Chapter 3

Genomic Best Linear Unbiased Prediction (GBLUP)

Wie in der Einführung im Kapitel 1.3.2 beschrieben wurde, werden genomische Zuchtwerte häufig in zwei Schritten geschätzt. Nicht so bei der Methode GBLUP. GBLUP ist ein sogenanntes **Single-Step** Verfahren, in welchem die genomischen Zuchtwerte in einem Schritt geschätzt werden.

Schon in der traditionellen Zuchtwertschätzung mit dem BLUP-Tiermodell werden mehr unbekannten Parameter (Zuchtwerte) geschätzt oder vorhergesagt, als dass wir Beobachtungen im Datensatz haben. Dies ist möglich, da wir die Covarianz-Matrix der zu schätzenden zufälligen Zuchtwerte aufgrund der als bekannt angenommenen additiv-genetische Verwandtschaftsmatrix A modellieren. Über die Verknüpfung der zu schätzenden Zuchtwerte aufgrund der additiv-genetischen Verwandtschaftsmatrix werden Zuchtwerte von Tieren ohne Beobachtungen als Funktionen der Beobachtung von verwandten Tieren geschätzt. Somit eignen sich BLUP-basierte Schätzverfahren grundsätzlich für die Art von Datenanalysen, wie wir sie in der Schätzung von genomischen Zuchtwerten vorfinden. Daher ist es naheliegend ein BLUP-basiertes Schätzverfahren auch für die Schätzung der genomischen Zuchtwerte zu verwenden. Wie das gemacht wird, soll in diesem Kapitel beschrieben werden.

3.1 Einführung in GBLUP

Die Methode GBLUP verwendet sogenannte **genomische Verwandtschaftsbeziehungen** zur Schätzung des genetischen Potentials von Individuen. Die genomische Verwandtschaft wird aufgrund von DNA-Informationen geschätzt. Die genomische Verwandtschaftsmatrix G definiert die Kovarianz zwischen Individuen aufgrund von Ähnlichkeiten auf dem Niveau der genomischen Information. Diese Definition der Kovarianz steht im Gegensatz zur Kovarianz aufgrund der erwarteten Ähnlichkeit aufgrund von Pedigrees in der traditionellen Verwandtschaftsmatrix A.

3.2 DNA Marker

Die Verfügbarkeit von Markerkarten in genügender Dichte verteilt über das ganze Genom, ermöglicht die Schätzung des genetischen Potentials (in Form von genomischen Zuchtwerten) von Individuen basierend auf der beobachteten Markerinformation. Das aktuell am weitesten verbreitete Markermodell wird als `Single Nucleotide Polymorphism` (SNP) bezeichnet. Bei den SNPs handelt es sich um Positionen im Genom, an welchen Unterschiede zwischen Individuen einer Population bei einer einzelnen DNA-Base auftreten. Nehmen wir an, dass Genorte (QTL), welche wichtig sind für die Expression von interessanten phänotypischen Eigenschaften, sich in der Nähe von beobachteten SNPs befinden, können wir aufgrund von Kopplung-

sungleichgewichten zwischen SNPs und QTL eine statistische Beziehung zwischen phänotypischen Werten und SNP-Genotypen modellieren.

Für die Schätzung der SNP-Effekte auf phänotypische Ausprägungen stehen verschiedene statistische Verfahren zur Verfügung. Als einfachste Methode können wir die Regression auf einzelne SNP-Marker bezeichnen. Das Problem der Single-Marker Regression liegt in der im Vergleich zur Anzahl der Beobachtungen sehr grossen Anzahl an SNP-Markern. Eine Lösung dieses Problems besteht darin, dass die SNP-Effekte im linearen Modell als zufällige Effekte aufgefasst werden. Dadurch wird das einfache lineare Regressionsmodell in ein lineares gemischtes Modell verwandelt. Aus der traditionellen Zuchtwertschätzung ((Hofer, 1990) und (von Rohr, 2016)) kennen wir die Mischmodellgleichungen, welche uns Schätzwerte für fixe und zufällige Effekte mit BLUE- und BLUP-Eigenschaften liefern. Bei diesem Verfahren werden aber allen SNPs der gleiche Varianzanteil zugesprochen. Die totale genetische Varianz wird somit auf alle SNPs gleichmässig aufgeteilt. Bayes'sche Methoden, wie sie in Kapitel 5 beschrieben sind, erlauben es verschiedenen genomischen Regionen verschiedene Varianzanteile zuzuordnen.

3.3 Markerinformationen in BLUP-Verfahren

Parameterschätzverfahren, welche zu Schätzungen für fixe Effekte mit `Best Linear Unbiased Estimation` (BLUE)-Eigenschaften und zu Vorhersagen von zufälligen Effekten mit `Best Linear Unbiased Prediction` (BLUP)-Eigenschaften führen werden hier kurz als **BLUP-Verfahren** bezeichnet. Die traditionelle Zuchtwertschätzung basiert auf Pedigreeinformationen zur Schätzung der Covarianz zwischen zufälligen Effekten (Zuchtwerten) von verwandten Tieren. Die Covarianz kann aber auch aufgrund von Markerinformation geschätzt werden. Die aufgrund von SNPs erstellte Matrix wird als `Genomic Relationship Matrix` (GRM) bezeichnet.

Wir besprechen hier zwei verschiedene Ansätze

1. `Ridge Regression` (RR)-BLUP
2. `Genomic BLUP` (GBLUP)

3.3.1 Ridge Regression (RR) BLUP

Diese Methode wurde von (Meuwissen et al., 2001) und (Habier et al., 2007) untersucht. Dabei haben die Autoren das folgende Modell angenommen.

$$y = 1_n\mu + Wq + e \tag{3.1}$$

wobei

y	Vektor der Länge n mit Beobachtungen
μ	allgemeines Mittel
q	Vektor der Länge m mit zufälligen SNP-Effekten
W	Matrix der Dimension $n \times m$, welche SNP-Genotypen codiert
e	Vektor der Länge n mit zufälligen Resteffekten

Die Genotypen an jedem SNP-Locus in Matrix W werden mit 0, 1 und 2 codiert. Diese Codes repräsentieren die Anzahl an SNP-Allelen mit positiver Wirkung. Die SNP-Effekte werden als zufällig betrachtet. Das Modell in (3.1) ist also ein gemischtes lineares Modell. Die unbekannten Parameter werden mit Hilfe von Mischmodellgleichungen geschätzt. Der geschätzte genomische Zuchtwert entspricht der Summe der SNP-Effekte über alle SNP-Loci. Die durch die SNP-Effekte erklärte Varianz ist gleich $WW^T\sigma_q^2$. Die Varianz der Resteffekte wird als $I\sigma_e^2$ angenommen. Somit entspricht die Covarianz zwischen Beobachtungen $WW^T\sigma_q^2 + I\sigma_e^2$. In RR-BLUP wird die Varianz an allen SNPs als konstant angenommen.

3.3.2 Genomic BLUP (GBLUP)

Die zweite Methode, welche genomische Information berücksichtigt verwendet die genomische Verwandtschaftsmatrix G anstelle der additiv genetischen Verwandtschaftsmatrix A in einem gemischten linearen Modell. Dieser Ansatz wird als **gBLUP** bezeichnet. Das Modell zur Umsetzung von gBLUP lautet

$$y = Xb + Zg + e \tag{3.2}$$

wobei

y	Vektor der Länge n mit Beobachtungen
b	Vektor der Länge r mit fixen Effekten
X	Inzidenzmatrix zur Verknüpfung von b und y
g	Vektor der Länge t mit zufälligen genetischen Effekten
Z	Inzidenzmatrix zur Verknüpfung von g und y
e	Vektor der Länge n mit zufälligen Resteffekten

Der Vektor g enthält zufällige genetische Effekte für alle Tiere, welche typisiert sind. Diese Tiere können Beobachtungen aufweisen oder nicht. Die Tiere mit Beobachtungen und Typisierungsergebnisse werden allgemein als Trainings- oder Referenzpopulation bezeichnet. Die Tiere ohne phänotypische Beobachtung mit Typisierungsergebnissen bilden das Testset, für welche die genomischen Zuchtwerte geschätzt werden sollen. Die Varianz $var(g) = G * \sigma_g^2$ wobei G der genomischen Verwandtschaftsmatrix entspricht. Die zufälligen Resteffekte werden als unabhängig angenommen mit der Covarianzmatrix $var(e) = I * \sigma_e^2$.

GBLUP hat drei wichtige Vorteile im Vergleich zu RR-BLUP.

1. Die Dimension der genetischen Effekte in GBLUP beträgt $n \times n$, wobei n die Anzahl Tiere sind. In RR-BLUP beträgt diese Dimension $m \times m$, wobei m der Anzahl an SNP-Markern entspricht. Somit ist GBLUP effizienter im Hinblick auf Rechenresourcen.
2. Die Genauigkeiten der genomischen Zuchtwerte können bei GBLUP analog zu den Genauigkeiten der Zuchtwerte im BLUP-Tiermodell berechnet werden.
3. GBLUP kann mit Pedigree-basierten Informationen zu den sogenannten **single step** Verfahren kombiniert werden (Misztal et al., 2009).

In (Fernando et al., 2016) haben die Autoren gezeigt, dass wenn die Anzahl typisierter Tiere grösser ist als die Anzahl SNP-Loci, die GBLUP-Modelle gewisse Nachteile aufweisen, welche bei den RR-BLUP-Modellen nicht existieren.

3.4 Genomische Verwandtschaftsmatrix (GRM)

Die Covarianz zwischen den gentischen Effekten g im Modell (3.2) wird über die genomische Verwandtschaftsmatrix G ausgedrückt. Analog zum BLUP-Tiermodell, soll die Covarianz der genetischen Effekte als Produkt der genomischen Verwandtschaftsmatrix G mal die Varianzkomponente σ_g^2 dargestellt werden.

3.4.1 Herleitung der GRM

Als erstes stellt sich die Frage, wie wir die genetischen Effekte g überhaupt definieren sollen. Auf dieser Definition von g aufbauend können wir uns anschliessend überlegen, wie die Matrix G aufgestellt werden kann. Die folgenden Eigenschaften für die genetischen Effekte g und für die GRM G sollen gelten.

1. Die genetischen Effekte g sollen der Summe aller SNP-Effekte q entsprechen.
2. Die genetischen Effekte g sollen nicht als Absolute Werte sondern analog zu den Zuchtwerten aus dem BLUP-Tiermodell als Abweichungen von einer festgelegten Basis definiert sein, das heisst der Erwartungswert $E[g] = 0$.
3. Wie schon erwähnt soll die Covarianz der genetischen Effekte g dem Produkt aus GRM G und der Varianzkomponente σ_g^2 entsprechen, d.h. $var(g) = G * \sigma_g^2$.

4. Die GRM G soll ähnlich wie die additive Verwandtschaftsmatrix A aussehen, d.h. die Diagnoalelemente sollen um 1 liegen und auf der Offdiagonalen sollen hohe Werte mit genetisch ähnlichen Tieren assoziiert werden.

Als Informationsquellen für die Definition von g und zum Aufstellen der GRM G haben wir die SNP-Markerinformationen zur Verfügung. Das hier vorgestellte Material basiert auf den Arbeiten von (VanRaden, 2008) und von (Gianola et al., 2009). In den folgenden Unterabschnitten wollen wir die Konsequenzen der oben aufgelisteten Eigenschaften analysieren und daraus die GRM G aufstellen.

3.4.2 Genetische Effekte als Summe der SNP-Effekte

Basierend auf der SNP-Markerinformation können wir Effekte für die einzelnen SNP-Marker schätzen. Wir nehmen hier also an, dass wir den Vektor q kennen. Formell bedeutet die Eigenschaft, dass die genetischen Effekte g als Summe der SNP-Effekte q dargestellte werden können, dass es eine Matrix U gibt, für welche gilt, dass

$$g = U * q \tag{3.3}$$

wobei an dieser Stelle die Matrix U noch unbekannt ist. Wir werden diese anhand der nächsten Eigenschaft bestimmten.

3.4.3 Genetische Effekte als Abweichungen

Die genetischen Effekte sollen, analog zu den Zuchtwerten aus dem BLUP-Tiermodell als Abweichungen von einer festgelegten Basis definiert werden. Die Basis stellt den Nullpunkt der genetischen Effekte dar. Somit kommen die einzelnen Genetischen Effekte g_i für Tier i aus einer Verteilung mit Erwartungswert $E[g_i] = 0$. Die Frage ist nun, wie muss die Matrix U aussehen, dass unabhängig vom Vektor q der Erwartungswert der Komponenten von g_i gleich null ist.

Betrachten wir die Zufallsvariable w mit den SNP-Genotypencodes in Matrix W im RR-BLUP-Modell (3.1) und nehmen wir an, das Hardy-Weinberg-Gleichgewicht gelte für alle SNP-Genorte, dann sind folgende Realisierungswerte für w möglich

$$w = \begin{cases} 0 & \text{mit Wahrscheinlichkeit} \quad (1-p)^2 \\ 1 & \text{mit Wahrscheinlichkeit} \quad 2p(1-p) \\ 2 & \text{mit Wahrscheinlichkeit} \quad p^2 \end{cases} \tag{3.4}$$

Der Erwartungswert der Zufallsvariablen w entspricht

$$E[w] = 0 * (1-p)^2 + 1 * 2p(1-p) + 2 * p^2 = 2p \tag{3.5}$$

Die Matrix U aus (3.3) berechnen wir als Differenz zwischen den Matrizen W und P. Die Matrix P besteht aus Kolonnenvektoren $1_n 2p_j$ der Länge n für die entsprechenden SNPs $j = 1, \ldots, m$. Für einen bestimmten SNP-Locus j folgen aufgrund der Definition der Matrix U die folgenden Codierungen der SNP-Genotypen. Dabei nehmen wir an, dass das Allel $(G_2)_j$ einerseits mit Frequenz p_j auftritt und andererseits dasjenige Allel mit der positiven Wirkung ist. Die nachfolgende Tabelle gibt eine Übersicht über die genotypischen Werte und der Codierung der Genotypen in der Matrix U für die drei auftretenden Genotypen am SNP-Locus j.

Genotypen	Genotypischer Wert	Codierung in Matrix U
$(G_1G_1)_j$	$-2p_jq_j$	$-2p_j$
$(G_1G_2)_j$	$(1-2p_j)q_j$	$1-2p_j$
$(G_2G_2)_j$	$(2-2p_j)q_j$	$2-2p_j$

Hier ist es wichtig zu bemerken, dass q_j für die j-te Komponente im Vektor q aus dem Model (3.3) steht.

Unter der Annahme des Hardy-Weinberg-Gleichgewichts für die Genotypen am SNP-Locus j können wir den Erwartungswert für den genetischen Effekt g am SNP-Locus j berechnen. Der Erwartungswert wird über alle Komponenten des Vektors g der genetischen Effekte berechnet.

$$\begin{aligned} E[g]_j &= \left[(1-p_j)^2(-2p_j) + 2p_j(1-p_j)(1-2p_j) + p_j^2(2-2p_j)\right] q_j \\ &= \left[(1-2p_j+p_j^2)(-2p_j) + (2p_j-2p_j^2)(1-2p_j) + p_j^2(2-2p_j)\right] q_j \\ &= \left[-2p_j + 4p_j^2 - 2p_j^3 + 2p_j - 4p_j^2 - 2p_j^2 + 4p_j^3 + 2p_j^2 - 2p_j^3\right] q_j \\ &= 0 \end{aligned} \tag{3.6}$$

Die gleiche Herleitung lässt sich für jeden SNP Locus $j = 1, \ldots, m$ machen. Somit ist der Erwartungswert über alle genetischen Komponenten g_i gleich null, unabhängig von den SNP-Effekten q_j am Locus j. Somit haben wir die Matrix U, welche aus den SNP-Effekten q in (3.1) die genetischen Effekte g in (3.2) berechnet, gefunden.

3.4.4 Alternative Codierung

Anstelle der verwendeten Codierungen von 0, 1 und 2 für die SNP-Genotypen können diese auch mit -1, 0 und 1 codiert werden. Diese Codierung wird in (VanRaden, 2008) zum Aufstellen der GRM G verwendet. Auch die Autoren des Papers (Gianola et al., 2009) verwenden diese alternative Codierung. Am prinzipiellen Vorgehen ändert sich aber dadurch nichts. Nur die verwendeten Werte in den Matrizen sind anders. Die Matrix W unter der alternativen Codierung besteht aus Werten -1, 0 und 1. Die Zufallsvariable w, welche wir in (3.4) definiert und zur Bestimmung der Matrix P verwendet hatten, sieht unter der alternativen Codierung wie folgt aus.

$$w = \begin{cases} -1 & \text{mit Wahrscheinlichkeit} \quad (1-p)^2 \\ 0 & \text{mit Wahrscheinlichkeit} \quad 2p(1-p) \\ 1 & \text{mit Wahrscheinlichkeit} \quad p^2 \end{cases} \tag{3.7}$$

Der Erwartungswert für w unter der alternativen Codierung ist dann

$$E[w] = (-1) * (1-p)^2 + 0 * 2p(1-p) + 1 * p^2 = -1 + 2p - p^2 + p^2 = 2p - 1 = 2(p - 0.5) \tag{3.8}$$

Die Kolonne j der Matrix P besteht aus dem Kolonnenvektoren $1_n * 2(p_j - 0.5)$. Die Matrix U (welche der Matrix Z in (VanRaden, 2008) entspricht) kann wird berechnet als

$$U = W - P \tag{3.9}$$

Die Elemente der Matrix U der Genotypen am SNP-Locus j unter der alternativen Codierung lauten

Genotypen	Codierung in Matrix U
$(G_1G_1)_j$	$-1 - 2(p_j - 0.5)$
$(G_1G_2)_j$	$-2(p_j - 0.5)$
$(G_2G_2)_j$	$1 - 2(p_j - 0.5)$

Der Erwartungswert der Komponenten der genetischen Effekte g aufgrund des SNP-Locus j beträgt dann

$$\begin{aligned} E[g]_j &= \left[(1-p_j)^2(-1-2(p_j-0.5)) + 2p_j(1-p_j)(-2(p_j-0.5)) + p_j^2(1-2(p_j-0.5))\right] q_j \\ &= \left[-2\,p_j^3 + 4\,p_j^2 - 2\,p_j + 4\,p_j^3 - 6\,p_j^2 + 2\,p_j - 2\,p_j^3 + 2\,p_j^2\right] q_j \\ &= 0 \end{aligned} \tag{3.10}$$

3.4.5 Varianz der genetischen Effekte

Aufgrund der postulierten Eigenschaften soll die Varianz $var(g)$ der genetischen Effekte berechnet werden als das Produkt der GRM G mal die gemeinsame Varianzkomponente σ_g^2. Als Formel geschrieben lautet der Ausdruck für $var(g)$

$$var(g) = G * \sigma_g^2. \tag{3.11}$$

Aus der Theorie folgt unter Berücksichtigung von (3.3), dass

$$var(g) = U * var(q) * U^T$$

ist. Für die Varianz der SNP-Effekte nehmen wir an, dass $var(q) = I\sigma_q^2$. Dies bedeutet, dass sich der Ausdruck für die Kovarianz der genetischen Effekte g vereinfacht zu $var(g) = UU^T\sigma_q^2$.

Im Paper von (Gianola et al., 2009) wurde σ_g^2 aus σ_q^2 hergeleitet. Das Resultat dieser Herleitung lautet

$$\sigma_g^2 = 2\sum_j^m p_j(1-p_j)\sigma_q^2 \tag{3.12}$$

Fassen wir alle diese Beziehungen zusammen, können wir die verschiedenen Ausdrücke für die Kovarianz der genetischen Effekte gleichsetzen.

$$var(g) = G * \sigma_g^2 = UU^T\sigma_q^2 \tag{3.13}$$

Ersetzen wir σ_g^2 in (3.13) durch den Ausdruck in (3.12) so folgt

$$G * 2\sum_j^m p_j(1-p_j)\sigma_q^2 = UU^T\sigma_q^2 \tag{3.14}$$

Aus der Gleichung in (3.14) kann die genomische Verwandtschaftsmatrix G berechnet werden als

$$G = \frac{UU^T}{2\sum_j p_j(1-p_j)} \tag{3.15}$$

3.4.6 R-Code zur Berechnung der GRM

Im Kapitel 13 von (Clark and van der Werf, 2013) ist ein R-Programm zur Berechnung der genomischen Verwandtschaftsmatrix aufgeführt. Wir wollen dieses Programm analysieren und präsentieren auch eine alternative Art der Berechnung der genomischen Verwandtschaftsmatrix.

```
# Making the genomic relationship matrix
nmarkers <- 1000
# Reading SNP-Genotype codes (0,1,2) from file
data <- matrix(scan("genotypes.txt"),ncol = nmarkers,byrow = TRUE)
# Initialisation of variables
sumpq = 0
freq = dim(data)[1]
P = freq
lamda = ncol(data)
# Compute allele frequencies and column vectors of matrix P
for(i in 1:ncol(data)){
```

```
  (freq[i]  <- ((mean(data[,i])/2)))
  (P[i] = (2*(freq[i]-0.5)))
  (sumpq = sumpq+(freq[i]*(1-freq[i])))
}
# Transform coding from (0,1,2) to (-1,0,1) and subtract P
Z <- data
for(i in 1:nrow(data)){
  for(j in 1:ncol(data)){
    (Z[i,j] <- ((data[i,j]-1)-(P[j])))
  }
}
# Compute GRM
Zt = t(Z)
ZtZ = Z%*%Zt
G = ZtZ/(2*sumpq)
G
```

```
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,]  0.8978768 -0.2330233 -0.2286366 -0.2067029 -0.2295139
## [2,] -0.2330233  0.8469907 -0.2014388 -0.2189858 -0.1935427
## [3,] -0.2286366 -0.2014388  0.9171785 -0.2453062 -0.2417968
## [4,] -0.2067029 -0.2189858 -0.2453062  0.9215652 -0.2505703
## [5,] -0.2295139 -0.1935427 -0.2417968 -0.2505703  0.9154238
```

Das gezeigte Programm zur Berechnung der genomischen Verwandtschaftsmatrix aus (Clark and van der Werf, 2013) kann vereinfacht werden. Insbesondere können die `for`-loops durch einfachere Funktionen und vektorisierte Berechnungen ersetzt werden. Das folgende Programm berechnet die gleiche genomische Verwandtschafsmatrix.

```
# Making the genomic relationship matrix
nmarkers <- 1000
# Reading SNP-Genotype codes (0,1,2) from file
data <- matrix(scan("genotypes.txt"),ncol = nmarkers,byrow = TRUE)
# Compute allele frequencies, column vectors of matrix P and sum of frequency products
freq <- apply(data, 2, mean) / 2
P <- 2 * (freq - 0.5)
sumpq <- sum(freq*(1-freq))
# Compute matrix Z by first changing the coding from (0,1,2)
# to (-1,0,1) and then subtracting matrix P
Z <- data - 1 - matrix(P, nrow = nrow(data), ncol = ncol(data), byrow = TRUE)
# Z%*%Zt is replaced by tcrossprod(Z)
(matGrm <- tcrossprod(Z)/(2*sumpq))
```

```
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,]  0.8978768 -0.2330233 -0.2286366 -0.2067029 -0.2295139
## [2,] -0.2330233  0.8469907 -0.2014388 -0.2189858 -0.1935427
## [3,] -0.2286366 -0.2014388  0.9171785 -0.2453062 -0.2417968
## [4,] -0.2067029 -0.2189858 -0.2453062  0.9215652 -0.2505703
## [5,] -0.2295139 -0.1935427 -0.2417968 -0.2505703  0.9154238
```

Wir können überprüfen, ob die beiden Matrizen identisch sind

```
all.equal(G, matGrm)
```

```
## [1] TRUE
```

Das oben gezeigte alternative Programm zur Berechnung der genomischen Verwandtschaftsmatrix kann

so verwendet werden. Da R eine funktionale Programmiersprache ist, soll es unser Ziel sein, möglichst viele Aufgaben in Funktionen zu kapseln. Die folgende Funktion `computeMatGrm()` berechnet aufgrund der Datenmatrix die genomischen Verwandtschaftsmatrix auf.

```
#' Compute genomic relationship matrix based on data matrix
computeMatGrm <- function(pmatData) {
  # Allele frequencies, column vector of P and sum of frequency products
  freq <- apply(pmatData, 2, mean) / 2
  P <- 2 * (freq - 0.5)
  sumpq <- sum(freq*(1-freq))
  # Changing the coding from (0,1,2) to (-1,0,1) and subtract matrix P
  Z <- data - 1 - matrix(P, nrow = nrow(pmatData),
                         ncol = ncol(pmatData),
                         byrow = TRUE)
  # Z%*%Zt is replaced by tcrossprod(Z)
  return(tcrossprod(Z)/(2*sumpq))
}

# Computing the genomic relationship matrix using the above defined function
nmarkers <- 1000
# Reading SNP-Genotype codes (0,1,2) from file
data <- matrix(scan("genotypes.txt"),ncol = nmarkers,byrow = TRUE)
# calling the function
(matGrmFunc <- computeMatGrm(pmatData = data))
```

```
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,]  0.8978768 -0.2330233 -0.2286366 -0.2067029 -0.2295139
## [2,] -0.2330233  0.8469907 -0.2014388 -0.2189858 -0.1935427
## [3,] -0.2286366 -0.2014388  0.9171785 -0.2453062 -0.2417968
## [4,] -0.2067029 -0.2189858 -0.2453062  0.9215652 -0.2505703
## [5,] -0.2295139 -0.1935427 -0.2417968 -0.2505703  0.9154238
```

```
# checking the result
all.equal(G, matGrmFunc)
```

```
## [1] TRUE
```

3.5 Wie GBLUP funktioniert

Die genomischen Verwandtschaftsmatrix erlaubt es uns die Daten der Tiere mit Beobachtung mit den Tieren ohne Beobachtungen aber mit SNP-Genotypen zu verlinken. Die Inverse G^{-1} der genomischen Verwandtschaftsmatrix wird zur Modellierung der Kovarianzstruktur zwischen den genetischen Effekten und somit zwischen den genomischen Zuchtwerten verwendet. Wie schon beim BLUP-Tiermodell werden auch bei GBLUP die unbekannten Effekte mit Mischmodellgleichungen geschätzt. Für das GBLUP-Modell in (3.2) lauten die Mischmodellgleichungen, wie folgt.

$$\begin{bmatrix} X^TX & X^TZ & 0 \\ Z^TX & Z^TZ + G^{11} & G^{12} \\ 0 & G^{21} & G^{22} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \\ 0 \end{bmatrix}$$

G^{11} steht für den Teil der Inversen G^{-1}, der zu den Tieren mit phänotypischen Beobachtungen gehört. Analog dazu entspricht G^{22} dem Teil von G^{-1}, der die Tiere ohne Beobachtungen mit nur SNP-Genotypen beinhaltet. Die Teile G^{12} und G^{21} verknüpfen die Tiere mit und ohne Beobachtungen. Bei den oben gezeigten Mischmodellgleichungen wird ein Verhältnis der Varianzkomponenten σ_g^2 und σ_e^2 von 1 angenommen.

Aufgrund der letzten Zeile der GBLUP-Mischmodellgleichungen ist ersichtlich, dass die genomischen Zuchtwerte $\hat{g}_2$ der Tiere ohne phänotypische Beobachtungen aufgrund der Schätzungen $\hat{g}_1$ der genomischen Zuchtwerte der Tiere mit Beobachtungen vorhergesagt werden können.

$$\hat{g}_2 = -\left(G^{22}\right)^{-1} G^{21} \hat{g}_1 \tag{3.16}$$

Die Beziehung in (3.16) wird auch als genomische Regression der Zuchtwerte der Tiere ohne Beobachtung auf die genomischen Zuchtwerte der Tiere mit Beobachtung bezeichnet.

3.6 Ein Beispiel für die Verwendung von GBLUP

Nehmen wir an, dass wir fünf Tiere haben in einem Pedigree. Von diese fünf Tieren haben nur vier eine phänotypische Beobachtung. Wir wollen genomische Informationen aller fünf Tiere verwenden, damit wir den genomischen Zuchtwert des fünften Tieres schätzen können. Das folgende Pedigree symbolisiert die Verwandtschaftsbeziehungen zwischen den fünf Tieren und wird als gegeben angenommen.

```
##   sire  dam
## 1 <NA> <NA>
## 2    1 <NA>
## 3    1 <NA>
## 4 <NA> <NA>
## 5    1 <NA>
```

Aus dem oben gezeigten Pedigree können wir die folgenden genetisch-additive Verwandtschaftsmatrix A ableiten.

$$A = \begin{bmatrix} 1.00 & 0.50 & 0.50 & 0.00 & 0.50 \\ 0.50 & 1.00 & 0.25 & 0.00 & 0.25 \\ 0.50 & 0.25 & 1.00 & 0.00 & 0.25 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.50 & 0.25 & 0.25 & 0.00 & 1.00 \end{bmatrix}$$

Unter der Annahme, dass die fixen Effekte bekannt sind und dass die Beobachtungen y als Abweichungen von einem Mittel definiert sind, dann kann der Vektor $\hat{u}$ der geschätzten Zuchtwerte aus den Mischmodellgleichungen berechnet werden als

$$\hat{u} = (Z^T Z + A^{-1})^{-1} Z^T y$$

Für unser Beispiel ist die Designmatrix Z, welche die Zuchtwerte u zu den Beobachtungen y verlinkt, wie folgt definiert.

$$Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Somit beträgt die Projektionsmatrix P, welche die Beobachtungen y auf die geschätzten Zuchtwerte $\hat{u}$ abbildet

$$P = (Z^T Z + A^{-1})^{-1} Z^T$$

$$P = \begin{bmatrix} 0.4375 & 0.1250 & 0.1250 & 0.0000 \\ 0.1250 & 0.4643 & 0.0357 & 0.0000 \\ 0.1250 & 0.0357 & 0.4643 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.5000 \\ 0.2188 & 0.0625 & 0.0625 & 0.0000 \end{bmatrix}$$

Schaut man sich die erste Zeile der Matrix P an, so können wir sehen, aus welchen Beobachtungen der geschätzte Zuchtwert $\hat{u}_1$ des Tieres 1 aus den Beobachtungen zusammengesetzt ist. Auf der fünften Zeile können wir sehen, wie der geschätzte Zuchtwert $\hat{u}_5$ zusammengesetzt ist. Vergleicht man die beiden geschätzten Zuchtwerte so resultiert, dass

$$\hat{u}_5 = 0.5\hat{u}_1$$

Nehmen wir die folgende genomische Verwandtschaftsmatrix G als bekannt an und setzen den Ansatz aus dem Kapitel 3.5 so erhalten wir das folgende Resultat.

$$G = \begin{bmatrix} 1.000 & 0.500 & 0.500 & 0.020 & 0.500 \\ 0.500 & 1.000 & 0.200 & 0.015 & 0.200 \\ 0.500 & 0.200 & 1.000 & 0.025 & 0.300 \\ 0.020 & 0.015 & 0.025 & 1.000 & 0.200 \\ 0.500 & 0.200 & 0.300 & 0.020 & 1.000 \end{bmatrix}$$

Für die Schätzung des genomischen Zuchtwertes g_5 für das Tier 5 mit der Formel (3.16) brauchen wir die Matrizen G^{22} und G^{21}, welche wir aus der Inversen G^{-1} von G herauslesen können. Als erstes berechnen wir also G^{-1}

$$G^{-1} = \begin{bmatrix} 2.0505 & -0.7563 & -0.6723 & 0.0006 & -0.6724 \\ -0.7574 & 1.3447 & 0.0841 & -0.0089 & 0.0863 \\ -0.6747 & 0.0845 & 1.3446 & -0.0198 & -0.0790 \\ 0.1220 & -0.0240 & -0.0046 & 1.0031 & -0.2554 \\ -0.6738 & 0.0843 & -0.0840 & -0.0126 & 1.3478 \end{bmatrix}$$

Somit ist

$$G^{22} = \begin{bmatrix} 1.3478 \end{bmatrix}$$

und

$$(G^{21})^T = \begin{bmatrix} -0.6738 \\ 0.0843 \\ -0.0840 \\ -0.0126 \end{bmatrix}$$

Somit sind die Koeffizienten für die Schätzung von g_5

```
## [1]  0.49994639 -0.06257539  0.06230733  0.00938202
```

Somit beträgt

$$\hat{g}_5 = 0.4999464\hat{g}_1 + (-0.0625754\hat{g}_2) + (0.0623073\hat{g}_3) + (0.009382\hat{g}_4)$$