From Estimation to Prediction of Genomic Variances: Allowing for Linkage Disequilibrium and Unbiasedness

Inauguraldissertation

zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften der Universität Mannheim

vorgelegt von

Nicholas Michael Schreck
aus Aschaffenburg

Mannheim 2018

Dekan: Dr. Bernd Lübcke, Universität Mannheim
Referent: Professor Dr. Martin Schlather, Universität Mannheim
Korreferent: Professor Dr. Hans-Peter Piepho, Universität Hohenheim

Tag der mündlichen Prüfung: 19. Juni 2018

Abstract

In Chapter 1 of this thesis, we briefly summarize the theory about the genetic variance from quantitative genetics, the genomic variance in linear regression models, and relate these quantities to the "Missing Heritability".
In Chapter 2, we introduce novel concepts of estimating the genomic variance in accordance to quantitative genetics theory, i.e. the source of the genetic variability stems from the variability in marker-genotypes and not from the randomness of the marker effects. We distinguish the analysis between Fixed Effect Models, Bayesian Regression Models and Random Regression Models. We adapt the estimators for the genomic variance to the specific model set-ups and show that the resulting quantities explicitly include the contribution of linkage disequilibrium.
We substantiate our theoretical findings in simulations studies in Chapter 3 by showing that our approach enables a reduction of the "Missing Heritability".

Zusammenfassung

Im ersten Kapitel dieser Dissertation fassen wir theoretische Resultate zu der genetischen Varianz aus der quantitativen Genetik und Ergebnisse über die genomische Varianz in linearen Regressionsmodellen zusammen. Außerdem stellen wir den Zusammenhang zu der sogenannten „Missing Heritability" her.
Im zweiten Kapitel führen wir neuartige Konzepte zur Schätzung der genomischen Varianz ein, die im Einklang mit der Theorie aus der quantitativen Genetik stehen. Das bedeutet, dass als Ursache der genetischen Variabilität die Variabilität in den Genotypen der Marker benutzt wird und nicht die Zufälligkeit der Effekte dieser Marker. Wir unterscheiden Regressionsmodelle mit fixen Effekten, Bayesianische Regressionsmodelle und Regressionsmodelle mit zufälligen Effekten. Wir passen die Schätzer für die genomische Varianz an die spezifischen Modellvoraussetzungen an und zeigen, dass die so erhaltenen Schätzer explizit den Beitrag des Kopplungsungleichgewichtes enthalten.
Diese theoretischen Resultate werden in Kapitel 3 in Simulationsstudien untermauert. Dabei zeigen wir, dass unser Vorgehen eine Verringerung der „Missing Heritability" ermöglicht.

Acknowledgements

For what is a man, what has he got
If not himself, then he has naught
To say the things he truly feels
And not the words of one who kneels
The record shows I took the blows
And did it my way.

From "My Way" - Frank Sinatra

I wish to express my gratitude to the people who supported me – scientifically and personally – during the challenging time of writing this thesis.

First of all, I would like to express my most sincere gratitude to my supervisor Prof. Dr. Martin Schlather. He enabled me to become a doctoral student and was always available whenever I needed guidance and advice. He got me started on the subject of the estimation of genomic variances which proved to be a stroke of luck for me. Most importantly, he put his faith and trust in me and my work even in times when others clearly did not.
I am particular thankful to Prof. Dr. Hans-Peter Piepho for his great effort, insight and expertise that assisted the research for this thesis.
Special thanks go to Anja Gilliar for taking care of all administrative work and for covering our backs.

I am grateful to Jonas Brehmer, Maximilian Bögner, Stella Dohn, Maximilian Gierlich and Torsten Pook for helpful comments on earlier versions of this thesis, and to them and the rest of the "Gang" for their loyal friendship over the past 6–7 years. Thanks to Eike for challenging me on and beyond the tennis court.

I am deeply indebted to my parents for their love and support throughout my whole life. They have always enabled me to focus on my education and made my academic carrier possible. Finally, I owe my deepest gratitude to Juliane for her unconditional support and love. Words cannot meet her contribution and dedication in every situation of life.

I gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft, DFG, project number SCHL 1865/4, as well as the 'RTG 1953 - Statistical Modeling of Complex Systems and Processes'.

Contents

Introduction

> I know it's trivial, but I've forgotten why.
>
> Allan Guth

Genetics as a subarea of biology is the science of heredity of genetic characteristics, dealing with resemblances and differences of related organisms resulting from the interaction of their genes and the environment. It is concerned with the transmission of hereditary dispositions from parents to their progeny as well as variations of these phenomena. This area of research originated in Gregor Mendel's crossing experiments on peas around 1865 and has applications in the breeding of improved yielding plants and animals, and genetic testing, for instance (Lex, 1999). Enormous advances in gene technology (blotting, chromatography, electrophoresis, polymerase chain reaction, DNA-sequencing, ...) and detecting technologies (genetic chips, sensor-technology, ...) from the mid 1980's on are intrinsically tied to the step from the investigation of single genes to the analysis of the whole nucleotide sequence of genomes (entire set of genetic material), sequences generally in the range of billions of nucleotide pairs. Genomics denotes the systematic analysis of the complete genome, or all active genes, of an organism and includes the creation of gene libraries, genetic mapping of genes on chromosomes, sequencing as well as sequence analysis, also known as "decoding" of the DNA (Lex, 1999; Bickel et al., 2009; Becker, 2011). Genomics aims at improving the understanding of evolution of organisms, evolutionary biological processes as well as the emergence of diseases (Lex, 1999). In the year 1990 the Human Genomes Project started with the goal of sequencing the whole human genome and successfully finished in 2003 with a reported sequence coverage of 99% of the human genome to an accuracy of 99.99% (Bickel et al., 2009). Nowadays, there are many fully-sequenced model organisms, for example the flower *Arabidopsis thaliana* (The 1001 Genomes Consortium, 2016), which are the basis for many genomic studies.

The continuously sinking sequencing and data-storing costs have shifted the focus in genomics from the acquisition of sequence data to functional aspects like the interaction of genes and the analysis of their (biological) function (Bickel et al., 2009). The occurrence of a quantitative characteristic (phenotype) often depends on the interaction of several genes, whose DNA-sequence and specific function is not known. Instead of genes, one speaks in general of quantitative trait loci (QTL), which stands for locations (in the following loci) on chromosomes that influence the heredity of quantitative traits (Becker, 2011). Molecular biological methods enable to specifically select genotypes with favored alleles such that the

overall aim in many studies is to link the phenotype to the genotype, which is often connected with enormous amounts of complex and noisy genomic data. This high-dimensional data offers many opportunities but also entails many challenges (Fan et al., 2014). The recognition of subtle patterns and heterogeneity become possible but they are tied to noise accumulation, spurious correlations and measurement errors (Fan et al., 2014). The demand for high-dimensional statistical models including tools for prediction and inference, where the number of parameter is of much larger order than the sample size, is definitely given (Bickel et al., 2009; Fan et al., 2014). Possible coping methods include supervised learning (Hastie et al., 2008) such as dimension reduction, variable selection and regularization approaches (Hoerl and Kennard, 1970; Tibshirani, 1996; Fan and Lv, 2008; Bühlmann and van de Geer, 2011), mixed modeling (Henderson, 1984; Searle et al., 1992) and Bayesian statistics (Gelman et al., 2014).

In quantitative genetics, the additive genetic variance is defined as the variance of the additive genetic value in a population. It is the chief cause of resemblance between relatives and therefore the most important determinant of the response of a population to selection (Falconer and Mackay, 1996). In addition to that, the additive variance is a main component of the (narrow-sense) heritability, which is defined as the proportion of the phenotypic variance that can be explained by the additive variance of the genotypic value (Falconer and Mackay, 1996). The heritability is one of the main objectives in many genetic studies and is eminent, amongst other things, for the prediction of the response to selection in the breeder's equation (Piepho and Moehring, 2007; Hill, 2010).
The genomic variance, the genomic equivalent to the genetic variance, is defined as the variance of a trait that can be explained by a linear regression on a set of markers (de los Campos et al., 2015). Many authors have been chasing what is sometimes coined "missing heritability" (Maher, 2008) which means that only a fraction of the "true" genetic variance can be captured by a regression on influential loci.
Approaches for the estimation of the genomic variance include single-marker fixed effect regression in genome-wide association studies (Maher, 2008), a joint fit of all common markers in a method termed genome-wide complex trait analysis (GCTA) GREML (Yang et al., 2011), and using the posterior distribution of marker effects in Bayesian regression (Lehermeier et al., 2017).
Recently, there has been a general discussion whether estimators for the genomic variance account for linkage disequilibrium (LD) between markers, which is defined as the covariance between additive effects of marker pairs (Bulmer, 1971). Some authors argue that estimators similar to GCTA-GREML lack the contribution of LD (de los Campos et al., 2015; Kumar et al., 2015, 2016; Lehermeier et al., 2017) whereas others (Yang et al., 2016) resolutely disagree.

In Chapter 1 of this thesis we introduce relevant definitions and aspects of quantitative genetics and genomics. We elaborate on the "missing heritability" and existing estimation approaches of the genomic variance in the literature. In

Chapter 2 we investigate the additive genomic variance in linear regression models within the framework of quantitative genetics. This connection is reflected in the fixed effect model (FEM), where the regression parameter β is deterministic and the genomic variability comes in only through the randomness of the marker content. However, unbiased FEM's like ordinary-least-squares (OLS) are ill-posed in genomic datasets that are characterized by their high-dimensionality. As a remedy, Bayesian regression models (BRM) and random effect models (REM) are often used. In these models, though, the effect vector β is defined as a random variable and therefore these models do not lie within the classical framework of quantitative genetics. We investigate the expression for the genomic variance in FEM, BRM and REM and notice that, in general, the genomic variance strongly depends on the assumptions for the effect vector. We show that it is necessary to consider the genomic variance as a random quantity and not as a fixed population parameter in these model set-ups. In BRM, this results in the estimation of the posterior expectation of the genomic variance. In REM, we show that up-to-now, the genomic variance has been estimated as a parameter of the marginal, i.e. unconditional, model (e.g. GCTA-GREML). By strictly conditioning on the effect vector as in BRM, we constitute a paradigm shift from the estimation of the marginal genomic variance to the prediction of the random conditional genomic variance, which is structurally in perfect accordance to the posterior genomic variance in BRM. Inspired by the prediction of random effects (or in equivalent terminology: the estimation of the realized values of random effects) introduced by Henderson (1984) at the beginning of his chapter on prediction of random variables, we call our procedure the prediction of the genomic variance in REM. To this end, we introduce a mathematically founded best unbiased predictor for the genomic variance that is adapted to the specified model assumptions.
We take on the above mentioned critique that GCTA-GREML neglects the contribution of LD due to the diagonal covariance structure of the marginal β (Kumar et al., 2015, 2016). We show that the conditional genomic variance explicitly accounts for LD and remarkably reduces the "missing heritability". In addition to that, the difference of the novel predictor and the estimator of the marginal genomic variance in REM can be used as an indicator for the contribution of LD to the genomic variance. In general, the conditional genomic variance in REM is structurally similar to the genomic variance in FEM and therefore has an interpretation close to the classical genetic variance from quantitative genetics. Chapter 2 as well as Chapter 3, where we substantiate our theoretical findings by an exemplary simulation study based on the commonly used dataset on 1814 mice, are fully based on the *bioRxiv*-manuscript Schreck and Schlather (2018).

1. Preliminaries

> Life is the only art that we are required to practice without preparation, and without being allowed the preliminary trials, the failures and botches, that are essential for training.
>
> Lewis Mumford

1.1. Quantitative Genetics

This section is based on Falconer and Mackay (1996) with some extensions of my own.

Quantitative genetics deals with the inheritance of differences between individuals in a population, whose genetic constitution can be specified by the nature and the count of each genotype (exact genetic fixture by an individual set of genes). We only consider diploid organisms, i.e. organisms with two complete sets of chromosomes, which also defines the number of possible alleles for the genes.
The phenotypic value P, the degree of a certain characteristic of an individual, can be separated into the genetic value G and environmental deviations E

$$P = G + E, \tag{1.1}$$

where $\mathbb{E}[E] = 0$.

As a toy example, we firstly assume that the genome is made up of one single gene A with corresponding alleles A_1 and A_2. As a consequence, there exist three possible genotypes, namely the homozygotes A_1A_1 and A_2A_2 as well as the heterozygote A_1A_2. We assign the homozygote A_1A_1 the value $a \in \mathbb{R}$ and the other homozygote the value $-a$. The value of the heterozygote A_1A_2 equals the dominance deviation $d \in \mathbb{R}$ representing the effect of putting genes together to genotypes. The discrepancy d is caused by interactions of the alleles of a gene (within locus interactions). From now on we assume absence of dominance, i.e. $d = 0$.
Summarized, the genetic, or genotypic, value G of an individual depends on its

genotype as well as the assigned corresponding effect:

$$G := \begin{cases} a, & \text{if genotype } \mathrm{A_1A_1} \\ 0, & \text{if genotype } \mathrm{A_1A_2} \\ -a, & \text{if genotype } \mathrm{A_2A_2}. \end{cases} \tag{1.2}$$

We denote the frequency of allele A_1 in a population by p and the frequency of allele A_2 in the same population by q, so that $p+q=1$. The mean genotypic value in the population then equals $a(p-q)$. The genetic, or genotypic, variance is defined as the variance of the genetic value G. Heritability in the broad sense is the relative importance of heredity in determining phenotypic values, i.e. the extent to which individual phenotypes are determined by the genotype. The narrow-sense heritability h^2, in the following only the heritability h^2, is the extent to which phenotypes are determined by genes transmitted from parents, or the relative importance of the additive variance. In the absence of dominance, the genetic variance V at one locus equals

$$V := \mathrm{Var}(G) = p^2a^2 + q^2a^2 - a^2(p-q)^2 = 2pqa^2. \tag{1.3}$$

The additive genetic value G is considered as a random variable in (1.3), but the source of variation is not clearly specified.
Let us model the genotype at locus A as a discrete random variable X with three realizations by using the arbitrary equidistant coding $\mathrm{A_1A_1}\widehat{=}2$, $\mathrm{A_1A_2}\widehat{=}1$ and $\mathrm{A_2A_2}\widehat{=}0$ with the corresponding allele frequencies:

$$X = \begin{cases} 2, & \text{with } P(X=2) = p^2 \\ 1, & \text{with } P(X=1) = 2pq \\ 0, & \text{with } P(X=0) = q^2. \end{cases} \tag{1.4}$$

The random variable X has expectation $\mathbb{E}[X] = 2p$ and variance $\mathrm{Var}(X) = 2pq$. The genetic value G as in (1.2) can be expressed as

$$G = a(X-1).$$

The effect a is also called the effect of allele substitution because the effect of a genotype increases by the amount a for every additional allele $\mathrm{A_1}$. The genetic variance equals

$$V = \mathrm{Var}(G) = \mathrm{Var}(X)\, a^2 = 2pqa^2, \tag{1.5}$$

which coincides with (1.3) and is independent of linear shifts of the random variable X. Scaling of X by the constant b can be absorbed by the effect size a, and a/b then denotes the effect of allele substitution.

We extend this analysis to the multi-loci case by assuming that genomes are made-up of k bi-allelic genes $\mathrm{A, B, C, ...}$ with corresponding alleles $\mathrm{A_1, A_2, B_1, B_2, C_1, C_2, ...}$ In accordance to the single-locus case, there exist three

possible genotypes with their corresponding genotypic values a_j, allele frequencies p_j and genetic values G_j, $j = 1, ..., k$, at each locus. Usually, the total genetic value G of an individual is partitioned into additive ($\sum_{j=1}^{k} G_j$), dominance (D) and interaction (I) contributions

$$G = \sum_{j=1}^{k} G_j + D + I, \tag{1.6}$$

where the list of possible effects determining G in (1.6) is not exhaustive. Deviations from the purely additive combination of the single genotypic values of the loci are called interactions or epistasis deviations (I). Although non-additive genomic variation exists, most of the genetic variation can be explained by the additive model, such that it is sufficient to investigate the additive genetic variance (Hill et al., 2008). Specifically, epistasis I is only important on the gene-level but not for genetic variances (Hill et al., 2008), and Zhu et al. (2015) show that for human complex traits dominance variation D contributes little. In this thesis we assume that genes act additively within each locus ($D = 0$) and between loci ($I = 0$). Consequently, we can write G as the sum of the genotypic values of the different loci and calculate as in Bulmer (1971)

$$V := \mathrm{Var}(G) = \mathrm{Var}\left(\sum_{j=1}^{k} G_j\right) = \sum_{j=1}^{k} \mathrm{Var}(G_j) + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} \mathrm{Cov}(G_i, G_j), \tag{1.7}$$

where the sum of the variances of genetic values of the single loci is called genic variance. Linkage disequilibrium (LD) is defined as the covariance between the additive effects of the genes (Bulmer, 1971) and the contribution of LD is zero under random mating and the absence of selection (Bulmer, 1971). LD is an important factor for the genetic variance, especially when departing from random mating and Hardy-Weinberg equilibrium, which is generally the case in breeding (Hill et al., 2008; Dempfle, 2018), for instance.
We model the genotypes of the single loci as a random k-vector with an arbitrary covariance structure and an equidistant coding of the genotypes of the loci as in (1.4). Then, the genetic value G can be expressed as

$$G = \sum_{j=1}^{k} G_j = \sum_{j=1}^{k} a_j (X_j - 1). \tag{1.8}$$

The genetic variance (1.7) equals

$$V = \sum_{j=1}^{k} a_j \mathrm{Var}(X_j) + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} a_i a_j \mathrm{Cov}(X_i, X_j), \tag{1.9}$$

independent of linear shifts of the X_j, $j = 1, ..., k$. This implies that the variability of the genetic value is caused by the variability in the genotypes, and their (fixed)

effects a_j, $j = 1, ..., k$, act as factors, see (1.9). The difference of individuals in their genetic values is caused by the inter-individual differences in allele genotypes at QTL (Gianola et al., 2009; de los Campos et al., 2015). This conclusion is going to be a vital foundation for the analysis in Chapter 2.

1.2. Genomics

The effects of quantitative trait loci (QTL) are often very small and disguised by environmental noise and cannot be detected directly in the phenotypes. Possible solutions include the use of molecular markers that are linked to the object of interest. Markers are DNA-sequences that are not necessarily the DNA-sequences of the genes for a certain trait but "close" to it and are inherited together with the genes (Becker, 2011). In order to make inferences about the QTL using the markers, it is necessary that the markers are in LD with the QTL, i.e. that the distribution of the alleles for the markers and the QTL are not independent. Molecular markers can be produced by several methods. They all have in common that differences between genotypes are made visible in their DNA-sequence. The so-called Single Nucleotide Polymorphisms (SNP) are the most effective sort of markers because every variation in the nucleotide sequence of the DNA can be detected (Bickel et al., 2009; Becker, 2011).

We assume that the genome is mapped with $p \in \mathbb{N}$ markers. The phenotype of n individuals is regressed on the marker-data in order to make investigations and inferences on the contribution of the markers to the phenotype. We consider the phenotype-genotype regression model

$$y = \mu + g + \varepsilon \tag{1.10}$$

as the genomic equivalent to (1.1). By μ we denote a fixed intercept column-n-vector with equal entries, g is the n-vector of genomic values and $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, $i = 1, ..., n$, denotes the environmental deviations.
In accordance to the genetic characteristics in Section 1.1, the genomic value g is defined as the sum of the genomic values at each marker. We denote by $\mathbf{X}$ the $n \times p$ design matrix coding the genotypes of the markers similar to the coding of the genotypes of the QTL in (1.4), and by β we denote the p-vector of marker effects. Then, the genomic values can be separated into the coded genotype of the single markers and their corresponding effects:

$$g := \mathbf{X}\beta. \tag{1.11}$$

Model (1.10) is called linear equivalent model (Henderson, 1984) to the "standard" additive linear regression model

$$y = \mu + \mathbf{X}\beta + \varepsilon := \mu + \left(\sum_{j=1}^{p} x_{ij}\beta_j \right)_{i=1,...,n} + \varepsilon, \tag{1.12}$$

if y in (1.10) equals y in (1.12) in distribution. We consider mean-centered data: $\sum_{i=1}^{n} x_{ij} = 0$ for $j = 1, ..., p$, in order to be consistent with the literature. If necessary we adapt model (1.10) accordingly such that the equivalence remains. We consider deviations from the mean-centering assumptions in Appendix A.5.

The genomic variance is defined as the variance of the phenotypes that can be explained by a linear regression on a set of markers (de los Campos et al., 2015), the mean trace of the variances of the genomic values g_i, $i = 1, ..., n$, in model (1.10)

$$V^{\text{equi}} := \frac{1}{n}\text{tr}\big(\text{Cov}(g)\big), \tag{1.13}$$

or equivalently

$$V^{\text{real}} := \frac{1}{n}\text{tr}\big(\text{Cov}(\mathbf{X}\beta)\big) \tag{1.14}$$

in model (1.12). In accordance to Section 1.1, the genomic heritability is defined as the proportion of the genomic variance of the trait variance.

Statistical theory mainly focuses on model (1.12). For instance, it is possible to use fixed-effect regression like ordinary-least squares (OLS), see Appendix A.1. In this case, the genomic variance-covariance matrix $\text{Cov}(\mathbf{X}\beta)$ would constantly equal 0. Genomic data are usually high-dimensional, caused by the increasing number of markers p compared to the relatively small number of sequenced individuals n ($p \gg n$). Consequently, the matrix $\mathbf{X}$ is not of full column rank p and fitting a unique OLS model is ill-posed (Bühlmann and van de Geer, 2011).
Possible solutions include single-step regressions methods that are often used in genome-wide association studies (GWAS) to execute variable selection on the basis of p-values, and penalized estimation methods like Ridge Regression (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996), Sure Independence Screening (Fan and Lv, 2008) and many others. Variable selection is only meaningful with sparsity assumptions (Bühlmann and van de Geer, 2011) on the effect vector β, which in most applications contradicts the infinitesimal model with the assumption that all effects are small and that the number of QTL's p tends to infinity (Bulmer, 1971).
Random effect models (REM), in which the effect vector β is assigned a normal distribution, see Appendix A.3, are widely used especially in animal breeding (Henderson, 1984). Due to the paper of Meuwissen et al. (2001) the usage of Bayesian regression models (BRM) has strongly increased in genomics. Similar to the REM, the basic idea of adjustment in BRM's, see Appendix A.2, is to express uncertainty of the effect vector β by assigning it a prior distribution. Then, by adapting to the data by means of its likelihood, one attains the posterior distribution of the effect vector.
Both the BRM and REM are widely used in genomic applications. The variance of the effect vector β has the Bayesian interpretation of expressing uncertainty about the true but unknown effect value of the specific locus (variance of 0 only means

that there is no uncertainty, not that the effect is null). The frequentist interpretation of the variance of the effects β is as the variance in a conceptual sampling scheme where the effects are drawn at random from a population of loci (Gianola et al., 2009). Despite that, the genomic variance, $\text{Cov}(\mathbf{X}\beta)$, in BRM and REM is assumed to be caused by the randomness of the marker effects β. As a consequence, the genomic variance is not in accordance with the genetic variance from Section 1.1. We believe that this is the chief cause for the "missing heritability".

1.3. The "Missing Heritability"

In general, the expression "missing heritability" refers to the difference of the genetic and the genomic variance. Possible causes of the "missing heritability" include imperfect LD between the QTL and the markers, various genetic effects that cannot be captured by a linear regression on markers, inadequate theory and biased estimators. We focus on the difference of the additive genetic and genomic variance under the structural assumptions made in Section 1.1 and Section 1.2, particularly the linear and additive gene action, see (1.6). We review methods to estimate the genomic variance widely used in literature and connect them to the "missing heritability". This section is partly based on the *bioRxiv*-manuscript Schreck and Schlather (2018).

To begin with, researchers have used GWAS in order to find QTL by using single-marker fixed effect regression combined with variable selection based on p-values. After having added the estimated corresponding genomic variances of the single statistically significant loci, they asserted that they could only account for a fraction of the "true" genetic variance. For instance, Maher (2008) found that only 5% instead of the widely accepted heritability estimate of 80% of human height could be explained. Golan et al. (2014) state that the "true" genetic variance is generally underestimated when applying variable selection to genomic datasets which are typically characterized by their high-dimensionality, where the number of variables p is much larger than the number of observations n.

It is well known that a lot of traits are influenced by many genes and that at least some loci with tiny effects are missed when using variable selection or even single-marker regression models. As a consequence, Yang et al. (2010) decided to fit all common markers jointly using genomic best linear unbiased prediction (GBLUP), where they assume the genomic value g, see (1.11), to vary at random in the equivalent model (1.10). The n-vector of genomic values is assumed to follow a normal distribution:

$$g \sim \mathcal{N}\Big(0, \sigma_g^2 \mathbf{G}\Big). \tag{1.15}$$

In accordance with the infinitesimal model (Bulmer, 1971), the mean of the marginal genomic value is set to 0. The only unknown in the variance-covariance

matrix $\sigma_g^2\mathbf{G}$ of g is the parameter σ_g^2. The matrix $\mathbf{G}$ is a genomic relationship matrix (GRM), expressing the relationship of the n individuals under consideration. Often, the mean trace of $\mathbf{G}$ is normalized to equal 1 (VanRaden, 2008; Yang et al., 2010, 2011). Then, the genomic variance, see (1.13), is calculated as the mean of the genomic variances of the single individuals:

$$V_{\mathrm{r}}^{\mathrm{equi}} := \frac{1}{n}\mathrm{tr}\big(\mathrm{Cov}(g)\big) = \frac{1}{n}\mathrm{tr}\big(\sigma_g^2\mathbf{G}\big) = \sigma_g^2. \tag{1.16}$$

We note that in this approach, the variance component σ_g^2 of the linear model (1.10) equals the genomic variance. This variance component, however, exists due to model assumptions that do not reflect quantitative genetics theory from Section 1.1.
The variance component σ_g^2 is estimated by restricted maximum likelihood (REML), see Appendix A.3, in an approach termed genome-wide complex trait analysis genomic restricted maximum likelihood (GCTA-GREML) (Yang et al., 2010, 2011). The estimated equivalent genomic variance $\hat{V}_{\mathrm{r}}^{\mathrm{equi}}$ equals

$$\hat{V}_{\mathrm{r}}^{\mathrm{equi}} = \frac{1}{n}\mathrm{tr}(\mathbf{G})\hat{\sigma}_g^2 = \hat{\sigma}_g^2. \tag{1.17}$$

Yang et al. (2010) show that using this approach to quantify the combined effect of all SNP's in one regression model explains a larger part of the heritability than only using certain variants quantified by GWAS methods. They illustrate their results on the dataset on human height by pointing out that they can explain a heritability of about 45%. They conclude that the main reason for the remaining missing heritability is incomplete LD of causal variants with the genotyped SNPs, which refers to the general difference of genetic and genomic variances.
Kumar et al. (2015, 2016) criticize GCTA-GREML because of the assumption that the estimated GRM $\mathbf{G}$ is treated as a fixed quantity without sampling errors, although the GRM is actually a realization of an underlying stochastic process. In addition to that, we notice that the estimator (1.17) strongly depends on the specific form of $\mathbf{G}$ that determines the variance-covariance structure of the genomic effects g. Modifications or different constructions of this matrix lead to differences in estimated genomic variances (Legarra, 2015). This is not a favorable behavior and contradicts assumptions about the independence of the genetic variance with respect to coding, see Section 1.1.
Some authors argue that estimators similar to GCTA-GREML lack the contribution of LD (de los Campos et al., 2015; Kumar et al., 2015, 2016; Lehermeier et al., 2017) whereas others (Yang et al., 2016) resolutely disagree. More specifically, Kumar et al. (2015, 2016) state that in GCTA-GREML the contributions of the p markers to the phenotypic value are assumed to be independent normally distributed random variables with equal variances. Thus, they claim that the random contribution made by each marker is not correlated with the random contributions made by any other marker which leads to a negligence of the contribution of LD to the genomic variance. In order to clarify this critique, we consider the "original" linear model (1.12) in which the genotypic

value g is split up into the marker-genotype matrix $\mathbf{X}$ and its effects β. This corresponds to the equivalent model (1.10) by defining

$$\sigma_\beta^2 \mathbf{X}\mathbf{X}^\top = \frac{1}{p}\mathbf{X}\mathbf{X}^\top (p\sigma_\beta^2) =: \mathbf{G}\sigma_g^2, \tag{1.18}$$

where $\sigma_{\mathrm{g}}^2 := p\sigma_\beta^2$ and $\mathbf{G} := \frac{1}{p}\mathbf{X}\mathbf{X}^\top$ (VanRaden, 2008; Yang et al., 2010, 2011). We calculate the genomic variance-covariance matrix as

$$\mathrm{Cov}(\mathbf{X}\beta) = \mathbf{X}\mathbf{X}^\top \sigma_\beta^2. \tag{1.19}$$

The genomic variance (1.14) is estimated using (1.19):

$$\begin{aligned} \hat{V}_{\mathrm{r}}^{\mathrm{real}} = \frac{1}{n}\mathrm{tr}\Big(\widehat{\mathrm{Cov}}(\mathbf{X}\beta)\Big) &= \frac{1}{n}\mathrm{tr}\Big(\mathbf{X}\mathbf{X}^\top\Big)\hat{\sigma}_\beta^2 \\ &\approx \hat{\sigma}_\beta^2 \mathrm{tr}\Big(\hat{\Sigma}_X\Big) = \hat{\sigma}_\beta^2 \sum_{j=1}^{p} \widehat{\mathrm{Var}}(X_j), \end{aligned} \tag{1.20}$$

where we have used trace-properties and the unbiased method-of-moments estimator

$$\hat{\Sigma}_X := \frac{1}{n-1}\mathbf{X}^\top\mathbf{X} \tag{1.21}$$

for the variance-covariance matrix of the marker genotypes Σ_X (more details in Chapter 2). We conclude that the empirical variances of the marker genotypes influence the estimated genomic variance in this model, but due to the assumptions on the structure of the unconditional distribution of β, the empirical covariances between the markers do not contribute to the genomic variance in (1.20). This explicitly contradicts formula (1.9) in Section 1.1 and leads to the general assertions that estimators of this sort neglect the contribution of LD.

In a study on the model plant *Arabidopsis thaliana* (The 1001 Genomes Consortium, 2016), Lehermeier et al. (2017) use Bayesian ridge regression (BRR) to relate the phenotype flowering time to the genomic data. In BRR the effect vector β, the variance of β and the residual variance σ_ε^2 are assigned prior distributions, for more details see Appendix A.2. The authors derive an empirical sample variance of the genomic value g in (1.11) as

$$n^{-1}\beta^\top \mathbf{X}^\top \mathbf{X}\beta, \tag{1.22}$$

which, for mean-centered data, resembles an empirical genomic version of the genetic variance (1.9). They estimate (1.22) using the Markov Chain Monte Carlo (MCMC) samples $\big(\hat{\beta}^{(m)}\big)_{m=1,\ldots,M}$ in the estimator

$$\widehat{W}_{\mathrm{Post}} := \frac{1}{M}\sum_{m=1}^{M} \Big(\hat{\beta}^{(m)}\Big)^\top \hat{\Sigma}_X \hat{\beta}^{(m)} \tag{1.23}$$

which according to Lehermeier et al. (2017) draws samples from the posterior distribution of (1.22). The authors show that this estimator explains a larger proportion of the phenotypic variance than estimators like (1.17) and (1.20) (VanRaden, 2008; Yang et al., 2010, 2011) derived in REM's because $\widehat{W}_{\text{Post}}$ explicitly includes the contribution of LD. It is not clear how the estimator (1.23) was derived and what its corresponding theoretical value is. If β is treated as a random variable, then calculating a sample variance in (1.22) results in a random variable again and it is not clear why the theoretical variance of β was not used to derive a theoretical genomic variance term. This approach implicitly treats the genomic variance in BRM as a random variable and makes inferences about its posterior mean. All in all, despite lacking theoretical justification, this approach constitutes a very important step towards estimation approaches for the genomic variance including the contribution of LD that are based on MCMC sampling.

2. Genomic Variances

> There is nothing so practical as a good theory.
>
> Kurt Lewin

This chapter is mainly based on the *bioRxiv*-manuscript Schreck and Schlather (2018).

In Section 1.1 we have seen that the source of variation in the genetic value of an individual is the uncertainty in allele content at the QTL, whereas the genotypic effects are deterministic population parameters and therefore possess no variance, see (1.9). Contrary to that, in the regression analysis on markers in genomics in Section 1.2, the genomic variance is caused by the model assumptions for the effect vector. As a consequence, we have seen in Section 1.3 that the contribution of LD to the genomic variance is mostly neglected and "missing heritability" is created, especially in REM.

In this chapter, we execute a mathematically rigorous analysis of the estimation of the genomic variance with respect to the definitions in quantitative genetics. We first transfer the methodological assumptions from quantitative genetics to genomics, namely that the variability in the genomic value of an individual stems from the uncertainty in allele content of the markers, whereas the effects of the marker genotypes are assumed to be deterministic. In order to do so, we consider the basic additive linear model

$$Y = \mu + X\beta + \varepsilon, \tag{2.1}$$

where Y is the phenotype of a random individual, μ is a deterministic intercept and β is a p-vector of marker effects. The random allele content at the markers is coded by the random row-p-vector X similar to (1.4) with expectation $\mathbb{E}[X] = 0$ in order to be consistent with Section 1.3. In Appendix A.5 we consider deviations from this assumption. The covariance matrix of X is denoted by Σ_X. The residual ε is assumed to be independent of $X\beta$ and normally distributed with mean 0 and variance σ_ε^2.
In the realized model (1.12) one considers n independent realizations of (Y, X) without differences in the regression analysis (Bühlmann and van de Geer, 2011). Empirical mean-centering approximates the assumptions in the theoretical model (Bühlmann and van de Geer, 2011).

There is no theoretical equivalent form of (2.1) that induces model (1.10) similar to how model (2.1) induces (1.12). Therefore, we base the analysis in this section on model (2.1) and in the end transfer the resulting estimators and predictors to the equivalent realized model using relationship of the models (1.10) and (1.12).

To be consistent with Section 1.2, the additive genomic variance V is defined as the variance of the genomic value $X\beta$ which consists of the inter-individual differences in allele content at the markers as well as the effects of the markers themselves (de los Campos et al., 2015):

$$V := \mathrm{Var}(X\beta). \tag{2.2}$$

Due to independence of $X\beta$ and ε we can separate the phenotypic variance σ^2_Y in the genomic variance V and into the residual variance σ^2_ε:

$$\sigma^2_Y = V + \sigma^2_\varepsilon. \tag{2.3}$$

In Section 2.1, we define the genomic variance in the FEM as the genomic equivalent of the genetic variance (1.9). In Section 2.2 we provide the theoretical foundations for the estimation of the posterior genomic variance similar to that in Lehermeier et al. (2017). In Section 2.3 we improve on the estimation of the genomic variance in REM as in (1.17) and (1.20) by introducing the novel concept of the prediction of the genomic variance.

2.1. Fixed Effect Model (FEM)

We consider β in model (2.1) as a p-vector of fixed effects here, i.e. as a deterministic population parameter. Consequently, we calculate the genomic variance V defined in (2.2) as

$$\begin{aligned} V_\mathrm{f} := \mathrm{Var}(X\beta) &= \beta^\top \Sigma_X \beta \\ &= \underbrace{\sum_{j=1}^{p} \beta_j^2 \mathrm{Var}(X_j)}_{=:V_\mathrm{f}^\mathrm{g}} + \underbrace{\sum_{i=1}^{p} \sum_{\substack{j=1 \\ j\neq i}}^{p} \beta_i \beta_j \mathrm{Cov}(X_i, X_j)}_{=:V_\mathrm{f}^\mathrm{LD}}, \end{aligned} \tag{2.4}$$

which is the genomic equivalent of the genetic variance for multiple QTL in quantitative genetics, see (1.9).
We have split up the genomic variance in the FEM into the additive locus-specific variance V_f^g (also called genic variance) and the contribution V_f^LD of LD between different markers to the genomic variance. The genomic variance V_f in (2.4) is a weighted sum of the variances of the single marker content and the covariance between the content of the markers, where the weights are given by products of the elements of the fixed effect vector β. It is important to notice that in the FEM the genomic variance would constantly equal 0 if the analysis was based on the conditional model (1.12), see also Section 1.2.

The genomic variance V_f is a quadratic form in the unknown elements of β and simply plugging an unbiased estimator $\hat{\beta}$ for β and an unbiased estimator $\hat{\Sigma}_X$ for Σ_X into (2.4) leads to the biased estimator

$$\hat{V}_\mathrm{f}^{\mathrm{bias}} = \hat{\beta}^\top \hat{\Sigma}_X \hat{\beta} \tag{2.5}$$

for the genomic variance V_f. It has expectation

$$\begin{aligned}
\mathbb{E}\Big[\hat{V}_\mathrm{f}^{\mathrm{bias}}\Big] &= \mathbb{E}\Big[\hat{\beta}^\top \hat{\Sigma}_X \hat{\beta}\Big] \qquad (2.6)\\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbb{E}\Big[\hat{\sigma}_{ij}^X \hat{\beta}_i \hat{\beta}_j\Big] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \Big[\mathrm{Cov}\Big(\hat{\sigma}_{ij}^X, \hat{\beta}_i\hat{\beta}_j\Big) + \mathbb{E}\Big[\hat{\sigma}_{ij}^X\Big]\mathbb{E}\Big[\hat{\beta}_i\hat{\beta}_j\Big]\Big] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{Cov}\Big(\hat{\sigma}_{ij}^X, \hat{\beta}_i\hat{\beta}_j\Big) + \sigma_{ij}^X\Big[\sigma_{ij}^{\hat{\beta}} + \mathbb{E}\Big[\hat{\beta}_i\Big]\mathbb{E}\Big[\hat{\beta}_j\Big]\Big] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \Big(\sigma_{ij}^X \beta_i\beta_j + \sigma_{ij}^X \sigma_{ij}^{\hat{\beta}}\Big) + \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{Cov}\Big(\hat{\sigma}_{ij}^X, \hat{\beta}_i\hat{\beta}_j\Big) \\
&= \beta^\top \Sigma_X \beta + \mathrm{tr}\Big(\Sigma_X \Sigma_{\hat{\beta}}\Big) + \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{Cov}\Big(\hat{\sigma}_{ij}^X, \hat{\beta}_i\hat{\beta}_j\Big), \qquad (2.7)
\end{aligned}$$

where we denote the covariance between the random variables Z_i and Z_j by $\sigma_{ij}^Z := \mathrm{Cov}(Z_i, Z_j)$. The plug-in estimator $\hat{V}_\mathrm{f}^{\mathrm{bias}}$ contains second order products of the random variables $\hat{\beta}_j$, $j = 1, ..., p$, and is therefore biased by the amount

$$\begin{aligned}
\mathrm{Bias}\Big(\hat{V}_\mathrm{f}^{\mathrm{bias}}\Big) &:= \mathbb{E}\Big[\hat{V}_\mathrm{f}^{\mathrm{bias}}\Big] - V_\mathrm{f} \\
&\overset{(2.7),(2.4)}{=} \mathrm{tr}\Big(\Sigma_X \Sigma_{\hat{\beta}}\Big) + \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{Cov}\Big(\hat{\sigma}_{ij}^X, \hat{\beta}_i\hat{\beta}_j\Big), \qquad (2.8)
\end{aligned}$$

where only $\mathrm{tr}\big(\Sigma_X \Sigma_{\hat{\beta}}\big)$ is amenable to estimation.
Consequently, we correct for the covariance of the estimator $\hat{\beta}$ by defining

$$\hat{V}_\mathrm{f} := \hat{\beta}^\top \hat{\Sigma}_X \hat{\beta} - \mathrm{tr}\Big(\hat{\Sigma}_X \hat{\Sigma}_{\hat{\beta}}\Big) \tag{2.9}$$

as a less biased estimator for V_f, where $\hat{\Sigma}_{\hat{\beta}}$ denotes an unbiased estimator for the variance-covariance matrix $\Sigma_{\hat{\beta}} := \mathrm{Cov}\big(\hat{\beta}\big)$ of $\hat{\beta}$.

We investigate the bias-correction term $\mathrm{tr}\big(\hat{\Sigma}_X \hat{\Sigma}_{\hat{\beta}}\big)$ and find that

$$\begin{aligned}
\mathbb{E}\Big[\mathrm{tr}\Big(\hat{\Sigma}_X \hat{\Sigma}_{\hat{\beta}}\Big)\Big] &= \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbb{E}\Big[\hat{\sigma}_{ij}^{X}\hat{\sigma}_{ij}^{\hat{\beta}}\Big] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\sigma}_{ij}^{\hat{\beta}}\Big) + \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbb{E}\Big[\hat{\sigma}_{ij}^{X}\Big]\mathbb{E}\Big[\hat{\sigma}_{ij}^{\hat{\beta}}\Big] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\sigma}_{ij}^{\hat{\beta}}\Big) + \sum_{i=1}^{p}\sum_{j=1}^{p} \sigma_{ij}^{X}\sigma_{ij}^{\hat{\beta}} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\sigma}_{ij}^{\hat{\beta}}\Big) + \mathrm{tr}\Big(\Sigma_X \Sigma_{\hat{\beta}}\Big). \qquad (2.10)
\end{aligned}$$

Then, we examine the estimator $\hat{V}_{\mathrm{f}}$:

$$\begin{aligned}
\mathbb{E}\Big[\hat{V}_{\mathrm{f}}\Big] &= \mathbb{E}\Big[\hat{\beta}^{\top}\hat{\Sigma}_X\hat{\beta}\Big] - \mathbb{E}\Big[\mathrm{tr}\Big(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\Big)\Big] \\
&\overset{(2.7),(2.10)}{=} \beta^{\top}\Sigma_X\beta + \mathrm{tr}\Big(\Sigma_X\Sigma_{\hat{\beta}}\Big) + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\beta}_i\hat{\beta}_j\Big) \\
&\quad - \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\sigma}_{ij}^{\hat{\beta}}\Big) - \mathrm{tr}\Big(\Sigma_X\Sigma_{\hat{\beta}}\Big) \\
&= \beta^{\top}\Sigma_X\beta + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\beta}_i\hat{\beta}_j - \hat{\sigma}_{ij}^{\hat{\beta}}\Big). \qquad (2.11)
\end{aligned}$$

The estimator $\hat{V}_{\mathrm{f}}$ is biased by the amount

$$\begin{aligned}
\mathrm{Bias}\Big(\hat{V}_{\mathrm{f}}\Big) &:= \mathbb{E}\Big[\hat{V}_{\mathrm{f}}\Big] - V_{\mathrm{f}} \\
&\overset{(2.11),(2.4)}{=} \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\beta}_i\hat{\beta}_j - \hat{\sigma}_{ij}^{\hat{\beta}}\Big). \qquad (2.12)
\end{aligned}$$

It is not clear that the bias of $\hat{V}_{\mathrm{f}}$ in (2.12) is smaller than the bias of $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ in (2.8). But, the bias in (2.12) is caused only by dependencies between the unbiased plug-in estimators $\hat{\Sigma}_X$, $\hat{\beta}$ and $\hat{\Sigma}_{\hat{\beta}}$. If they are pairwise uncorrelated, the estimator V_{f} is unbiased, whereas $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ would still be biased by the amount $\mathrm{tr}(\Sigma_X\Sigma_{\hat{\beta}})$.
We call estimators "nearly unbiased", if they are biased only due to correlations between plug-in estimators.

In the case that the mean-centered realized design matrix $\mathbf{X}$ is of full column rank $p < n$ we can uniquely fit the realized linear model (1.12) using OLS.
As an outcome we obtain the estimated effect vector $\hat{\beta}$, its estimated covariance matrix $\hat{\Sigma}_{\hat{\beta}}$ and the estimator for the residual variance $\hat{\sigma}_{\varepsilon}^2$, see Appendix A.1.

Plugging these quantities into $\hat{V}_\mathrm{f}$ from (2.9) we obtain an improved estimator for the genomic variance V_f in (2.4). We first calculate

$$\begin{aligned}\mathrm{tr}\Big(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\Big) &\overset{(1.21),(A.2)}{=} \mathrm{tr}\bigg(\frac{1}{n-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\hat{\sigma}_\varepsilon^2\bigg)\\ &= \frac{1}{n-1}\,\hat{\sigma}_\varepsilon^2\,\mathrm{tr}(\mathbb{1}_{p\times p})\\ &= \frac{p}{n-1}\,\hat{\sigma}_\varepsilon^2. \end{aligned} \tag{2.13}$$

We express the nearly unbiased estimator $\hat{V}_\mathrm{f}$ in OLS as

$$\begin{aligned}\hat{V}_\mathrm{f} &\overset{(2.9)}{=} \hat{\beta}^\top\hat{\Sigma}_X\hat{\beta} - \mathrm{tr}\Big(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\beta}}\Big)\\ &\overset{(2.13)}{=} y^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\frac{1}{n-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y - \frac{p}{n-1}\hat{\sigma}_\varepsilon^2\\ &\overset{(A.1)}{=} \frac{1}{n-1}y^\top\mathbf{H}y - \frac{p}{n-1}\frac{1}{n-(p+1)}\Big[y^\top(1-\mathbf{H})y + \hat{\mu}^\top\hat{\mu} - 2y^\top(1-\mathbf{H})\hat{\mu}\Big]\\ &= \frac{1}{n-1}y^\top\mathbf{H}y + \Big(\frac{1}{n-1} - \frac{1}{n-(p+1)}\Big)\Big[y^\top(1-\mathbf{H})y + \hat{\mu}^\top\hat{\mu} - 2y^\top\hat{\mu}\Big]\\ &= \frac{1}{n-1}y^\top\mathbf{H}y + \frac{1}{n-1}\Big[y^\top(1-\mathbf{H})y + \hat{\mu}^\top\hat{\mu} - 2y^\top\hat{\mu}\Big] - \hat{\sigma}_\varepsilon^2\\ &= \frac{1}{n-1}y^\top y + \frac{n}{n-1}\bar{y}^2 - \frac{2}{n-1}\bar{y}\sum_{i=1}^n y_i - \hat{\sigma}_\varepsilon^2\\ &= \hat{\sigma}_y^2 - \hat{\sigma}_\varepsilon^2. \end{aligned}$$

We obtain the exact empirical variance decomposition

$$\hat{\sigma}_y^2 = \hat{V}_\mathrm{f} + \hat{\sigma}_\varepsilon^2 \tag{2.14}$$

in the OLS model which resembles the theoretical variance decomposition (2.3) in model (2.1). This implies that we can expect the improved estimator for the genomic variance and the estimator for the residual variance to sum up exactly to the phenotypic variance regardless of the data considered. When using the OLS method to fit a linear model, using the less biased estimator $\hat{V}_\mathrm{f}$, see (2.9), to estimate the genomic variance contribution of all markers is equivalent to simply subtracting the residual variance from the phenotypic variance.

In high-dimensional datasets the OLS method does either not lead to unique solutions or leads to estimated effects with large variances. Fixed effect solutions with lower variance include penalized regression methods. These estimates are obtained as the solution to an optimization problem that balances goodness of fit and model complexity and is of the general form

$$\hat{\beta} = \arg\min_\beta\{L(y,\beta) + \lambda J(\beta)\},$$

where $L(y, \beta)$ is a loss function that measures the lack of fit of the model to the data, $J(\beta)$ is a measure of model complexity and $\lambda \geq 0$ is a regularization parameter controlling the trade-off between goodness of fit and model complexity. These models do not produce unbiased estimators for the effect vector $\hat{\beta}$, which is required in the derivation of V_{f}, and most of the time it is not possible to explicitly correct for the bias. The theory for the FEM is consistent with quantitative genetics and the genomic variance (2.4) explicitly includes the contribution of LD. As the FEM is not appropriate in most genomic datasets, we extend the analysis of this section to the BRM and the REM, which are both commonly used in genomics.

2.2. Bayesian Regression Model (BRM)

Due to the paper of Meuwissen et al. (2001) the usage of Bayesian methods has strongly increased in quantitative genetics. The high-dimensionality of genomic data necessitates some way of regularization. The basic idea of adjustment in Bayesian regression models is to express uncertainty of the effect vector β by assigning it a prior distribution. Then, by adapting to the data by means of its likelihood, one attains the posterior distribution of the effect vector.
We consider the linear model (2.1) again where β possesses the prior distribution $p(\beta)$ with prior expectation μ_β (often chosen as 0) and prior variance-covariance matrix Σ_β. The specific form of the distribution of β is not relevant for the following analysis. The genomic variance V given by (2.2) equals

$$\begin{aligned}
V_{\mathrm{b}} &:= \mathrm{Var}(X\beta) \\
&= \mathrm{Var}\big(\mathbb{E}[X\beta \,|\, \beta]\big) + \mathbb{E}\big[\mathrm{Var}(X\beta \,|\, \beta)\big] \\
&= \mathrm{Var}\big(\mathbb{E}[X]\beta\big) + \mathbb{E}\Big[\beta^\top \Sigma_X \beta\Big] \\
&= \mathbb{E}[X]\Sigma_\beta \mathbb{E}[X]^\top + \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbb{E}\Big[\sigma_{ij}^X \beta_i \beta_j\Big] \\
&\overset{\mathbb{E}[X]=0}{=} \sum_{i=1}^{p}\sum_{j=1}^{p} \sigma_{ij}^X \Big(\sigma_{ij}^\beta + \mathbb{E}[\beta_i]\mathbb{E}[\beta_j]\Big) \\
&= \mathrm{tr}(\Sigma_X \Sigma_\beta) + \mu_\beta^\top \Sigma_X \mu_\beta. \qquad (2.15)
\end{aligned}$$

This expression for the genomic variance is meaningless because we can arbitrarily strongly influence it by the choice of the prior expectation and prior variance-covariance matrix. Instead, we require the genomic variance in BRM to move away from the prior assumptions by adapting to the data. In order to enable this Bayesian learning, we consider the variance of the genomic value $X\beta$ conditional on the effect vector β:

$$W := \mathrm{Var}(X\beta \,|\, \beta) = \beta^\top \Sigma_X \beta, \qquad (2.16)$$

which is a quadratic form in the effect vector β and consistent with (2.4). By assigning β a prior distribution, the genomic variance W (2.16) is assigned a prior distribution with prior expectation:

$$\begin{aligned}\mathbb{E}[W] &= \mathbb{E}\Big[\operatorname{tr}\Big(\Sigma_X \beta\beta^\top\Big)\Big] \\ &= \operatorname{tr}\Big(\Sigma_X \mathbb{E}\Big[\beta\beta^\top\Big]\Big) \\ &= \operatorname{tr}\Big(\Sigma_X \Big(\operatorname{Cov}(\beta) + \mathbb{E}\Big[\beta\Big]\mathbb{E}\Big[\beta^\top\Big]\Big)\Big) \\ &= \operatorname{tr}(\Sigma_X \Sigma_\beta) + \mu_\beta^\top \Sigma_X \mu_\beta \\ &= V_\mathrm{b}.\end{aligned}$$

Investigations in the BRM in the conditional linear model (1.12) are performed on the posterior distribution of β by adapting to the phenotypic data y. We use characteristics of the posterior distribution $p(\beta|y)$ of β to infer the posterior distribution of the genomic variance W given by (2.16), or equivalently the posterior distribution of the quadratic form W of β. We define the fixed posterior mean W_b of the genomic variance W as

$$\begin{aligned}W_\mathrm{b} &:= \mathbb{E}[W \,|\, y] \\ &\overset{(2.16)}{=} \operatorname{tr}\Big(\Sigma_X \mathbb{E}\Big[\beta\beta^\top \,\Big|\, y\Big]\Big) \\ &= \operatorname{tr}\Big(\Sigma_X \Big(\mathbb{E}\Big[\beta \,\Big|\, y\Big]\mathbb{E}\Big[\beta^\top \,\Big|\, y\Big] + \operatorname{Cov}(\beta \,|\, y)\Big)\Big) \\ &= \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + \operatorname{tr}\Big(\Sigma_X \Sigma_{\beta|y}\Big),\end{aligned} \tag{2.17}$$

which comprises the posterior expectation $\mu_{\beta|y} := \mathbb{E}[\beta \,|\, y]$ and the posterior variance-covariance matrix $\Sigma_{\beta|y} := \operatorname{Var}(\beta \,|\, y)$ of β. Consequently, we need not explicitly sample from the posterior distribution of W because it is enough to infer the posterior first and second moment of β.
The expression W_b structurally resembles the prior expectation V_b of W but includes the posterior mean as well as the posterior covariance of β instead of the prior moments. Structurally, the expressions for W and W_b resemble the genomic variance V_f given by (2.4) in the FEM in Section 2.1. Furthermore, they explicitly include the contribution of LD, where the role of the weights for the covariance terms of X $\big($formerly played by $\beta_i\beta_j$, $i \neq j$, in V_f, see equation (2.4)$\big)$ is taken over by the off-diagonal elements of the matrix of the posterior second moments $\mathbb{E}\big[\beta\beta^\top \,\big|\, y\big]$ of β. Hence, W_b can be split up in the genic variance and a part including the contribution of LD similar to V_f into (2.4).

There are many different approaches to fit the conditional model (1.12) in BRM that mainly differ in the choice of the prior distribution for the effect vector β. Analysis is always done on the posterior distribution of β from which samples are drawn using MCMC methods, for instance. Then, the fixed characteristics of the (posterior) effect vector can be estimated using for example the mean value and

the empirical variance of the resulting Markov chain.
In this context, we denote the sequence of the Markov chain of the estimated effects, after discarding the burn-in iterations and after thinning the chain, by the sequence of p-vectors $\left(\hat{\beta}^{(m)}\right)_{m=1,\ldots,M}$. These vectors are draws from the distribution $p(\beta|y)$. We express the quantities $\mu_{\beta|y}$ and $\Sigma_{\beta|y}$ by their empirical counterparts defined in the Appendix A.2, namely the estimated posterior mean $\hat{\mu}_{\beta|y}$ in (A.3) of β and the estimated posterior covariance $\hat{\Sigma}_{\beta|y}$ in (A.4). We propose to plug (A.3) and (A.4) into the estimator

$$\widehat{W}_{\mathrm{b}} := \underbrace{\hat{\mu}_{\beta|y}^{\top}\hat{\Sigma}_X\hat{\mu}_{\beta|y} - \mathrm{tr}\Big(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\mu}_{\beta|y}}\Big)}_{\widehat{W}_{\mathrm{b}}^{(1)}} + \underbrace{\mathrm{tr}\Big(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}\Big)}_{\widehat{W}_{\mathrm{b}}^{(2)}}, \tag{2.18}$$

for the mean of the posterior genomic variance W_{b}, see (2.17), in BRM, where $\hat{\Sigma}_{\hat{\mu}_{\beta|y}}$ denotes an unbiased estimator for the covariance $\Sigma_{\hat{\mu}_{\beta|y}}$, see (A.6), of the estimated effects $\hat{\mu}_{\beta|y}$.

The first part of expression (2.18), $\widehat{W}_{\mathrm{b}}^{(1)}$, is similar to $\hat{V}_{\mathrm{f}}$ such that we calculate similar to (2.11):

$$\begin{aligned}\mathbb{E}\Big[\widehat{W}_{\mathrm{b}}^{(1)}\Big] =& \mu_{\beta|y}^{\top}\Sigma_X\mu_{\beta|y} \\ &+ \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, (\hat{\mu}_{\beta|y})_i(\hat{\mu}_{\beta|y})_j - \hat{\sigma}_{ij}^{\hat{\mu}_{\beta|y}}\Big).\end{aligned}$$

We derive the expectation of the second part of expression (2.18), $\widehat{W}_{\mathrm{b}}^{(2)}$, as

$$\begin{aligned}\mathbb{E}\Big[\widehat{W}_{\mathrm{b}}^{(2)}\Big] &= \mathbb{E}\Big[\mathrm{tr}\big(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}\big)\Big] \\ &= \mathbb{E}\Big[\sum_{i=1}^{p}\sum_{j=1}^{p}\hat{\sigma}_{ij}^{X}\hat{\sigma}_{ij}^{\beta|y}\Big] \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\sigma}_{ij}^{\beta|y}\Big) + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}\Big[\hat{\sigma}_{ij}^{X}\Big]\mathbb{E}\Big[\hat{\sigma}_{ij}^{\beta|y}\Big] \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\sigma}_{ij}^{\beta|y}\Big) + \mathrm{tr}(\Sigma_X\Sigma_{\beta|y}).\end{aligned}$$

Combining these results, we find

$$\begin{aligned}\text{Bias}\Big(\widehat{W}_{\text{b}}\Big) &:= \mathbb{E}\Big[\widehat{W}_{\text{b}}\Big] - W_{\text{b}} \\ &= \mathbb{E}\Big[\widehat{W}_{\text{b}}^{(1)}\Big] + \mathbb{E}\Big[\widehat{W}_{\text{b}}^{(2)}\Big] - W_{\text{b}} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p} \text{Cov}\Big(\hat{\sigma}_{ij}^{X}, \hat{\sigma}_{ij}^{\beta|y} + (\hat{\mu}_{\beta|y})_i(\hat{\mu}_{\beta|y})_j - \hat{\sigma}_{ij}^{\hat{\mu}_{\beta|y}}\Big) \\ &\quad + \sum_{i=1}^{p}\sum_{j=1}^{p} \text{Cov}\Big(\hat{\sigma}_{ij}^{X}\Big).\end{aligned}$$

The remaining bias of the estimator $\widehat{W}_{\text{b}}$ vanishes if the estimators $\hat{\sigma}_{ij}^{X}$ and $\hat{\sigma}_{ij}^{\beta|y} - \hat{\sigma}_{ij}^{\hat{\mu}_{\beta|y}} + (\hat{\mu}_{\beta|y})_i(\hat{\mu}_{\beta|y})_j$ themselves are pairwise uncorrelated for all $i, j = 1, ..., p$.

We express the nearly unbiased estimator $\widehat{W}_{\text{b}}$ on the basis of MCMC realizations as

$$\widehat{W}_{\text{b}} \overset{(2.18),(A.7)}{\approx} \hat{\mu}_{\beta|y}^{\top}\hat{\Sigma}_X\hat{\mu}_{\beta|y} + \Big(1 - \frac{1}{M}\Big)\text{tr}\big(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}\big). \tag{2.19}$$

Plugging $\hat{\mu}_{\beta|y}$, see (A.3), and $\hat{\Sigma}_{\beta|y}$, see (A.4), into (2.19), we obtain:

$$\begin{aligned}\widehat{W}_{\text{b}} &\approx \hat{\mu}_{\beta|y}^{\top}\hat{\Sigma}_X\hat{\mu}_{\beta|y} + (1 - \frac{1}{M})\text{tr}(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}) \\ &= \Big(\frac{1}{M}\sum_{m=1}^{M}\big(\hat{\beta}^{(m)}\big)^{\top}\Big)\hat{\Sigma}_X\Big(\frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}\Big) \\ &\quad + \frac{M-1}{M}\Big[\frac{1}{M-1}\sum_{m=1}^{M}\big(\hat{\beta}^{(m)}\big)^{\top}\hat{\Sigma}_X\hat{\beta}^{(m)} \\ &\quad - \frac{1}{M(M-1)}\sum_{k=1}^{M}\sum_{m=1}^{M}\big(\hat{\beta}^{(k)}\big)^{\top}\hat{\Sigma}_X\hat{\beta}^{(m)}\Big] \\ &= \frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}^{(m)})^{\top}\hat{\Sigma}_X\hat{\beta}^{(m)} = \widehat{W}_{\text{Post}}.\end{aligned} \tag{1.23}$$

The estimator (1.23) can explicitly be interpreted as an estimator for the posterior mean of the genomic variance W in (2.16), in which the empirical mean is calculated using the realizations of W in every MCMC sample. The estimator $\widehat{W}_{\text{Post}}$ approximately equals the nearly unbiased estimator $\widehat{W}_{\text{b}}$ for the mean of the posterior genomic variance W_{b}. The estimator $\widehat{W}_{\text{Post}}$ is called M2 in Lehermeier et al. (2017), see also Section 1.3, and it had already been mentioned that this approach draws inferences on the posterior distribution of the genomic variance. By introducing the random genomic variance W with posterior mean W_{b} for the BRM in this section, we have laid the theoretical foundation and justification to

use (1.23) as a nearly unbiased estimator for the posterior mean of the genomic variance.

In this Bayesian approach, the expression W in (2.16) can also be interpreted as a parameter to be inferred. In order to do so, and to express uncertainty about its true value, we assign it a prior distribution. In our case this automatically happens by making inferences on the effect vector β. This prior knowledge about the genomic variance is expressed by the prior genomic variance V_{b} and then this belief is updated using the data. However, the prior belief influences the adaption to the data, which is only unproblematic in cases of perfect Bayesian learning.
After arriving at the posterior distribution, the value of the parameter can be inferred using some characteristic of the posterior distribution. In the next section we investigate the frequentist counterpart of the BRM, where the genomic variance will be purely treated as a random variable.

2.3. Random Effect Model (REM)

The effect vector β in model (2.1) is assumed to be a normally distributed random variable with mean 0 and diagonal variance-covariance matrix with equal variances σ^2_β, which is equivalent to modeling the single p components of β as independent random variables $\beta_j \sim \mathcal{N}(0, \sigma^2_\beta)$, $j = 1, ..., p$.
We obtain the marginal genomic variance V in model (2.1) defined in (2.2) as

$$\begin{aligned}
V_{\mathrm{r}} &= \mathrm{Var}(X\beta) \\
&= \mathrm{Var}\big(\mathbb{E}[X\beta \,|\, \beta]\big) + \mathbb{E}\big[\mathrm{Var}(X\beta \,|\, \beta)\big] \\
&= \mathrm{Var}\big(\mathbb{E}[X]\beta\big) + \mathbb{E}\Big[\beta^\top \Sigma_X \beta\Big] \\
&= \sigma^2_\beta \, \mathbb{E}[X]\mathbb{E}[X]^\top + \mathbb{E}\Big[\mathrm{tr}\Big(\Sigma_X \beta\beta^\top\Big)\Big] \\
&\overset{\mathbb{E}[X]=0}{=} \mathrm{tr}\Big(\Sigma_X \mathbb{E}\Big[\beta\beta^\top\Big]\Big) \\
&= \mathbb{E}[\beta]^\top \Sigma_X \mathbb{E}[\beta] + \mathrm{tr}(\Sigma_X \Sigma_\beta) \\
&= \sigma^2_\beta \mathrm{tr}(\Sigma_X) = \sigma^2_\beta \sum_{j=1}^{p} \mathrm{Var}(X_j). \qquad (2.20)
\end{aligned}$$

The specific form of the genomic variance V_{r} depends very much on the assumptions for the first and second moment of the distribution of β.
Subsequently, the marginal genomic variance V_{r} in (2.20) can be estimated by

$$\hat{V}_{\mathrm{r}} = \hat{\sigma}^2_\beta \mathrm{tr}(\hat{\Sigma}_X) = \frac{1}{n-1} \hat{\sigma}^2_\beta \mathrm{tr}\Big(\mathbf{X}^\top \mathbf{X}\Big) \qquad (2.21)$$

after obtaining an unbiased estimator $\hat{\sigma}^2_\beta$ for the variance component σ^2_β, see Appendix A.3. We calculate

$$\begin{aligned}\mathbb{E}\Big[\hat{V}_\mathrm{r}\Big] &= \mathbb{E}\Big[\hat{\sigma}^2_\beta \mathrm{tr}\big(\hat{\Sigma}_X\big)\Big] \\ &= \mathrm{Cov}\Big(\hat{\sigma}^2_\beta, \mathrm{tr}\big(\hat{\Sigma}_X\big)\Big) + \mathbb{E}\Big[\hat{\sigma}^2_\beta\Big]\mathbb{E}\Big[\mathrm{tr}\big(\hat{\Sigma}_X\big)\Big] \\ &= \sigma^2_\beta \mathrm{tr}(\Sigma_X) + \mathrm{Cov}\Big(\hat{\sigma}^2_\beta, \mathrm{tr}\big(\hat{\Sigma}_X\big)\Big).\end{aligned}$$

We conclude that $\hat{V}_\mathrm{r}$ is a nearly unbiased estimator for the genomic variance V_r, and is bias-free if the estimators $\hat{\sigma}^2_\beta$ and $\hat{\Sigma}_X$ are uncorrelated.

The estimators $\hat{V}_\mathrm{r}^\mathrm{real}$ and $\hat{V}_\mathrm{r}^\mathrm{equi}$ derived in Chapter 1.3 for the genomic variances based on the realized model (1.12) or the equivalent model (1.10), respectively, equal the estimator $\hat{V}_\mathrm{r}$ given by (2.21).
No matter which of these equivalent approaches to estimate the marginal genomic variance V_r in (2.20) is used, they are similar to the first part of expression V_f (2.4), namely $\sum_{j=1}^p \beta_j^2 \mathrm{Var}(X_j)$. But instead of weighting the variances of the allele content by different components of the (fixed) effect vector β, the weights in V_r, see (2.20), equal the variance component σ^2_β for every locus. More strikingly, the covariances between the different loci take no part in V_r in (2.20) but they do so in V_f in (2.4).
Nevertheless, it is not clear how strong LD is involved in the estimation of $\hat{\sigma}^2_\beta$ or $\hat{\sigma}^2_g$ in the REML equations and implicitly influences the estimates $\hat{V}_\mathrm{r}$ (2.21), $\hat{V}_\mathrm{r}^\mathrm{real}$ (1.20) and $\hat{V}_\mathrm{r}^\mathrm{equi}$ (1.17).

The assumptions on the marginal distribution of β (especially on its covariance structure) are very influential and cause the marginal genomic variance V_r in (2.20) to be unsatisfactory. This is similar to the genomic variance V_b in (2.15) in Section 2.2 that is arbitrarily strongly influenced by the prior moments of β.
Analogously to Section 2.2, we consider the genomic variance V conditionally on the effect vector β

$$W := \mathrm{Var}(X\beta \,|\, \beta) = \beta^\top \Sigma_X \beta = \mathrm{tr}\Big(\Sigma_X \beta\beta^\top\Big), \tag{2.16}$$

which is a quadratic form in the normally distributed effect vector β. This random variable has expectation

$$\begin{aligned}\mathbb{E}[W] &= \mathbb{E}\Big[\mathrm{tr}\Big(\Sigma_X \beta\beta^\top\Big)\Big] \\ &= \mathrm{tr}\Big(\Sigma_X \mathbb{E}\Big[\beta\beta^\top\Big]\Big) \\ &= \mathrm{tr}\Big(\Sigma_X \Big(\mathbb{E}\Big[\beta\Big]\mathbb{E}\Big[\beta^\top\Big] + \mathrm{Cov}(\beta)\Big)\Big) \\ &= \mathrm{tr}(\Sigma_X \sigma^2_\beta \mathbb{1}_{p\times p}) \\ &= \sigma^2_\beta \sum_{j=1}^p \mathrm{Var}(X_j) = V_\mathrm{r}.\end{aligned}$$

Investigations on the random variable W have to be done similar to investigations on the random effect β in REM, namely by a strict conditioning on the phenotypic data y in accordance to the prediction (Henderson, 1984) of the effect vector β, where the BLUP $\mu_{\beta|y} := \mathbb{E}[\beta \,|\, y]$ of β is given by the conditional expectation of β on y (Searle et al., 1992), see also Appendix A.3.

We define an unbiased predictor for the random genomic variance W in (2.16) as the expectation of the random variable W conditional on the data y

$$W_{\mathrm{r}} := \mathbb{E}[W \,|\, y] = \mathrm{tr}\Big(\Sigma_X \mathbb{E}\Big[\beta\beta^\top \,\Big|\, y\Big]\Big) = \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + \mathrm{tr}(\Sigma_X \Sigma_{\beta|y}), \tag{2.22}$$

which can be expressed by the conditional variance-covariance matrix $\Sigma_{\beta|y} := \mathrm{Cov}(\beta \,|\, y)$ of β, additional to the BLUP $\mu_{\beta|y}$.
The predictor W_{r} is by definition unbiased for the random variable W, if $\mathbb{E}[W_{\mathrm{r}}] = \mathbb{E}[W]$, which holds per construction:

$$\mathbb{E}[W_{\mathrm{r}}] = \mathbb{E}\big[\mathbb{E}[W \,|\, y]\big] = \mathbb{E}[W] = V_{\mathrm{r}}.$$

In addition to that, the predictor W_{r} is the minimum mean-squared-error (MMSE) predictor as a function in y for the random variable W:

Assume that we have an arbitrary predictor $g(y)$ as a measurable function in y for the random variable W. Then, it holds that

$$\begin{aligned}
\mathbb{E}\Big[\big(W - g(y)\big)^2 \,\Big|\, y\Big] &= \mathbb{E}\Big[\big(W - W_{\mathrm{r}} + W_{\mathrm{r}} - g(y)\big)^2 \,\Big|\, y\Big] \\
&= \mathbb{E}\Big[\big(W - W_{\mathrm{r}}\big)^2 \,\Big|\, y\Big] + \mathbb{E}\Big[\big(W_{\mathrm{r}} - g(y)\big)^2 \,\Big|\, y\Big] \\
&\quad + 2\big(W_{\mathrm{r}} - g(y)\big)\mathbb{E}\Big[\big(W - W_{\mathrm{r}}\big) \,\Big|\, y\Big] \\
&= \mathbb{E}\Big[\big(W - W_{\mathrm{r}}\big)^2 \,\Big|\, y\Big] + \big(W_{\mathrm{r}} - g(y)\big)^2.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{E}\Big[\big(W - g(y)\big)^2\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\big(W - g(y)\big)^2 \,\Big|\, y\Big]\Big] \\
&= \mathbb{E}\Big[\big(W - W_{\mathrm{r}}\big)^2\Big] + \mathbb{E}\Big[\big(W_{\mathrm{r}} - g(y)\big)^2\Big] \\
&\geq \mathbb{E}\Big[\big(W - W_{\mathrm{r}}\big)^2\Big],
\end{aligned}$$

where equality holds if $g(y) = W_{\mathrm{r}}$. This analysis is analogous to the one for the BLUP $\mu_{\beta|y}$, see Appendix A.3, which is the MMSE predictor as a function in y for the random effect β. We conclude that W_{r} is the optimal predictor for W with respect to MSE.

The predictor W_{r} for the genomic variance W is structurally in perfect accordance with the posterior genomic variance W_{b} in (2.17) and consequently has the same

interpretation as V_f, see (2.4), similar to the genetic variance in quantitative genetics. Most importantly, opposed to the marginal genomic variance V_r in (2.20), the predicted genomic variance W_r includes the contribution of LD similar to V_f in (2.4). This is achieved by weighting the covariances of X with the off-diagonals of the matrix of the conditional second moment $\mathbb{E}\big[\beta\beta^\top \,\big|\, y\big]$ of β. Hence, W_r can be split up in the genic variance and a part including the contribution of LD similar to V_f in (2.4).
The marginal covariance structure $\sigma^2_\beta \mathbb{1}_{p\times p}$ of β in V_r in (2.20), where its components are independent with equal variances, changes drastically when considering the conditional covariance structure $\Sigma_{\beta|y}$ of β, see (A.15). In this conditional approach, the single components of $\beta|y$ are not equally and independently distributed, but possess an arbitrary covariance structure by adapting to the data by means of the likelihood of the data similar to the posterior covariance $\Sigma_{\beta|y}$ in Section 2.2. Consequently, we tackle one of the central points of critique of GCTA-GREML issued by Kumar et al. (2015, 2016) by introducing the concept of the prediction of conditional genomic variance.

We plug (A.15) into W_r in (2.22) and obtain an analytical relationship to its expectation (the marginal genomic variance V_r):

$$\begin{aligned} W_\mathrm{r} &= \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + \mathrm{tr}(\Sigma_X \Sigma_{\beta|y}) \\ &= \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} + \sigma^2_\beta \mathrm{tr}(\Sigma_X) - \mathrm{tr}(\Sigma_X \sigma^2_\beta \mathbf{X}^\top \tilde{\Sigma} \mathbf{X} \sigma^2_\beta) \\ &= V_\mathrm{r} + \mu_{\beta|y}^\top \Sigma_X \mu_{\beta|y} - \mathrm{tr}(\Sigma_X \sigma^2_\beta \mathbf{X}^\top \tilde{\Sigma} \mathbf{X} \sigma^2_\beta). \end{aligned} \tag{2.23}$$

We replace the variance components σ^2_β and σ^2_ε in (2.23) by unbiased estimators and plug them into W_r:

$$\begin{aligned} \widehat{W}_\mathrm{r} &= \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} + \mathrm{tr}(\hat{\Sigma}_X \hat{\Sigma}_{\beta|y}), \\ &= \hat{V}_\mathrm{r} + \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} - \hat{\sigma}^4_\beta \mathrm{tr}\big(\hat{\Sigma}_X \mathbf{X}^\top \hat{\tilde{\Sigma}} \mathbf{X}\big). \end{aligned} \tag{2.24}$$

We make the important note that the unbiasedness of the predictor $\widehat{W}_\mathrm{r}$ can only be given conditional on the estimated variance components $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_\varepsilon$ because of dependencies between these estimators and y. This problem is common in REM

and also holds true for the BLUP, see Appendix A.3. We calculate

$$\begin{aligned}
\mathbb{E}\Big[\widehat{W}_{\mathrm{r}} \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] &= \mathbb{E}\Big[\hat{V}_{\mathrm{r}} + \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} - \hat{\sigma}_\beta^4 \mathrm{tr}\big(\hat{\Sigma}_X \mathbf{X}^\top \hat{\tilde{\Sigma}} \mathbf{X}\big) \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] \\
&= \hat{\sigma}_\beta^2 \mathbb{E}\Big[\mathrm{tr}(\hat{\Sigma}_X) \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] + \mathbb{E}\Big[\hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y} \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] \\
&\quad - \hat{\sigma}_\beta^4 \mathrm{tr}\big(\mathbb{E}\Big[\hat{\Sigma}_X \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] \mathbf{X}^\top \hat{\tilde{\Sigma}} \mathbf{X}\big) \\
&\overset{(2.7)}{=} \hat{\sigma}_\beta^2 \mathrm{tr}(\Sigma_X) + \mathbb{E}\Big[\hat{\mu}_{\beta|y} \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big]^\top \Sigma_X \mathbb{E}\Big[\hat{\mu}_{\beta|y} \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] \\
&\quad + \mathrm{tr}\Big(\Sigma_X \mathrm{Cov}\Big(\hat{\mu}_{\beta|y} \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big)\Big) - \hat{\sigma}_\beta^4 \mathrm{tr}\big(\Sigma_X \mathbf{X}^\top \hat{\tilde{\Sigma}} \mathbf{X}\big) \\
&\quad + \sum_{i=1}^p \sum_{j=1}^p \mathrm{Cov}\Big(\hat{\sigma}_{ij}^X, (\hat{\mu}_{\beta|y})_i (\hat{\mu}_{\beta|y})_j \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big) \\
&\overset{(A.19),(A.18)}{=} \hat{\sigma}_\beta^2 \mathrm{tr}(\Sigma_X) + \sum_{i=1}^p \sum_{j=1}^p \mathrm{Cov}\Big(\hat{\sigma}_{ij}^X, (\hat{\mu}_{\beta|y})_i (\hat{\mu}_{\beta|y})_j \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big) \\
&= \mathbb{E}\Big[W \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] + \sum_{i=1}^p \sum_{j=1}^p \mathrm{Cov}\Big(\hat{\sigma}_{ij}^X, (\hat{\mu}_{\beta|y})_i (\hat{\mu}_{\beta|y})_j \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big),
\end{aligned}$$

such that we can assert nearly (conditional) unbiasedness.

In the equivalent linear model (1.10) there is no theoretical analogue to model (2.1). Therefore, we perform the analysis in the model (2.1) and perform the estimation and prediction in model (1.12). Afterwards, we transfer the results to the equivalent model.
We can express the GBLUP $\mu_{g|y} := \mathbb{E}[g \,|\, y]$ as well as the conditional variance-covariance matrix $\Sigma_{g|y} := \mathrm{Cov}(g \,|\, y)$ using characteristics from the "original" linear model (1.12), see (A.20) and (A.21). We calculate

$$\frac{1}{n-1} \hat{\mu}_{g|y}^\top \hat{\mu}_{g|y} = \frac{1}{n-1} \hat{\mu}_{\beta|y}^\top \mathbf{X}^\top \mathbf{X} \hat{\mu}_{\beta|y} = \hat{\mu}_{\beta|y}^\top \hat{\Sigma}_X \hat{\mu}_{\beta|y}$$

and

$$\begin{aligned}
\frac{1}{n-1} \mathrm{tr}\Big(\hat{\Sigma}_{g|y}\Big) &= \frac{1}{n-1} \mathrm{tr}\Big(\mathbf{X} \hat{\Sigma}_{\mu_{\beta|y}} \mathbf{X}^\top\Big) \\
&= \frac{1}{n-1} \mathrm{tr}\Big(\mathbf{X}^\top \mathbf{X} \hat{\Sigma}_{\mu_{\beta|y}}\Big) \\
&= \mathrm{tr}\Big(\hat{\Sigma}_X \hat{\Sigma}_{\mu_{\beta|y}}\Big).
\end{aligned}$$

Consequently, we define

$$\widehat{W}_{\mathrm{r}}^{\mathrm{equi}} := \frac{1}{n-1} \hat{\mu}_{g|y}^\top \hat{\mu}_{g|y} + \frac{1}{n-1} \mathrm{tr}\big(\hat{\Sigma}_{g|y}\big) \tag{2.25}$$

in the linear model (1.12) as the analogous predictor to $\widehat{W}_{\mathrm{r}}$, see (2.24).

By considering the genomic variance in REM as random it becomes consistent with quantitative genetic theory, see Section 1.1, and explicitly includes the contribution of LD. The predictor for this random variable can be applied in the "standard" linear model as well as the equivalent linear model. We also bridge the gap between the estimation of the posterior mean of the genomic variance in BRM and estimation of the marginal variance in REM that has been observed in Lehermeier et al. (2017) and sketched in Section 1.3.
In Appendix A.4 we extend the considerations of this section to mixed effect models (MEM).

3. Empirical Analysis

> In theory there is no difference
> between theory and practice.
> In practice there is.
>
> Yogi Berra

We illustrate the theoretical results from Chapter 2. We performed all calculations with the free software R (R Development Core Team, 2017). This chapter is mainly based on the *bioRxiv*-manuscript Schreck and Schlather (2018).

3.1. Preparation of Datasets

3.1.1. Data Availability

For the simulation studies in Section 3.2 we considered the mice dataset that comes with the *R*-package "BGLR" (Perez and de los Campos, 2014). The data originally stem from an experiment from Valdar et al. (2006a,b) in a mice population. The dataset contains $p = 10346$ polymorphic markers that were measured in $n = 1814$ mice. The trait under consideration was body mass index (BMI) and body length (BL). In order to compare the estimators from the FEM with the ones from BRM and REM we created a second dataset (the reduced mice dataset) where we included only the first $\tilde{p} = 0.6\,n \approx 1088$ markers, such that $\tilde{p} < n$ holds true. This is not the best way to fully assess the structure of the genome (if the markers are ordered we probably consider the markers on the first chromosome(s) only). However, this was not the main goal of our investigation. We used the $n \times p$ ($n \times \tilde{p}$) matrix $\mathbf{X}$ coding the marker content from the mice (reduced mice) dataset to obtain a realistic LD-structure for the further analysis. In order to obtain modified datasets with different QTL-to-marker densities we assigned k out of the p ($\tilde{p}$) markers to be QTL. We denote by X_k the restriction of the marker content data to the (designated) QTL content. For each k, we calculated the covariance matrix Σ_{X_k} applying the method of moments estimator (1.21) to the QTL content data $\mathbf{X}_k$ with all individuals. We attributed each designated QTL with a corresponding "true" (fixed) effect k-vector β_k. Then, we calculated the V_k as

$$V_k = \beta^\top \Sigma_{X_k} \beta, \tag{3.1}$$

because it resembles the genetic variance, see (1.9) defined in Section 1.1.
It has been claimed that the main source of the "missing heritability" is imperfect

LD between markers and QTL (Yang et al., 2010) which we exclude by explicitly assigning markers to be QTL. In addition to that, the genomic variance under consideration is purely additive and the variance-covariance matrix of the QTL content is given. Consequently, the performance of the estimators and of the predictor depends only on their ability to represent the genomic variance V_k for all k.

In order to investigate the estimation (prediction) procedures for each k for different levels of heritabilities

$$h_k^2 := \frac{V_k}{V_k + \sigma_\varepsilon^2} \in \{0.2, 0.5, 0.8\},$$

we set the error variance σ_ε^2 equal to 1 and multiplied the "true" effect vector β_k by the constant c_k, where

$$c_k^2 := \frac{h_k^2}{1 - h_k^2} \frac{\sigma_\varepsilon^2}{V_k}.$$

This results in considering genetic variances of $V_k \in \{0.25, 1, 4\}$ for each QTL-marker ratio k/p ($k/\tilde{p}$). Therefore, we can drop the dependence of V_k and h_k^2 on k and in the following use only V and h^2. We drew n realizations of ε from a normal distribution with mean 0 and standard deviation $\sigma_\varepsilon = 1$, calculated the phenotypic values y_k using the additive linear model (1.12), and hence obtained several modified genomic datasets with phenotypic and genotypic values for each V and h^2.

Additional to the full mice dataset with phenotypes BMI and BL, we used the publicly available historical wheat dataset that comes with the *R*-package "BGLR" (Perez and de los Campos, 2014) for the analysis in Section 3.3. The data originally stems from CIMMYT's Global Wheat Program and consists of $n = 599$ lines of wheat where the trait under consideration was average grain yield. The phenotypes are divided up into four basic target sets of environments designated as Wheat I, Wheat II, Wheat III and Wheat IV. The lines where genotyped using Diversity Array Technology and after removing markers with allele frequencies lower than 0.05 we were left with $p = 1279$ polymorphism markers. More information on the dataset can be found in Perez and de los Campos (2014).
In addition to that, we analyzed a population of $n = 1057$ fully sequenced Arabidopsis lines for which phenotypes and genotypes are also publicly available by the effort of the Arabidopsis 1001 Genomes project
(The 1001 Genomes Consortium, 2016). The lines represent natural inbred lines and we examined the same trait, namely flowering time at 10°C (FT10), and the same $p = 193697$ SNP markers that were used in Lehermeier et al. (2017).

3.1.2. Model-fitting and Genomic Variance Calculation

Given the phenotypic and genotypic data described in Subsection 3.1.1, we fitted the OLS model using the R-function "lm" and obtained the estimated effect vector

$\hat{\beta}$ as well as the estimated error variances $\hat{\sigma}^2_\varepsilon$. We used these in order to calculate the biased estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ in (2.5) as well as the nearly unbiased estimator $\hat{V}_{\mathrm{f}}$ in (2.9). The OLS method is not appropriate in applications where the number of markers p is larger than the number of individuals n. Therefore we applied this method only to the reduced mice dataset after removing any collinearity.
We fitted the BRR model with the function "BGLR" with the specification of the model equal to "BRR" in the R-package "BGLR" (Perez and de los Campos, 2014). We decided to use 30000 iterations of the Markov chain and discarded the first 10000 as burn-in, after we had exemplarily checked the convergence of the resulting Markov chain and asserted convergence in every case. We kept only every fifth realization of the remaining chain in order to obtain approximate independence. This left us with $M = 4000$ state values that are assumed to be representative of the posterior distribution. As a result of the application we obtained estimators $\hat{\mu}$ for the intercept, $\hat{\sigma}^2_\varepsilon$ for the residual variance, and a $M \times p$ ($M \times \tilde{p}$) matrix with realizations of the estimated effect vector $\hat{\beta}^{(m)}$, $m = 1, ..., M$, in every state m of the considered Markov chain. We plugged these into $\widehat{W}_{\mathrm{Post}}$, see (1.23), in order to calculate an estimator for the posterior expectation of the genomic variance W_{b} defined in (2.17). BRR is conditionally on the variance components equivalent to the BLUP model, for more details see Appendix A.2 and Appendix A.3. Consequently, we expect similar results for both methods.
We fitted the GBLUP model in its equivalent form (1.10) as in Section 2.3 by using the R-package "sommer" (Covarrubias-Pazaran, 2017) and in particular its function "mmer". We obtained the predicted effects $\hat{\mu}_{g|y}$ and the estimated variance components $\hat{\sigma}^2_g$ and $\hat{\sigma}^2_\varepsilon$. We used these quantities in order to calculate the estimator $\hat{V}_{\mathrm{r}}^{\mathrm{equi}}$ in (1.17) for V_{r} and the predictor $\widehat{W}_{\mathrm{r}}^{\mathrm{equi}}$ in (2.25) for the conditional genomic variance W_{r}. Despite the explicit implementation of $\hat{V}_{\mathrm{r}}^{\mathrm{equi}}$ and $\widehat{W}_{\mathrm{r}}^{\mathrm{equi}}$ we use the equivalent quantities $\hat{V}_{\mathrm{r}}$, see (2.21), and $\widehat{W}_{\mathrm{r}}$, see (2.24), to describe the simulation studies in order to emphasize the derivation using the stochastic data-generating process X.

3.1.3. Performance Indexes

We compared each estimator $\hat{V}$ for the genomic variance V with respect to the absolute value of its relative bias

$$\mathrm{rBias}(\hat{V}) := \frac{|\mathbb{E}[\hat{V}] - V|}{V}, \tag{3.2}$$

and its relative root-mean-squared-error

$$\mathrm{rRMSE}(\hat{V}) := \sqrt{\frac{\mathbb{E}\big[(\hat{V} - V)^2\big]}{V}}. \tag{3.3}$$

For the analysis in Subsection 3.2.2 we define the relative contribution rLD of LD to the genomic variance V as

$$\mathrm{rLD}(V) := \frac{\sum_{i=1}^{p} \sum_{j=1, j\neq i}^{p} \beta_i^{(k)} \beta_j^{(k)} \mathrm{Cov}(X_i^{(k)}, X_j^{(k)})}{V}, \tag{3.4}$$

and the indicator I_{r} in REM for the contribution of LD to the genomic variance as

$$I_{\mathrm{r}} := \frac{\widehat{W}_{\mathrm{r}} - \hat{V}_{\mathrm{r}}}{\widehat{W}_{\mathrm{r}}}. \tag{3.5}$$

3.2. Simulation Studies

3.2.1. Variation of Observational Data

We randomly selected k QTL as described in Subsection 3.1.1 and fixed them for the further analysis. We chose the number k for the reduced mice dataset from the set $K_{\mathrm{rm}} := \{10, 50, 100, 200, 500, 1000\}$ and for the mice dataset from the set $K_{\mathrm{m}} := \{10, 100, 500, 1000, 2000, 5000, 10000\}$. For practical reasons of creating effect vectors with shapes of realizations of normal distributions or the heavier-tailed gamma distribution, we chose the "true" effect vector β_k as a realization (i.e. fixed value) according to the distributions depicted in Table 3.1. Formally, we considered an unknown data-generating process X with n realized p-vectors contained in the design matrices $\mathbf{X}$. We randomly selected $\tilde{n} = 0.8\,n$ out of the n realizations (individuals) 500 times for each combination of k and h^2 which imitates drawing from the data-generating process X. In each iteration, we calculated the estimators and the predictor in the OLS, BRR and (G)BLUP-models as described in Subsection 3.1.2.

The estimation performance of the biased estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ compared to the improved estimator $\hat{V}_{\mathrm{f}}$ from FEM in the reduced mice dataset is depicted in Figure 3.1 for a heritability of 0.2 ($V = 0.25$). The biased estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ performs drastically worse than the improved estimator $\hat{V}_{\mathrm{f}}$. This behavior of $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ is very similar for all considered h^2 which emphasizes the importance of the bias-correction in the FEM. For reasons of clarity we abstain from depicting the estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ in the further analysis.

We compared the performance of the remaining estimators and the predictor for the genomic variance in the reduced mice dataset for $h^2 = 0.2$ in Figure 3.2, for $h^2 = 0.5$ in Figure 3.3 and for $h^2 = 0.8$ in Figure 3.4. The estimated variances are averaged over the 500 realizations and are depicted in subject to the number of QTL k which also determines the QTL-marker ratios $k/\tilde{p}$.
The bias-corrected estimator $\hat{V}_{\mathrm{f}}$ given by (2.9) performs best and is very close to the "true" value of the genomic variance for all levels of heritabilities h^2 and numbers of QTL k.
The estimator $\widehat{W}_{\mathrm{b}}$, see (2.18), from the BRM overestimates the "true" genomic variance for $h^2 = 0.2$ for about over 10%. The performance of the estimator improves with larger heritability and for $h^2 = 0.8$ the estimator is very close to the "true" value for all k. A possible reason for the overestimation by $\widehat{W}_{\mathrm{b}}$ is that the model-fit in general could be poor such that the plugged-in state values are not representative of the posterior distribution, although the MCMC-algorithm had converged.

The estimation performance of $\hat{V}_{\mathrm{r}}$ given by (2.21) depends on the QTL-marker ratio. The underestimation of $\hat{V}_{\mathrm{r}}$ drastically increases with increasing number of QTL's k, whereas for a small QTL-marker ratio, the estimator $\hat{V}_{\mathrm{r}}$ tends to overestimate the genetic variance. The performance of the estimator strongly declines with increasing heritability, such that for $h^2 = 0.2$ the relative bias amounts to about 4%, for $h^2 = 0.5$ to 5%-15% and for $h^2 = 0.8$ to 5%-20%. The novel predictor $\widehat{W}_{\mathrm{r}}$ defined in (2.24) from the REM overestimates the "true" genomic variance for $h^2 = 0.2$ but nevertheless performs better than the estimators from the REM and the BRM. The predictor $\widehat{W}_{\mathrm{r}}$ performs relatively independent of the QTL-marker ratio and its performance advantage upon $\hat{V}_{\mathrm{r}}$ increases with increasing h^2. Although the "true" genomic variance is calculated according to the FEM, the performance of $\widehat{W}_{\mathrm{r}}$ can more than compete with the estimators $\hat{V}_{\mathrm{f}}$ from FEM and $\widehat{W}_{\mathrm{b}}$ from the BRM. We put special emphasis on the performance improvement of the novel predictor $\widehat{W}_{\mathrm{r}}$ versus the estimator $\hat{V}_{\mathrm{r}}$ in the case of higher heritability (Figure 3.4). This resembles the study of the "missing heritability" (Maher, 2008; Yang et al., 2010) and the novel predictor remarkably reduced the "missing heritability" in REM in our simulation study. The number of covariances that contribute to the genomic variance V_k depends quadratically $(k^2 - k)$ on k and we draw the conclusion that the increasing bias of $\hat{V}_{\mathrm{r}}$ in (2.21) with increasing k is due to the quadratic increase in the number of missed covariances. In contrast to that, the estimators $\hat{V}_{\mathrm{f}}$ in (2.9), $\widehat{W}_{\mathrm{b}}$ in (2.18) and the predictor $\widehat{W}_{\mathrm{r}}$ in (2.24), whose theoretical counterparts are in accordance to the genetic variance, fluctuate around the "true" value of the genomic variance independent on the number of covariances.

The performance of the estimators and the predictor from BRM and REM in the full mice dataset is very similar to the performance in the reduced mice dataset such that we can also assert the improved performance of $\widehat{W}_{\mathrm{b}}$ and $\widehat{W}_{\mathrm{r}}$ in the case of $p \gg n$. In addition to that, we compared the estimators and the predictor with respect to relative root-mean-squared-error, see (3.3), and assert similar behavior as when investigating the estimation bias. We conclude that treating the genomic variance as random is also advantageous with respect to the precision of the estimators and the predictor. All additional figures can be found in Appendix B.1.

Table 3.1.: Sources of Effect vector β in Subsection 3.2.1

K	β
10	$(1, 0.3, -0.5, 5, -2.4, 0.1, -0.6, 1.3, -2, -1.7)^\top$
50	$\mathcal{U}[-2.6, 3]$
100	$\mathcal{G}(0.1, 5)$
200	$\mathcal{N}(0.1, 0.38^2)$
500	$\mathcal{N}(0.2, 1)$
1000	$\mathcal{G}(0.03, 8)$
2000	$\mathcal{N}(0.1, 0.38^2)$
5000	$\mathcal{G}(0.03, 8)$
10000	$\mathcal{N}(0.1, 1)$

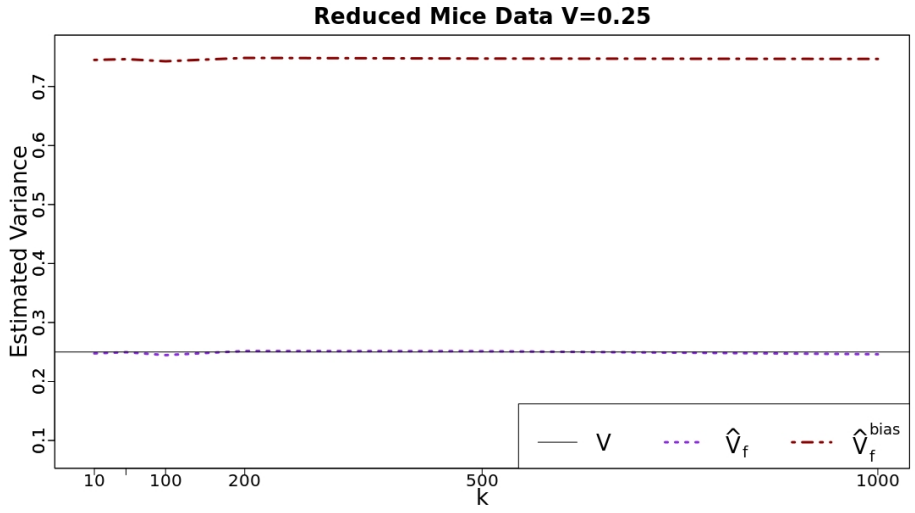

Figure 3.1.: Estimated variance in the FEM (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed numbers of markers $\tilde{p} = 1088$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The estimator $\hat{V}_\mathrm{f}$ performs remarkably better than the biased estimator $\hat{V}_\mathrm{f}^{\mathrm{bias}}$ and is very close to V independently of the QTL-to-marker ratio.

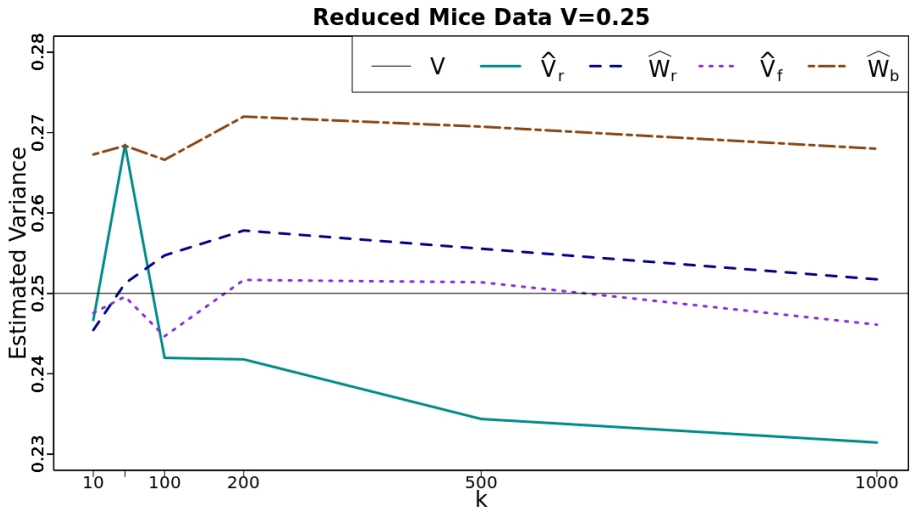

Figure 3.2.: Estimated variance (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed numbers of markers $\tilde{p} = 1088$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The estimator $\hat{V}_\mathrm{f}$ from FEM performs best followed by the predictor $\widehat{W}_\mathrm{r}$ for the conditional genomic variance in REM which slightly overestimates V. The estimator $\hat{V}_\mathrm{r}$ underestimates V and the bias of the estimators increases with k. The estimator for the posterior genomic variance $\widehat{W}_\mathrm{b}$ constantly overestimates V by around 10%.

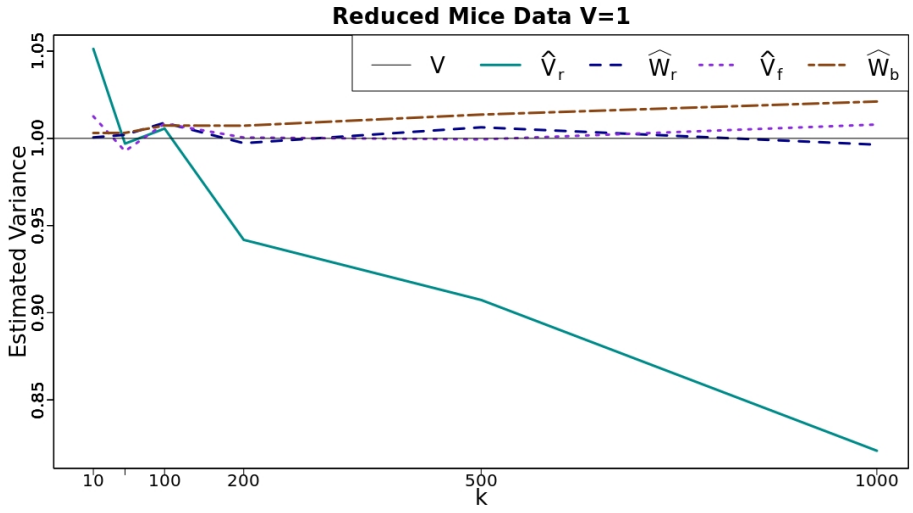

Figure 3.3.: Estimated variance (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed numbers of markers $\tilde{p} = 1088$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The estimator $\hat{V}_\mathrm{f}$ from the FEM and the predictor $\widehat{W}_\mathrm{r}$ from the REM are very close to the "true" V for all k. The estimator $\widehat{W}_\mathrm{b}$ for the posterior mean of genomic variance also performs well but slightly overestimates V with increasing k. The estimator $\hat{V}_\mathrm{r}$ drastically underestimates V and the bias of the estimator strongly increases with k.

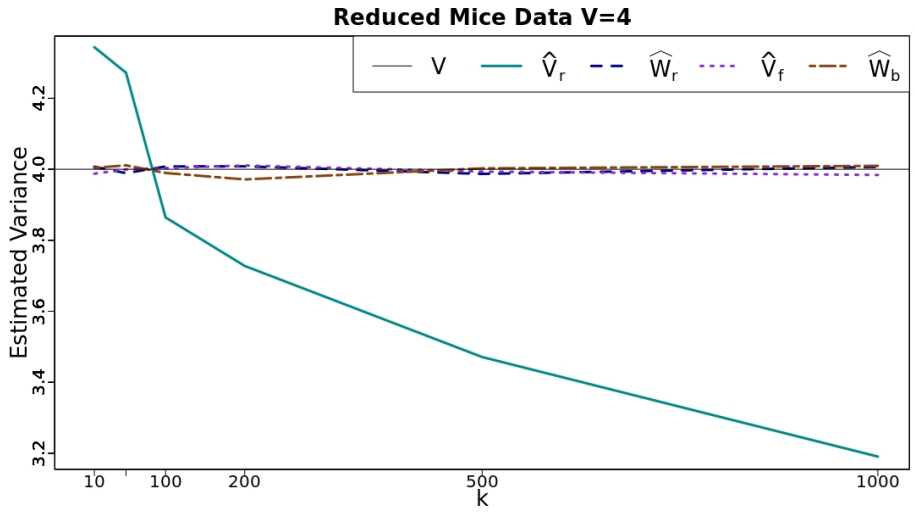

Figure 3.4.: Estimated variance (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed numbers of markers $\tilde{p} = 1088$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The estimator $\hat{V}_\mathrm{f}$ from the FEM, the predictor $\widehat{W}_\mathrm{r}$ from the REM and the estimator $\widehat{W}_\mathrm{b}$ are constantly very close to V for all QTL-to-marker ratios $k/\tilde{p}$. The estimator $\hat{V}_\mathrm{r}$ from the REM drastically underestimates V and the bias of the estimator strongly increases with k.

3.2.2. Variation of QTL-Allocations

In Subsection 3.2.1 we investigated the performance of the estimators and the predictor of the genomic variance for a fixed QTL-allocation and varying observations. Hence, it is possible that the conclusions made depend strongly on the specific QTL-allocation and the corresponding implied LD-structure, and cannot be generalized. Consequently, we considered the whole dataset of individuals and conducted the analysis in this section for different QTL-allocations for each level of heratibility and number of QTL k. In order to do so, we undertook 2000 iterations of randomly choosing the actual QTL-allocations for every level of heritability h^2 and each number of QTL $k \in K$, where $K_\mathrm{m} = \{10, 100, 500, 1000, 2000, ..., 10000\}$ for the mice dataset and $K_\mathrm{rm} = \{10, 50, 100, 200, ..., 1000\}$ for the reduced mice dataset. We used $\beta = (1, ..., 1)_k$ as the "true" effect vector in V_k, see (3.1), prior to scaling by c_k, in order to weight all locus-specific variances as well as all disequilibrium covariances equally.

We compared the performance of the estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$ in (2.5) to the improved estimator $\hat{V}_\mathrm{f}$ in (2.9) in the reduced mice dataset in Figure 3.5. Similar to Subsection 3.2.1 we notice that the bias-corrected estimator $\hat{V}_\mathrm{f}$ behaves much better than the estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$. In addition to that, $\hat{V}_\mathrm{f}$ fluctuates around the value of $V = 0.25$ for all k, which indicates that the performance is independent of the QTL-marker ratio. We observed similar behavior for $h^2 = 0.5$ and $h^2 = 0.8$ and,

for reasons of clarity, made these figures available in Appendix B.2.

In Figures 3.6 ($h^2 = 0.2$), 3.7 ($h^2 = 0.5$) and 3.8 ($h^2 = 0.8$) we depict the average of the estimators and the predictor over all considered QTL-allocations in the reduced mice dataset for different number of QTL k for a fixed number of markers $\tilde{p} = 1088$.
We notice that the behavior of all considered quantities over the range of k's is more bumpy compared to the analysis for a fixed QTL-allocation. This indicates that the QTL-allocation influences the estimators and the predictor. The general conduct of the estimator $\hat{V}_{\mathrm{f}}$, see (2.9), the estimator $\widehat{W}_{\mathrm{b}}$, see (2.18) and the predictor $\widehat{W}_{\mathrm{r}}$, see (2.24), is similar and independent of the level of heritability, as we notice that these quantities have spikes and slabs for the same k (same QTL-allocations) for each h^2. This indicates that $\hat{V}_{\mathrm{f}}$, $\widehat{W}_{\mathrm{b}}$, and $\widehat{W}_{\mathrm{r}}$ are in accordance and confirms that they can be used to estimate the genomic variance in accordance with quantitative genetic theory.
The estimator $\hat{V}_{\mathrm{f}}$ fluctuates around the "true" value of the genomic variance, whereas the estimator $\widehat{W}_{\mathrm{b}}$ constantly overestimates V for small h^2 as in Subsection 3.2.1. The predictor $\widehat{W}_{\mathrm{r}}$ fluctuates around the "true" value of the genomic variance for $h^2 = 0.2$ and slightly overestimates for larger heritabilities, but performs at least as good as $\hat{V}_{\mathrm{f}}$ and $\widehat{W}_{\mathrm{b}}$. The estimator $\hat{V}_{\mathrm{r}}$ from REM underestimates the "true" value of the genomic variance in all cases where the bias increases with increasing k regardless of h^2. Compared to the behavior in Subsection 3.2.1 where only one QTL-allocation was examined, the estimator $\hat{V}_{\mathrm{r}}$ underestimates V also for small k. The difference to the novel predictor $\widehat{W}_{\mathrm{r}}$ is striking. Especially for $h^2 = 0.8$ the estimator $\hat{V}_{\mathrm{r}}$ accounts for less than half of the genetic variance, which is in accordance with observations of the "missing heritability" (Maher, 2008; Yang et al., 2010). The missed covariances increase quadratically in k which explains the increasing bias of the estimator $\hat{V}_{\mathrm{r}}$. This simulation study indicates that the novel predictor $\widehat{W}_{\mathrm{r}}$ in (2.24) as well as the estimator $\widehat{W}_{\mathrm{b}}$ in (2.18) are possible solutions to the "missing heritability", and we conclude that this is due to their explicit inclusion of LD.

For each level of heritability h^2 and each number of QTL k we considered 2000 different QTL-allocations and each of them defines a specific LD-structure. Consequently, the "true" value of the genomic variance for each QTL-allocation can be distinguished by a different relative contribution of LD to V as defined in rLD in (3.4). We depict the empirical covariance of this relative contribution of LD with the value of $\hat{V}_{\mathrm{r}}$, $\widehat{W}_{\mathrm{r}}$, the indicator I_{r} given by (3.5), the relative bias (3.2) of $\hat{V}_{\mathrm{r}}$ and the relative bias of $\widehat{W}_{\mathrm{r}}$ for each h^2 and k in Figure 3.9.
The correlation of $\hat{V}_{\mathrm{r}}$ with the relative contribution of LD is negative (about -0.75) which indicates that the larger the contribution of LD, the smaller the estimator becomes. This is clearly contrasted by the novel predictor $\widehat{W}_{\mathrm{r}}$ which is approximately uncorrelated with the contribution of LD. In addition to that, the relative bias of $\hat{V}_{\mathrm{r}}$ is positively correlated (about 0.75) with the relative contribution of LD which demonstrates that the larger the contribution of LD, the

larger the bias of the estimator becomes. This is once again contrasted by the relative bias of $\widehat{W}_{\mathrm{r}}$ that is approximately uncorrelated to the contribution of LD. Strikingly, the empirical correlation of the indicator I_{r}, which can be calculated using only $\hat{V}_{\mathrm{r}}$ and $\widehat{W}_{\mathrm{r}}$, is positively correlated with the relative contribution of LD to the genomic variance. As a consequence, I_{r} constitutes a novel approximation of the relative contribution of LD to the genomic variance.

In addition to the analysis for the reduced mice dataset, we compared the estimators and the predictor in the full mice dataset where $p \gg n$. The performance of the estimator $\widehat{W}_{\mathrm{b}}$, $\hat{V}_{\mathrm{r}}$, and the predictor $\widehat{W}_{\mathrm{r}}$ are very similar to the performance in the reduced mice dataset. For reasons of clarity, we made the corresponding figures available in Appendix B.2.

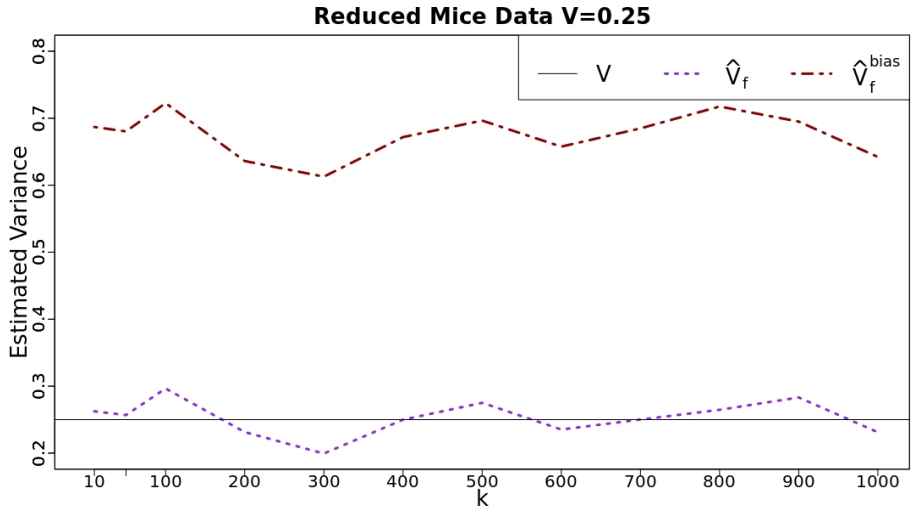

Figure 3.5.: Estimated variance in the FEM (mean value over different QTL-allocations) in the reduced mice dataset for different numbers of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The estimator $\hat{V}_{\mathrm{f}}$ performs remarkably better than the biased estimator $\hat{V}_{\mathrm{f}}^{\mathrm{bias}}$ and is very close to V independently of the QTL-to-marker ratio.

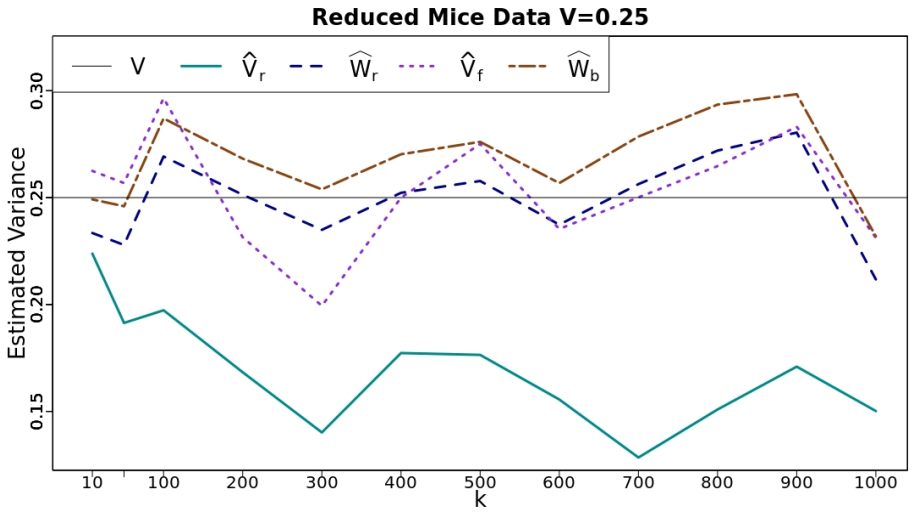

Figure 3.6.: Estimated variance (mean value over different QTL-allocations) in the reduced mice dataset for different numbers of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The estimator $\hat{V}_\mathrm{f}$ from the FEM performs similar to the predictor $\widehat{W}_\mathrm{r}$ from the REM and they are both close to the true V of 0.25. The estimator $\widehat{W}_\mathrm{b}$ from the BRM performs solidly but constantly sightly overestimates the "true" genomic variance. The estimator $\hat{V}_\mathrm{r}$ from the REM underestimates V by around 40% and the bias of the estimator tends to increase with the number of QTL k.

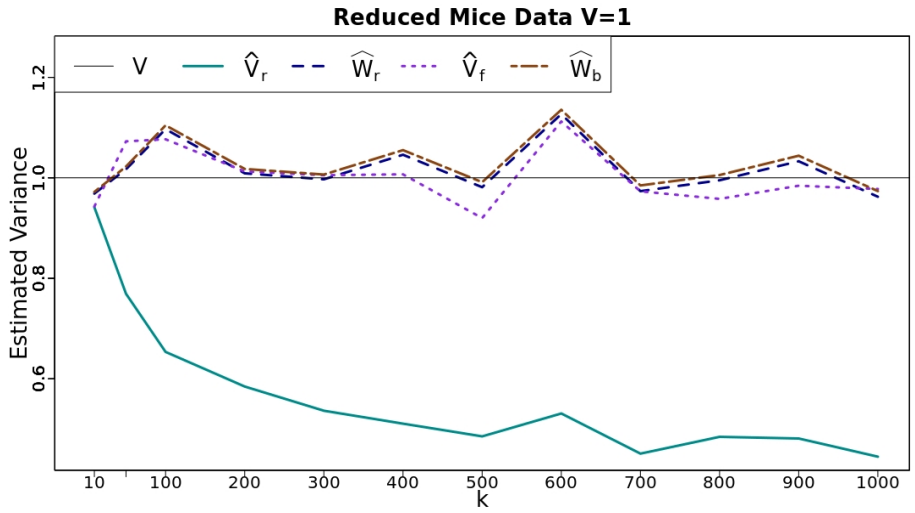

Figure 3.7.: Estimated variance (mean value over different QTL-allocations) in the reduced mice dataset for different numbers of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The estimator $\hat{V}_\mathrm{f}$ from the FEM performs similar to the predictor $\widehat{W}_\mathrm{r}$ from the REM and the estimator $\widehat{W}_\mathrm{b}$ from the BRM and they are all very close to V. The estimator $\hat{V}_\mathrm{r}$ from the REM underestimates V increasingly with the number of QTL k and by at least 40% starting at a QTL-to-marker ratio of 10%.

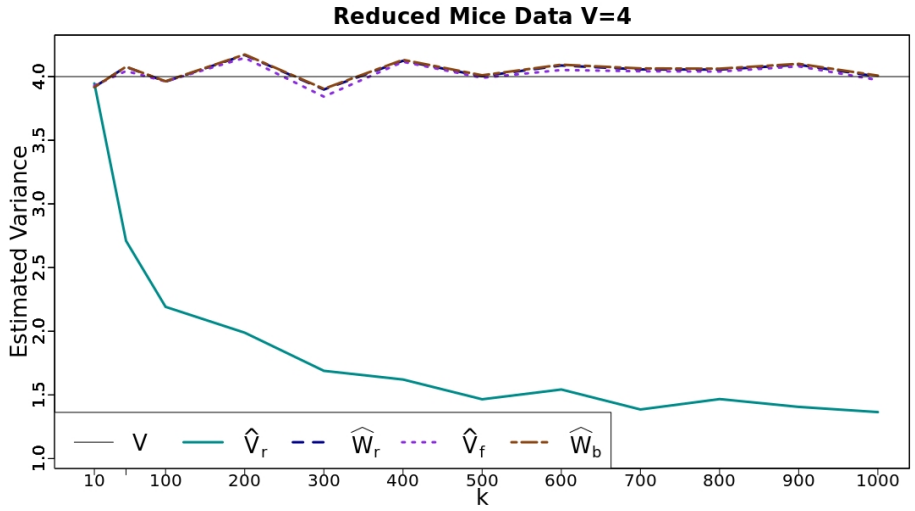

Figure 3.8.: Estimated variance (mean value over different QTL-allocations) in the reduced mice dataset for different numbers of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The estimator $\hat{V}_\mathrm{f}$ from the FEM, the predictor $\widehat{W}_\mathrm{r}$ from the REM and the estimator $\widehat{W}_\mathrm{b}$ from the BRM are very close to the true V of 4. The estimator $\hat{V}_\mathrm{r}$ from the REM drastically underestimates V and the bias of the estimator tends to increase with the number of QTL's k. For a large QTL-to-marker ratio, $\hat{V}_\mathrm{r}$ can only recover about 40% of the genetic variance.

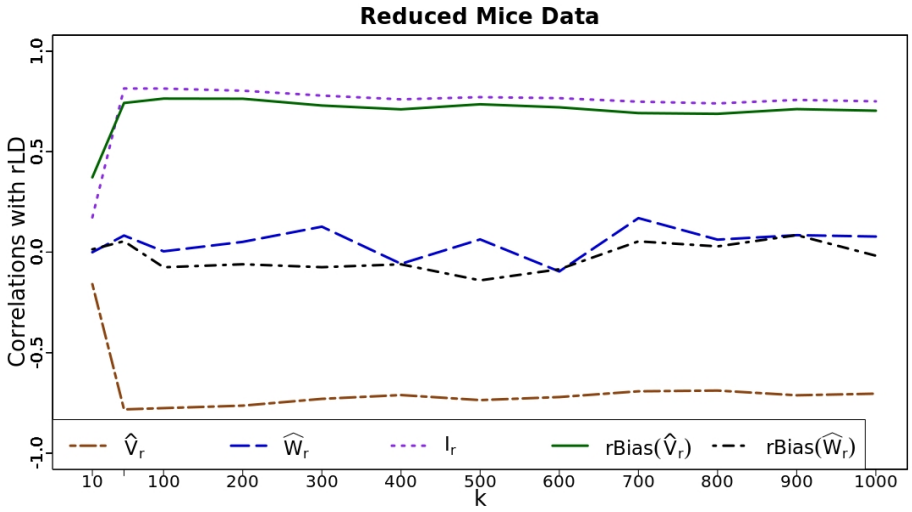

Figure 3.9.: Empirical correlations with the relative contribution of LD to the "true" genomic variance in the reduced mice dataset for different numbers of QTL k for fixed number of markers $\tilde{p} \approx 1088$. Values are averaged over the different levels of $h^2 \in \{0.2, 0.5, 0.8\}$. The correlation of the estimator $\hat{V}_\mathrm{r}$ with the relative contribution of LD is about -0.7 except for $k = 10$, whereas the correlation of the predictor $\widehat{W}_\mathrm{r}$ fluctuates around 0. The correlation of the relative bias of $\hat{V}_\mathrm{r}$ is about 0.7 except for $k = 10$ which indicates that the larger the contribution of LD to the genomic variance, the larger the bias of $\hat{V}_\mathrm{r}$ becomes. Contrary to that, the bias of the predictor $\widehat{W}_\mathrm{r}$ is approximately uncorrelated to the relative contribution of LD. The quantity I_r is positively correlated $(0.7 - 0.8)$ to the relative contribution of LD which makes it an usable indicator for the relative contribution of LD to the genomic variance.

3.3. Applications to Genomic Datasets

In Section 3.2 we noticed that the estimator $\hat{V}_{\mathrm{r}}$ underestimated the genetic variance most of the time, whereas the predictor $\widehat{W}_{\mathrm{r}}$ was very close to the "true" value, even in the high-dimensional set-up. We traced that back to the negligence of the contribution of LD to the genetic variance (1.9). If we take a closer look at the formula of the genetic variance and its genomic counterpart, we notice that the sign and the size of the contribution of LD depends also on the elements of the weighting vector β. As a consequence, it is also possible that the contribution of LD reduces the genomic variance, such that the estimator $\hat{V}_{\mathrm{r}}$ overestimates the genomic variance. The analysis in Subsection 3.2.2 has been done using the effect vector $\beta = (1, ..., 1)_k^\top$ in order to weight the contribution of each locus equally. But when different weighting vectors are used, no statement of the performance of the estimator and the predictor in applications can yet be made.

In this section, we apply the estimator $\hat{V}_{\mathrm{r}}$ and the predictor $\widehat{W}_{\mathrm{r}}$ to the genomic datasets with their corresponding traits introduced in Subsection 3.1.1. In this application, we cannot make any statements about whether the estimator $\hat{V}_{\mathrm{r}}$ or the predictor $\widehat{W}_{\mathrm{r}}$ is less biased because we do not know the value of the true genetic variance.
Similar to the analysis in Lehermeier et al. (2017), we standardize the phenotypic variance to equal 1 in order to be able to judge the variance decomposition of the phenotypic variance in the genomic variance and the residual variance, see (2.3).

In Table 3.2 we depict the estimated genomic variance $\hat{V}_{\mathrm{r}}$ in the first column, the predicted genomic variance $\widehat{W}_{\mathrm{r}}$ in the second column as well as the estimated residual variance $\hat{\sigma}_\varepsilon^2$ in the last column for the different datasets and traits.
For the Arabidopsis dataset in the first row we notice that the predictor is about twice the size of the estimator. While the sum of $\hat{V}_{\mathrm{r}}$ and $\hat{\sigma}_\varepsilon^2$ equals about 0.55, the sum of $\widehat{W}_{\mathrm{r}}$ and $\hat{\sigma}_\varepsilon^2$ is very close to the phenotypic variance 1. Thus, the predictor captures a larger amount of the phenotypic variance, whereas the estimator misses a large part of it. In accordance with the overall theme of this thesis, we carefully conclude that this remaining genomic variance is due to the contribution of LD.
We notice in all seven datasets that the sum of the predictor $\widehat{W}_{\mathrm{r}}$ and the estimated residual variance $\hat{\sigma}_\varepsilon^2$ is very close to the phenotypic variance, whereas the sum of $\hat{V}_{\mathrm{r}}$ and $\hat{\sigma}_\varepsilon^2$ does not equal the phenotypic variance. In the mice datasets, BMI and BL, as well as the wheat datasets the estimator $\hat{V}_{\mathrm{r}}$ is larger than the predictor $\widehat{W}_{\mathrm{r}}$ and therefore explains a larger part of the phenotypic variance. But the sum of $\hat{V}_{\mathrm{r}}$ and $\hat{\sigma}_\varepsilon^2$ in these cases exceeds the phenotypic variance. We conclude that in these particular cases, the contribution of LD to the genomic variance (sum of the covariances between the loci weighted by the effect vector β) is negative.

Consequently, it is also possible that estimators similar to $\hat{V}_{\mathrm{r}}$ overestimate the genomic variance. This entails problems in the definition of the heritability, namely that the quotient of the genomic variance and the phenotypic variance

differs from the quotient of the genomic variance and the sum of the genomic and residual variance.
The predictor $\widehat{W}_{\mathrm{r}}$ also corrects for a negative contribution of LD and behaves in accordance with the decomposition of the phenotypic variance, such that the heritability can be uniquely defined.
This analysis also indicates that in purely random effect models the unbiased predictor for the genomic variance equals the phenotypic variance minus the residual variance.

Table 3.2.: Empirical Variance Decompositions

	$\hat{V}_{\mathrm{r}}$	$\widehat{W}_{\mathrm{r}}$	$\hat{\sigma}^2_{\varepsilon}$
FT10	0.47486	0.92603	0.07389
BMI	0.22284	0.18301	0.81702
BL	0.37405	0.29840	0.70163
Wheat I	0.60417	0.45908	0.54094
Wheat II	0.53626	0.43503	0.56500
Wheat III	0.43330	0.34803	0.65204
Wheat IV	0.49061	0.40896	0.59112

Concluding Remarks

Would it help?

Rudolf Abel in "Bridge of Spies" as a response to "Do you never worry?"

The original aim of my research was a full and mathematically rigorous analysis of the genomic variance in phenotype-genotype regression models. The main reason for that was the lack of a solid theoretical basis on which to build the estimation of genomic variances, which led to various estimators in many different models, producing different results. This brought about discussions about the unbiasedness of the estimators, about the "missing heritability", and about whether estimators respect for the contribution of linkage disequilibrium.

Results

The key to the results in Chapter 2 was to accommodate the theory of the genomic variance with quantitative genetics from Section 1.1. We believe that the randomness of the effect vector β in Bayesian regression and random effect models should be considered as a way of regularization and not as the source of genomic variation. We changed the source of variation from the marker effects to the marker content. This enabled an estimation of a genomic variance different from 0 in the FEM and resulted in treating the genomic variance in the BRM and the REM as a random variable in the effect vector. The expectation of this random variable with respect to y equals the marginal, or prior, genomic variance. By adapting to the actual data, we move away from these assumptions. The posterior distribution of this random variable is investigated in the Bayesian model, whereas in the random effect model it resulted in the prediction of the genomic variance, which is in accordance with the prediction of the random effects themselves.

We noticed that the expression for the genomic variance as a fixed population parameter strongly depends on the model assumptions for the effect vector β. Consequently, it is important to distinguish the analysis of the genomic variance in the FEM, the BRM and the REM.
The genomic variance V_{f} in the FEM in Section 2.1 is the genomic equivalent of the genetic variance in quantitative genetics from Section 1.1 and explicitly includes the contribution of LD. We derived the nearly unbiased estimator $\hat{V}_{\mathrm{f}}$ in (2.9).
The genomic variance V_{b} as a population parameter in the BRM in Section 2.2

proved to be meaningless because of its dependence on characteristics of the prior distribution of the effect vector. In order to include characteristics of the posterior distribution of β, we proposed to consider the genomic variance as the random variable W with prior expectation V_{b}. Estimation in this model resulted in the posterior expectation W_{b}, see (2.17), which includes the contribution of LD and has an interpretation similar to the genomic variance in the FEM, even in high-dimensional genomic datasets. We laid the theoretical foundations for the posterior genomic variance in BRM, whose mean had already been estimated without a study of the random variable under consideration, for instance in Lehermeier et al. (2017).
The genomic variance in REM has been treated as the parameter V_{r} given by (2.20), and popular estimation methods (e.g. GCTA-GREML) are based on the marginal covariance matrix of the effect vector β which leads to a negligence of the contribution of LD. In perfect accordance with the BRM, we introduced the novel concept of the random genomic variance W, see (2.16), in REM by conditioning on the effect vector β. We derived a nearly unbiased predictor $\widehat{W}_{\mathrm{r}}$ in (2.24) for the random genomic variance in REM that is based on the covariance of the conditional distribution of β given the data y. By adapting to the data, this approach explicitly allowed for the contribution of LD and remarkably reduced the "missing heritability" of $\hat{V}_{\mathrm{r}}$ in REM.
We illustrated our theoretical results in simulation studies in Section 3.2 as well as on full genomic datasets in Section 3.3. We stated that the novel predictor $\widehat{W}_{\mathrm{r}}$ performs drastically superior to the estimator $\hat{V}_{\mathrm{r}}$ and performs at least as good as the estimator $\widehat{W}_{\mathrm{b}}$ used for the mean of the posterior genomic variance in BRM. We introduced an innovative indicator I_{r}, see (3.5), of the contribution of LD to the genomic variance by comparing the estimator $\hat{V}_{\mathrm{r}}$ and the predictor $\widehat{W}_{\mathrm{r}}$. This added to the conclusion that the improved performance of the novel predictor $\widehat{W}_{\mathrm{r}}$ compared to the estimator $\hat{V}_{\mathrm{r}}$ is caused by the inclusion of LD.

Discussion

The additive genetic variance and the narrow-sense heritability are clearly and uniquely defined in quantitative genetics, but nevertheless estimation procedures for the genomic variance give different results (Chen, 2016). The estimation of the genomic variance varies, especially in REM, even when using the same marker data to calculate different genomic relationship matrices (Legarra, 2015; Fernando et al., 2017). We substantiate that in the Appendix A.5, where we showed that transformations of the input marker-matrix $\mathbf{X}$ change the estimate of the genomic variance when using estimators (e.g. GCTA-GREML) based on the marginal genomic variance V_{r} defined in (2.21). This contradicts the genetic variance in Section 1.1 that is independent of the coding of the genotypes. In addition to that, Kumar et al. (2015, 2016) state that the GRM in GCTA-GREML is an estimate of the underlying data-generating process but is treated as a fixed quantity, which makes the calculation of the genomic variance as in (1.19) invalid.

Contrary to that, we built our analysis of genomic variances on the data-generating process of the marker data by considering X as a random vector in model (2.1). The resulting expression for the genomic variance in Chapter 2, $\beta^\top \Sigma_X \beta$, showed to be independent of the centering of the marker data, regardless of whether β is considered as a fixed population parameter or as a random variable (which makes the genomic variance a random variable).
This tackles yet another central point of critique on GCTA-GREML issued by Kumar et al. (2015, 2016), namely that the single marker effects are treated as independent random variables with equal variances. Conditionally on the phenotypic data, the random contribution of each marker is not independent any more, see (A.15).
Our approach of treating the marker content X as random and considering the genomic variance W in BRM and REM as a random variable conditional on the effect vector can be considered as an extension of the genomic variance from the FEM to high-dimensional datasets. Consequently, our approach constitutes the genomic equivalent of the genetic variance in high-dimensional datasets also, which is intrinsically tied to an explicit contribution of LD to the genomic variance.

In the theoretical expression of the genomic variance V_r, LD does not contribute. However, when using the REML algorithm to estimate the variance component σ^2_β, LD takes a part and consequently also influences estimators similar to GCTA-GREML. Nevertheless, as we have noticed in Figure 3.9, the bias of $\hat{V}_\mathrm{r}$ is still very much correlated with the contribution of LD.

The estimation and the prediction of the effects β in high-dimensional datasets using the BRM and the REM is executed by adapting to the data by means of its likelihood, which possibly results in an over-adjustment. As a consequence, estimating the posterior mean of the conditional variance in BRM and predicting the conditional genomic variance in REM bears the risk of over-adjustment to the data.
The simulation studies in Section 3.2 have been performed under a very simplistic model and excluded the influence of imperfect LD between the markers and the QTL, for instance. This removed one of the main sources of the "missing heritability" claimed in literature, see Yang et al. (2010). The stability of the novel predictor $\widehat{W}_\mathrm{r}$ as well as of the estimator of the posterior mean $\widehat{W}_\mathrm{b}$ has still to be further tested in more complex scenarios with different LD-structures between markers and with imperfect LD between markers and QTL.
The application to full genomic dataset in Section 3.3 proved that it is also possible that the estimator $\hat{V}_\mathrm{r}$ overestimates the genetic variance, namely in cases that the weighted sum of covariances between the marker genotypes becomes negative. The predictor $\widehat{W}_\mathrm{r}$ is designed to correct for the negative sum of weighted covariances in these scenarios likewise.

The nearly exact empirical decomposition of the phenotypic variance in the predictor $\widehat{W}_\mathrm{r}$ and the residual variance $\hat{\sigma}^2_\varepsilon$ in Table 3.2 indicates that some sort of

analytical decomposition similar to the theoretical variance counterpart in (2.3) is also possible for the random genomic variance W. This implies that there might be yet another derivation and interpretation of the random genomic variance.

The genomic variance is tied to a regression on markers, whereas the genetic variance is connected with QTL. Consequently, imperfect LD between the markers and the QTL causes possible under- or overestimation of the genetic variance. The simulation studies in Section 3.2 have shown that when the QTL are contained in the set of markers, the FEM as well as the BRM and the REM are capable of producing good results for the genomic variance, if the corresponding model-specific estimators and predictors are used.
The REM and the BRM are connected by the special case in which a normal prior with equal variances is chosen for the elements of the effect vector β, see Appendix A.2 and Appendix A.3. Bayesian methods are more flexible in that many different prior distributions for the effect vector can be chosen. Bayesian hierarchical models enable an assignment of probabilities to the variance components in the model, such that it is in principal possible to obtain arbitrary complex expressions for the genomic variance in BRM. In BRM we make inferences about the posterior distribution of the genomic variance W, see (2.16). It is also possible to use a more robust characteristic like the median instead of the mean of the posterior distribution. Both the BRM and the REM are connected to penalized regression models which are themselves members of the family of FEM's. Carving out theses connections possibly presents another interesting research object.

A. Linear Regression Models

A.1. Ordinary Least Squares (OLS)

In this section we give the most important analytical results for ordinary least-squares (OLS), which is particularly relevant for Section 2.1. Similar results can, for instance, be found in Hastie et al. (2008), Izenman (2008), or Wakefield (2013).

The OLS method is based on the realized (or conditional on X) model

$$y = \mu + \mathbf{X}\beta + \varepsilon := \mu + \left(\sum_{j=1}^{p} x_{ij}\beta_j \right)_{i=1,\dots,n} + \varepsilon, \tag{1.12}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{1}_{n\times n})$. Denote by $\tilde{\mathbf{X}}$ the $n \times (p+1)$ matrix $(1_n, \mathbf{X})$, where 1_n is the column-n-vector with every entry equal to 1. Then, the estimator that minimizes the residual sum of squares $\varepsilon^\top \varepsilon$ is called OLS estimator and is given by

$$\left(\hat{\mu}, \hat{\beta}^\top\right)^\top = \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^\top y.$$

For mean-centered data, i.e. $\sum_{i=1}^{n} x_{ij} = 0$ for all $j = 1, ..., p$, it holds that

$$\left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\right)^{-1} = \begin{pmatrix} n^{-1} & 0 \\ 0 & (\mathbf{X}^\top \mathbf{X})^{-1} \end{pmatrix}.$$

Consequently,

$$\left(\hat{\mu}, \hat{\beta}^\top\right)^\top = \left(\bar{y}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top y\right).$$

The OLS estimator $\hat{\beta}$ for β is distributed as

$$\hat{\beta} \sim \mathcal{N}\big(\beta, \sigma_\varepsilon^2 (\mathbf{X}^\top \mathbf{X})^{-1}\big)$$

and due to the Gauß-Markov Theorem it is the best linear unbiased estimator (BLUE) for β (linear function of y, unbiased, and smallest variance among all linear unbiased estimators).

An unbiased estimator $\hat{\sigma}_\varepsilon^2$ for the residual variance σ_ε^2 is given by

$$\hat{\sigma}_\varepsilon^2 := \frac{\left(y - \mathbf{X}\hat{\beta} - 1_n\hat{\mu}\right)^\top \left(y - \mathbf{X}\hat{\beta} - 1_n\hat{\mu}\right)}{n - p - 1},$$

which can also be expressed as

$$\begin{aligned}\hat{\sigma}_\varepsilon^2 &= \frac{1}{n-(p+1)}\Big[y^\top(1-\mathbf{H})y + n\hat{\mu}^2 - 2y^\top(1-\mathbf{H})1_n\hat{\mu}\Big] \\ &= \frac{1}{n-(p+1)}\Big[y^\top(1-\mathbf{H})y + n\hat{\mu}^2 - 2n\bar{y}\hat{\mu}\Big],\end{aligned} \tag{A.1}$$

where the hat-matrix $\mathbf{H}$ is defined as

$$\mathbf{H} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

It holds that

$$\mathbf{H}\hat{\mu} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top 1_n\hat{\mu} = 0$$

because of the column-wise mean-centering of $\mathbf{X}$.
Subsequently, an unbiased estimator $\hat{\Sigma}_{\hat{\beta}}$ for the variance of $\hat{\beta}$ in OLS is given by:

$$\hat{\Sigma}_{\hat{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\hat{\sigma}_\varepsilon^2. \tag{A.2}$$

A.2. BRM's with Markov Chain Monte Carlo

In this section we sketch some results from Bayesian regression models in combination with Markov Chain Monte Carlo (MCMC) estimates that are important for the calculation of the genomic variance in Section 2.2. We partly build on the *bioRxiv*-manuscript Schreck and Schlather (2018).

Full Bayesian regression models are based on the model (2.1) and include a distribution for X with parameter ψ such that there is a joint likelihood of the data $p(X, Y|\beta, \psi)$ combined with a prior distribution $p(\psi, \beta)$ for the parameters (Gelman et al., 2014). However, by assuming prior independence of the parameters determining $p(Y|X, \beta)$ and the parameters ψ determining $p(X|\psi)$ leads to the factorization $p(\psi, \beta) = p(\psi)p(\beta)$. Consequently, the full conditional posteriori $p(\psi, \beta|X, Y)$ can be expressed as the product $p(\psi|X)p(\beta|X, Y)$ such that the second factor can be analyzed by itself in a standard regression model without loss of information: $p(\beta|X, Y) \approx p(\beta)p(Y|X, \beta)$ (Gelman et al., 2014).
Model (1.12) can be considered as n draws of (Y, X). The effect vector β in

$$y = \mu + \mathbf{X}\beta + \varepsilon := \mu + \left(\sum_{j=1}^{p} x_{ij}\beta_j\right)_{i=1,\dots,n} + \varepsilon, \tag{1.12}$$

is assigned a prior distribution $p(\beta)$ with finite prior expectation $\mu_\beta := \mathbb{E}[\beta]$ and finite prior variance-covariance matrix $\Sigma_\beta := \mathrm{Cov}(\beta)$. We leave the form of the distribution $p(\beta)$ unspecified in this general approach.

After the assignment of prior distributions to parameters of interest in model (1.12), their posterior distribution is investigated, usually computationally with a

simulation algorithm such as MCMC (Gelman et al., 2014). As a result of the application, we obtain the Markov chain sequence of p-vectors $\left(\hat{\beta}^{(m)}\right)_{m=1,...,M}$ which are assumed to be draws from the posterior distribution $p(\beta|y)$ after discarding the burn-in iterations and after thinning the chain. We use the empirical mean

$$\hat{\mu}_{\beta|y} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)} \tag{A.3}$$

as an unbiased estimator for the posterior expectation $\mu_{\beta|y}$ and the empirical variance

$$\hat{\Sigma}_{\beta|y} = \frac{1}{M-1}\sum_{m=1}^{M}\hat{\beta}^{(m)}\left(\hat{\beta}^{(m)}\right)^{\top} - \frac{1}{M(M-1)}\sum_{k=1}^{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}\left(\hat{\beta}^{(k)}\right)^{\top} \tag{A.4}$$

as an estimator for the posterior covariance $\Sigma_{\beta|y}$. In order to calculate $\widehat{W}_{\mathrm{b}}$, see (2.18), we still need an empirical expression for the covariance $\Sigma_{\hat{\mu}_{\beta|y}}$ of the estimated effects.
It holds for all $k, m \in \{1, ..., M\}$, $k \neq m$, that

$$\mathrm{Cov}\left(\hat{\beta}^{(m)}, \hat{\beta}^{(k)}\right) \approx 0, \tag{A.5}$$

because we have thinned the MCMC sample in order to obtain an approximately independent chain. We find

$$\begin{aligned}
\Sigma_{\hat{\mu}_{\beta|y}} &:= \mathrm{Cov}\left(\hat{\mu}_{\beta|y}\right) \\
&\overset{(A.3)}{=} \mathrm{Cov}\left(\frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}, \frac{1}{M}\sum_{k=1}^{M}\hat{\beta}^{(k)}\right) \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{k=1}^{M}\mathrm{Cov}\left(\hat{\beta}^{(m)}, \hat{\beta}^{(k)}\right) \\
&= \frac{1}{M^2}\left[\sum_{m=1}^{M}\mathrm{Cov}\left(\hat{\beta}^{(m)}\right) + \sum_{m=1}^{M}\sum_{\substack{k=1 \\ k\neq m}}^{M}\mathrm{Cov}\left(\hat{\beta}^{(m)}, \hat{\beta}^{(k)}\right)\right] \\
&\overset{(A.5)}{\approx} \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{Cov}\left(\hat{\beta}^{(m)}\right) \\
&= \frac{1}{M}\Sigma_{\beta|y},
\end{aligned} \tag{A.6}$$

where the last equation is due to the fact that all samples $\hat{\beta}^{(m)}, m = 1, ..., M$, left in the chain are representative of the posterior distribution. Thus,

$$\mathrm{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\hat{\mu}_{\beta|y}}\right) \overset{(A.6)}{\approx} \frac{1}{M}\mathrm{tr}\left(\hat{\Sigma}_X\hat{\Sigma}_{\beta|y}\right), \tag{A.7}$$

where the approximation in (A.7) is more precise, the closer the chain is to independence.

Bayesian models offer a great flexibility in the choice of the prior distribution (Gianola et al., 2009). It is also possible to assign prior distributions to the variance component σ_ε^2 as well as hierarchically to parameters of already chosen prior distributions. However, the prior influence can overwhelm the data for $p \to \infty$ (Leon-Novelo and Casella, 2011), depending on the specification of the prior. Choosing the burn-in iterations as well as appropriate thinning of the chain is an important task (Givens and Hoeting, 2013). Maybe even more important, convergence analysis and diagnostics play a major role after the application of a simulation algorithm and before using the results (Congdon, 2006).

Bayesian Ridge Regression (BRR) results from the assumption of independent normal priors with identical variances σ_β^2 for the components of the effect vector:

$$p(\beta_j|\sigma_\beta^2) \sim \mathcal{N}(0, \sigma_\beta^2), \quad j = 1, ..., p, \tag{A.8}$$

together with inverse-gamma priors for the variance components

$$\sigma_\beta^2 \sim \text{IG}(a_\beta, b_\beta) \tag{A.9}$$

and

$$\sigma_\varepsilon^2 \sim \text{IG}(a_\varepsilon, b_\varepsilon). \tag{A.10}$$

A random variable Z follows the inverse-gamma distribution $\text{IG}(a, b)$ with shape $a > 0$ and scale $\beta > 0$, if $1/Z \sim \mathcal{G}(a, b)$.

The inference in BRM is always done on the posterior distributions, or on the full conditionals, respectively. For the special case of BRR, Kneib et al. (2011) derived:

$$\begin{aligned}
&\beta|\sigma_\beta^2, \sigma_\varepsilon^2 \sim \mathcal{N}(\mu_{\beta|y}, \Sigma_{\beta|y}) \\
&\quad \mu_{\beta|y} = \Sigma_{\beta|y} \frac{1}{\sigma_\varepsilon^2} \mathbf{X}^\top (y - \mu) \qquad \text{(A.11)} \\
&\quad \Sigma_{\beta|y} = \left(\frac{1}{\sigma_\varepsilon^2} \mathbf{X}^\top \mathbf{X} + \frac{1}{\sigma_\beta^2} \mathbb{1}_{p \times p} \right)^{-1} \qquad \text{(A.12)} \\
&\quad \sigma_\beta^2|\beta \sim \text{IG}\Big(a_\beta + 0.5p, b_\beta + 0.5\beta^\top \beta \Big) \\
&\quad \sigma_\varepsilon^2|\beta \sim \text{IG}\Big(a_\varepsilon + 0.5n, b_\varepsilon + 0.5(y - \mu - \mathbf{X}\beta)^\top (y - \mu - \mathbf{X}\beta) \Big).
\end{aligned}$$

Drawing samples iteratively from these full conditionals illustrates the processing of MCMC-algorithms.
We are going to see in Appendix A.3 that, conditionally on the variance components σ_β^2 and σ_ε^2, the BRR model is the Bayesian equivalent to the frequentist BLUP-method.

A.3. Best Linear Unbiased Prediction (BLUP)

In this section we give the most important analytical results for the BLUP method and connect it to the genomic BLUP (GBLUP) in the equivalent model. We mainly build on Henderson (1984) and Searle et al. (1992), and partly on the *bioRxiv*-manuscript Schreck and Schlather (2018).

In the model

$$y = \mu + \mathbf{X}\beta + \varepsilon := \mu + \Big(\sum_{j=1}^{p} x_{ij}\beta_j\Big)_{i=1,\dots,n} + \varepsilon, \tag{1.12}$$

it holds that $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{1}_{n\times n})$ independently of the effect p-vector β which is normally distributed with mean $\mu_\beta = 0$ and finite marginal variance-covariance matrix $\sigma_\beta^2 \mathbb{1}_{p\times p}$

$$\beta \sim \mathcal{N}(0, \sigma_\beta^2 \mathbb{1}_{p\times p}),$$

such that

$$y \sim \mathcal{N}\Big(\mu, \underbrace{\mathbf{X}\mathbf{X}^\top \sigma_\beta^2 + \sigma_\varepsilon^2 \mathbb{1}_{n\times n}}_{:=\tilde{\Sigma}^{-1}}\Big). \tag{A.13}$$

It is not possible to estimate β because it is a random variable. Henderson (1984) introduced the concept of the prediction of β, which refers to the estimation of the realized values of the random effects. Common approaches to find a BLUP for β are based on the mixed model equations (Henderson, 1984) or in general on maximizing the conditional likelihood of y.

The joint distribution of y and β equals

$$\begin{pmatrix} y \\ \beta \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \tilde{\Sigma}^{-1} & \sigma_\beta^2 \mathbf{X} \\ \sigma_\beta^2 \mathbf{X}^\top & \sigma_\beta^2 \mathbb{1}_{p\times p} \end{pmatrix} \right].$$

We obtain

$$\beta | y \sim \mathcal{N}\Big(\sigma_\beta^2 \mathbf{X}^\top \tilde{\Sigma}(y-\mu), \sigma_\beta^2 \mathbb{1}_{p\times p} - \sigma_\beta^2 \mathbf{X}^\top \tilde{\Sigma} \mathbf{X} \sigma_\beta^2\Big)$$

because of the joint normal distribution (Kotz et al., 2000). The BLUP $\mu_{\beta|y}$ for β is defined as $\mu_{\beta|y} := \mathbb{E}[\beta \,|\, y]$ (Searle et al., 1992) such that we obtain

$$\mu_{\beta|y} := \mathbb{E}[\beta \,|\, y] = \sigma_\beta^2 \mathbf{X}^\top \tilde{\Sigma}(y-\mu). \tag{A.14}$$

The unbiasedness of the predictor $\mu_{\beta|y}$ is seen very quickly, because

$$\mathbb{E}[\mu_{\beta|y}] = \mathbb{E}\Big[\mathbb{E}[\beta \,|\, y]\Big] = \mathbb{E}[\beta] = 0.$$

In addition to that, the predictor $\mu_{\beta|y}$ is the optimal predictor as a function in y under MSE considerations. Assume that $g(y)$ is a measurable function in y. Then,

$$\begin{aligned}\mathbb{E}\Big[\big(\beta - g(y)\big)^2 \,\Big|\, y\Big] =& \mathbb{E}\Big[\big(\beta - \mu_{\beta|y} + \mu_{\beta|y} - g(y)\big)^2 \,\Big|\, y\Big] \\ =& \mathbb{E}\Big[\big(\beta - \mu_{\beta|y}\big)^2 \,\Big|\, y\Big] + \mathbb{E}\Big[\big(\mu_{\beta|y} - g(y)\big)^2 \,\Big|\, y\Big] \\ &+ 2\big(\mu_{\beta|y} - g(y)\big)\mathbb{E}\Big[\big(\beta - \mu_{\beta|y}\big) \,\Big|\, y\Big] \\ =& \mathbb{E}\Big[\big(\beta - \mu_{\beta|y}\big)^2 \,\Big|\, y\Big] + \big(\mu_{\beta|y} - g(y)\big)^2.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}\Big[\big(\beta - g(y)\big)^2\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\big(\beta - g(y)\big)^2 \,\Big|\, y\Big]\Big] \\ &= \mathbb{E}\Big[\big(\beta - \mu_{\beta|y}\big)^2\Big] + \mathbb{E}\Big[\big(\mu_{\beta|y} - g(y)\big)^2\Big] \\ &\geq \mathbb{E}\Big[\big(\beta - \mu_{\beta|y}\big)^2\Big].\end{aligned}$$

The variance-covariance matrix $\Sigma_{\beta|y}$ of the conditional distribution of β equals

$$\Sigma_{\beta|y} := \mathrm{Cov}(\beta \,|\, y) = \sigma_\beta^2 \mathbb{1}_{p\times p} - \sigma_\beta^2 \mathbf{X}^\top \tilde{\Sigma} \mathbf{X} \sigma_\beta^2. \tag{A.15}$$

The Sherman-Morrison formula, or the more general Woodbury matrix identity (Henderson and Searle, 1981), for a nonsingular matrix A and matrices U, B, V with suitable dimensions equals

$$(A - UD^{-1}V)^{-1} = A^{-1} + A^{-1}U(D - VA^{-1}U)^{-1}VA^{-1}. \tag{A.16}$$

Setting $A = \frac{1}{\sigma_\beta^2}\mathbb{1}_{p\times p}$, $U = -\mathbf{X}^\top$, $D^{-1} = \frac{1}{\sigma_\varepsilon^2}\mathbb{1}_{n\times n}$ and $V = \mathbf{X}$ enables an equivalent expression of (A.15)

$$\Sigma_{\beta|y} = \left(\frac{1}{\sigma_\varepsilon^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\sigma_\beta^2}\mathbb{1}_{p\times p}\right)^{-1}, \tag{A.17}$$

which is the expression for the fully conditional posterior covariance matrix in (A.12) of β in BRR, see Appendix A.2. In addition to that, the full conditional posterior mean of β in BRR, see (A.11), can be reformulated using the equivalence of (A.15) and (A.17)

$$\begin{aligned}\mu_{\beta|y}^{\mathrm{BRR}} &\overset{(A.11)}{=} \Sigma_{\beta|y}\frac{1}{\sigma_\varepsilon^2}\mathbf{X}^\top(y-\mu) \\ &\overset{(A.15)}{=} \frac{1}{\sigma_\varepsilon^2}\Big[\sigma_\beta^2\mathbf{X}^\top - \sigma_\beta^2\mathbf{X}^\top\tilde{\Sigma}\mathbf{X}\mathbf{X}^\top\sigma_\beta^2\Big](y-\mu) \\ &= \sigma_\beta^2\mathbf{X}^\top\tilde{\Sigma}\left[\frac{1}{\sigma_\varepsilon^2}\tilde{\Sigma}^{-1} - \mathbf{X}\mathbf{X}^\top\frac{\sigma_\beta^2}{\sigma_\varepsilon^2}\right](y-\mu) \\ &\overset{(A.13)}{=} \sigma_\beta^2\mathbf{X}^\top\tilde{\Sigma}(y-\mu) \\ &\overset{(A.14)}{=} \mu_{\beta|y}^{\mathrm{BLUP}},\end{aligned}$$

such that we can state equivalence between the full conditional posterior mean of β in BRR and the predictor for β in BLUP.

The variance of the BLUP equals

$$\mathrm{Cov}(\mu_{\beta|y}) = \mathrm{Cov}(\beta) - \mathbb{E}\big[\mathrm{Cov}(\beta \,|\, y)\big] = \sigma_\beta^2 \mathbf{X}^\top \tilde{\Sigma} \mathbf{X} \sigma_\beta^2. \tag{A.18}$$

The variance components σ_ε^2 and σ_β^2 in model (1.12) in REM are usually estimated using restricted maximum likelihood (REML), and are in general less biased than maximum-likelihood estimates (Patterson and Thompson, 1971; Corbeil and Searle, 1976; Searle et al., 1992). For balanced data, REML estimates for the variance components equals the ANOVA estimates which are known to be unbiased (Patterson and Thompson, 1971; Searle et al., 1992).

After inserting estimators for the variance components σ_ε^2 and σ_β^2, the unbiasedness of $\hat{\mu}_{\beta|y} = \widehat{\mathbb{E}}[\beta \,|\, y] = \hat{\sigma}_\beta^2 \mathbf{X}^\top \hat{\tilde{\Sigma}}(y - \hat{\mu})$ can only be asserted conditionally on the estimated variance components:

$$\begin{aligned}
\mathbb{E}\Big[\hat{\mu}_{\beta|y} \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] &= \mathbb{E}\Big[\hat{\sigma}_\beta^2 \mathbf{X}^\top \hat{\tilde{\Sigma}}(y - \hat{\mu}) \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] \\
&= \hat{\sigma}_\beta^2 \mathbf{X}^\top \hat{\tilde{\Sigma}} \mathbb{E}\Big[y - \hat{\mu} \,\Big|\, \hat{\sigma}_\beta^2, \hat{\sigma}_\varepsilon^2\Big] \\
&= 0 = \mathbb{E}[\beta].
\end{aligned} \tag{A.19}$$

In the equivalent model

$$y = \mu + g + \varepsilon \tag{1.10}$$

it holds that

$$g \sim \mathcal{N}(0, \sigma_g^2 \mathbf{G})$$

with

$$\sigma_\beta^2 \mathbf{X}\mathbf{X}^\top = \frac{1}{p} \mathbf{X}\mathbf{X}^\top (p\sigma_\beta^2) =: \mathbf{G}\sigma_g^2, \tag{1.18}$$

where $\sigma_{\mathrm{g}}^2 := p\sigma_\beta^2$ and $\mathbf{G} := \frac{1}{p}\mathbf{X}\mathbf{X}^\top$. Consequently,

$$g \overset{\mathrm{d}}{=} \mathbf{X}\beta,$$

such that the genomic best linear unbiased predictor (GBLUP) for g equals

$$\mu_{g|y} := \mathbb{E}[g \,|\, y] = \mathbb{E}[\mathbf{X}\beta \,|\, y] = \mathbf{X}\mu_{\beta|y} \overset{(A.14),(1.18)}{=} \sigma_g^2 \mathbf{G}(\mathbf{G}\sigma_g^2 + \sigma_\varepsilon^2 \mathbb{1}_{n\times n})^{-1}(y - \mu). \tag{A.20}$$

The conditional variance-covariance matrix of g is obtained as

$$\begin{aligned}\Sigma_{g|y} &:= \mathrm{Cov}(g \,|\, y) \\ &= \mathbf{X}\mathrm{Cov}(\mu_{\beta|y})\mathbf{X}^\top \\ &= \mathbf{X}\Sigma_{\mu_{\beta|y}}\mathbf{X}^\top \\ &\overset{(1.18),(A.15)}{=} \sigma_g^2\mathbf{G} - \sigma_g^2\mathbf{G}(\mathbf{G}\sigma_g^2 + \sigma_\varepsilon^2 \mathbb{1}_{n\times n})^{-1}\mathbf{G}\sigma_g^2. \end{aligned} \tag{A.21}$$

This points out the relationship between the conditional moments of the effects in the BLUP-model and the GBLUP-model.

A.4. Mixed-Effect Model (MEM)

This section is fully based on the *bioRxiv*-manuscript Schreck and Schlather (2018).

Up-to-now we have considered random effect models only. We extend model (2.1) by including a fixed effect Zf which results in a mixed effect model (MEM) of the form

$$Y = Zf + X\beta + \varepsilon,$$

where f is a k-vector of fixed effects as in section 2.1, β is a p-vector of random effects as in section 2.3, Z is a random k row-vector and X is a random p-row-vector. We assume that Zf and ε as well as $X\beta$ and ε are independent. We calculate

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}(Zf + X\beta + \varepsilon) \\ &= \mathrm{Var}(Zf) + \mathrm{Var}(X\beta) + 2\mathrm{Cov}(Zf, X\beta) + \sigma_\varepsilon^2.\end{aligned}$$

Inferences on the additive genomic variance of the fixed effect Zf can be done as in Section 2.1 and inferences on the additive genomic variance of the random effect $X\beta$ can be done as in Section 2.3. If one is interested in the contribution of LD between fixed effects and random effects, e.g. when including single important markers as fixed effects in the MEM, we propose to predict the random conditional covariance

$$\mathrm{Cov}(Zf, X\beta \,|\, \beta) = f^\top \mathrm{Cov}(Z, X)\beta, \tag{A.22}$$

similar to the random genomic variance W in (2.16). We introduce

$$\hat{f}\hat{\Sigma}_{ZX}\hat{\mu}_{\beta|y} \tag{A.23}$$

as a predictor for (A.22), where

$$\hat{f} = (\mathbf{Z}^\top \hat{\hat{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top \hat{\hat{\Sigma}} y \tag{A.24}$$

is the BLUE of f with conditional covariance

$$\begin{aligned}\mathrm{Cov}\Big(\hat{f}\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big) &= \mathrm{Cov}\Big((\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\tilde{\Sigma}}y\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big)\\ &= (\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathrm{Cov}\Big(y\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big)\hat{\tilde{\Sigma}}\mathbf{Z}(\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{Z})^{-1}\\ &= (\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{Z})^{-1}. \qquad \text{(A.25)}\end{aligned}$$

We calculate:

$$\begin{aligned}\mathbb{E}\Big[\hat{f}\hat{\Sigma}_{ZX}\hat{\mu}_{\beta|y}\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big] &\overset{(2.7)}{=} \mathrm{tr}\Big(\Sigma_{XZ}\mathrm{Cov}\Big(\hat{f},\hat{\mu}_{\beta|y}\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big)\Big)\\ &\quad+\sum_{i=1}^{k}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{ZX},\hat{f}_i(\hat{\mu}_{\beta|y})_j\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big)\\ &= \sum_{i=1}^{k}\sum_{j=1}^{p}\mathrm{Cov}\Big(\hat{\sigma}_{ij}^{ZX},\hat{f}_i(\hat{\mu}_{\beta|y})_j\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big),\end{aligned}$$

because

$$\begin{aligned}\mathrm{Cov}\Big(\hat{f},\hat{\mu}_{\beta|y}\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big) &= \mathrm{Cov}\Big(\hat{f},\hat{\sigma}_\beta^2\mathbf{X}^\top\hat{\tilde{\Sigma}}y\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big)\\ &\quad-\mathrm{Cov}\Big(\hat{f},\hat{\sigma}_\beta^2\mathbf{X}^\top\hat{\tilde{\Sigma}}\mathbf{Z}\hat{f}\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big)\\ &\overset{(A.24)}{=}\mathrm{Cov}\Big((\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\tilde{\Sigma}}y,\hat{\sigma}_\beta^2\mathbf{X}^\top\hat{\tilde{\Sigma}}y\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big)\\ &\quad-\hat{\sigma}_\beta^2\mathrm{Cov}\Big(\hat{f}\,\Big|\,\hat{\sigma}_\beta^2,\hat{\sigma}_\varepsilon^2\Big)\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{X}\\ &\overset{(A.25)}{=}\hat{\sigma}_\beta^2(\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\tilde{\Sigma}}\hat{\tilde{\Sigma}}^{-1}\hat{\tilde{\Sigma}}\mathbf{X}\\ &\quad-\hat{\sigma}_\beta^2(\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\tilde{\Sigma}}\mathbf{X}\\ &= 0.\end{aligned}$$

The predictor in (A.23) is nearly unbiased for (A.22) given unbiased estimators $\hat{\sigma}_\beta^2$ and $\hat{\sigma}_\varepsilon^2$, because the random covariance in (A.22) has expectation 0.

A.5. Notes on the Mean-centering of X

This section is fully based on the *bioRxiv*-manuscript Schreck and Schlather (2018).

In model (2.1) in Chapter 2 we consider X to be a random row vector with expectation 0. If we depart from that assumption and consider $\tilde{X}$ with $\mathbb{E}[\tilde{X}] \neq 0$ and $\mathrm{Cov}(\tilde{X}) = \Sigma_X$ instead of X, we reformulate model (2.1) based on $\tilde{X}$ as

$$\begin{aligned}Y = \mu + \tilde{X}\beta + \varepsilon &= \mu + \big(\tilde{X} - \mathbb{E}[\tilde{X}]\big)\beta + \mathbb{E}[\tilde{X}]\beta + \varepsilon\\ &\overset{\mathrm{d}}{=} \mu + X\beta + \mathbb{E}[\tilde{X}]\beta + \varepsilon.\end{aligned}$$

In the FEM (β deterministic) the fixed term $\mathbb{E}[\tilde{X}]\beta$ is absorbed by the intercept such that

$$Y = \tilde{\mu} + X\beta + \varepsilon,$$

with $\tilde{\mu} = \mu + \mathbb{E}[\tilde{X}]\beta$. We obtain the linear model (2.1) with mean-centered data but different (fixed) intercept. Consequently, the genomic variance V_{f} in the FEM, see (2.4), is unchanged whether we consider mean-centered allele content X or not ($\tilde{X}$):

$$\mathrm{Var}(X\beta) = \beta^\top \Sigma_X \beta = \mathrm{Var}(\tilde{X}\beta).$$

In BRM and REM, where $\beta \sim (\mu_\beta, \Sigma_\beta)$ is a random variable, the term $\mathbb{E}[\tilde{X}]\beta$ is a random variable itself and is absorbed by the residual instead of the intercept

$$\begin{aligned} Y &= \mu + \tilde{X}\beta + \varepsilon \\ &= \mu + \big(\tilde{X} - \mathbb{E}[\tilde{X}]\big)\beta + \mathbb{E}[\tilde{X}]\beta + \varepsilon \\ &\overset{\mathrm{d}}{=} \mu + X\beta + \tilde{\varepsilon}, \end{aligned}$$

where $\tilde{\varepsilon} \sim (0, \sigma_\varepsilon^2 + \mathbb{E}[X]\Sigma_\beta\mathbb{E}[X]^\top)$. However, β is no longer independent of $\tilde{\varepsilon}$. For the genomic variance V, see (2.2), in BRM and REM it makes a difference whether we consider the mean-centered X or $\tilde{X}$ because:

$$\begin{aligned} \mathrm{Var}\big(\tilde{X}\beta\big) &= \mathrm{Var}_\beta\Big(\mathbb{E}\big[\tilde{X}\beta \,|\, \beta\big]\Big) + \mathbb{E}_\beta\Big[\mathrm{Var}\big(\tilde{X}\beta \,|\, \beta\big)\Big] \\ &= \mathrm{Var}_\beta\Big(\mathbb{E}\big[\tilde{X}\big]\beta\Big) + \mathbb{E}_\beta\Big[\beta^\top\Sigma_X\beta\Big] \\ &= \mathbb{E}\Big[\tilde{X}\Big]\Sigma_\beta\mathbb{E}\Big[\tilde{X}\Big]^\top + \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbb{E}\Big[\sigma_{ij}^X\beta_i\beta_j\Big] \\ &= \mathbb{E}\big[\tilde{X}\big]\Sigma_\beta\mathbb{E}\big[\tilde{X}\big]^\top + \mathrm{tr}(\Sigma_X\Sigma_\beta) + \mu_\beta^\top\Sigma_X\mu_\beta \\ &\neq \mathrm{tr}(\Sigma_X\Sigma_\beta) + \mu_\beta^\top\Sigma_X\mu_\beta = \mathrm{Var}(X\beta). \end{aligned}$$

This is consistent with the approach in Section 1.3 based on the realized model (1.12) where the genomic variance in REM is estimated based on $\mathbf{X}$

$$\mathrm{Cov}(\mathbf{X}\beta) = \mathbf{X}\mathbf{X}^\top\sigma_\beta^2 \tag{1.19}$$

or when using GRM's in the equivalent model (1.10)

$$\sigma_\beta^2\mathbf{X}\mathbf{X}^\top = \frac{1}{p}\mathbf{X}\mathbf{X}^\top(p\sigma_\beta^2) =: \mathbf{G}\sigma_g^2. \tag{1.18}$$

The genomic variance in these models clearly depends on whether using mean-centered matrices $\mathbf{X}$ or not, especially in the equivalent model transformations of $\mathbf{X}$ change the variance-covariance matrix of g. The GRM's are generally based on mean-centered matrices which is the reason why we have based

the main analysis of this thesis on the mean-centered approach.
The random genomic variance W, defined in (2.16), however, does not depend on centering

$$W := \mathrm{Var}(X\beta \,|\, \beta) = \beta^\top \Sigma_X \beta = \mathrm{tr}(\Sigma_X \beta\beta^\top) = \mathrm{Var}(\tilde{X}\beta \,|\, \beta), \tag{2.16}$$

and is therefore consistent with the genetic variance in Section 1.1 with respect to the independence to the coding of the QTL genotypes, see for instance (1.4).

A scaling of X by the p-vector b can, similarly to Section 1.1, be absorbed by the effect vector β, which then has to be redefined as β/b.

B. Figures

B.1. Variation of Observations

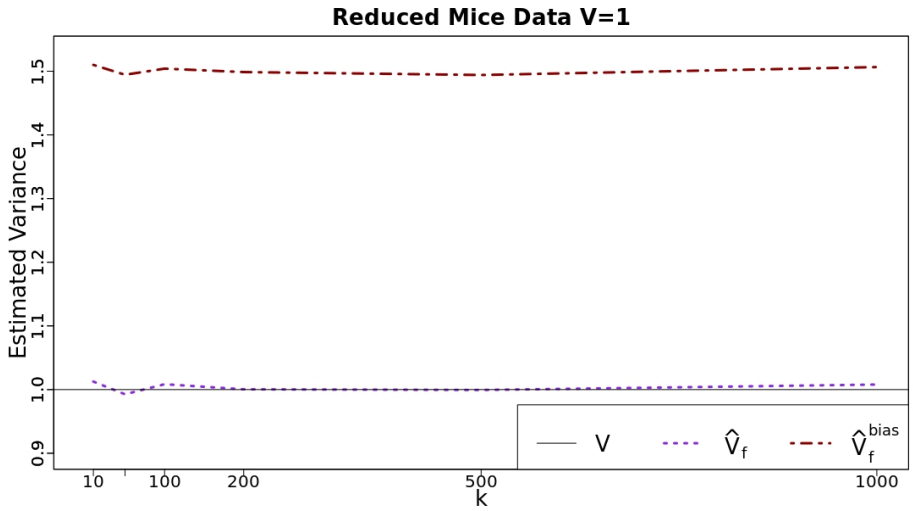

Figure B.1.: Estimated variance in the FEM (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The estimator $\hat{V}_\mathrm{f}$ performs remarkably better than the biased estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$ and is very close to V independently of the QTL-to-marker ratio.

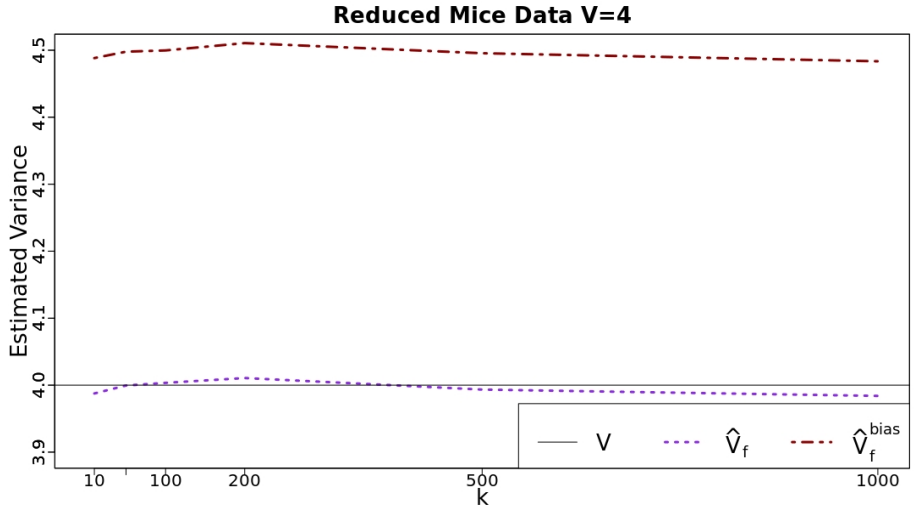

Figure B.2.: Estimated variance in the FEM (mean value over iterations of subsets of individuals) in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The estimator $\hat{V}_\mathrm{f}$ performs remarkably better than the biased estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$ and is very close to V independently of the QTL-to-marker ratio.

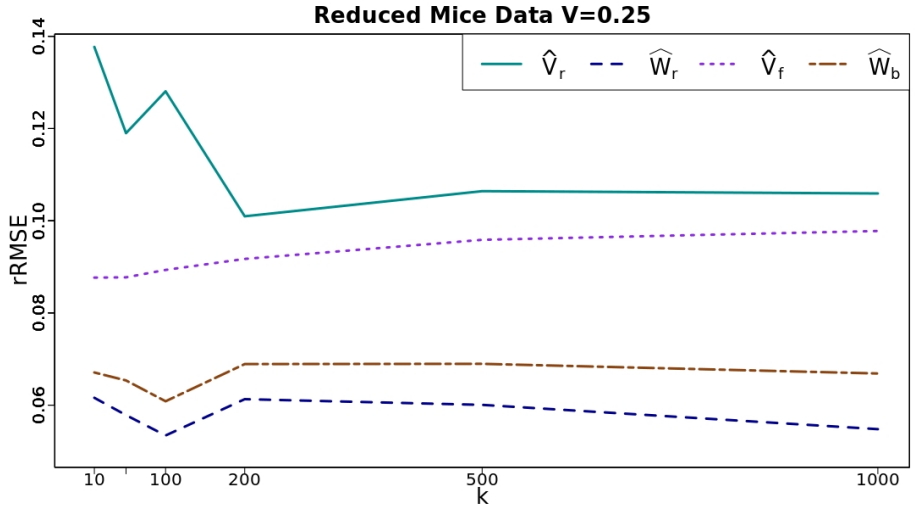

Figure B.3.: Relative root-mean-squared-error over iterations of subsets of individuals in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The predictor $\widehat{W}_\mathrm{r}$ in the REM and the estimator $\widehat{W}_\mathrm{b}$ in the BRM perform best with an rRMSE of around 6%. The rRMSE of the estimator $\hat{V}_\mathrm{f}$ from FEM is about 3% larger and the rRMSE of the estimator $\hat{V}_\mathrm{r}$ in the REM is largest and larger than the rRMSE of $\widehat{W}_\mathrm{r}$ by up to the factor 2.

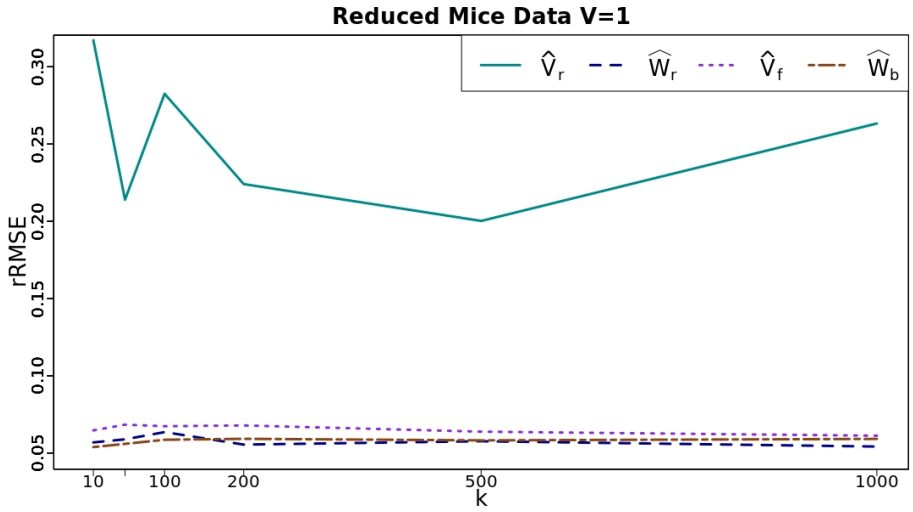

Figure B.4.: Relative root-mean-squared-error over iterations of subsets of individuals in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The predictor $\widehat{W}_\mathrm{r}$ from the REM, the estimator $\widehat{W}_\mathrm{b}$ from the BRM and the estimator $\hat{V}_\mathrm{f}$ from FEM perform best whereas $\hat{V}_\mathrm{r}$ from the REM is largest and larger than the rRMSE of $\widehat{W}_\mathrm{r}$ approximately by a factor of 4.

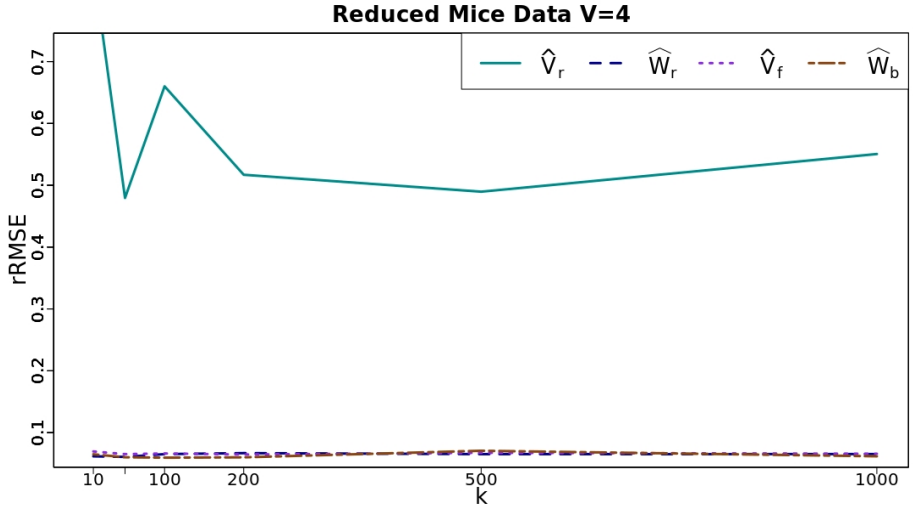

Figure B.5.: Relative root-mean-squared-error over iterations of subsets of individuals in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The predictor $\widehat{W}_\mathrm{r}$ from the REM, the estimator $\widehat{W}_\mathrm{b}$ from the BRM and the estimator $\hat{V}_\mathrm{f}$ from FEM perform best whereas $\hat{V}_\mathrm{r}$ from the REM is drastically larger.

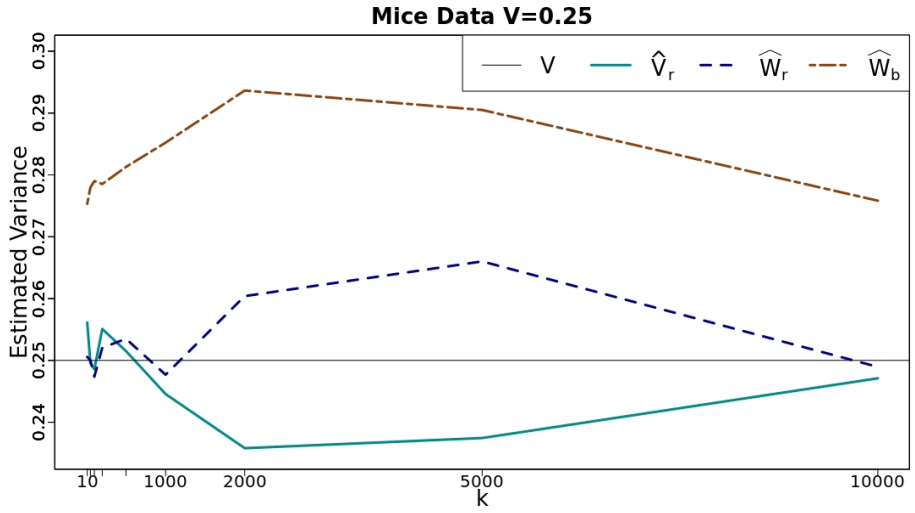

Figure B.6.: Estimated variance (mean value over iterations of subsets of individuals) in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The predictor $\widehat{W}_r$ from the REM overestimates V approximately to the same extend that the estimator $\hat{V}_r$ underestimates V. The estimator $\widehat{W}_b$ from BRM performs worst and overestimates V by at least 12%.

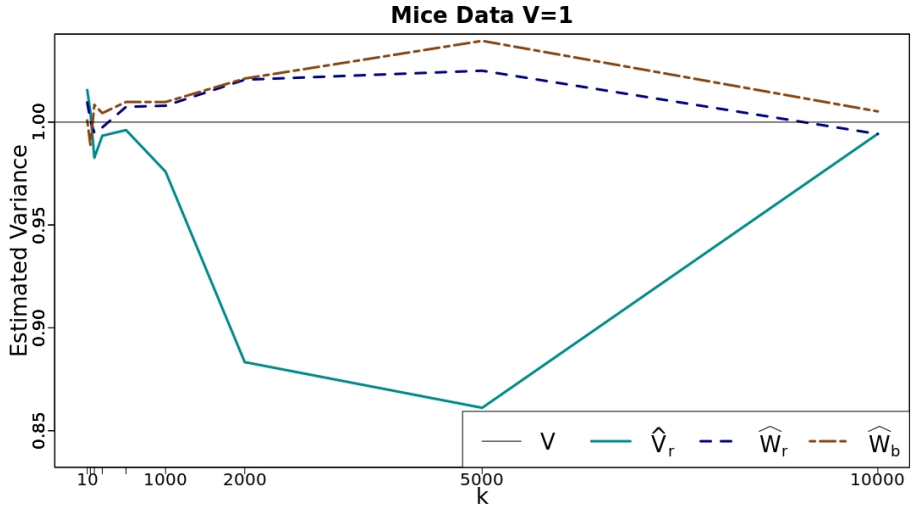

Figure B.7.: Estimated variance (mean value over iterations of subsets of individuals) in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The predictor $\widehat{W}_r$ from the REM performs best but slightly overestimates V. The estimator $\hat{V}_r$ from the REM underestimates V by a large margin and the bias of the estimator is worst for a medium QTL-to-marker ratio. The estimator $\widehat{W}_b$ from the BRM constantly overestimates V but performs very similar to the predictor $\widehat{W}_r$ from REM.

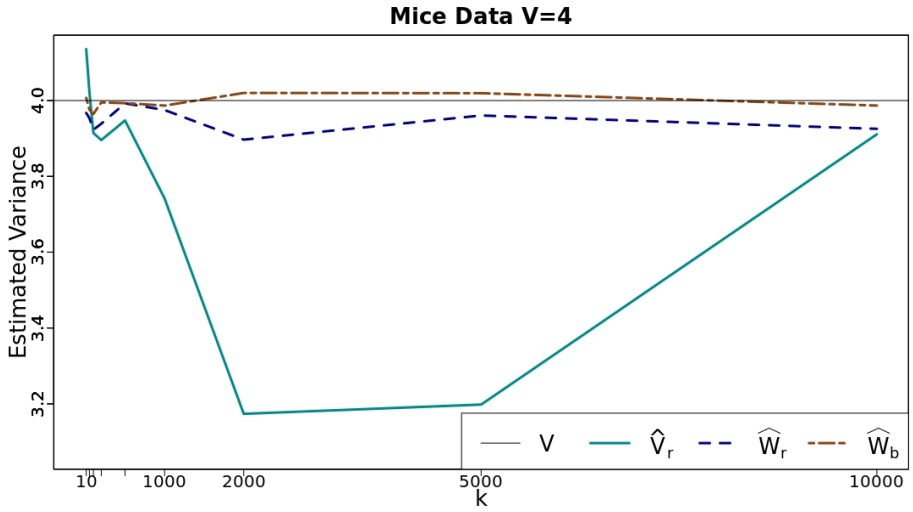

Figure B.8.: Estimated variance (mean value over iterations of subsets of individuals) in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The estimator $\widehat{W}_\mathrm{b}$ from the BRM performs best and is very close to the "true" V. The predictor $\widehat{W}_\mathrm{r}$ from the REM slightly underestimates V whereas the estimator $\hat{V}_\mathrm{r}$ from the REM drastically underestimates V and the bias of the estimator is worst for a medium QTL-to-marker ratio.

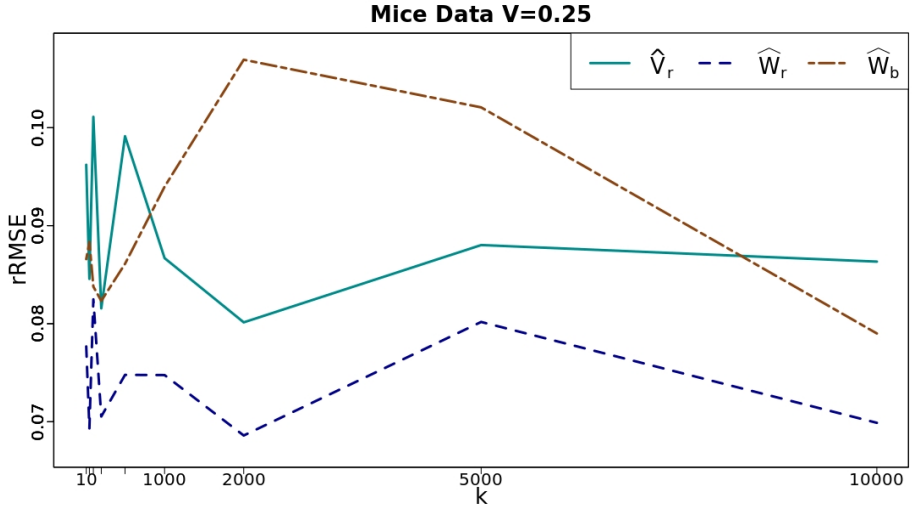

Figure B.9.: Relative root-mean-squared-error over iterations of subsets of individuals in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The predictor $\widehat{W}_\mathrm{r}$ from the REM performs best with an rRMSE of 7%-8% followed by the estimator $\hat{V}_\mathrm{r}$ with an rRMSE of 8%-10%. The estimator $\widehat{W}_\mathrm{b}$ perform worst for medium QTL-to-marker ratio.

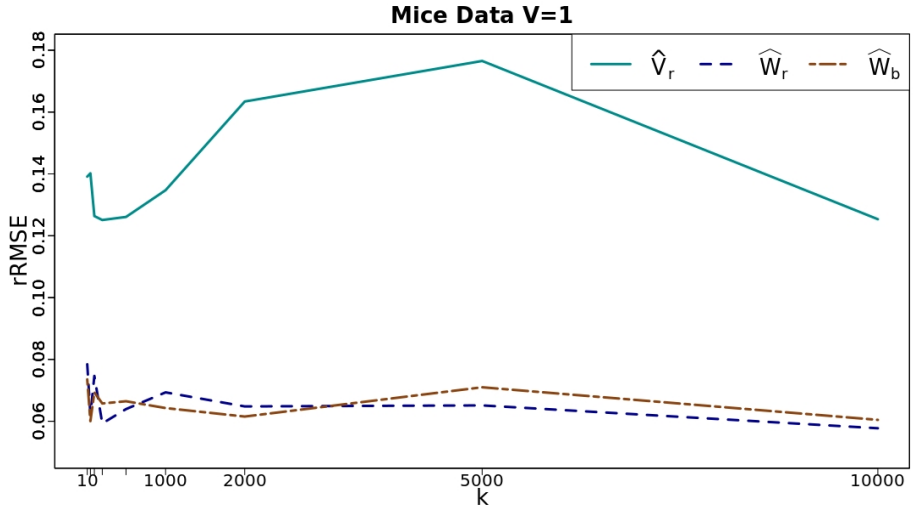

Figure B.10.: Relative root-mean-squared-error over iterations of subsets of individuals in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The predictor $\widehat{W}_\mathrm{r}$ from the REM and the estimator $\widehat{W}_\mathrm{b}$ from the BRM perform best with an rRMSE of around 6% whereas the rRMSE of $\hat{V}_\mathrm{r}$ is severely larger by up to a factor of 3.

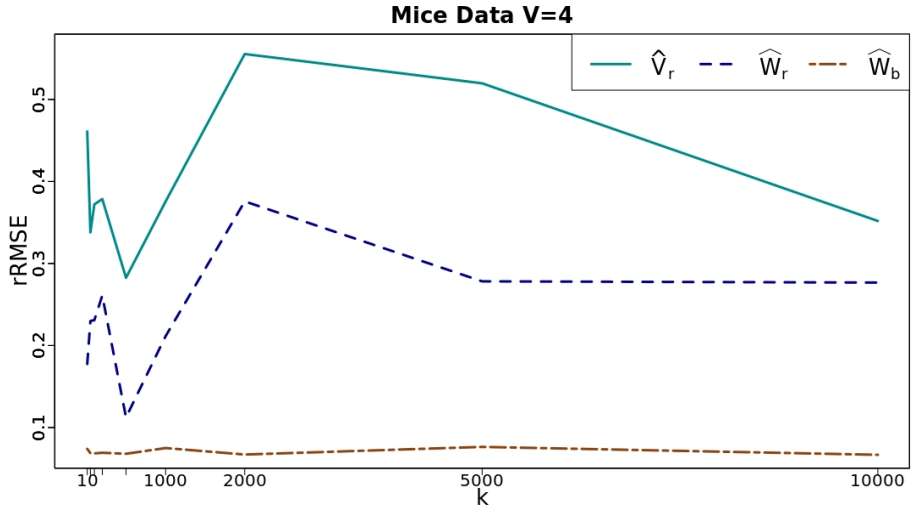

Figure B.11.: Relative root-mean-squared-error over iterations of subsets of individuals in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The estimator $\widehat{W}_\mathrm{b}$ from the BRM perform best with an rRMSE of around 6%-8% whereas the predictor $\widehat{W}_\mathrm{r}$ performs worse but still better than the estimator $\hat{V}_\mathrm{r}$.

B.2. Variation of QTL-Allocations

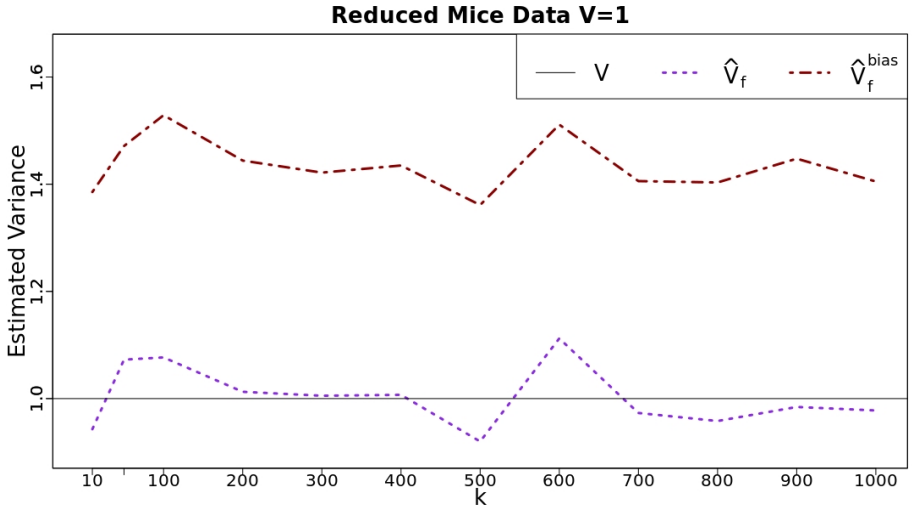

Figure B.12.: Estimated variance in FEM (mean value over different QTL allocations) in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The estimator $\hat{V}_\mathrm{f}$ performs remarkably better than the biased estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$ and is very close to V independently of the QTL-to-marker ratio.

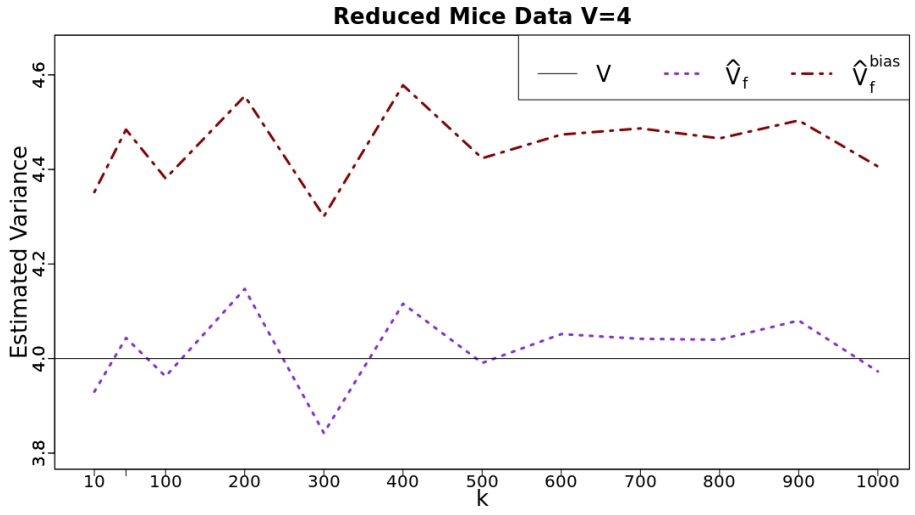

Figure B.13.: Estimated variance in FEM (mean value over different QTL allocations) in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The estimator $\hat{V}_\mathrm{f}$ performs remarkably better than the biased estimator $\hat{V}_\mathrm{f}^\mathrm{bias}$ and is very close to V independently of the QTL-to-marker ratio.

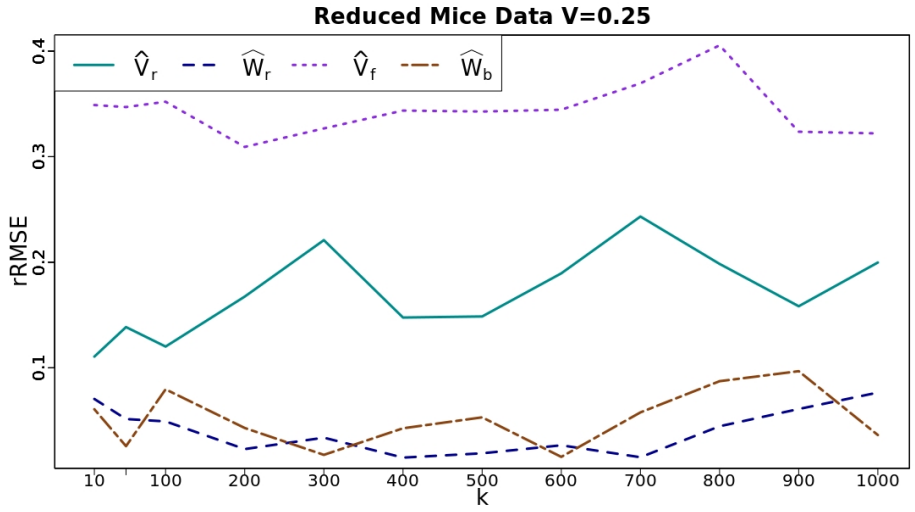

Figure B.14.: Relative root-mean-squared-error over different QTL allocations in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The "true" genomic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The predictor $\widehat{W}_\mathrm{r}$ for the conditional genomic variance in REM and the estimator for the posterior genomic variance $\widehat{W}_\mathrm{b}$ perform best. The rRMSE of the estimator $\hat{V}_\mathrm{r}$ for the marginal genomic variance is larger than the rRMSE of $\widehat{W}_\mathrm{r}$ by more than the factor 2. The rRMSE of the estimator $\hat{V}_\mathrm{f}$ from FEM is largest with an average value of about 35%.

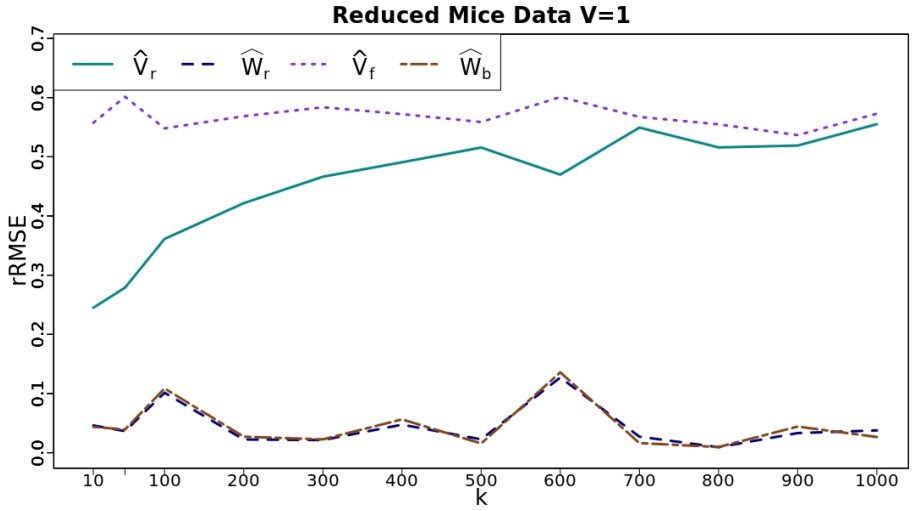

Figure B.15.: Relative root-mean-squared-error over different QTL allocations in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The "true" genomic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The predictor $\widehat{W}_\mathrm{r}$ for the conditional genomic variance in REM and the estimator for the posterior genomic variance $\widehat{W}_\mathrm{b}$ perform best. The rRMSE of the estimator $\hat{V}_\mathrm{r}$ for the marginal genomic variance is significantly larger than the rRMSE of $\widehat{W}_\mathrm{r}$. The rRMSE of the estimator $\hat{V}_\mathrm{f}$ from FEM is largest with an average value of about 60%.

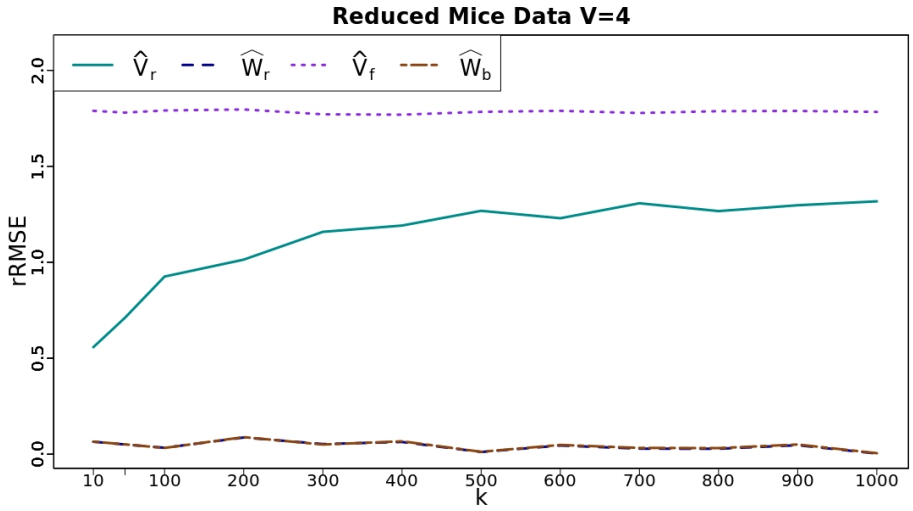

Figure B.16.: Relative root-mean-squared-error over different QTL allocations in the reduced mice dataset for different number of QTL k and fixed number of markers $\tilde{p} = 1088$. The "true" genomic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The predictor $\widehat{W}_\mathrm{r}$ for the conditional genomic variance in REM and the estimator for the posterior genomic variance $\widehat{W}_\mathrm{b}$ perform best. The rRMSE of the estimator $\hat{V}_\mathrm{r}$ for the marginal genomic variance is drastically larger than the rRMSE of $\widehat{W}_\mathrm{r}$. The rRMSE of the estimator $\hat{V}_\mathrm{f}$ from FEM is largest by a large amount.

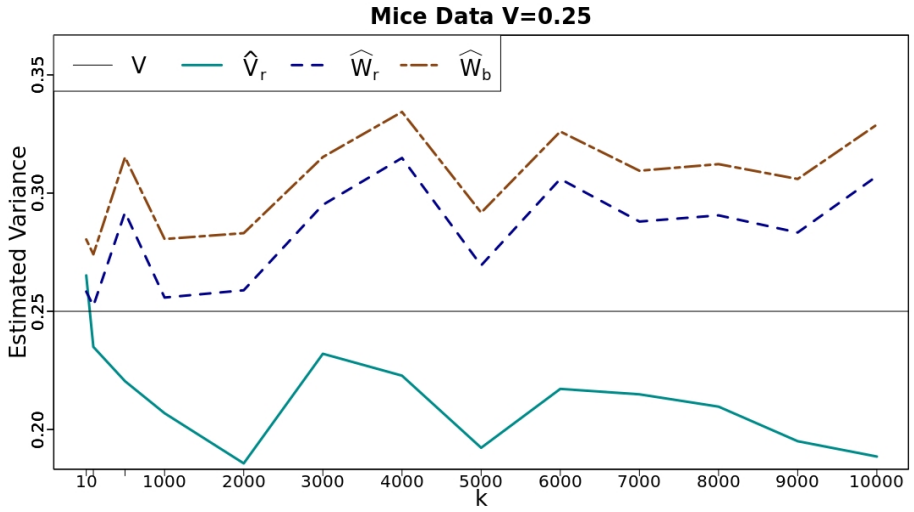

Figure B.17.: Estimated variance (mean value over different QTL allocations) in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The predictor $\widehat{W}_\mathrm{r}$ from the REM and the estimator $\widehat{W}_\mathrm{b}$ from the BRM overestimate V whereas the estimator $\hat{V}_\mathrm{r}$ from REM underestimates V by approximately the same extend.

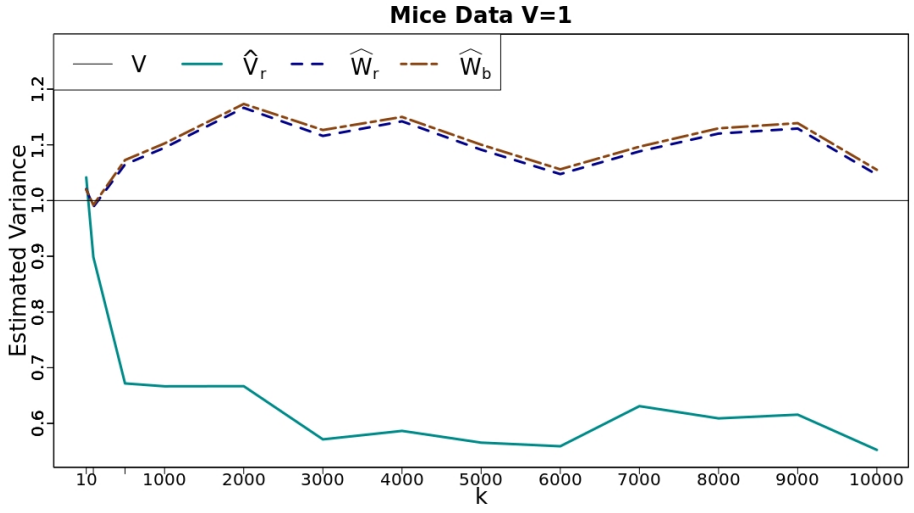

Figure B.18.: Estimated variance (mean value over different QTL allocations) in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The predictor $\widehat{W}_\mathrm{r}$ from the REM and the estimator $\widehat{W}_\mathrm{b}$ from the BRM perform best but overestimate V by about 10%. They perform remarkably better than the estimator $\hat{V}_\mathrm{r}$ from the REM which drastically underestimates V by over 30%-40%.

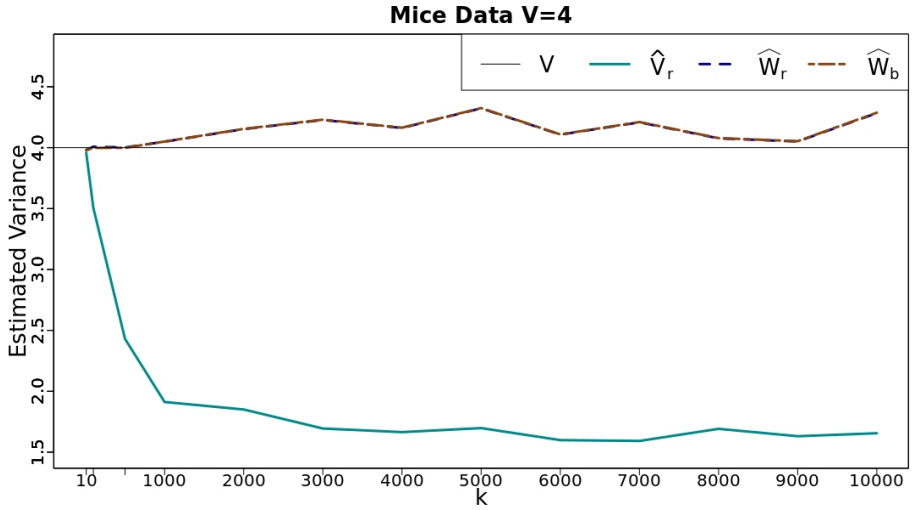

Figure B.19.: Estimated variance (mean value over different QTL allocations) in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The predictor $\widehat{W}_\mathrm{r}$ from the REM and the estimator $\widehat{W}_\mathrm{b}$ from the BRM perform best and are both very close to the "true" V although they slightly overestimate it. They perform remarkably better than the estimator $\hat{V}_\mathrm{r}$ from REM which drastically underestimates V by over 50%. This constitutes a striking example for the missing heritability of $\hat{V}_\mathrm{r}$.

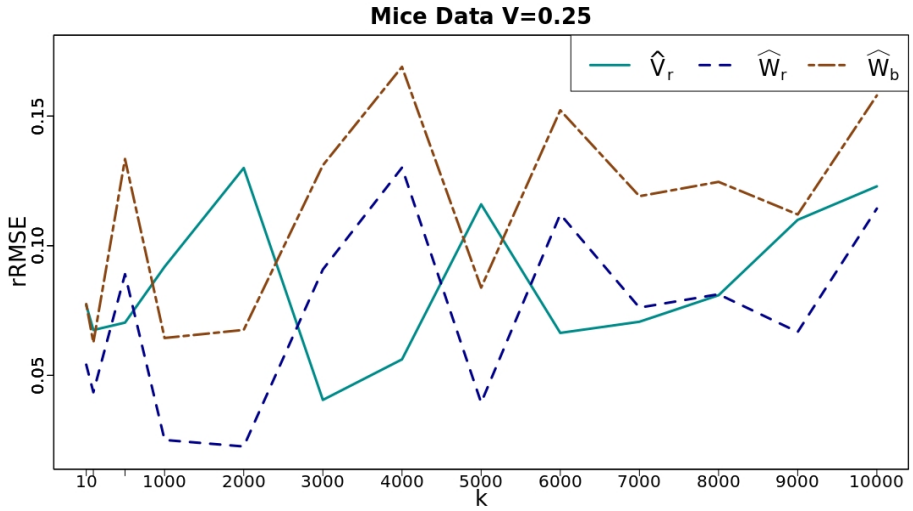

Figure B.20.: Relative root-mean-squared-error over different QTL allocations in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 0.25 for all k which resembles a heritability h^2 of 0.2. The rRMSE of the predictor $\widehat{W}_\mathrm{r}$, the estimator $\widehat{W}_\mathrm{b}$ and the estimator $\hat{V}_\mathrm{r}$ fluctuate heavily in k, where the spikes and slabs of $\widehat{W}_\mathrm{r}$ and $\widehat{W}_\mathrm{b}$ are in accordance.

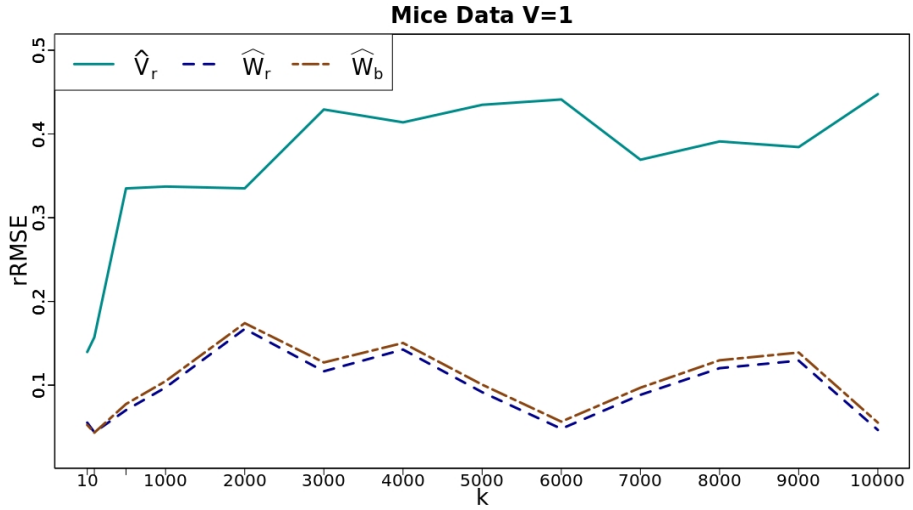

Figure B.21.: Relative root-mean-squared-error over different QTL allocations in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 1 for all k which resembles a heritability h^2 of 0.5. The predictor $\widehat{W}_r$ for the conditional genomic variance in REM and the estimator for the posterior genomic variance $\widehat{W}_b$ perform best whereas the rRMSE of $\hat{V}_r$ increases in k in the beginning and is larger than the rRMSE of $\widehat{W}_r$.

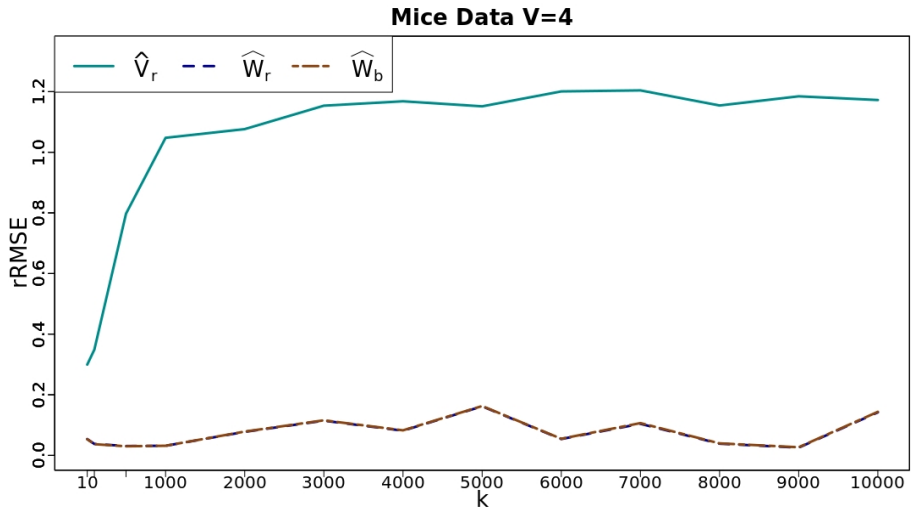

Figure B.22.: Relative root-mean-squared-error over different QTL allocations in the mice dataset for different number of QTL k and fixed number of markers $p = 10346$. The genetic variance V equals 4 for all k which resembles a heritability h^2 of 0.8. The predictor $\widehat{W}_r$ for the conditional genomic variance in REM and the estimator for the posterior genomic variance $\widehat{W}_b$ perform best whereas the rRMSE of $\hat{V}_r$ increases in k in the beginning and is larger than the rRMSE of $\widehat{W}_r$ by an average factor of more than 10.

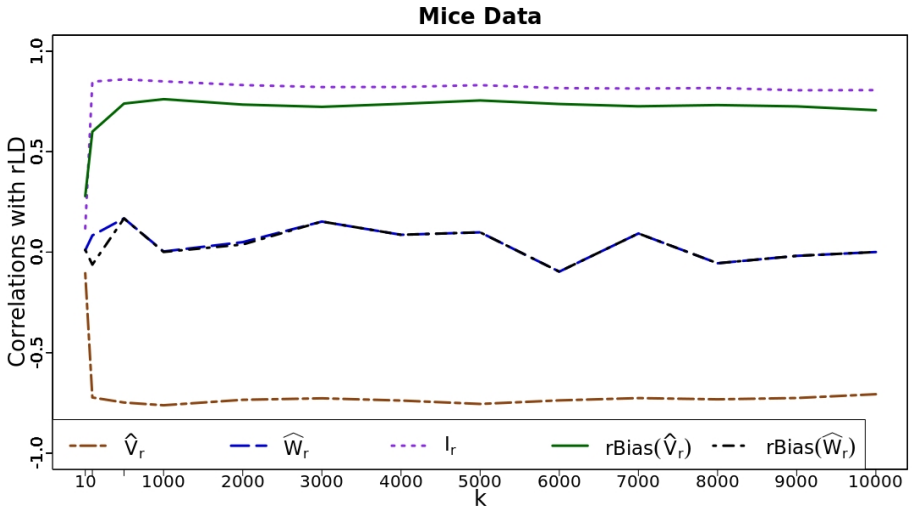

Figure B.23.: Empirical correlations with the relative contribution of LD to the genetic variance in the mice dataset for different number of QTL's k for fixed number of markers $p = 10346$. Values are averaged over the different levels of $h^2 \in \{0.2, 0.5, 0.8\}$. The correlation of the estimator $\hat{V}_r$ with the relative contribution of LD is about -0.7 except for $k = 10$, whereas the correlation of the predictor $\widehat{W}_r$ fluctuates around 0. The correlation of the relative bias of $\hat{V}_r$ is about 0.7 except for $k = 10$ which indicates that the larger the contribution of LD to the genomic variance, the larger the bias of $\hat{V}_r$ becomes. Contrary to that, the bias of the predictor $\widehat{W}_r$ is approximately uncorrelated to the relative contribution of LD. The quantity I_r is positively correlated (0.9) to the relative contribution of LD which makes it an usable indicator for the relative contribution of LD to the genomic variance.

References

Lexikon der Biologie. Heidelberg: Spektrum Akad Verlag, 1999. URL `https://www.spektrum.de/lexikon/biologie/`.

H. Becker. *Pflanzenzüchtung*. Ulmer, 2nd edition, 2011.

P. J. Bickel, J. B. Brown, H. Huang, and Q. Li. An overview of recent developments in genomics and the statistical methods that bear on them. Technical report, University of Berkeley, 2009.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer Series in Statistics, 2011.

M. G. Bulmer. The effect of selection on genetic variability. *American Naturalist*, 105:201–211, 1971.

G. B. Chen. On the reconciliation of missing heritability for genome-wide association studies. *European Journal of Human Genetics*, 24:1810–1816, 2016.

P. Congdon. *Bayesian Statistical Modelling*. John Wiley & Sons, 2nd edition, 2006.

R. R. Corbeil and S. R. Searle. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Techometrics*, 18(1):31–38, 1976.

G. Covarrubias-Pazaran. *Solving Mixed Model Equations in R*, 2017.

G. de los Campos, D. Sorensen, and D. Gianola. Genomic heritability: What is it? *PLoS Genetics*, 11:e1005048, 2015.

L. Dempfle. Personal Communication, January 2018.

D. S. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics*. Pearson, 4th edition, 1996.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70:849–911, 2008.

J. Fan, F. Han, and H. Liu. Challenges of big data analysis. *National Science Review*, 1(2):293–314, 2014.

R. Fernando, H. Cheng, X. Sun, and D. J. Garrick. A comparison of identity-by-descent and identity-by-state matrices that are used for genetic evaluation and estimation of variance components. *Journal of Animal Breeding and Genomics*, 134:213–223, 2017.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2014.

D. Gianola, G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. Additive genetic variability and the bayesian alphabet. *Genetics*, 183:347–363, 2009.

G. H. Givens and J. A. Hoeting. *Computational Statistics*. Wiley, 2nd edition, 2013.

D. Golan, E. S. Lander, and S. Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111:E5272–E5281, 2014.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2008.

C. R. Henderson. *Applications of Linear Models in Animal Breeding*. University of Guelph, 1984.

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60, 1981.

W. G. Hill. Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B*, 365:73–85, 2010.

W. G. Hill, M. E. Goddard, and P. M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4(2):1–10, 2008.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Techometrics*, 12(1):55–67, 1970.

A. J. Izenman. *Modern Multivariate Statistical Techniques*. Springer, 1st edition, 2008.

T. Kneib, S. Konrath, and L. Fahrmeir. High dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance. *Applied Statistics*, 60(1):51–70, 2011.

S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions*. Wiley, 2nd edition, 2000.

S. K. Kumar, M. W. Feldman, D. H. Rehkopf, and S. Tuljapurkar. Limitations of GCTA as a solution to the missing heritability problem. *PNAS*, pages E61–E70, 2015.

S. K. Kumar, M. W. Feldman, D. H. Rehkopf, and S. Tuljapurkar. Respone to "commentary on limitations of GCTA as a solution to the missing heritability problem". *bioRxiv*: `http://dx.doi.org/10.1101/039594`, 2016.

A. Legarra. Comparing estimates of genetic variance across different relationship models. *Theoretical Population Biology*, 107:26–30, 2015.

C. Lehermeier, G. de los Campos, V. Wimmer, and C-C. Schön. Genomic variance estimates: With or without disequlibrium covariances? *Journal of Animal Breeding and Genomics*, 134:232–241, 2017.

L. Leon-Novelo and G. Casella. Prior influence in linear regression when the number of covariates increases to infinity. *Elsevier Statistics and Probability Letters*, 82, 2011.

B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(6): 18–21, 2008.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

P. Perez and G. de los Campos. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, 198:483–495, 2014.

H-P. Piepho and J. Moehring. Computing heritability and selection response from unbalanced plant breeding trials. *Genetics*, 177(3):1881–1888, 2007.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

N. Schreck and M. Schlather. From estimation to prediction of genomic variances: Allowing for linkage disequilibrium and unbiasedness. *bioRxiv*: `http://dx.doi.org/10.1101/282343`, 2018.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. Wiley Interscience, 1992.

The 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, 166:481–491, 2016.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. P. Rawlins, R. Mott, and J. Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38 (8):879–887, 2006a.

W. Valdar, L. C. Solberg, D. Gauguier, W. O. Cookson, J. N. P. Rawlins, R. Mott, and J. Flint. Genetic and environmental effects on complex traits in mice. *Genetics*, 174(2):959–984, 2006b.

P. M. VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008.

J. Wakefield. *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics, 2013.

J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *National Genetics*, 42(7):565–569, 2010.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88:76–82, 2011.

J. Yang, S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher. Commentary on "Limitations of GCTA as a solution to the missing heritability problem". *bioRxiv*: `http://dx.doi.org/10.1101/036574`, 2016.

Z. Zhu, A. Bakshi, A. A. E. Vinkhuyzen, G. Hemani, S. H. Lee, I. M. Nolte, J. V. can Vliet-Ostaptchouk, H. Snieder, The LifeLines Cohort Study, T. Esko, L. Milani, R. Mägi, A. Metspalu, W. G. Hill, B. S. Weir, M. E. Goddard, P. M. Visscher, and J. Yang. Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics*, 96:377–385, 2015.

List of Abbreviations and Symbols

Abbreviations

BL	Body Length
BLUE	Best Linear Unbiased Estimation/ Estimator
BLUP	Best Linear Unbiased Prediction/ Predictor
BMI	Body Mass Index
BRM	Bayesian Regression Model(s)
BRR	Bayesian Ridge Regression
FEM	Fixed Effect Model(s)
GBLUP	Genomic Best Linear Unbiased Prediction/ Predictor
GCTA-GREML	Genome-wide Complex Trait Analysis Genomic REML
GRM	Genomic Relationship Matrix
GWAS	Genome-wide Association Studies
LD	Linkage Disequilibrium
MCMC	Markov-Chain Monte Carlo
MSE	Mean-squared-error
OLS	Ordinary Least Squares
QTL	Quantitative Trait Locus/ Loci
REM	Random Effect Model(s)
REML	Restricted Maximum Likelihood
SNP	Single Nucleotide Polymorphism
rBias	relative Bias
rLD	relative contribution of LD to the genomic variance
rRMSE	relative root-mean-squared-error

Genetics

E	Environmental deviations
G	Genetic/ genomic value
P	Phenotypic value
V	Variance of the genetic/ genomic value

Genomics

$\mathbf{G}$	$n \times n$ genomic relationship matrix
$\hat{V}$	Estimated version of variable V
V^{equi}	Marginal genomic variance in equivalent model
V^{real}	Marginal genomic variance in realized model
V_{b}	Prior genomic variance in the BRM
V_{f}	Genomic variance in the FEM
V_{r}	Marginal genomic variance in the REM
W	Random genomic variance
W_{b}	Posterior mean of W in BRM
W_{r}	Predictor for W in REM
X	Stochastic p-vector of marker genotypes
$\mathbf{X}$	$n \times p$ matrix of marker genotypes
Y	Phenotypic value of random individual
g	n-vector of genomic values, equals $X\beta$
h^2	Narrow-sense heritability
y	n-vector of phenotypic values
β	Effect of marker-allele substitution
ε	Environmental deviations
μ	Intercept
μ_β	Mean of β
$\mu_{\beta\|y}$	BLUP for β
$\mu_{g\|y}$	GBLUP for g
σ^2_β	Marginal variance of random effect β
σ^2_ε	Variance of ε
σ^2_g	"Genomic" variance in GBLUP
σ^2_{Y}	Phenotypic variance

Miscellaneous

$\mathcal{G}(a, b)$	Gamma distribution with shape $a > 0$ and scale $b > 0$
$\mathbf{H}$	hat-matrix
I_{r}	Indicator for the relative contribution of LD to genomic variance in REM
$\text{IG}(a, b)$	Inverse-gamma distribution with shape $a > 0$ and scale $b > 0$
K_{m}	Set containing the numbers of QTL for the mice dataset

K_{rm}	Set containing the numbers of QTL for the reduced mice dataset
M	Number of MCMC samples after burn-in and thinning
k	Number of QTL
n	Number of individuals/ observations
p	Number of markers
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
Σ_Z	Variance-Covariance matrix of a random vector Z
$\mathbb{1}_{p\times p}$	$p \times p$ identity matrix
1_n	Column-n-vector with every entry equal to 1