Applied Statistical Methods in Animal Sciences

Peter von Rohr

2021-02-19

Contents

Preface

This document contains the course notes for

751-7602-00L Applied Statistical Methods in Animal Sciences.

General Developments

With the advent of **Big Data** (see [Wikipedia, 2019] and [Mashey, 1998] for a reference), it became clear that the importance of statistical methods to analyze the huge amounts of collected data would increase dramatically. Many modern statistical methods are only applicable due to the vast availability of cheap computing resources. The progress of the development that happens in the hardware manufacturing industry is often referred to by the term **Moore's Law**. This law was stated as a projection as early as 1965 by one of the founders of the Intel cooperation [Moore, 1965]. In a very general term, Moore's law says that the number of circuits that could be placed on a silicon waver would double every 18 months. In a derived version the law was interpreted in a way that the performance of computers would double every 18 months. Together with the high degree of automated production of the building blocks of a computer, the prices for a single unit of computation dropped dramatically. This development made it possible that the possibility to analyze large amounts of data with modern methods can be done by almost everyone. This created very many opportunities which are actively used by many business companies. Statistical methods used to be only used by academic researchers. Nowadays almost all important decisions in business companies are done based on supporting facts that are derived from analyzing market and customer data. With that it is clear that the importance of being able to use statistical methods to analyze data is almost ubiquitous and the knowledge of these methods can be very important in many different jobs or employments.

Where Does This Course Fit In?

This course gives a short introduction to a collection of statistical methods that I believe are relevant for a wide range of topics in Animal Sciences. These methods include

- Multiple Linear Least Squares Regression (MLLSR)
- Best Linear Unbiased Prediction (BLUP) which is called GBLUP when applied in the context of genomics
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Bayesian Estimation of Unknown Parameters (BEUP)

The above listed collection of statistical methods all happen to be illustrated around the same type of dataset. This dataset contains the genetic variants at many locations in the genome for a number of livestock breeding animals. Because there are many genetic locations considered in such a dataset and the locations are distributed across the complete genome, such a dataset is referred to as a **genomic** dataset. This type of dataset does appear in an area of livestock breeding which is called **Genomic Selection** (GS). GS was introduced in a seminal paper by [Meuwissen et al., 2001]. This very same paper is used as a building block to explain some of the statistical methods (MLLSR and BEUP) used in this course. Furthermore the same publication illustrates that some methods (MLLSR) are not suitable for analyzing certain aspects in a genomic dataset.

The time available for this course is just half a semester. This leaves very little time for the introduction of each topic. As a consequence of that each topic can only be presented very superficially and students are expected to work on their own during the exercise hours. Exercises consist of sets of problems related to each topic. Problems are often to be expected to be solved using the R programming language [R Core Team, 2018].

This version of the course corresponds to the fifth edition. With each additional iteration of the course, improvements are sought to be implemented. Hence any input from the students are greatly appreciated.

Course Objectives

The students are familiar with the properties of multiple linear regression and they are able to analyze simple data sets using regression methods. The students know why multiple linear regression cannot be used for problems where the number of parameters exceeds the number of observations. One such problem is the prediction of genomic breeding values used in genomic selection. The students know alternative statistical methods that can be applied in situations where the number of parameters is larger than the number of observations. Examples of such methods are BLUP-based approaches, Bayesian procedures

and LASSO. The students are able to solve simple exercise problems applying BLUP-based approaches, LASSO and BEUP. The students are expected to use the statistical language and environment R [R Core Team, 2018].

Prerequisites

Because the data that is used in this course comes from genetics, a basic level of quantitative genetics is useful for this course. All statistical models will be presented in matrix-vector notation, hence some basics of linear algebra helps in understanding the presented material. Introductory chapters to both subjects (quantitative genetics and linear algebra) are included in these course notes, but will not be discussed during the lecture. These chapters are prepared for students who feel that they need more background. But this material is left for self-studying.

Chapter 1

Introduction

According to Wikipedia [Wikipedia, 2019], the term `Big Data` has been used since the 1990s. Some credit was given to John Mashey [Mashey, 1998] for popularizing the term. Nowadays `Big Data` is used in connection with large companies, social media or governments which collect massive amounts of data. This data is then used to infer certain conclusions about behaviors of customers, or followers or voters. The following subsections show a few examples of `Big Data`-applications.

1.1 US Presidential Campaigns

The presidential election campaigns of Barack Obama were examples of how `Big Data` was used to access behaviors of voters [Issenberg, 2013].

1.2 Health Care

A different example is the use of `Big Data` in health care. An overview of the use of `Big Data` in health care is given in [Adibuzzaman et al., 2017]. The collected health data is most likely not only used by research but also by insurance companies.

1.3 Face Recognition

The Swiss TV news show `10 vor 10` showed on the 7th Feb. 2020 how a data journalist managed to build a face recognition system. The general idea how this system works is shown in Figure 1.1.
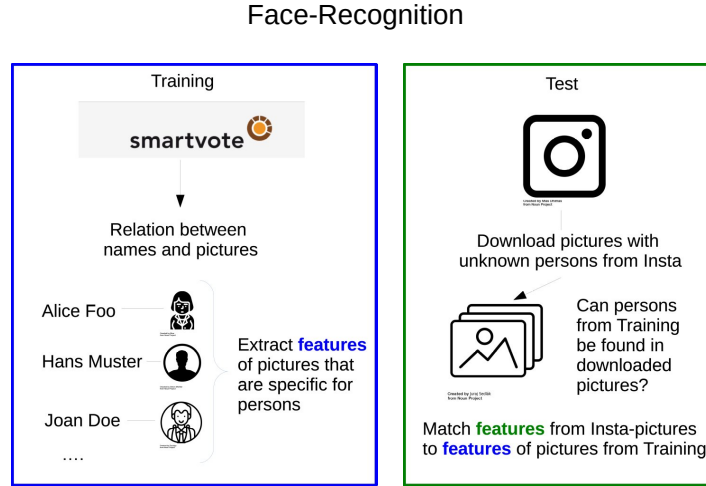
Face-Recognition

Figure 1.1: Design of Face Recognition System

The main goal of the face recognition system was to be able to identify certain persons, in this case the politicians that had a picture on the platform 'smartvote', on random pictures obtained from the social media platform Instagram. The data used for the face recognition system can be split into two parts.

1. **Training** set. The training set consists of pictures showing different politicians. This dataset was downloaded from the politics platform smartvote. The training set is used to establish a fixed relation between pictures of politicians and their names. Furthermore the pictures of the politicians are also used to extract characteristic features that are different among the different pictures of the different politicians. Most of these features consist of numbers which describe the faces shown on the pictures. Examples of such features might be the surface of the face, the surface of the hairs shown in the picture, the length of the mouth, etc.

2. **Test** set: The test set consists of 230000 publicly available pictures on the platform Instagram. The content of these pictures is a priori unknown. The question that the face recognition system tries to answer is whether it is possible to identify any of the politicians from the training set on any of the Instagram-pictures. This question is answered by a comparison of the features extracted from the instagram pictures to the features that were obtained from the pictures of the training set. If the feature comparison results in a match, the system suggests that we found a given person on

one of the instagram pictures.

The complete story about the face recognition system is available under https://www.srf.ch/news/schweiz/automatische-gesichtserkennung-so-einfach-ist-es-eine-ueberwachungsmaschine-zu-bauen.

1.4   Feed Intake and Behavior Traits of Cows

In the recent past technologies based on computer vision have been introduced into agricultural applications. Two examples of such applications are

1. Estimation of feed intake of cows based on video data as described by [Chizzotti et al., 2015].
2. An ongoing EU-Interreg project called "SESAM" aims at predicting basic behavior traits from data obtained from sensors an from video recordings.

1.5   Conclusions from Examples

The above shown examples demonstrate that data can be used for very different purposes. Using just one source of data does in most cases not give a lot of insights. But when different sources of information are combined, they can be used to make certain predictions that influences our daily lives. Hence this kind of development is becoming a general interest to all of us. In what follows, we try to show that some of these methods have been applied for a long time in the area of animal science and especially in livestock breeding.

1.6   Traditional Livestock Breeding

In livestock breeding the statistical analyses that are used together with `Big Data` technologies have long been applied to predict breeding values for livestock populations. The process of breeding value prediction uses statistical methods to assess the genetic potential of breeding animals in a population. The data used to predict the breeding values are collected mainly for quality control or management purposes. The prediction of breeding values can be viewed as a side product. In the area of cattle breeding, data collection consists of rather complex flows of information. The flow of information is shown in Figure 1.2.

1.7   Genomic Selection

The data flow shown in Figure 1.2 contains the traditional evaluation of data to result in predicted breeding values. But it is missing the newest development
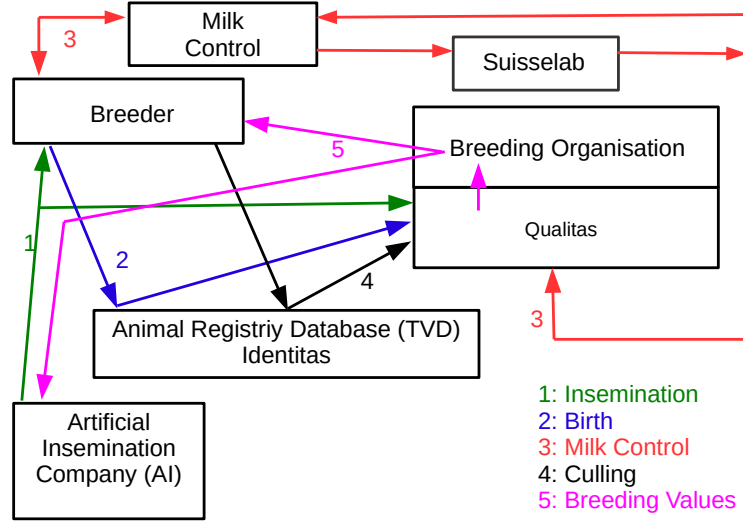
Figure 1.2: Data Flow in an Animal Breeding Program

in the breeding industry. This development is known as `Genomic Selection`
(GS). GS was introduced by the work of [Meuwissen et al., 2001]. The methods presented by [Meuwissen et al., 2001] were only introduced into practical
breeding programs when [Schaeffer, 2006] showed the tremendous potential of
saving costs for breeding programs. The use of **genomic** information for the
assessment of the genetic potential of all breeding animals represents the core
of the evaluation approach presented by [Meuwissen et al., 2001]. The term
`genomic` is used because genetic markers which are evenly spaced over the complete genome are used as information source. Single Nucleotide Polymorphisms
(SNP) are the most widely used marker model nowadays. SNPs are single positions in the genome that occur in different variants in the whole population.
A description on how to identify SNPs in a population is given in [Czech et al.,
2018]. Potential use cases of SNPs are outlined by [Seidel, Jr., 2010] and [Pant
et al., 2012]. The genetic configuration of an SNP in a given population is shown
in Figure 1.3.

These SNPs can occur anywhere in the genome which means they can be observed in coding regions, in non-coding regions as well as in regulatory regions.
In genomic selection, we are working with a large set of SNPs that are distributed over the complete genome. Hence some of the SNPs will be located
close to genetic positions that are important for the expression of quantitative
traits of interest. Such genetic positions which are related to quantitative traits
are often called `Quantitative Trait Loci` (QTL). QTL themselves are difficult
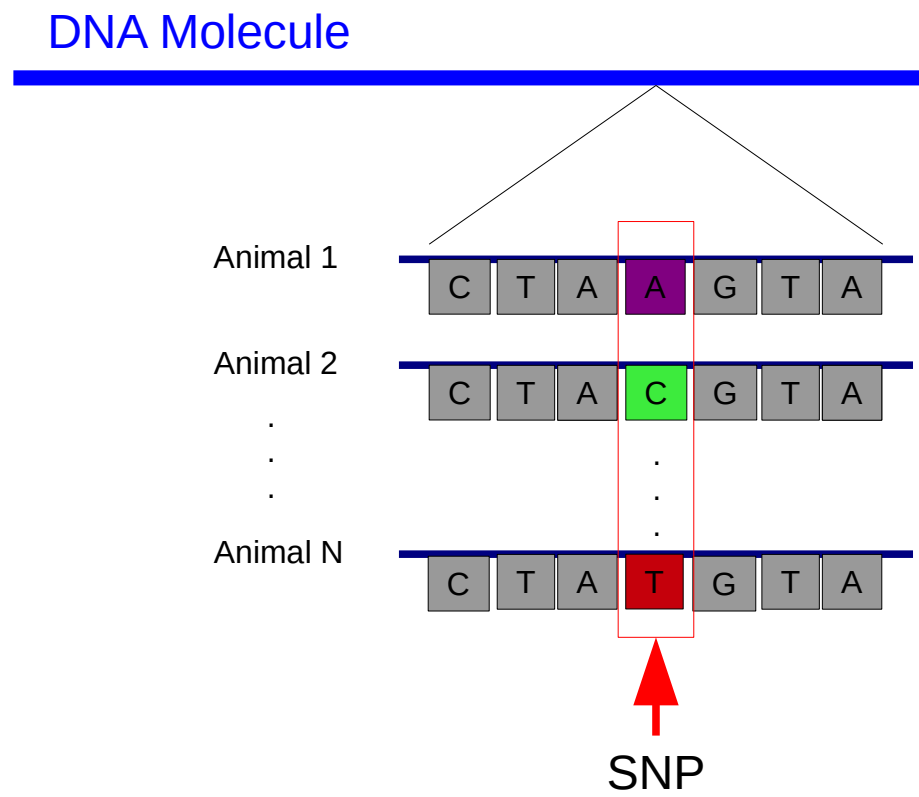to detect and their inheritance is often manifested in complex modes. But due

Figure 1.3: Genetic Configuration of a Single Nucleotide Polymorphism (SNP)

to the likely occurrence of several SNPs in the close proximity of a QTL, the inheritance of QTL alleles and of surrounding SNP alleles will not be independent due to linkage between SNPs and QTL. Such a linkage scenario between two SNPs flanking a QTL is shown in Figure 1.4.
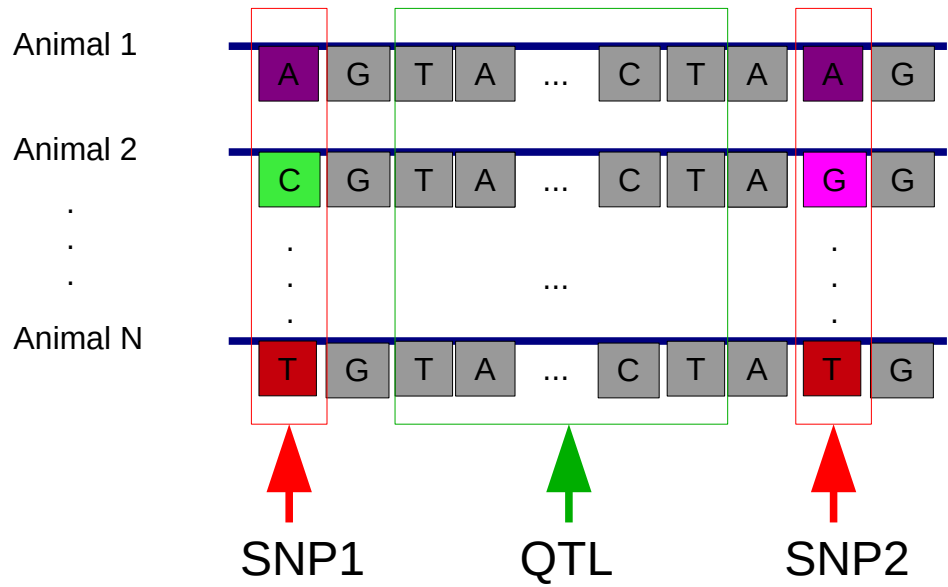
Figure 1.4: Two SNPs flanking a QTL

Although the QTL is likely to span a range of many positions on the chromosome, we can still assume the QTL to be bi-allelic with alleles Q_1 and Q_2. In theory, any SNP position can have four different alleles according to the four different bases. But when looking at different SNPs in real-world populations, most of them only show two alleles. Hence, for the two SNPs flanking the QTL shown in Figure 1.4 they also have just two alleles $SNP1_1$, $SNP1_2$, $SNP2_1$ and $SNP2_2$. In genetics the dependency of the inheritance of neighboring loci (marker or QTL) is referred to as `linkage disequilibrium` (LD). This means that any joint allele frequency $Pr(SNP1_i, Q_j, SNP2_k)$ does not correspond to

the product of the single allele frequencies of the two SNPs ($SNP1$ and $SNP2$) and the QTL. In a formula this can be written as

$$Pr(SNP1_i, Q_j, SNP2_k) \neq Pr(SNP1_i) * Pr(Q_j) * Pr(SNP2_k) \qquad (1.1)$$

Assuming that the QTL allele Q_1 is favorable for the expression of a given trait of interest and using the fact of LD as expressed in (1.1), the alleles of $SNP1$ and $SNP2$ which occur more frequently together with Q_1 are therefore also related to favorable expression levels of the trait of interest. In real breeding populations, the position of the QTL is unknown. But because we know the allelic configuration of a large number of SNP loci from many breeding animals, we can reliably relate SNP alleles and favorable expression levels of traits of interest.

1.8   Mono-Genic Model

In quantitative genetics, the so-called mono-genic or single-locus model allows us to quantify the genetic potential of breeding animals in terms of breeding values. The standard reference in quantitative genetics in which also the mono-genic model is described is [Falconer and Mackay, 1996]. For a single locus, the breeding value depends on the allele frequencies at that locus and on the additive substitution effect which is often called α. The mono-genic model for any given SNP locus in relation to the level of expression of a given trait of interest can be visualized in the following Figure 1.5.

In a real breeding population, we assume that the effect of all loci linked to the SNPs are purely additive. Hence any values for d are all zero. As a consequence of that the breeding values at any given SNP position only depend on the allele frequencies of the SNP and the a values at every SNP. The overall breeding value of a given animal is computed as the sum of all locus-specific breeding values. This overall breeding value is called **genomic breeding value** (GBV). In order to get an estimate of such a GBV, we have to estimate all a values at any SNP position. This estimation procedure can be done in one of the following two ways.

1. Two step approach
2. Single step approach

1.9   Two Step Approach

In the two step approach the estimation of the a-values and the computation of the GBVs are done in two separate steps. For the estimation of the a values for all SNPs, a reference population is defined. In dairy cattle breeding this reference population consists of all male breeding animals. In the recent past,
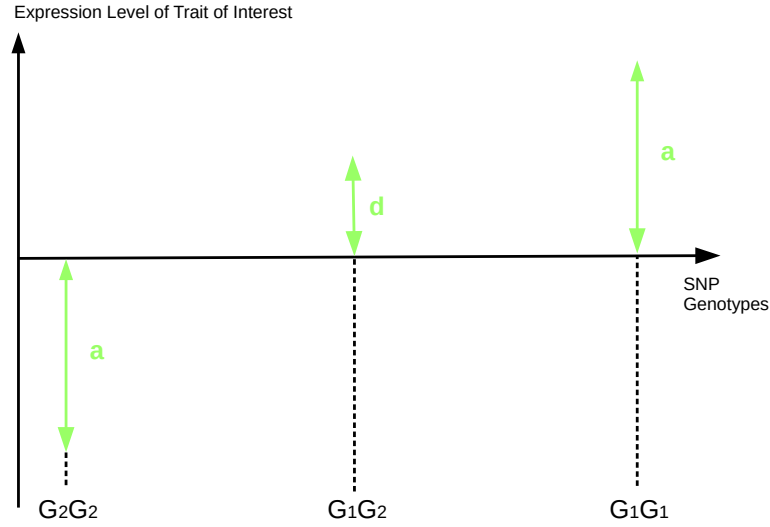
Figure 1.5: Single-Locus Model for a Quantitative Trait

the reference population has been augmented continuously with female animals. The animals in the reference population are all genotyped and they also all have phenotypic measurements[1] for the trait of interest. The estimation of the a values amounts to estimating fixed effects in a linear model. We will see in later chapters of this course what methods are available to estimate these parameters.

In the second step the estimates for all the a values are used to compute the GBVs for all animals with genomic SNP information also for those outside of the reference population. The Figure 1.6 tries to summarize the process graphically.

The big advantage of the two step method is that once we have defined a good reference population which yields reliable estimates for the a values, the computation of the GBV is a simple computation of just summing up the a contributions with the correct sign determined by the SNP genotypes of the animals for which the GBVs should be determined. All animals with SNP genotypes can get GBV values. The difficult part in the two step approach is to define a reliable reference population and to determine good phenotypic measurements (y).

---

[1]Whenever phenotypic measurements are not available, traditionally predicted breeding values are transformed back into pseudo phenotypes which are then used to estimate a values.
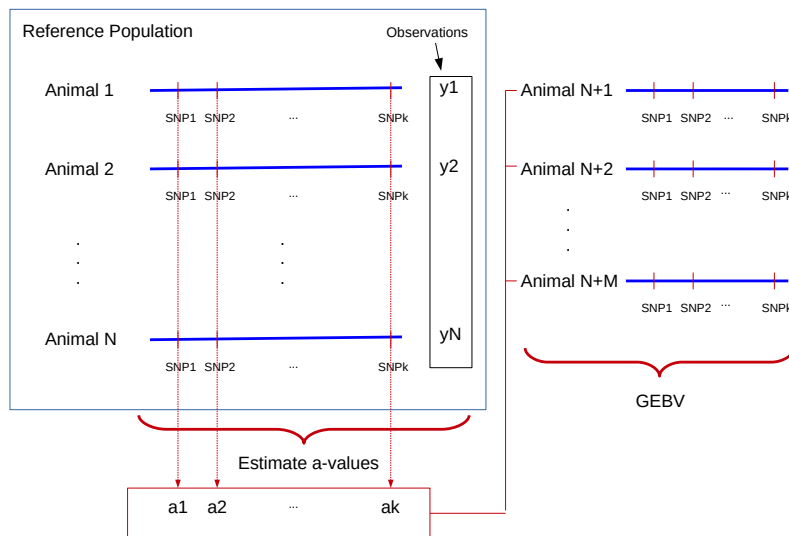
Figure 1.6: Two Step Approach To Estimate Genomic Breeding Values

1.10 Single Step Approach

The estimation of the a values and the prediction of the genomic breeding values is done in one step using linear mixed effects models. In this single step evaluation animals with and without genomic information can get predicted genomic breeding values in a single analysis. One possibility to get to this predicted breeding values is via the use of `Genomic BLUP` (GBLUP). This will be the topic of a complete chapter in this course. The problem with the single step approach is to get an estimate of the covariance between animals with and without genomic information. This is a problem of ongoing research.

1.11 Summary

The main difference between traditional predictions of breeding values using a BLUP animal model and the prediction of GBV is that the former uses the so called **infinitesimal** model to assess the genetic potential and the latter uses sufficiently dense genomic information and uses a **polygenic** model. This difference is illustrated in Figure 1.7.

In the remaining chapters, different approaches for the prediction of GBVs are described. Chapter 2 gives a description of the fixed linear effects model and how it was tried to be used for GBV prediction by [Meuwissen et al., 2001].
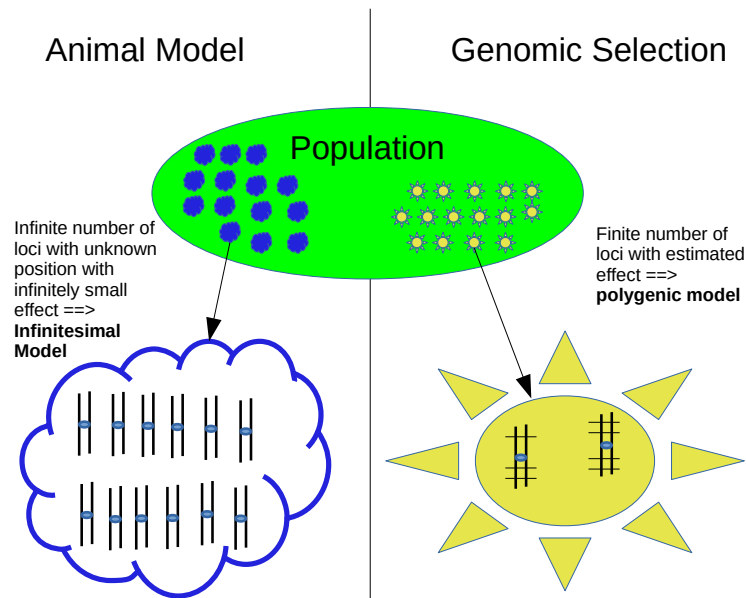
Figure 1.7: Infinitesimal Versus Polygenic Model

Chapter 3 introduces BLUP methodology in the context of predicting GBVs. In Chapter 4 the method called LASSO is introduced. Interestingly enough, this method is used very seldom in the area of animal breeding. Last but not least, Chapter 5 makes an excursion into Bayesian estimation approaches. The Bayesian methods are important because they are used in practical breeding programs of Swiss Dairy cattle.

Appendix A

Introduction To Linear Algebra

Linear Algebra is a large area from which we only need the following three topics

1. Vectors
2. Matrices and
3. Systems of linear equations.

A.1 Glimpse Ahead

The central topic of this course is the prediction of breeding values. Most approaches to predict breeding values require the solution of large systems of linear equations. These systems of equations are written down using vectors and matrices. Hence the three mentioned topics are important to understand at a level that they can be used as tools for the prediction of breeding values.

A.2 Vectors

The material of this section is largely based on the video tutorial (https://youtu. be/fNk_zzaMoSs) from [3blue1brown, 2016]. We try to give a summarized transcript of the video. The vector is the fundamental building block of linear algebra. There are three different but related concepts about what vectors are. We call them

1. the physics perspective
2. the computer science perspective and
3. the mathematics perspective.

The mathematics perspective tries to provide a very general concept, saying that anything can be a vector as long as, one can add two vectors or a vector can be multiplied by a factor and the result of both operations is a vector again. For what we want to use vectors for in the context of livestock breeding and genomics, the mathematics perspective is not so useful, hence we ignore it from now on.
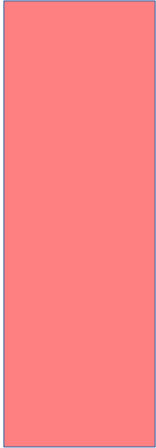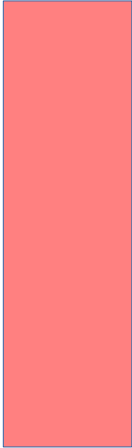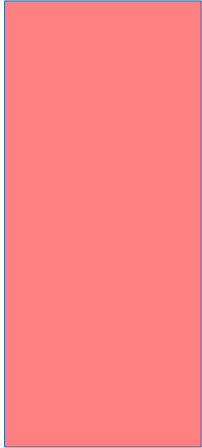
A.2.1   Physics Perspective

The physics perspective is that vectors are arrows with a certain **length** and a **direction** they are pointing to. As long as length and direction are the same, the arrows can be moved around and they are still the same vector. Different arrows with the same length and the same direction are called **representatives** of the same vector. Vectors that are in a flat plane are called two-dimensional. Those who are sitting in the same Euclidean space that we are all living in, are called three-dimensional.

Representative 1
of vector w

Vector v

Vector w

Representative 2
of vector w
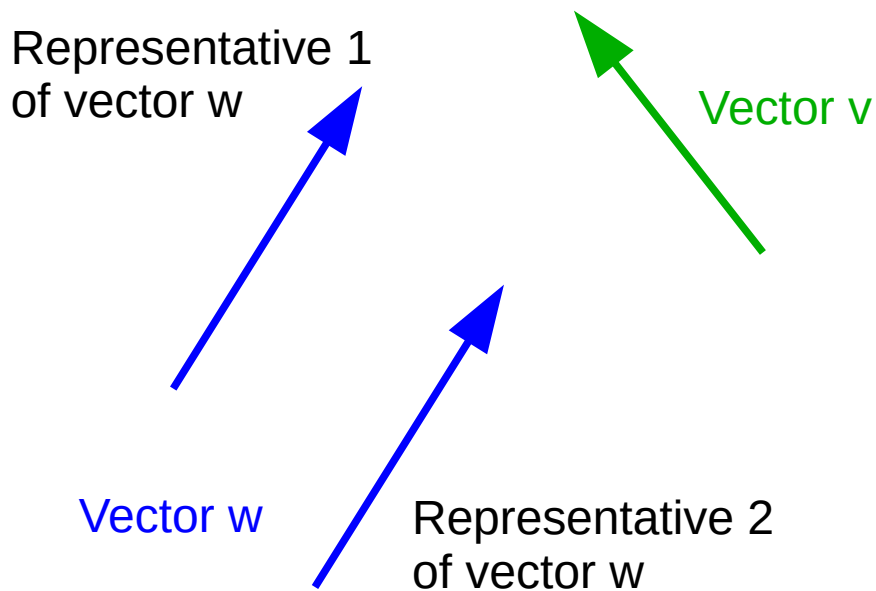
A.2.2 Computer Science Perspective

In the computer science perspective vectors are ordered list of numbers. Later we will see that vectors can also contain more general objects like strings. As an example, we assume that we are analyzing carcasses and the only thing we know about a carcass is its slaughter-weight (SW) and its price (P). The different carcasses can then be represented by a pair of numbers the first being the slaughter-weight and the second being the price. It is important to note here, that the order of the number matters. In terms of vectors, here each carcass is represented by a two-dimensional vector.

	Carcass 1	Carcass 2	Carcass 3
Weight	290 kg	265 kg	320 kg
Price	1943 Fr	1829 Fr	1984 Fr

A.2.3 Geometric Context

Some basic properties of vectors are introduced using the geometric context, that a vector is an arrow located in a certain coordinate system with its tail sitting at the origin of the coordinate system. This is a little bit different from

the physics perspective (see A.2.1) where the arrow can sit anywhere in space. In linear algebra it is almost always the case that vectors are rooted at the origin. Once we understand the properties of vectors in the context of arrows in space, we can then translate these properties to the list-of-numbers point of view (see A.2.2) considering the coordinates of the vectors.

A.2.4 Coordinate System

It is important to introduce the coordinate system, because this will be the basis of the correspondence between the two perspectives of linear algebra. For the moment, we focus on two dimensions. The horizontal line is called the x-axis and the vertical line is called the y-axis. The place where the two lines intersect is called the origin. An arbitrary length is chosen to represent 1. The coordinates of a vector is a pair of numbers that give instructions for how to get from the tail of that vector at the origin to its tip. The first number tells you how far to walk along the x-axis (positive numbers indicating rightward motion, negative numbers indicating leftward motion) and the second number tell you how far to walk parallel to the y-axis (positive numbers indicating upward motion, negative numbers indicating downward motion).
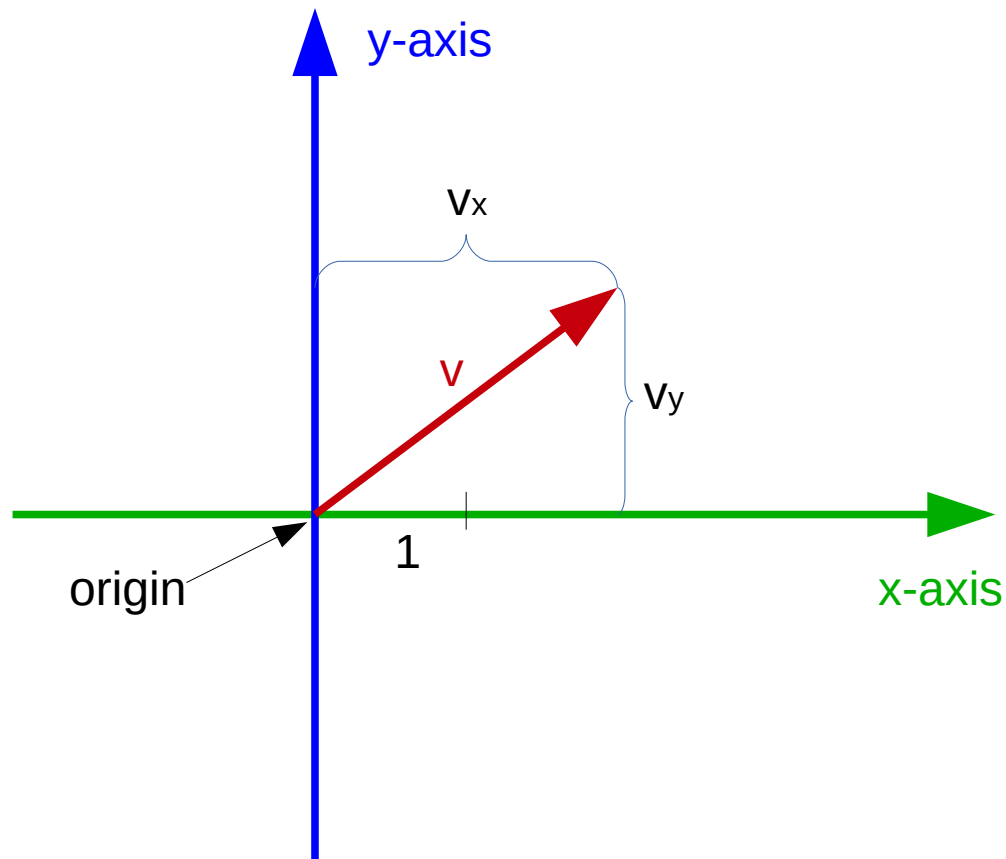
A.2.5 Vector Operations

The vectors by themselves can be pretty interesting objects, but they get really useful when considering some operations that we can perform on them. Here we consider three basic operations.

1. addition
2. multiplication by a scalar number and
3. dot product

A.2.5.1 Addition

Let us assume, we have two vectors v and w. To add these two vectors, move the second one such that its tail sits at the tip of the first one. Then draw a

new vector from the tail of the first one to the tip of the second one. The new vector corresponds to the sum of the two vectors (Figure A.1).
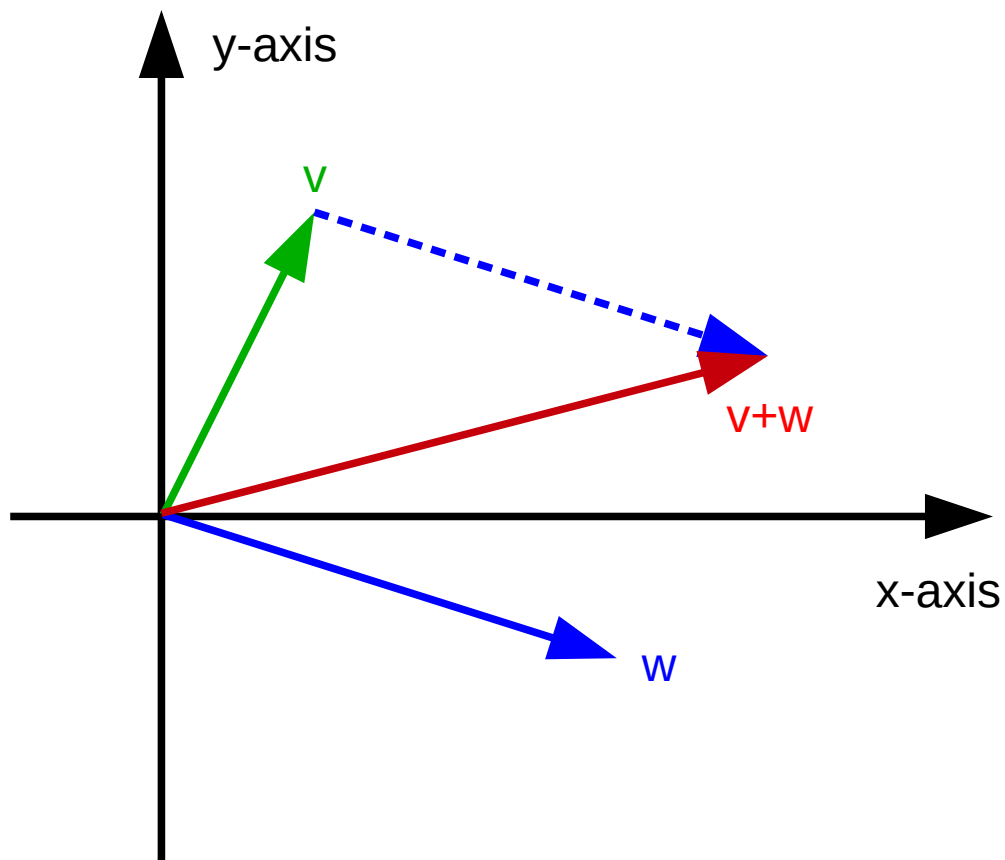
Figure A.1: Addition of two vectors

Numerically, vector addition corresponds to summing up each of the coordinates individually. Hence if we have two vectors v and w with their coordinates given as

$$v = \left[\begin{array}{c} v_x \\ v_y \end{array} \right], \ w = \left[\begin{array}{c} w_x \\ w_y \end{array} \right]$$

then the sum $v + w$ has coordinates

$$v + w = \left[\begin{array}{c} v_x + w_x \\ v_y + w_y \end{array} \right]$$

A.2.5.2 Multiplication by a Scalar Number

This operation is best understood by looking at a few examples. If we take the number 2 and multiply it by a certain vector v, this means that we stretch out the vector v such that it is 2 times as long as the original vector. Multiplication of a vector with positive numbers does not change the direction of the vector. Multiplying a vector v with a negative number like -0.5 then the direction gets flipped around and then squished by 0.5.
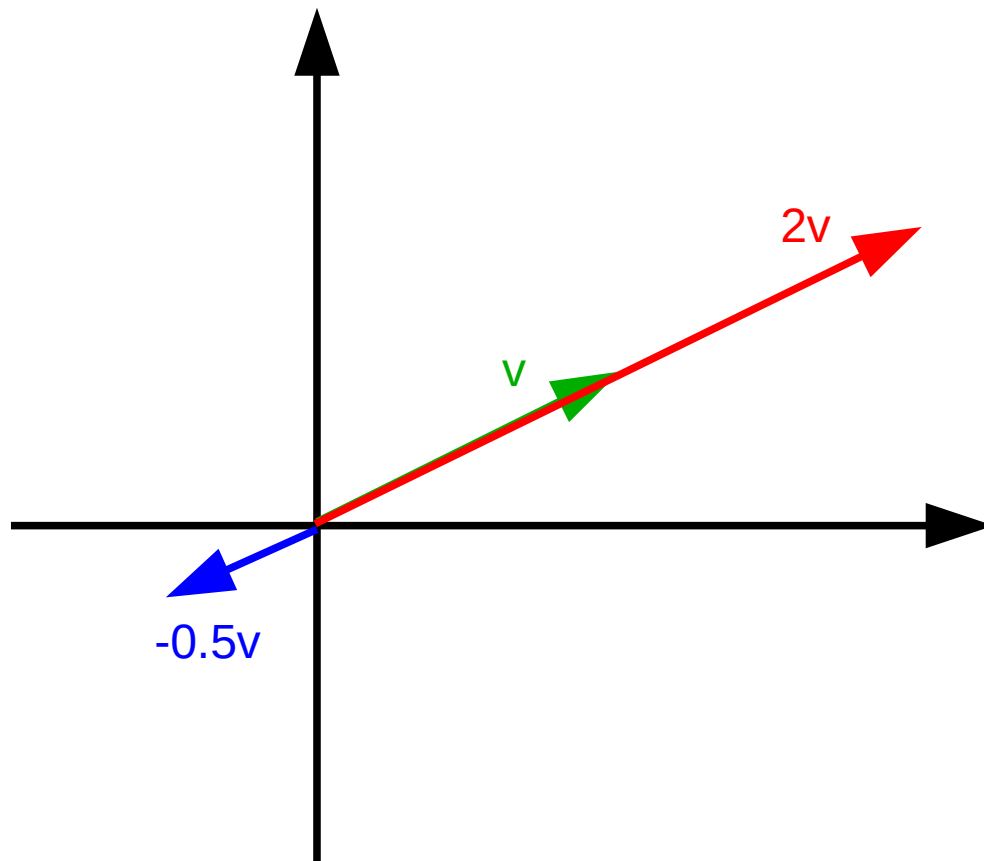
The operation of multiplying a vector by a given number, like 2 or -0.5 is also called **scaling** and that is the reason why in linear algebra the numbers like 2 and -0.5 are called **scalar** numbers or just scalars. Numerically, stretching a vector by a given number like 2, corresponds to multiplying each of the coordinate components by that factor 2. For a vector v with coordinate components v_x and v_y, the vector $2v$ has coordinates $2v_x$ and $2v_y$

$$v = \left[\begin{array}{c} v_x \\ v_y \end{array} \right] \quad , \quad 2v = \left[\begin{array}{c} 2v_x \\ 2v_y \end{array} \right]$$

A.2.5.3 Dot Product

The dot product is explained in a different video that can be seen on https: //youtu.be/LyGKycYT2v0. Numerically, if you have two vectors of the same dimension, meaning two lists of numbers of the same length, e.g. v and w then their dot product $v \cdot w$ can be computed by pairing up all of the coordinates, multiplying these pairs together and adding the result. So the vectors

$$v = \left[\begin{array}{c} v_x \\ v_y \end{array} \right] \quad \text{and } w = \left[\begin{array}{c} w_x \\ w_y \end{array} \right]$$

their dot product $v \cdot w$ then is computed as

$$v \cdot w = v_x * w_x + v_y * w_y$$

A.3   Matrices

The introduction to the topic of matrices is available from https://youtu.be/ kYB8IZa5AuE and https://youtu.be/XkY2DOUCWMU. An $m \times n$ matrix is a table-like object of $m * n$ numbers arranged in m rows and n columns. In general the $m \times n$ matrix A has the following structure.

$$A = \left[\begin{array}{cccc} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & ... & a_{mn} \end{array} \right]$$

The $m*n$ numbers inside of the square brackets are called elements of the matrix. The element of matrix A that is in row i and in column j is called a_{ij} or $(A)_{ij}$. As an example

$$A = \left[\begin{array}{ccc} 2 & 3 & 1 \\ 5 & 1 & 2 \end{array} \right]$$

is a 2×3 matrix. In the first row the second element corresponds to $(A)_{12} = a_{12} = 3$. An $n \times n$ matrix (i.e. a matrix with equal numbers of rows and columns) is called a **quadratic** matrix. Two matrices A and B are called **equal**, if they have the same number of rows and columns and if the corresponding elements are the same, i.e.

$$(A)_{ij} = (B)_{ij} \text{ for all i and j}$$

A.3.1 Special Matrices

The following matrices are special and are used in special cases.

- **Nullmatrix**: The $m \times n$ matrix 0 is called Nullmatrix, if each element is equal to zero.
- **Upper Triangular Matrix**: The square matrix R is called upper triangular matrix, if $(R)_{ij} = 0$ for $i > j$.
- **Lower Triangular Matrix**: The square matrix L is called lower triangular matrix, if $(L)_{ij} = 0$ for $i < j$.
- **Diagonal Matrix**: The square matrix D is called diagonal matrix, if $(D)_{ij} = 0$ for $i \neq j$.
- **Identity Matrix**: The diagonal matrix I is called identity matrix, if all diagonal elements $(I)_{ii} = 1$.
- **Column Vector**: A $m \times 1$ matrix is often called a column vector.
- **Row Vector**: A $1 \times n$ matrix is is often called a row vector.

A.3.2 Matrix Operations

The following operations with matrices are defined.

A.3.2.1 Addition

For two $m \times n$ matrices A and B, their sum $A + B$ is again a $m \times n$ matrix with each element corresponding to the sum of the corresponding elements from A and B. Hence, we can write

$$(A + B)_{ij} = (A)_{ij} + (B)_{ij} \text{ for all i and j}$$

A.3.2.2 Multiplication with a Number

A $m \times n$ matrix A is multiplied by a number α by multiplying every element $(A)_{ij}$ of A with α. The result $\alpha * A$ is computed as $(\alpha * A)_{ij} = \alpha * (A)_{ij}$ for all i and j.

A.3.2.3 Multiplication of two Matrices

Given a $m \times n$ matrix A and a $n \times p$ matrix B, their matrix product AB is a $m \times p$ matrix with

$$(AB)_{ij} = \sum_{k=1}^{n} (A)_{ik} * (B)_{kj} = (A)_{i1} * (B)_{1j} + (A)_{i2} * (B)_{2j} + ... + (A)_{in} * (B)_{nj}$$

A.3.2.4 Laws of Matrix Operations

- **Commutativity**: For two $m \times n$ matrices A and B the addition is commutative, i.e. $A + B = B + A$.
- **Associativity of addition**: For $m \times n$ matrices A, B and C, the addition is associative, i.e., $A + (B + C) = (A + B) + C$
- **Associativity of multiplication**: For a $m \times n$ matrix A, a $n \times p$ matrix B and a $p \times q$ matrix C, the multiplication is associative, i.e., $A(BC) = (AB)C$
- **Distributivity**: For $m \times n$ matrices A and B and $n \times p$ matrices C and D, the distributive law holds, i.e., $(A+B)C = AC + BC$ and $A(C+D) = AC + AD$

A.3.2.5 Matrix Transpose

Given a $m \times n$ matrix A, then the $n \times m$ matrix A^T is called its **transpose**, if $(A^T)_{ij} = A_{ji}$. The matrix A is called **symmetric**, if $A = A^T$. For every matrix A the transpose of the transpose is the matrix itself, i.e., $(A^T)^T = A$. For any $m \times n$ matrices A and B, the transpose $(A + B)^T$ of their sum $(A + B)$ is computed as

$$(A + B)^T = A^T + B^T$$

For every $m \times n$ matrix A and every $n \times p$ matrix B, it holds that

$$(AB)^T = B^T A^T$$

A.3.2.6 Inverse of a Matrix

In this section, we are looking at square matrices. The **inverse** X of a square matrix A is defined as the square matrix that satisfies the condition $AX = I$. If the inverse matrix X exists, then the matrix A is called invertable. If X does not exist, A is called singular. If the inverse of a matrix A exists, it is uniquely determined and we call it A^{-1}.

Let us assume two invertable $n \times n$ matrices A and B, then the following equations hold

1. $A^{-1}A = I$

2. A^{-1} is invertable and $(A^{-1})^{-1} = A$
3. I is invertable and $I^{-1} = I$
4. AB is invertable and $(AB)^{-1} = B^{-1}A^{-1}$
5. A^T is invertable and $(A^T)^{-1} = (A^{-1})^T$

For every square matrix A, the following statements are equivalent.

1. A is invertable
2. The system of equations $Ax = b$ is solvable for every b.
3. The system of equations $Ax = 0$ has only the trivial solution $x = 0$.

A.3.2.7 Orthogonal Matrices

A square matrix A is called **orthogonal**, if the condition $A^T A = I$ holds. For two orthogonal matrices A and B, the following statements hold.

1. A is invertable and $A^{-1} = A^T$
2. A^{-1} is orthogonal
3. AB is orthogonal
4. I is orthogonal

A.4 Systems Of Equations

Systems of linear equations are introduced based on [Nipp and Stoffer, 2002] and [Searle, 1971]. Solving systems of linear equations is one of the fundamental tasks of linear algebra. We start with a general example of a system of linear equations which is given as

$$
\begin{aligned}
x_1 + 2x_2 &= 5 \\
2x_1 + 3x_2 &= 8
\end{aligned}
\tag{A.1}
$$

In (A.1) we are given a system of linear equations with two equations and two unknowns x_1 and x_2. The aim is to find numeric values for x_1 and x_2 such that both equations are satisfied. Inserting the values $x_1 = 1$ and $x_2 = 2$ into the above equations show that they are both satisfied. Hence the set $L = \{x_1 = 1, x_2 = 2\}$ consisting of the values for x_1 and x_2 that satisfy both equations is called a solution or a solution set for the above shown equations.

In general, a linear system of equations consists of m equations and n unknowns. In the example (A.1), $m = 2$ and $n = 2$.

The example in (A.2) does not have any solutions.

$$x_1 + x_2 = 4$$
$$2x_1 + 2x_2 = 5 \tag{A.2}$$

This can be seen, that if the first equation in (A.2) is multiplied by 2, we get $2x_1 + 2x_2 = 8$ which contradicts the second equation shown in (A.2).

A system with $m = 2$ equations and $n = 3$ unknowns in shown in (A.3).

$$x_1 - x_2 + x_3 = 2$$
$$2x_1 + x_2 - x_3 = 4 \tag{A.3}$$

There are infinitely many solutions consisting of $x_1 = 2$, $x_2 = \alpha$ and $x_3 = \alpha$ for any real number α.

The examples in (A.1), (A.2) and (A.3) already show all possible cases that may occur when solving linear systems of equations. The question is how to determine the set of all solutions of a system of linear equations.

A.4.1   Matrix-Vector Notation

So far, we have written systems of linear equations explicitly in the sense that every equation was written on one line. For small systems this is not a problem. But when the number of equations (m) and the number of unknowns (n) get very large, the explicit notation is no longer feasible. Hence, we need a notation that can also be used for large systems of equations. The so-called matrix-vector notation provides an efficient way to write down large systems of equations very efficiently.

We return to the example given by (A.1) and we define the matrix A to be

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix},$$

the vector x to be

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

and the vector y to be

$$y = \begin{bmatrix} 5 \\ 8 \end{bmatrix},$$

With these definitions, we can write the system of equations given in (A.1) using matrix-vector notation as

$$A \cdot x = y \tag{A.4}$$

A.5 Solving Systems of Linear Equations

If matrix A in (A.4) is not singular, i.e. the inverse Matrix A^{-1} of A does exist, the solution x to (A.4) can be written as $x = A^{-1}y$. This result is obtained by pre-multiplying both sides of (A.4) with A^{-1} and since a matrix times its inverse results in the identity matrix I, the solution is obtained as

$$\begin{aligned} A \cdot x &= y \\ A^{-1} \cdot A \cdot x &= A^{-1} \cdot y \\ I \cdot x &= A^{-1} \cdot y \\ x &= A^{-1} \cdot y \end{aligned} \tag{A.5}$$

For systems of equations with a singular matrix A, solutions can be found, if the equations are **consistent**. The linear equations $Ax = y$ are consistent, if any linear relationship existing among the rows of A also exist among the corresponding elements of y. As a simple example, the equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 21 \end{bmatrix}$$

are consistent. In the matrix on the left the second row corresponds to three times the first row and in the vector on the right, the second element is also three times the first element. In contrast the equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 24 \end{bmatrix}$$

are not consistent. From this example, we can already see that non-consistent equations do not have any solutions. But consistent equations $Ax = y$ have a solution which can be written as $x = Gy$ if and only if, $AGA = A$ which means that G is a so-called generalized inverse of A. The matrix G is often written as A^-. The proof of this statement is given on page 9 of [Searle, 1971].

Appendix B

Basics in Quantitative Genetics

As already mentioned in section **??**, the central dogma of molecular biology tells us that the genotype is the basics of any phenotypic expression. The genotype of an individual is composed of a number of genes which are also called **loci**. In this section, we start with the simplest possible genetic architecture where the genotype is composed by just one locus. The connection between the genotype and the phenotype is modeled according to equation (**??**). The phenotype is assumed to be a quantitative trait. That means we are not looking at binary or categorical traits. Categorical traits can just take a limited number of different levels. Examples of categorical traits are the horn status in cattle or certain color characteristics. Quantitative traits do not take discrete levels but they show specific distributions.

B.1 Single Locus - Quantitative Trait

In Livestock there are not many examples where a quantitative trait is influenced by just one locus. But this case helps in understanding the foundation of more complex genetic architectures. We start by looking at the following idealized population (Figure B.1).

B.1.1 Terminology

The different genetic variants that are present at our Locus G are called **alleles**. When looking at all individuals in the population for our locus, we have two different alleles G_1 and G_2. Hence, we call the locus G to be a **bi-allelic** locus. In any given individual of the population, the two alleles of the locus G together
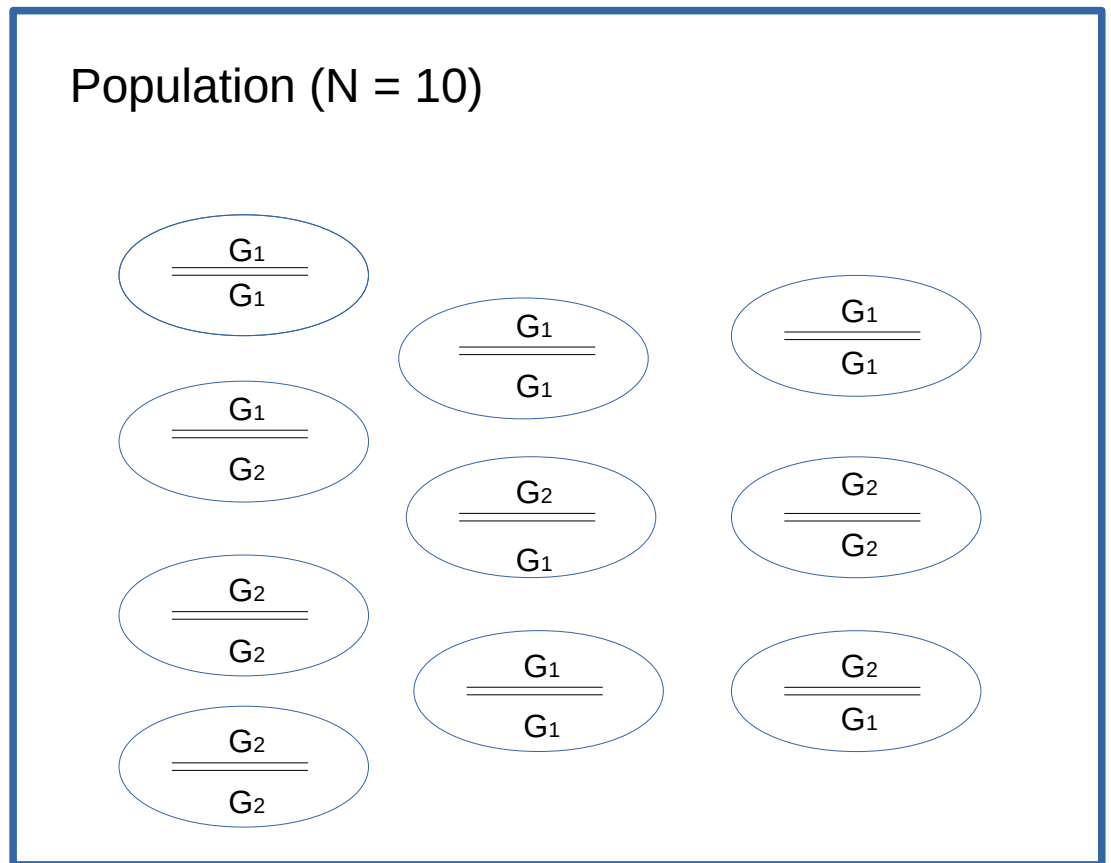
Figure B.1: Idealized Population With A Single Locus

are called the individuals **genotype**. All possible combinations of the two alleles at the locus G leads to a total number of three genotypes. It is important to mention that the order of the alleles in a given genotype is not important. Hence, G_1G_2 and G_2G_1 are the same genotype. The two genotypes G_1G_1 and G_2G_2 are called **homozygous** and the genotype G_1G_2 is called **heterozygous**.

B.2   Frequencies

To be able to characterize our population with respect to the locus of interest, we are first looking at some frequencies. These are measures of how often a certain allele or genotype does occur in our population. For our example population shown in Figure B.1, the **genotype frequencies** are

$$f(G_1G_1) = \frac{4}{10} = 0.4$$
$$f(G_1G_2) = \frac{3}{10} = 0.3$$
$$f(G_2G_2) = \frac{3}{10} = 0.3 \tag{B.1}$$

The **allele frequencies** can be determined either by counting or they can be computed from the genotype frequencies.

$$f(G_1) = f(G_1G_1) + \frac{1}{2} * f(G_1G_2) = 0.55$$
$$f(G_2) = f(G_2G_2) + \frac{1}{2} * f(G_1G_2) = 0.45 \tag{B.2}$$

B.3   Hardy-Weinberg Equilibrium

The Hardy-Weinberg equilibrium is the central law of how allele frequencies and genotype frequencies are related in an idealized population. Given the allele frequencies

$$f(G_1) = p$$
$$f(G_2) = q = 1 - p \tag{B.3}$$

During mating, we assume that in an idealized population alleles are combined independently. This leads to the genotype frequencies shown in Table B.1.

Summing up the heterozygous frequencies leads to

Table B.1: Genotype Frequencies under Hardy-Weinberg equilibrium

Alleles	G_1	G_2
G_1	$f(G_1 G_1) = p^2$	$f(G_1 G_2) = p * q$
G_2	$f(G_1 G_2) = p * q$	$f(G_2 G_2) = q^2$

Table B.2: Values for all Genotypes

Variable	Genotype	Values
V_{11}	$G_1 G_1$	a
V_{12}	$G_1 G_2$	d
V_{22}	$G_2 G_2$	-a

$$f(G_1 G_1) = p^2$$
$$f(G_1 G_2) = 2pq$$
$$f(G_2 G_2) = q^2 \tag{B.4}$$

Comparing these expected genotype frequencies in a idealized population under the Hardy-Weinberg equilibrium to what we found for the small example population in Figure B.1, we can clearly say that the small example population is not in Hardy-Weinberg equilibrium.

B.4   Value and Mean

Our goal is still to improve our population at the genetic level. The term improvement implies the need for a quantitative assessment of our trait of interest. Furthermore, we have to be able to associate the genotypes in the population to the quantitative values of our trait.

B.4.1   Genotypic Values

The values V_{ij} to each genotype $G_i G_j$ are assigned as shown in Figure B.2.

The origin of the genotypic values is placed in the middle between the two homozygous genotypes $G_2 G_2$ and $G_1 G_1$. Here we are assuming that G_1 is the favorable allele. This leads to values of $+a$ for genotype $G_1 G_1$ and of $-a$ for genotype $G_2 G_2$. The value of genotype $G_1 G_2$ is set to d and is called dominance deviation. Table B.2 summarizes the values for all genotypes.
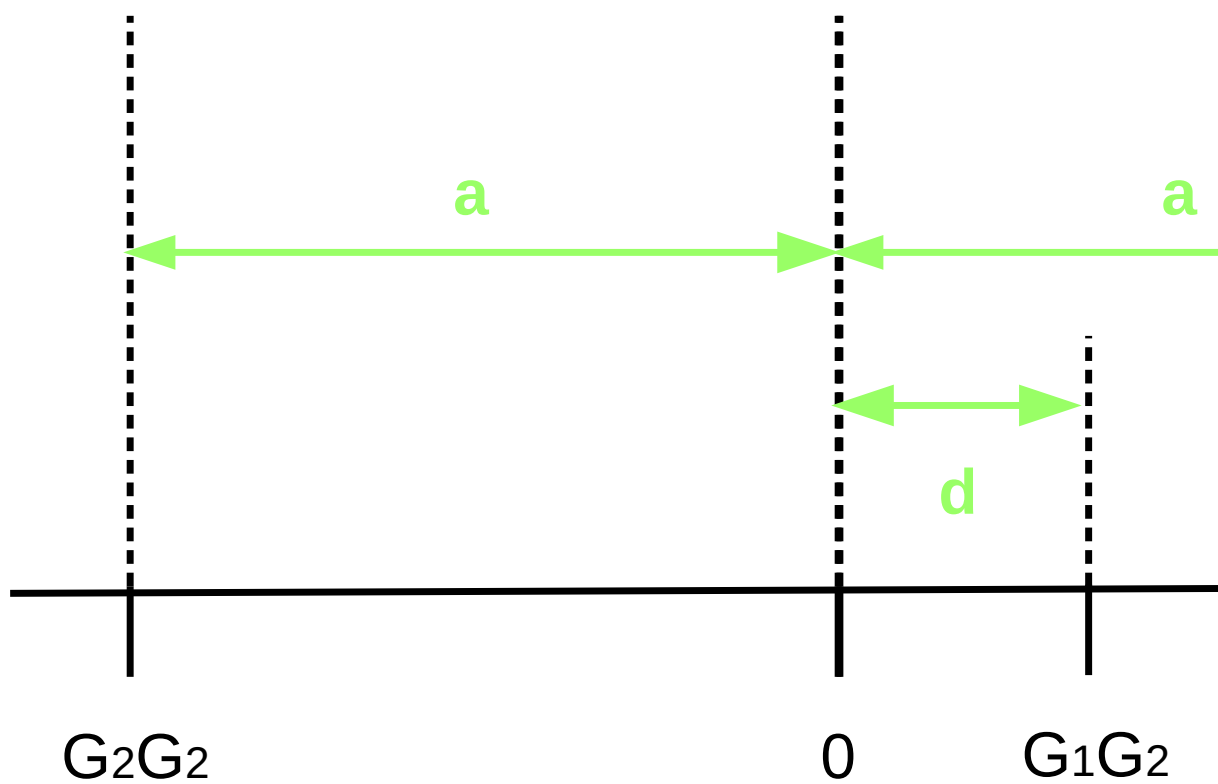
Figure B.2: Genotypic Values

B.4.2   Population Mean

For the complete population, we can compute the **population mean** (μ) of all values at the locus G. This mean corresponds to the expected value and is computed as

$$
\begin{aligned}
\mu &= V_{11} * f(G_1 G_1) + V_{12} * f(G_1 G_2) + V_{22} * f(G_2 G_2) \\
&= a * p^2 + d * 2pq + (-a) * q^2 \\
&= (p - q)a + 2pqd
\end{aligned}
\tag{B.5}
$$

The population mean depends on the values a and d and on the allele frequencies p and q. The larger the difference between p and q the more influence the value a has in μ, because for very different p and q the product $2pq$ is very small. On the other hand, if $p = q = 0.5$, then $\mu = 0.5d$. For loci with $d = 0$, the population mean $\mu = (p - q)a$ and hence, if in addition we have $p = q$, then $\mu = 0$.

B.4.3   Breeding Values

The term **breeding value** is defined as shown in Definition B.1.

Definition B.1 (Breeding Value). The breeding value of an animal i is defined as two times the difference between the mean value of offsprings of animal i and the population mean.

Applying this definition and using the parameters that we have computed so far leads to the following formulas for the breeding value of an animal with a certain genotype.

B.4.3.1   Breeding value for $G_1 G_1$

Assume that we have a given parent S with a genotype $G_1 G_1$ and we want to compute its breeding value. Let us further suppose that our single parent S is mated to a potentially infinite number of animals from the idealized population, then we can deduce the following mean genotypic value for the offspring of parent S.

	Mates of S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_1) = 1$	$f(G_1 G_1) = p$	$f(G_1 G_2) = q$

Because parent S has genotype G_1G_1, the frequency $f(G_1)$ of a G_1 allele coming from S is 1 and the frequency $f(G_2)$ of a G_2 allele is 0. The expected genetic value (μ_{11}) of the offspring of animal S can be computed as

$$\mu_{11} = p * a + q * d \tag{B.6}$$

Applying definition B.1, we can compute the breeding value (BV_{11}) for animal S as shown in equation (B.7) while using the results given by equations (B.6) and (B.5).

$$
\begin{aligned}
BV_{11} &= 2 * (\mu_{11} - \mu) \\
&= 2 \left(pa + qd - [(p - q)a + 2pqd] \right) \\
&= 2 \left(pa + qd - (p - q)a - 2pqd \right) \\
&= 2 \left(qd + qa - 2pqd \right) \\
&= 2 \left(qa + qd(1 - 2p) \right) \\
&= 2q \left(a + d(1 - 2p) \right) \\
&= 2q \left(a + (q - p)d \right) \tag{B.7}
\end{aligned}
$$

Breeding values for parents with genotypes G_2G_2 and G_1G_2 are derived analogously.

B.4.3.2 Breeding value for G_2G_2

First, we determine the expected genotypic value for offsprings of a parent S with genotype G_2G_2

	Mates of parent S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_2) = 1$	$f(G_1G_2) = p$	$f(G_2G_2) = q$

The expected genetic value (μ_{22}) of the offspring of animal S can be computed as

$$\mu_{22} = pd - qa \tag{B.8}$$

The breeding value BV_{22} corresponds to

$$
\begin{aligned}
BV_{22} &= 2 * (\mu_{22} - \mu) \\
&= 2\,(pd - qa - [(p - q)a + 2pqd]) \\
&= 2\,(pd - qa - (p - q)a - 2pqd) \\
&= 2\,(pd - pa - 2pqd) \\
&= 2\,(-pa + p(1 - 2q)d) \\
&= -2p\,(a + (q - p)d)
\end{aligned}
\tag{B.9}
$$

B.4.3.3 Breeding value for $G_1 G_2$

The genotype frequencies of the offsprings of a parent S with a genotype $G_1 G_2$ is determined in the following table.

	Mates of parent S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_1) = 0.5$	$f(G_1 G_1) = 0.5p$	$f(G_1 G_2) = 0.5q$
$f(G_2) = 0.5$	$f(G_1 G_2) = 0.5p$	$f(G_2 G_2) = 0.5q$

The expected mean genotypic value of the offsprings of parent S with genotype $G_1 G_2$ is computed as

$$
\mu_{12} = 0.5pa + 0.5d - 0.5qa = 0.5\,[(p - q)a + d]
\tag{B.10}
$$

The breeding value BV_{12} corresponds to

$$
\begin{aligned}
ZW_{12} &= 2 * (\mu_{12} - \mu) \\
&= 2\,(0.5(p - q)a + 0.5d - [(p - q)a + 2pqd]) \\
&= 2\,(0.5pa - 0.5qa + 0.5d - pa + qa - 2pqd) \\
&= 2\,(0.5(q - p)a + (0.5 - 2pq)d) \\
&= (q - p)a + (1 - 4pq)d \\
&= (q - p)a + (p^2 + 2pq + q^2 - 4pq)d \\
&= (q - p)a + (p^2 - 2pq + q^2)d \\
&= (q - p)a + (q - p)^2 d \\
&= (q - p)\,[a + (q - p)d]
\end{aligned}
\tag{B.11}
$$

B.4.3.4 Summary of Breeding Values

The term $a + (q - p)d$ appears in all three breeding values. We replace this term by α and summarize the results in the following table.

Genotype	Breeding Value
G_1G_1	$2q\alpha$
G_1G_2	$(q - p)\alpha$
G_2G_2	$-2p\alpha$

B.4.4 Allele Substitution

Comparing the genotype G_2G_2 with the genotype G_1G2, one of the differences is in the number of G_1-alleles. G_2G_2 has zero G_1-alleles and G_1G_2 has one G_1-allele. They also have different breeding values. Because the breeding values are to be used to assess the value of a given genotype, any difference between the breeding values BV_{12} and B_{22} can be associated to that additional G_1 allele in G_1G_2 compared to G_2G_2.

The computation of the difference between the breeding value BV_{12} and B_{22} results in

$$
\begin{aligned}
BV12 - BV_{22} &= (q - p)\alpha - (-2p\alpha) \\
&= (q - p)\alpha + 2p\alpha \\
&= (q - p + 2p)\alpha \\
&= (q + p)\alpha \\
&= \alpha
\end{aligned}
\tag{B.12}
$$

The analogous computation can be done by comparing the breeding values BV_{11} and BV_{12}.

$$
\begin{aligned}
BV_{11} - BV_{12} &= & 2q\alpha - (q - p)\alpha \\
&= & (2q - (q - p))\,\alpha \\
&= & \alpha
\end{aligned}
\tag{B.13}
$$

The breeding values themselves depend on the allele frequencies. But the differences between breeding values are linear while replacing G_2 alleles by G_1 alleles. This replacement is also called allele-substitution and the term *alpha* is called **allele-substitution effect**.

B.4.5   Dominance Deviation

When looking at the difference between the genotypic value V_{ij} and the breeding value BV_{ij} for each of the three genotypes, we get the following results.

$$
\begin{aligned}
V_{11} - BV_{11} &= a - 2q\alpha \\
&= a - 2q\left[a + (q-p)d\right] \\
&= a - 2qa - 2q(q-p)d \\
&= a(1 - 2q) - 2q^2 d + 2pqd \\
&= \left[(p-q)a + 2pqd\right] - 2q^2 d \\
&= \mu + D_{11}
\end{aligned}
\tag{B.14}
$$

$$
\begin{aligned}
V_{12} - BV12 &= d - (q-p)\alpha \\
&= d - (q-p)\left[a + (q-p)d\right] \\
&= \left[(p-q)a + 2pqd\right] + 2pqd \\
&= \mu + D_{12}
\end{aligned}
\tag{B.15}
$$

$$
\begin{aligned}
V_{22} - BV_{22} &= -a - (-2p\alpha) \\
&= -a + 2p\left[a + (q-p)d\right] \\
&= \left[(p-q)a + 2pqd\right] - 2p^2 d \\
&= \mu + D_{22}
\end{aligned}
$$

The difference all contain the population mean μ plus a certain deviation. This deviation term is called **dominance deviation**.

B.4.6   Summary of Values

The following table summarizes all genotypic values all breeding values and the dominance deviations.

Genotyp $G_i G_j$	genotypic value V_{ij}	Breeding Value BV_{ij}	Dominance Deviation D_{ij}
$G_1 G_1$	a	$2q\alpha$	$-2q^2 d$
$G_1 G_2$	d	$(q-p)\alpha$	$2pqd$
$G_2 G_2$	$-a$	$-2p\alpha$	$-2p^2 d$

The formulas in the above shown table assume that G_1 is the favorable allele with frequency $f(G_1) = p$. The allele frequency of G_2 is $f(G_2) = q$. Since we have a bi-allelic locus $p + q = 1$.

Based on the definition of dominance deviation, the genotypic values V_{ij} can be decomposed into the components population mean (μ), breeding value (BV_{ij}) and dominance deviation (D_{ij}) according to equation (B.16).

$$V_{ij} = \mu + BV_{ij} + D_{ij} \tag{B.16}$$

Taking expected values on both sides of equation (B.16) and knowing that the population mean μ was defined as the expected value of the genotypic values in the population, i.e. $E[V] = \mu$, it follows that the expected values of both the breeding values and the dominance deviations must be 0. More formally, we have

$$\begin{aligned} E[V] &= E[\mu + BV + D] \\ &= E[\mu] + E[BV] + E[D] \\ &= \mu \end{aligned} \tag{B.17}$$

From the last line in equation (B.17), it follows that $E[BV] = E[D] = 0$. This also shows that both breeding values and dominance deviations are defined as deviation from a given mean.

B.5 Variances

The population mean μ and derived from that the breeding values were defined as expected values. Their main purpose is to assess the state of a given population with respect to a certain genetic locus and its effect on a phenotypic trait of interest. One of our primary goals in livestock breeding is to improve the populations at the genetic level through the means of selection and mating. Selection of potential parents that produce offspring that are closer to our breeding goals is only possible, if the selection candidates show a certain level of variation in the traits that we are interested in. In populations where there is no variation which means that all individuals are exactly at the same level, it is not possible to select potential parents for the next generation.

In statistics the measure that is most often used to assess variation in a certain population is called **variance**. For any given discrete random variable X the variance is defined as the second central moment of X which is computed as shown in equation (B.18).

$$Var[X] = \sum_{x_i \in \mathcal{X}} (x_i - \mu_X)^2 * f(x_i) \tag{B.18}$$

where $\mathcal{X}$:   set of all possible x-values

          $f(x_i)$ probability that x assumes the value of x_i

          μ_X expected value $E\left[X\right]$ of X

In this section we will be focusing on separating the obtained variances into different components according to their causative sources. Applying the definition of variance given in equation (B.18) to the genotypic values V_{ij}, we obtain the following expression.

$$
\begin{aligned}
\sigma_G^2 = Var\left[V\right] = & (V_{11} - \mu)^2 * f(G_1 G_1) \\
+ \; & (V_{12} - \mu)^2 * f(G_1 G_2) \\
+ \; & (V_{22} - \mu)^2 * f(G_2 G_2)
\end{aligned}
\tag{B.19}
$$

where $\mu = (p - q)a + 2pqd$ the population mean.

Based on the decomposition of the genotypic value V_{ij} given in (B.16), the difference between V_{ij} and μ can be written as the sum of the breeding value and the dominance deviation. Then σ_G^2 can be written as

$$
\begin{aligned}
\sigma_G^2 = Var\left[V\right] = & (BV_{11} + D_{11})^2 * f(G_1 G_1) \\
+ \; & (BV_{12} + D_{12})^2 * f(G_1 G_2) \\
+ \; & (BV_{22} + D_{22})^2 * f(G_2 G_2)
\end{aligned}
\tag{B.20}
$$

Inserting the expressions for the breeding values BV_{ij} and for the dominance deviation D_{ij} found earlier and simplifying the equation leads to the result in (B.21). A more detailed derivation of σ_G^2 is given in the appendix (B.8) of this chapter.

$$
\begin{aligned}
\sigma_G^2 &= 2pq\alpha^2 + (2pqd)^2 \\
&= \sigma_A^2 + \sigma_D^2
\end{aligned}
\tag{B.21}
$$

The formula in equation (B.21) shows that σ_G^2 consists of two components. The first component σ_A^2 is called the **genetic additive variance** and the second component σ_D^2 is termed **dominance variance**. As shown in equation (B.25) σ_A^2 corresponds to the variance of the breeding values. Because we have already seen that the breeding values are additive in the number of favorable alleles, σ_A^2 is called genetic additive variance. Because σ_D^2 corresponds to the variance of the dominance deviation effects (see equation (B.27)) it is called dominance variance.

B.6 Extension To More Loci

When only a single locus is considered, the genotypic values (V_{ij}) can be decomposed according to equation (B.16) into population mean, breeding value and dominance deviation. When a genotype refers to more than one locus, the genotypic value may contain an additional deviation caused by non-additive combination effects.

B.6.1 Epistatic Interaction

Let V_A be the genotypic value of locus A and V_B denote the genotypic value of a second locus B, then the total aggregate genotypic value V attributed to both loci A and B can be written as

$$V = V_A + V_B + I_{AB} \tag{B.22}$$

where I_{AB} is the deviation from additive combination of these genotypic values. When computing the population mean earlier in this chapter, we assumed that I was zero for all combinations of genotypes. If I is not zero for any combination of genes at different loci, those genes are said to **interact** with each other or to exhibit **epistasis**. The deviation I is called interaction deviation or epistatic deviation. If I is zero, the genes are called to act additively between loci. Hence *additive action* may mean different things. When referring to one locus, it means absence of dominance. When referring to different loci, it means absence of epistasis.

Interaction between loci may occur between pairs or between higher numbers of different loci. The complex nature of higher order interactions, i.e., interactions between higher number of loci does not need to concern us. Because in the aggregate genotypic value V, interaction deviations of all sorts are treated together in an overall interaction deviation I. This leads to the following generalized form of decomposing the overall aggregate genotype V for the case of multiple loci affecting a certain trait of interest.

$$V = \mu + U + D + I \tag{B.23}$$

where U is the sum of the breeding values attributable to the separate loci and D is the sum of all dominance deviations. For our purposes in livestock breeding where we want to assess the genetic potential of a selection candidate to be a parent of offspring forming the next generation, the **breeding value** is the most important quantity. The breeding value is of primary importance because a given parent passes a random sample of its alleles to its offspring. We have seen in section B.4.4 that breeding values are additive in the number of favorable alleles. Hence the more favorable alleles a given parent passes to its offspring the higher the breeding value of this parent.

On the other hand, the dominance deviation measures the effect of a certain genotype occurring in an individual and the interaction deviation estimates the effects of combining different genotypes at different loci in the genome. But because parents do not pass complete genotypes nor do they pass stretches of DNA with several loci, but only a random collection of its alleles, it is really the breeding value that is of primary importance in assessing the genetic potential of a given selection candidate.

B.6.2 Interaction Variance

If genotypes at different loci show epistatic interaction effects as described in section B.6.1, the interactions give rise to an additional variance component called V_I, which is the variance of interaction deviations. This new variance component V_I can be further decomposed into sub-components. The first level of sub-components is according to the number of loci that are considered. Two-way interactions involve two loci, three-way interactions consider three loci and in general n-way interactions arise from n different loci. The next level of subdivision is according to whether they include additive effects, dominance deviations or both.

In general it can be said that for practical purposes, interaction effects explain only a very small amount of the overall variation. As already mentioned in section @ref(#epistatic-interaction) for livestock breeding, we are mostly interested in the additive effects. This is also true when looking at the variance components. Hence dominance variance and variances of interaction deviations are not used very often in practical livestock breeding application.

B.7 Genetic Models

In this chapter, we have seen how to model the genetic basis of a quantitative trait when a single locus affects the trait of interest. We call this a single-locus model. When several loci have an effect on a certain trait, then we talk about a **polygenic model**. Letting the number of loci affecting a certain phenotype tend to infinity, the resulting model is called **infinitesimal model**.

From a statistical point of view, the breeding values in an infinitesimal model are considered as a random effect with a known distribution. Due to the central limit theorem, this distribution is assumed to be a normal distribution. The central limit theorem says that the distribution of any sum of a large number of very small effects converges to a normal distribution. For our case where a given trait of interest is thought to be influenced by a large number of genetic loci each having a small effect, the sum of the breeding values of all loci together can be approximated by a normal distribution. Figure (B.3) shows the distribution for a sum of 10, 100 and 1000 components each. The histograms show a better

approximation to the normal distribution the larger the number of components considered in the sum.
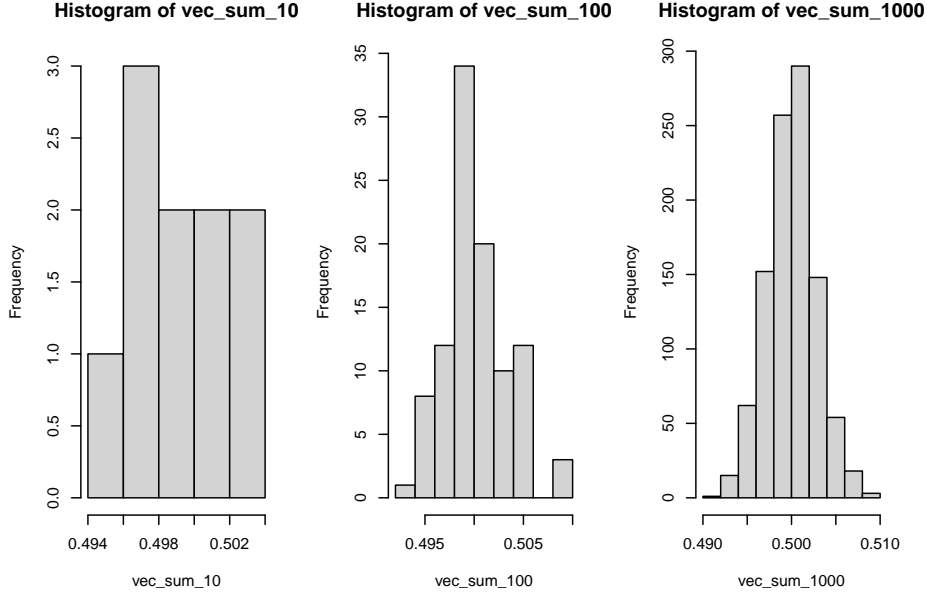

Figure B.3: Distribution of Sums of Different Numbers of Components

B.7.1   Model Usage In Routine Evaluations

Traditional prediction of breeding values before the introduction of genomic selection is based on the infinitesimal model. When genomic selection was introduced which takes into account the information from a large number of loci, genomic breeding values are estimated using a polygenic model.

B.8   Appendix: Derivations

This section shows how the genetic variance in equation (B.21) is computed.

$$
\begin{aligned}
\sigma_G^2 &= (BV_{11} + D_{11})^2 * p^2 \\
&+ (BV_{12} + D_{12})^2 * 2pq \\
&+ (BV_{22} + D_{22})^2 * q^2 \\
&= (2q\alpha - 2q^2 d)^2 * p^2 \\
&+ ((q - p)\alpha + 2pqd)^2 * 2pq \\
&+ (-2p\alpha - 2p^2 d)^2 * q^2 \\
&= (4q^2\alpha^2 - 8q^3 d\alpha + 4q^4 d^2) * p^2 \\
&+ (q^2\alpha^2 - 2pq\alpha^2 + p^2\alpha^2 - 4(q - p)pqd\alpha + 4p^2 q^2 d^2) * 2pq \\
&+ (4p^2\alpha^2 + 8p^3 d\alpha + 4p^4\alpha^2) * q^2 \\
&= 4p^2 q^2\alpha^2 - 8p^2 q^3 d\alpha + 4p^2 q^4 d^2 \\
&+ 2pq^3\alpha^2 - 4p^2 q^2\alpha^2 + 2p^3 q\alpha^2 \\
&- 8p^3 q^2 d\alpha + 8p^2 q^3 d\alpha + 8p^3 q^3 d^2 \\
&+ 4p^2 q^2\alpha^2 + 8p^3 q^2 d\alpha + 4p^4 q^2 d^2 \\
&= 4p^2 q^2\alpha^2 + 4p^2 q^4 d^2 \\
&+ 2pq^3\alpha^2 + 2p^3 q\alpha^2 \\
&+ 8p^3 q^3 d^2 \\
&+ 4p^4 q^2 d^2 \\
&= 2pq\alpha^2 (p^2 + 2pq + q^2) \\
&+ (2pqd)^2 (p^2 + 2pq + q^2) \\
&= 2pq\alpha^2 + (2pqd)^2 \\
&= \sigma_A^2 + \sigma_D^2 \tag{B.24}
\end{aligned}
$$

From the last two lines of (B.24) it follows that $\sigma_A^2 = 2pq\alpha^2$ and $\sigma_D^2 = (2pqd)^2$. It can be shown that σ_A^2 corresponds to the squared breeding values times the associated genotype frequencies. Because the expected values of the breeding values is zero, σ_A^2 is equivalent to the variance of the breeding values.

$$
\begin{aligned}
\sigma_A^2 &= Var[BV] = (BV_{11} - E[BV])^2 * f(G_1 G_1) + (BV_{12} - E[BV])^2 * f(G_1 G_2) + (BV_{22} - E[BV])^2 * \\
&= BV_{11}^2 * f(G_1 G_1) + BV_{12}^2 * f(G_1 G_2) + BV_{22}^2 * f(G_2 G_2) \\
&= (2q\alpha)^2 * p^2 + ((q - p)\alpha)^2 * 2pq + (-2p\alpha)^2 * q^2 \\
&= 4p^2 q^2\alpha^2 + (q^2\alpha^2 - 2pq\alpha^2 + p^2\alpha^2) * 2pq + 4p^2 q^2\alpha^2 \\
&= 8p^2 q^2\alpha^2 + 2pq^3\alpha^2 - 4p^2 q^2\alpha^2 + 2p^3 q\alpha^2 \\
&= 4p^2 q^2\alpha^2 + 2pq^3\alpha^2 + 2p^3 q\alpha^2 \\
&= 2pq\alpha^2 (2pq + q^2 + p^2) \\
&= 2pq\alpha^2 \tag{B.25}
\end{aligned}
$$

In the above derivation in (B.25) of the variance of the breeding values, we were using the fact that the expected value $E[BV] = 0$. This can be shown more formally as follows

$$
\begin{aligned}
E[BV] &= BV_{11} * f(G_1 G_1) + BV_{12} * f(G_1 G_2) + BV_{22} * f(G_2 G_2) \\
&= 2q\alpha * p^2 + (q - p)\alpha * 2pq + (-2p\alpha) * q^2 \\
&= 2p^2 q\alpha + 2pq^2\alpha - 2p^2 q\alpha - 2pq^2\alpha \\
&= 0
\end{aligned}
\tag{B.26}
$$

Similarly to (B.25) we can show that σ_D^2 corresponds to the squared dominance deviations times the frequencies of the corresponding genotypes. That is the reason why σ_D^2 is called dominance variance.

$$
\begin{aligned}
\sigma_D^2 &= D_{11}^2 * f(G_1 G_1) + D_{12}^2 * f(G_1 G_2) + D_{22}^2 * f(G_2 G_2) \\
&= (-2q^2 d)^2 * p^2 + (2pqd)^2 * 2pq + (-2p^2 d)^2 * q^2 \\
&= 4p^2 q^4 d^2 + 8p^3 q^3 d^2 + 4p^4 q^2 d^2 \\
&= 4p^2 q^2 d^2 \left(q^2 + 2pq + p^2\right) \\
&= 4p^2 q^2 d^2
\end{aligned}
\tag{B.27}
$$

Bibliography

3blue1brown. What are vectors, 2016. URL https://youtu.be/fNk{_}zzaMoSs.

M Adibuzzaman, P DeLaurentis, J Hill, and B D Benneyworth. Big data in healthcare - the promises, challenges and opportunities from a research perspective: A case study with a model database. *AMIA Annu Symp Proc*, 2017:384–392, 2017. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pubmed/29854102.

M. L. Chizzotti, F. S. Machado, E. E.L. Valente, L. G.R. Pereira, M. M. Campos, T. R. Tomich, S. G. Coelho, and M. N. Ribas. Technical note: Validation of a system for monitoring individual feeding behavior and individual feed intake in dairy cattle. *Journal of Dairy Science*, 98(5):3438–3442, 2015. ISSN 15253198. doi: 10.3168/jds.2014-8925. URL http://dx.doi.org/10.3168/jds.2014-8925.

Bartosz Czech, Magdalena Fraszczak, Magda Mielczarek, and Joanna Szyda. Identification and annotation of breed-specific single nucleotide polymorphisms in Bos taurus genomes. *PLoS ONE*, 13(6):1–9, 2018. ISSN 15498328. doi: 10.1109/TCSI.2014.2341116. URL https://doi.org/10.1371/journal.pone.0198419.

D. S. Falconer and Trudy F. C. Mackay. *Introduction to Quantitative Genetics*. Addison Wesley Longman Limited, Essex, 4 edition, 1996. ISBN 0582-24302-5.

Sasha Issenberg. How Obama's Team Used Big Data to Rally Voters. *MIT Technology Review*, 116(1):38–49, 2013.

John R Mashey. Big Data ... and the Next Wave of InfraStress, 1998. URL http://static.usenix.org/event/usenix99/invited{_}talks/mashey.pdf.

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.

G E Moore. Cramming more components onto integrated circuits. *Electronics*, 38, 1965.

Kaspar Nipp and Daniel Stoffer. *Lineare Algebra*. vdf Hochschulverlag, Zurich, 5 edition, 2002. ISBN 3 7281 2818 X.

Sameer D. Pant, Flavio S. Schenkel, Chris P. Verschoor, and Niel A. Karrow. Use of breed-specific single nucleotide polymorphisms to discriminate between Holstein and Jersey dairy cattle breeds. *Animal Biotechnology*, 23(1):1–10, 2012. ISSN 10495398. doi: 10.1080/10495398.2012.636224. URL https://doi. org/10.1080/10495398.2012.636224.

R Core Team. R: A language and environment for statistical computing., 2018. ISSN 16000706. URL https://www.r-project.org/.

L. R. Schaeffer. Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet*, 123(123):218–223, 2006. ISSN 09312668. doi: 10.1111/j. 1439-0388.2006.00595.x.

S R Searle. *Linear Models*. John Wiley & Sons, New York, wiley clas edition, 1971. ISBN 0-471-18499-3.

G. E. Seidel, Jr. Brief introduction to whole-genome selection in cattle using single nucleotide polymorphisms. *Reproduction Fertility and Development*, 22(1):138–144, 2010. ISSN 1031-3613. doi: 10.1071/RD09220. URL https: //doi.org/10.1071/RD09220.

Contributors Wikipedia. Big data - Wikipedia, 2019. URL https://en.wikipedia. org/w/index.php?title=Big{_}data{&}oldid=881938730.