Methoden des Elastic Net zur sparsamen Variablenselektion und deren Anwendung in der Genetik

Masterarbeit
von
Claudia Stuckart

Betreuung: Prof. Dr. Volker Schmid (LMU)
Prof. Dr. Chris-Carolin Schön (TUM)
Valentin Wimmer (TUM)

Ludwig-Maximilians-Universität München
Fakultät für Mathematik, Informatik und Statistik
Institut für Statistik

18. Dezember 2012

Zusammenfassung

In dieser Arbeit werden verschiedene Bayesianische und frequentistische Regressionsmodelle auf ihre Eignung zur Vorhersage in Züchtungsprogrammen untersucht. Die Modelle, die dabei betrachtet werden, sind das Bayesianische Ridge, das Bayesianische Lasso, das Bayesianische Elastic Net, die frequentistischen Analoga und das Generalisierte Elastic Net. Die Sensibilität des Bayesianischen Elastic Net wird über verschiedene Szenarien der Wahl der Hyperparameter für die Priori-Verteilungen untersucht. Die Vorhersagegenauigkeit aller penalisierten Modelle wird über Kreuzvalidierungen geprüft. Angewendet werden die Regressionsmodelle auf experimentelle Daten zu Arabidopsis thaliana und vier quantitativen Merkmalen mit unterschiedlicher genetischer Architektur. Es zeigt sich, dass das Bayesianische Elastic Net teilweise sensibel auf die Wahl der Hyperparameter reagiert. Die Vorhersagegenauigkeit der Methoden unterscheidet sich für die verschiedenen Merkmale im Allgemeinen gering. Die neuesten Modelle, das Bayesianische Elastic Net und das Generalisierte Elastic Net, sind bezüglich ihrer Vorhersagegenauigkeit nicht signifikant besser als die etablierten Methoden.

Inhaltsverzeichnis

1 Einführung

Die Genetik ist ein Gebiet der Wissenschaft, dessen Bedeutung seit Mitte des 19. Jahrhunderts, ausgelöst durch Gregor Mendel, stark an Bedeutung gewonnen hat. Durch Fortschritte in der makroskopischen, mikroskopischen und molekularen Forschung sind viele Fragen über Organismen geklärt und dennoch befindet sich die Erforschung des Genoms erst im Anfangsstadium. Klar ist soweit, dass der Genotyp eines Organismus den Phänotyp, also das Erscheinungsbild des Organismus, bestimmt.
Die Analyse des Erbguts ist insbesondere im Anwendungsgebiet der Pflanzenzüchtung und Selektion von großer Bedeutung (Jannink *et al.*, 2010). Durch die Vorhersage der quantitativen Merkmale alleine basierend auf den genetischen Eigenschaften wäre es nicht mehr nötig erst die Ernte abzuwarten, um die Qualität und den Ertrag von Nutzpflanzen zu beurteilen. Basierend auf den frühzeitigen Erkenntnissen könnte der Selektionszyklus beschleunigt werden und damit der genetische Fortschritt schneller ablaufen.
Die genetische Erbinformation ist in der DNA beziehungsweise in den Chromosomen gespeichert. Bei vielen Pflanzen liegt der Chromosomensatz und somit auch jedes Gen doppelt (diploid) vor. Dies impliziert jedoch nicht, dass die Nukleotidensequenz, welche das Gen definiert, bei homologen Chromosomen identisch ist. Die unterschiedlichen Formen eines Gens werden als Allele bezeichnet. Das Auftreten genetisch unterschiedlicher Phänotypen in einer Population, bedingt durch die Allele einer Gens wird als Polymorphismus bezeichnet. Der häufigste Polymorphismus im Genom ist der Einzelnukleotid-Polymorphismus (engl.: single nucleotide polymorphism, SNP), also die Variation eines einzelnen Nukleotids. Darunter versteht man zum Beispiel den einzelnen Basenaustausch von Adenin und Thymin oder den einzelnen Basenaustausch von Cytosin und Guanin. Alle SNPs haben zwei Allele. Eine detaillierte Beschreibung der allgemeinen Genetik und der Molekulargenetik wird von Knust und Janning (2008) und Griffiths *et al.* (2012) gegeben.
Ziel dieser Arbeit ist es, den gemeinsamen Einfluss von vielen SNPs auf die quantitativen Merkmale zu untersuchen. Die SNPs können drei Ausprägungen aufweisen. Diese sind zum einen die homozygoten Ausprägungen mit entweder zwei dominanten oder zwei rezessiven Allelen und zum anderen die heterozygote Ausprägung mit einen dominanten und einem rezessiven Allel. Die SNPs werden jedoch nicht als kategoriale Einflussgrößen in ein Regressionsmodell aufgenommen, sondern deren Ausprägungen so rekodiert, dass die Ausprägung einer Einflussgröße die Anzahl der seltenen Allele in dem SNP ist. Die Einflussgrößen sind somit metrisch.
In dieser Arbeit werden die öffentlich verfügbaren Daten zu dem Modellorganismus *Arabidopsis thaliana* (L.) untersucht. Als Merkmale werden die

Pflanzenhöhe, die Wachstumsrate, die Zeit bis zum Schossbeginn und die Zeit zwischen Schossbeginn und Blütezeit betrachtet. Diese Merkmale werden stetig gemessen und gehen dementsprechend als metrisch Zielvariablen in das Modell ein.
Für die Untersuchung des Einflusses von SNPs auf die quantitativen Merkmale wird, auf Grund der metrischen Zielvariablen, ein lineares Regressionsmodell unterstellt. Die unbekannten Regressionskoeffizenten lassen sich im Allgemeinen durch Kleinste-Quadrate (KQ) Schätzung oder Maximum-Likelihood (ML) Schätzung bestimmen (Fahrmeir *et al.*, 2007). Der KQ Schätzer ist erwartungstreu und unter allen linearen erwartungstreuen Schätzern, jener mit der kleinsten Varianz (Gauß Markov Theorem). Je kleiner die Varianz eines Schätzers ist, desto genauer ist meist die Schätzung und desto besser ist die Vorhersagegenauigkeit (Hastie *et al.*, 2009). Bei der Schätzung durch die KQ Methode bleiben alle Variablen in dem Modell. Es findet keine Variablenselektion statt. Bei vielen Variablen im Modell ist die Interpretation schwierig, da die Interpretation von einzelnen Koeffizienten immer unter der Bedingung 'festhalten der anderen Variable' erfolgt. Ein weiterer Nachteil der KQ Methode ist, dass der Schätzer nur existiert, falls in der Schätzgleichung keine singulären Matrizen vorkommen. Damit keine Singularitätsprobleme auftreten müssen mehr Beobachtungen als Variablen vorliegen und es darf keine exakte lineare Abhängigkeit zwischen den Variablen bestehen. Sind die Variablen nicht exakt linear abhängig, sondern hoch korreliert, kann die Varianz der Schätzung extrem groß werden.
Die erhobenen molekulargenetischen Daten der Arabidopsis umfassen $n = 426$ Individuen und $p = 1260$ SNPs. Somit ist die Anzahl der Prädiktoren wesentlich größer als die Anzahl der Beobachtungen. Es resultiert das sogenannte $p \gg n$-Problem, welches unter anderem von Fan und Lv (2008) erläutert wird. Die Spaltendimension der Designmatrix ist im Vergleich zur Zeilendimension sehr groß. Dadurch treten bei der KQ Schätzung Singularitätsprobleme auf und eine Schätzung der Parameter ist nicht mehr möglich. Die Prädiktoren weisen, allein schon auf Grund der hohen Anzahl an Einflussgrößen, eine Korrelation auf. Aber nicht nur die Dimension von SNP Daten verursacht eine Korrelation, sondern auch die inhaltliche Beschaffenheit der genetischen Kopplung. Mit Kopplung wird die Assoziation von Genen auf dem gleichen Chromosom bezeichnet, welche zur gemeinsamen Vererbung der entsprechenden Merkmale führt (Knust und Janning, 2008). Bei Genen, deren gemeinsame Allelverteilung nicht zufällig ist, spricht man von einem Kopplungsungleichgewicht (engl.: linkage disequilibrium, LD) (Griffiths *et al.*, 2012). Das Kopplungsungleichgewicht führt zu Kollinearität. Dadurch werden einflussreiche und nicht einflussreiche Pädiktoren nicht immer als solche erkannt.

Für die Schätzung eines Regressionsmodells bei Daten mit einem $p \gg n$-Problem existieren diverse Ansätze. Allgemeine Ziele dieser Ansätze sind Vorhersagegenauigkeit und gute Interpretierbarkeit des Modells. Die Vorhersagegenauigkeit kann durch Schätzer mit einer geringen Varianz und die gute Interpretierbarkeit durch Variablenselektion erreicht werden (Hastie *et al.*, 2009).
Hoerl und Kennard (1970a,b) führten die Ridge Regression ein. Dies ist ein Penalisierungsverfahren, bei dem die resultierenden Schätzer eine geringe Varianz aufweisen. Die Schrumpfung der Regressionskoeffizienten erfolgt über einen Penalisierungsterm. Die Stärke der Penalisierung wird über den Penalisierungsparameter des Penalisierungsterms gesteuert. Der Ridge Schätzer weist den sogenannten Gruppierungseffekt (engl.: grouping effect) auf. Als Gruppierungseffekt wird der Effekt bezeichnet, dass korrelierte Einflussgrößen ähnliche Schätzer erhalten. Jedoch findet bei diesem Verfahren keine Variablenselektion statt. Eine Methode, bei der neben der Schrumpfung von Parametern zusätzlich Variablen selektiert werden, wurde von Tibshirani (1996) vorgeschlagen und wird als Kleinster Absoluter Schrumpfungs- und Selektionsoperator (engl.: Least absolute shrinkage und selection operator, Lasso) bezeichnet. Der Lasso Schätzer weist im Vergleich zum Ridge Schätzer keinen Gruppierungseffekt auf. Eine Kombination des Ridge und Lasso Verfahrens ist das Elastic Net (Zou und Hastie, 2005). Dieses soll die Vorteile von Variablenselektion und Gruppierungseffekt vereinen. In diesen frequentistischen Methoden werden die Penalisierungsparameter über eine Kreuzvalidierung bestimmt.
Diese Penalisierungsverfahren können ebenfalls Bayesianisch formuliert werden. So beschreiben Fahrmeir *et al.* (2010) das Bayesianische Ridge, Park und Casella (2008) das Bayesianisches Lasso und Li und Lin (2010) das Bayesianische Elastic Net. Die Bayesianischen Methoden weisen alle Vorteile der frequentistischen Methoden auf und erlauben zusätzlich Vorwissen über die Parameter und Penalisierungsparameter in das Modell aufzunehmen.
Ishwaran und Rao (2011) erweiterten das Elastic Net zu dem sogenannten Generalisierten Elastic Net.
Weitere Modelle für die Inferenz in $p \gg n$-Situationen, welche in dieser Arbeit nicht näher betrachtet werden, jedoch in der Literatur viel Anwendung finden, sind die Modelle BayesA und BayesB (Meuwissen *et al.*, 2001). Desweiteren resultiert, basierend auf einem linearen gemischten Modell, der BLUP (best linear unbiased predictor) Schätzer (Henderson, 1984). Dieser Schätzer wird unter anderem in den Studien von Fernando und Grossman (1989) und Meuwissen und Goddard (1996) verwendet.
In dieser Arbeit wird untersucht, ob die Elastic Net Methoden bessere Ergebnisse liefern als die Lasso und Ridge Methoden und ob die Bayesianischen

Methoden den frequentistischen Methoden überlegen sind. Für die Beurteilung werden verschiedene Gütekriterien herangezogen. Die Gütekriterien sind die Korrelation zwischen den wahren und angepassten Werten, die Anzahl der effektiven Parameter, das Devianz Informationskriterium (engl.: deviance information criterion, DIC) und Kreuzvalidierungen mit verschiedenen Kriterien. Desweiteren werden die Unterschiede für Merkmale mit verschiedener genetischer Architektur untersucht.
Die weitere Arbeit gliedert sich wie folgt. Die statistischen Methoden werden in Kapitel 2 vorgestellt. Das Kapitel 2 ist unterteilt in vier Teilkapitel. In dem ersten Teilkapitel wird das Ridge, das Lasso und das Elastic Net Modell geschildert. In dem zweiten Teilkapitel wird die Bayes Inferenz eingeführt und in dem dritten Teilkapitel das Bayesianische Ridge, das Bayesianische Lasso und das Bayesianische Elastic Net beschrieben. Die Erläuterung des Generalisierten Elastic Net erfolgt in dem vierten Teilkapitel. In Kapitel 3 werden die Daten der Arabidopsis deskriptiv und explorativ analysiert. Die Anwendung aller vorgestellten Modelle auf diese Daten erfolgt in Kapitel 4. In Kapitel 5 werden die Ergebnisse dieser Arbeit zusammengefasst und diskutiert.

2 Penalisierte lineare Modelle

In diesem Kapitel werden verschiedene penalisierte lineare Modelle vorgestellt. Alle diese Verfahren sind Erweiterungen des multiplen linearen Regressionsmodells. Die Daten liegen in der Form $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$ vor. Dabei ist $y_i, i = 1, \ldots, n$ der Phänotyp des Individuums i und $x_{ij}, j = 1, \ldots, p$ die Anzahl des seltenen Allels in SNP j. Das multiple lineare Regressionsmodell ist definiert durch (Fahrmeir *et al.*, 2007):

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & & \ddots & \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

wobei die Störgrößen homoskedastisch sind und einer Normalverteilung mit Erwartungswert $\mathbf{0}$ und Varianz σ^2 folgen: $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$.
Die unbekannten Regressionskoeffizenten $\boldsymbol{\beta}^\top = (\beta_0, \ldots, \beta_p)$ lassen sich im Allgemeinen durch Kleinste-Quadrate Schätzung oder Maximum-Likelihood Schätzung bestimmen (Fahrmeir *et al.*, 2007):

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{ML} &= \underset{\beta}{\operatorname{argmax}}\ \mathrm{l}(\boldsymbol{\beta}, \sigma^2) = \underset{\beta}{\operatorname{argmax}} \left\{ -\log((2\pi\sigma^2)^{n/2}) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} \\ &= \underset{\beta}{\operatorname{argmin}} \underbrace{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}_{Residuenquadratsumme} = \hat{\boldsymbol{\beta}}^{KQ}. \end{aligned}$$

Der KQ Schätzer ist zwar unverzerrt, hat aber bei Kollinearität der Variablen eine große Varianz und ist in einer $p > n$-Situation auf Grund der Nichtinvertierbarkeit von $\mathbf{X}^\top\mathbf{X}$ nicht schätzbar. Verzerrte Schätzer, welche auch in $p > n$-Situationen berechnet werden können, sind penalisierte Likelihood Schätzer. Diese lassen sich allgemein wie folgt darstellen:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\beta}{\operatorname{argmax}}\{\mathrm{l}(\boldsymbol{\beta}) - \mathrm{pen}(\boldsymbol{\beta})\} \\ &= \underset{\beta}{\operatorname{argmin}}\{-\mathrm{l}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta})\}, \end{aligned}$$

wobei mit $\mathrm{l}(\boldsymbol{\beta})$ die logarithmierte Likelihood und mit $\mathrm{pen}(\boldsymbol{\beta})$ der Penalisierungsterm bezeichnet wird. Um die Stärke der Penalisierung zu regulieren beinhaltet der Penalisierungsterm den sogenannten Penalisierungsparameter. Der Penalisierungsparameter kann über Kreuzvalidierung oder bayesianisch über eine Priori-Verteilung geschätzt werden.

2.1 Aufbau des Elastic Net

In diesem Teilkapitel wird in Abschnitt 2.1.1 die Methode der Ridge Schätzung erklärt. In dem Abschnitt 2.1.2 wird das Verfahren Lasso vorgestellt. Der Naive Elastic Net Schätzer und dessen Verbesserung, der Elastic Net Schätzer, werden in den Abschnitten 2.1.3 und 2.1.4 erläutert. Die Schätzungen für das Lasso und Elastic Net erfolgen iterativ. Nur im orthogonalen Design lassen sich alle Schätzer konkret formulieren. Die Betrachtung der Schätzer im orthogonalen Design erfolgt in Abschnitt 2.1.5.

2.1.1 Ridge

Hoerl und Kennard (1970a,b) führten die Ridge Regression ein. Die Ridge Regression liefert einen Schätzer, welcher die Residuenquadratsumme minimiert und dessen Länge beschränkt ist (L_2 Penalisierung):

$$\hat{\boldsymbol{\beta}}^R = \underset{\beta}{\operatorname{argmin}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \quad \text{u.d.B.} \quad \sum_{j=1}^{p} \beta_j^2 \leq t.$$

Bei der Ridge Regression werden die KQ Koeffizienten kontinuierlich gegen Null geschrumpft. Je kleiner der Anpassungsparameter (engl.: tuning paramter) t, desto stärker ist die Schrumpfung. Die Schätzer werden jedoch nie exakt gleich Null. Äquivalent ist die penalisierte Schreibweise in Matrixform:

$$\hat{\boldsymbol{\beta}}^R = \underset{\beta}{\operatorname{argmin}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \right\},$$

mit dem Penalisierungsterm $\text{pen}(\boldsymbol{\beta}) = \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}$. Je größer der Parameter λ, desto stärker ist die Schrumpfung der Koeffizienten gegen Null. Die Parameter λ und t haben eine eineindeutige Beziehung, sind jedoch nicht gleich. Die Lösung der Ridge Regression ist einfach darstellbar durch:

$$\hat{\boldsymbol{\beta}}^R = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}. \tag{2.1}$$

Die Motivation zur Einführung des Ridge Schätzers war die Problematik der Kollinearität zu lösen. Sind Variablen hoch korreliert, so ist die Varianz der durch die KQ Methode geschätzten Koeffizienten extrem hoch. Die Vorhersagegenauigkeit des resultieren Modells kann dann schlecht sein. Der Ridge Schätzer reduziert die Varianz der Regressionskoeffizienten, ist jedoch nicht mehr unverzerrt (Batah und Gore, 2009). Die Reduktion der Varianz führt in der Regel zu einer Verbesserung der Vorhersagegenauigkeit. Um eine hohe Varianz durch Kollinearität zu vermeiden, werden bei dem Ridge

Schätzer zu der Diagonalen von $\mathbf{X}^\top\mathbf{X}$ Konstanten λ (Gleichung (2.1)) addiert. Durch die Addition wird sicher gestellt, dass $\mathbf{X}^\top\mathbf{X}$ selbst bei einer $p \gg n$-Datengrundlage immer invertierbar ist, die Varianz nicht zu groß wird und der Ridge Schätzer existiert. Die Varianz bei der Ridge Regression ist immer kleiner als die Varianz bei der KQ Schätzung (Miller, 2002). Wird λ so gewählt, dass die Verzerrung klein ist, dann hat der Ridge Schätzer einen kleineren Mittleren Quadratischen Fehler (engl.: mean squared error, MSE) als der KQ Schätzer (Hoerl und Kennard, 1970b).
Da bei der Ridge Regression die Koeffizienten nie exakt auf Null geschätzt werden, findet keine Variablenselektion statt. Dadurch gibt es keine Verbesserung in der Interpretierbarkeit. Ein Charakteristikum der Ridge Regression ist der Gruppierungseffekt. Der Gruppierungseffekt quantifiziert den Unterschied zwischen zwei Regressionskoeffizienten über eine Funktion des Korrelationskoeffizienten der zugehörigen Variablen. Für gleiche Kovariablen, also Kovariablen mit einer Korrelation von Eins, werden dieselben Regressionskoeffizienten geschätzt (Zou und Hastie, 2005).

2.1.2 Lasso

In diesem Abschnitt wird ein Schätzverfahren eingeführt, das den Vorteil der Koeffizientenschrumpfung der Ridge Regression beibehält und zusätzlich eine Variablenselektion beinhaltet. Die Koeffizienten sollen geschrumpft werden und auch Schumpfungen auf exakt Null stattfinden. Auf diesem Weg soll Vorhersagegenauigkeit und gute Interpretierbarkeit erreicht werden.
Der Schätzer für den dies zutrifft ist der Lasso Schätzer, welcher von Tibshirani (1996) definiert wurde. Das Lasso minimiert die Residuenquadratsumme unter einer Nebenbedingung (L_1 Penalisierung):

$$\hat{\boldsymbol{\beta}}^L = \underset{\beta}{\operatorname{argmin}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \quad \text{u.d.B.} \quad \sum_{j=1}^{p} |\beta_j| \leq t.$$

Dies ist äquivalent zur penalisierten Maximum-Likelihood Schätzung:

$$\hat{\boldsymbol{\beta}}^L = \underset{\beta}{\operatorname{argmin}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

mit dem Penalisierungsterm $\text{pen}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|$. Durch die Bestrafung ist die Lösung nicht linear in $\boldsymbol{y}$. Eine geschlossene Lösung ist im Allgemeinen nicht möglich. Die iterative Lösung durch den LARS Algorithmus wird von Efron *et al.* (2004) beschrieben.
Der Lasso Parameter $t \geq 0$ kontrolliert die Stärke der Schrumpfung. Je kleiner t, desto stärker ist die Schrumpfung. Ist $t < \sum |\hat{\beta}_j^{KQ}|$, so werden die KQ

Regressoren gegen und gleich Null geschrumpft. Ist jedoch $t \geq \sum |\hat{\beta}_j^{KQ}|$, so ist der Lasso Schätzer gleich dem KQ Schätzer. Der Lasso Parameter t ist folglich sinnvoll gewählt mit $t \in [0, \sum |\hat{\beta}_j^{KQ}|]$. Oft wird der standardisierte Lasso Parameter $s = t / \sum |\hat{\beta}_j^{KQ}| \in [0, 1]$ betrachtet. Der Parameter λ, der in eineindeutiger Beziehung zu t steht, hat die umgekehrte Wirkung wie t. Je größer λ, desto stärker ist die Schrumpfung. Bei $\lambda = 0$ ist der Lasso Schätzer gleich dem KQ Schätzer.
Auf Grund der Schrumpfung der KQ Regressoren mittels eines Parameters, λ oder t, hat der Lasso Schätzer eine kleinere Varianz als der KQ Schätzer. Folglich verbessert sich, wie bei der Ridge Regression, bei einer sinnvollen Wahl des Penalisierungsparameters die Vorhersagegenauigkeit. Da die Koeffizienten auf exakt gleich Null geschrumpft werden können, findet eine Variablenselektion statt (Abschnitt 2.1.5). Trotz der vielen Vorteile des Lasso sollten auch die Nachteile nicht unerwähnt bleiben. Ist die Anzahl der Prädiktoren größer als die Anzahl der Beobachtungen, $p > n$, werden maximal n Koeffizienten selektiert. Die Variablenselektion ist daher eingeschränkt. Bei gruppierten Variablen beziehungsweise stark korrelierten Variablen tendiert das Lasso dazu, aus einer Gruppe eine beliebige Variable zu wählen und ignoriert die anderen Variablen der Gruppe (Zou und Hastie, 2005). Im Vergleich zu dem Ridge Schätzer existiert kein Gruppierungseffekt. Dies hat zum einen den Vorteil eines sparsameren Modells aber zum anderen den Nachteil, dass die Selektion bei korrelierten Variablen mit einer gewissen Beliebigkeit verbunden sein kann.

2.1.3 Naiver Elastic Net Schätzer

Sowohl der Lasso Schätzer als auch der Ridge Schätzer weisen Vor- und Nachteile auf. Basierend auf den Vorteilen dieser Schätzer entsteht die Idee eines Schätzverfahrens mit einer kleinen Varianz der Schätzer, uneingeschränkter Variablenselektion und einem Gruppierungseffekt. Ein Verfahren, welches dies realisiert, ist das Naive Elastic Net von Zou und Hastie (2005).
Das Naive Elastic Net verwendet sowohl die L_1 Penalisierung des Lasso als auch die L_2 Penalisierung des Ridge. Das Naive Elastic Net minimiert die Residuenquadratsumme unter Nebenbedingung:

$$\hat{\boldsymbol{\beta}}^{NEN} = \underset{\beta}{\operatorname{argmin}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \quad \text{u.d.B.} \quad (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t.$$

Bei $\alpha = 1$ entspricht der Naive Elastic Net Schätzer dem Ridge Schätzer und bei $\alpha = 0$ dem Lasso Schätzer. Die Minimierung der Residuenquadratsumme unter Nebenbedingung ist äquivalent zur penalisierten Maximum-Likelihood

Schätzung mit dem Penalisierungsterm $\text{pen}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$:

$$\hat{\boldsymbol{\beta}}^{NEN} = \underset{\beta}{\text{argmin}} \left\{ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\},$$

wobei $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$.
Zou und Hastie (2005) zeigen durch eine Umformulierung des Naiven Elastic Net, dass die positiven Eigenschaften des Lasso erhalten bleiben und die Selektion von mehr als p Variablen möglich ist. Dafür wird basierend auf $(\boldsymbol{y}, \mathbf{X})$ ein künstlicher Datensatz $(\boldsymbol{y}^*, \mathbf{X}^*)$ erzeugt:

$$\mathbf{X}^* = (1 + \lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I} \end{pmatrix}, \boldsymbol{y}^* = \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix},$$

wobei $\dim(\mathbf{X}^*) = (n+p) \times p$ und $\dim(\boldsymbol{y}^*) = (n+p) \times 1$. Mit $\gamma = \lambda_1/\sqrt{1+\lambda_2}$ und $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$ lässt sich der Naive Elastic Net Schätzer die folgt berechnen:

$$\hat{\boldsymbol{\beta}}^* = \underset{\beta^*}{\text{argmin}} \left\{ (\boldsymbol{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*)^\top (\boldsymbol{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*) + \gamma \sum_{j=1}^{p} |\beta_j^*| \right\},$$

$$\hat{\boldsymbol{\beta}}^{NEN} = \frac{1}{\sqrt{1+\lambda_2}} \hat{\boldsymbol{\beta}}^*.$$

Die Berechnung von $\hat{\boldsymbol{\beta}}^*$ weist genau dieselbe Struktur wie die Berechnung von $\hat{\boldsymbol{\beta}}^L$ auf. In der Lasso Schätzung ist der Rang des Designmatrix $\mathbf{X}$ gleich n und daher maximal die Selektion von n Prädiktoren möglich. Bei dem Naiven Elastic Net Schätzer weist die Designmatrix $\mathbf{X}^*$ den Rang $\text{rg}(\mathbf{X}^*) = p$ auf und daher können bei der Naiven Elastic Net Schätzung bis zu p Variablen selektiert werden.
Der Gruppierungseffekt des Ridge Schätzers existiert ebenfalls für den Naiven Elastic Net Schätzer. Dies wird von Zou und Hastie (2005) gezeigt. Ein Gruppierungseffekt liegt immer bei streng konvexen Penalisierungsfunktionen vor. Die Penalisierungsfunktion des Lasso ist konvex, aber nicht streng konvex. Die Penalisierungsfunktionen des Ridge und des Naiven Elastic Net sind streng konvex (Abschnitt 2.1.5). Bei Lasso existiert somit kein Gruppierungseffekt und bei Ridge und dem Naiven Elastic Net schon.
Der Nachteil der Inferenz des Naiven Elastic Net ist der Effekt der Doppel-Schrumpfung. Die Schätzung der Parameter erfolgt in zwei Stufen. Zuerst werden für feste Werte von λ_2, die Koeffizienten der Ridge Regression geschätzt und anschließend die Lasso Schätzung ausgeführt. Es werden somit zwei Schrumpfungsmethoden angewandt und dies führt zu einer zusätzlichen Verzerrung. Ein verbesserter Schätzer, welcher dieses Problem nicht aufweist, ist der Elastic Net Schätzer.

2.1.4 Elastic Net Schätzer

Der Elastic Net Schätzer wird von Zou und Hastie (2005) definiert und stellt eine verbesserte Form des Naiven Elastic Net dar. Der Naive Elastic Net basiert auf der bereits erläuterten Schätzung von:

$$\hat{\boldsymbol{\beta}}^* = \underset{\beta^*}{\operatorname{argmin}} \left\{ (\boldsymbol{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*)^\top (\boldsymbol{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*) + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \sum_{j=1}^{p} |\beta_j^*| \right\}.$$

Daraus resultiert der Elastic Net Schätzer:

$$\hat{\boldsymbol{\beta}}^{EN} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^*,$$

welcher eine Reskalierung des Naiven Elastic Net Schätzers ist:

$$\hat{\boldsymbol{\beta}}^{EN} = (1+\lambda_2)\hat{\boldsymbol{\beta}}^{NEN}.$$

Diese Reskalierung impliziert, dass die positiven Charakteristika des Naiven Elastic Net für das Elastic Net erhalten bleiben. Für den Elastic Net Schätzer existiert der Effekt der Doppel-Schrumpfung nicht.
Eine iterative Schätzung des Elastic Net ist über den LARS-EN Algorithmus (Zou und Hastie, 2005) möglich.

2.1.5 Orthogonales Design und Geometrie im $\mathbb{R}^2$

Ein orthogonales Design liegt vor, falls $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$. Bei diesem Design lassen sich die Lösungen der Methoden Ridge, Lasso und Naives Elastic Net explizit darstellen. Tabelle 1 zeigt, dass jede Methode eine einfache Transformation des KQ Schätzers ist (Zou und Hastie, 2005). Dabei ist $z^+ = z$ für $z > 0$ und $z^+ = 0$ für $z < 0$. Die entsprechenden Schätzer sind in der Ab-

Tabelle 1: Ridge, Lasso und Elastic Net Schätzer im Orthogonalen Design

Verfahren	Schätzer	pen(β_j)
Ridge	$\hat{\beta}_j^R = \frac{1}{1+\lambda_2}\hat{\beta}_j^{KQ}$	$\lambda\beta_j^2$
Lasso	$\hat{\beta}_j^L = (\lvert\hat{\beta}_j^{KQ}\rvert - \lambda_1/2)^+ \cdot \operatorname{sign}(\hat{\beta}_j^{KQ})$	$\lambda\lvert\beta_j\rvert$
Naives Elastic Net	$\hat{\beta}_j^{NEN} = \frac{(\lvert\hat{\beta}_j^{KQ}\rvert-\lambda_1/2)^+}{1+\lambda_2}\operatorname{sign}(\hat{\beta}_j^{KQ})$	$\lambda_1\lvert\beta_j\rvert + \lambda_2\beta_j^2$

bildung 1 eingezeichnet. Bei der Ridge Regression findet eine proportionale Schrumpfung mittels des konstanten Faktors $\frac{1}{1+\lambda_2}$ statt. Bei Lasso wird der

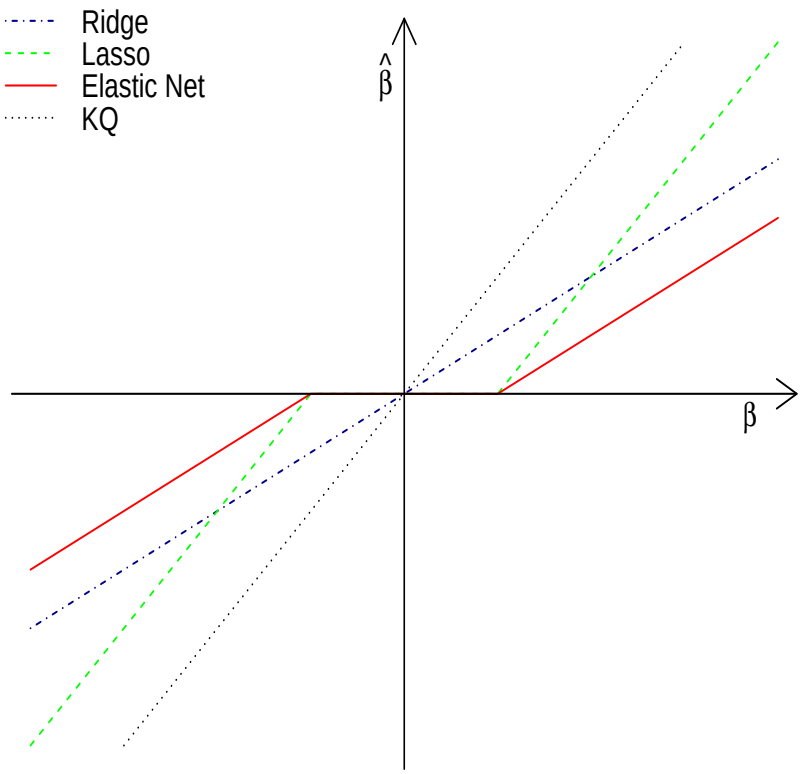

Abbildung 1: Ridge, Lasso und Elastic Net Schätzer mit $\lambda_1 = 2$, $\lambda_2 = 1$ (Zou und Hastie, 2005)

KQ Schätzer um den konstanten Faktor $\lambda_1/2$ verschoben und das Vorzeichen des KQ Schätzers beibehalten. Die Verschiebung endet sobald der Betrag des KQ Schätzers kleiner $\lambda_1/2$ ist. Dann ist $\hat{\beta}_j^L = 0$. Bei der Naiven Elastic Net Schätzung wird der KQ Schätzer mit dem Faktor $\frac{1}{1+\lambda_2}$ geschrumpft und um $\frac{\lambda_1/2}{1+\lambda_2}$ verschoben. Für $|\hat{\beta}_j^{KQ}| < \lambda_1/2$ ist der Schätzer gleich Null.
Der Ridge, Lasso und Naiver Elastic Net Schätzer können im $\mathbb{R}^2$ auch ganz allgemein grafisch dargestellt werden. Alle diese Schätzer minimieren die Residuenquadratsumme $\sum_i (y_i - \sum_j \beta_j x_{ij})^2$ unter einer Nebenbedingung. Die Residuenquadratsumme lässt sich umformen zu (Tibshirani, 1996):

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{KQ})^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{KQ}) + \text{const.} \tag{2.2}$$

Diese Funktion hat elliptische Konturen um $\hat{\boldsymbol{\beta}}^{KQ}$. Für verschiedene Werte der Konstante ergeben sich verschiedene Ellipsen.
Die Nebenbedingungen der Verfahren im $\mathbb{R}^2$ sind in der Abbildung 2 aufgeführt. Die Penalisierungsfunktion des Ridge ist ein Kreis und streng konvex. Bei Lasso ist die Penalisierungsfunktion quadratisch und konvex, jedoch nicht streng konvex. Die Funktion ist in den Achsenschnittpunkten nicht differenzierbar. Die Penalisierungsfunktion des Naiven Elastic Net liegt erwartungsgemäß zwischen der Ridge und Lasso Penalisierung. Die Funktion ist streng konvex und in den Achsenschnittpunkten nicht differenzierbar. Aus der strengen Konvexität der Ridge und Elastic Net Penalisierungsfunktion folgt der Gruppierungseffekt dieser Schätzverfahren. Aus der Nicht-Differenzierbarkeit in den Achsenschnittpunkten folgt die Variablenselektion des Lasso und Elastic Net. Grafisch ist der Koeffizientenschätzer der Punkt, an dem die Ellipse der Residuenquadratsumme (2.2) die Penalisierungsfunktion berührt.

2.2 Bayesianische Inferenz

In dem Teilkapitel 2.1 wurden die Methoden Ridge, Lasso und Elastic Net frequentistisch eingeführt. Diese Modelle können auch mittels Bayes Inferenz geschätzt werden. Die Bayes Inferenz wird in diesem Teilkapitel beschrieben. In Abschnitt 2.2.1 werden die Bayesianischen Punktschätzer und deren Vertrauensintervalle vorgestellt. Die Annahmen, welche a priori für die Schätzung eines Bayesianischen Modells benötigt werden, werden in Abschnitt 2.2.2 erklärt. Die Beschreibung des Bayesianischen linearen Modells erfolgt in Abschnitt 2.2.3. Die Schätzung der Parameter von Bayesianischen Modellen erfolgt über Markov Chain Monte Carlo (MCMC) Methoden (Abschnitt 2.2.4). Die Beurteilung der Modellgüte kann über die in Abschnitt 2.2.5 beschriebenen Methoden erfolgen.

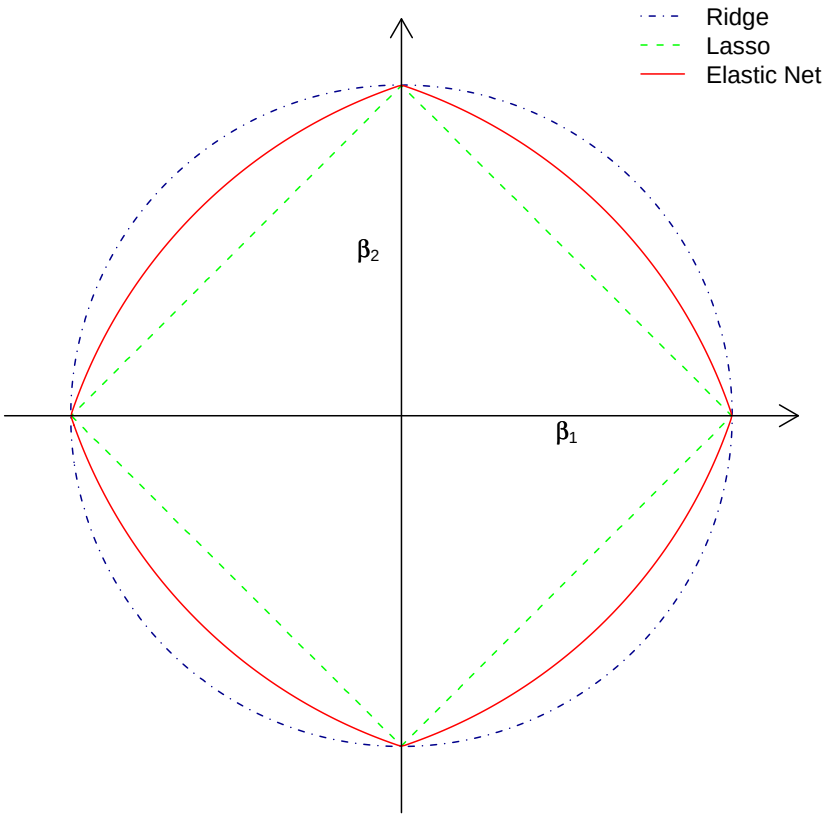

Abbildung 2: Penalisierungsfunktion des Ridge, Lasso und Elastic Net im $\mathbb{R}^2$ (Zou und Hastie, 2005)

Rüger (1999) und Fahrmeir *et al.* (2007) geben eine Einführung in die Bayes Inferenz. Die Bayes Inferenz stützt sich auf ein Theorem ihres Namensgebers Thomas Bayes, dem Satz von Bayes:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{p(\boldsymbol{y})} \\ &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \\ &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \end{aligned}$$

Dabei werden sowohl die unbekannten Parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ als auch die Beobachtungen $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ bedingt auf die Parameter als Zufallszahlen betrachtet und dementsprechend Verteilungen für diese angenommen. Dabei steht $p(\boldsymbol{y}|\boldsymbol{\theta})$ für die Datenverteilung, $p(\boldsymbol{\theta})$ für die Priori-Verteilung und $p(\boldsymbol{\theta}|\boldsymbol{y})$ für die Posteriori-Verteilung. Der Term $1/p(\boldsymbol{y})$ ist eine Konstante bezüglich $\boldsymbol{\theta}$ und die Datenverteilung ist proportional zur Likelihood $\mathrm{L}(\boldsymbol{\theta})$. Es gilt entsprechend die folgende Proportionalität:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto \mathrm{L}(\boldsymbol{\theta})p(\boldsymbol{\theta}). \end{aligned}$$

In der Bayes Inferenz kann die Möglichkeit eine Priori-Verteilung zu spezifizieren sowohl Vorteil als auch Nachteil sein. Bei einer realitätsnahen Wahl der Priori-Verteilung verbessert sich die Aussagekraft des Modells. Bei einer Fehlspezifikation der Priori-Verteilung hingegen kann das resultierende Modell die wahren zugrundeliegenden Sachverhalte eventuell nur unzureichend beschreiben. Die Berechnungszeit ist in der Bayes Inferenz für den Fall, dass die Posteriori-Verteilung unbekannt ist, langsamer als die Berechnungszeit in der frequentistischen Inferenz.

2.2.1 Punktschätzer und Vertrauensintervalle

Die Parameterschätzung in der Bayesianischen Inferenz beruht auf der Posteriori-Verteilung. Es existieren drei mögliche Punktschätzer. Diese sind der Posteriori-Erwartungswert, der Posteriori-Modus und der Posteriori-Median. Der Posteriori-Erwartungswert ist definiert durch:

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{y}) = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} = c \cdot \int \boldsymbol{\theta}\, p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

und der Posteriori-Modus wird bestimmt über:

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\boldsymbol{y}) = \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

Diese Schätzer geben aber keine Auskunft über die Schätzgenauigkeit. Die Genauigkeit der Schätzungen kann über die Vertrauensintervalle erfasst werden. Das $(1-\alpha)$-Vertrauensintervall ist wie folgt definiert:

$$\mathrm{P}(\boldsymbol{\theta} \in C|\boldsymbol{y}) \geq 1-\alpha,$$

beziehungsweise:

$$\int_{C(\boldsymbol{y})} p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} = 1-\alpha\,,$$

wobei der Vertrauensbereich C eine Teilmenge des Parameterraums $\boldsymbol{\Theta}$ ist. Der Vertrauensbereich für $\boldsymbol{\theta}$ ist also so definiert, dass $1-\alpha$ die Posteriori Wahrscheinlichkeit ist, dass $\boldsymbol{\theta} \in C(\boldsymbol{y})$. Es kann folglich bei der Bayes Inferenz eine direkte Wahrscheinlichkeitsaussage über die Parameter getroffen werden. Der Zusammenhang zwischen dem Bayesianischen Posteriori-Modus Schätzer und dem frequentistischen Maximum-Likelihood Schätzer kann wie folgt erläutert werden:

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}} \{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\} = \underset{\beta}{\operatorname{argmax}} \{\mathrm{l}(\boldsymbol{\theta}) - \mathrm{pen}(\boldsymbol{\theta})\},$$

mit $\mathrm{l}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta})$ und $\mathrm{pen}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$.

2.2.2 Priori Annahmen

Zur Berechnung der Posteriori-Verteilung und somit der Parameterschätzer muss das Beobachtungsmodell und die Priori-Verteilung für die unbekannten Parameter spezifiziert werden. In der Priori-Verteilung soll das Vorwissen über die Parameter abgebildet werden. Häufig gewählte Priori-Verteilungen sind die flache Priori und die konjugierte Priori.
Die flache Priori entspricht einer Gleichverteilung des Parameters auf dem Parameterraum $\boldsymbol{\Theta}$ und ist konstant bezüglich des Parameters. Je nach vorliegendem Parameterraum können dies impropere Verteilungen sein, welche keine echten Wahrscheinlichkeitsverteilungen darstellen. Flache Priori-Verteilungen drücken ein *a priori Nichtwissen* (Rüger, 1999) aus. Es existieren neben der Gleichverteilung auch noch andere Verteilungen, die ein a priori Nichtwissen signalisieren. Genaueres Vorgehen und Beispiele für nichtinformative Priori-Verteilungen werden von Rüger (1999) beschrieben.
Eine Alternative sind die konjugierten Priori-Verteilungen. Eine Priori-Verteilung wird als zu einer Datenverteilung konjugiert bezeichnet, falls die daraus folgende Posteriori-Verteilung zum selben Verteilungstyp gehört wie die Priori-Verteilung. Durch die Annahme einer konjugierten Priori ist die Verteilungsfamilie der Posteriori-Verteilung bekannt. Die Parameterschätzung

ist bei einem bekannten Posteriori-Verteilungstyp einfacher, da die Integration und das Ziehen von Zufallszahlen aus einer bekannten Verteilung für die MCMC-Methoden (Abschnitt 2.2.4) zumeist implementiert sind.
Die hierarchischen Modelle des Bayesianischen Ridge, des Bayesianischen Lasso und des Bayesianischen Elastic Net sind sowohl über konjugierte als auch über nichtinformative Priori-Verteilungen definiert.

2.2.3 Bayesianisches lineares Modell

Das multiple lineare Regressionsmodell kann nicht nur frequentistisch sondern auch analog Bayesianisch formuliert werden. Im Bayesianischen Ansatz wird die Zielgröße als bedingte Verteilung der Parameter formuliert:

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

und somit folgt für die bedingte Verteilung der Zielvariablen:

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})).$$

Die unbekannten Parameter $\boldsymbol{\beta}$ und σ^2 wurden in den frequentistischen Modellen als fest angenommen. Im Bayesianischen Ansatz werden die Parameter als Zufallsvariablen angesehen auf Grund dessen auch Verteilungen für diese angenommen. Die gemeinsame Priori-Verteilung der unbekannten Parameter wird berechnet über:

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2).$$

Bei einer Normalverteilungsannahme für das Beobachtungsmodell ist eine konjugierte Priori-Verteilung für $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ die Normal-Inverse Chi-Quadrat-Verteilung (Fahrmeir *et al.*, 2007). Für die gemeinsame Posteriori-Verteilung:

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \cdot p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)$$

erhält man die Dichte einer Normal-Inverse Chi-Quadrat-Verteilung.

2.2.4 Markov Chain Monte Carlo Methoden

Die Posteriori-Verteilung kann analytisch und numerisch unzugänglich sein, sodass direkt keine Statistik der Posteriori-Verteilung berechnet werden kann. Eine iterative Lösung dieses Problems bieten die MCMC-Methoden. Eine Beschreibung der MCMC-Methoden wird von Robert und Casella (2004) und Fahrmeir *et al.* (2007) gegeben. Diese Methoden beruhen auf der Simulation von Zufallszahlen aus der Posteriori-Verteilung. Basierend auf der Verteilung

der Zufallszahlen können die Statistiken berechnet werden. Die Berechnung der Zufallszahlen aus der Posteriori-Verteilung erfolgt jedoch ohne direktes Ziehen aus der Posteriori-Verteilung. Stattdessen wird eine ergodische Markov Kette erzeugt, deren stationäre Verteilung die Posteriori-Verteilung ist. Die Markov-Kette konvergiert dann in Verteilung gegen die Posteriori-Verteilung. Um sicher zu stellen, Zufallszahlen aus einer akzeptabel approximierten Posteriori-Verteilung zu erhalten, sollte eine gewisse Konvergenzphase (engl.: burn in) gewährt werden. Die Glieder einer Markov Kette sind per Definition voneinander abhängig. Um möglichst unabhängige Stichproben aus der Markov Kette zu analysieren, kann die Markov Kette ausgedünnt werden, indem beispielsweise nur jede zwanzigste Ziehung berücksichtigt wird.
Die bekanntesten MCMC-Methoden sind der Metropolis-Hastings-Algorithmus von Metropolis *et al.* (1953) und Hastings (1970) und der Gibbs-Sampler von Geman und Geman (1984). Diese werden im Folgenden näher beschrieben.

Metropolis-Hastings-Algorithmus
Der Metropolis-Hastings-Algorithmus erzeugt wie im Folgenden beschrieben Zufallszahlen aus der Posteriori-Verteilung (Fahrmeir *et al.*, 2007):

1. Wähle einen Startwert $\boldsymbol{\theta}^{(0)}$ und die Anzahl der Iterationen T. Setze $t = 1$.

2. Ziehe eine Zufallszahl $\boldsymbol{\theta}^*$ aus der Vorschlagsdichte $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ und akzeptiere diese als neuen Zustand $\boldsymbol{\theta}^{(t)}$ mit Wahrscheinlichkeit $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$, anderenfalls setze $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

3. Falls $t = T$ beende den Algorithmus, ansonsten setze $t = t + 1$ und fahre fort mit 2.

Innerhalb des Algorithmus wird nicht unmittelbar aus der Posteriori-Verteilung gezogen, sondern aus einer Vorschlagsdichte $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$. Die Vorschlagsdichte ist von dem aktuellen Zustand $\boldsymbol{\theta}^{(t-1)}$ abhängig und sollte so gewählt werden, dass aus ihr leicht Zufallszahlen gezogen werden können. Die vorgeschlagenen Ziehungen $\boldsymbol{\theta}^*$ werden jeweils mit der Akzeptanzwahrscheinlichkeit von

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = \min\left\{\frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})\, q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y})\, q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}, 1\right\}$$

als neue Ziehungen angenommen. Auf diese Weise wird eine Markov Kette generiert. Die Zufallszahlen $\boldsymbol{\theta}^{(t_0+1)}, \ldots, \boldsymbol{\theta}^{(T)}$ können nach der Konvergenzphase $t_0 > 0$ als Stichprobe aus der Posteriori-Verteilung $p(\boldsymbol{\theta}|\boldsymbol{y})$ betrachtet werden. Die hintereinander gezogenen Zufallszahlen sollten möglichst un-

abhängig voneinander sein, sodass die benötigte Anzahl an Stichproben zur genauen Schätzung der Posteriori Eigenschaften gering ist.

Gibbs-Sampler
Der Gibbs-Sampler stellt eine Alternative zum Metropolis-Hastings-Algorithmus dar. Dieser Algorithmus ist insbesondere dann dem Metropolis-Hastings-Algorithmus vorzuziehen, wenn der Parametervektor hochdimensional ist. Der Gibbs-Sampler setzt voraus, dass die vollständig bedingten Dichten bekannt sind. Bei den Gibbs-Sampler geht der Vektor $\boldsymbol{\theta}$ nicht im Ganzen sondern über die S Teilvektoren $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S$ ein.
Der Gibbs-Sampler simuliert auf folgende Weise Zufallszahlen der Posteriori-Verteilung (Fahrmeir *et al.*, 2007):

1. Wähle Startwerte $\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_S^{(0)}$ und die Anzahl der Iterationen T. Setze $t = 1$.

2. Für $s = 1, \ldots, S$: Ziehe Zufallszahlen $\boldsymbol{\theta}_s^{(t)}$ aus der vollständig bedingten Dichte
$$p(\boldsymbol{\theta}_s | \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_{s+1}^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)}, \boldsymbol{y}).$$
Man beachte, dass in der Bedingung jeweils die momentan aktuellen Zustände verwendet werden.

3. Falls $t = T$ beende den Algorithmus, ansonsten setze $t = t + 1$ und fahre fort mit 2.

Innerhalb des Algorithmus wird nicht direkt aus der Posteriori-Verteilung gezogen, sondern aus den vollständig bedingten Dichten $p(\boldsymbol{\theta}_1|\cdot), \ldots p(\boldsymbol{\theta}_S|\cdot)$. Die Zufallszahlen $\boldsymbol{\theta}_s^{(t_0+1)}, \ldots, \boldsymbol{\theta}_s^{(T)}$ können nach der Konvergenzphase t_0 als Stichproben aus der Marginalverteilung von $\boldsymbol{\theta}_s|\boldsymbol{y}$ betrachtet werden. Im Vergleich zum Metropolis-Hastings-Algorithmus wird keine der Ziehungen verworfen, beziehungsweise liegt die Akzeptanzwahrscheinlichkeit hier bei Eins.

2.2.5 Modellkomplexität und Modellanpassung

Bayesianische Modelle können über das Devianz Informationskriterium und die Anzahl effektiver Parameter verglichen und beurteilt werden. Für den Vergleich Bayesianischer und frequentistischer Modelle eignet sich eine Kreuzvalidierung mit den Kriterien Korrelation und Mittlerer Quadratischer Fehler. Diese Komplexitäts- und Anpassungskriterien werden in dem folgenden Abschnitt vorgestellt.
Das Devianz Informationskriterium wurde von Spiegelhalter *et al.* (2002) zur Beurteilung der Modellgüte von Bayesianischen Modellen eingeführt. Es

basiert, wie die meisten Informationskriterien, auf der gleichzeitigen Betrachtung der Modellanpassung und der Modellkomplexität. Die Komplexität wird über die Anzahl der effektiven Parameter p_D spezifiziert:

$$\begin{aligned} p_D &= E_{\theta|y}[-2\log p(\boldsymbol{y}|\boldsymbol{\theta})] + 2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}) \\ &= E_{\theta|y}\left[-2\log \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta})}\right] + 2\log \frac{p(\bar{\boldsymbol{\theta}}|\boldsymbol{y})}{p(\bar{\boldsymbol{\theta}})}. \end{aligned}$$

Dabei steht $\bar{\boldsymbol{\theta}}$ für den Posteriori-Erwartungswert der Parameter $E(\boldsymbol{\theta}|\boldsymbol{y})$. Alternativ könnte auch der Posteriori-Modus oder Median gewählt werden. Die Anzahl der effektiven Parameter lässt sich auch über die unstandardisierte Devianz $D(\boldsymbol{\theta}) = -2\log p(\boldsymbol{y}|\boldsymbol{\theta})$ als die Differenz zwischen der erwarteten Devianz und der Devianz des Erwartungswerts berechnen:

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}).$$

Die erwartete Posteriori-Devianz kann als Maß für die Bayesianische Modellanpassung verwendet werden. Zusammen mit der Anzahl der effektiven Parameter resultiert das Devianz Informationskriterium:

$$\begin{aligned} \text{DIC} &= \overline{D(\boldsymbol{\theta})} + p_D \\ &= 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) \\ &= D(\bar{\boldsymbol{\theta}}) + 2p_D. \end{aligned}$$

Bei der Inferenz basierend auf Bayesianischen Modellen mit MCMC-Methoden ist das DIC schnell und einfach berechenbar. Es seien $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ die Zufallszahlen aus der Posteriori-Verteilung, welche während eines MCMC-Algorithmus gezogen wurden. Die erwartete Posteriori-Devianz wird über den Mittelwert der Devianzen der Zufallsstichproben $\overline{D(\boldsymbol{\theta})} = \frac{1}{T}\sum_{t=1}^{T} D(\boldsymbol{\theta}^{(t)})$ und der Posteriori-Erwartungswert der Parameter über den Mittelwert $\bar{\boldsymbol{\theta}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\theta}^{(t)}$ geschätzt. Das Devianz Informationskriterium, basierend auf den Posteriori Stichproben des MCMC-Algorithmus, errechnet sich dementsprechend über:

$$\text{DIC} = 2 \cdot \frac{1}{T}\sum_{t=1}^{T} D(\boldsymbol{\theta}^{(t)}) - D\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\theta}^{(t)}\right).$$

Ein Modell mit einem kleinen DIC-Wert ist einem Modell mit einem größeren DIC-Wert vorzuziehen (Spiegelhalter *et al.*, 2002).
Das DIC und p_D sind nur für den Vergleich Bayesianischer Modelle geeignet und geben keine Information über die Vorhersagegenauigkeit des Modells. Um die Vorhersagegüte frequentistischer und Bayesianischer Modelle zu bestimmen können die Kriterien Korrelation und Mittlerer Quadratischer Fehler herangezogen werden.

Im Folgenden wird die Vorhersagegenauigkeit des Modells, gemessen als die Korrelation zwischen den realen Daten $\boldsymbol{y}$ und den durch das Modell prognostizierten Werten $\hat{\boldsymbol{y}}$, betrachtet. Der Bravais-Pearson Korrelationskoeffizient ist wie folgt definiert (Fahrmeir *et al.*, 2003):

$$\rho = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{\hat{y}})^2}}.$$

Da ein Regressionsmodell speziell an die vorliegenden Daten angepasst wird sollte zur Beurteilung der Vorhersagegüte nicht der volle Datensatz analysiert werden, sondern die Daten in Trainings- und Validierungsdaten getrennt werden. Auf den Trainingsdaten wird das Regressionsmodell angepasst und mittels dieses Modells eine Vorhersage für die Validierungsdaten durchgeführt. Anschließend können die realen Werte der Validierungsdaten mit den prognostizierten Werten der Validierungsdaten verglichen werden. Um ein valides Ergebnis zu erhalten sollte dies mehrfach für verschiedene Validierungsdatensätze durchgeführt werden. Als systematische Methodik empfiehlt sich die Kreuzvalidierung (engl.: cross-validation, CV), welche unter anderem von Fahrmeir *et al.* (2007) und Hastie *et al.* (2009) beschrieben wird. Im Weiteren wird kurz und allgemein die K-fache Kreuzvalidierung beschrieben:

[a] Zerlegung der Daten in K Teildatensätze circa gleicher Größe.

[b] 1. Teildatensatz = Validierungsstichprobe, Parameterschätzung basierend auf den 2.-K. Teildatensätzen, Daten der Validierungsstichprobe prognostizieren, Prognosemaß z.B. Korrelation ρ_k berechnen.

[c] Jeweils 2. bis K. Teildatensatz als Validierungsstichprobe verwenden.

[d] Berechnung der Vorhersagegenauigkeit als $\bar{\rho} = \frac{1}{K} \sum_{k=1}^{K} \rho_k$ mit dem Prognosemaß ρ_k.

Ein Modell mit einem größeren $\bar{\rho}$-Wert ist einem Modell mit einem kleineren $\bar{\rho}$-Wert vorzuziehen. Alternativ zur Korrelation kann der Mittlere Quadratische Fehler betrachtet werden. Dieser ist definiert durch:

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Hastie *et al.* (2009) schlagen vor, die Daten in einen Trainingsdatensatz von 50%, einen Validierungsdatensatz von 25% und einen Testdatensatz von 25% aufzuteilen. Auf dem Trainingsdatensatz wird das Modell angepasst. Sei das Maß für die Vorhersagegenauigkeit der Mittlere Quadratische Fehler, so wird

der Validierungsdatensatz verwendet um den Vorhersagefehler für die Modellselektion zu schätzen. Mittels des Testdatensatzes wird der Generalisierungsfehler des Endmodells bestimmt.

2.3 Aufbau des Bayesianischen Elastic Net

Das Bayesianische Elastic Net ist eine Kombination des Bayesianischen Ridge und des Bayesianischen Lasso. Alle drei Ansätze basieren auf dem Bayesianischen linearen Modell, welches in dem Abschnitt 2.2.3 erläutert wurde. Die Beschreibung des Bayesianischen Ridge erfolgt in dem Abschnitt 2.3.1 und die des Bayesianischen Lasso findet in Abschnitt 2.3.2 statt. Vorschläge für die Wahl der Hyperparameter werden in Abschnitt 2.3.3 gegeben. Das Bayesianische Elastic Net wird in Abschnitt 2.3.4 erläutert.

2.3.1 Bayesianisches Ridge

Die Bayesianische Ridge Regression wird von Fahrmeir *et al.* (2010) und Pérez *et al.* (2010) beschrieben. Die Likelihood entspricht einer Normalverteilung (Abschnitt 2.2.3):

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma_\epsilon^2) = \prod_{i=1}^{n} \mathrm{N}\left(y_i| \sum_{j=1}^{p} x_{ij}\beta_j, \sigma_\epsilon^2\right).$$

Der Penalisierungsterm der frequentistischen Betrachtung $\mathrm{pen}(\boldsymbol{\beta}) = \lambda \cdot \boldsymbol{\beta}^\top \boldsymbol{\beta}$ ist äquivalent zur Normalverteilungspriori für $\boldsymbol{\beta}$ mit Erwartungswert 0 und Varianz σ_β^2 für jeden Marker:

$$p(\boldsymbol{\beta}|\sigma_\beta^2) = \prod_{j=1}^{p} \mathrm{N}(\beta_j|0, \sigma_\beta^2).$$

Die Information dieser Priori-Verteilung steigt an je kleiner die Varianz σ_β^2 ist. Die gemeinsame Posteriori-Verteilung der unbekannten Parameter wird im Allgemeinen errechnet über:

$$p(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\beta^2|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma_\epsilon^2) \cdot p(\boldsymbol{\beta}|\sigma_\beta^2) \cdot p(\sigma_\beta^2) \cdot p(\sigma_\epsilon^2).$$

Bei der Annahme von Konstanten für die Hyperparameter entspricht der Posteriori-Modus und Erwartungswert von $\boldsymbol{\beta}$ dem frequentistischen Ridge Schätzer:

$$\mathrm{E}(\boldsymbol{\beta}|\boldsymbol{y}) = \hat{\boldsymbol{\beta}}^R = (\mathbf{X}^\top\mathbf{X} + \frac{\sigma_\epsilon^2}{\sigma_\beta^2}\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{y},$$

mit $\lambda = \sigma_\epsilon^2/\sigma_\beta^2$. Die Bayesianische Betrachtung erlaubt zusätzlich die Spezifizierung von Priori-Verteilungen für die Parameter σ_ϵ^2 und σ_β^2. Die konjugierte Verteilung zur Normalverteilung ist die Inverse χ^2-Verteilung. A priori werden deshalb die folgenden Priori-Verteilungen spezifiziert:

$$p(\sigma_\epsilon^2) = \chi^{-2}(\sigma_\epsilon^2|df_\epsilon, S_\epsilon),$$
$$p(\sigma_\beta^2) = \chi^{-2}(\sigma_\beta^2|df_\beta, S_\beta),$$

mit den Freiheitsgraden df_ϵ und df_β und den Skalierungsparametern S_ϵ und S_β.

2.3.2 Bayesianisches Lasso

Park und Casella (2008), de los Campos *et al.* (2009) und Fahrmeir *et al.* (2010) definieren das Bayesianische Lasso. Der Penalisierungsterm der frequentistischen Betrachtung $\text{pen}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p |\beta_j|$ ist äquivalent zur Laplace Priori-Verteilung für $\beta_j, j = 1, \ldots, p$ (Fahrmeir *et al.*, 2010):

$$\beta_j|\lambda \sim \text{Laplace}(0, \lambda)$$

und

$$p(\boldsymbol{\beta}|\lambda) = \prod_{j=1}^p \frac{\lambda}{2}\exp(-\lambda|\beta_j|) \propto \exp(-\lambda \sum_{j=1}^p |\beta_j|).$$

Die Laplace Verteilung hat mehr Masse direkt um Null und mehr Masse in den Enden als die Normalverteilung. Dadurch tendiert das Bayesianische Lasso dazu kleine Effekte stärker und größe Effekte schwächer zu schrumpfen als das Bayesianische Ridge. Bei der Annahme von Konstanten für die Hyperparameter entspricht die Posteriori-Modus Schätzung der frequentistischen Lasso Schätzung:

$$p(\boldsymbol{\beta}, \lambda|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j|\right) \cdot const.$$

Die Laplaceverteilung kann als skalierte Mischung von Normalverteilungen mit einer Exponentialverteilung als Mischungsdichte formuliert werden (Park und Casella, 2008):

$$\frac{\lambda}{2}\exp(-\lambda|\beta_j|) = \int_0^\infty \left[\frac{\exp(-(\beta_j^2/2\sigma_j^2))}{\sqrt{2\pi\sigma_j^2}}\right]\left[\frac{\lambda^2}{2}\exp(-\frac{\lambda^2}{2}\sigma_j^2)\right] d\sigma_j^2.$$

Das hierarchische Modell lautet damit:

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma_\epsilon^2) &= \prod_{i=1}^{n} \mathrm{N}(y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_\epsilon^2), \\
p(\boldsymbol{\beta}|\sigma_\epsilon^2, \boldsymbol{\tau}^2) &= \prod_{j=1}^{p} \mathrm{N}(\beta_j|0, \tau_j^2\sigma_\epsilon^2), \\
p(\sigma_\epsilon^2) &= \chi^{-2}(\sigma_\epsilon^2|df_\epsilon, S), \\
p(\boldsymbol{\tau}^2|\lambda) &= \prod_{j=1}^{p} \mathrm{Exp}(\tau_j^2|\lambda), \\
p(\lambda^2) &= \mathrm{Ga}(\lambda^2|\alpha_1, \alpha_2).
\end{aligned}
$$

Durch die markerspezifische Varianz $\tau_j^2\sigma_\epsilon^2$ der bedingten Priori-Verteilung $p(\boldsymbol{\beta}|\sigma_\epsilon^2, \boldsymbol{\tau}^2)$ wird, im Vergleich zum Bayesianischen Ridge, eine markerspezifische Schrumpfung der Koeffizientenschätzer erlaubt. Je kleiner der τ_j Parameter desto informativer ist die Priori-Verteilung. Für die Priori-Verteilung von $p(\lambda^2)$ kann anstelle der von Park und Casella (2008) vorgeschlagenen Gammaverteilung auch eine Betaverteilung verwendet werden (de los Campos *et al.*, 2009).
Im Bayesianischen Lasso ist, im Vergleich zum frequentistischen Lasso die Anzahl der selektierbaren Prädiktoren nicht durch die Anzahl der Beobachtungen beschränkt (de los Campos *et al.*, 2009).

2.3.3 Wahl der Hyperparameter

Für die Wahl der Hyperparameter des Bayesianischen Lasso und des Bayesianischen Ridge schlagen Pérez *et al.* (2010) die sogenannten optimalen Parameter vor. Die Wahl der optimalen Parameter stützt sich auf die Heritabilität (Griffiths *et al.*, 2012). Die Heritabilität basiert auf der Annahme, dass die phänotypische Ausprägung (P) von den genotypischen Ausprägungen (G) und der Umwelt (ϵ) abhängt. Die phänotypische Varianz σ_G^2 ist, falls keine Genotyp-Umwelt-Interaktion vorliegt, durch die Summe der genotypischen Varianz und der Umwelt-Varianz definiert: $\sigma_P^2 = \sigma_G^2 + \sigma_\epsilon^2$. Die Heritabilität h^2, also der Anteil der genotypischen Varianz an der phänotypischen Varianz, ist definiert durch:

$$
h^2 = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_\epsilon^2}. \tag{2.3}
$$

Durch Umformen der Gleichungen (2.3) ergibt sich für die genotypische Varianz die Schätzung $\sigma_G^2 = \sigma_P^2 \cdot h^2$ und für die nicht genetische Umwelt-Varianz

folgt:

$$\sigma_\epsilon^2 = \sigma_P^2(1 - h^2). \tag{2.4}$$

Der Skalierungsparameter der inversen χ^2-Priori von σ_ϵ^2 wird geschätzt durch:

$$S_\epsilon = \mathrm{E}(\sigma_\epsilon^2) \cdot (df_\epsilon + 2),$$

wobei die Schätzung des Erwartungswert von σ_ϵ^2 über die nicht genetische Umwelt-Varianz (2.4) erfolgt und für den Freiheitsgrad $df_\epsilon = 4.1$ angenommen wird. Bei der inversen χ^2-Priori von σ_β^2 lautet der Schätzer des Skalierungsparameters:

$$S_\beta = \frac{\sigma_G^2(df_\beta + 2)}{\sum_j \bar{x}_j^2},$$

wobei für den Freiheitsgrad $df_\beta = 4.1$ angenommen wird. Der optimale Parameter λ wird wie folgt geschätzt:

$$\lambda = \sqrt{2\frac{1 - h^2}{h^2}\sum_j \bar{x}_j^2}. \tag{2.5}$$

2.3.4 Bayesianischer Elastic Net Schätzer

Das Bayesianische Elastic Net kombiniert die Methode des Bayesianischen Ridge und des Bayesianischen Lasso. Durch die Kombination dieser Methoden besitzt das Bayesianische Elastic Net sowohl die positiven Eigenschaften des Lasso als auch die Vorteile des Ridge. Li und Lin (2010) definieren den Bayesianischen Elastic Net Schätzer und beschreiben die Inferenz dieser Methode.
Der Penalisierungsterm der frequentistischen Betrachtung:

$$\mathrm{pen}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

ist äquivalent zur Kombination der Normalverteilungspriori und Laplaceverteilungspriori für $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\sigma^2) \propto \exp\left\{ -\frac{1}{2\sigma^2}(\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2) \right\}.$$

Wird eine Konstante für die Priori-Verteilung $p(\sigma^2)$ angenommen, so entspricht der Posteriori-Modus Schätzer dem frequentistischen Elastic Net Schätzer. Li und Lin (2010) spezifizieren die nichtinformative Priori-Verteilung

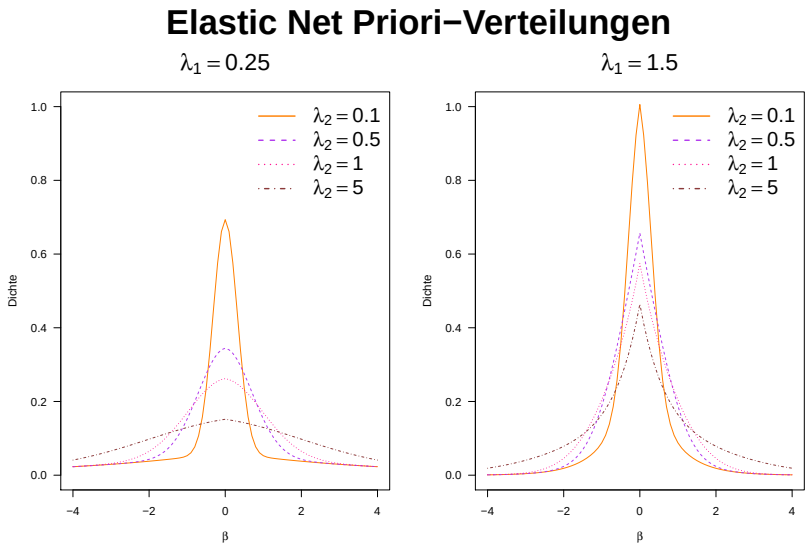

Abbildung 3: Vergleich der Priori-Verteilungen des Bayesianischen Elastic Net für verschiedene Werte von λ_1 und λ_2

$p(\sigma^2) = 1/\sigma^2$. In der Abbildung 3 werden Priori-Verteilungen für $\boldsymbol{\beta}$ des Bayesianischen Elastic Net für verschiedene Werte von λ_1 und λ_2 gezeigt. Je größer der λ_1 Parameter oder je größer der λ_2 Parameter, desto mehr Masse der Priori-Verteilung konzentriert sich um Null und desto stärker ist die Penalisierung. Einen Vergleich der Priori-Verteilungen des Bayesianischen Ridge, des Bayesianischen Lasso und des Bayesianischen Elastic Net zeigt Abbildung 4. Die Priori-Verteilung der Ridge Schätzung ist in Null differenzierbar. Bei dem Lasso und dem Elastic Net sind die Priori-Verteilungen nicht in Null differenzierbar. Dies führt dazu, dass bei der Ridge Schätzung keine Variablenselektion erfolgt und bei der Lasso beziehungsweise Elastic Net Schätzung hingegen schon. Die Priori-Verteilung des Elastic Net ist flacher als die Priori-Verteilung des Lasso. Die Variablenselektion ist bei dem Lasso stärker als bei dem Elastic Net.
Die marginale Posteriori-Verteilung für $\boldsymbol{\beta}$ ist wie folgt definiert:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \int_0^\infty p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \cdot p(\boldsymbol{\beta}|\sigma^2) \cdot p(\sigma^2) d\sigma^2 =$$

$$\int_0^\infty \frac{C(\lambda_1, \lambda_2, \sigma^2)}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2}{2\sigma^2} \right\} p(\sigma^2)\, d\sigma^2,$$

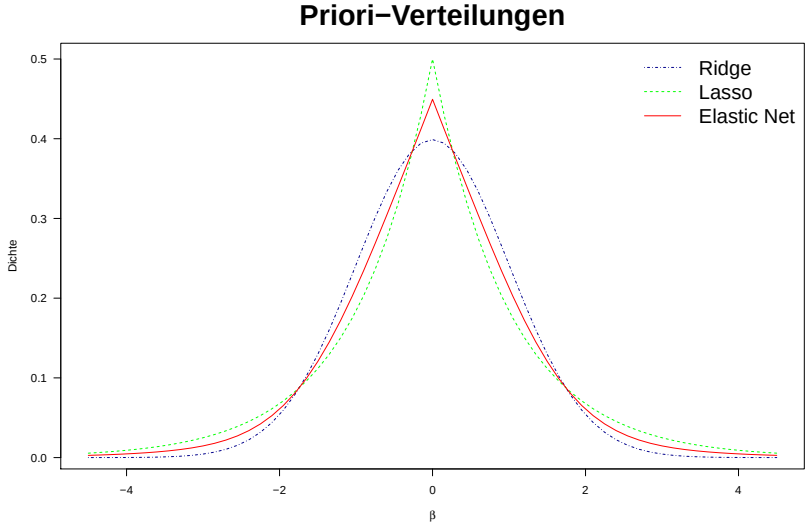

Abbildung 4: Vergleich der Priori-Verteilungen des Bayesianischen Ridge, des Bayesianischen Lasso und des Bayesianischen Elastic Net für $\lambda_1 = 1$, $\lambda_2 = 1$

wobei $C(\lambda_1, \lambda_2, \sigma^2)$ eine Normalisierungskonstante darstellt. Eine geschlossene Darstellung der Posteriori Schätzer basierend auf deren marginaler Posteriori-Verteilung ist nicht immer möglich. Für die Inferenz kann deswegen der Gibbs-Sampler verwendet werden. Im Gibbs-Sampler werden die vollständig bedingten Dichten der Parameter verwendet. Auf Grund der $|\beta_j|$ würde aus dem hier vorgestellten hierarchischen Modell eine unbekannte vollständig bedingte Verteilung folgen. Deshalb schlagen Li und Lin (2010) ein anderes hierarchisches Modell vor, welches auf einem Umformulierung der Priori-Verteilung $p(\boldsymbol{\beta}|\sigma^2)$ beruht:

$$C(\lambda_1, \lambda_2, \sigma^2) \prod_{j=1}^{p} \int_1^\infty \sqrt{\frac{t}{t-1}} \exp\left\{ -\frac{\beta_j^2}{2} \left(\frac{\lambda_2}{\sigma^2} \frac{t}{t-1} \right) \right\} t^{-1/2} \exp\left(-\frac{1}{2\sigma^2} \frac{\lambda_1^2}{4\lambda_2} t \right) dt.$$

Dies zeigt, dass die Verteilung von $\beta_j|\sigma^2$ als eine Mischung von Normalverteilungen $\mathrm{N}(0, \sigma^2(t-1)/(\lambda_2 t))$ darstellt werden kann, wobei die Mischverteilung eine auf $(1, \infty)$-trunkierte Gammaverteilung mit Gestaltparameter 0.5 und Skalierungsparameter $8\lambda_2\sigma^2/\lambda_1$ ist. Daraus resultiert folgendes hierarchisches

Modell:

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) &= \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}), \\
p(\boldsymbol{\beta}|\boldsymbol{\tau}, \sigma^2) &= \prod_{j=1}^{p} \mathrm{N}\left(0, \left(\frac{\lambda_2}{\sigma^2}\frac{\tau_j}{\tau_j - 1}\right)^{-1}\right), \\
p(\boldsymbol{\tau}|\sigma^2) &= \prod_{j=1}^{p} \mathrm{TG}\left(\frac{1}{2}, \frac{8\lambda_2\sigma^2}{\lambda_1^2}, (1, \infty)\right), \\
p(\sigma^2) &= \frac{1}{\sigma^2}.
\end{aligned}
$$

Die vollständig bedingten Dichten sind folgende:

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \boldsymbol{\tau}) &= \mathrm{N}(\mathbf{A}^{-1}\mathbf{X}^\top\boldsymbol{y}, \sigma^2\mathbf{A}^{-1}) \text{ , mit} \\
&\qquad \mathbf{A} = \mathbf{X}^\top\mathbf{X} + \lambda_2 \mathrm{diag}\left(\frac{\tau_1}{\tau_1 - 1}, \ldots, \frac{\tau_p}{\tau_p - 1}\right), \\
p((\boldsymbol{\tau} - \mathbf{1})|\boldsymbol{y}, \sigma^2, \boldsymbol{\beta}) &= \prod_{j=1}^{p} \mathrm{GIG}\left(\lambda = \frac{1}{2}, \psi = \frac{\lambda_1}{4\lambda_2\sigma^2}, \chi = \frac{\lambda_2\beta_j^2}{\sigma^2}\right), \\
p(\sigma^2|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau}) &= \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+p+1} \left\{\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\right\}^{-p} \cdot \\
&\qquad \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_2\sum_{j=1}^{p}\frac{\tau_j}{\tau_j - 1}\beta_j^2 + \frac{\lambda_1^2}{4\lambda_2}\sum_{j=1}^{p}\tau_j\right\}\right],
\end{aligned}
$$

wobei $\Gamma_u(\alpha, x) = \int_x^\infty t^{\alpha-1}e^{-t}dt$ und $\mathrm{GIG}(\lambda, \psi, \chi)$ die generalisierte inverse Gammaverteilung mit der Dichte:

$$
p(x|\lambda, \chi, \psi) = \frac{(\psi/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\psi\chi})}x^{\lambda-1}\exp\left\{-\frac{1}{2}(\chi x^{-1} + \psi x)\right\}
$$

und $K_\lambda(\cdot)$ die modifizierte Bessel Funktion mit Ordnung λ.

Hofmarcher *et al.* (2011) schlagen vor für das Bayesianische Elastic Net eine $\boldsymbol{\beta}$ Priori-Verteilung zu wählen, welche einer spike & slab Mischung entspricht. Für jedes β_j wird eine Mischverteilung aus einer Punktmasse auf Null $\mathbf{I}_0$ und der üblichen β_j Priori-Verteilung angenommen:

$$
p(\beta_j|\gamma_j, \tau_j, \sigma^2) \sim (1 - \gamma_j) \cdot \mathbf{I}_0 + \gamma_j \cdot p(\beta_j|\tau_j, \sigma^2)
$$

und für γ_j wird eine Bernoulli Priori-Verteilung gewählt: $p(\gamma_j) = Be(\underline{\gamma})$, wobei $\underline{\gamma} = \bar{p}/p$. Dabei kann $\bar{p}$ als die a priori erwartete Anzahl der Parameter

ungleich Null interpretiert werden. Auf diese Weise ist es möglich, dass ein a priori Wissen über die Modellgröße in die Inferenz eingeht. Sollte kein a priori Wissen über die Modellgröße vorliegen, so resultiert mit $\bar{p} = p$ das klassische Bayesianische Elastic Net. Bei der Inferenz der Arabidopsis liegt kein Vorwissen über die Modellgröße vor.

Wahl der Penalisierungsparameter
Die Wahl der Penalisierungsparamter entscheidet über die Form der Priori-Verteilung von $\boldsymbol{\beta}$ und ist somit sehr wichtig für die Inferenz. Die Wahl der Penalisierungsparameter wird empirisch und iterativ über den Monte Carlo EM Algorithmus (Casella, 2001) getroffen. Dieser Algorithmus maximiert approximativ die marginale Likelihood. Grundidee in der Penalisierungsparameterschätzung durch den Monte Carlo EM Algorithmus ist es $\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2$ als fehlende Daten und (λ_1, λ_2) als feste Parameter zu behandeln. Die Likelihood, ohne Konstanten bezüglich der festen Parameter, ist folgende (Li und Lin, 2010):

$$\lambda_1^p \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+p+1} \left\{\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\right\}^{-p} \prod_{j=1}^{p} \left(\frac{1}{\tau_j - 1}\right)^{1/2} \cdot$$
$$\exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_2 \sum_{j=1}^{p} \frac{\tau_j}{\tau_j - 1}\beta_j^2 + \frac{\lambda_1}{4\lambda_2}\sum_{j=1}^{p}\tau_j\right\}\right]$$

und die logarithmierte Likelihood entsprechend:

$$p\log(\lambda_1) - p\log\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right) - \frac{\lambda_2}{2\sigma^2}\sum_{j=1}^{p}\frac{\tau_j}{\tau_j - 1}\beta_j^2 - \frac{1}{2\sigma^2}\frac{\lambda_1^2}{4\lambda_2}\sum_{j=1}^{p}\tau_j.$$

Die auf $\lambda^{(k-1)} = (\lambda_1^{(k-1)}, \lambda_2^{(k-1)})$ und Y bedingte logarithmierte Likelihood im k-ten Schritt des Monte Carlo EM Algorithmus lautet wie folgt:

$$Q(\lambda|\lambda^{(k-1)}) = p\log(\lambda_1) - p\mathrm{E}\left[\log\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)|\lambda^{(k-1)}, Y\right] -$$
$$\frac{\lambda_2}{2}\sum_{j=1}^{p}\mathrm{E}\left[\frac{\tau_j}{\tau_j - 1}\frac{\beta_j^2}{\sigma^2}|\lambda^{(k-1)}, Y\right] - \frac{\lambda_1^2}{8\lambda_2}\sum_{j=1}^{p}\mathrm{E}\left[\frac{\tau_j}{\sigma^2}|\lambda^{(k-1)}, Y\right] + const =$$
$$= R(\lambda|\lambda^{(k-1)}) + const.$$

Dies ist der E-Schritt des EM-Algorithmus. In dem M-Schritt wird $R(\lambda|\lambda^{(k-1)})$ maximiert:

$$\frac{dR}{d\lambda_1} = \frac{p}{\lambda_1} + \frac{p\lambda_1}{4\lambda_2}\mathrm{E}\left[\left\{\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\right\}^{-1}\phi\left(\frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\frac{1}{\sigma^2}|\lambda^{(k-1)}, Y\right] - \frac{\lambda_1}{4\lambda_2}\sum_{j=1}^{p}\mathrm{E}\left[\frac{\tau_j}{\sigma^2}|\lambda^{(k-1)}, Y\right],$$

$$\frac{dR}{d\lambda_2} = -\frac{p\lambda_1^2}{8\lambda_2^2}\mathrm{E}\left[\left\{\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right\}^{-1}\phi\left(\frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\frac{1}{\sigma^2}|\lambda^{(k-1)}, Y\right] - \frac{1}{2}\sum_{j=1}^{p}\mathrm{E}\left[\frac{\tau_j}{\tau_j - 1}\frac{\beta_j^2}{\sigma^2}|\lambda^{(k-1)}, Y\right] + \frac{\lambda_1^2}{8\lambda_2^2}\sum_{j=1}^{p}\mathrm{E}\left[\frac{\tau_j}{\sigma^2}|\lambda^{(k-1)}, Y\right].$$

mit $\phi(t) = t^{-1/2}e^{-t}$.
Eine andere Möglichkeit die Penalisierungsparameter zu spezifizieren ist es, Priori-Verteilungen für diese zu wählen. Folgende Priori-Verteilungen könnten hierfür gewählt werden (Li und Lin, 2010):

$$p(\lambda_1^2) = \mathrm{Ga}(a, b),$$
$$p(\lambda_2) = \mathrm{GIG}(1, c, d).$$

Hofmarcher *et al.* (2011) verwenden in ihrer Datenauswertung ebenfalls Priori-Verteilungen für die Penalisierungsparameter, welche jedoch so gewählt werden, dass der Lasso und Ridge Parameter in einem gewissen Zusammenhang stehen. Hierfür werden die Penalisierungsparameter λ_1 und λ_2 so reparametrisiert, dass $\lambda_1 = \alpha \cdot \lambda$ und $\lambda_2 = (1 - \alpha)\lambda$. Für α wird a priori eine auf $(0, 1)$-trunkierte Normalverteilung mit Erwartungswert 0.5 und Varianz 0.000001 angenommen. Dies scheint vorerst eine strenge Annahme zu sein. Ob dies tatsächlich eine strenge Einschränkung ist wird in der Inferenz (Kapitel 4) über verschiedene Annahmen für den Erwartungswert und die Varianz überprüft. Für λ^2 wird a priori die Gamma-Priori $p(\lambda^2) = \mathrm{Ga}(0.1, 0.1)$ spezifiziert. Desweiteren wird a priori für den Intercept die Priori-Verteilung $p(\mu) = \mathrm{N}(0, 0.000001)$ angenommen. Diese Normalverteilungspriori hat fast die gesamte Masse auf Null. Diese Modellierung ist adäquat für standardisierte Größen. Desweiteren wird folgende Priori-Verteilung der Varianz definiert: $p(\sigma^2) = \mathrm{Ga}(0.001, 0.001)$.
Durch die simultane Schätzung von λ_1 und λ_2 im Bayesianischen Ansatz tritt, im Vergleich zur frequentistischen Schätzung, kein Effekt der Doppel-Schrumpfung auf.

2.4 Generalisiertes Elastic Net

Das Generalisierte Elastic Net wurde von Ishwaran und Rao (2011) eingeführt und ist eine geeignete Inferenzmethode bei hochdimensionalen Problemen. Zur Definition und Implementierung des Generalisierten Elastic Net werden Bayesianische Modell Mittelwert (engl.: bayesian model average, BMA) Schätzer genutzt.
In Abschnitt 2.1.1 wurde der Ridge Schätzer erläutert. Dieser Schätzer kann auch allgemeiner mittels individueller Penalisierungsparamter für jeden Koeffizientenschätzer wie folgt formuliert werden:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GR} &= \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\{(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})+\sum_{j=1}^{p}\lambda_j\beta_j^2\}\\ &= (\mathbf{X}^\top\mathbf{X}+\boldsymbol{\Lambda})^{-1}\mathbf{X}^\top\boldsymbol{y}. \qquad (2.6)\end{aligned}$$

Dabei wird mit $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_j\}_{j=1}^p$, $\lambda_j > 0$ die Ridge Matrix der Penalisierungsparameter bezeichnet. Diese verallgemeinerte Form der Ridge Methode wird als Generalisierte Ridge Regression (engl.: generalized ridge regression, GRR) bezeichnet. Ishwaran und Rao (2011) zeigen, dass der Generalisierte Ridge Schätzer eine Schrumpfung der Regressionskoeffizienten auf exakt Null zulässt. Dies könnte den Generalisierten Ridge Schätzer zu einer adäquaten Methode in $p \gg n$-Situationen machen. Der Generalisierte Ridge Schätzer (2.6) kann auch über $\mathbf{X}_* = \mathbf{X}\boldsymbol{\Lambda}$ als reskalierter Ridge Schätzer dargestellt werden:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GR} &= \boldsymbol{\Lambda}^{-1/2}(\boldsymbol{\Lambda}^{-1/2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Lambda}^{-1/2}+\mathbf{I}_p)^{-1}\boldsymbol{\Lambda}^{-1/2}\mathbf{X}^\top\boldsymbol{y}\\ &= \boldsymbol{\Lambda}^{-1/2}(\mathbf{X}_*^\top\mathbf{X}_*+\mathbf{I}_p)^{-1}\mathbf{X}_*^\top\boldsymbol{y}\\ &= \boldsymbol{\Lambda}^{-1/2}\hat{\boldsymbol{\beta}}_R^*,\end{aligned}$$

wobei $\hat{\boldsymbol{\beta}}_R^* = (\mathbf{X}_*^\top\mathbf{X}_*+\mathbf{I}_p)^{-1}\mathbf{X}_*^\top\boldsymbol{y}$ der Ridge Schätzer mit der Design-Matrix $\mathbf{X}_*$ und $\lambda = 1$. Anhand der Geometrie des Generalisierten Ridge Schätzers zeigen Ishwaran und Rao (2011), dass der Schätzer $\hat{\boldsymbol{\beta}}_{GR}$ effektiv bei der Variablenselektion in $p \gg n$-Situationen ist.
Ideale Variablenselektion, also die korrekte Identifikation aller wahren Nullkoeffizienten in den Steigungsparametern, kann für den Generalisierten Ridge Schätzer nur garantiert werden, falls die wahre Anzahl an nicht-Nullkoeffizienten deutlich kleiner als die Anzahl der Beobachtungen ist (Ishwaran und Rao, 2011). Gesetz des Falles, dass das wahre Modell mehr wahre nicht-Nullkoeffizienten als Beobachtungen enthält, sollte eine Linearkombination von Generalisierten Ridge Schätzer verwendet werden. Diese Linearkombination der Schätzer wird als gewichteter Generalisierter Ridge Schätzer (engl.:

weighted generalized Ridge estimator, WGRR) bezeichnet. Bei dem von Ishwaran und Rao (2011) geschilderten Bayesianischen Modell zur Berechnung des WGRR Schätzers resultiert der Spezialfall von BMA Schätzern.

In die Schätzung des Generalisierten Elastic Net gehen, zur Verkürzung der Rechenzeit, nur die größten Effekte des BMA Schätzers ein. Die Variablen werden nach ihren Absolutwerten des BMA Schätzers geordnet und es wird eine Designmatrix $\mathbf{X}^*$ definiert, welche nur jene geordneten Variablen beinhaltet, für welche die BMA Effekte streng positiv sind. Die Spaltendimension von $\mathbf{X}^*$ sei K. Bayesianische Schätzer, wie der BMA Schätzer, benötigen Ad-hoc Methoden zur Variablenselektion. Diese sind beispielsweise das Vertrauensintervall-Kriterium und das skalierte Umgebungs-Kriterium (Li und Lin, 2010). Für die Inferenz der Arabidopsis (Kapitel 4) werden anstelle der Ad-hoc Methoden einfach nur die größten K Effekte des BMA Schätzers selektiert. Für K wird in Kapitel 4 die Anzahl der Individuen n gewählt. Die Designmatrix und die Zielgröße seien standardisiert.
Das Generalisierte Elastic Net stellt eine Verallgemeinerung des Elastic Net dar. Wie für das Generalisierte Ridge werden individuelle Parameter für die L_2 Penalisierung spezifiziert. Der Generalisierte Elastic Net Schätzer ist wie folgt definiert:

$$\hat{\boldsymbol{\beta}}^*_{GEN} = \underset{\beta\in\mathbb{R}^K}{\operatorname{argmin}} \left\{ (\boldsymbol{y} - \mathbf{X}^*\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}^*\boldsymbol{\beta}) + \sum_{k=1}^{K} \lambda_k \beta_k^2 + \lambda_0 \sum_{k=1}^{K} |\beta_k| \right\}, \quad (2.7)$$

wobei $(\lambda_k)_{k=1}^K$ und λ_0 feste, positive Parameter sind. Analog zum Elastic Net (Abschnitt 2.1.3) kann gezeigt werden, dass es sich bei der Berechnung des Generalisierten Elastic Net Schätzers um ein L_1 Optimierungsproblem handelt:

$$\hat{\boldsymbol{\beta}}^*_{GEN} = \underset{\beta\in\mathbb{R}^K}{\operatorname{argmin}} \left\{ (\boldsymbol{y}_A - \mathbf{X}^*_A\boldsymbol{\beta})^\top (\boldsymbol{y}_A - \mathbf{X}^*_A\boldsymbol{\beta}) + \lambda_0 \sum_{k=1}^{K} |\beta_k| \right\},$$

mit

$$\mathbf{X}^*_A = \begin{pmatrix} \mathbf{X}^* \\ \boldsymbol{\Lambda}^{1/2} \end{pmatrix}_{(n+K)\times K}, \boldsymbol{y}_A = \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix}_{n+K},$$

und $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_k\}_{k=1}^K$. Grafisch entspricht dies der Minimierung des Ellipsoids um den Generalisierten Ridge Schätzer $\hat{\boldsymbol{\beta}}^*_{GR} = (\mathbf{X}^{*\top}\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}^{*\top}\boldsymbol{y}$ unter der Nebenbedingung $\sum_{k=1}^K |\beta_k| < L$ für ein $L > 0$.
Die Berechnung des Generalisierten Elastic Net Schätzers in der praktischen Anwendung erfolgt in drei Schritten:

1. Berechnung der $(\lambda_k)_{k=1}^K$ Penalisierungsparameter durch:

$$\lambda_k = \frac{|a_k|}{K^{-1}\sum_{k=1}^K |a_k|} \cdot \frac{\sqrt{n}}{|\hat{\boldsymbol{\beta}}^*_{BMA,k}|}, k = 1, \ldots, K$$

   wobei $a_k = \sqrt{n}(\mathbf{X}^{*\top}\boldsymbol{y})_k - (\mathbf{X}^{*\top}\mathbf{X}^*\hat{\boldsymbol{\beta}}^*_{BMA})_k$ und $\hat{\boldsymbol{\beta}}^*_{BMA}$ der BMA Schätzer.

2. $(\lambda_k)_{k=1}^K$ Parameter als fest annehmen und Erstellung des λ_0-Lösungspfades der Schätzgleichung (2.7) durch den LARS Algorithmus (Efron *et al.*, 2004).

3. Finales Modell mit der Pfadlösung wählen, welches das Modell mit dem kleinsten Akaike Informationskriterium liefert.

Ishwaran und Rao (2011) beweisen, dass für den Generalisierten Elastic Net Schätzer die Fan-Li *Oracle Property* (Fan und Li, 2001) gilt. Die *Oracle Property* besagt, dass ein sparsamer und asymptotisch normalverteilter Schätzer dieselbe Grenzverteilung besitzt wie der KQ Schätzer beschränkt auf die wahren nicht-Nullkoeffizienten. Das bedeutet, dass die Methodik genauso gut funktioniert, als wenn das wahre Modell schon zuvor bekannt gewesen wäre.

3 Beschreibung der Arabidopsis thaliana Daten

In diesem Kapitel werden die Genotypen und Phänotypen des Datensatzes zu Arabidopsis deskriptiv und explorativ analysiert. Die Daten sind Teil der MAGIC (Multiparent Advanced Generation Inter-Cross)-Population (Kover *et al.*, 2009). Basierend auf 19 Arabidopsis Stämmen wurden diverse Kreuzungen durchgeführt um ein weites genetisches Spektrum zu erhalten. Weitere Details der Datenerhebung werden von Kover *et al.* (2009) gegeben. Die resultierenden Daten sind öffentlich erhältlich auf `http://spud.well.ox.ac.uk/arabidopsis/`.
Die phänotypischen Merkmale, welche im Weiteren betrachtet werden, sind die Anzahl an Tagen zwischen dem Schossbeginn und der Blütezeit, die Anzahl an Tagen bis zum Schossbeginn, die absolute Höhe der Pflanzen in Zentimetern und die Wachstumsrate. Die Wachstumsrate wird errechnet als das Residuum einer einfachen linearen Regression, wobei die Einflussgröße die Anzahl an Blättern am Tag 28 nach Säen der Saat und die Zielgröße die Anzahl der Tage bis zur Keimung ist. Die Verteilungen der Anzahl der Tage zwischen Schossbeginn und Blütezeit und der Anzahl der Tage bis zum Schossbeginn sind linkssteil. Um symmetrischere Verteilungen zu erhalten werden diese Variablen zukünftig ausschließlich logarithmiert betrachtet. Die Verteilungen der Merkmale der 426 phänotypisierten Individuen werden über univariate Histogramme und der Zusammenhang der Merkmale über bivariate Streudiagramme und über bivariate Korrelationen in Abbildung 5 dargestellt. Die Merkmale sind paarweise signifikant positiv korreliert. Die Zeit zwischen dem Schossbeginn und der Blütezeit, die Pflanzenhöhe und die Wachstumsrate gleichen visuell einer Normalverteilung. Der Kolmogorov-Smirnov-Test auf Normalverteilung lehnt die Normalverteilung jedoch für alle Phänotypen ab. Der Grund ist aus den QQ-Plots in Abbildung 6 ersichtlich. Im Zentrum der Daten stimmen die theoretischen Quantile der Normalverteilung mit den empirischen Quantilen überein. An den Rändern der Verteilungen treten jedoch Abweichungen auf.
Genotypisiert wurden die Arabidopsis Individuen mit 1260 SNPs auf fünf Chromosomen. Der Anteil fehlender Werte in der Marker-Matrix beläuft sich auf 2.93%. Für die weitere Auswertung werden nur jene SNPs betrachtet, deren Anteil an fehlenden Werten kleiner als 10% ist und deren Häufigkeit des seltenen Allels (engl.: minor allele frequency, MAF) größer als 5% ist. Damit reduziert sich der Anteil der fehlenden Werte auf 1.64%. Auf Grund des geringen Anteils werden die fehlenden Werte gemäß der Randverteilung der SNP Ausprägungen ersetzt. Für die Analyse verbleiben 1073 SNPs. Es sind

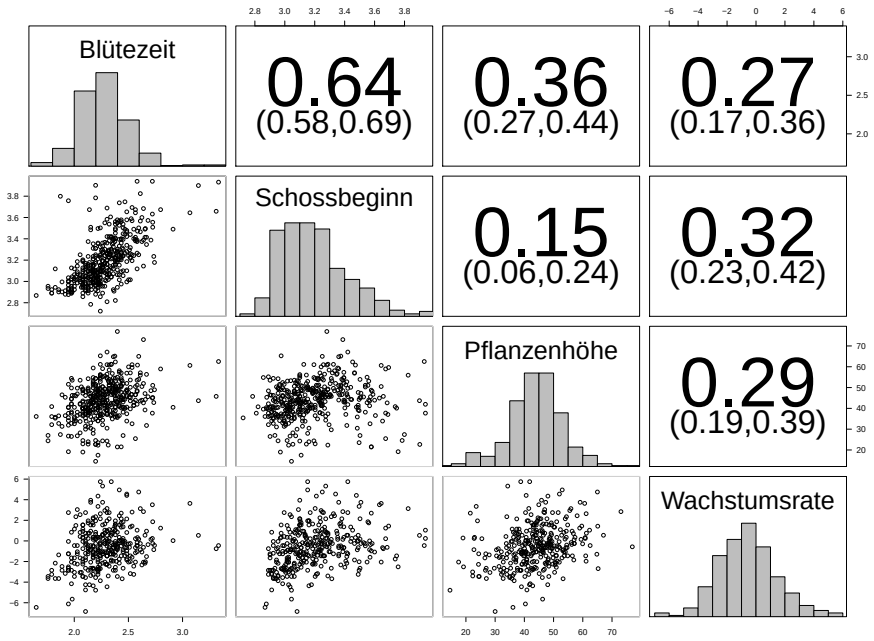

Abbildung 5: Histogramme, Streudiagramme, Korrelationen und Konfidenzintervalle der Korrelationen für die Merkmale: Log. Anzahl an Tagen zwischen Schossbeginn und Blütezeit, Log. Anzahl an Tagen bis Schossbeginn, Pflanzenhöhe in Zentimetern und Wachstumsrate

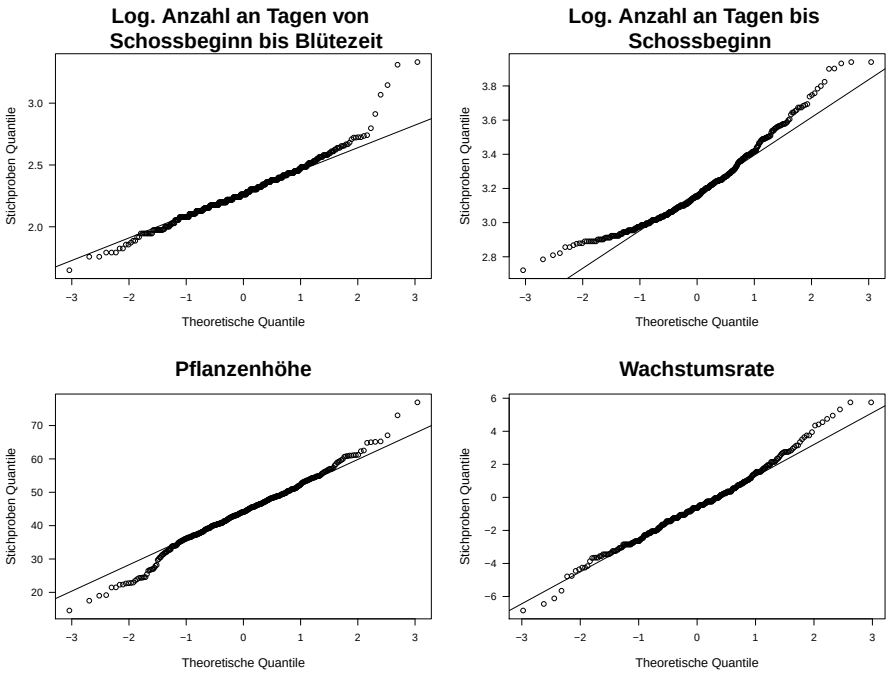

Abbildung 6: Normal-QQ-Plot für die Merkmale: Log. Anzahl an Tagen zwischen Schossbeginn und Blütezeit, Log. Anzahl an Tagen bis Schossbeginn, Pflanzenhöhe in Zentimetern und Wachstumsrate

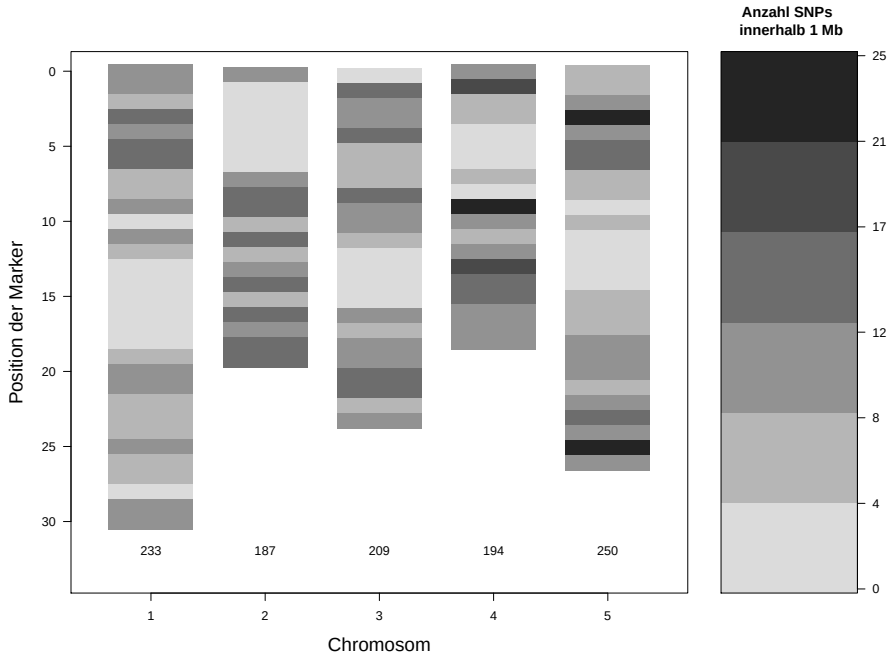

Abbildung 7: Markerdichte auf den fünf Chromosomen

nur homozygote Genotypen vertreten, da es sich bei den MAGIC-Individuen um Inzuchtlinien handelt. Die relative Häufigkeit des seltenen Allels beträgt gemittelt über alle Marker 23.8%. Die kleinste relative Häufigkeit des seltenen Allels ist 5.1% und die größte Häufigkeit des seltenen Alles beträgt 49.9%.
In der Abbildung 7 ist die Dichte der Marker über der Anzahl der SNP-Marker pro 1 Mb abgetragen. Die durchschnittliche Distanz der Marker beträgt 0.11 Mb. Die kleinste Distanz liegt bei 66 bp und die maximale Distanz liegt bei 1.81 Mb.
Die Abhängigkeitsstruktur der SNPs kann über das Kopplungsungleichgewicht untersucht werden. Ein Kopplungsungleichgewicht liegt vor, falls die Allele auf verschiedenen Loci voneinander abhängig sind. Bei der Betrachtung von zwei Allelen und zwei Loci seien die relativen Häufigkeiten der Allele p_A, p_a, p_B und p_b. Die entsprechenden möglichen Haplotypen sind folglich AB, Ab, aB und ab. Diese treten mit den relativen Häufigkeiten p_{AB}, p_{Ab}, p_{aB} und p_{ab} auf. Bei einem Kopplungsgleichgewicht und den relativen Häufigkeiten $p_A = p_a = p_B = p_b = 0.5$ würden die Haplotypen je mit einer Wahrscheinlichkeit von 0.25 auftreten. Das Kopplungsungleichgewicht wird gemessen über die Differenz der tatsächlichen Häufigkeit und der erwarteten Häufigkeit unter der Unabhängigkeitshypothese (Lewontin und Kojima,

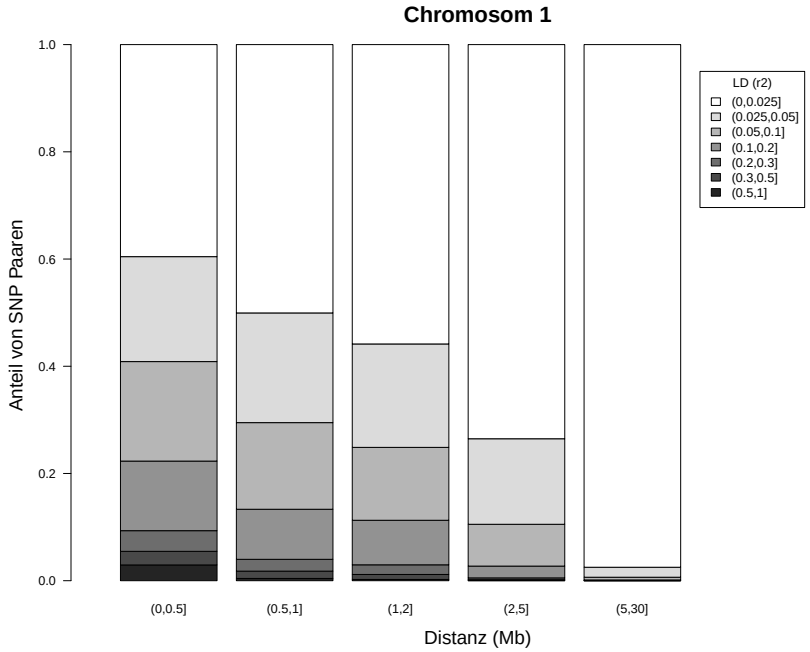

Abbildung 8: Kopplungsungleichgewicht

1960): $D = p_{AB} - p_A p_B$. Eine Skalierung des Kopplungsungleichgewicht wird von Hill und Robertson (1968) vorgeschlagen:

$$r^2 = \frac{D^2}{p_A p_B p_a p_b} \in [0, 1].$$

Bei $r^2 = 0$ liegt ein Kopplungsgleichgewicht und bei $r^2 = 1$ eine vollständige Kopplung vor.
Das mittlere Kopplungsungleichgewicht liegt bei allen paarweisen Markervergleichen bei 0.017. Die Standardabweichung des Kopplungsungleichgewicht beträgt 0.049. Der Abfall des Kopplungsungleichgewicht der SNPs mit steigender Distanz der SNPs ist in Abbildung 8 für Chromosom Eins dargestellt. Das Auftreten des Kopplungsungleichgewicht in Abhängigkeit der Distanz auf den Chromosomen Zwei bis Fünf gleicht dem Kopplungsungleichgewicht auf Chromosom Eins stark. Bei einer Distanz kleiner als 0.5 Mb tritt bei circa 10% der SNP Paare ein skaliertes Kopplungsungleichgewicht größer als 0.2 auf. Bei einer Distanz größer als 5 Mb ist das skalierte Kopplungsungleichgewicht fast immer kleiner als 0.2.
Um einen ersten Anhaltspunkt zu erhalten wie stark der Einfluss der SNPs auf die phänotypischen Merkmale ist, werden jeweils einfache lineare Regressionen eines SNPs auf den Phänotyp berechnet und der negative logarith-

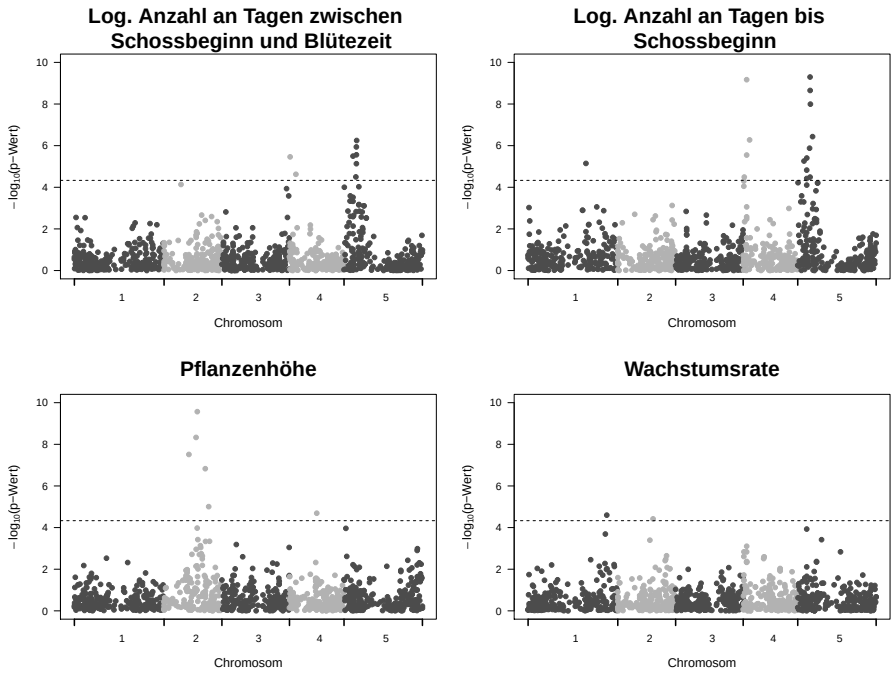

Abbildung 9: Manhattan-Plot der negativen logarithmierten p-Werte und die Bonferroni-Schranke: $-log_{10}(0.05/1073) = 4.33$

mierte p-Wert zur Basis Zehn des Steigungsparameters grafisch gegen die SNPs abgetragen. Die sogenannten Manhattan-Plots sind in Abbildung 9 zu sehen. Die Struktur der Manhattan-Plots unterscheidet sich zwischen den Merkmalen. Daher wird vermutet, dass ihnen eine unterschiedliche genetische Architektur unterliegt. Bei der Wachstumsrate ist der negative logarithmierte p-Wert über alle SNPs eher gleichmäßig. Da es sich hier um ein multiples Testproblem, handelt muss für die Fehlerwahrscheinlichkeit eine Bonferroni-Korrektur (Fahrmeir *et al.*, 2003) durchgeführt werden. Zum Niveau 0.0046% sind zwei Effekte signifikant. Bei der Pflanzenhöhe sind einige signifikante lineare Effekte auf Chromosom zwei und bei der Zeit zwischen Schossbeginn und Blütezeit und der Zeit bis zum Schossbeginn sind einige signifikante lineare Effekte auf dem vierten und fünften Chromosom zu erkennen.
Bei der Inferenz werden zur Wahl der optimalen Parameter Schätzungen für die Heritabilität benötigt. Die Schätzwerte für die Heritabilität werden von Kover *et al.* (2009) übernommen. Die geschätzte Heritabilität für die Wachstumsrate beträgt 0.22. Für die Anzahl der Tage bis zum Schossbeginn wird die Heritabilität auf 0.72 und für die Anzahl der Tage zwischen Schossbeginn und Blütezeit auf 0.40 geschätzt. Von Kover *et al.* (2009) wird keine Heritabilitätsschätzung der Pflanzenhöhe angegeben. Dieser Schätzer wird über

ein Gemischtes Modell berechnet (Kover *et al.*, 2009) und beträgt 0.54.
Für die Beschreibung der Datengrundlage wurde die statistische Software R (R Development Core Team, 2012) und insbesondere das Paket synbreed (Wimmer *et al.*, 2012) verwendet.

4 Ergebnisse der Arabidopsis thaliana Inferenz

In diesem Kapitel werden die Regressionen der genetischen Marker auf die Phänotypen mit Hilfe der vorgestellten Methoden durchgeführt. Für alle Regressionen werden die Zielvariablen standardisiert.
Zunächst wird in Abschnitt 4.1 die Sensitivität des Bayesianischen Elastic Net mit der Parametrisierung nach Hofmarcher *et al.* (2011) bei verschiedenen Priori Annahmen überprüft. Die dafür verwendeten Maße sind die Korrelation zwischen den realen Werten der Zielvariable $\boldsymbol{y}$ und den Werten der Modellanpassung $\hat{\boldsymbol{y}}$, die Anzahl der effektiven Parameter und das Devianz Informationskriterium. Ein direkter Vergleich der Modellanpassung bei verschiedenen Priori Annahmen erfolgt über die Korrelation der angepassten Werte. Die Konvergenz der Schätzer wird über die Konvergenzpfade von λ_1, λ_2, α, σ, μ, $\boldsymbol{\beta}$ und $\boldsymbol{\tau}$, welche jeweils im Anhang aufgeführt sind, untersucht. Da 1073 $\boldsymbol{\beta}$-Koeffizienten und 1073 $\boldsymbol{\tau}$-Koeffizienten geschätzt werden, sind für diese Koeffizienten exemplarisch jeweils nur drei Konvergenzpfade abgebildet.
Für alle Bayesianischen Modelle werden 15000 Iterationen durchgeführt und eine Konvergenzphase von 7500 Iterationen gewählt. Desweiteren wird nur jede zehnte Beobachtung für die Auswertung berücksichtigt.
Verglichen wird das Bayesianische Elastic Net in dem Abschnitt 4.2 mit dem Bayesianischen Lasso, dem Bayesianischen Ridge, den entsprechenden frequentistischen Methoden und dem Generalisierten Elastic Net. Die Modellgüte des Bayesianischen Lasso, des Bayesianischen Ridge und des Bayesianischen Elastic Net wird über das Devianz Informationskriterium, die Anzahl der effektiven Parameter und die Korrelation zwischen realen und geschätzten phänotypischen Ausprägungen beurteilt. Der Vergleich der Bayesianischen Methoden mit dem Ridge, Lasso, Elastic Net und Generalisierten Elastic Net erfolgt über die Anzahl der effektiven Parameter und über die Korrelation zwischen realen und angepassten Werten. Die Vorhersagegüte dieser sieben Regressionsverfahren wird anhand einer fünffachen Kreuzvalidierung mit drei Wiederholungen über den Mittleren Quadratischen Fehler und die Korrelation zwischen wahren und prognostizierten Werten bestimmt.
Alle Methoden sind in der statistische Software R implementiert. Für das Ridge, Lasso und Elastic Net wird das Paket glmnet (Friedman *et al.*, 2010) und für das Bayesianischen Lasso beziehungsweise das Bayesianischen Ridge das Paket BLR (de los Campos und Rodriguez, 2012) verwendet. Die Berechnung des Bayesianischen Elastic Net erfolgt in Kombination der Software R und der Software JAGS (Version 3.2.0) unter Verwendung des Pakets R2jags

(Su und Yajima, 2012). In dem Paket spikeslab (Ishwaran *et al.*, 2010a,b) ist das Generalisierten Elastic Net implementiert.

4.1 Robustheit des Bayesianischen Elastic Net

Im Folgenden wird für das Bayesianische Elastic Net die hierarchische Formulierung von Hofmarcher *et al.* (2011) verwendet. Dabei ist $\lambda_1 = \alpha\lambda$ und $\lambda_2 = (1-\alpha)\lambda$. Für die Verteilung von α wird a priori eine auf $(0, 1)$-trunkierte Normalverteilung angenommen. Um die Stabilität der Schätzung zu untersuchen, werden je drei verschiedene Erwartungswerte und Varianzen für die trunkierte Normalverteilung verwendet. Für den Erwartungswert werden die Priori Werte 0.1, 0.5, 0.9 und für die Varianz die Priori Werte 0.000001, 0.0001, 0.01 spezifiziert.
Von Hofmarcher *et al.* (2011) wird für λ^2 a priori die Gamma Priori-Verteilung $\text{Ga}(0.1, 0.1)$ gewählt. Diese Wahl der Hyperparameter für die Gammaverteilung führt dazu, dass bei der Regression auf die Pflanzenhöhe und die Wachstumsrate kein Effekt der Schrumpfung vorliegt. Dies spiegelt sich unter anderem in unrealistisch hohen Schätzwerten für die Anzahl der effektiven Parameter wider. Als Alternative für die von Hofmarcher *et al.* (2011) vorgeschlagenen Hyperparameter werden in dieser Arbeit für das Bayesianische Elastic Net dieselben Gestalt- und Maßparameter (a_{shape}, a_{rate}) wie für die λ^2 Priori-Verteilung des Bayesianischen Lasso gewählt. Es wird der optimale λ Parameter, entsprechend der Gleichung (2.5), berechnet und anschließend die Gestalt- und Maßparameter so gewählt, dass die Dichte für λ ihr Maximum im optimalen λ hat. In der Abbildung 10 sind die Priori Dichten für λ aufgeführt.
Vorweg ist anzumerken, dass vier der 36 betrachteten Regressionsmodelle auf Grund numerischer Probleme nicht berechnet werden können. Mehrere Priori-Verteilungen des Bayesianischen Modells sind trunkiert und die Priori-Verteilung von σ^2 konvergiert von rechts bei Null gegen unendlich. Basierend auf numerischen Ungenauigkeiten können an den Rändern der Verteilungen nicht zulässige Werte entstehen.
In Tabelle 2 sind die Korrelationen zwischen realen und angepassten Werten $cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$, die Anzahl der effektiven Parameter p_D und das Devianz Informationskriterium DIC für alle Zielvariablen und für verschiedene Priori Annahmen aufgeführt. Die Korrelation ist durchwegs für alle Zielgrößen und Priori Annahmen größer als 0.8. Dies spricht für eine gute Anpassung des Modells an die Daten. Bei drei der vier Zielgrößen liegt das kleinste DIC und bei allen Zielgrößen das kleinste p_D vor, falls $\text{Var}(\alpha) = 0.01$ gewählt wird, jedoch immer für unterschiedliche Wahlen von $\text{E}(\alpha)$. Eine größere Varianz könnte mehr Flexibiliät der Schätzung erlauben. Das p_D ist ein Schätzwert,

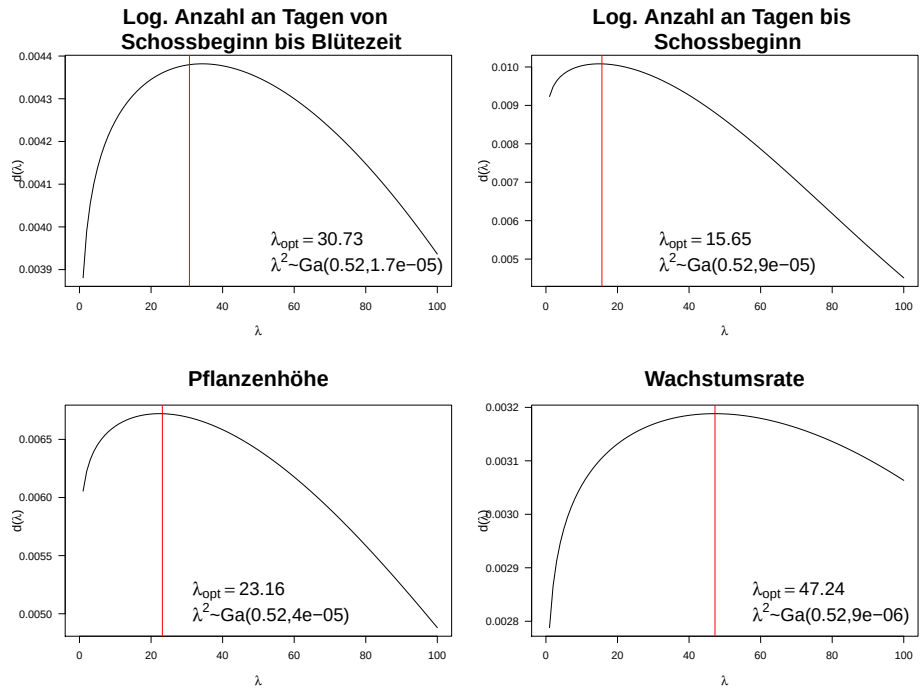

Abbildung 10: Priori Dichte für λ, wobei für die Priori-Verteilung von λ^2 die optimalen Paramter gewählt wurden; $p(\lambda|a_{shape}, a_{rate}) = \text{Ga}(\lambda^2|a_{shape}, a_{rate}) \cdot 2\lambda$

Tabelle 2: Untersuchung der Sensitivität des Bayesianischen Elastic Net bei verschiedenen Priori-Annahmen für α gemessen an $cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$, p_D, DIC (- =nicht berechenbare Modelle)

Log. Anzahl an Tagen zwischen Schossbeginn und Blütezeit									
	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$			p_D			DIC		
E(α) / Var(α)	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
0.01	0.8235	0.8164	0.8198	786	594	537	1807	1625	1564
0.0001	0.8116	0.8190	-	564	651	-	1603	1678	-
0.000001	-	0.8164	0.8198	-	594	706	-	1625	1734
Log. Anzahl an Tagen bis Schossbeginn									
	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$			p_D			DIC		
E(α) / Var(α)	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
0.01	0.8884	0.8804	0.8821	768	791	887	1647	1688	1781
0.0001	0.8889	0.8809	-	822	763	-	1697	1659	-
0.000001	0.8834	0.8804	-	779	791	-	1667	1688	-
Pflanzenhöhe									
	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$			p_D			DIC		
E(α) / Var(α)	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
0.01	0.8590	0.8650	0.8662	631	560	670	1571	1488	1592
0.0001	0.8676	0.8666	0.8647	566	752	617	1485	1673	1544
0.000001	0.8644	0.8666	0.8677	845	752	665	1776	1673	1586
Wachstumsrate									
	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$			p_D			DIC		
E(α) / Var(α)	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
0.01	0.8338	0.8379	0.8383	623	708	581	1471	1552	1421
0.0001	0.8440	0.8420	0.8335	595	623	1001	1424	1456	1856
0.000001	0.8471	0.8379	0.8312	697	708	612	1529	1552	1463

der als die Anzahl der effektiven Parameter interpretiert wird und nur als Richtwert betrachtet werden sollte. Die Unterschiede der betrachteten Maße sind innerhalb der Zielgrößen bei verschiedenen Priori Wahlen nicht auffällig groß.
In Tabelle 3 sind die Korrelationen der angepassten Werte $\hat{\boldsymbol{y}}$ der Bayesianischen Elastic Net Modelle mit verschiedenen Priori Annahmen aufgeführt. Alle berechneten Korrelationen sind signifikant positiv und größer als 0.99. Folglich sind die prognostizierten Werte bei verschiedenen Priori Annahmen sehr ähnlich und somit robust. Für weitere Analysen, wie zum Beispiel die Betrachtung der Vorhersagegenauigkeit, ist es ausreichend nur mit einer Wahl der Hyperparameter für die Priori-Verteilung von α fortzufahren.
Eine explizite Betrachtung des Posteriori-Mittelwerts, des 2.5%- und 97.5%-Quantils für α erfolgt anhand Tabelle 4. Je höher der a priori Erwartungswert für α gewählt wird, desto höher ist auch der Posteriori Schätzer für α. Die Posteriori Schätzer für α bei den Priori Annahmen $E(\alpha) = 0.1$ und $E(\alpha) = 0.5$ sind zumeist deutlich größer als die gewählten Hyperparameter. Es ist sowohl der Einfluss der Daten als auch der Einfluss der Priori-Verteilungen ersichtlich. Die a priori Wahl der Varianz von α beeinflusst sowohl den Posteriori-Mittelwert als auch das 95%-Intervall des Schätzers. Ein Systematik des Einflusses ist nicht zu erkennen.
Basierend auf Abbildung 9 würde man, außer für die Wachstumsrate, eine Dominanz des Lasso Parameters erwarten. Für die Merkmale Zeit zwischen Schossbeginn und Blütezeit, Zeit bis zum Schossbeginn und Pflanzenhöhe weisen einige SNPs einen signifikanten Effekt auf. Diese sollten durch eine Variablenselektion in dem Modell verbleiben. Falls fast alle SNPs, wie bei der Wachstumsrate, nicht signifikant sind scheint eine gleichmäßige Schrumpfung der Koeffizienten angemessen. Da $\alpha = 1$ dem Lasso Modell entspricht wird für die Zeit zwischen Schossbeginn und Blütezeit, Zeit bis zum Schossbeginn und der Pflanzenhöhe ein hoher Posteriori Schätzer für α erwartet. Diese Erwartungshaltung wird durch die Posteriori Schätzer bestätigt. Für die Wachstumsrate wäre eher ein Gleichgewicht der Parameter oder eine Dominanz des Ridge Parameters zu erwarten. Bei den Regressionen auf die Wachstumsrate dominiert auch der Lasso Parameter. Dies könnte darauf schließen lassen, dass bereits sehr wenige signifikante Effekte in der Einzelmarkerregression, zu einer Dominanz des Lasso Parameters führen.
Im Anhang dieser Arbeit sind die Konvergenzpfade des Bayesianischen Elastic Net bei verschiedenen Priori Annahmen dargestellt. In den Abbildungen 11 bis 19 sind exemplarisch die Pfade der Zielvariable Pflanzenhöhe bei 15000 Iterationen und Beachtung von nur jedem zehnten Kettenelement dargestellt. In den Abbildungen 20 bis 23 ist exemplarisch für jede Zielgröße und der Priori-Verteilung $\alpha \sim N(0.5, 0.0001)$ ein Konvergenzpfad mit 50000

Tabelle 3: Untersuchung der Sensitivität des Bayesianischen Elastic Net bei verschiedenen Priori-Annahmen für α gemessen über die Korrelation zwischen den angepassten Werten $\hat{\boldsymbol{y}}$ (-=nicht berechenbare Modelle)

Log. Anzahl an Tagen zwischen Schossbeginn und Blütezeit									
	E(α)	0.1		0.5			0.9		
E(α)	Var(α)	0.0001	0.01	0.000001	0.0001	0.01	0.000001	0.0001	0.01
0.1	0.000001	-	-	-	-	-	-	-	-
	0.0001		0.9984	0.9987	0.9987	0.9987	0.9986	-	0.9986
	0.01			0.9986	0.9988	0.9986	0.9989	-	0.9988
0.5	0.000001				0.9988	1.0000	0.9987	-	0.9988
	0.0001					0.9988	0.9987	-	0.9988
	0.01						0.9987	-	0.9988
0.9	0.000001							-	0.9986
	0.0001							-	-
Log. Anzahl an Tagen bis zum Schossbeginn									
	E(α)	0.1		0.5			0.9		
E(α)	Var(α)	0.0001	0.01	0.000001	0.0001	0.01	0.000001	0.0001	0.01
0.1	0.000001	0.9993	0.9992	0.9994	0.9994	0.9994	-	-	0.9993
	0.0001		0.9993	0.9991	0.9992	0.9991	-	-	0.9992
	0.01			0.9991	0.9992	0.9991	-	-	0.9990
0.5	0.000001				0.9995	1.0000	-	-	0.9993
	0.0001					0.9995	-	-	0.9994
	0.01						-	-	0.9993
0.9	0.000001							-	-
	0.0001								-
Pflanzenhöhe									
	E(α)	0.1		0.5			0.9		
E(α)	Var(α)	0.0001	0.01	0.000001	0.0001	0.01	0.000001	0.0001	0.01
0.1	0.000001	0.9991	0.9991	0.9991	0.9991	0.9992	0.9991	0.9991	0.9992
	0.0001		0.9987	0.9992	0.9992	0.9990	0.9993	0.9991	0.9992
	0.01			0.9988	0.9988	0.9989	0.9986	0.9988	0.9987
0.5	0.000001				1.0000	0.9990	0.9993	0.9993	0.9993
	0.0001					0.9990	0.9993	0.9993	0.9993
	0.01						0.9989	0.9989	0.9991
0.9	0.000001							0.9993	0.9993
	0.001								0.9993
Wachstumsrate									
	E(α)	0.1		0.5			0.9		
E(α)	Var(α)	0.0001	0.01	0.000001	0.0001	0.01	0.000001	0.0001	0.01
0.1	0.000001	0.9985	0.9979	0.9982	0.9982	0.9982	0.9974	0.9977	0.9982
	0.0001		0.9980	0.9982	0.9979	0.9982	0.9977	0.9980	0.9983
	0.01			0.9984	0.9979	0.9984	0.9984	0.9982	0.9986
0.5	0.000001				0.9982	1.0000	0.9981	0.9982	0.9982
	0.0001					0.9982	0.9976	0.9977	0.9981
	0.01						0.9981	0.9982	0.9982
0.9	0.000001							0.9983	0.9980
	0.001								0.9981

Tabelle 4: Posteriori-Mittelwert und 2.5%-, 97.5%-Quantile für den α-Parameter des Bayesianischen Elastic Net (-=nicht berechenbare Modelle)

Log. Anzahl an Tagen zwischen Schossbeginn und Blütezeit			
$E(\alpha)$ / $Var(\alpha)$	0.1	0.5	0.9
0.01	0.649(0.452,0.812)	0.846(0.789,0.895)	0.995(0.982,1.000)
0.0001	0.733(0.420,0.913)	0.781(0.622,0.951)	-
0.000001	-	0.846(0.789,0.895)	0.970(0.913,0.995)
Log. Anzahl an Tagen bis Schossbeginn			
$E(\alpha)$ / $Var(\alpha)$	0.1	0.5	0.9
0.01	0.594(0.400,0.789)	0.892(0.739,0.966)	0.972(0.943,0.989)
0.0001	0.836(0.677,0.917)	0.797(0.615,0.970)	-
0.000001	0.821(0.676,0.928)	0.892(0.739,0.966)	-
Pflanzenhöhe			
$E(\alpha)$ / $Var(\alpha)$	0.1	0.5	0.9
0.01	0.515(0.322,0.658)	0.661(0.471,0.867)	0.907(0.857,0.938)
0.0001	0.942(0.827,0.990)	0.947(0.881,0.988)	0.963(0.930,0.990)
0.000001	0.700(0.440,0.888)	0.947(0.881,0.988)	0.980(0.962,0.994)
Wachstumsrate			
$E(\alpha)$ / $Var(\alpha)$	0.1	0.5	0.9
0.01	0.518(0.297,0.793)	0.955(0.918,0.988)	0.996(0.991,0.999)
0.0001	0.775(0.564,0.896)	0.723(0.478,0.967)	0.918(0.839,0.974)
0.000001	0.751(0.542,0.920)	0.955(0.918,0.988)	0.987(0.975,0.996)

Iterationen und Verwendung jedes zehnten Kettenelements dargestellt.
Exemplarisch wurden jeweils die ersten drei $\boldsymbol{\beta}$-Koeffizienten und die ersten drei $\boldsymbol{\tau}$-Koeffizienten ausgewählt. Die nicht aufgeführten Konvergenzpfade sind von derselben Struktur wie die aufgeführten Pfade. Die Konvergenz ist augenscheinlich für den Intercept-Parameter μ und die Steigungsparameter $\boldsymbol{\beta}$ bei allen Zielgrößen und für alle Erwartungswert- und Varianzannahmen gegeben. Dies stellt eine gute Grundlage zur Vorhersage der phänotypischen Werte dar. Der λ_1- und λ_2-Parameter konvergieren nicht. Der Konvergenzpfad bewegt sich unstrukturiert und die Parameter sind nicht identifizierbar. Das Konvergenzverhalten des α-Parameters ist stark von den Priori Annahmen abhängig.

4.2 Methodenvergleich

Da sich im vorherigen Abschnitt das Bayesianische Elastic Net als robust bezüglich der Modellanpassung, gegenüber verschiedenen Priori Annahmen für α, erwiesen hat, wird im Weiteren dieses Abschnitts nur das Bayesianische Elastic Net mit der Priori-Verteilung $\alpha \sim \mathrm{N}(0.5, 0.001)$ betrachtet. Für die Hyperparameter des Bayesianischen Ridge und des Bayesianischen Lasso werden die optimalen Parameter (Abschnitt 2.3.3) gewählt.
Der Methodenvergleich erfolgt über die Korrelation zwischen realen und angepassten Werten $cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$, die Anzahl effektiver Parameter p_{eff}, das Devianz Informationskriterium und über eine Kreuzvalidierung mit den Kriterien Mittlerer Quadratischer Fehler $\mathrm{MSE}_{\mathrm{CV}}$ und Korrelation zwischen wahren und prognostizierten Werten $cor(\boldsymbol{y}, \hat{\boldsymbol{y}})_{\mathrm{CV}}$. Die Anzahl der effektiven Parameter wird in den Bayesianischen Modellen über das p_D geschätzt. Bei dem Lasso, Elastic Net und Generalisierten Elastic Net wird die Anzahl der nicht-Nullkoeffizienten des geschätzten Modells angegeben. Für das Ridge erfolgt die Schätzung der effektiven Parameter über die Freiheitsgrade $p_{eff} = df_{Ridge} = \mathrm{spur}(\mathbf{X}(\mathbf{X}^\top \mathrm{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^\top)$ (Hastie *et al.*, 2009). In Tabelle 5 sind die Resultate für die Gütekriterien der Modellschätzungen aufgeführt. In den Abbildungen 28 und 29 sind die Ergebnisse der Kreuzvalidierungen grafisch über Boxplots dargestellt.
Unter den Bayesianschen Methoden resultiert für das Elastic Net bei den Zielvariablen Pflanzenhöhe und Wachstumsrate die größte mittlere Korrelation und der kleinste Mittlere Quadratische Fehler. Für die Zielgröße Zeit zwischen Schossbeginn und Blütezeit liefert das Bayesianische Ridge und für die Zielgröße Zeit bis zum Schossbeginn das Bayesianische Lasso das beste mittlere Ergebnis unter den Bayesianischen Modellen. Allerdings liefern die frequentistischen Varianten häufig geringfügig bessere Ergebnisse.

Tabelle 5: Vergleich des Bayesianischen Elastic Net (BEN), Bayesianischen Lasso (BL), Bayesianischen Ridge (BR), Generalisierten Elastic Net (GEN), Elastic Net (EN), Lasso (L), Ridge (R); Standardabweichung bei CV in Klammern

Log. Anzahl an Tagen zwischen Schossbeginn und Blütezeit					
	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$	p_{eff}	DIC	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})_{\text{CV}}$	MSE_{CV}
BEN	0.8190	651	1678	0.325(0.092)	0.907(0.146)
BL	0.8278	115	1131	0.327(0.094)	0.907(0.15)
BR	0.8109	102	1139	**0.328(0.096)**	**0.905(0.146)**
GEN	0.6469	47		0.313(0.103)	0.912(0.156)
EN	0.6892	82		0.24 (0.077)	0.950(0.167)
L	0.6776	75		0.219(0.087)	0.957(0.17)
R	0.7935	412		0.322(0.092)	0.906(0.147)
Log. Anzahl an Tagen bis zum Schossbeginn					
	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$	p_{eff}	DIC	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})_{\text{CV}}$	MSE_{CV}
BEN	0.8809	763	1659	0.482(0.077)	0.776(0.108)
BL	0.8928	163	1029	0.486(0.076)	0.774(0.109)
BR	0.8774	148	1055	0.464(0.77)	0.794(0.108)
GEN	0.7917	84		0.510(0.066)	0.746(0.108)
EN	0.8226	121		**0.513(0.061)**	**0.744(0.093)**
L	0.8182	104		0.51 (0.058)	0.748(0.094)
R	0.8524	418		0.464(0.076)	0.789(0.102)
Pflanzenhöhe					
	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$	p_{eff}	DIC	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})_{\text{CV}}$	MSE_{CV}
BEN	0.8666	752	1673	0.443(0.099)	0.813(0.127)
BL	0.7499	27	1228	0.401(0.107)	0.951(0.157)
BR	0.9125	210	1439	0.399(0.091)	0.877(0.129)
GEN	0.6484	33		**0.458(0.078)**	**0.785(0.106)**
EN	0.7863	110		0.446(0.097)	0.81 (0.132)
L	0.7701	84		0.46 (0.093)	0.796(0.13)
R	0.8487	417		0.416(0.102)	0.837(0.126)
Wachstumsrate					
	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$	p_{eff}	DIC	$cor(\boldsymbol{y}, \hat{\boldsymbol{y}})_{\text{CV}}$	MSE_{CV}
BEN	0.8420	623	1456	**0.338(0.119)**	**0.912(0.158)**
BL	0.8238	76	942	0.321(0.118)	0.951(0.179)
BR	0.7949	49	955	0.322(0.122)	0.954(0.164)
GEN	0.6328	37		0.249(0.094)	0.951(0.153)
EN	0.5915	34		0.238(0.076)	0.962(0.149)
L	0.5728	28		0.237(0.075)	0.96 (0.147)
R	0.8164	334		0.334(0.122)	0.915(0.157)

Bei Betrachtung aller Methoden liefert für jede Zielgröße eine andere Methode das beste mittlere Ergebnis bezüglich der Vorhersagequalität. Wird neben dem arithmetischen Mittel der Korrelation und des Mittleren Quadratischen Fehlers auch die Streuung jener Größen betrachtet, so ist ersichtlich, dass die Methoden sich zumeist nicht relevant unterscheiden. Die Darstellung über die Boxplots und der Kolmogorov-Smirnov-Test Test zeigen, dass die Kreuzvalidierungsergebnisse keiner Normalverteilung folgen. Für den Test auf Unterschiede wird deshalb der nichtparametrische Kruskal-Wallis Test gewählt. Die Vorhersagegüte gemessen über die Korrelation unterscheidet sich für die penalisierten Modelle mit den Zielgrößen Planzenhöhe und Zeit bis zum Schossbeginn nicht signifikant. Für die Anzahl an Tagen zwischen Schossbeginn und Blütezeit unterscheiden sich die Methoden Bayesianisches Elastic Net, Bayesianisches Lasso, Bayesianisches Ridge, Generalisiertes Elastic Net und Ridge bezüglich der Korrelation nicht. Das Elastic Net und Lasso liefern kleinere Korrelationen und unterscheiden sich signifikant von den anderen Methoden. Für die Regressionen auf die Wachstumsrate unterscheidet sich die Vorhersagegüte gemessen über die Korrelation signifikant. Gleich gute Modelle zur Vorhersage liefern das Bayesiansiche Elastic Net, das Bayesianische Lasso, das Bayesianische Ridge und das Ridge. Das Generalisierte Elastic Net, Elastic Net und Lasso liefern kleinere Korrelationen und unterscheiden sich signifikant von den anderen Methoden. Wird die Vorhersagegüte nicht über die Korrelation sondern über den Mittleren Quadratischen Fehler gemessen, so unterscheiden sich die penalisierten Modelle bei den Zielgrößen Zeit zwischen Schossbeginn und Blütezeit, Zeit bis zum Schossbeginn und Wachstumsrate nicht signifikant. Für die Pflanzenhöhe unterscheiden sich die Methoden, mit Ausnahme des Bayesianischen Lasso nicht signifikant. Das Bayesianische Lasso weist einen höheren Mittleren Quadratischen Fehler auf. Zusammenfassend wird festgestellt, dass für jede Zielgröße die Annahme der Gleichheit der Methoden entweder durch das Kriterium der Korrelation oder das Kriterium MSE beibehalten wird. Für keine der Zielgrößen wird die Gleichheit durch beide Kriterien abgelehnt.
Die Korrelation zwischen wahren und angepassten Werten $cor(\boldsymbol{y}, \hat{\boldsymbol{y}})$ ist zumeist bei den Bayesianischen Methoden größer als bei den frequentistischen Methoden. Übergreifend betrachtet liefert das Bayesianische Elastic Net eine genauso gute Modellanpassung wie das Bayesianische Ridge oder das Bayesianische Lasso. Eine sehr gute Modellanpassung birgt auch immer die Gefahr einer Überanpassung des Modells und somit einer schlechter Vorhersagegenauigkeit. Die Vorhersagegenauigkeit wurde über eine Kreuzvalidierung überprüft und kann als gut eingestuft werden. Für die untersuchten Modelle liegt keine Überanpassung vor.
Die Anzahl der nicht-Nullkoeffizienten des Lasso, Elastic Net und Generali-

sierten Elastic Net unterscheiden sich untereinander und über die verschiedenen Zielgrößen stark. Das Generalisierte Elastic Net weist bei drei der Zielvariablen die kleinste Anzahl an nicht-Nullkoeffizienten auf und liefert das sparsamste Modell. Die Anzahl der effektiven Parameter der Ridge Schätzung liegt höher als die Anzahl der nicht-Nullkoeffizienten der anderen frequentistischen Methoden. Das Devianz Informationskriterium und die Anzahl der effektiven Parameter sind für das Bayesianische Elastic Net höher als für das Bayesianische Lasso und Bayesianische Ridge. Da das p_D nur Schätzwerte sind, sollten sie für die Beurteilung der Modellgüte nicht überbewertet werden. Die Ergebnisse der Kreuzvalidierung hingegen haben sehr hohes Gewicht um die Modelle bezüglich der praktischen Anwendung in Züchtungsprogrammen zu beurteilen.
Der Schrumpfungseffekt der penalisierten Modelle ist am besten grafisch zu erkennen und zu beurteilen. In den Abbildung 24 bis 27 im Anhang sind in der ersten Spalte Manhattan-Plots mit den SNP Effekten dargestellt. Bei dem Bayesianischen Ridge werden die Effekte in der Regel gleichmäßiger geschrumpft als bei dem Bayesianischen Lasso oder Bayesianischen Elastic Net. Die Effektgrößen des Bayesianischen Elastic Net sind bei drei der vier Variablen denen des Bayesianischen Lasso ausgesprochen ähnlich. Die Spannweite der Effekte der Bayesianischen Modelle ist kleiner als die des Lasso, Elastic Net und Generalisierten Elastic Net. Die Struktur der Manhattanplots bezüglich der einflussreichen und nicht einflussreichen Effekte ist jeweils für die quantitativen Merkmale bei allen Methoden ähnlich.
In der zweiten Spalten der Abbildungen 24- 27 sind die Streudiagramme zwischen den SNP Effekten aus einer nichtsimultanen Schätzung durch ein lineares Modell und den SNP Effekten der penalisierten Modelle abgebildet. Der allgemeine Effekt der Schrumpfung der penalisierten Verfahren ist deutlich zu erkennen. Die Steigung der Regressionsgerade der SNP Effekte der Einzelmarkerregression auf die SNP Effekte der penalisierten Schätzung kann als Maß für die Stärke der Schrumpfung interpretiert werden. Die Schrumpfung kann durchaus als stark bezeichnet werden. Über alle Pflanzenmerkmale hinweg kann kein Aussage darüber getroffen werden welches Verfahren im Allgemeinen zur stärksten Schrumpfung führt.

4.3 Kritik am Bayesianischen Elastic Net

Bei der Berechnung des Bayesianischen Elastic Net können auf Grund numerischer Probleme nicht alle Modelle angepasst werden. Die Konvergenz des μ-Parameters und der Steigungsparameter ist immer zufriedenstellend. Die Identifizierbarkeit der Schrumpfungsparameter ist hingegen, sowohl bei 15000 als auch bei 50000 Iterationen, nicht immer gegeben. Eine der Kernide-

en des Bayesianischen Elastic Net mit der α Parametrisierung ist es flexibel zu gestalten und zu erkennen, ob der Lasso und der Ridge Parameter dominiert. Bei der verwendeten Anzahl an Iterationen und der Konvergenzphase ist sowohl eine Abhängigkeit des Posteriori Schätzers für α von der Wahl der Priori-Verteilung als auch der deutliche Einfluss der Daten zu erkennen. Das ursprüngliche Ziel der flexiblen Modellierung hingehend zum Lasso oder zum Ridge kann nicht erreicht werden. Bei Verwendung der Kreuzvalidierung zur Beurteilung der Vorhersagegüte kann keine relevante Verbesserung des Bayesianischen Elastic Net gegenüber den anderen Methoden festgestellt werden. Desweiteren sind die computationalen Berechnungszeiten für das Bayesianische Elastic sehr hoch. Es gibt keinen relevanten Zusatznutzen des Bayesianischen Elastic Net gegenüber den etablierten Methoden in der genetischen Vorhersage.

5 Diskussion

In dieser Arbeit wurden Bayesianische und frequentistische Regressionsmodelle auf ihre Eignung zur Vorhersage in Züchtungsprogrammen untersucht. Die Modelle, die dabei betrachtet wurden, sind das Bayesianische Ridge, Lasso, Elastic Net, die frequentistischen Analoga und das Generalisierte Elastic Net. Diese Modelle gehören der Klasse der penalisierten linearen Modelle an und erlauben die Inferenz auch in $p \gg n$-Situationen.
Die Anwendung der Modelle erfolgte auf die genotypischen und vier phänotypische Merkmale der Pflanze Arabidopsis. Die betrachteten phänotypischen Merkmale waren die Wachstumsrate, die Pflanzenhöhe, die Zeit bis zum Schossbeginn und die Zeit zwischen Schossbeginn und Blütezeit. Die Anzahl der SNPs ($p = 1260$) war größer als die Anzahl der untersuchten Individuen ($n = 426$).
Über eine Einzelmarkerregression wurde ein erster Eindruck über die Stärke der SNP Effekte gegeben. Für alle Pflanzenmerkmale lagen signifikante Effekte vor. Bei der Variable Wachstumsrate waren die Effektstärken zueinander ähnlicher als bei den anderen Variablen. Für die Wachstumsrate würde man eine gleichmäßige Schrumpfung der Effekte erwarten. Dies spricht für die Anwendung der Ridge Regression. Bei den anderen drei phänotypischen Merkmalen waren die Unterschiede der Effektstärken groß. Angemessen scheint auf Grund dessen eine Variablenselektion und somit die Anwendung des Lasso Verfahrens.
Das Bayesianische Elastic Net kombiniert das Bayesianische Lasso und Bayesianische Ridge. Über die spezielle Parametrisierung des Bayesianischen Elastic Net nach Hofmarcher *et al.* (2011) soll ersichtlich sein, ob das Lasso oder das Ridge dominiert. Für alle Variablen dominierte der Lasso Parameter. Bei drei der vier Zielgrößen entspricht dies den Erwartungshaltungen aus den Einzelmarkerregressionen. Der Grund könnte sein, dass bereits wenige signifikante Effekte der Einzelmarkerregression zu einer Dominanz des Lasso führen.
Desweiteren stellt sich die Frage, ob die SNPs mit signifikanten Effekten in der Einzelmarkerregression auch bei der simultanen Schätzung die größten Effekte aufweisen. Bei der Untersuchung des Einflusses der genotypischen Merkmale auf das phänotypische Merkmal Anteil der schwarzen Fellfarbe bei Rindern über die Methode BayesA (Meuwissen *et al.*, 2001) zeigt sich in der Studie von Hayes *et al.* (2010) ein Zusammenhang zwischen der Einzelmarkerregression und den penalisierten Effekten. Die SNPs mit signifikanten Effekten in der Einzelmarkerregression weisen auch große Effekte in der Regression mit BayesA auf. Dieser Zusammenhang zeigte sich ebenfalls bei den in dieser Arbeit vorgestellten Methoden und deren Anwendung auf die Daten

der Arabidopsis. Allgemein wird bei großen beziehungsweise kleinen Effekten in der Einzelregression davon ausgegangen, dass diese Effekte auch bei einer simultanen Schätzung groß bezeihungsweise klein sind. Auf Grund von Kollinearität ist dies jedoch nicht immer zutreffen. Ein solcher Effekt von Kollinearität ist in den vorliegenden Regressionen nicht zu erkennen.
Der Ort des Genoms auf dem ein Gen liegt, welches auf ein quantitatives Merkmal wirkt, wird als *quantitativ trait locus* (QTL) bezeichnet (Griffiths *et al.*, 2012). Kover *et al.* (2009) geben für Arabidopsis die Positionen auf den Chromosomen an bei denen ein QTL identifiziert wurde. Diese entsprechen auch den Positionen auf denen die Effekte der vorgestellten penalisierten Regressionen groß waren.
Für Bayesianische Modelle werden im Vergleich zu frequentistischen Modellen a priori Annahmen über die Parameter getroffen. Um die Sensitivität des Bayesianischen Elastic Net gegenüber der Wahl der Hyperparameter zu untersuchen wurden verschiedene Hyperparameter gewählt und die resultierenden Modelle anhand der Kriterien Anzahl der effektiven Parameter, Devianz Informationskriterium und Korrelation zwischen realen und angepassten Werten miteinander verglichen. Die Wahl der Parameter der Priori-Verteilung von λ^2 ist sehr bedeutend für sinnvolle Regressionsergebnisse, da das Bayesianische Elastic Net sensibel auf die Hyperparameterwahl für λ^2 reagierte. Eine angemessene Wahl stellen die optimalen Parameter entsprechend Pérez *et al.* (2010) dar. Die Wahl der Hyperparameter der Anteilsvariable α für den Lasso und Ridge Parameter führte nur zu kleineren Unterschieden in den Inferezergebnissen. Die Anpassung des Modells an die Daten war durchwegs gut und die angepassten Werte bei verschiedenen Wahlen der Hyperparameter waren sehr ähnlich. Der Konvergenz des Intercept und der Steigungsparameter war immer gegeben. Bezüglich dieser Parameter ist das Bayesianische Elastic Net robust bei den verschiedenen Wahlen der Hyperparameter für die Priori-Verteilung von α. Für den Lasso Parameter und den Ridge Parameter war die Konvergenz nicht gewährleistet. Der Lasso Parameter hatte für alle Zielgrößen eine Dominanz gegenüber den Ridge Parameter. Zukünftig könnte untersucht werden, ob die Konvergenzprobleme der Penalisierungsparameter durch die Wahl anderer Priori-Verteilungen lösbar sind.
Die Vorhersagegenauigkeit aller Methoden wurde über eine Kreuzvalidierung mit den Gütekriterien Korrelation und Mittlerer Quadratischer Fehler überprüft. Für keine Zielgröße wurde die Methodengleichheit sowohl über die Korrelation als auch über den Mittleren Quadratischen Fehler abgelehnt. Es kann nicht eindeutig gezeigt werden, dass es einen Unterschied der Methoden gibt.
Penalisierte Regressionsmodelle bewirken eine Schrumpfung der Parameter. Zur visuellen Darstellung des Schrumpfungseffekts wurden die Effekte der

Einzelmarkerregression gegen die Effekte der penalisierten Regressionsmodelle abgetragen. Diese Schrumpfung war über alle Modelle und Phänotypen deutlich zu erkennen. Es ist übergreifend keine Aussage darüber zu treffen, welches Modell im Allgemeinen am stärksten schrumpft.
Riedelsheimer *et al.* (2012b) analysieren den Einfluss des Genom der Maispflanze, bei dem ein sehr hohes Kopplungsungleichgewicht vorliegt, auf die Metaboliten der Maispflanze mit den Methoden Lasso, Ridge und Elastic Net. Bei Arabidopsis lag, insbesondere ab einer Distanz von 5 Mb, ein schwaches Kopplungsungleichgewicht vor. Das Kopplungsungleichgewicht des Mais (Riedelsheimer *et al.*, 2012a) ist stärker als das Kopplungsungleichgewicht der Arabidopsis. Ein Vergleich der Effektstärken zeigt, dass der Gruppierungseffekt des Ridge bei Mais deutlich stärker ist als bei Arabidopsis. Dies bestätigt, dass ein hohes Kopplungsungleichgewicht zu einer hohen Korrelation der Kovariablen und somit zu einem starken Gruppierungseffekt bei der Ridge Regression führt (Zou und Hastie, 2005).
Die Anzahl der nicht-Nullparameter lag bei dem Generalisierten Elastic Net vergleichsweise niedrig. Das Generalisierte Elastic Net besitzt für $n \to \infty$, im Vergleich zu den anderen vorgestellten Modellen, die Eigenschaft der Fan-Li *oracle property* (Ishwaran und Rao, 2011). Dementsprechend könnte die Modellgröße des Generalisierten Elastic Net die unbekannte Wahrheit am besten widerspiegeln. Ob die Modellgröße des Generalisierten Elastic Net der Anzahl der wahren nicht-Nullkoeffizienten entspricht kann auf Grund dessen, dass es sich um eine experimentelle Datengrundlage handelt nicht nachgewiesen werden.
Resende Jr *et al.* (2012) vergleichen die Methoden Ridge, BayesA, BayesB und das Bayesianische Lasso nach Legarra *et al.* (2011) bezüglich der Vorhersagegenauigkeit zur genomischen Vorhersage und kommen zu dem Ergebnis, dass sich die Methoden nur geringfügig unterscheiden. Ein weiterer Methodenvergleich wird von Heslot *et al.* (2012) durchgeführt. Dabei werden neun Datensätze und elf Methoden bezüglich der genomischen Vorhersage analysiert. Die betrachteten Methoden sind unter anderem das Ridge, Elastic Net, Bayesianische Lasso, BayesA, BayesB, Gewichtete Bayesianische Schrumpfung (Hayashi und Iwata, 2010), E-Bayes (XU und HU, 2011) und *Machine-learning* Methoden (Breiman, 2001, Drucker *et al.*, 1997, Gardner und Dorling, 1998). Die mittlere Vorhersagegüte der Methoden ist sehr ähnlich. Auch die lineare Kombination verschiedener Modelle führt zu keiner Verbesserung der Genauigkeit.
Zusammenfassend wird festgestellt, dass die Elastic Net Methoden keine signifikant besseren Ergebnisse liefern als die Ridge und Lasso Methoden und dass die Bayesianischen Methoden den frequentistischen Methoden nicht immer überlegen sind.

Möglicherweise liefern nicht-lineare Modelle oder Interaktionen eine Verbesserung der Vorhersagegenauigkeit. Dies gilt es in Zukunft zu untersuchen.

Anhang

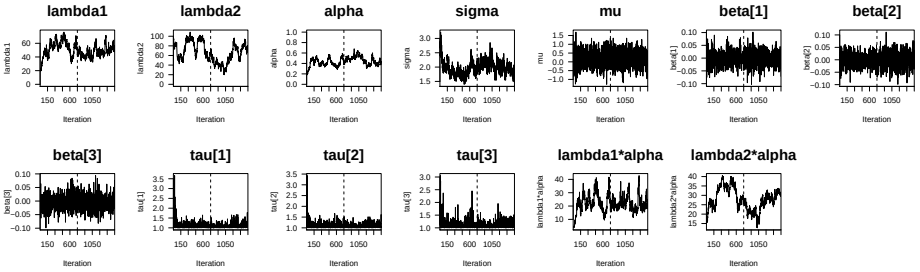

Abbildung 11: Pflanzenhöhe, E(α)=0.1,Var(α)=0.000001

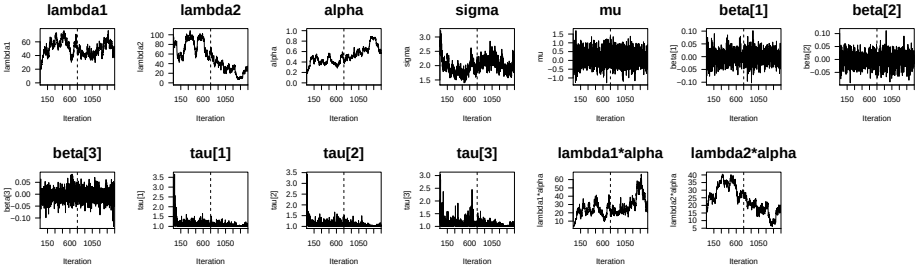

Abbildung 12: Pflanzenhöhe, E(α)=0.1,Var(α)=0.0001

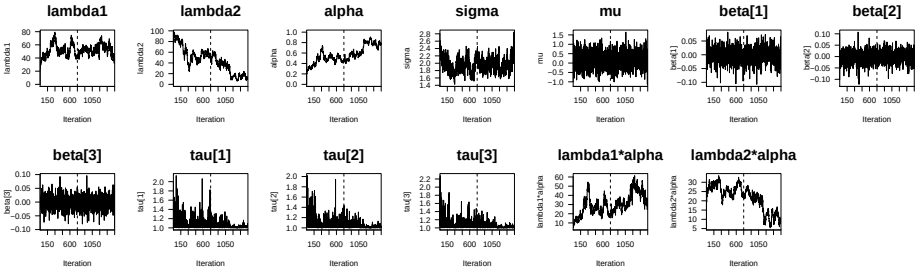

Abbildung 13: Pflanzenhöhe, E(α)=0.1,Var(α)=0.01

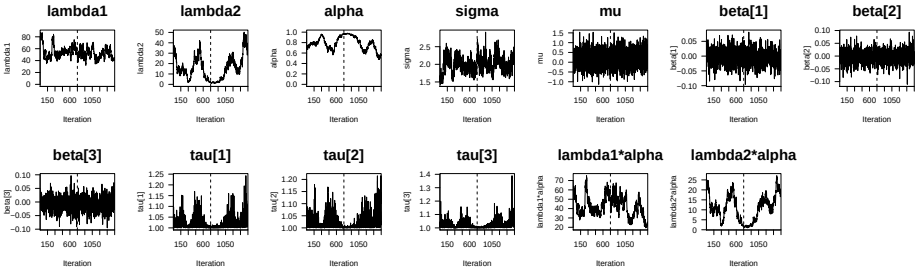

Abbildung 14: Pflanzenhöhe, E(α)=0.5,Var(α)=0.000001

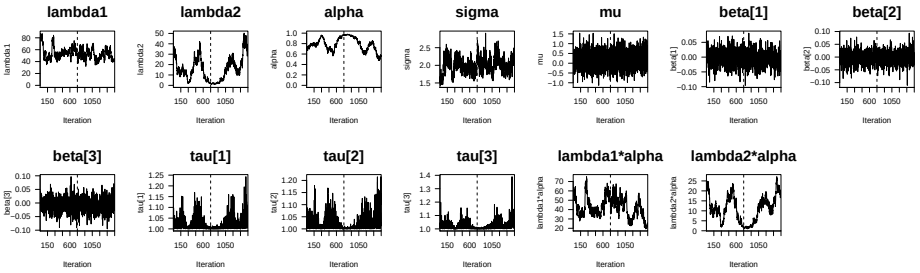

Abbildung 15: Pflanzenhöhe, E(α)=0.5,Var(α)=0.0001

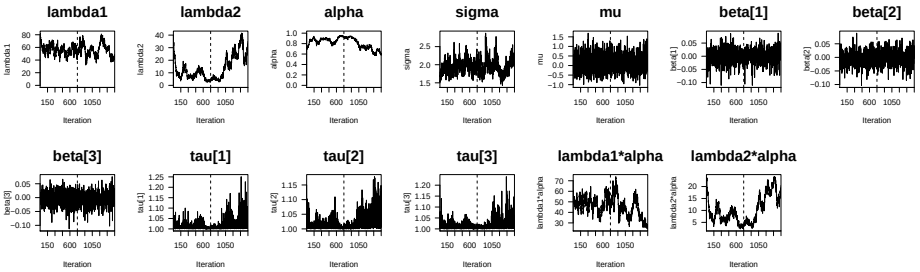

Abbildung 16: Pflanzenhöhe, E(α)=0.5,Var(α)=0.01

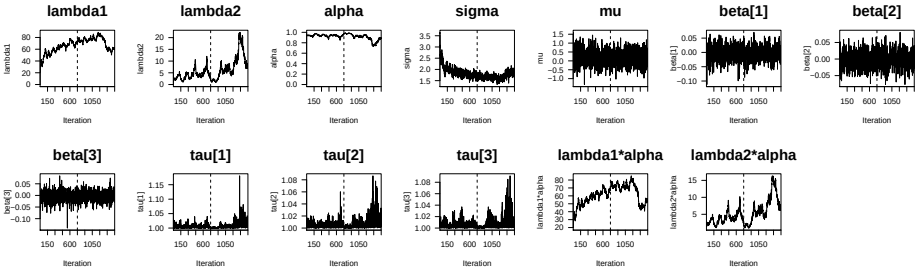

Abbildung 17: Pflanzenhöhe, E(α)=0.9,Var(α)=0.000001

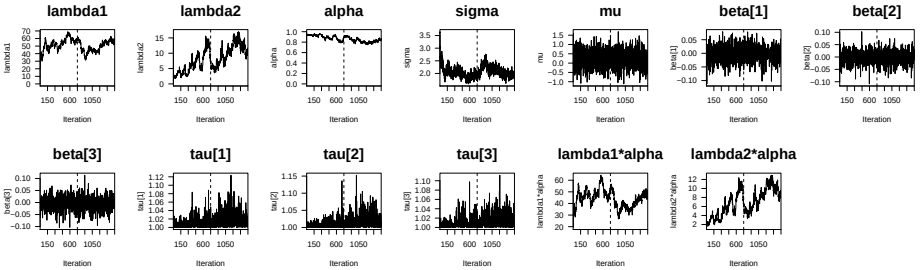

Abbildung 18: Pflanzenhöhe, E(α)=0.9,Var(α)=0.0001

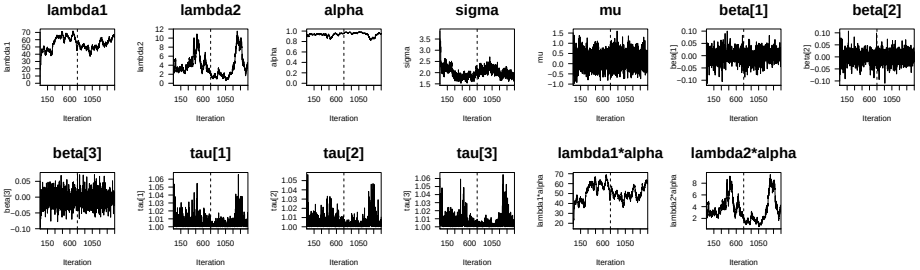

Abbildung 19: Pflanzenhöhe, E(α)=0.9,Var(α)=0.01

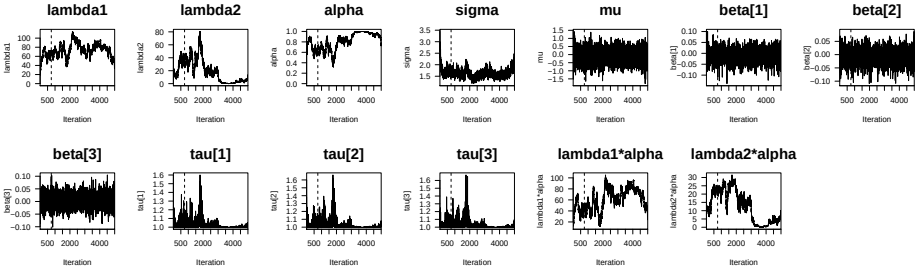

Abbildung 20: Log. Anzahl an Tagen zwischen Schossbeginn und Blütezeit, E(α)=0.5,Var(α)=0.0001, 50000 Iterationen

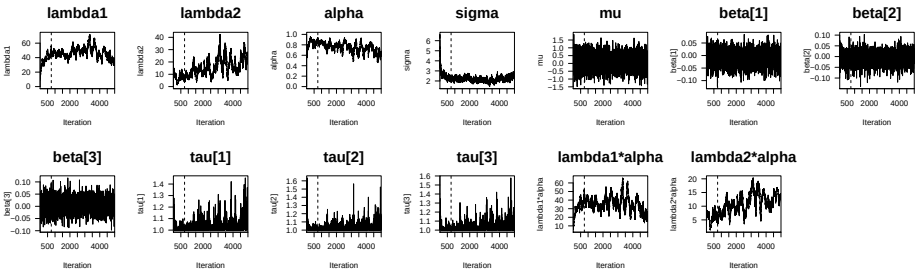

Abbildung 21: Log. Anzahl an Tagen bis zum Schossbeginn, E(α)=0.5,Var(α)=0.0001, 50000 Iterationen

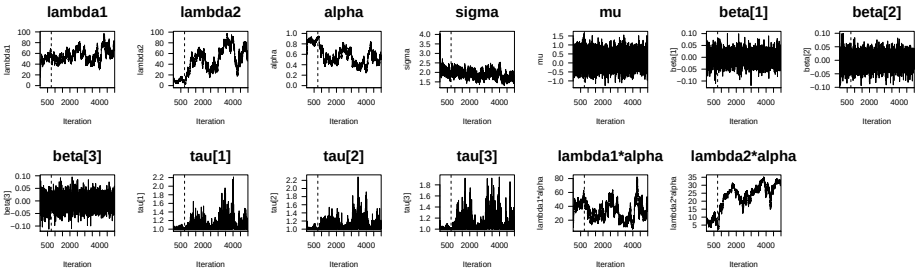

Abbildung 22: Pflanzenhöhe, E(α)=0.5,Var(α)=0.0001, 50000 Iterationen

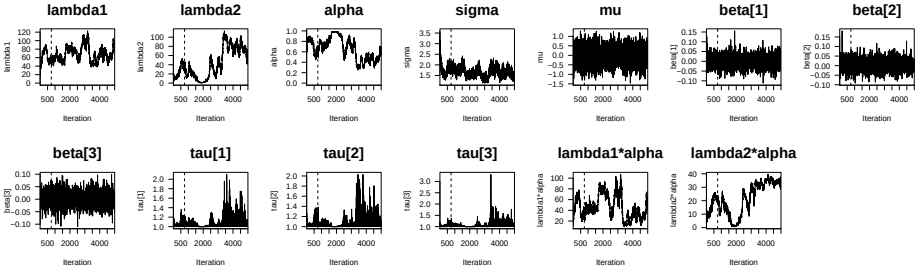

Abbildung 23: Wachstumsrate, E(α)=0.5,Var(α)=0.0001, 50000 Iterationen

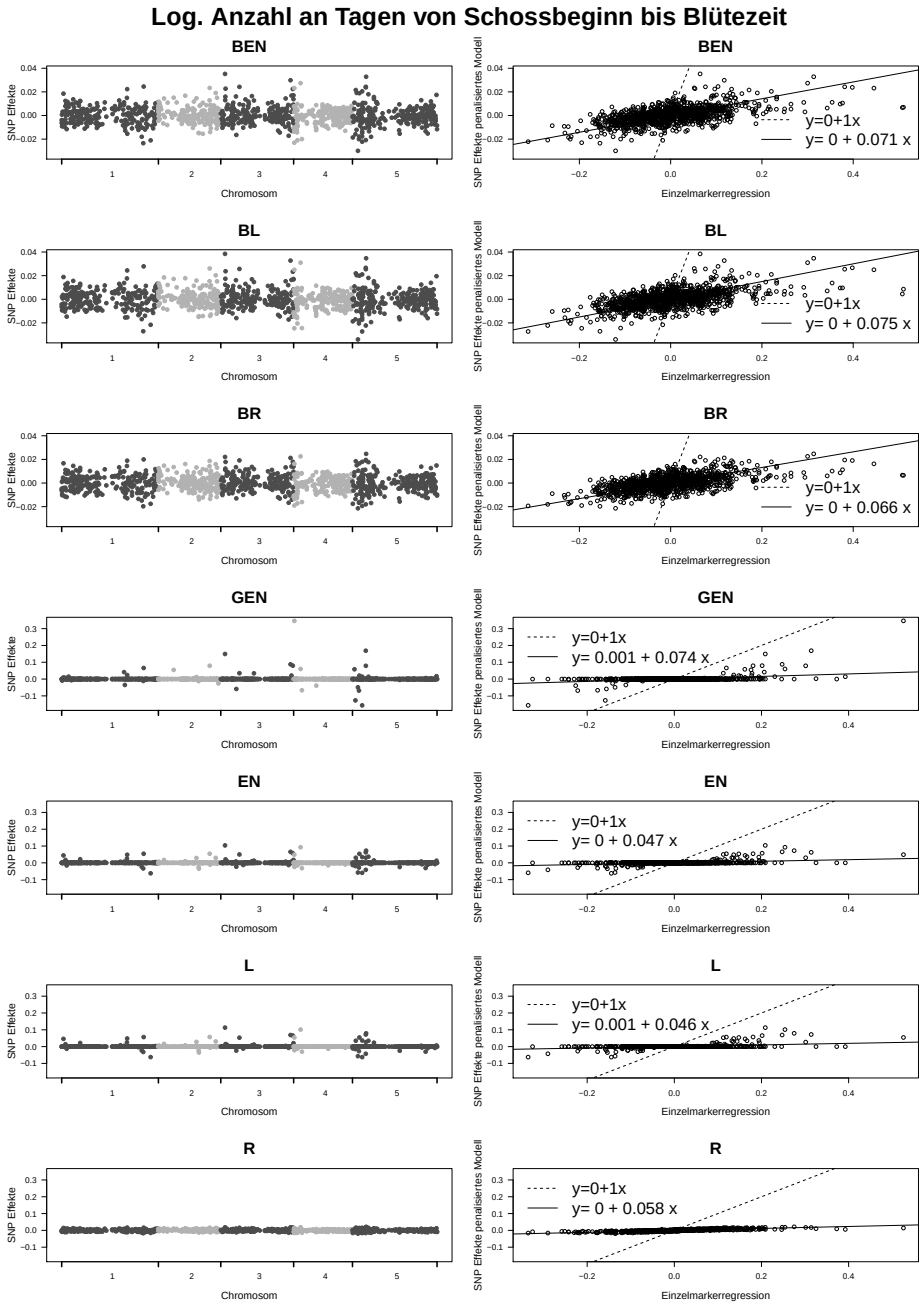

Abbildung 24: Logarithmierte Zeit zwischen Schossbeginn und Blütezeit, Manhattan-Plot der SNP Effekte und Streudiagramm der SNP Effekte der penalisierten Modelle und der nicht simultanen SNP Effekte eines linearen Modells

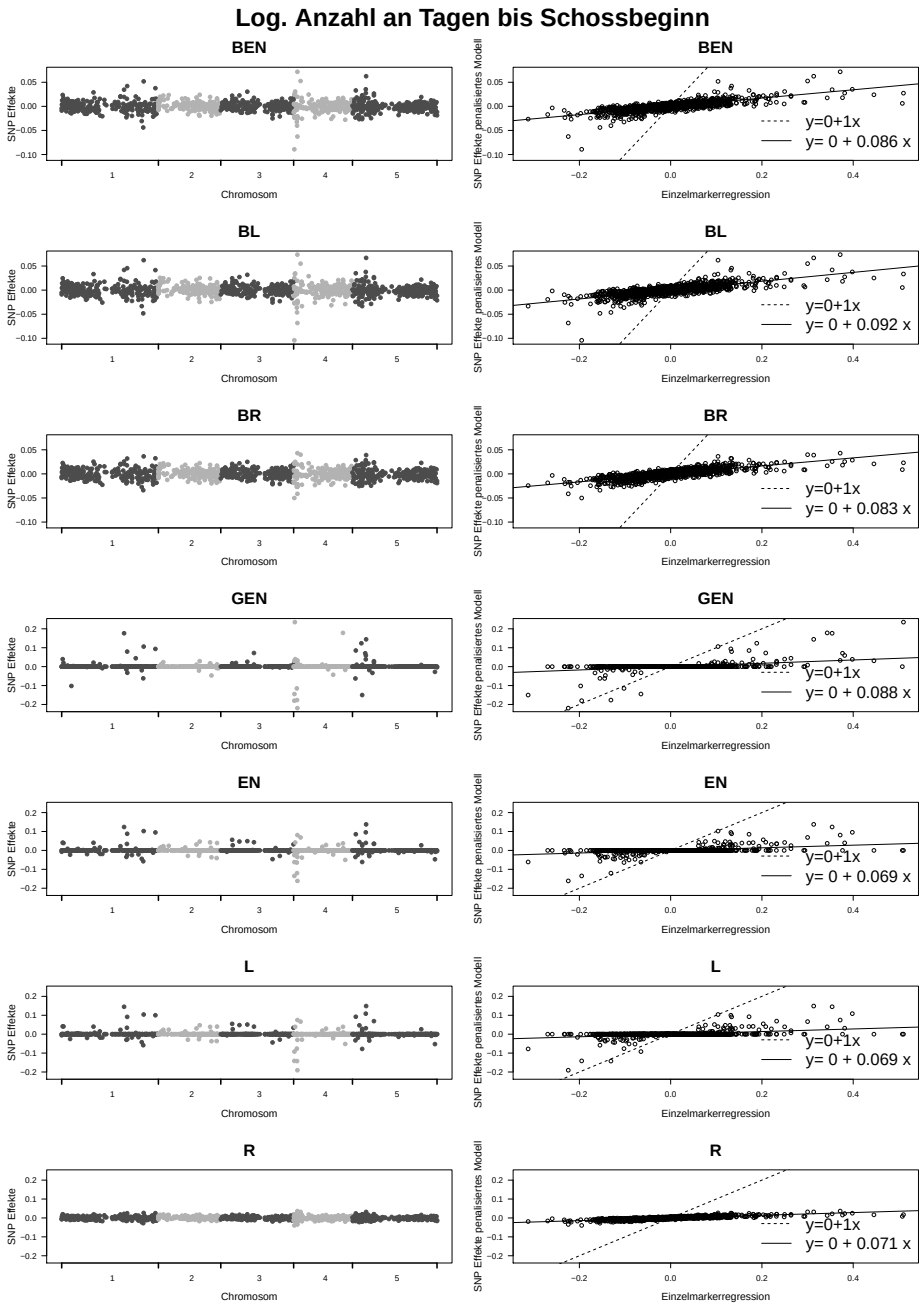

Abbildung 25: Log. Anzahl an Tagen bis zum Schossbeginn, Manhattan-Plot der SNP Effekte und Streudiagramm der SNP Effekte der penalisierten Modelle und der nicht simultanen SNP Effekte eines linearen Modells

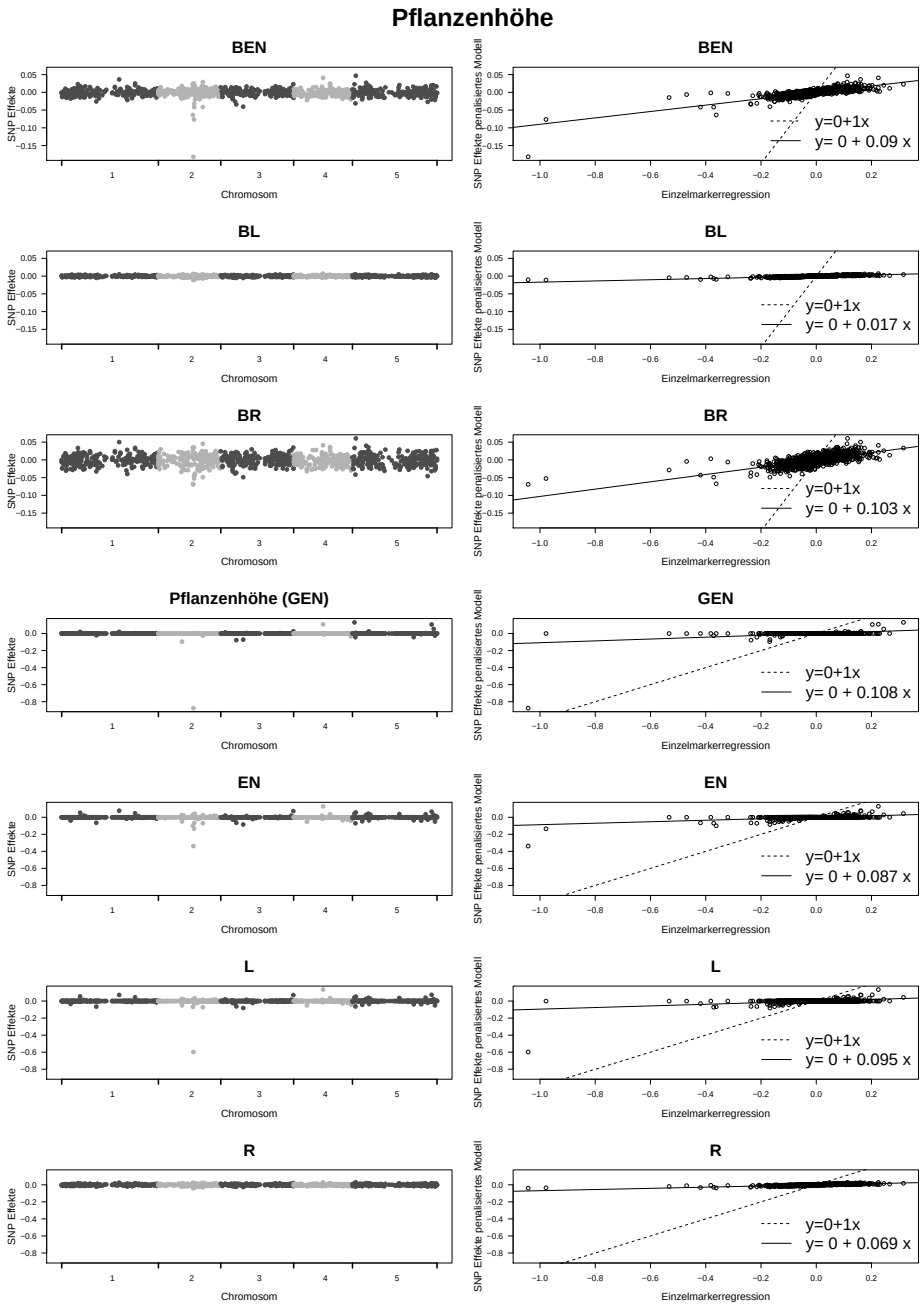

Abbildung 26: Pflanzenhöhe, Manhattan-Plot der SNP Effekte und Streudiagramm der SNP Effekte der penalisierten Modelle und der nicht simultanen SNP Effekte eines linearen Modells

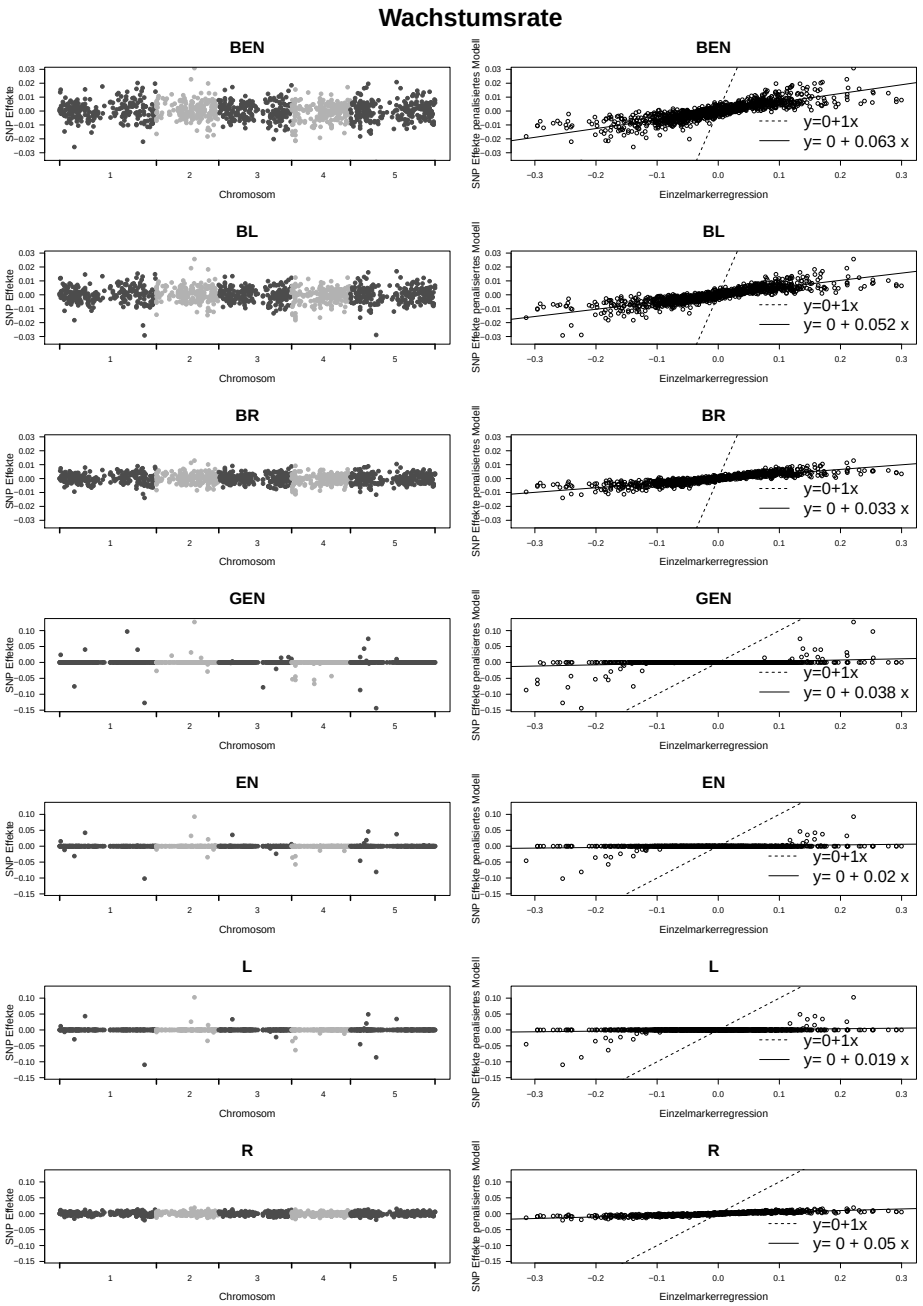

Abbildung 27: Wachstumsrate, Manhattan-Plot der SNP Effekte und Streudiagramm der SNP Effekte der penalisierten Modelle und der nicht simultanen SNP Effekte eines linearen Modells

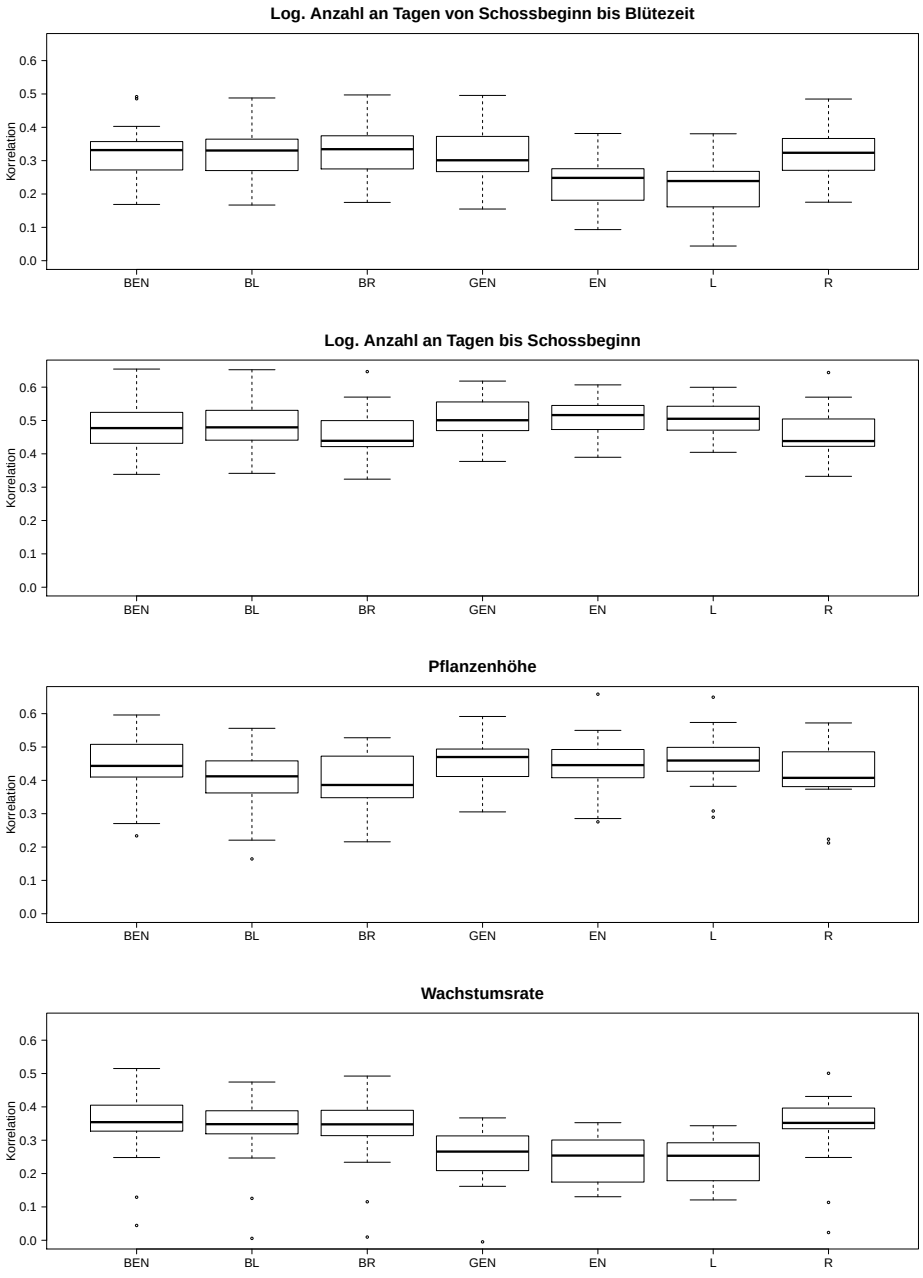

Abbildung 28: Prognosegenauigkeit penalisierter Methoden gemessen über eine fünffache Kreuzvalidierung mit drei Wiederholungen und dem Kriterium Korrelation

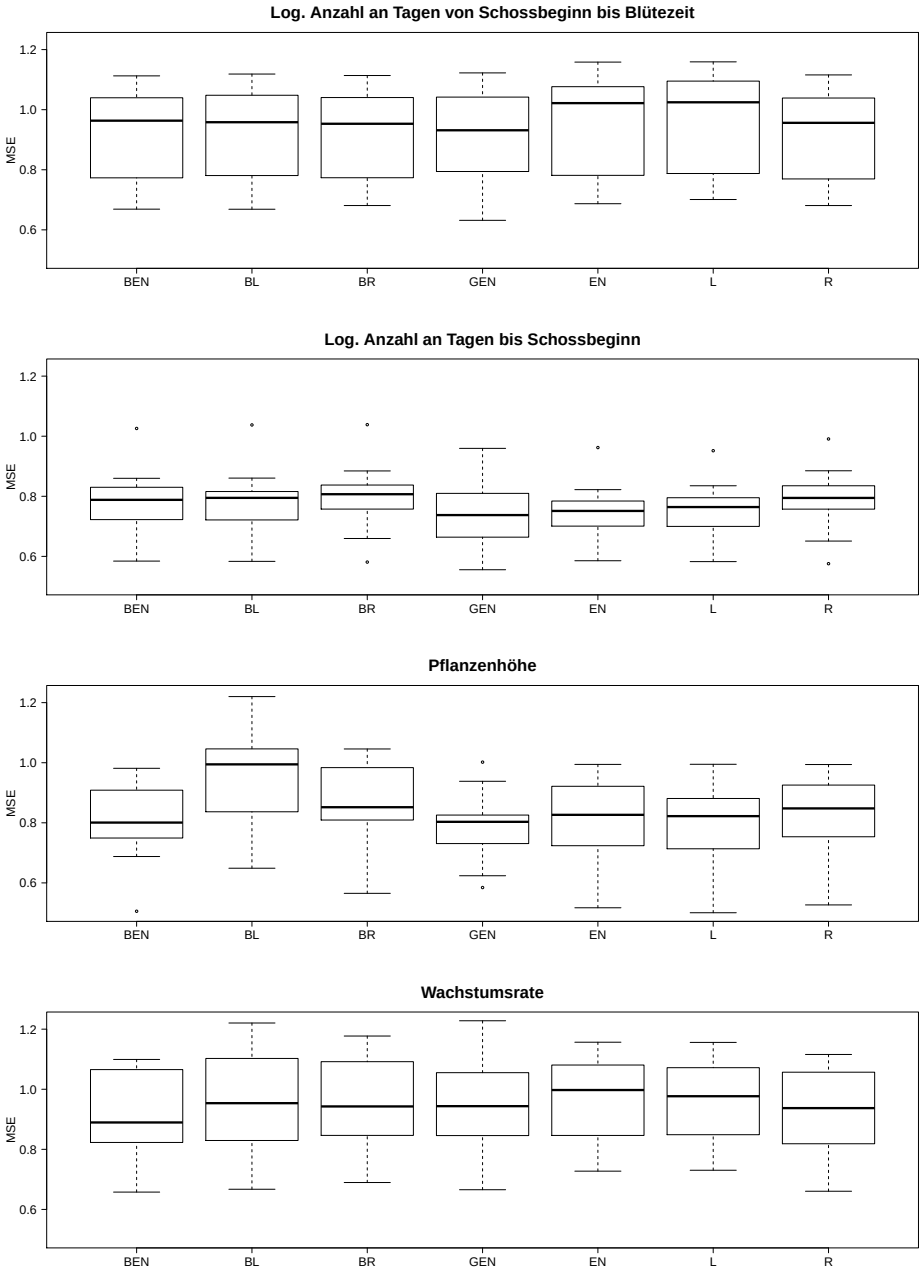

Abbildung 29: Prognosegenauigkeit penalisierter Methoden gemessen über eine dreifache Kreuzvalidierung mit drei Wiederholungen und dem Kriterium Mittlerer Quadratischer Fehler

Literatur

Batah F, Gore S (2009). "Ridge regression estimator: combining unbiased and ordinary Ridge regression methods of estimation." *Surveys in Mathematics and its Applications*, **4**, 99–109.

Breiman L (2001). "Random forests." *Machine learning*, **45**(1), 5–32.

Casella G (2001). "Empirical Bayes Gibbs sampling." *Biostatistics*, **2**(4), 485–500.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes J (2009). "Predicting quantitative traits with regression models for dense molecular markers and pedigree." *Genetics*, **182**(1), 375–385.

de los Campos G, Rodriguez PP (2012). *BLR: Bayesian Linear Regression*. R package version 1.3, URL `http://CRAN.R-project.org/package=BLR`.

Drucker H, Burges C, Kaufman L, Smola A, Vapnik V (1997). "Support vector regression machines." *Advances in neural information processing systems*, **9**, 155–161.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). "Least angle regression." *The Annals of statistics*, **32**(2), 407–499.

Fahrmeir L, Kneib T, Konrath S (2010). "Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection." *Statistics and Computing*, **20**(2), 203–219.

Fahrmeir L, Kneib T, Lang S (2007). *Regression: Modelle, Methoden und Anwendungen*. Springer.

Fahrmeir L, Künstler R, Pigeot I, Tutz G (2003). *Statistik: Der Weg zur Datenanalyse*. Springer.

Fan J, Li R (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Fan J, Lv J (2008). "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(5), 849–911.

Fernando R, Grossman R (1989). "Marker-assisted selection using best linear unbiased prediction." *Genetics Selection Evolution*, **21**, 467–477.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22.

Gardner M, Dorling S (1998). "Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences." *Atmospheric environment*, **32**(14-15), 2627–2636.

Geman S, Geman D (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6), 721–741.

Griffiths A, Wessler S, Carroll S, Doebley J (2012). *Introduction to genetic analysis*. W.H. Freeman.

Hastie T, Tibshirani R, Friedman J (2009). *The elements of statistical learning*, volume 2. Springer Series in Statistics.

Hastings W (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, **57**(1), 97–109.

Hayashi T, Iwata H (2010). "EM algorithm for Bayesian estimation of genomic breeding values." *BMC genetics*, **11**(1), 3.

Hayes B, Pryce J, Chamberlain A, Bowman P, Goddard M (2010). "Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits." *PLoS Genetics*, **6**(9), e1001139.

Henderson C (1984). "Applications of linear models in animal breeding." *University of Guelph.*

Heslot N, Yang H, Sorrells M, Jannink J (2012). "Genomic selection in plant breeding: A comparison of models." *Crop Science*, **52**(1), 146–160.

Hill W, Robertson A (1968). "Linkage disequilibrium in finite populations." *TAG Theoretical and Applied Genetics*, **38**(6), 226–231.

Hoerl A, Kennard R (1970a). "Ridge regression: applications to nonorthogonal problems." *Technometrics*, **12**(1), 69–82.

Hoerl A, Kennard R (1970b). "Ridge regression: Biased estimation for non-orthogonal problems." *Technometrics*, **12**(1), 55–67.

Hofmarcher P, Cuaresma JC, Grün B, Hornik K (2011). "Fishing Economic Growth Determinants Using Bayesian Elastic Nets." *Report 113*, Research Report Series, Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien. URL `http://epub.wu.ac.at/3213/`.

Ishwaran H, Kogalur U, Rao J (2010a). "spikeslab: Prediction and Variable Selection Using Spike and Slab Regression." *The R Journal*, **2**, 68–73.

Ishwaran H, Kogalur U, Rao J (2010b). *spikeslab: Prediction and variable selection using spike and slab regression*. R package version 1.1.2, URL `http://CRAN.R-project.org`.

Ishwaran H, Rao J (2011). "Generalized ridge regression: Geometry and computational solutions when p is larger than n." *Technical Report*.

Jannink J, Lorenz A, Iwata H (2010). "Genomic selection in plant breeding: from theory to practice." *Briefings in Functional Genomics*, **9**(2), 166–177.

Knust E, Janning W (2008). *Genetik: Allgemeine Genetik-Molekulare Genetik-Entwicklungsgenetik*. Thieme.

Kover P, Valdar W, Trakalo J, Scarcelli N, Ehrenreich I, Purugganan M, Durrant C, Mott R (2009). "A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana." *PLoS Genetics*, **5**(7), e1000551.

Legarra A, Robert-Granié C, Croiseau P, Guillaume F, Fritz S, *et al.* (2011). "Improved Lasso for genomic selection." *Genetics research*, **93**(1), 77.

Lewontin R, Kojima K (1960). "The evolutionary dynamics of complex polymorphisms." *Evolution*, **14**(4), 458–472.

Li Q, Lin N (2010). "The Bayesian elastic net." *Bayesian Analysis*, **5**(1), 151–170.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953). "Equation of state calculations by fast computing machines." *The journal of chemical physics*, **21**, 1087.

Meuwissen T, Goddard M (1996). "The use of marker haplotypes in animal breeding schemes." *Genetics Selection Evolution*, **28**, 161–176.

Meuwissen T, Hayes B, Goddard M (2001). "Prediction of total genetic value using genome-wide dense marker maps." *Genetics*, **157**(4), 1819–1829.

Miller A (2002). *Subset selection in regression*. Chapman & Hall/CRC.

Park T, Casella G (2008). "The bayesian lasso." *Journal of the American Statistical Association*, **103**(482), 681–686.

Pérez P, de los Campos G, Crossa J, Gianola D (2010). "Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression package in R." *The plant genome*, **3**(2), 106–116.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

Resende Jr M, Muñoz P, Resende M, Garrick D, Fernando R, Davis J, Jokela E, Martin T, Peter G, Kirst M (2012). "Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.)." *Genetics*, **190**(4), 1503–1510.

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger A (2012a). "Genomic and metabolic prediction of complex heterotic traits in hybrid maize." *Nature genetics*, **44**(2), 217–220.

Riedelsheimer C, Technow F, Melchinger A (2012b). "Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines." *BMC genomics*, **13**, 452.

Robert C, Casella G (2004). *Monte Carlo statistical methods*. Springer Verlag.

Rüger B (1999). *Test-und Schätztheorie 1. Grundlagen*, volume 1. Oldenbourg Wissenschaftsverlag.

Spiegelhalter D, Best N, Carlin B, Van Der Linde A (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.

Su YS, Yajima M (2012). *R2jags: A Package for Running jags from R*. R package version 0.03-07, URL `http://CRAN.R-project.org/package=R2jags`.

Tibshirani R (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Wimmer V, Albrecht T, Auinger HJ, Schoen CC (2012). "synbreed: a framework for the analysis of genomic prediction data using R." *Bioinformatics*, **28**(15), 2086–2087.

XU S, HU Z (2011). "Methods of plant breeding in the genome era." *Genetics Research*, **92**(5), 423.

Zou H, Hastie T (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

Eigenständigkeitserklärung

Hiermit erkläre ich, Claudia Stuckart, dass es sich bei der vorliegenden Masterarbeit um eine selbständig verfasste Arbeit handelt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden.

München, den 18. Dezember 2012

Claudia Stuckart