ratios described in text.

5.3 Case Study 2: Multiple Pig Traits, SNP Genotype Data

5.3.1 Description of Dataset

We applied methods described in Chapters 3 and 4 to a pig dataset containing genotypes, phenotypes, and estimated breeding values. provided by PIC (a Genus company) as a common dataset for genomic analysis for livestock populations, used in a number of publications for comparing genomic selection and prediction methods. The dataset is described by Cleveland et al. (2012), and the following text contains some numerical and factual statements quoted directly from that description.

The dataset contains 3534 individuals with high-density genotypes, phenotypes, and estimated breeding values for five traits. The individuals are related, and a pedigree is available, which includes non-genotyped parents and grandparents of the genotyped animals, for a total of 6473 pedigree animals.

Cleveland et al. (2012) state that accuracy of (traditional) prediction with respect to the second trait was initially used to select the animals to be genotyped. Table 5.7 summarized the properties of the five phenotypes. The heritability values h^2 and additive variances σ^2_A are given by Cleveland et al. (2012) as calculated by pedigree methods, and the μ and raw (sample) σ are given and readily verified. Note that because of relatedness, the raw variance is distinct from the population variance, without regard to finite sample biases;

this is particularly true for the second phenotype. We solve for the population variance,

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \rightarrow \sigma_P^2 = \frac{\sigma_A^2}{h^2} \tag{5.1}$$

and compute the ratios of raw to implied population variance and standard deviation.

There are 52,842 total SNPs, encoded as real numbers from 0 to 2, with interpolation in case of genotyping ambiguity. This dataset, designed for comparison of predictive methods, has the feature that SNPs are given in arbitrary order, without chromosome label. For this analysis, we treat each SNP as a distinct locus. IBF states are computed using the interpolation procedure in Section 3.2.2.

Of the 6473 animals, 3534 are genotyped; of these, 2715 have the T2 phenotype recorded, and 2314 have all five phenotypes recorded. We split these 2314 into filtered sets F_0 and F_1, with 1157 animals each, by odd/even sequential animal identifier. Smoothed histograms and cross-plots of the traits, using set F_0, are given in Figure 5.8.

5.3.2 Model Fitting

The training set is used to fit the variance components using the procedures of Chapter 4. We fitted the variance components with Equal and Variable weights across loci, using the Additive and Random Ordered Strict Dominance models. Variable variance components were obtained by regression of trait 2 onto each SNP, using the method of Section 4.3.

Prediction and effect estimation (the procedures of Chapter 3) are done in the validation set, using the variance components discovered in the training set. The fGRM based heritability estimates are obtained directly from the training set; the bottom-up heritability estimates are obtained in the validation set, using the variance components discovered in the training set. Allele frequencies are calculated in-sample, and Hardy-Weinberg equilibrium is assumed.

Table 5.8 illustrates the likelihood gain (relative to the noise-only model) from fitting variance components under each suggested model for each available trait. The low heritability trait $t1$ fails to fit a nonzero genetic variance component with any model. The use of variable variance components relative to trait $t2$ is a benefit in fitting trait $t2$ but not the other traits. Two traits favor the Additive model, and two favor the Random Order Strict

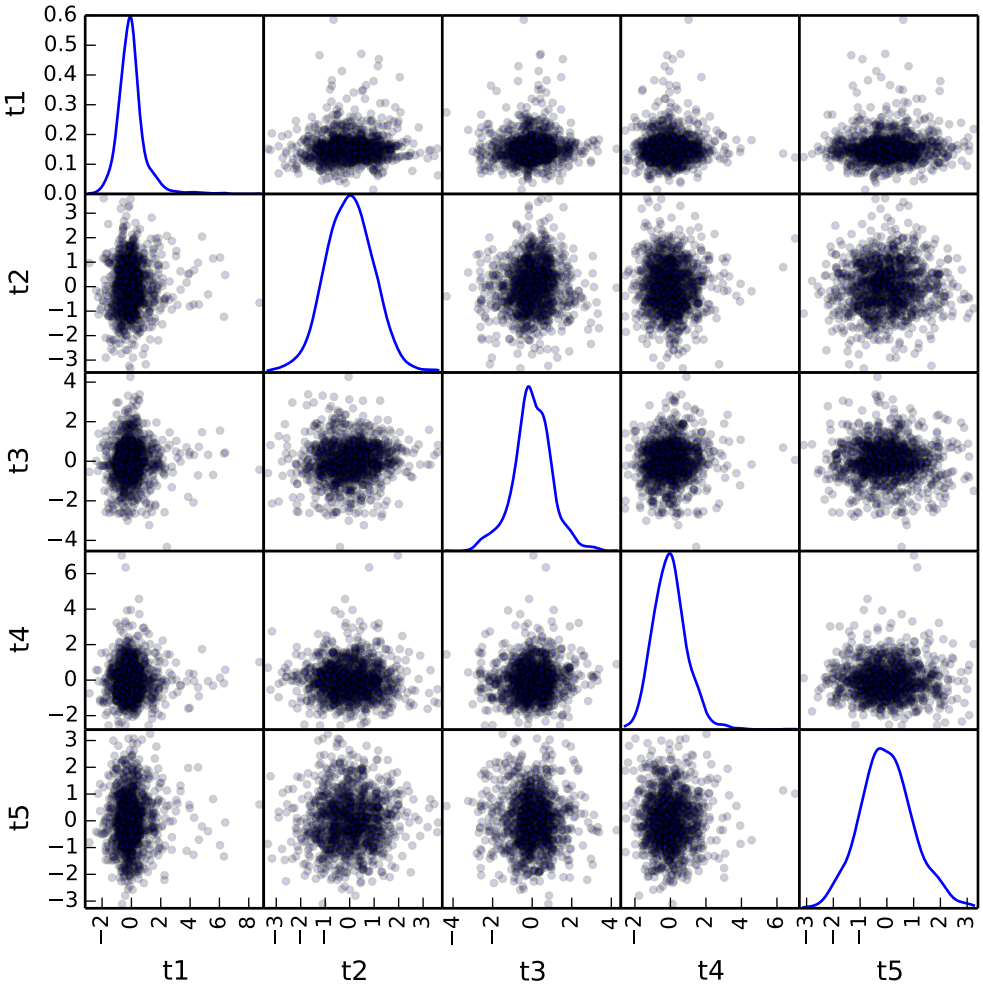


Figure 5.8: Normalized distributions for the five pig phenotypes.

Dominance model. The difference in likelihood between Additive and Dominance models is less than that between Variable and Equal weights.

5.3.3 Prediction

We perform a prediction procedure for each of the five traits, using the methodology of Section 3.4.2. Prediction is performed on a "leave one out" basis; that is, each of the N individuals in the validation sample is predicted using the phenotypes and genotypes of the other $N-1$ individuals. The calculation uses each of the sets of variance components (i.e. IBF parameters) given above.

We perform a regression of the observed onto the predicted phenotype. If the prediction is unbiased as expected, the regression should yield the line $y = 0 + 1x$ We compute the R^2 of this regression, which serves as the out of sample goodness of fit measure. Figures 5.9 and 5.10 present these results. Predictive performance is not obviously related to the comparative heritability of the traits, and appears little affected by the choice of model or weighting. In all cases, p-values from F-tests of regression coefficients appropriately support the unbiased predictor hypothesis $y = 1x + 0$ and oppose the noise-only model, $y = 0x + k$.

5.3.4 Heritability

Heritability is estimated by two methods. We use the variance component estimates to evaluate the scaling of the fGRM (Section 3.4.5), and estimate the heritability-related variance components directly. The results, compared in with the pedigree-based estimates given by Cleveland et al. (2012), are presented in Table 5.9. In the bottom-up approach, we follow Section 3.4.4. This is a reconstruction for the variable weight, dominance model. The results, in aggregate and for two individual SNP's representing a low and a high contribution to heritability. are presented in Table 5.10. Our two methods are consistent with one another, but differ from the pedigree-based estimates given by Cleveland et al. (2012). The maximum error bound on the heritability, formed by summing standard deviations (assuming worst-case case correlation) rather than variances (assuming independence) of errors across loci, is high relative to the $0-1$ range of the heritability parameter, and easily

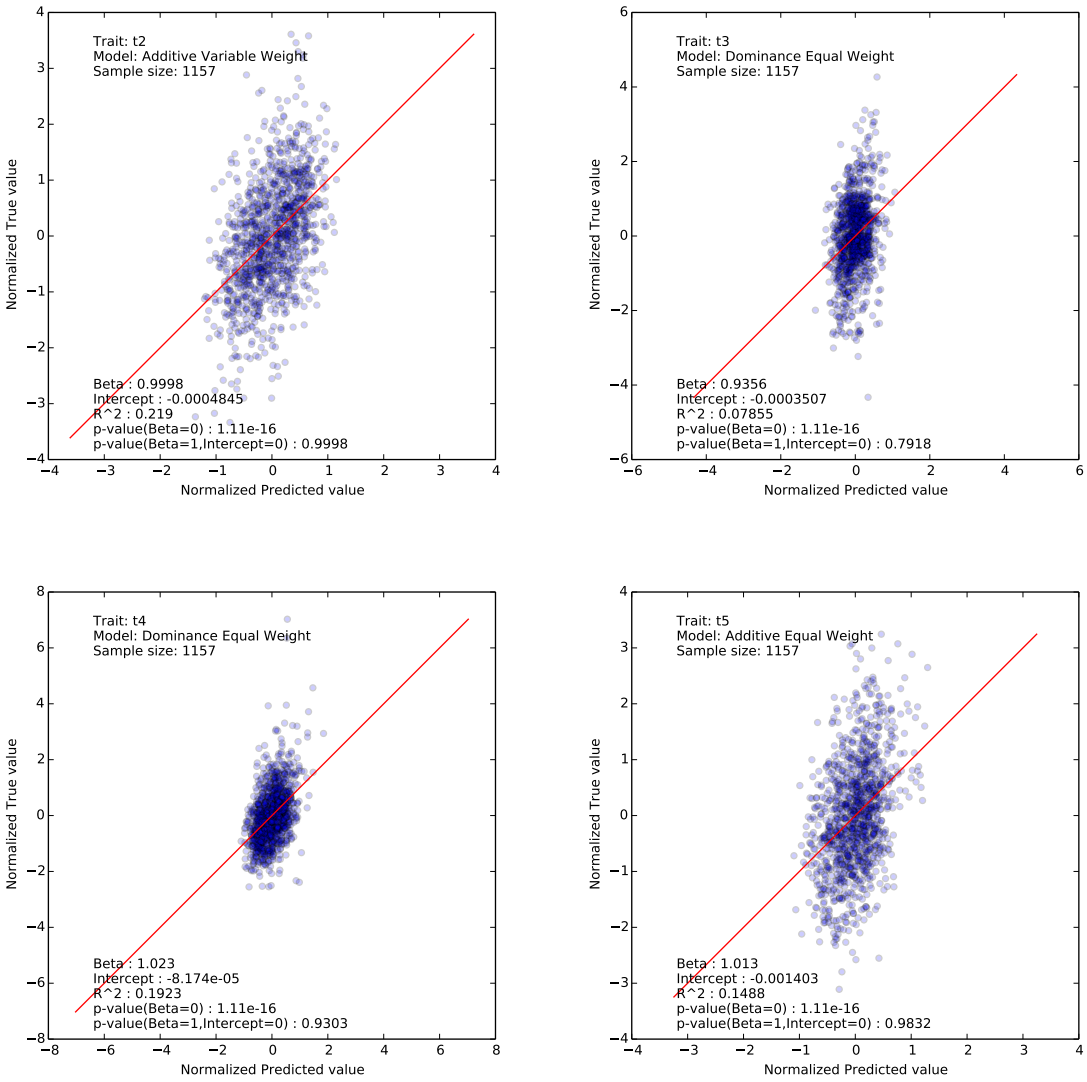


Figure 5.9: Comparison of prediction performance for multiple pig traits using best fitting model for each trait.

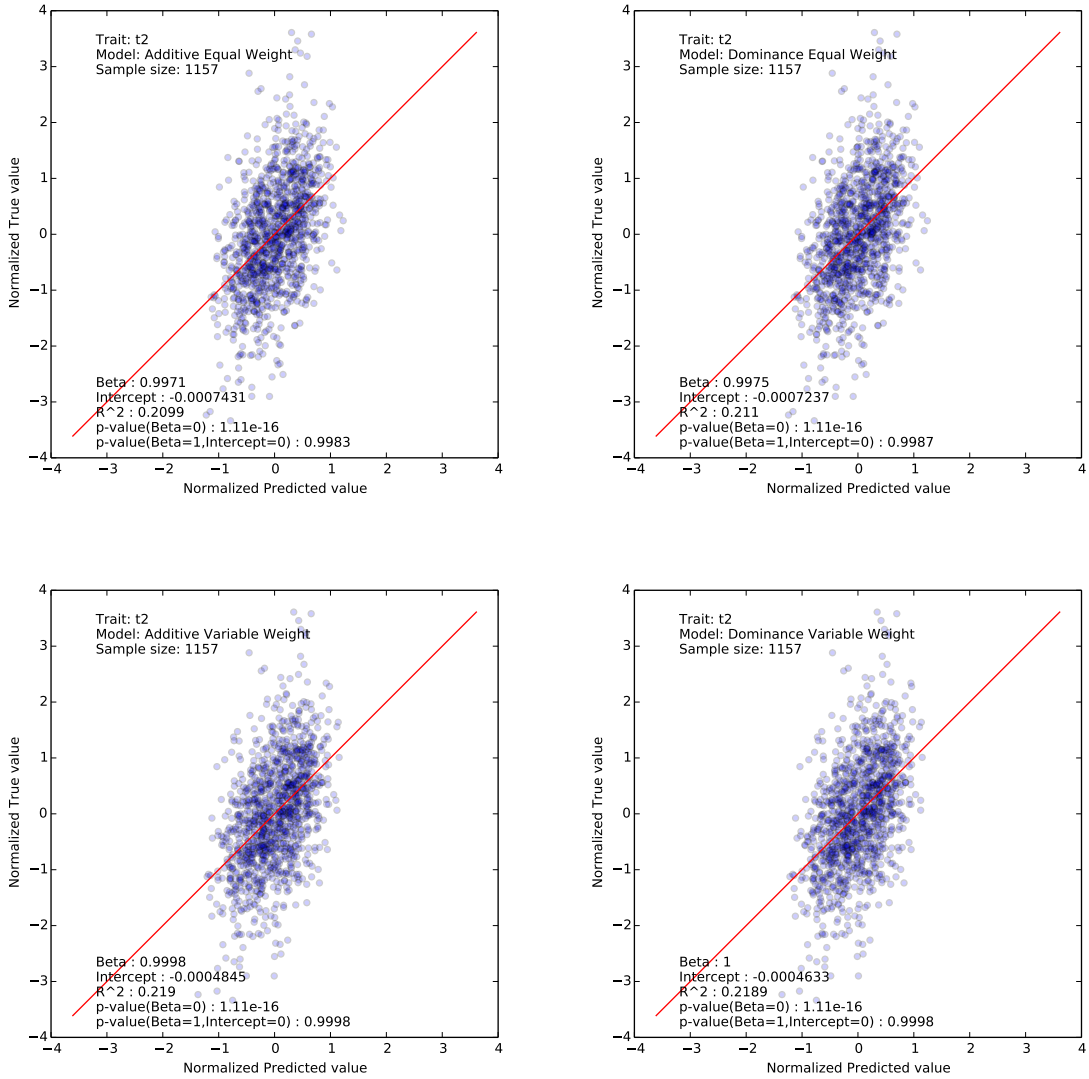
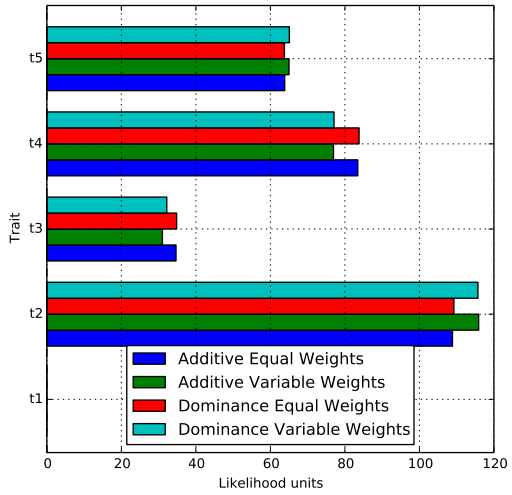


Figure 5.10: Comparison of prediction performance for pig trait $t2$ for various models.



Trait	t1	t2	t3	t4	t5
Model					
Additive Equal Weights	0.0	108.88	34.67	83.46	**63.84**
Additive Variable Weights	0.0	**115.9**	31.02	76.94	65.0
Dominance Equal Weights	0.0	109.25	**34.84**	**83.8**	63.76
Dominance Variable Weights	0.0	115.7	32.19	77.08	65.09

Table 5.8: Log Likelihood gain (genetic component vs. noise only) for each trait and model in the pig dataset.

accounts for any discrepancy.

	T1	T2	T3	T4	T5
Additive Equal Weights	0.0%	28.0%	23.8%	33.1%	29.7%
Additive Variable Weights	0.0%	27.0%	22.7%	31.0%	31.1%
Dominance Equal Weights	0.0%	28.1%	23.7%	33.4%	29.9%
Dominance Variable Weights	0.0%	26.9%	24.1%	30.5%	31.1%
Pedigree Estimate	7.0%	16.0%	38.0%	58.0%	62.0%

Table 5.9: fGRM Heritability estimates in the pig dataset; pedigree based estimates from Cleveland et al. (2012).

Small Effect SNP (#0)	σ^2_A	σ^2_D	σ^2_G
Explained	7.187×10^{-10}	3.102×10^{-9}	3.821×10^{-9}
Unexplained	1.842×10^{-5}	1.193×10^{-6}	1.961×10^{-5}
Total	1.842×10^{-5}	1.196×10^{-6}	1.962×10^{-5}
StDev of Total	2.605×10^{-5}	1.691×10^{-6}	2.617×10^{-5}
Large Effect SNP (#12464)	σ^2_A	σ^2_D	σ^2_G
Explained	8.052×10^{-7}	5.778×10^{-8}	8.052×10^{-7}
Unexplained	6.200×10^{-5}	4.433×10^{-6}	6.603×10^{-5}
Total	6.200×10^{-5}	4.490×10^{-6}	6.683×10^{-5}
StDev of Total	8.777×10^{-5}	6.350×10^{-6}	8.857×10^{-5}
Cumulative Heritability	σ^2_A	σ^2_D	σ^2_G
Explained	0.0035	3.600×10^{-5}	0.0035
Unexplained	0.2667	0.0178	0.2845
Total	0.2702	0.0178	0.2880
Max StDev of Total	0.3820	0.0252	0.3835

Table 5.10: Bottom-up heritability reconstruction for pig trait $t2$.

5.4 Case Study 3: Human Height, Multiple Alleles

5.4.1 Description of Dataset

Genotype data used for human height data have been sourced from the Health Professionals Follow-up Study (HPFS), (Rimm et al., 1991), containing male subjects, and the Nurses' Health Study (NHS) (Colditz and Hankinson, 2005), containing female subjects. The genotyping was performed as part of the GENEVA Genes and Environment Initiatives in Type 2 Diabetes (Nurses' Health Study/Health Professionals Follow-up Study), a GENEVA (Gene Environment Association Studies) project. Height data are included in both sets alongside other anthropometric traits as part of an effort to measure covariates relevant to cardiovascular and related disease. Yang et al. (2011) have previously analyzed this dataset for height.

5.4.2 Model Fitting

Genotypes The individuals were genotyped using the Affymetrix Genome-Wide Human 6.0 array with 909,622 SNP probes. Imputation to the resolution of the 1000 genomes dataset (1000 Genomes Project Consortium, 2012) was performed using BEAGLE (Browning and Browning, 2009) by the GENEVA project. Our starting point was this imputed genotype data. Partition into gene scale regions and the formation of IBF indices from SNP data is described in Section 3.2.4.

Phenotypes We performed analyses on male and female samples separately, avoiding the question of sex adjustment for height. Height phenotypes were zero-centered and adjusted for age by linear regression. As a sensitivity check, calculations were also performed without adjustment for age, and with linear and quadratic adjustment for age, with no observable difference in key results or significance levels. Heights were reported in centimeters; however, it appears most measurements were originally made in inches, as the raw histogram of heights is consistent with integer inches converted to centimeters and rounded. Figure 5.11 displays the (normalized) height trait, together with (normalized) BMI and (normalized) weight traits in the same dataset.

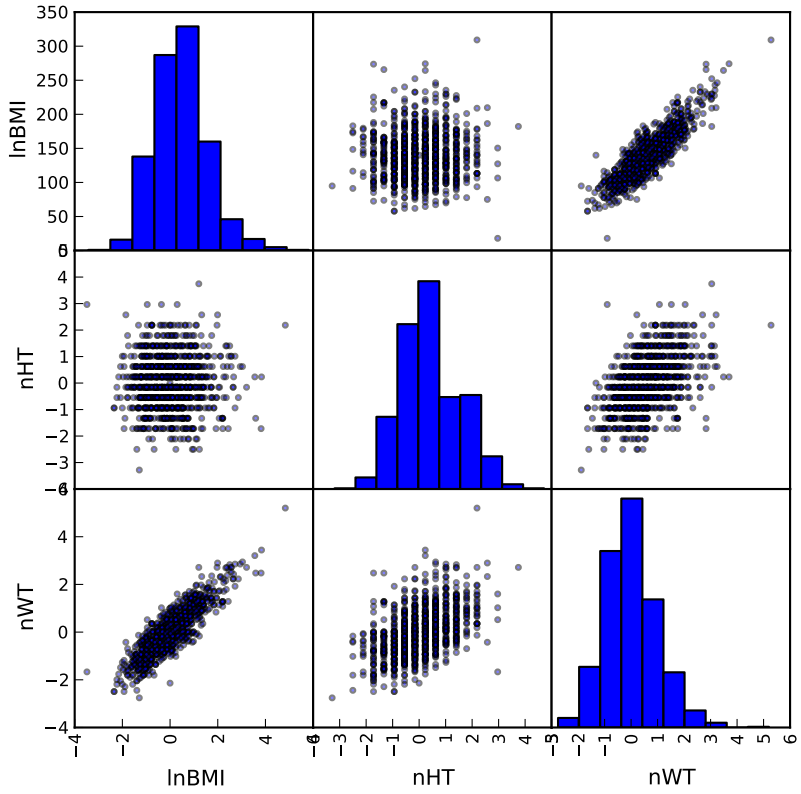


Figure 5.11: Normalized distributions for human phenotypes (BMI, height, weight).

Models Fitted In the analysis described below, we used the first 1000 individuals from the male sample, by identifier order. An equally weighted variance component model was fitted for height. Unlike height, variance component models did not fit nonzero genetic components for BMI or weight, using the same set of 1000 nominally unrelated individuals. Body Mass Index (BMI) and weight are not presumed to be as heritable as height, if only because of their variability with age.

5.4.3 Prediction

The prediction methodology, given IBF matrices and parameters, exactly mirrors that of the SNP based pig dataset, though the methods of constructing variance components and evaluating IBF states differ. The results are given in Figure 5.12. The choice of model, or correction for age, does not appear to meaningfully affect predictive performance as measured by R^2. In all cases, p-values from F-tests of regression coefficients appropriately support the unbiased predictor hypothesis $y = 1x + 0$ and oppose the noise-only model, $y = 0x + k$.

5.4.4 Heritability

Figure 5.13 demonstrates the decomposition by locus of the expected (Unexplained) heritability. As with the pig traits, the explained heritability is negligible.

The Additive model points estimates a total heritability of 19.41%. The Random Order, Strict Dominance model estimates a narrow-sense (additive) heritability of 20.39%, plus a dominance variance of 4.49% for a total genetic variance of 24.88%. These values are around half of those of Yang et al. (2010). We are, however, using only protein coding regions, and the fraction of variability captured is analogous to that attributed to protein coding regions by Yang et al. (2011). We note that the human trait has much lower predictive performance relative to a similar estimated heritability and sample size of the pig traits. The pig population is highly related, and the human population nominally unrelated, and there is no theoretical reason to expect R^2 to be indifferent to population structure. For example, high values of R^2 for human height in highly related populations are demonstrated by Makowsky et al. (2011).

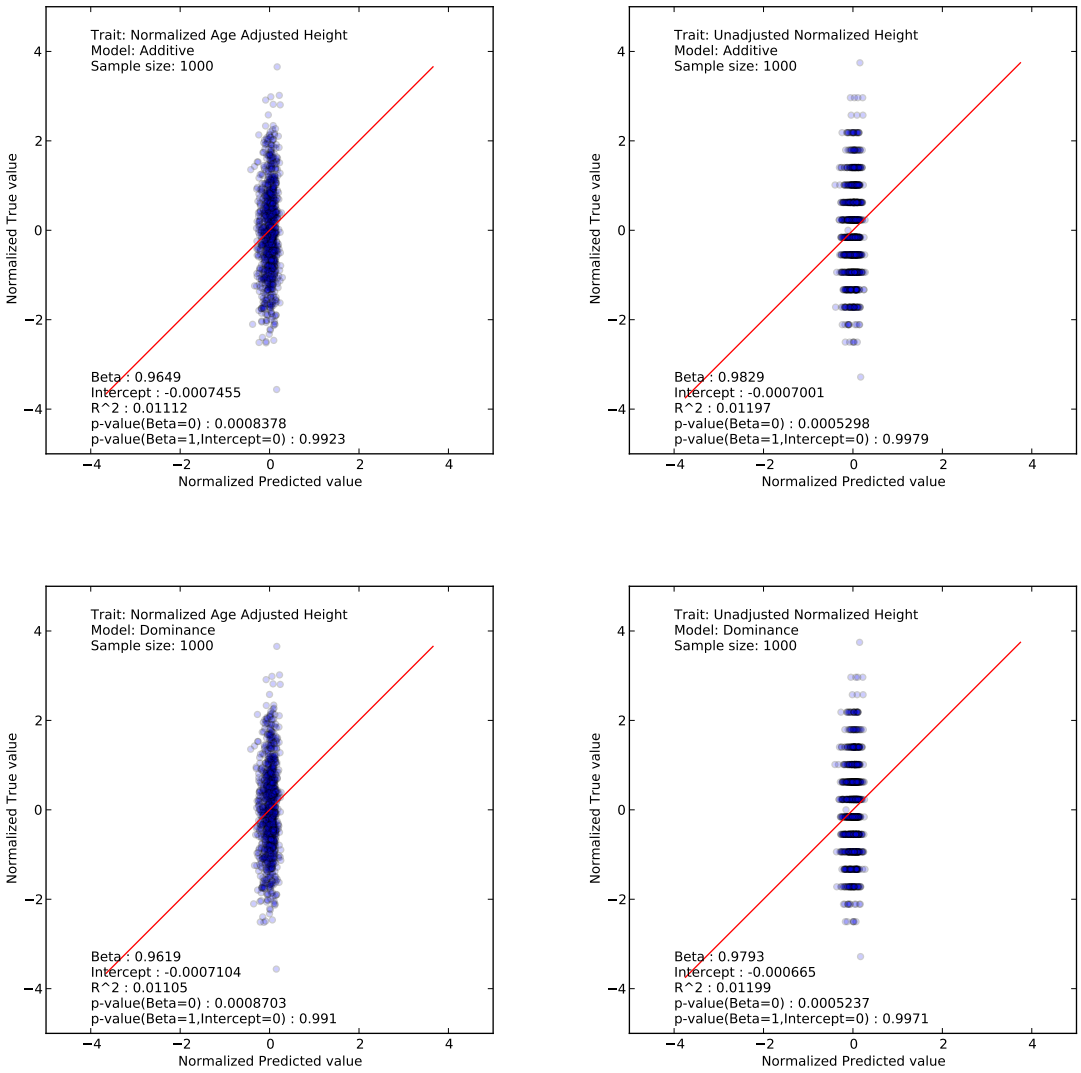


Figure 5.12: Comparison of prediction performance for human height for various models.

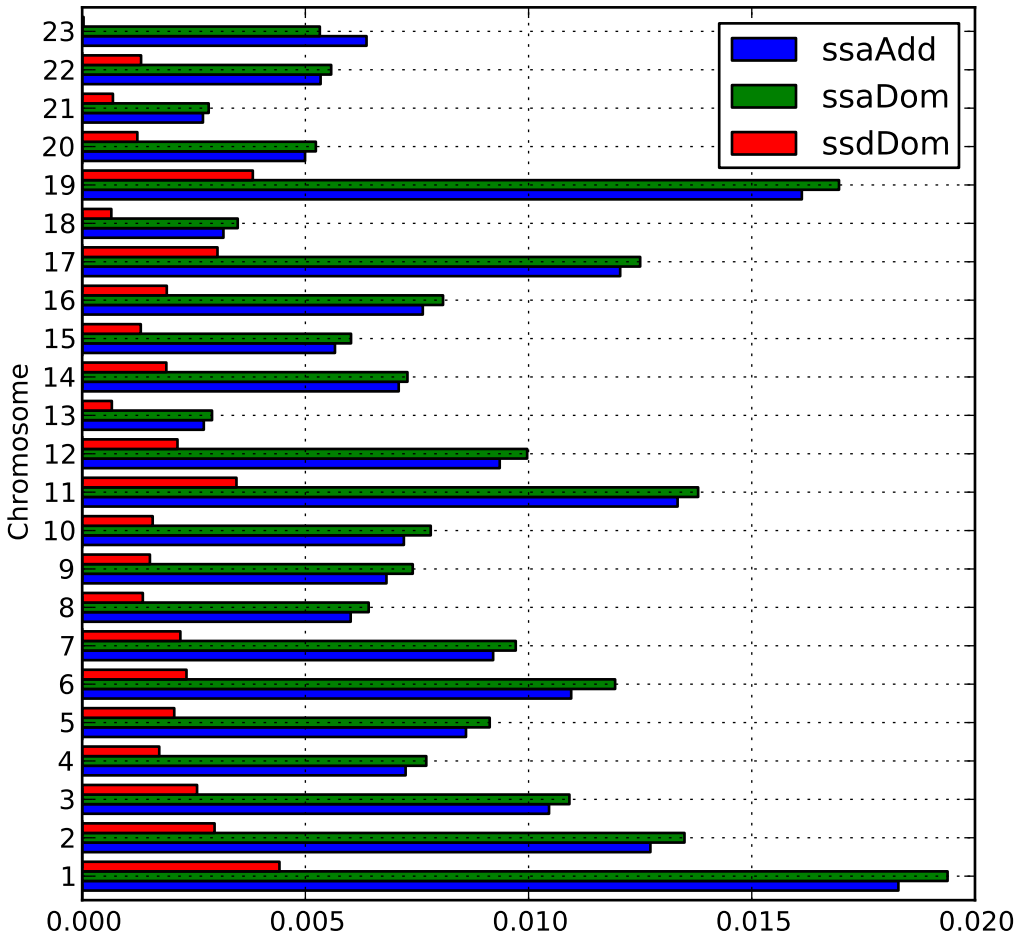


Figure 5.13: Human Height: Decomposition of Heritability by Chromosome. SSA is additive variance (σ^2_A), and SSD is dominance variance (σ^2_A), under the additive (Add) or random order strict dominance (Dom) model. Chromosome 23 is the X chromosome (no dominance variance estimated for males).

Chapter 6

THEORY EXTENSIONS: COMBINATORIAL APPROACHES TO DOMINANCE AND EPISTASIS

6.1 Introduction: Nonlinearity and the Natural Scale

In this chapter we consider the general phenomenon of nonlinearity in quantitative genetics. Henceforth we have only considered dominance, nonlinearity within a diploid gene. Epistasis is nonlinearity among multiple genes. We distinguish between the non-controversial assertion that epistasis in the biological sense, at the level of genes interacting to produce the phenotype of an individual, is common, and the controversy around the relative importance of epistasis as a statistical phenomenon useful in describing variation at the population level. There is an ongoing debate between the advocates (Zuk et al., 2012) and critics (Hill et al., 2008) of an important role for epistatic genetic variance in general complex traits. The canonical example of a human anthropometric trait thought to have high epistatic genetic variance is the Total Dermal Ridge Count (Heath et al., 1984), a property of fingerprints. Fingerprints are typically nearly identical in monozygotic twins, but similarities between the fingerprints of other relatives do not follow a simple pattern. Therefore, σ^2_G is high and σ^2_A is low.

The general genotype-phenotype relationship

$$Phenotype = f\left(Genotype, Environment\right)$$

can be specialized by the assumption of no Gene-Environment interaction, or the additive separability of contributions from genotype and environment,

$$Phenotype = f_G\left(Genotype\right) + f_E\left(Environment\right)$$

Further, the assumption of no epistasis implies additive separability over loci:

$$Phenotype = \sum_{locus\ l} f_{Gl}\left(\left(Genotype\right)_l\right) + f_E\left(Environment\right)$$

The assumption of no dominance is equivalent to additive separability of the two diploid alleles within the same locus. Diploid symmetry implies that the functions mapping the two alleles at the same locus to their real-valued contributions are the same:

$$Phenotype = \sum_{locus\ l} \left[f_{Al}\left(Allele_{1l}\right) + f_{Al}\left(Allele_{2l}\right) \right] + f_E\left(Environment\right)$$

Such additive separability depends on the scale of measurement of the trait. If we observe a trait T without epistasis or dominance, measuring on a different scale, say $\ln T$ or T^3, would violate additivity for any nontrivial choice of T, producing a biologically equivalent trait with epistasis or dominance. To what extent is epistasis a mathematical artifact of the choice of scale, as opposed to an inherent biological property?

First, we sketch a model of the classical approach to epistasis, connecting it with the distinction between IBD, IBS, and IBF discussed in Chapter 2. This leads us to the idea of correcting for scale with a curvature adjustment. From a statistical perspective, given a complex polygenic traits, with networks too big to untangle individual interactions, and many small interacting effects, such a curvature adjustment is a "second order effect" for deviations due to epistasis.

Next, we introduce the combinatorial approach to dominance and epistasis. Ideally, the question of whether epistasis or dominance exists as a biological phenomenon should be considered by a method invariant to monotonic transformation. This leads us to consider ordinal, rather than quantitative, traits. By specifying traits as an ordering relation or ranking over genotypes, we can ask questions about traits without reference to the scale on which they are measured. Can we choose a *natural scale* on which the trait is additive, that is, on which dominance and epistasis do not exist? Under what assumptions can such a scale be chosen exactly or approximately?

Multiple theories within quantitative genetics are more formulated with respect to additive models, and thus apply to the natural scale. The natural scale is the scale on which evolution operates in the sense of Fisher's Fundamental Theorem of Natural Selection and the Breeder's Equation. In the applied field of Genomic Selection, breeding value prediction and phenotype prediction methods work the same way under an exactly additive model, and differ when dominance is present. In the GWAS context, additivity has the implication

that effects attributed to alleles do not differ due to allele frequencies; thus effects measured on an arbitrary scale should in principle vary from population to population, but not effects measured on the natural scale.

We apply the combinatorial approach with similar methods to both dominance and epistasis. We demonstrate the role of *directional consistency of substitution* assumption and illustrate the combinatorially large number of discrete epistatic architectures which may be classified according to biologically relevant properties.

6.2 Epistasis from a Population Perspective

A more general concept of a natural scale would involve additivity of environmental, as well as genetic, contributions, and therefore homoscedasticity of the environmental contributions. We will address some of the complications that arise due to environmental effects under scale transformations, but the results we derive are most readily interpreted for fully genetic traits with negligible environmental components. Throughout this chapter we will refer to genotypic values, averaged in some way over repeated realizations of environmental variability, rather than phenotypic trait values.

6.2.1 The Correlation Function and Correlation Decay

In classical quantitative genetics, under the Kempthorne-Cockerham decomposition (Kempthorne, 1954; Cockerham, 1954), $\sigma_G(x,y)$, the genetic covariance between individuals x and y, is decomposed into additive (A^n) and dominance (D^m) covariance terms,

$$\sigma_G(x,y) = \sum (2\Theta_{xy})^n \Delta_{xy}^m \sigma^2_{A^n D^m} \tag{6.1}$$

where $r = 2\Theta_{xy}$ is the expected fraction of genome shared IBD, with values between 0 and 1, and Δ_{xy}^m is a fraternity coefficient, the probability that the two individuals share both alleles IBD. For simplicity we assume no inbreeding throughout this section. Ignoring dominance terms:

$$\sigma_G(x,y) = \sum (2\Theta_{xy})^n \sigma^2_{A^n} = \sum r^n \sigma^2_{A^n} \tag{6.2}$$

The individual variance is decomposed:

$$\sigma^2_G = \sigma_G(x,x) = \sum \sigma^2_{A^n} = \sigma^2_A + \sigma^2_{AA} + \sigma^2_{AAA} + \dots \tag{6.3}$$

From the covariance we express the genotypic correlation (that is, the correlation between genotypic values, defined e.g. as median phenotypic values for a given genotype) as a function of r, the (expected) fraction of genome shared:

$$\rho(r) = \frac{\sum r^n \sigma^2_{A^n}}{\sum \sigma^2_{A^n}} \tag{6.4}$$

Here $\rho(0) = 0$ and $\rho(1) = 1$; with positive $\sigma^2_{A^n}$, the function is monotonic. In the linear (non-epistatic) case, if $\sigma^2_G = \sigma^2_A$, $\rho(r) = r$; if $\sigma^2_G = \sigma^2_{AA}$ then $\rho(r) = r^2$.

In the pedigree context, correlation as a function of genetic distance could also be written as a function of generational distance. By doing so we use directly observable quantities, and abstract away from Mendelian concepts of inheritance embedded in the Cockerham-Kempthorne expansion. Such an abstraction is appropriate when comparing non-Mendelian patterns of resemblance, due to environmental or epigenetic causes. Thus for generational distance g, where $g = 0$ is self, $g = 1$ is parent-offspring, etc, we write

$$r = 2^{-g} \tag{6.5}$$

Then, in the case of no epistasis:

$$\rho(r) = r = \rho_g(g) = 2^{-g} \tag{6.6}$$

and the general case:

$$\rho(r) = \rho_g(g) = \rho_g\left(-\frac{\ln r}{\ln 2}\right) \tag{6.7}$$

This, in a restricted form, excluding fraternal relations, is again the Cockerham-Kempthorne expansion, and the power series coefficients on r^n in this expression correspond to the relative contributions of $\sigma^2_{A^n}$.

6.2.2 Epistasis as Curvature under IBD, IBS, and IBF

Epistasis can thus be interpreted as the nonlinearity, or more narrowly convexity, of the relationship between a fractional measure of genetic or genomic similarity, such as *relatedness*, and *genotypic value correlation*. The absence of epistasis is a linear relationship between genetic similarity and correlation, as in the case $\rho(r) = r$ above.

More generally, relatedness can be:

1. Pedigree relatedness.

   In the classical case, relatedness measures can be interpreted as probabilities of IBD gene sharing. Θ_{xy} is the probability that, choosing independently and equiprobably one of paired chromosomes from each of the two individuals x and y, the two chromosomes are IBD a given fixed point locus. In the absence of inbreeding, this probability is constrained between 0% and 50%. Thus $r = 2\Theta_{xy}$ is constrained between 0% and 100%, with 100% meaning monozygotic twin or clone, and 0% meaning unrelated relative to a fixed set of founders. The phenotypic correlation is relative to the founding population, and the same set of founders.

2. Realized relatedness (inferred IBD).

   Inferred IBD in the sense of Browning and Browning (2012) leads to relationship estimates constrained between 0% and 100% in the absence of inbreeding. Kinship estimation using the GRM, using the Powell et al. (2010) interpretation, allows negative measures of relatedness, calibrating relatedness of 0 as typically nominally unrelated within the current population. Nominal relatedness values below 0 and above 1 are difficult to interpret either as correlations or probabilities, but for typical close pedigree relatives, such as parent-offspring with $r = 0.5$ or first cousins at $r = 0.125$, the scale of GRM-derived realized relatedness measures matches its classical counterpart.

3. Fraction of genome shared IBS.

   As with the other measures of genetic similarity, it is easy to interpret monozygotic twins or clones as having 100% genome shared IBS. Outside of the artificial examples, such as the ones we will construct below, it is difficult to fix the 0% end of the scale; members of the same species share the overwhelming majority of their DNA sequence, so any IBS method must choose some set of polymorphic sites over which to evaluate sequence similarity, setting the zero point to no allele sharing on that designated set of sites. This choice is inherently arbitrary. Likewise, there is no single obvious approach for dealing with dipoloidy in the IBS context.

4. Weighted identity by function (IBF).

   The IBF framework defines a covariance between pairs of individuals on the basis of functional similarity between their genotypes. A complete genetic architecture of a trait under the IBF model is given by specifying, at each locus, a locus weight and a set of IBF parameters ($\sigma_{AA}/\sigma_{AB}, \rho_3, \rho_2$ as specified in Section 2.3.2) related to the amount of dominance. Empirically, the weights and dominance patterns are determined by variance component fitting procedures. The locus weights are related to the importance of the locus to the specific trait, rather than its size in base pairs or recombination distance. At each locus, identity states are translated to contributions to the covariance in accordance with the IBF parameters. The covariance is the weighted sum of these locus contributions. Turning the covariance into a correlation by normalizing by dividing by the individual variance, we can define a 0% to 100% IBF-based shared genome scale. 100% IBF sharing has the meaning of complete diploid identity; 0% means no identical alleles at relevant loci, using the IBF definition of functional allele identity and approach to locus classification. The correlation cannot go below zero because of the positive semidefinite properties of the IBF matrices.

6.2.3 Models of Nonlinearity

To illustrate the nonlinearity phenomenon, we will build up increasingly complex epistatic models with interaction terms. These are models of effect correlation ρ given fraction of shared genome s. As such they blur the lines between IBD, IBS, and IBF concepts. This conceptual looseness simplifies the analysis, but limits the ability to generalize the results to a more realistic setting. The models include a number of simplifying assumptions, including symmetry among loci.

Random Effect Model, No Epistasis (Haploid) There are L loci. Each of the $\binom{L}{1} = L$ loci has an additive effect $N\left(0, \sigma^2\right)$. Total effect variance is $L\sigma^2$. If the two individuals share L_S loci IBS, they have covariance $L_S\sigma^2$.

$$\rho(s) = \frac{L_S}{L} = s \tag{6.8}$$

Random Effect Model, Pair Interactions (Haploid) There are L loci.

Each of the $\binom{L}{1}$ loci has an additive effect $N\left(0, \sigma^2\right)$.

There are also P pairs of loci with additive epistatic effect $N\left(0, \sigma_P^2\right)$ for each pair.

Total effect variance is $L\sigma^2 + P\sigma_P^2$.

If the two individuals share L_S loci and P_S IBS, they have covariance $L_S\sigma^2 + P_S\sigma_P^2$.

$$\rho\left(s\right) = \rho\left(\frac{L_S}{L}\right) = \frac{L_S\sigma^2 + \mathrm{E}[P_S|L_S]\sigma_P^2}{L\sigma^2 + P\sigma_P^2} \tag{6.9}$$

Random Effect Epistasis (Haploid) There are L loci.

Each of the $\binom{L}{1}$ loci has an effect $N\left(0, \sigma^2\right)$.

Each of the $\binom{L}{2}$ pairwise interactions has an effect $N\left(0, \sigma^2 A\right)$.

Each of the $\binom{L}{k}$ k-way interactions has an effect of size $N\left(0, \sigma^2 A^{k-1}\right)$. Thus A is a parameter governing the decay in the scale of interaction terms; each unit increase in k, the order of the interaction, corresponds to the variance of the random variable generating the Equivalently, the scale of the interaction terms is multiplied by $\sqrt{A}$.

The total effect has variance:

$$\begin{aligned}
&\left[\binom{L}{1}A^1 + \binom{L}{2}A^2 + \binom{L}{3}A^3 \ldots \binom{L}{L}A^L\right] A^{-1}\sigma^2 \\
&= \left[\left[\binom{L}{0}A^0 + \binom{L}{1}A^1 + \binom{L}{2}A^2 + \binom{L}{3}A^3 \ldots \binom{L}{L}A^L\right] - 1\right] A^{-1}\sigma^2 \\
&= \left[(1+A)^L - 1\right] A^{-1}\sigma^2
\end{aligned} \tag{6.10}$$

If there are a specific L_S loci shared between two individuals, their covariance is

$$\mathrm{Cov}\left(G_1, G_2\right) = \left[(1+A)^{L_s} - 1\right] A^{-1}\sigma^2 \tag{6.11}$$

Then

$$\mathrm{Corr}\left(G_1, G_2\right) = \frac{\left[(1+A)^{L_S} - 1\right] A^{-1}\sigma^2}{\left[(1+A)^L - 1\right] A^{-1}\sigma^2} = \frac{(1+A)^{L_S} - 1}{(1+A)^L - 1} = \frac{\left((1+A)^L\right)^{L_S/L} - 1}{(1+A)^L - 1} \tag{6.12}$$

$$\rho(s) = \frac{e^{as} - 1}{e^a - 1} \tag{6.13}$$

where $s = L_S/L$ is the shared genome fraction and the fraction

$$a = L \ln(1 + A) \approx LA \tag{6.14}$$

We plot $\rho(s)$ as Figure 6.1 for several values of a. In the limit as $a \to 0$,

$$\rho(s) = \frac{e^{as} - 1}{e^a - 1} = s + \frac{\left(s^2 - s\right) a}{2} + \cdots \to s \tag{6.15}$$

which is the linear model with no epistatic effects.

6.2.4 Binomial Sampling

We have derived the above in the IBS setting, sharing a fixed number or fraction of loci. In the IBD setting, the fraction shared is random, and the kinship expresses the expectation of the true number of shared alleles. Assuming independence, consider the case where L_R loci are sampled at random out of L:

$$L_S \sim \text{Binomial}\left(\frac{L_R}{L}, L\right) \tag{6.16}$$

$$\text{Corr}(G_1, G_2) = \frac{(1 + A)^{L_S} - 1}{(1 + A)^L - 1} \tag{6.17}$$

The MGF of Binomial(p, n) is $\text{E}(\exp[tX]) = \left(1 - p + pe^t\right)^n$, so

$$\text{E}[\text{Corr}(G_1, G_2)] = \frac{\left(1 - \frac{L_R}{L} + \frac{L_R}{L}(1 + A)\right)^L - 1}{(1 + A)^L - 1} = \frac{\left(1 + A\frac{L_R}{L}\right)^L - 1}{(1 + A)^L - 1} \tag{6.18}$$

Using the shared genome parametrization, $s = L_R/L$ is the shared genome fraction and the fraction $a = L \ln(1 + A) \approx LA$,

$$\rho(s) = \frac{(1 + As)^L - 1}{(1 + A)^L - 1} = \frac{\left(1 + \left[\exp\frac{a}{L} - 1\right] s\right)^L - 1}{e^a - 1} \tag{6.19}$$

The binomial sampling method thus depends of L, converging to the IBS expression (equation 6.13) as $L \to \infty$.

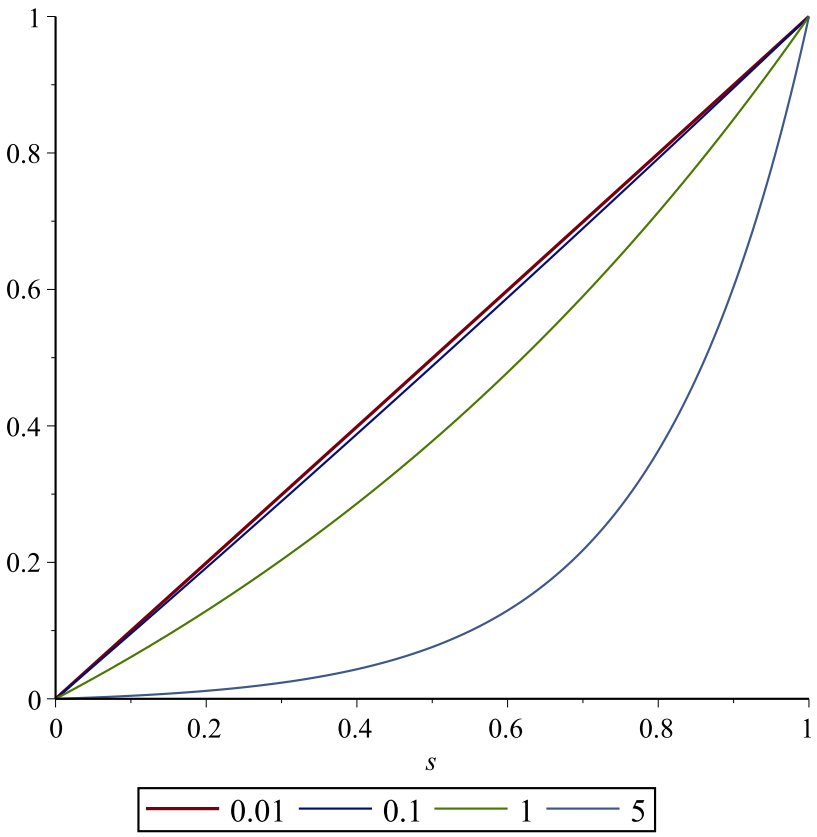


Figure 6.1: Correlation due to epistasis as a function of shared genome; colors are various values of a.

6.2.5 Implications of the Relationship Between Correlation and Shared Genome

The specific artificial model of higher order epistatic effect does not matter as much as the idea that epistasis manifests itself, in several contexts, as the curvature of the relationship between shared genome and correlation between related individuals' phenotypes. Models of covariance between relatives as described in Chapter 2, whether based on IBD or IBS or IBF concepts, and in particular the GRM, are linear in one or another measure of shared genome. Incorporating a nonlinear transformation, by fitting a nonlinear curvature adjustment for epistasis, for example fitting parameter a in 6.13 with the limit $a = 0$ as an untransformed linear shared genome, adds a parameter to potentially improve predictive performance. This sets a context for the relationship between kernels in the sense of Gianola and van Kaam (2008) capturing epistatic effects, and their non-epistatic counterparts. Curvature adjustment of the correlation function does not translate directly to a transformation of the trait to the natural scale; however, if a natural scale exists, the appropriate model of the trait on the natural scale does not have a curvature adjustment.

6.3 Ordinal Traits and Dominance

6.3.1 Phenotypes, Genotypic values, and Genotype order

Our method is based on the ordering of genotypes according to their genotypic values. We start with the assumption that the ranking of genotypes according to genotypic value is known.

We will be considering quantitative traits, but the ranking model corresponds better to a disease risk associated with a genotype, a single number, rather than a classical quantitative trait like height, in which a genotype maps to a mean with a standard deviation associated with environmental variability, A genotypic value is defined as the average of the phenotype values for a given genotype, taken over potential realizations of the trait under environmental variability. This definition is problematic when we are considering nonlinear transformations of the measurement scale of the phenotype. There is no guarantee that order of genotypic values is preserved under such transformations.

Table 6.1 illustrates this effect. We observe four phenotype values for each genotype; or,

	Phenotype Values					Average	Rank
AA	80	90	100	132		100.5	1
AB	80	100	100	120		100.0	2
BB	94	100	100	104		99.5	3
	Log Phenotype Values					Average	Rank
AA	4.382	4.500	4.605	4.883		4.592	3
AB	4.382	4.605	4.605	4.787		4.595	2
BB	4.543	4.605	4.605	4.644		4.600	1

Table 6.1: Counterexample to rank consistency of genotypic values under monotonic transformation.

alternatively, the phenotype distribution given each genotype is discrete and consists of four equiprobable values. Three genotypes are ranked $AA \succ AB \succ BB$ based on average trait value before transformation (upper table). The ordering is reversed to $BB \succ AB \succ AA$ by logarithmic transformation (lower table).

A sufficient condition for the ranking to be consistent is for the environmental variability to be homoscedastic, or additively independent. That is, on some scale, not necessarily the scale on which the trait is additive, the phenotype P can be decomposed into $P = G + E$, where G is the genotypic value and E is environmental contribution, a random variable with a distribution that does not depend on G. Without this condition, genotypes cannot generally be ranked in a consistent scale-invariant way using average phenotypic value.

If, contrary to convention, we define genotypic value as the median, rather than mean, phenotypic value, we will not encounter this problem for continuous traits. The median would not be appropriate for discrete traits encoded as continuous variables, such as diseases with binary (0 or 1) status. For discrete traits, a genotypic value ranking can be associated with a risk or hazard rate, parameters which, though subject to nonlinear transformation with respect to e.g. time scale, maintain a consistent ranking.

6.3.2 Diploid Constraints and Directional Consistency of Substitution

In the single locus, dominance context, we can identify several implications of additivity which have a biological interpretation but are invariant to a monotonic transformation of the trait's scale.

1. No overdominance (no cases such as $AA \prec BB \prec AB$)

2. No complete dominance (no cases such as $AA \prec AB \approx BB$)

3. Directional consistency of substitution ($CA \prec CB$ implies $DA \prec DB$)

Each of these features is implied by additivity, and therefore necessary for the trait to be additive on some scale. The broader question is under what circumstances are these conditions sufficient for additivity.

Consider a single diploid locus with alleles A, B, C, D. We will refer to four alleles labelled by letters for convenience of exposition; the generalization to an arbitrary number of alleles is straightforward. Each pair of alleles is associated with a trait value, and we are given an ordering relation $\succeq$ on those value. It will be convenient to assume that each genotype has a unique trait value; that is, that the genotypes can be ranked with no ties, forming a complete order relation $\succ$. We will write, e.g. $AB \succ AC$ for genotype AB has trait value higher than that of genotype AC. Assume a complete ordering; that is, all trait values are distinct (e.g. either $AD \succ BC$ or $AD \prec BC$ but not $AD \approx BC$) except matching heterozygotes ($AB \approx BA$).

Without loss of generality, we order the allele labels alphabetically according to their homozygotes,

$$AA \prec BB \prec CC \prec DD \tag{6.20}$$

The principle of non-overdominance is that for any $P, Q : PP \prec QQ$,

$$PP \prec PQ \prec QQ \tag{6.21}$$

By using strict inequalities, we also rule out complete dominance ($PQ \approx QQ$); this is consistent with the assumption of a distinct value for each genotype.

We introduce the principle of directional consistency of substitution (DCS): substituting an allele R for Q has either a positive or a negative effect, regardless of the identity of the unchanged allele. That is,

$$PQ \succ PR \rightarrow SQ \succ SR \tag{6.22}$$

Directional consistency of substitution implies non-overdominance because either $PP \succ PQ$ implying $QP \succ QQ$, or $PQ \succ PP$ implying $QQ \succ QP$. We can say DCS generalizes the idea of non-overdominance to multiple alleles.

Directional consistency of substitution implies an ordering on the individual alleles A, B, C that is the same as the ordering of homozygote genotypic values. If $PP \prec QQ$, then substitution of Q for P has a postitive effect.

Under DCS, the matrix of cardinal genotypic values expressed on any scale,

$$F = \begin{bmatrix} f_{AA} & f_{AB} & f_{AC} & f_{AD} \\ f_{AB} & f_{BB} & f_{BC} & f_{BD} \\ f_{AC} & f_{BC} & f_{CC} & f_{CD} \\ f_{AD} & f_{BD} & f_{CD} & f_{DD} \end{bmatrix} \tag{6.23}$$

is monotonically increasing both along its rows and along its columns. Directional consistency and non-overdominance can be expressed as partial orderings on the set of genotypes (given a fixed order of homozygotes), as illustrated in Figure 6.2.

6.3.3 Formulating the Linearization Question

We wish to discover the circumstances under which it is possible to express an arbitrary trait on a linear scale. That is, given a diploid trait with genotypic values T_{ij} on some arbitrary scale (for example, on a rank scale from 1 to $\binom{n}{2} + \binom{n}{1} = \frac{n^2+n}{2}$, the number of distinct genotypes for n alleles), is there a strictly increasing function f such that there exist $\alpha_1 \ldots \alpha_n$ for which

$$f\left(T_{ij}\right) = \alpha_i + \alpha_j \tag{6.24}$$

A relaxed formulation of the problem is: does there exist such a function f that is non-decreasing, but not trivially a constant? The constant-valued function $f\left(\cdot\right) = c$ is always a non-decreasing solution; a nontrivial solution would vary over the domain of $f\left(\cdot\right)$. Additivity

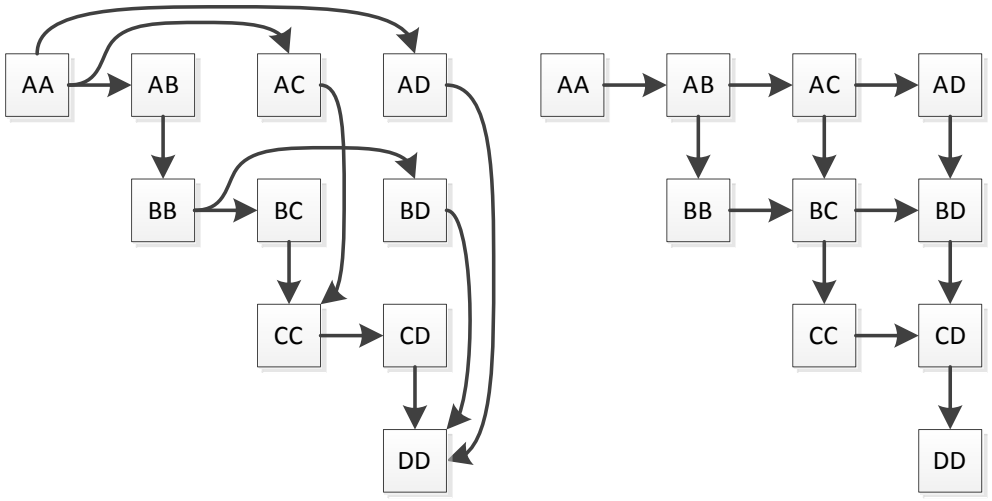


Figure 6.2: Ordering relations constraints for single diploid locus. Left: no overdominance. Right: directional consistency of substitution.

	A	B	C	D	E
A	1	2	3	5	8
B	2	4	5	7	***10***
C	3	5	6	8	11
D	5	7	8	***9***	12
E	8	***10***	11	12	13

Table 6.2: Trait with directional consistency of substitution, which provably cannot be transformed to additivity.

clearly implies directional consistency of substitution; therefore directional consistency is a necessary condition for additivity. But is it a sufficient condition, or a sufficient condition for some approximate solution?

6.3.4 Counterexample

A transformation to additivity is not always possible. We can illustrate a counterexample most readily if we forego the constraint requiring the genotypic values to be distinct. Suppose a trait T_{ij} for five alleles has the ranking given by Table 6.2.

This matrix is monotonically increasing both along rows and columns, and therefore represents a directionally consistent trait. Suppose the trait was linearizable; that is, there is an increasing function f such that $f(T_{AB}) = A + B$, etc. Then

$$A + D = B + C \tag{6.25}$$

$$C + D = A + E \tag{6.26}$$

Adding the two equations and subtracting $A + C$,

$$D + D = B + E \tag{6.27}$$

This implies $f(9) = f(10)$. We cannot find an additive solution with an increasing function f, though this particular contradiction does not prevent us from finding a non-decreasing solution, with $f(t)$ increasing over $t < 9$ and $t > 10$ and constant for $9 \leq t \leq 10$.

6.3.5 Traversal Trees

The ordering constraint on the genotypes of a single locus system imposed by non-overdominance or directional consistency of substitution, as in Figure 6.2, forms a directed acyclic graph (DAG). Any complete traversal of this DAG is an ordering of the nodes (genotypes) consistent with the ordering constraint, and therefore a trait architecture. For example,

$$AA \prec AD \prec AC \prec AB \prec BB \prec BC \prec BD \prec CC \prec CD \prec DD$$

is an ordering that satisfies non-overdominance, but not directional consistency of substitution (e.g. $AC \prec AB$ but $BB \prec BC$).

We developed an algorithm to transform the ordering constraint DAG to the *Path DAG*. This is a DAG with a single root and a single leaf. Each path from the root to the leaf corresponds corresponds to a traversal of the original DAG, with edges of the Path DAG corresponding to nodes of the original DAG (genotypes), and nodes of the Path DAG labeled with the number of distinct paths from that node to the leaf. This allows the enumeration, and therefore equally weighted sampling, of the paths through the original DAG. Figures 6.3 and 6.4 demonstrate the original and Path DAG for directional consistency of substitution for single locus diploid systems with between 2 and 7 alleles. The growth in number of directionally consistent orderings can be observed from the labeled number of paths from the root of the Path DAG.

6.3.6 Natural Scale Recovery

Given a genotype ordering, for example

$$AA \prec AB \prec BB \prec AC \prec BC \prec CC$$

we can find numerical values for a natural scale, exactly or approximately, by a linear program. We will introduce three sets of variables:

1. Genotype variables: AA, AB, BB, AC, BC, CC

   Each represents the genotypic value (e.g. median phenotype) corresponding to each genotype on a particular scale.

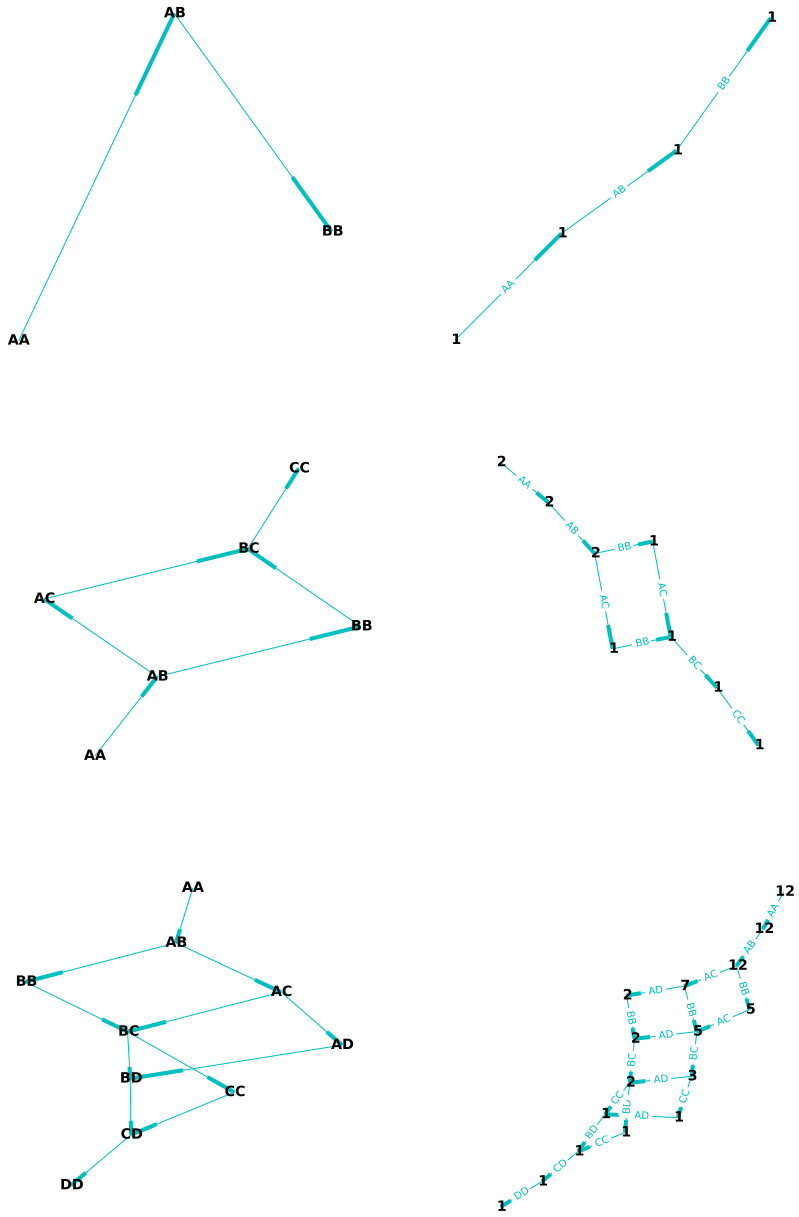


Figure 6.3: Dominance patterns for $n = 2, 3, 4$. Graph on the left is the allele ordering imposed on genotypes by directional consistency of substitution. Graph on the right is the Path DAG (Section 6.3.5).

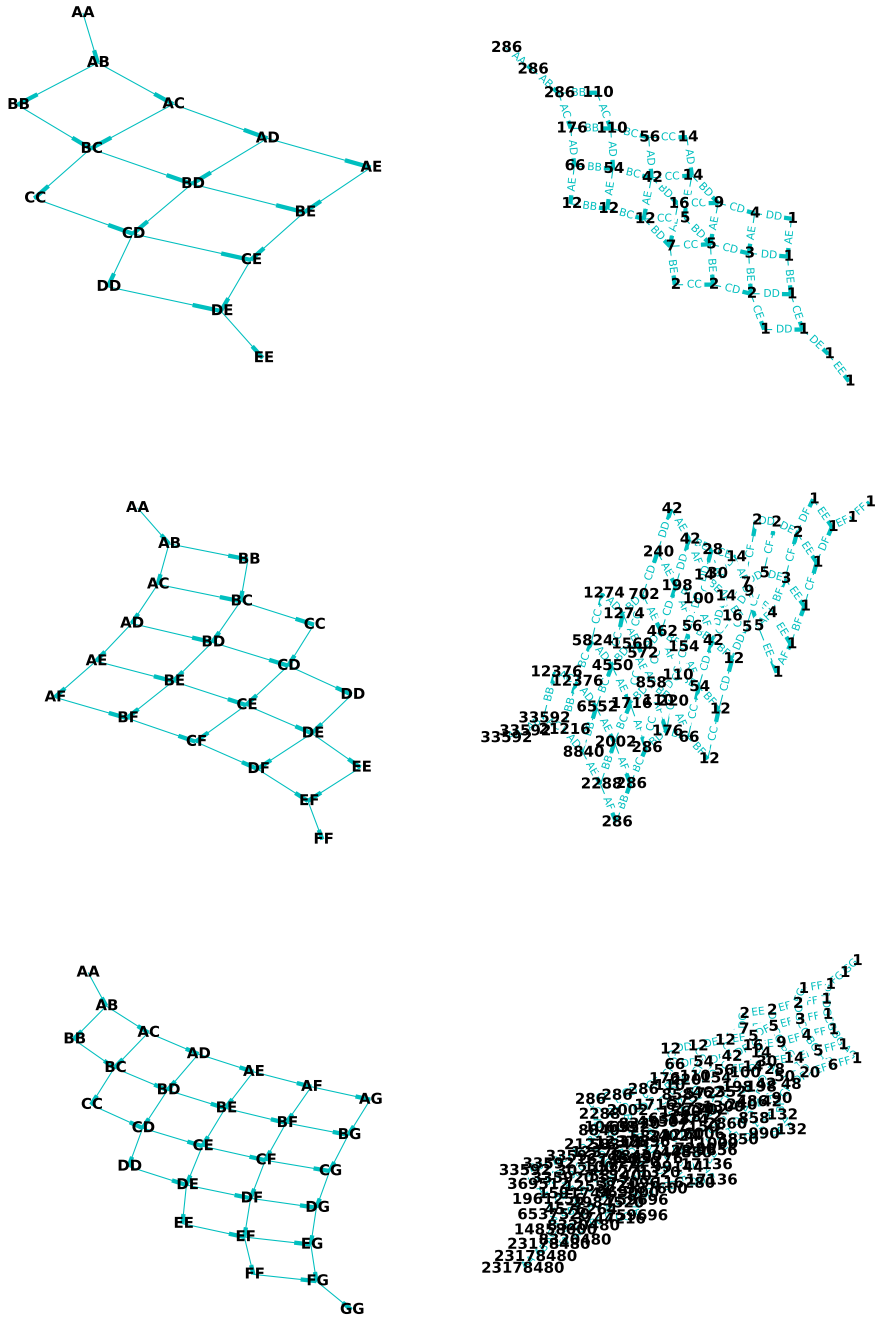


Figure 6.4: Dominance patterns for $n = 5, 6, 7$.

2. Allele variables: A, B, C

   Each represents the additive contribution of an allele under the assumption of additivity; for example $A + B = AB$.

3. Gap variables: $g_{AA}^{AB}, g_{AB}^{BB}, g_{BB}^{AC}, g_{AC}^{BC}, g_{BC}^{CC}$

   One gap variable is introduced for each non-redundant inequality in the specified ordering. A gap variable is binary, taking on the values 0 or 1, depending on whether the inequality is satisfied strictly or nonstrictly. These variables are used both in the objective and in constraints to design a linear program that seeks an additive solution where all inequalities are satisfied in the strict sense, but will fall back to an imperfect solution.

The linear program then has the form

maximize $g_{AA}^{AB} + g_{AB}^{BB} + g_{BB}^{AC} + g_{AC}^{BC} + g_{BC}^{CC}$ subject to

1. Domain: $A, B, C \geq 0$ (can be simplified to $A = 0$)

2. Linearity:

   $AA = A + A$, $AB = A + B$, $BB = B + B$, $AC = A + C$, $BC = B + C$, $CC = C + C$

3. Gap variables: $0 \leq g_{AA}^{AB}, g_{AB}^{BB}, g_{BB}^{AC}, g_{AC}^{BC}, g_{BC}^{CC} \leq 1$

4. Order:

   $AB - AA \geq g_{AA}^{AB}$, $BB - AA \geq g_{AB}^{BB}$, $AC - BB \geq g_{BB}^{AC}$, $BC - AC \geq g_{AC}^{BC}$, $CC - BC \geq g_{BC}^{CC}$

This "accordion" linear program maximizes the number of gaps of the form g_{AA}^{AB}, assigning the value of 1 to each strict inequality satisfied, and 0 to each inequality relaxed into an equality. If it can, the program will make all inequalities strict; at worst, it will turn all inequalities into equalities, producing the trivial constant solution. The use of linear programming and optimization in translating order constraints to a cardinal scale can be traced

to Kruskal (1964) and the development of the non-metric version of the multi-dimensional scaling framework.

6.3.7 Solvability of Dominance Systems

We enumerated all possible single locus systems for 2, 3 and 4 alleles, up to permutation, and displayed the solutions as 6.5. For up to 4 alleles, all but two possible order architectures are solvable, that is, have a natural scale that respects the strict inequalities of the ordering constraints. The two partially solvable orderings have relaxed or partial solutions; that is, they can be solved by turning some, but not all, or the strict inequalities into nonstrict inequalities. We tabulated results for higher numbers of alleles as Table 6.3. As the number of alleles rises, the fraction of architectures that either have relaxed solutions or no solutions increases.

6.4 Ordinal Traits and Epistasis

6.4.1 Directional Consistency in Epistasis

The concepts of the dominance framework developed above carry over directly to the case of epistasis over multiple loci, but the number of combinatorial cases, and both the computational and notational difficulties, increase considerably.

Consider a haploid organism with biallelic genotypes of k loci. The haploid genotype may be represented by a binary k-tuple with a genotypic value function of the form $f : \{0,1\}^n \rightarrow \mathcal{R}$, mapping the haploid genotype to genotypic value, such as the median phenotype without an environmental contribution. Where unambiguous, we will use the letter notation $AbCDe$ or ACD for haploid genotype $(1,0,1,1,0)$, and $abcde$ or 0 for the specific haploid genotype $(0,0,0,0,0)$. We assume a complete order on the haploid genotypes, as the possibility of ties is algebraically inconvenient. Without the loss of generality we will order the loci, and choose which alleles to call 1 or uppercase, so that $E \succ D \succ C \succ B \succ A \succ 0$.

For the multi-locus case we must distinguish between weak (or 1^{st} order) directional consistency and strong, complete, or general directional consistency. Intermediate order i directional consistency can also be defined.

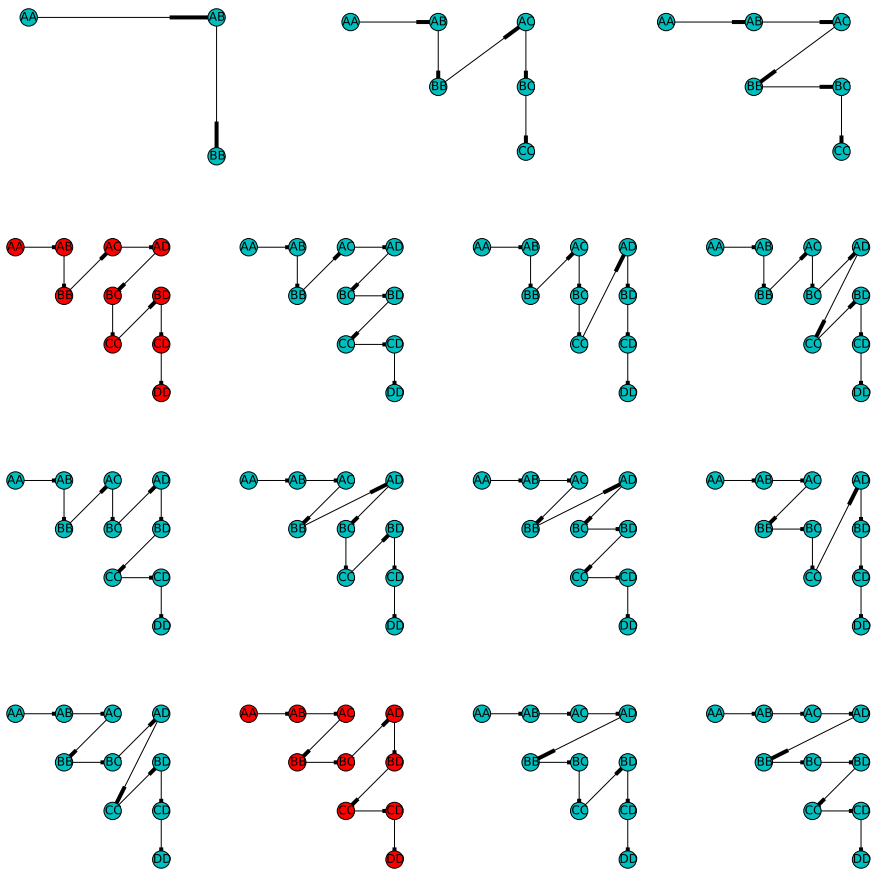
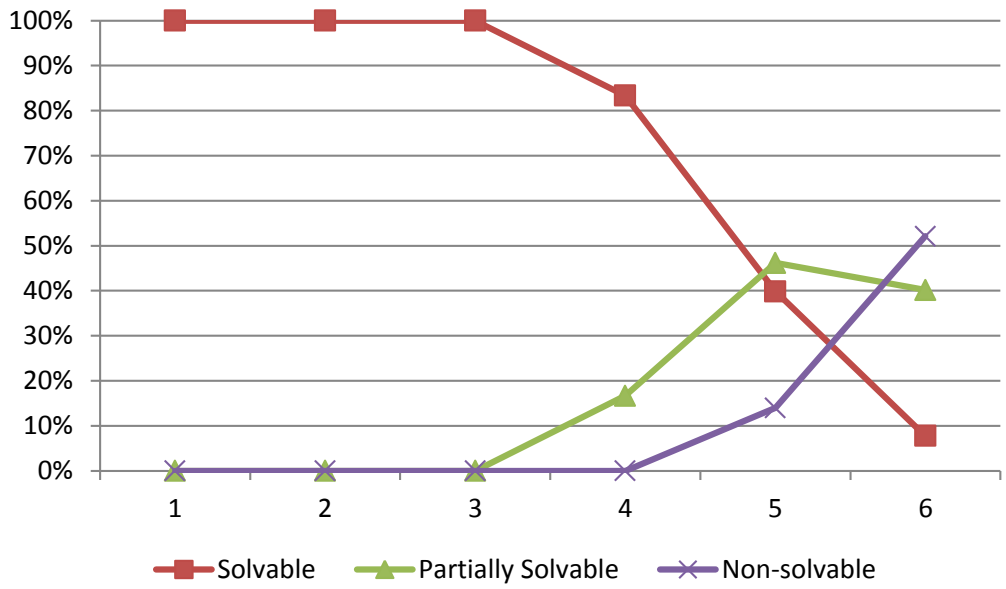


Figure 6.5: Patterns of dominance, up to label permutation, for 2, 3, and 4 alleles. Solvable systems are in green; partially solvable in red.



n	Total	Solvable	Partially	NonSolvable
1	1	1	0	0
2	1	1	0	0
3	2	2	0	0
4	12	10	2	0
5	286	114	132	40
6	33592	2608	13488	17496

Table 6.3: Solvability of Dominance Systems for $n = 1 \ldots 6$.

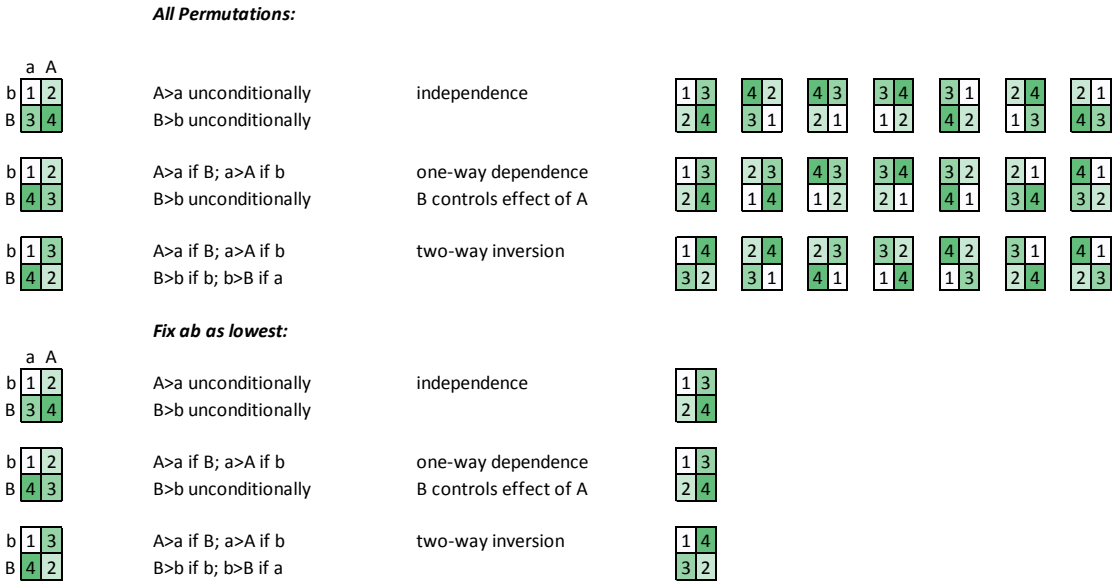


Table 6.4: General 2×2 epistatic order combinations.

Weak directional consistency is defined the invariance in the direction of the effect of the substitution at a single locus on the genotype at other loci. For example, if the substitution $c \to C|abde$ is positive, that is, if $abCde \succ abcde$, then so is the substitution $c \to C|AbDe$.

Order i directional consistency extends this to substitution at i loci at a time. For example, under second order directional consistency, if the substitution $bD \to Bd|ace$ is positive, so is $bD \to Bd|aCE$.

Strong directional consistency is directional consistency of all orders.

If the genotype is diploid, a single substitution may occur on either strand; thus a diploid, single locus model with weak directional consistency is equivalent to the directional consistency model of dominance. Additivity implies strong directional consistency, and strong directional consistency implies weak directional consistency, but neither converse holds. Weak directional consistency as an ordering constraint can be represented as a DAG over genotypes for a fixed ordering of single "uppercase" locus genotypes, but strong directional

```
abcd Abcd aBcd ABcd abCd AbCd aBCd ABCd abcD AbcD aBcD ABcD abCD AbCD aBCD ABCD
strict: [1, 2, 4, 8]

abcd Abcd aBcd ABcd abCd AbCd aBCd ABCd abcD AbcD aBcD abCD ABcD AbCD aBCD ABCD
strict: Infeasible
relaxed: [1, 2, 3, 6]
abcd Abcd aBcd ABcd abCd AbCd aBCd ABCd abcD AbcD aBcD abCD ABcD AbCD aBCD ABCD
{'~~~d': ['ABc', 'abC'], '~~~D': ['abC', 'ABc']}

abcd Abcd aBcd ABcd abCd AbCd aBCd abcD AbcD ABCd aBcD ABcD abCD AbCD aBCD ABCD
strict: Infeasible
relaxed: [0, 1, 2, 3]
abcd Abcd aBcd ABcd abCd AbCd aBCd abcD AbcD ABCd aBcD ABcD abCD AbCD aBCD ABCD
{'A~~~': ['bcD', 'BCd'], 'a~~~': ['BCd', 'bcD']}

abcd Abcd aBcd ABcd abCd AbCd aBCd abcD AbcD ABCd aBcD abCD ABcD AbCD aBCD ABCD
strict: Infeasible
relaxed: [1, 2, 3, 5]
abcd Abcd aBcd ABcd abCd AbCd aBCd abcD AbcD ABCd aBcD abCD ABcD AbCD aBCD ABCD
{'A~~~': ['bcD', 'BCd'], 'a~~~': ['BCd', 'bcD']}
{'~~~d': ['ABc', 'abC'], '~~~D': ['abC', 'ABc']}

abcd Abcd aBcd ABcd abCd AbCd aBCd abcD AbcD aBcD abCD ABCd ABcD AbCD aBCD ABCD
strict: Infeasible
relaxed: Infeasible
abcd Abcd aBcd ABcd abCd AbCd aBCd abcD AbcD aBcD abCD ABCd ABcD AbCD aBCD ABCD
{'A~~~': ['bcD', 'BCd'], 'a~~~': ['BCd', 'bcD']}
{'~B~~': ['acD', 'ACd'], '~b~~': ['ACd', 'acD']}
{'~~c~': ['ABd', 'abD'], '~~C~': ['abD', 'ABd']}
{'~~~d': ['ABc', 'abC'], '~~~D': ['abC', 'ABc']}

abcd Abcd aBcd ABcd abCd AbCd aBCd abcD AbcD aBcD ABCd ABcD abCD AbCD aBCD ABCD
strict: Infeasible
relaxed: Infeasible
abcd Abcd aBcd ABcd abCd AbCd aBCd abcD AbcD aBcD ABCd ABcD abCD AbCD aBCD ABCD
{'A~~~': ['bcD', 'BCd'], 'a~~~': ['BCd', 'bcD']}
{'~B~~': ['acD', 'ACd'], '~b~~': ['ACd', 'acD']}
```

Table 6.5: Output of Violation Classification Algorithm.

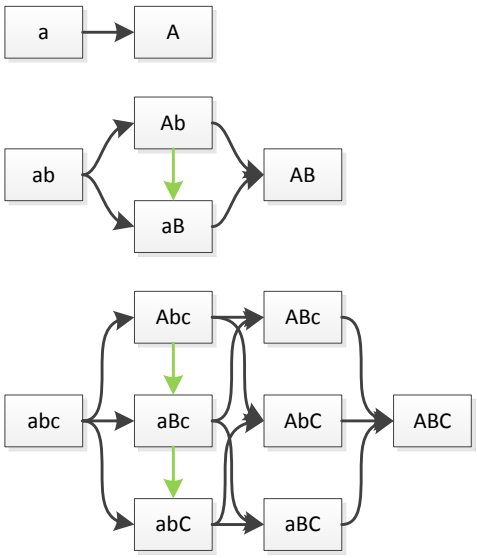


Figure 6.6: Order constraints on epistatic systems with weak directional consistency for 1, 2, or 3 loci. Green indicates ordering of loci imposed by label permutation.

consistency cannot. Directional consistency, in a relatively weak or strong form, is still a weaker and more biologically plausible hypothesis about a given trait than additivity. We plot the order constraints due to weak directional consistency as Figures 6.6 and 6.7.

6.4.2 Deviations from Directional Consistency

Consider 2×2 tables assigning rank values to the 2-locus, haploid system $\{A, a\} \times \{B, b\}$. There are $4! = 24$ possible orderings of the 4 cells of the 2×2 table. Without loss of generality, within-locus allelic types can be assigned to make ab the lowest ranked value, leaving $3! = 6$ orderings. Likewise the labeling of A and B can ensure $Ab \prec aB$. Only $3!/2 = 3$ distinct rankings remain. For 2 loci, this problem has a decomposition into 3 cases; but with even 3 loci, the number of cases increase to 840. Table 6.6 illustrates this phenomenon. For a system with n (biallelic haploid) loci, there are 2^n genotypes, and therefore $2^n!$ possible orders. Eliminating patterns equivalent up to symmetry, within-locus

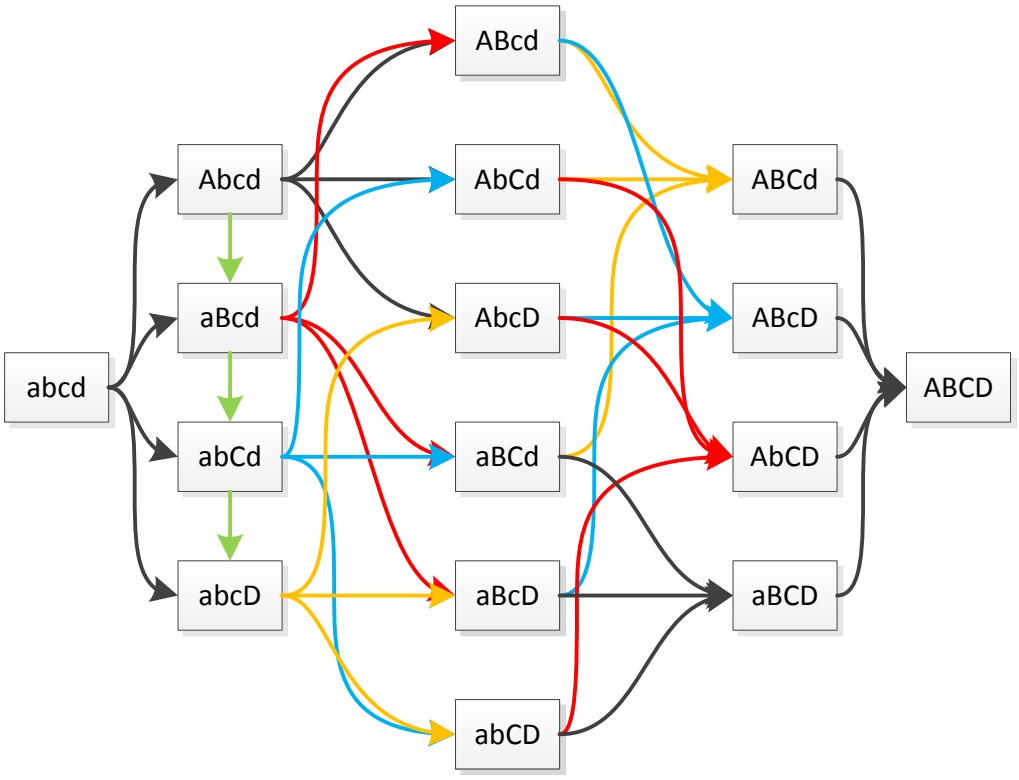


Figure 6.7: Order constraints on epistatic systems with weak directional consistency for 4 loci. Green indicates ordering of loci imposed by label permutation. Other colors arbitrary, for contrast only.

allelic types can be assigned to make $ab\ldots$ the lowest ranked value, and the labeling of $A, B, \ldots$ chosen to ensure $Ab\cdots \prec aB\ldots$. This reduces this count of distinct patterns by a factor of $2^n n!$.

Table 6.4 illustrates the 3 patterns within the 2 locus system:

1. Pattern Z, directional consistency: $ab \prec Ab \prec aB \prec AB$

2. Pattern Π, one-way inversion: $ab \prec Ab \prec AB \prec aB$

3. Pattern X, two-way inversion: $ab \prec AB \prec Ab \prec aB$

Pattern Z is the previously considered directionally consistent form. $a \prec A$ and $b \prec B$, regardless of the state of the other gene. In Pattern X, switching either gene affects the other. The asymmetric character of pattern Π is the biologically interesting case, where one gene appears to control the action of another; this corresponds to the "robustness gene" hypothesis. One of the genes, A, can switch the direction of its estimated effect in the GWAS context, depending on the population frequency of B.

Consider a robustness gene R controlling a downstream client gene C. When the robustness gene is in the active state, the phenotype range is narrow, $f(Rc) = 99$ and $f(RC) = 101$; in the passive state, the phenotype range is wide, $f(rc) = 80$ and $f(rC) = 120$. Thus, the ordering is $rc \prec Rc \prec RC \prec rC$ and the pattern is Π. Conditional on either R or r, $c \prec C$; but $r \prec R|c$ and $R \prec r|C$.

Note that even though biologically we would say R controls C, the direction of effect of R depends on C, not the other way around. In the GWAS setting, the effect of the $c \to C$ substitution is always positive. The effect of the $r \to R$ substitution changes sign depending on the population frequency of C. Thus, even though the R gene can be interpreted as more important, the C gene can be easier to detect by GWAS, and easier to replicate across different populations.

6.4.3 Algorithmic Solutions

We developed an algorithm to solve epistatic order problems for natural scale, exactly or approximately, by linear programming, as well as to detect violations of directional

n	Possible orders ($2^n!$)	Up to symmetry
1	2	1
2	24	3
3	40320	840
4	2.09×10^{13}	5.45×10^{10}
5	2.63×10^{35}	6.85×10^{31}
6	1.27×10^{89}	2.75×10^{84}
7	3.86×10^{215}	5.98×10^{209}

Table 6.6: Counts of general $n \times n$ epistatic order combinations.

consistency. For selected orderings, the output is illustrated as Table 6.5. As an example showing how to read this table, take the second block. The first row is an ordering of haploid genotypes, in increasing order from *abcd* to *ABCD*. A strict solution, with no relaxation of strict inequalities, is impossible; but the maximum number of inequalities are satisfied in the best nontrivial solution, assigning $A = 1, B = 2, C = 3, D = 6$. The last row specifies the violation of directional consistency in the order given: conditioning on either d or D, the ordering or *ABc* and *abC* flips, a violation of third order consistency.

6.4.4 Separability Theory and General Result

Separability, or additive utility, theory, concerns the conditions under which a weak order over tuples can be expressed as the order of sums of real-valued functions of the components of the tuples. The key early result in separability theory are due to Debreu (1960), in the context of a topological method with continuous variables applied to economic utility theory, and Luce and Tukey (1964) applied to measurement theory. Modern research efforts on the subject are those of Wakker (1989) and Gonzales (1996).

Our interest is in additivity over finite, discrete sets. We follow a result given as Theorem 4.1 in Fishburn (1970) in the utility theory context, translating terminology to the genetic. That result has three forms (A, B, and C), which apply to different equivalence relations

within the partial order (equality, equivalence, and indifference). The proof of this result can be based on, equivalently, the Theorem of the Alternative, Farkas' Lemma, or the Separating Hyperplane Theorem.

Translating the notation to our setting, consider genotypes as multisets of alleles. This covers all cases of interest: single locus diploid as in the dominance model, and haploid or diploid multi-locus epistasis model. Each locus $i = 1 \ldots L$ has a disjoint set of alleles V_i. A diploid genotype over L loci contains $2L$ alleles, two from each V_i, possibly the same allele repeated twice. The genotypic value function is given by a ranking $\succ$ over all realizable genotypes, without ties. The additivity question takes the form: under what circumstances can we assign real values to the alleles, $f(\cdot) : \bigcup_i V_i \to \Re$, so that for all genotypes X and Y,

$$X \succ Y \text{ iff } \sum_{x \in X} f(x) > \sum_{y \in Y} f(y) \tag{6.28}$$

Following the result in Fishburn (1970), a linear solution $f(\cdot)$ exists if and only if:

There *does not* exist a paired list of genotypes $A_1 \ldots A_n, B_1 \ldots B_n$ such that

1. The multiset unions of the two lists are the same, $\bigcup_i A_i = \bigcup_i B_i$.

2. All $A_i \succeq B_i$, and at least one $A_i \succ B_i$; without ties, this simplifies to all $A_i \succ B_i$.

We can state the principle in several different ways.

1. In the non-epistatic case, if two populations have identical total gene content, one cannot dominate the other.

2. For an additive trait, reshuffling genotypes among individuals cannot increase the average trait value in a population.

3. A trait possesses essential, non-removable epistasis if and only if, by moving alleles between individuals within the population without adding or removing any, the trait values of some individuals can be increased with none decreased.

In Section 6.3.4, we showed that a specific trait cannot be additive. The argument we used was a special case of the general result in this section. In the terminology of this section a linear solution for that trait does not exist, because there exist paired lists of genotypes, $[AD, CD, BE]$ and $[BC, AE, DD]$, meeting the two criteria:

1. The multiset unions of the two lists are the same, $\bigcup_i A_i = \bigcup_i B_i = \{A, B, C, D, D, E\}$.

2. All $A_i \succeq B_i$, and at least one $A_i \succ B_i$; here $AD \approx BC$ and $CD \approx AE$ but $BE \succ DD$.

If, for a given trait specified by an ordering, we can find such a counterexample (i.e. paired lists of genotypes satisfying the two criteria), assuming the existence of an additive representation readily leads to a contradiction. The result referred to in this section is that the non-existence of such a counterexample is not only necessary, but sufficient for the existence of an additive representation. Finding a counterexample by exhaustive search is not a simple problem, due to the number of combinations for even moderate numbers of loci.

Chapter 7

METHODOLOGY EXTENSIONS: LINKAGE DISEQUILIBRIUM AND THE GENOMIC RELATIONSHIP MATRIX

7.1 Introduction

The Genomic Relationship Matrix (e.g. Hayes et al., 2009; VanRaden, 2008) and extension methods (e.g. Thornton et al., 2012) can be interpreted as estimating kinship for each of the n^2 pairs formed by n individuals genotyped at L polymorphic, biallelic SNPs. Analogous methods have recently been proposed (Vitezica et al., 2013) for the estimation of the coefficient of fraternity. This approach can in principle extend the method to the estimation of other probabilities of pairwise identity by descent, in particular the Jacquard coefficients. The interpretation of GRM-derived relatedness estimates as IBD probabilities is made difficult by the possibility of negative estimates. Powell et al. (2010) discusses this question in light of the related questions of choosing an appropriate reference (founder) population and estimating allele frequencies appropriate to this population.

Existing approaches, implicitly or explicitly, treat the L SNP loci as independent, or unlinked and in linkage equilibrium, and invariant to permutation of locus order.

The two phenomena necessarily neglected by such approximations are linkage and linkage disequilibrium. The two terms are sometimes blurred in the literature; we will define them in a way that emphasizes the contrast. Linkage is absent when alleles at different loci segregate independently during a breeding event. Linkage disequilibrium is absent (i.e. linkage equilibrium is attained) if, when sampling nominally unrelated individuals from *a population*, observations of the allele states at two loci are statistically independent. Linkage disequilibrium may be present for loci whether or not they are on the same chromosome, and may be due to any phenomena of population history. Linkage is biologically excluded for loci on different chromosomes.

The method we develop deals with linkage disequilibrium, and does not explicitly ref-

erence linkage. We select a subset of SNPs (a nonnegative weighting, with some SNPs excluded by being assigned a weight of 0). The choice of weights is optimal in minimizing the variance of the kinship estimate under particular circumstances, while retaining the unbiasedness property of the base GRM kinship estimator.

Model misspecification ignoring the presence of linkage disequilibrium may both reduce the efficiency of estimation, and introduce biases in accounting for cryptic relatedness. Linkage disequilibrium, like allele frequencies, varies among subpopulations. Both population structure and admixture can manifest in the form of LD. Unlike other phenomena that can be neglected in approximations and overwhelmed by an increase in the quality and resolution of genotype data, LD becomes increasingly problematic with dense (e.g. 1,000,000 SNP) panels. Dense panels contain physically close SNPs, which form blocks within haplotypes which have never been broken up in the history of one branch of the population. It is not clear that increasing SNP density, especially approaching full genome sequence by introducing increasingly rare variants, will thus result in better kinship estimates without appropriate LD adjustment.

Aside from the importance of linkage disequilibrium as a general phenomenon in all areas of genetic analysis, we can identify some distinct phenomena related to the use of the GRM with modern SNP panels:

1. If several loci are highly linked and in perfect linkage disequilibrium, and in effect assort as though they were one, we should not "double count" them. If several nearby loci that always assort together are in a particular identity state, that is less rare, and less informative about relatedness, than if distant loci, which do not always assort together, happen to be in the same identity state. This is especially relevant with highly dense SNP panels and structured (i.e. distantly related groups of close relatives) populations. When there are selectively favored alleles of a single gene, nearby nonfunctional SNP markers may be swept along forming haplotype blocks.

   A perfect LD block, a set of SNPs with pairwise $R^2 = 1$, should be treated as a single SNP. This can be accomplished either by arbitrarily choosing one of the SNPs to represent the block, or by downweighting each SNP in the block to add up to the

effect size of one independent SNP.

The method's treatment of an approximately perfect LD, with some imperfect correlation $R^2 \approx 1$, should approach such behavior.

2. The choice of SNP coverage may be driven by factors such as technological feasibility and the availability of information about polymorphisms from prior studies. In the human context, this may involve specific biases such as ethnic and geographic biases of ascertainment. Different genome regions have different densities of SNP coverage in any panel, however perfectly constructed.

   An optimal selection or weighting of SNP's would be expected to cancel out the most overt effects of such arbitrariness. Individual SNPs in a densely covered region receives a comparatively lower weight, to account for redundancy, and correspondingly, the few SNPs representing a sparsely covered region are weighted higher. The formal LD adjustment corresponds to the intuitive notion of equalizing weights across the genome map. Such equalization is neither with respect to recombination distance nor physical distance, but would approach either under simple conditions.

3. Rare (in the low minor allele frequency sense) SNPs, which are increasingly common (in the sense of making up a large fraction of the list of all SNPs) in dense high-coverage panels approaching full genome sequence, can be highly influential. The effect of the improbable rare homozygote is particularly large, and the estimator is sensitive to genotyping errors that falsely indicate this rare homozygote. Without genotyping error, we demonstrate that such SNPs can account for a disproportionate amount of the kurtosis of the genome-wide estimator.

   This effect is especially a problem when many rare correlated SNPs occur together in a block, as is the case if one family in the sample shares a small segment with ancestry distant from the rest of the sample. The SNP's in the segment of distant ancestry would appear rare if allele frequencies were estimated for the sample's primary population. The estimator will be sensitive to just how many SNPs in such a block

happen to be included in a panel. While we do not propose a way to combine a penalty for rare allele frequency with an LD adjustment, our LD adjustment does resolve the problem of redundant loci, whether or not they are rare.

7.2 The GRM estimator

The single-locus GRM kinship estimator is

$$X_l = c_{ijl} = \frac{(x_{il} - 2p_l)(x_{jl} - 2p_l)}{2p_l(1 - p_l)} \tag{7.1}$$

Here i and j identify two individuals (the M in GRM, standing for Matrix, refers to the matrix of these evaluated across pairs of individuals). The locus is designated by l, and p_l is the major allele frequency, a population parameter which may be estimated as a raw sample frequency or by methods accounting for e.g. known relatedness or population structure. The x_{il} and x_{jl} are the genotypes encoded as major allele dosages, 0, 1, or 2, for the respective individuals.

Defining

$$z_{il} = \frac{x_{il} - 2p_l}{\sqrt{2p_l(1 - p_l)}} \tag{7.2}$$

we can write the normalized (correlation) form

$$X_l = z_{il} z_{jl} \tag{7.3}$$

Suppose the population at locus l deviates from Hardy-Weinberg equilibrium to an extent captured in inbreeding coefficient F, in the classical (e.g. Wright, 1921) sense; that is, the marginal distribution of x_{il} is

$$\Pr(x_{il} = 0) = (1 - F)(1 - p_l)^2 + F(1 - p_l) \tag{7.4}$$

$$\Pr(x_{il} = 1) = (1 - F)\,2p_l(1 - p_l) \tag{7.5}$$

$$\Pr(x_{il} = 2) = (1 - F)\,{p_l}^2 + Fp_l \tag{7.6}$$

We will evaluate the properties of the estimator as applied to unrelated individuals. We make the simplifying assumption that individuals are independent and identically distributed. In particular, $x_{il} \perp x_{jl}$ and $x_{il} \perp x_{jm}$. To introduce linkage disequilibrium, we

specifically do not assume that $x_{il} \perp x_{im}$; there is independence among individuals but not along the genome. We have assumed an inbreeding coefficient F, as a parameter of the joint distribution of two SNPs within the same individual. We assume independence between individuals. This would not be the case, for example, in a Wright-Fisher model where individuals are pairs of alleles in a finite population. This approach resolves, implicitly, the ambiguity between linkage disequilibrium and cryptic relatedness, by categorizing everything captured by the covariance structure as linkage disequilibrium.

Following classical results on inbreeding, we can compute the moments of the elementary variables:

$$\mathrm{E}\left[x_{il}\right] = (1-F)\,2p_l\,(1-p_l) + 2\,(1-F)\,{p_l}^2 + 2Fp_l = 2p_l \tag{7.7}$$

$$\mathrm{E}\left[z_{il}\right] = 0 \tag{7.8}$$

$$\mathrm{Var}\left[z_{il}\right] = \mathrm{E}\left[z_{il}^2\right] = \frac{1}{2p_l\,(1-p_l)} \left[\begin{array}{c} \left[(1-F)\,(1-p_l)^2 + F\,(1-p_l)\right]\,[0-2p_l]^2 + \\ (1-F)\,2p_l\,(1-p_l)\,[1-2p_l]^2 + \\ \left[(1-F)\,{p_l}^2 + Fp_l\right]\,[2-2p_l]^2 \end{array} \right] = 1+F \tag{7.9}$$

Applying the definition of covariance formally,

$$\mathrm{Cov}\left(X_l, X_m\right) = \mathrm{E}\left[X_l X_m\right] = \mathrm{E}\left[z_{il} z_{jl} z_{im} z_{jm}\right] \tag{7.10}$$

Taking $l = m$,

$$\mathrm{Var}\left(X_l\right) = \mathrm{E}\left[z_{il} z_{jl} z_{il} z_{jl}\right] = \mathrm{E}\left[z_{il} z_{il} z_{jl} z_{jl}\right] = \mathrm{E}\left[z_{il} z_{il}\right]\mathrm{E}\left[z_{jl} z_{jl}\right] = (1+F)^2 \tag{7.11}$$

Exploiting the assumed independence among individuals,

$$\begin{aligned} \mathrm{Cov}\left(X_l, X_m\right) &= \mathrm{E}\left[z_{il} z_{im} z_{jl} z_{jm}\right] = \mathrm{E}\left[z_{il} z_{im}\right]\mathrm{E}\left[z_{jl} z_{jm}\right] \\ &= \sigma_{z_{il}} \sigma_{z_{il}} \rho_{z_{il};z_{im}} \sigma_{z_{jl}} \sigma_{z_{jl}} \rho_{z_{jl};z_{jm}} \end{aligned} \tag{7.12}$$

By the iid assumption, the distribution of z_{il} and the pairwise joint distribution (z_{il}, z_{im}) does not depend on the choice of i. Since z_{il} is affine in x_{il}, $\rho_{z_{il};z_{im}} = \rho_{x_{il};x_{im}}$, and is simply ρ_{lm}, the dosage correlation between the two loci. Further, $\sigma_{z_{il}} = \sqrt{\mathrm{Var}\left[z_{il}\right]} = \sqrt{1+F}$, giving

$$\mathrm{Cov}\left(X_l, X_m\right) = \sigma_{z_{il}} \sigma_{z_{il}} \rho_{z_{il};z_{im}} \sigma_{z_{jl}} \sigma_{z_{jl}} \rho_{z_{jl};z_{jm}} = (1+F)^2 \rho_{lm}^2 \tag{7.13}$$

Conveniently, ρ_{lm} only enters the expression in a squared form, so the covariance is invariant of the choice of the minor allele, and is always non-negative. Since $(1+F)^2$ enters the optimization only as a scaling factor, we do not actually need to estimate F to proceed. A correlation matrix is positive semidefinite; squaring a matrix memberwise preserves the positive semidefinite property (Schur, 1911, p. 14, Theorem VII).

We assume that the squared correlation/LD matrix $\mathbf{R} = \left[\rho_{lm}^2\right]_{lm}$ is known. The literature on covariance and correlation matrix estimation is extensive, and includes sparse, penalized, and rank limited methods. Given the large number of loci in a typical problem, we can approach the problem hierarchically, starting with treating the matrix as blockwise diagonal.

7.3 Weighted estimates

Across L loci, where L is large (tens of thousands to millions) for a current generation SNP array, the multi-locus, ensemble, or genome-wide estimator of kinship has the form

$$Y = \sum_{l=1}^{L} w_l X_l \tag{7.14}$$

As X_l is unbiased, Y is unbiased so long as $\sum_{l=1}^{L} w_l = 1$. The two typical weightings (VanRaden, 2008) are

$$A = \frac{1}{L}\sum_{l=1}^{L} \frac{(x_{il}-2p_l)(x_{jl}-2p_l)}{2p_l(1-p_l)} = \sum_{l=1}^{L}\frac{1}{L}X_l \tag{7.15}$$

$$B = \frac{1}{L}\frac{\sum_{l=1}^{L}(x_{il}-2p_l)(x_{jl}-2p_l)}{\sum_{l=1}^{L}2p_l(1-p_l)} = \sum_{l=1}^{L}\left[\frac{2p_l(1-p_l)}{L\sum_{m=1}^{L}2p_m(1-p_m)}\right]X_l \tag{7.16}$$

Both of these can be considered as weighted combinations of X_l; A weighs all loci equally, and B weighs loci proportionately to $p_l(1-p_l)$. The latter choice emphasizes SNPs with MAF near $\frac{1}{2}$, and downweights rare alleles. In Section 7.6 we will show that rare alleles have a disproportionally high contribution to the kurtosis of the estimator, and this weighting has the desirable characteristic of penalizing kurtosis.

Using the moment expressions derived above, we can derive the variance of Y for a given

set of weights w_l, for the case of estimating kinship between independent individuals:

$$\mathrm{Var}\,(Y) = (1+F)^2 \sum_{a=1}^{L}\sum_{b=1}^{L} w_a w_b R_{ab} \tag{7.17}$$

By choosing weights that minimize this variance, we find the estimator that most efficiently confirms that unrelated individuals are unrelated. These weights are thus not guaranteed to be optimally efficient for kinships other than 0.

7.4 Optimization

7.4.1 General and Non-Negative Solutions

We wish to find a set of locus weights w_l that solves the optimization problem

$$\min_{w} \sum_{a=1}^{L}\sum_{b=1}^{L} w_a w_b \mathbf{R}_{ab} \;\; : w_l \geq 0, \sum_{l=1}^{L} w_l = 1 \tag{7.18}$$

Imposing non-negativity has desirable characteristics, such as sparsity (some weights will be set to 0), interpretability, reduced sensitivity to noise, and ensuring that $[Y]_{ij}$, considered as a matrix over pairs of individuals, is positive semidefinite. Nonetheless, allowing negative weights is mathematically plausible. The problem can also be formulated and solved, without the non-negativity constraint, by Lagrange multipliers:

$$\mathbf{w'Rw} - \lambda\left(\mathbf{w'1} - 1\right) = \mathbf{0} \tag{7.19}$$

differentiating to solve for the optimal λ and $\mathbf{w}$:

$$2\mathbf{Rw} = \lambda\mathbf{1} \rightarrow \mathbf{w} = \frac{\lambda}{2}\mathbf{R}^{+}1 \tag{7.20}$$

$$\mathbf{1'w} = 1 = \frac{\lambda}{2}1'\mathbf{R}^{+}\mathbf{1} \rightarrow \frac{\lambda}{2} = \frac{1}{\mathbf{1'R}^{+}\mathbf{1}} \tag{7.21}$$

giving the solution

$$\mathbf{w} = \frac{\mathbf{R}^{+}\mathbf{1}}{\mathbf{1'R}^{+}\mathbf{1}} \tag{7.22}$$

The unconstrained solution is non-negative in simple cases, such as symmetric or diagonal matrices $\mathbf{R}$.

7.4.2 Diagonal Solution

Although the only form of $\mathbf{R}$ we have mentioned has been a squared correlation matrix, it is convenient to consider $\mathbf{R}$ as an arbitrary positive semidefinite matrix. If R is a general diagonal matrix, the solution is inverse variance (precision) weighting:

$$\mathbf{w} = \frac{\mathbf{R}^{+}\mathbf{1}}{\mathbf{1}'\mathbf{R}^{+}\mathbf{1}} \rightarrow \mathbf{w}_i = \frac{(\mathbf{R}_{ii})^{-1}}{\sum_{i=\mathbf{1}}^{L}(\mathbf{R}_{ii})^{-1}} \tag{7.23}$$

7.4.3 Quadratic Programming Solution

To avoid the computational overhead of a general purpose quadratic programming solver, the special form of the optimization problem can be exploited. The optimization is a quadratic program over a simplex, a combination of an equality and multiple inequality constraint.

To solve with a box (specifically, non-negativity) constraint, we first observe that a constrained solution for any positive value of $\sum_{l=1}^{L}\mathbf{w}_l$, not just 1, is acceptable up to rescaling. Choosing an arbitrary constraint of the form $\sum_{l=1}^{L} w_l = C$ is not the same as removing that constraint altogether; without the constraint, the optimal solution is simply $\mathbf{w} = \mathbf{0}$. We choose, arbitrarily, $\lambda = 1$, which corresponds to an optimal solution for some unknown C, and solve the quadratic problem without equality constraints,

$$\min_{\mathbf{w}} \mathbf{w}'\mathbf{R}\mathbf{w} - \mathbf{w}'\mathbf{1} \ : \mathbf{w}_l \geq 0 \tag{7.24}$$

This yields a set of unnormalized weights; rescaling them to $\mathbf{w}'\mathbf{1} = 1$ gives the solution.

7.4.4 Blockwise Diagonal Solution

Suppose the matrix $\mathbf{R}$ is blockwise diagonal. We solve each block $\mathbf{R}_{(i)}$ for the restricted $\mathbf{w}_{(i)}$, such that $\mathbf{1}'\mathbf{w}_{(i)} = 1$ and $V_{(i)} = \mathbf{w}'_{(i)}\mathbf{R}\mathbf{w}_{(i)}$ is minimized. Then, analogously to the diagonal solution, the global optimum $\mathbf{w}$ is the concatenation of weighed block optimal weights,

$$\frac{V_{(i)}^{-1}}{\sum_{(j)} V_{(j)}^{-1}}\ \mathbf{w}_{(i)}$$

7.5 Properties of Solutions and Degrees of Freedom

Can we account for the number of degrees of freedom yielded by the weighting scheme? If we have 100 nearby SNPs which are completely linked, and in complete linkage disequilibrium, we have less information than if we have 10 SNPs on different chromosomes which assort independently and are not associated due to population structure. We formalize this notion by constructing a measure of effective independent factors.

When there are $L = n$ loci in linkage equilibrium, the optimal weighting is the equal weighting $w_l = \frac{1}{n}$ by an application of equation 7.23, and $\mathbf{R} = \mathbf{I}$, so the optimal ensemble estimator variance is

$$\mathbf{w}'\mathbf{R}\mathbf{w} = \left(\frac{1}{n}\mathbf{1}\right)' \left(\frac{1}{n}\mathbf{1}\right) = n\frac{1}{n^2} = \frac{1}{n} \tag{7.25}$$

We define the equivalent number of degrees of freedom by analogy with this case, as the number of independent SNPs necessary to achieve as low a variance as that achieved by given weighting $\mathbf{w}$:

$$df\left(\mathbf{w}, \mathbf{R}\right) = \frac{1}{\mathbf{w}'\mathbf{R}\mathbf{w}} \tag{7.26}$$

For a block in perfect linkage disequilibrium, $\mathbf{R} = \mathbf{1}\mathbf{1}'$, so regardless of the choice of $\mathbf{w}$,

$$df\left(\mathbf{w}, \mathbf{R}\right) = \frac{\mathbf{1}}{\mathbf{w}'\mathbf{R}\mathbf{w}} = \frac{\mathbf{1}}{\mathbf{w}'\mathbf{1}\mathbf{1}'\mathbf{w}} = \frac{\mathbf{1}}{\left(\mathbf{w}'\mathbf{1}\right)\left(\mathbf{1}'\mathbf{w}\right)} = 1 \tag{7.27}$$

consistent with the intuition of a single degree of freedom. For the more general case where the unconstrained solution $\mathbf{w} = \frac{\mathbf{R}^+\mathbf{1}}{\mathbf{1}'\mathbf{R}^+\mathbf{1}}$ happens to be nonnegative,

$$df\left(\frac{\mathbf{R}^+\mathbf{1}}{\mathbf{1}'\mathbf{R}^+\mathbf{1}}, \mathbf{R}\right) = \frac{\mathbf{1}}{\mathbf{w}'\mathbf{R}\mathbf{w}} = \frac{\left(\mathbf{1}'\mathbf{R}^+\mathbf{1}\right)^2}{\mathbf{1}'\mathbf{R}^+\mathbf{R}\mathbf{R}^+\mathbf{1}} = \frac{\left(\mathbf{1}'\mathbf{R}^+\mathbf{1}\right)^2}{\mathbf{1}'\mathbf{R}^+\mathbf{1}} = \mathbf{1}'\mathbf{R}^+\mathbf{1} \tag{7.28}$$

Consider the singular value decomposition of $\mathbf{R}$,

$$\mathbf{R} = \lambda_1\mathbf{u}_1\mathbf{u}_1' + \lambda_2\mathbf{u}_2\mathbf{u}_2' \ldots \tag{7.29}$$

$$\mathbf{1}'\mathbf{R}^+\mathbf{1} = \mathbf{1}'\left(\lambda_1^{-1}\mathbf{u}_1\mathbf{u}_1' + \lambda_2^{-1}\mathbf{u}_2\mathbf{u}_2' \ldots\right)\mathbf{1} = \lambda_1^{-1}\left(\mathbf{u}_1'\mathbf{1}\right)^2 + \lambda_2^{-1}\left(\mathbf{u}_2'\mathbf{1}\right)^2 \ldots \tag{7.30}$$

Perfect LD blocks of size n loci out of L correspond to eigenvalues of n and eigenvectors with $\frac{1}{\sqrt{n}}$ for each included locus, and 0 otherwise. Each block contributes one degree of freedom:

$$\lambda_1^{-1}\left(\mathbf{u}_1'\mathbf{1}\right)^2 = n^{-1}\left(\frac{1}{\sqrt{n}}n\right)^2 = 1 \tag{7.31}$$

Similarly, eigenvalues orthogonal to the vector **1** do not contribute to the degrees of freedom. All terms are positive, and to be large and positive, a term would need both a small eigenvalue, and a eigenvector with components that do not sum to near zero.

7.6 Skewness, Kurtosis, and Influential Outliers

By the method we used to derive the variance of X_l, we can find arbitrary moments:

$$\mathrm{Var}\left(X_l\right) = \mathrm{E}\left[z_{il}z_{jl}z_{il}z_{jl}\right] = \mathrm{E}\left[z_{il}z_{il}z_{jl}z_{jl}\right] = \mathrm{E}\left[z_{il}z_{il}\right]\mathrm{E}\left[z_{jl}z_{jl}\right] = \left(1+F\right)^2 \tag{7.32}$$

$$\mathrm{E}\left[X_l^k\right] = \mathrm{E}\left[\left(z_{il}z_{jl}\right)^k\right] = \mathrm{E}\left[z_{il}^k z_{jl}^k\right] = \left(\mathrm{E}\left[z_{il}^k\right]\right)^2 \tag{7.33}$$

The skewness and kurtosis are:

$$\mathrm{E}\left[X_l^2\right] = \left(1+F\right)^2 \tag{7.34}$$

$$\mathrm{E}\left[X_l^3\right] = \frac{\left(2p+6Fp-1-3F\right)^2}{2p\left(1-p\right)} = \frac{\left(1-2p\right)^2\left(1+3F\right)^2}{2p\left(1-p\right)} \tag{7.35}$$

$$\mathrm{E}\left[X_l^4\right] = \left[\frac{1+7F-24Fp\left(1-p\right)}{2p\left(1-p\right)}\right]^2 = \left[\frac{1+7F}{2\left(1-p\right)} + \frac{1+7F}{2p} - 12F\right]^2 \tag{7.36}$$

For $p \approx 0$,

$$\mathrm{E}\left[X_l^3\right] \approx \frac{\left(1+3F\right)^2}{2}\frac{1}{p} \tag{7.37}$$

$$\mathrm{E}\left[X_l^4\right] \approx \left[\frac{1+7F}{2}\right]^2\frac{1}{p^2} \tag{7.38}$$

For $F = 0$,

$$\mathrm{E}\left[X_l^3\right] = \frac{\left(1-2p\right)^2}{2p\left(1-p\right)} \tag{7.39}$$

$$\mathrm{E}\left[X_l^4\right] = \left[\frac{1}{2p\left(1-p\right)}\right]^2 \tag{7.40}$$

The contributions of individual terms to the global kurtosis can be high enough to affect genome-wide results; at moderate values of p and F, a kurtosis of 10^4 to 10^6 has the same order of magnitude as the L in a plausible study, as illustrated in Figure 7.1. Since a single SNP is weighted approximately $1/L$ in the genome estimator, and the kinship being estimated has scale (e.g. range or standard deviation) of approximately 1, a kurtosis of order of magnitude L is the scale at which a single SNP is influential for global estimate.

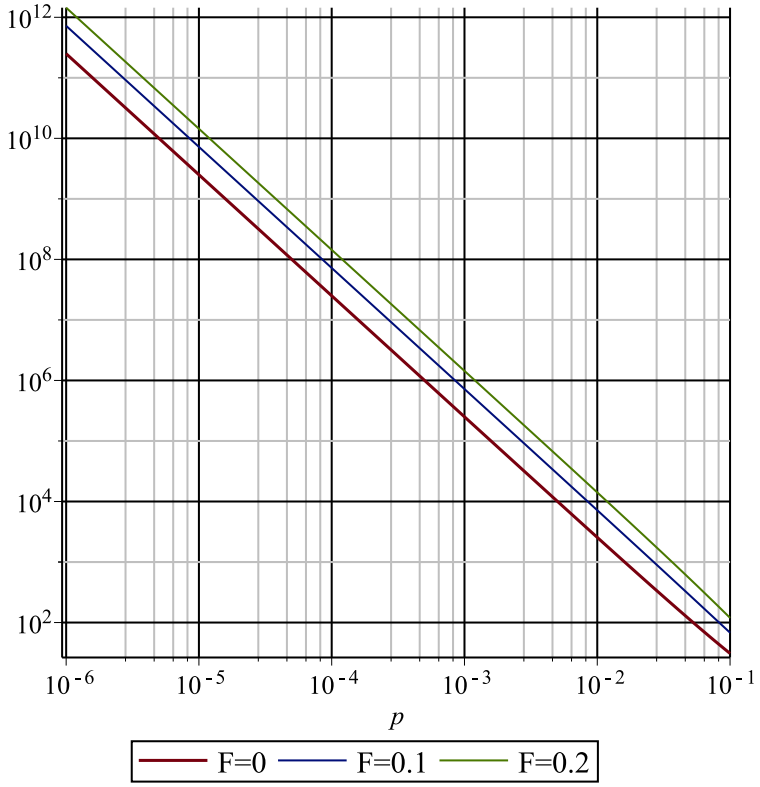


Figure 7.1: Kurtosis of the (single locus) GRM estimator as a function of minor allele frequency p for several values of F.

If we use $w_l \propto p_l\,(1-p_l)$, the weighting in equation 7.16, the contribution of individual loci to both skewness and kurtosis grows in $p\,(1-p)$:

$$w_l^3 \mathrm{E}\left[X_l^3\right] \propto p^2(1-p)^2(1-2p)^2 \tag{7.41}$$

$$w_l^4 \mathrm{E}\left[X_l^4\right] \propto p^2(1-p)^2 \tag{7.42}$$

As a compromise between equations 7.15 and 7.16, we can choose the milder penalty, $w_l \sim \sqrt{p_l\,(1-p_l)}$, equalizing the contributions of all loci to kurtosis:

$$w_l^3 \mathrm{E}\left[X_l^3\right] \propto \sqrt{p_l\,(1-p_l)}(1-2p)^2 \tag{7.43}$$

$$w_l^4 \mathrm{E}\left[X_l^4\right] \propto 1 \tag{7.44}$$

With a known F, the precise adjustment would be of the form

$$w_l \propto \sqrt[4]{\mathrm{E}\left[X_l^4\right]} = \sqrt{\frac{2p\,(1-p)}{1+7F-24Fp\,(1-p)}} \tag{7.45}$$

Chapter 8
CONCLUSION AND DISCUSSION

8.1 Summary of Contributions

The following are summary descriptions of the novel results and contributions of this thesis, arranged by chapter.

Chapter 2: Theory: Identity by Function This chapter introduces a number of concepts published as Sverdlov and Thompson (2013). We critique the probabilistic interpretations of existing (type 2 and type 3) models, and argue for a hybrid model of a population based effect generating process. We introduce the identity by function model, the concept of population neutrality, and the functional identity states. We develop the multivariate normal and affinity-effect generative models, results on positive definite constraints, and demonstrate the connections between IBF and the GRM as well as IBS methods. We also introduce the bottom-up heritability reconstruction method, and the Explained/Unexplained heritability decomposition.

Chapter 3: Methodology: Identity by Function We introduce the state identity array, a method for representing genetic similarity regardless of the specific genotype, as well as several types of genotype ambiguity. We apply the method to several missing or ambiguous genotype situations. The Dictionary method is an approach to imputation and phasing for inferring IBF genotypes. For this method we derive a result on asymptotic approximate estimation of the Ewens θ parameter. We also describe our approach to classifying functional genome regions and alleles. The tools we develop for phenotype prediction and effect inference are analogous to their classical counterparts. The method for bottom-up heritability construction is an implementation of the classical definition of heritability; our contribution is in its extension to random effect models, and in particular to the Explained/Unexplained decomposition.

Chapter 4: Variance Component Estimation We extend the Tipping (2001) mass variance component framework with a closed form solution to the single random effect regression, which we then apply to pig traits in the empirical results chapter. We extend the eigenvalue method for optimizing a variance component likelihoods originating with Thompson and Shaw (1992) to a hybrid SVD/Cholesky method which, by simultaneously diagonalizing two matrices, can be applied to general mass variance component problems, generalizing the Tipping/Faul method.

Chapter 5: Empirical Results In this chapter we apply the results in other chapters to simulated trait data, pig data, and human height data. Of particular interest is our emphasis on measuring the uncertainty of estimates, including the application of the Explained/Unexplained decomposition of heritability.

Chapter 6: Theory Extensions: Combinatorial Approaches to Dominance and Epistasis We construct the framework of ordinal quantitative traits, natural scale, and directional consistency of substitution, and apply it to dominance and epistasis. We also introduce the distinction between removable and essential epistasis, describe tools for describing deviations from directional consistency, and formulate the robustness gene hypothesis in terms of the ordinal framework. The mathematical connection between partial ordering relations and additively separable multivariate functions has previously been used in multiple fields, in particular decision theory and utility theory within economics, mathematical psychology, and measurement theory, as well as non-metric multi-dimensional scaling. Our methodology differs mathematically in the emphasis on discrete combinatorial structures, but the primary novelty is in the application of these mathematical tools to the genetic context.

Chapter 7: Methodology Extensions: Linkage Disequilibrium and the Genomic Relationship Matrix We modify the GRM to efficiently estimate kinship in the presence of linkage disequilibrium, and results on optimization for computing this estimate. By the analysis of the kurtosis of the GRM estimator, we provide an explanation for known phe-

nomena about estimator stability under different weightings and sensitivity to the pruning of rare alleles, with closed form results.

8.2 Discussion: Identity By Function Model

8.2.1 Model Capabilities

Our IBF method defines a novel probability interpretation for the phenotypic covariance between genotyped individuals on the basis of a population process. It is an adaptation of the Fisher (1918) linear model, additive across loci, but not within a locus, with many alleles at each locus. It allows, in principle, the direct estimation of all individual α_i and δ_{ij} terms of the Fisher model.

The IBF parameters define a genetic architecture, giving the trait-specific weights of the different loci, and the pattern of dominance at each locus. This architecture is separate from the effects of individual alleles at each locus, and of the population frequencies of those alleles. This architecture allows us to express covariance between the traits of individuals in terms of alleles that need not have previously been observed. Inferences may be made about individuals across populations with different allele frequencies, including alleles which occur in one population but not the other. This is crucial for full sequence genomic analysis, where the number of individual alleles at a locus may be large, and the genotype of any given individual may include alleles at multiple loci not previously encountered in the population.

8.2.2 Selection and Neutrality

The model as presented is strictly neutralist, and as such invites selectionist extensions. In particular, the model recognizes only a single class of alleles, and treats mutations between such alleles symmetrically. If we partition alleles into multiple functional classes (e.g. *normal* and *detrimental*) based on bioinformatics and population genetics, we can construct an extended set of IBF states and a set of covariances between such states, at the cost of a larger number of parameters.

The set of homozygotes and heterozygotes within each class requires one set of IBF parameters $\mu_{AA}, \mu_{AB}, \sigma^2_{AA}, \sigma^2_{AB}, \rho_2, \rho_3$ for each class. To these would be added cross-class

heterozygotes (normal-detrimental), requiring a mean and variance parameter for each class pairing, as well as a set of correlations (among cross-class hetorozygotes sharing alleles, and between cross-class heterozygotes and within-class homozygotes and heterozygotes). To use this kind of model for inference given genotype, we need not explicitly model the population process giving rise to the different mutations. Breaking the symmetry between designated allele classes would be sufficient. Analogous extensions of the state space can integrate sex chromosomes and copy number variants into the polygenic framework.

8.2.3 Diploidy and Dominance

The inherently diploid nature of the model, and the ability to treat the phenomenon of dominance coherently for a large number of alleles, allows for a reexamination of the relative importance of dominance effects. Existing models can be parametrized to incorporate dominance. Type 2 models can incorporate dominance through kernels designed to include allele interaction terms. In type 1 or type 3 models, the dominance genetic relationship matrix, or the matrix of coefficients of fraternity, coupled with the dominance variance component, can be incorporated into the covariance model together with the kinship matrix and the additive variance component. In our model, individual dominance effects have a covariance structure implied by the IBF parameters. This allows for an arbitrary allele count and estimation of both the individual dominance deviations and variance components.

Dominance has not received much attention in the animal breeding literature, which is focused on breeding value estimation, or in the GWAS literature, where the linear SNP association models predominate. Dominance at the functional locus or gene level, as a distinct biological phenomenon, may be more detectable than nonlinearity in SNP dosage. Furthermore, even if additive, rather than dominance, variance predominates at the population level, (e.g. Hill et al., 2008), dominance could be important for individual phenotype prediction, especially where rare alleles are collectively common.

In the discussion of combinatorial methods, we argue against the importance of dominance, even in the multi-allelic context, under the assumption of directional consistency of substitution, and biologically incomplete dominance. The appropriate question is not

whether such assumptions are universally true, and therefore whether dominance is universally irrelevant, but rather to what biological circumstances the two extremes described by the pro- and anti-dominance frameworks apply.

8.3 Conclusions and Further Research

8.3.1 Functional Quantitative Genetics

The integration of bioinformatic allele classification with a quantitative genetic model is a step toward *Functional Quantitative Genetics*. Plant and animal breeding applications have repeatedly demonstrated the predictive power of quantitative genetics over a variety of complex traits, but a quantitative genetic model does not elucidate the causal mechanisms of gene action. Sequence bioinformatics aids the understanding of individual gene function at the molecular level, but not quantitative effects at organism scale. The relationship between genotype and phenotype for complex polygenic traits must be sought by the reconciliation of the two approaches.

8.3.2 Missing Heritability

A complete accounting of the sources of uncertainty in the measurement of heritability, starting with the differences in its definitions enumerated in Section 1.2.2, allows empirical measurements of heritability to vary greatly across experiment designs. Pedigree-based *top-down* heritability measurements are a well-established methodology, though, as we emphasized, the consistency of such measurements across the choice of pedigrees is model-dependent. We developed at least three sources of uncertainty in *bottom-up* reconstruction:

1. At the genetic sequence level, we illustrated the possibility of ambiguity in the definition of the gene scale quantitative trait locus and partition of alleles into classes, and the relevance of this question to the reconstruction of heritability.

2. At the single gene level, we developed the decomposition of additive and total genetic variance into Explained and Unexplained components.

3. At the genome level, we found large upper bounds on our error estimates of heritability due to the possibility that single gene contributions to genetic variance need not aggregate as independent error terms.

Widely differing estimates of heritability can legitimately coexist, pointing to complexity of genetic phenomena at both the individual and the population scale. Heritability has historically proved its usefulness in the context of pedigree relatives, with unknown genotypes. Today, we seek to extend the concept either to distant, nominally unrelated relatives, or to complete genome sequences. In physics, color and temperature lose any traditional meaning at the subatomic scale. Pushed far outside its classical context, heritability too may experience a breakdown of conceptual validity, rather than the problems associated with statistical parameter estimation.

BIBLIOGRAPHY

1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nature Methods 7, 248–249.

Bishop, C.M., 2003. Bayesian Regression and Classification, in: Advances in Learning Theory: Methods, Models and Applications, IOS Press. pp. 267–285.

Browning, B.L., Browning, S.R., 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics 84, 210–223.

Browning, S.R., Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics 81, 1084–1097.

Browning, S.R., Browning, B.L., 2012. Identity by Descent Between Distant Relatives: Detection and Applications. Annual Review of Genetics 46, 617–633.

de los Campos, G., Gianola, D., Rosa, G.J.M., 2009. Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. Journal of Animal Science 87, 1883–1887.

Cleveland, M.A., Hickey, J.M., Forni, S., 2012. A common dataset for genomic analysis of livestock populations. G3: Genes-Genomes-Genetics 2, 429–435.

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

Cockerham, C.C., Weir, B.S., 1983. Variance of actual inbreeding. Theoretical Population Biology 23, 85–109.

Colditz, G.A., Hankinson, S.E., 2005. The Nurses' Health Study: lifestyle and health among women. Nature Reviews Cancer 5, 388–396.

Debreu, G., 1960. Topological methods in cardinal utility theory. Mathematical Methods in the Social Sciences 1959, 16–26.

Eding, J.H., Meuwissen, T.H., 2001. Marker based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding Genetics 118, 141–159.

Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., Stein, L., Durbin, R., Ashburner, M., 2005. The sequence ontology: a tool for the unification of genome annotations. Genome Biology 6, R44.

Ewens, W.J., 2004. Mathematical Population Genetics 1: I. Theoretical Introduction. Springer.

Falconer, D.S., MacKay, T.F.C., 1996. Introduction to Quantitative Genetics. Longmans Green, Harlow, Essex, UK.. 4 edition.

Faul, A.C., Tipping, M.E., 2001. Analysis of Sparse Bayesian Learning, in: Advances in Neural Information Processing Systems 14, MIT Press. pp. 383–389.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates.

Fishburn, P.C., 1970. Utility theory for decision making. Technical Report. DTIC Document.

Fisher, R.A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52, 399–433.

Fisher, R.A., 1941. Average excess and average effect of a gene substitution. Annals of Eugenics 11, 53–63.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girn, C.G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Khri, A.K., Keenan, S., Kulesha, E., Martin, F.J., Maurel, T., McLaren, W.M., Murphy, D.N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S.J., Vullo, A., Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T.J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D.R., Searle, S.M., 2014. Ensembl 2014. Nucleic Acids Research 42, D749–D755.

Galton, F., 1886. Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute of Great Britain and Ireland , 246–263.

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., Fernando, R., 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183, 347–363.

Gianola, D., van Kaam, J.B., 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178, 2289–2303.

Gonzales, C., 1996. Additive Utilities When Some Components Are Solvable And Others Are Not. Journal of Mathematical Psychology 40, 141–151.

Guan, Y., Stephens, M., 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics 5, 1780–1815.

Hayes, B.J., Goddard, M.E., 2008. Technical note: Prediction of breeding values using marker-derived relationship matrices. Journal of Animal Science 86, 2089–2092.

Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetical Research 91, 47–60.

Heath, A., Martin, N., Eaves, L., Loesch, D., 1984. Evidence for polygenic epistatic interactions in man? Genetics 106, 719–727.

Hill, W.G., Goddard, M.E., Visscher, P.M., 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genetics 4, e1000008.

Hill, W.G., Weir, B.S., et al., 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics Research 93, 47–64.

Jacquard, A., 1972. Genetic Information Given by a Relative. Biometrics 28, 1101–1114.

Kärkkäinen, H.P., Sillanpää, M.J., 2012. Back to Basics for Bayesian Model Building in Genomic Selection. Genetics 191, 969–987.

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

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., Hultman, C.M., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y.Y., Price, A.L., de Bakker, P.I.W., Purcell, S.M., Sunyaev, S.R., 2012. Exome sequencing and the genetic basis of complex traits. Nature Genetics 44, 623–630.

Kimura, M., Crow, J.F., 1964. The Number of Alleles That Can Be Maintained in a Finite Population. Genetics 49, 725–738.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 1–27.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al., 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832–838.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D., 2011. Fast linear mixed models for genome-wide association studies. Nature Methods 8, 833–835.

Luce, R.D., Tukey, J.W., 1964. Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of Mathematical Psychology 1, 1–27.

Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates.

MacKay, D.J.C., 1994. Bayesian Nonlinear Modeling for the Prediction Competition. ASHRAE Transactions 100, 1053–1062.

MacKay, D.J.C., 1996. Bayesian methods for backpropagation networks, in: Models of neural networks III. Springer, pp. 211–254.

Maher, B., 2008. The case of the missing heritability. Nature 456, 18–21.

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., de los Campos, G., 2011. Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genetics 7, e1002051.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al., 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753.

Mathai, A., Provost, S.B., 1992. Quadratic Forms in Random Variables. Statistics Series, Taylor & Francis.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Neal, R.M., 1995. Bayesian learning for neural networks. Ph.D. thesis. University of Toronto.

Nejati-Javaremi, A., Smith, C., Gibson, J.P., 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75, 1738–1745.

Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., Chatterjee, N., 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics 42, 570–575.

Powell, J.E., Visscher, P.M., Goddard, M.E., 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11, 800–805.

Rimm, E., Giovannucci, E., Willett, W., Colditz, G., Ascherio, A., Rosner, B., Stampfer, M., 1991. Prospective study of alcohol consumption and risk of coronary disease in men. The Lancet 338, 464 – 468.

Robinson, G.K., 1991. That BLUP is a good thing: The estimation of random effects. Statistical Science 6, 15–32.

Schur, J., 1911. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. Journal für die reine und Angewandte Mathematik 140, 1–28.

Sillanpää, M.J., 2011. On statistical methods for estimating heritability in wild populations. Molecular Ecology 20, 1324–1332.

Slatkin, M., 2009. Epigenetic inheritance and the missing heritability problem. Genetics 182, 845–850.

Sverdlov, S., Thompson, E.A., 2013. Correlation between relatives given complete genotypes: From identity by descent to identity by function. Theoretical Population Biology 88, 57–67.

Tal, O., Kisdi, E., Jablonka, E., 2010. Epigenetic contribution to covariance between relatives. Genetics 184, 1037–1050.

Thompson, E., Shaw, R., 1990. Pedigree analysis for quantitative traits: variance components without matrix inversion. Biometrics 46, 399–413.

Thompson, E., Shaw, R., 1992. Estimating polygenic models for multivariate data on large pedigrees. Genetics 131, 971–978.

Thompson, E.A., 1974. Gene identities and multiple relationships. Biometrics 30, pp. 667–680.

Thompson, E.A., 1975. The estimation of pairwise relationships. Annals of Human Genetics 39, 173–188.

Thompson, E.A., 2013. Identity by descent: Variability in meiosis, across genomes and in populations. Genetics: accepted.

Thornton, T., Tang, H., Hoffmann, T.J., Ochs-Balcom, H.M., Caan, B.J., Risch, N., 2012. Estimating kinship in admixed populations. The American Journal of Human Genetics 91, 122–138.

Tipping, M., 2001. Sparse Bayesian learning and the relevance vector machine. The Journal of Machine Learning Research 1, 211–244.

Tipping, M.E., Faul, A., 2003. Fast Marginal Likelihood Maximisation for Sparse Bayesian Models, in: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, pp. 1–8.

VanRaden, P., 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.

Visscher, P.M., 2009. Whole genome approaches to quantitative genetics. Genetica 136, 351–358.

Visscher, P.M., 2010. A Commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al.(2010). Twin Research and Human Genetics 13, 517–524.

Visscher, P.M., McEvoy, B., Yang, J., 2010. From Galton to GWAS: quantitative genetics of human height. Genetics Research 92, 371–379.

Vitezica, Z.G., Varona, L., Legarra, A., 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195, 1223–1230.

Wakker, P.P., 1989. Additive representations of preferences: A new foundation of decision analysis. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Weir, B.S., Anderson, A.D., Hepler, A.B., 2006. Genetic relatedness analysis: modern data and new challenges. Nature Reviews Genetics 7, 771–780.

Wipf, D.P., Nagarajan, S.S., 2007. A new view of automatic relevance determination, in: Advances in Neural Information Processing Systems, pp. 1625–1632.

Wipf, D.P., Rao, B.D., 2004. Sparse bayesian learning for basis selection. IEEE Transactions on Signal Processing 52, 2153–2164.

Wright, S., 1921. Systems of mating. II. The effects of inbreeding on the genetic composition of a population. Genetics 6, 124.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics 89, 82–93.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., Visscher, P.M., 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42, 565–569.

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al., 2011. Genome partitioning of genetic variation for complex traits using common snps. Nature Genetics 43, 519–525.

Zuk, O., Hechter, E., Sunyaev, S.R., Lander, E.S., 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences 109, 1193–1198.

Appendix A

CONTRIBUTIONS OF INDIVIDUAL LOCI TO HERITABILITY

A.1 General Case

In the most general case, following Lynch and Walsh (1998), the (bottom-up) additive variance of the trait value P explained by allele dosages $D_{i(l)}$ at locus l is

$$\sigma_A^2 = \sum_{l=1}^{L} \sum_{i} \alpha_{i(l)} \mathrm{Cov}\left(P, D_{i(l)}\right) = \sum_{l=1}^{L} \sum_{i} \alpha_{i(l)} \left(2 p_{i(l)} \alpha^*_{i(l)}\right) \tag{A.1}$$

The α_i and α_i^* are defined by Fisher (1918, 1941):

- Additive effect α_i is the regression coefficient of trait on allele dosage.
- Average excess α_i^* is the mean trait value for the subset of the population having at least one allele, minus the overall population mean trait value.

The definitions do not require random mating and linkage equilibrium; rather $\alpha_i = \alpha_i^*$ under those conditions. The definitions can be stated in terms of genotypic values, or in terms of population mean phenotypes for given genotype. The latter form is meaningful in the multi-locus case, as we can define a population mean at one locus, averaging over genotypes at other loci.

Despite the apparent generality of these definitions, this only works if we have defined what the loci and alleles are. The definition of additive effect and average excess is entirely statistical and with respect to the population; therefore, it is internally consistent regardless of its biological meaning. It is in the pattern resemblance between relatives (on the assumption of Mendelian inheritance, per Fisher (1918)) that the model becomes falsifiable.

A.2 Simplified Cases

Under random mating, $\alpha_{i(l)}{=}\alpha^*_{i(l)}$,

$$\sigma^2_A{=}\,2\sum_{l=1}^{L}\sum_{i} p_{i(l)}\alpha^2_{i(l)} \tag{A.2}$$

For the biallelic, additive case,

$$p_1\alpha_1{+}\,(1{-}p_1)\,\alpha_2{=}\,0 \tag{A.3}$$

$$\frac{-p_1}{1{-}p_1}\alpha_1{=}\alpha_2 \tag{A.4}$$

Using the substitution effect size $\beta = \alpha_1 - \alpha_2$

$$\alpha_2{=}\,-p_1\beta \tag{A.5}$$

$$\alpha_1{=}\,(1{-}p_1)\,\beta \tag{A.6}$$

$$\sigma^2_A{=}\,2\left[p_1\alpha_1^2{+}p_2\alpha_2^2\right] \tag{A.7}$$

$$=2\left[p_1(1{-}p_1)^2\beta^2{+}\,(1{-}p_1)\,p_1^2\beta^2\right] \tag{A.8}$$

$$=2p_1\,(1{-}p_1)\,[p_1{+}\,(1{-}p_1)]\,\beta^2 \tag{A.9}$$

$$=2p\,(1{-}p)\,\beta^2 \tag{A.10}$$

This is the popular form of variance due to a single locus, in the biallelic SNP/GWAS setting. The effect size, β, is biological, and does not depend on the population, whereas the allele frequency p does (note that we write p as the expression is symmetric with respect to p_1 and $p_2 = 1 - p_1$.

Appendix B

POSITIVE DEFINITE CONDITIONS

In order for the joint distribution to exist, the two correlations ρ_3 and ρ_2 must be such that the covariance matrix of all homozygote and heterozygote values $[G_{ii}; G_{ij}]$ is positive semidefinite. The joint covariance matrix has the form

$$
\begin{aligned}
&\mathrm{Cov}\left(\left(\left[G_{11}\ G_{22}\ G_{33}\ G_{44}\right];\left[G_{12}\ G_{13}\ G_{14}\ G_{23}\ G_{24}\ G_{34}\right]\right)^T\right) \\
&= \begin{bmatrix}
\begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1 \end{bmatrix}\sigma^2_{AA} & \begin{bmatrix} 1&1&1&0&0&0\\ 1&0&0&1&1&0\\ 0&1&0&1&0&1\\ 0&0&1&0&1&1 \end{bmatrix}\rho_3\sigma_{AB}\sigma_{AA} \\
\begin{bmatrix} 1&1&0&0\\ 1&0&1&0\\ 1&0&0&1\\ 0&1&1&0\\ 0&1&0&1\\ 0&0&1&1 \end{bmatrix}\rho_3\sigma_{AB}\sigma_{AA} & \begin{bmatrix} 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1 \end{bmatrix}\sigma^2_{AB} + \begin{bmatrix} 0&1&1&1&1&0\\ 1&0&1&1&0&1\\ 1&1&0&0&1&1\\ 1&1&0&0&1&1\\ 1&0&1&1&0&1\\ 0&1&1&1&1&0 \end{bmatrix}\rho_2\sigma^2_{AB}
\end{bmatrix}
\end{aligned}
\tag{B.1}
$$

Or, generally, using analogous (homozygotes, then row major) ordering of the G_{ij},

$$
\mathrm{Cov}\left(\begin{bmatrix} G_{ii} \\ G_{jk} \end{bmatrix}\right) = \mathrm{A} = \begin{pmatrix} \mathrm{I}\sigma^2_{AA} & \mathrm{M}^{(3)}_{i;jk}\rho_3\sigma_{AB}\sigma_{AA} \\ \mathrm{M}^{(3)T}_{i;jk}\rho_3\sigma_{AB}\sigma_{AA} & \mathrm{I}\sigma^2_{AB}+\mathrm{M}^{(2)}_{jk;lm}\rho_2\sigma^2_{AB} \end{pmatrix} \tag{B.2}
$$

Here $\mathrm{M}^{(3)}_{i;jk}$ and $\mathrm{M}^{(2)}_{jk;lm}$ are binary matrices, containing 1 when exactly one index matches between the column and row identifiers, and 0 otherwise.

To find the values of ρ_3 and ρ_2 for which the matrix is positive semidefinite, we observe that $\mathrm{I}\sigma^2_{AA}$ is positive definite. Then A is positive semidefinite if and only if the Schur complement of $\mathrm{I}\sigma^2_{AB}+\mathrm{M}^{(2)}_{jk;lm}\rho_2\sigma^2_{AB}$ in A is positive semidefinite. The Schur complement is

$$\begin{aligned}\mathrm{S} &= \mathrm{I}\sigma^2_{AB} + \mathrm{M}^{(2)}_{jk;lm}\rho_2\sigma^2_{AB} \\ &\quad - \left[\mathrm{M}^{(3)}_{i;jk}\rho_3\sigma_{AB}\sigma_{AA}\right]\left[\mathrm{I}\sigma^2_{AA}\right]^{-1}\left[\mathrm{M}^{(3)T}_{i;jk}\rho_3\sigma_{AB}\sigma_{AA}\right] \\ &= \sigma^2_{AB}\left[\mathrm{I} + \mathrm{M}^{(2)}_{jk;lm}\rho_2 - \mathrm{M}^{(3)}_{i;jk}\mathrm{M}^{(3)T}_{i;jk}\rho_3^2\right] \end{aligned} \tag{B.3}$$

We observe that

$$\mathrm{M}^{(3)}_{i;jk}\mathrm{M}^{(3)T}_{i;jk} = \mathrm{M}^{(2)}_{jk;lm} + 2\mathrm{I} \tag{B.4}$$

Then

$$\begin{aligned}\mathrm{S} &= \sigma^2_{AB}\left[\mathrm{I} + \left(\mathrm{M}^{(3)}_{i;jk}\mathrm{M}^{(3)T}_{i;jk} - 2\mathrm{I}\right)\rho_2 - \mathrm{M}^{(3)}_{i;jk}\mathrm{M}^{(3)T}_{i;jk}\rho_3^2\right] \\ &= \sigma^2_{AB}\left[(1-2\rho_2)\,\mathrm{I} + \mathrm{M}^{(3)}_{i;jk}\mathrm{M}^{(3)T}_{i;jk}\left(\rho_2 - \rho_3^2\right)\right] \end{aligned} \tag{B.5}$$

$\mathrm{M}^{(3)}_{i;jk}\mathrm{M}^{(3)T}_{i;jk}$ is positive semidefinite; it is not of full rank for nontrivial numbers of alleles; thus it has an eigenvalue of 0 and no negative eigenvalues. The eigenvalues of $\mathrm{M}^{(3)}_{i;jk}\mathrm{M}^{(3)T}_{i;jk}$ grow without bound with the size of the matrix since the vector 1 is always an eigenvector with an eigenvalue that grows with the count of heterozygotes. The eigenvalues of S are simply $\lambda' = (1-2\rho_2) + \left(\rho_2 - \rho_3^2\right)\lambda$ for all eigenvalues λ of $\mathrm{M}^{(3)}_{i;jk}\mathrm{M}^{(3)T}_{i;jk}$. We want to constrain the values of ρ_2 and ρ_3^2 so that any number of heterozygotes, S is positive semidefinite. We must choose $\rho_2 - \rho_3^2 \geq 0$ because otherwise we can find a large enough λ for a large number of heterozygotes to make $\lambda' < 0$. Likewise, we must choose $1 - 2\rho_2 \geq 0$, since otherwise $\lambda' < 0$ for $\lambda = 0$. Together these create the constraint:

$$\rho_3^2 \leq \rho_2 \leq \frac{1}{2} \tag{B.6}$$

Appendix C

DIPLOID ADDITIVE MODEL

For the diploid additive model we associate an additive effect with each allele, $C(\cdot) : \mathcal{A} \to \Re$, and define the genotypic value function $G(A_i, A_j) = C(A_i) + C(A_j)$; then, for distinct i, j, k, l,

$$\begin{aligned} \mathrm{E}[G_{ii}] &= \mathrm{E}[C(A_i)] + \mathrm{E}[C(A_i)] = 2\ \mu_C = \\ \mathrm{E}[G_{ij}] &= \mathrm{E}[C(A_i)] + \mathrm{E}[C(A_j)] = 2\ \mu_C \end{aligned} \tag{C.1}$$

This implies $\mu_{AA} = \mu_{AB} = 2\mu_C$.

$$\mathrm{Var}[G_{ii}] = \mathrm{Var}[2\ C(A_i)] = 4\ \mathrm{Var}[C(A_i)] = 4\ \sigma_C^2 \tag{C.2}$$

$$\mathrm{Var}[G_{ij}] = \mathrm{Var}[C(A_i)] + \mathrm{Var}[C(A_j)] = 2\ \sigma_C^2 \tag{C.3}$$

Substituting into the general model,

$$\mathrm{Var}[G_{ii}] = \sigma_{AA}^2 = 4\ \sigma_C^2 \tag{C.4}$$

$$\mathrm{Var}[G_{ij}] = \sigma_{AB}^2 = 2\ \sigma_C^2 \tag{C.5}$$

This implies $\sigma_{AA} = 2\ \sigma_C$ and $\sigma_{AB} = \sqrt{2}\ \sigma_C$, or $\sigma_{AB}/\sigma_{AA} = 1/\sqrt{2}$.

$$\mathrm{Cov}[G_{ii}, G_{ij}] = \mathrm{Cov}[2C(A_i), C(A_i)] + \mathrm{Cov}[2\ C(A_i), C(A_j)] = 2\ \sigma_C^2 \tag{C.6}$$

$$\begin{aligned} \mathrm{Cov}[G_{ij}, G_{ik}] &= \mathrm{Var}[C(A_i)] + \mathrm{Cov}[C(A_i), C(A_j)] \\ &\quad + \mathrm{Cov}[C(A_i), C(A_k)] + \mathrm{Cov}[C(A_j), C(A_k)] = \sigma_C^2 \end{aligned} \tag{C.7}$$

Substituting into the general model,

$$\text{Cov}\left[G_{ii}, G_{ij}\right] = \rho_3 \sigma_{AA} \sigma_{AB} = \rho_3 \; 2\sqrt{2} \; \sigma_C^2 = 2 \; \sigma_C^2 \tag{C.8}$$

$$\text{Cov}\left[G_{ij}, G_{ik}\right] = \rho_2 \sigma_{AB}^2 = \rho_2 \; 2 \; \sigma_C^2 = \sigma_C^2 \tag{C.9}$$

This implies $\rho_2 = 1/2$ and $\rho_3 = 1/\sqrt{2}$.

Appendix D

OVERDOMINANCE IN THE MULTIVARIATE NORMAL MODEL

For two distinct alleles F and G, we define the dominance coefficient Y in terms of the genotypic values FF, FG, and GG:

$$Y = \frac{(FG - FF) + (FG - GG)}{FF - GG} \tag{D.1}$$

A value of $|Y| > 1$ indicates overdominance, an ordering other than $GG \leq FG \leq FF$ or $FF \leq FG \leq GG$. We will derive the probability of overdominance, P_{OD}, for the multivariate normal model with $\mu_{AB} = \mu_{AA}$.

Consider a general multivariate normal distribution of the form:

$$\begin{bmatrix} A_0 \\ B \\ A_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_A^2 & \sigma_A\sigma_B\rho_{AB} & \sigma_A^2\rho_A \\ \sigma_A\sigma_B\rho_{AB} & \sigma_B^2 & \sigma_A\sigma_B\rho_{AB} \\ \sigma_A^2\rho_A & \sigma_A\sigma_B\rho_{AB} & \sigma_A^2 \end{bmatrix} \right) \tag{D.2}$$

$$\begin{aligned}
&Var\left(-A_1 + 2B - A_0\right) \\
&\quad = \sigma_A^2 + 4\sigma_B^2 + \sigma_A^2 + 2\left(-2\sigma_A\sigma_B\rho_{AB} + 1\sigma_A^2\rho_A - 2\sigma_A\sigma_B\rho_{AB}\right) \\
&\quad = 2\sigma_A^2 + 4\sigma_B^2 - 8\sigma_A\sigma_B\rho_{AB} + 2\sigma_A^2\rho_A \\
&\quad = 2\left(1 + \rho_A\right)\sigma_A^2 - 8\rho_{AB}\sigma_A\sigma_B + 4\sigma_B^2 \\
&Cov\left(-A_1 + 2B - A_0,\ A_1 - A_0\right) \\
&\quad = -\sigma_A^2 + 2\sigma_A^2\rho_{AB} - \sigma_A^2\rho_A - \left(-\sigma_A^2\rho_A + 2\sigma_A^2\rho_{AB} - \sigma_A^2\right) \\
&\quad = 0 \\
&Var\left(A_1 - A_0\right) \\
&\quad = 2\sigma_A^2 - 2\sigma_A^2\rho_A \\
&\quad = 2\left(1 - \rho_A\right)\sigma_A^2
\end{aligned}$$

The distribution of a ratio of zero mean uncorrelated normals $Y = N/D$ with respective variances σ_N^2 and σ_D^2 and correlation ρ is a Cauchy with pdf $p(y) = \frac{1}{\pi}\beta/\left(y^2+\beta^2\right)$ where $\beta = \sigma_N/\sigma_D$.

$$\sigma_N^2 = 2(1+\rho_A)\sigma_A^2 - 8\rho_{AB}\sigma_A\sigma_B + 4\sigma_B^2 \tag{D.3}$$

$$\sigma_D^2 = 2(1-\rho_A)\sigma_A^2 \tag{D.4}$$

In the overdominance case, $A_1 = FF$; $A_2 = GG$; $B = FG$; $\rho_{AB} = \rho_3$; $\rho_A = 0$; the variances of the heterozygote and homozygote are σ_{AB}^2 and σ_{AA}^2 respectively.

$$\beta = \frac{\sigma_N}{\sigma_D} = \sqrt{\frac{2(1+\rho_A)\sigma_A^2 - 8\rho_{AB}\sigma_A\sigma_B + 4\sigma_B^2}{2(1-\rho_A)\sigma_A^2}} \tag{D.5}$$

$$= \sqrt{\frac{2+2\rho_A - 8\rho_{AB}\frac{\sigma_B}{\sigma_A} + 4\frac{\sigma_B^2}{\sigma_A^2}}{2-2\rho_A}} \tag{D.6}$$

$$= \sqrt{\frac{2-4\rho_{AB}\frac{\sigma_B}{\sigma_A} + 2\frac{\sigma_B^2}{\sigma_A^2}}{1-\rho_A} - 1} \tag{D.7}$$

The Cauchy CDF gives

$$\Pr(|Y|>1) \;= \frac{2}{\pi}\tan^{-1}\beta \tag{D.8}$$

$$\beta = \sqrt{1 - 4\rho_3\frac{\sigma_{AB}}{\sigma_{AA}} + 2\frac{\sigma_{AB}^2}{\sigma_{AA}^2}} \tag{D.9}$$

This gives the probability of overdominance P_{OD}:

$$P_{OD} = \Pr(|Y|>1) \;= \frac{2}{\pi}\tan^{-1}\beta \;= \frac{2}{\pi}\tan^{-1}\sqrt{1 - 4\rho_3\frac{\sigma_{AB}}{\sigma_{AA}} + 2\frac{\sigma_{AB}^2}{\sigma_{AA}^2}} \tag{D.10}$$

In the additive case, $\rho_3 = \sigma_{AB}/\sigma_{AA} = 1/\sqrt{2}$; then $P_{OD} = \frac{2}{\pi}\tan^{-1}\sqrt{1-2+1} \;= 0$ as expected.

Appendix E

STRICT DOMINANCE, RANDOM ORDERING

The marginal distribution of the heterozygotes is the same as that of the homozygotes, since under random ordering a homozygote's distribution is an equiprobable mixture of its two homozygotes' equal distributions. Table E.1 lists the 6 equally probably orderings of the affinity variable S_i. Conditional on each ordering, each heterozygote is equal to one of the three homozygotes, as tabulated. Taking wlog $\sigma_{AB} = \sigma_{AA} = 1$, the ρ_2 and ρ_3 columns give the conditional covariances $\mathrm{Cov}(G_{AB}, G_{AC})$ and $\mathrm{Cov}(G_{AB}, G_{AA})$ respectively, which, when averaged over the 6 random orderings, give the unconditional covariances and therefore correlations ρ_2 and ρ_3.

Ordering:	G_{AB}	G_{AC}	G_{BC}	ρ_3	ρ_2
$S_A > S_B > S_C$	G_{AA}	G_{AA}	G_{BB}	1	1
$S_A > S_C > S_B$	G_{AA}	G_{AA}	G_{CC}	1	1
$S_B > S_A > S_C$	G_{BB}	G_{AA}	G_{BB}	0	0
$S_B > S_C > S_A$	G_{BB}	G_{CC}	G_{BB}	0	0
$S_C > S_A > S_B$	G_{AA}	G_{CC}	G_{CC}	1	0
$S_C > S_B > S_A$	G_{BB}	G_{CC}	G_{CC}	0	0

Table E.1: Affinity orderings, heterozygote-homozygote equivalences, and conditional correlations for the Strict Dominance, Random Ordering model.

VITA

Serge Sverdlov received a B.S. with Honors in Engineering & Applied Science and Economics from the California Institute of Technology in 2001. He was a Milken Scholar, Caltech Merit Scholar, and a Caltech/NSF Entrepreneurial Fellow, and received an Engineering Management Certificate from the Caltech Industrial Relations Center. He received an M.S. in Statistics from the University of Washington in 2011 and a Ph.D. in Statistics (Statistical Genetics track) in 2014.