- LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN -

INSTITUT FÜR STATISTIK

Modelluntersuchung bei Anwendung von Lasso auf Bootstrap-Stichproben

BACHELORARBEIT

ZUR ERLANGUNG DES AKADEMISCHEN GRADES
BACHELOR OF SCIENCE (B.SC.)

Gutachterin: Prof. Dr. Anne-Laure Boulesteix

Betreuerin: M.Sc. Silke Janitza

Autorin: Johanna Völkl

Abgabedatum: 07.08.2015

Abstract

Im Lasso-Verfahren wird die Größe der Regressionskoeffizienten so restringiert, dass der Effekt mancher Kovariablen auf Null geschätzt wird. Durch diese Verknüpfung von Variablenselektion und Schätzung der Regressionskoeffizienten bietet das Lasso-Verfahren besonders für $p \gg n$ eine gute Alternative zum weit verbreiteten KQ-Schätzer. Damit auch unter vergleichsweise schwachen Annahmen eine konsistente Schätzung hervorgeht, wurden in der Literatur Methoden vorgeschlagen, in denen das Lasso-Verfahren auf Bootstrap- beziehungsweise Subsampling-Stichproben durchgeführt wird. Zudem ist aus der Literatur bekannt, dass Modellselektionsverfahren angewandt auf Bootstrap-Stichproben meist sehr komplexe Modelle liefern. Dies soll auch für die Anwendung des Lasso-Verfahrens auf Bootstrap-Stichproben untersucht werden. Dazu werden die resultierenden Modelle basierend auf Bootstrap-Stichproben mit denen für Original-Datensätze verglichen. Zusätzliche werden die Modelle betrachtet, die bei der Anwendung vom Lasso-Verfahren auf Subsampling-Stichproben entstehen. Ziel dieser Arbeit ist es, zu prüfen, ob die Resampling-Methoden verbunden mit dem Lasso-Verfahren wünschenswerte Ergebnisse erzielen. Dazu werden nach einer theoretischen Einführung in das Lasso-Verfahren die Ergebnisse für Bootstrap und Subsampling basierend auf simulierten Daten gegenübergestellt. Hierbei werden die Modellkomplexität, die Inclusion Frequencies und die Prädiktionsgüte betrachtet. Alle durchgeführten Analysen basieren auf unkorrelierten, normalverteilten Variablen, die zuvor simuliert wurden.

Die vorgenommenen Auswertungen zeigen, dass keine Resampling-Methode klar zu bevorzugen ist. Für Modelle, die möglichst alle relevanten Variablen enthalten und gute Vorhersagen treffen sollten, sollten eher Bootstrap-Stichproben der Größe n verwendet werden. Einen Kompromiss zwischen Interpretierbarkeit und guter Prädiktionsgüte bietet Subsampling.

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Einleitung

Besonders in der Genetik übersteigt die Menge potentieller Einflussvariablen oftmals die Zahl der Beobachtungen um ein Vielfaches. Dieses Problem wird als $p \gg n$ bezeichnet, wobei p die Anzahl an Variablen und n die Anzahl an Beobachtungen beschreibt. Klassische statistische Methoden wie der Kleinste-Quadrate-Schätzer sind in diesem Fall bei der Schätzung der β-Koeffizienten nicht mehr stabil und somit ungeeignet. Ein weiteres Risiko birgt die Aufnahme aller p Kovariablen in ein Regressionsmodell. Dadurch könnte zwar eine sehr gute Modellanpassung an die zur Schätzung genutzten Daten erreicht werden, aber das Modell wäre aufgrund eines resultierenden Overfittings zur Prognose kaum geeignet. Darüber hinaus wäre es durch die Vielzahl an Kovariablen sehr schwer interpretierbar. (Fahrmeir et al., 2013; Bühlmann und van de Geer, 2011)

Eine komfortable Lösung dieser Probleme schlug Tibshirani (1996) mit dem sogenannte **L**east **A**bsolute **S**hrinkage and **S**election **O**perator, kurz Lasso, vor. Hierbei werden die absoluten Werte der β-Koeffizienten geschrumpft und simultan eine Variablenselektion durchgeführt. Diese Kombination aus Schätzung und Variablenselektion macht das Lasso-Verfahren für den Anwender sehr attraktiv. Als ungünstig erweist sich jedoch, dass bei Verwendung des Penalisierungsparameters, der den kleinsten Prädiktionsfehler liefert, neben den relevanten auch meist irrelevante Variablen nach der Selektion im Modell enthalten bleiben. Zudem stellt sich die Frage, ob prinzipiell ein Penalisierungsparameter existiert, für den die selektierten Variablen auch genau den relevanten entsprechen. Existiert solch ein Parameter, sodass für n gegen unendlich die Wahrscheinlichkeit, nur genau die relevanten Variablen zu selektieren, gegen 1 geht, so würde man die Selektion als konsistent bezeichnen. Dies ist jedoch nur unter vergleichsweise strengen Annahmen gegeben. (Bühlmann und van de Geer, 2011; Meinshausen und Bühlmann, 2006)

Um auch unter schwächeren Annahmen eine konsistente Schätzung erreichen zu können, wurden zahlreiche Modifikationen des ursprünglichen Verfahrens entwickelt. Eine Möglichkeit ist die Anwendung des Lasso-Verfahrens auf Bootstrap-Stichproben. Im Bootstrap-Verfahren werden durch zufällige Ziehungen aus dem Original-Datensatz mehrere Pseudo-Datensätze generiert. Diese Ziehungen erfolgen mit Zurücklegen. Alternativ können auch Pseudo-Datensätze über Subsampling, das heißt Ziehen ohne Zurücklegen, erzeugt werden. Somit stehen dem Anwender

zur Variablenselektion nicht nur einer, sondern mehrere Datensätze zur Verfügung. Dies ermöglicht nicht nur die Unsicherheit eines Selektionsverfahrens zu quantifizieren, sondern auch die Stabilität zu vergrößern und Konsistenz zu schaffen. So zeigt Bach (2008), dass unter bestimmten Annahmen das Lasso-Verfahren alle relevanten Variablen mit einer Wahrscheinlichkeit, die für n gegen unendlich gegen 1 geht, identifiziert. Währenddessen besteht für irrelevante Variablen lediglich eine echt positive Wahrscheinlichkeit, in das geschätzte Modell aufgenommen zu werden. Dementsprechend befinden sich nach Anwendung des Lasso-Verfahrens auf verschiedene Bootstrap-Stichproben die relevanten Variablen in allen Modellen, während die irrelevanten nur zufällig aufgenommen werden. Durch Betrachtung der Modelle aller Bootstrap-Stichproben können somit relevante Variablen identifiziert und eine konsistente Variablenselektion erreicht werden. Eine weitere Möglichkeit, das Lasso-Verfahren zu verbessern, begründen Meinshausen und Bühlmann (2010) mit ihrem Konzept der Stability Selection. Hierbei werden mittels Subsampling verschiedene Pseudo-Datensätze gebildet. Nach Anwendung des Lasso-Verfahrens auf jeden der Pseudo-Datensätze wird für verschiedene Penalisierungsparameter die empirische Wahrscheinlichkeit bestimmt, dass der Effekt einer bestimmten Variable ungleich Null geschätzt wird. Übersteigt diese Wahrscheinlichkeit einen gewählten Wert, so wird die jeweilige Variable als relevant angesehen. (Bühlmann und van de Geer, 2011; Henderson, 2005)

Motiviert durch diese Publikationen wird im Rahmen der vorliegenden Arbeit untersucht, wie sich das Lasso-Verfahren bei Anwendung auf verschiedene Resampling-Methoden verhält. Da aktuelle Veröffentlichungen (Janitza et al., in Druck; Binder und Schumacher, 2008) zeigen, dass Selektionsverfahren für Bootstrap-Stichproben tendenziell komplexere Modelle liefern als für Original-Datensätze, steht besonders ein Vergleich mit dem alternativen Subsampling im Vordergrund. Dazu werden 1 000 Datensätze generiert und daraus Bootstrap-Stichproben der Größen m und n sowie Subsamples der Größe m gezogen. Anschließend wird das Lasso-Verfahren auf alle Datensätze zur Variablenselektion angewandt. Hierfür wird für jeden Datensatz der Penalisierungsparameter λ verwendet, der den kleinsten Prädiktionsfehler liefert. Zur Evaluation der verschiedenen Resampling-Methoden wird die resultierende Modellkomplexität betrachtet. Diese beschreibt die Anzahl aufgenommener Variablen in einzelnen Modellen. Des Weiteren werden die Prädiktionsgüte und sogenannte Inclusion Frequencies für die verschiedenen Methoden verglichen. Die Inclusion Frequencies beschreiben hierbei für jede Variable den Anteil der Modelle, in denen ihr

Effekt ungleich Null geschätzt wird. Durch Untersuchung der verschiedenen Charakteristika können sowohl die Vorteile als auch Probleme der Anwendung von Lasso für verschiedene Resampling-Methoden analysiert wird. Diese können bei der Entwicklung neuer Methoden berücksichtigt werden. (De Bin et al., in Druck)

Die vorliegende Arbeit ist folgendermaßen gegliedert: In Kapitel 2 wird die verwendete Methodik vorgestellt. Dazu werden das Lasso-Verfahren ausführlich erläutert und die Resampling-Methoden Bootstrap und Subsampling vorgestellt. Anschließend erfolgt in Kapitel 3 eine Anwendung der vorgestellten Methoden auf simulierte Daten. Hierfür wird zunächst auf die Vorgehensweise bei der Datensimulation näher eingegangen. Die Darstellung und der Vergleich der verschiedenen Resampling-Methoden erfolgen im Anschluss. Abschließend werden in Kapitel 4 die vorliegenden Ergebnisse kritisch diskutiert und weiterführende Gedanken formuliert.

2. Methodik

Grundlage der Analysen dieser Arbeit ist das Lasso-Verfahren. Bevor dieses genauer ausgeführt wird, ist der im linearen Regressionsmodell standardmäßig verwendete KQ-Schätzer zu erläutern. Basierend auf diesem wird anschließend das Lasso-Verfahren motiviert. Des Weiteren werden mögliche Probleme des Lasso-Verfahrens und Lösungsansätze durch die Kombination mit Bootstrap beschrieben. Obwohl das Lasso-Verfahren prinzipiell auch für generalisierte lineare Modelle anwendbar ist, beschränken sich die folgenden Erläuterungen und Analysen lediglich auf klassische lineare Modelle.

2.1. KQ-Schätzer

Der Einfluss von p erklärenden Kovariablen $x_1, ..., x_p$ auf eine interessierende Variable y wird häufig durch ein multiples lineares Regressionsmodell der Form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, \quad i = 1, ..., n$$

dargestellt. Oftmals wird auch die alternative Matrixnotation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

mit dem Vektor der Zielgrößen $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ und dem der Störgrößen $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$ und der Designmatrix $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$ verwendet. Für das lineare Regressionsmodell müssen folgende Annahmen gelten:

1. Die Störgrößen sind im Mittel Null, d.h. $\mathrm{E}(\epsilon_i) = 0$
2. Die Störgrößen sind homoskedastisch und unkorreliert, d.h. $Cov(\epsilon) = \mathrm{E}(\epsilon\epsilon^T) = \sigma^2 I$
3. Die Störgrößen sind normalverteilt, d.h. $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
4. Die Designmatrix X besitzt vollen Spaltenrang, d.h. rg(X) = p

Zur Schätzung der unbekannten Regressionskoeffizienten β_j wird in der Regel der Kleinste-Quadrate-Schätzer, im folgenden als KQ-Schätzer bezeichnet, verwendet.

Hierbei wird zunächst die Summe der quadrierten Abweichungen gebildet, welche in Matrix-Schreibweise über

$$KQ(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \tag{1}$$

dargestellt wird. Diese wird anschließend minimiert, indem die erste Ableitung

$$\frac{\partial KQ(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \tag{2}$$

mit Null gleichgesetzt wird. Um zu gewährleisten, dass es sich hierbei um ein Minimum handelt, wird zusätzlich die zweite Ableitung

$$\frac{\partial^2 KQ(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta^T}} = 2\mathbf{X}^T\mathbf{X}$$

betrachtet. Durch die vierte Modellannahme kann direkt gefolgert werden, dass die Matrix $\mathbf{X}^T\mathbf{X}$ positiv definit ist. Somit wird eine Minimierung genau dann erreicht, wenn die Ableitung in (2) gleich Null ist. Da positive Definitheit Invertierbarkeit impliziert, ist das Minimierungsproblem eindeutig mit

$$\hat{\boldsymbol{\beta}}_{KQ} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

lösbar. (Fahrmeir et al., 2009)

2.2. Motivation Shrinkage-Verfahren

Der KQ-Schätzer ist erwartungstreu und damit unverzerrt. Zudem weist er laut Gauß-Markov-Theorem unter allen linearen erwartungstreuen Schätzern die kleinste Varianz auf. Somit gilt er als BLUE (Best Linear Unbiased Estimator), das heißt als bester linearer unverzerrter Schätzer. Kritisch wird die Verwendung des KQ-Schätzers allerdings, wenn Spalten der Designmatrix nicht linear unabhängig sind oder mehr Kovariablen als Beobachtungen ($p > n$) vorliegen. In diesen Fällen besitzt die Designmatrix keinen vollen Spaltenrang und $\mathbf{X}^T\mathbf{X}$ ist nicht invertierbar. Infolgedessen ist die Lösung des KQ-Schätzers nicht mehr eindeutig und die Varianz der resultierenden β-Schätzer steigt stark an. (Fahrmeir et al., 1996; Fahrmeir et al., 2013)

Um auch in solchen Situationen adäquate Schätzer zu erhalten, wurden sogenannte Shrinkage-Verfahren entwickelt. Diese nehmen eine Verzerrung des Schätzers in

Kauf, um eine eindeutige Lösung bestimmen zu können. Dazu wird im Vergleich zur herkömmlichen KQ-Schätzung noch ein zusätzlicher Penalisierungsterm pen($\boldsymbol{\beta}$) eingeführt. Insgesamt wird also der Schätzer

$$\hat{\boldsymbol{\beta}}_{PKQ} = \underset{\beta}{\operatorname{argmin}}\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}, \quad \text{mit pen}(\boldsymbol{\beta}) \leq t \tag{3}$$

gebildet. Der Penalisierungsterm pen($\boldsymbol{\beta}$), welcher ein Maß für die Komplexität des Vektors der Regressionskoeffizienten darstellt, wird durch die Konstante t in seiner Größe beschränkt. Er steuert den Ausgleich zwischen Varianz und Bias des Schätzers (Bias-Varianz-Trade-off) und ist so konstruiert, dass er für wachsende β-Koeffizienten ansteigt. Dementsprechend wird die Größe der β-Koeffizienten durch t restringiert und die Varianz im Vergleich zum KQ-Schätzer verkleinert. (Fahrmeir et al., 2013)

2.3. Überblick Lasso-Verfahren

Mögliche Formen des Shrinkage-Verfahrens sind die Ridge-Regression von Hoerl und Kennard (1970) und das von Tibshirani (1996) vorgestellte Lasso-Verfahren. Die Abkürzung Lasso steht hierbei für **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. Dieses Verfahren bildet die Grundlage der vorliegenden Arbeit und wird im Folgenden genauer erläutert.

2.3.1. Definition

Während für die Ridge-Regression die L_2 - Norm zur Penalisierung verwendet wird, greift man für das Lasso-Verfahren auf die L_1 - Norm zurück. Dementsprechend wird der penalisierte KQ-Schätzer für das Lasso-Verfahren durch

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\beta}{\operatorname{argmin}}\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}, \quad \text{mit } \sum_{j=1}^{p} |\beta_j| \leq t \tag{4}$$

dargestellt. Wie der Name des Verfahrens bereits erkennen lässt, werden hierbei die absoluten Werte der β-Koeffizienten durch die zusätzliche Restriktion geschrumpft und teilweise gleich Null gesetzt. Somit findet bei diesem Verfahren simultan zur Schätzung eine Variablenselektion statt. Eine weitere, äquivalente Darstellung des

Lasso-Problems ist

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\beta}{\operatorname{argmin}}\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|\}, \tag{5}$$

wobei $\lambda \geq 0$ einen Penalisierungsparameter bezeichnet. Die beiden Darstellungsformen (4) und (5) sind insofern äquivalent, dass für jedes $\lambda \in [0, \infty)$ ein $t \geq 0$ existiert, sodass beide Probleme die gleiche Lösung besitzen. (Leng et al., 2006; Fahrmeir et al., 2013)

An dieser Stelle gilt es zu beachten, dass sowohl in Gleichung (4) als auch in Gleichung (5) der Intercept β_0 nicht in den Penalisierungsterm mit eingeht. Andernfalls wäre die Schätzung der β-Koeffizienten von der Skalierung von Y abhängig. So würde eine Verschiebung aller Werte y_i um eine Konstante c nicht eine Verschiebung der Prädiktion um die gleiche Konstante, sondern eine veränderte Schätzung bewirken. Stattdessen werden im Vorhinein alle Kovariablen und der Response zentriert, sodass $\bar{y} = 0$ und $\bar{x} = 0$. Dies führt automatisch zu $\hat{\beta}_0 = 0$. Alternativ können auch nur die Variablen x_{ij} zentriert und der Intercept über $\hat{\beta}_0 = \bar{y}$ geschätzt werden. Da die resultierenden Schätzer keine Skaleninvarianz aufweisen, ist es zudem sinnvoll, die Kovariablen zu standardisieren. Aus Gründen der Übersichtlichkeit wird im Folgenden von standardisierten Kovariablen und einem zentrierten Response ausgegangen. (Hastie et al., 2009; Fahrmeir et al., 2013)

2.3.2. Allgemeine Eigenschaften

Wie bereits erläutert, besteht zwischen der Konstante t aus Gleichung (4) und dem Penalisierungsparameter $\lambda \geq 0$ aus Gleichung (5) eine eins-zu-eins Beziehung, sie sind jedoch nicht äquivalent. Beide Parameter steuern die Stärke der Penalisierung und dementsprechend auch den Grad der Schrumpfung. Wird λ sehr klein gewählt, so werden große Werte für $\hat{\beta}_j$ kaum bestraft und die Schätzer $\hat{\beta}_{j,Lasso}$ werden den KQ-Schätzern $\hat{\beta}_{j,KQ}$ sehr ähnlich sein. Wird λ hingegen sehr groß gewählt, so werden große Werte für $\hat{\beta}_j$ stärker bestraft. Die geschätzten Werte $\hat{\beta}_{j,Lasso}$ werden dementsprechend verhältnismäßig klein oder gleich Null sein. Für t ist das Schrumpfungsverhalten gegensätzlich. So führen große Werte von t zu einer schwachen und kleine Werte von t zu einer starken Schrumpfung. (Fahrmeir et al., 2013)

Abbildung 1 zeigt, basierend auf einem simulierten Datensatz, wie die Schätzung der

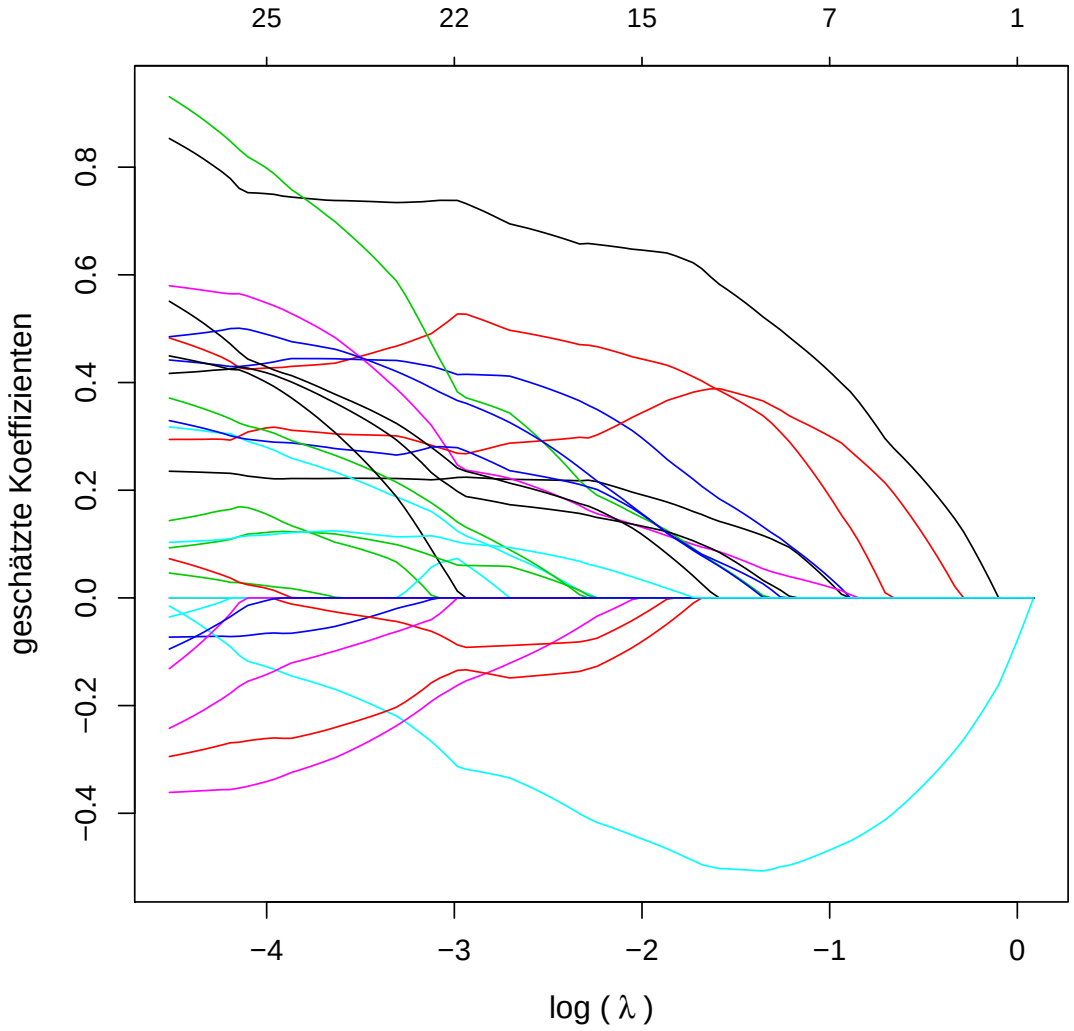

Abbildung 1: Veränderung von mit Hilfe des Lasso-Verfahrens geschätzten β-Koeffizienten in Abhängigkeit von $\log(\lambda)$

β-Koeffizienten in Abhängigkeit vom gewählten λ variieren kann. Hierbei wird die Darstellung der Veränderung eines einzelnen β-Koeffizienten als Regularisierungspfad bezeichnet. Aus Gründen der Übersichtlichkeit wird für λ häufig eine log-Skala verwendet. Da die Logarithmus-Funktion streng monoton steigend ist, bedeutet eine Zunahme von $\log(\lambda)$ auch eine Zunahme von λ. Wie zu erwarten werden β-Koeffizienten mit steigendem λ unterschiedlich schnell Richtung Null geschrumpft, bis schlussendlich alle Koeffizienten gleich Null geschätzt werden. Zusätzlich zum Penalisierungsparameter befindet sich noch eine weitere horizontale Achse in der Grafik. Diese gibt an, wie viele Koeffizienten sich noch im Modell befinden, das heißt ungleich Null geschätzt werden. Somit verschafft diese Darstellungsweise einen

schnellen Überblick, für welche Werte von λ die einzelnen Variablen im Modell enthalten sind. (Friedman et al., 2010)

Da im Lasso-Verfahren die absoluten Werte in den Penalisierungsterm eingehen, ist das penalisierte KQ-Kriterium aus Gleichung (5) nicht differenzierbar. Zur Bestimmung des Minimums müsste die Gleichung

$$2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + 2\mathbf{X}^T\mathbf{y} + \lambda \sum_{j=1}^{k} \text{sign}(\beta_j) = \mathbf{0}$$

gelöst werden. Diese Gleichung besitzt Sprungstellen und ist nur numerisch lösbar. Somit gibt es keine explizite Form für den Schätzer $\hat{\boldsymbol{\beta}}_{Lasso}$, sondern er muss rechnergestützt über spezielle Algorithmen bestimmt werden. Dementsprechend ist auch die Herleitung statistischer Kenngrößen, wie Varianz und Bias, vergleichsweise komplex. Im Vergleich zu dem KQ-Schätzer kann festgestellt werden, dass $\hat{\boldsymbol{\beta}}_{Lasso}$ zwar verzerrt ist, aber eine kleinere Varianz aufweist. (Fahrmeir et al., 2013)

2.3.3. Geometrische Eigenschaften

Im Vergleich zu anderen Shrinkage-Verfahren ist ein Vorteil von Lasso, dass β-Koeffizienten exakt gleich Null geschätzt werden können. Somit wird simultan zur Schätzung eine Variablenselektion durchgeführt. Dieses Verhalten wird im Folgenden für $p = 2$ mit Hilfe der geometrischen Eigenschaften des penalisierten KQ-Schätzer genauer erläutert. Eine Übertragung der Ergebnisse auf den mehrdimensionalen Fall ist ohne Probleme möglich. Es wird weiterhin von standardisierten Kovariablen und einem zentrierten Response ausgegangen, weshalb der Intercept nicht weiter betrachtet wird. (Fahrmeir et al., 2009)

Das KQ-Kriterium aus Gleichung (1) kann unter Vernachlässigung einer Konstante als quadratische Funktion von $\boldsymbol{\beta}$

$$LS(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

umformuliert werden. Als Lösung des Problems $LS(\boldsymbol{\beta}) = c$, für beliebige Konstanten c, resultieren für die Werte von $\boldsymbol{\beta}$ ellipsenförmige Konturlinien. Diese werden in Abbildung 2 dargestellt. Das Zentrum aller Ellipsen bildet der KQ-Schätzer $\hat{\boldsymbol{\beta}}_{KQ}$, das heißt der Schätzer, für den die Summe der quadratischen Abweichungen am kleinsten ist. Ein steigender Ellipsendurchmesser spricht für eine größere Abweichung.

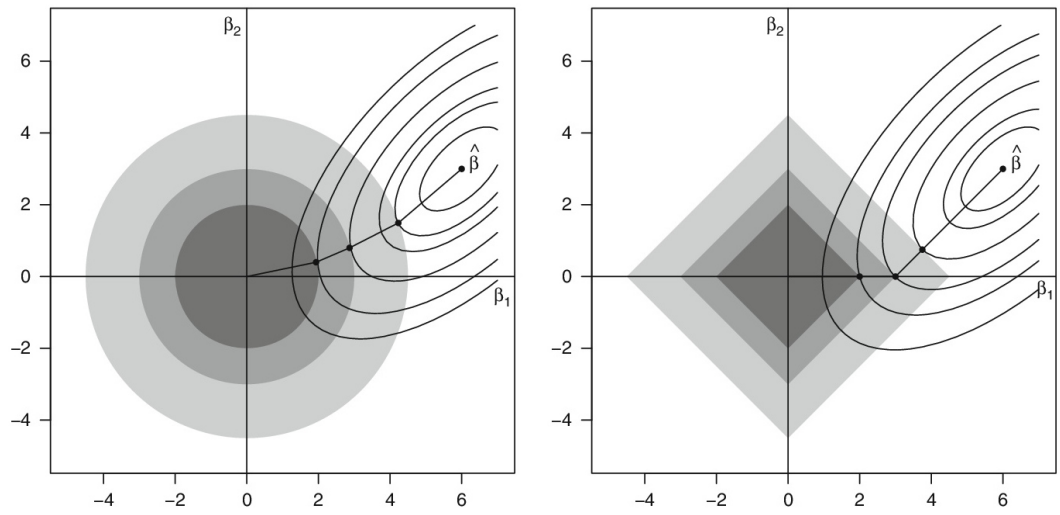

Abbildung 2: Geometrische Visualisierung des penalisierten KQ-Schätzers
links: Ridge-Regression; *rechts:* Lasso-Verfahren
(Fahrmeir et al., 2013)

Die spezifische Form der Ellipsen wird durch die Matrix $\mathbf{X}^T\mathbf{X}$ festgelegt. Zusätzlich befinden sich in Abbildung 2 Schattierungen um den Nullpunkt des Koordinatensystems. Diese stellen verschiedene Stufen der Restriktion für die β-Schätzer dar. Für das Lasso-Verfahren ist die Form der Restriktion $|\beta_1| + |\beta_2| \leq t$ die eines um 90° gedrehten Quadrates mit der Seitenlänge $\sqrt{2}\,t$. Im Vergleich dazu bildet die Restriktion $\beta_1^2 + \beta_2^2 \leq t$ der Ridge-Regression einen Kreis. Die Lösung des penalisierten Minimierungsproblems aus Gleichung (3) ist der Punkt, an dem die kleinstmögliche Konturlinie eine gewählte Restriktion berührt. (Fahrmeir et al., 2013)

Dementsprechend wird im Lasso-Verfahren einer der Koeffizienten auf Null geschätzt, wenn der Berührungspunkt genau an einer Ecke des Quadrates, das heißt auf einer Koordinatenachse liegt. Wird der Wert der Restriktionsparameter t ausreichend klein gewählt, so befinden sich die Berührungspunkte zwangsläufig auf einer oder mehreren Koordinatenachsen. Da, wie in Abbildung 2 ersichtlich, bei der Ridge-Regression keine Ecken als Berührungspunkte zur Verfügung stehen, ist es äußerst unwahrscheinlich, dass Koeffizienten auf Null geschätzt werden. Dies erklärt, weshalb die vorteilhafte Variablenselektion zwar beim Lasso-Verfahren, nicht aber bei der Ridge-Regression erfolgt. In Abbildung 2 werden diese Zusammenhänge anhand eines hypothetischen Werts für den KQ-Schätzer von $\hat{\boldsymbol{\beta}}_{KQ} = (6, 3)^T$ verdeutlicht. (Fahrmeir et al., 2013)

Durch die Standardisierung der Kovariablen befinden sich im zweidimensionalen Fall die Hauptachsen aller Ellipsen im 45°-Winkel zu den Koordinatenachsen. Somit liegen in diesem Fall die Berührungspunkte und folglich die Lösungen des Lasso-Schätzers im selben Quadranten wie die des KQ-Schätzers $\hat{\boldsymbol{\beta}}_{KQ}$. Dementsprechend sind die Vorzeichen der beiden Schätzer gleich. Für den mehrdimensionalen Fall ist diese Eigenschaft jedoch nicht gegeben. (Tibshirani, 1996)

2.3.4. Wahl des Penalisierungsparameters

Je nach Wahl des Penalisierungsparameters t kann die Schätzung der β-Koeffizienten stark variieren. Wird t größer oder gleich der Summe der absoluten KQ-Schätzer $t_0 = \sum_{j=1}^{p} |\hat{\beta}_{j,KQ}|$ gewählt, so ist die Lösung des Minimierungsproblems zwangsläufig $\hat{\boldsymbol{\beta}}_{Lasso} = \hat{\boldsymbol{\beta}}_{KQ}$. Werte von $t < t_0$ führen hingegen zu einer Schrumpfung der Koeffizienten gegen Null. So werden im Fall $t = \frac{t_0}{2}$ zur Schätzung der $\hat{\beta}_{j,Lasso}$ die $\hat{\beta}_{j,KQ}$ durchschnittlich um 50% geschrumpft. Um den Penalisierungsterm möglichst sinnvoll zu wählen, können verschiedene Verfahren angewandt werden. Dazu stellt Tibshirani (1996) drei verschiedene Methoden vor: die Kreuzvalidierung, die generalisierte Kreuzvalidierung und die analytische, unverzerrte Risikoschätzung. Im Rahmen dieser Arbeit wird nur auf die Kreuzvalidierung genauer eingegangen, da diese bei der späteren Analyse verwendet wird. (Hastie et al., 2009)

Zur Kreuzvalidierung wird der Datensatz zufällig in K gleich große Pseudo-Datensätze aufgeteilt. Eine gängige Wahl hierbei ist $K = 10$. Nun wird der erste Pseudo-Datensatz D_1 gewählt. Dieser wird für die anschließende Parameterschätzung ausgeschlossen, das heißt es wird eine Schätzung auf Basis der anderen $K - 1$ Pseudo-Datensätze durchgeführt. Die Schätzung von $\hat{\boldsymbol{\beta}}_{Lasso,-D_1}(\lambda)$, wobei „$-D_1$“ den ausgeschlossenen Pseudo-Datensatz darstellt, kann nun mit Hilfe des Pseudo-Datensatzes D_1 evaluiert werden. Dazu werden die wahren Werte des Response mit den gefitteten verglichen, das heißt der mittlere quadratische Fehler der Schätzung

$$CV(\lambda)_{D_1} = \frac{1}{|D_1|} \sum_{i \in D_1} (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_{Lasso,-D_1}(\lambda))^2$$

bestimmt. Diese Vorgehensweise wird für alle K Pseudo-Datensätze wiederholt, um das zugehörige $CV(\lambda)_{D_k}$ zu ermitteln. Somit kann anschließend der gesamte mittlere

quadratische Fehler (Mean Squared Error)

$$CV(\lambda) = \frac{1}{K}\sum_{k=1}^{K} CV(\lambda)_{D_k},$$

im Folgenden als MSE bezeichnet, berechnet werden. (Fahrmeir et al., 2009)

Zur Wahl eines optimalen λ wird dieses Verfahren für verschiedene Werte von λ wiederholt. Zur Veranschaulichung kann der MSE in Abhängigkeit von λ grafisch dargestellt werden. So zeigt Abbildung 3 anhand simulierter Daten, wie sich der MSE nach 10-facher Kreuzvalidierung in Abhängigkeit von λ verhalten kann. Hierbei kennzeichnen die roten Punkte den MSE, die grauen Markierungen dessen Standardabweichung für das jeweilige λ. Die Standardabweichung wird dazu im Allgemeinen über

$$\sigma = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(CV(\lambda)_{D_k} - CV(\lambda))^2}$$

berechnet. Zur Darstellung wurde aus Gründen der Übersichtlichkeit für λ, wie in Abbildung 1, eine Log-Skala verwendet. Der Wert von λ_{min} befindet sich an der Stelle, an der der MSE minimal wird und wird in der Regel für die Lasso-Schätzung gewählt. Eine weitere Möglichkeit ist es, λ_{se} als Penalisierungsparameter zu nutzen. Dieser bezeichnet den Wert, bei dem sich der MSE noch innerhalb einer Standardabweichung des minimalen Fehlers befindet, aber das Modell am stärksten restringiert wird. Die obere horizontale Achse gibt, wie in Abbildung 1, die Komplexität des Modells je nach Wahl des Penalisierungsparameters λ an. Dabei ist gut erkennbar, wie die Anzahl der ins Modell aufgenommenen Parameter mit steigendem λ sinkt. (Friedman et al., 2010)

2.3.5. Besonderheiten im orthonormalen Fall

Einen besonderen Fall zur Berechnung der β-Koeffizienten stellt der orthonormale dar. Hierbei ist die Designmatrix orthonormal, das heißt $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$. Die Konturlinien in Abbildung 2 wären nun kreisförmig. In diesem Ausnahmefall sind die β-Koeffizienten für Lasso explizit über die Gleichung

$$\hat{\beta}_{j,Lasso}(\lambda) = sign(\hat{\beta}_{j,KQ})\left[|\hat{\beta}_{j,KQ}| - \frac{\lambda}{2}\right]_+,$$

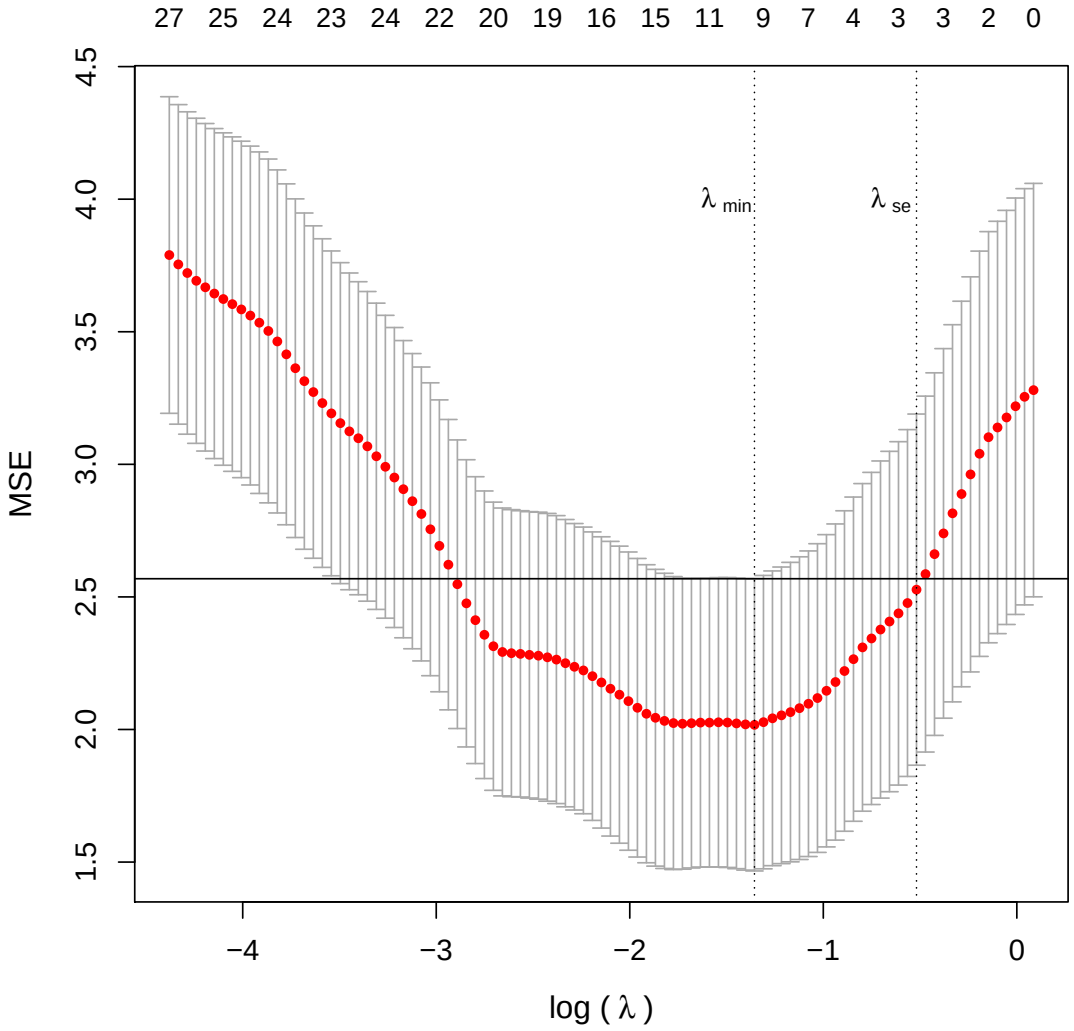

Abbildung 3: Berechnung der MSE für verschiedene Werte von λ
λ_{min} = minimaler MSE
λ_{se} = MSE innerhalb einer Standardabweichung

schätzbar, wobei $[x]_+ = \max(0, x)$. Diese Gleichung verdeutlicht das typische Verhalten des Lasso-Schätzers: Der KQ-Schätzer wird geschrumpft und ab einem bestimmten Wert für λ gleich Null geschätzt. Im orthonormalen Design befindet sich dieser Wert bei $|\hat{\beta}_{j,KQ}| \leq \frac{\lambda}{2}$. Abbildung 4 visualisiert dieses Schrumpfungsverhalten, indem der Lasso-Schätzer als Funktion des KQ-Schätzers beispielhaft für einen Wert von $\lambda = 2$ dargestellt wird. Da der KQ-Schätzer nicht geschrumpft wird, sondern sich selbst abbildet, stellt dieser eine Winkelhalbierende dar. Der Lasso-Schätzer bildet für $|\hat{\beta}_{j,KQ}| > \frac{\lambda}{2}$ eine um $\frac{\lambda}{2} = 1$ verschobene Gerade und ist sonst gleich Null. (Fahrmeir et al., 2013; Härdle und Simar, 2015)

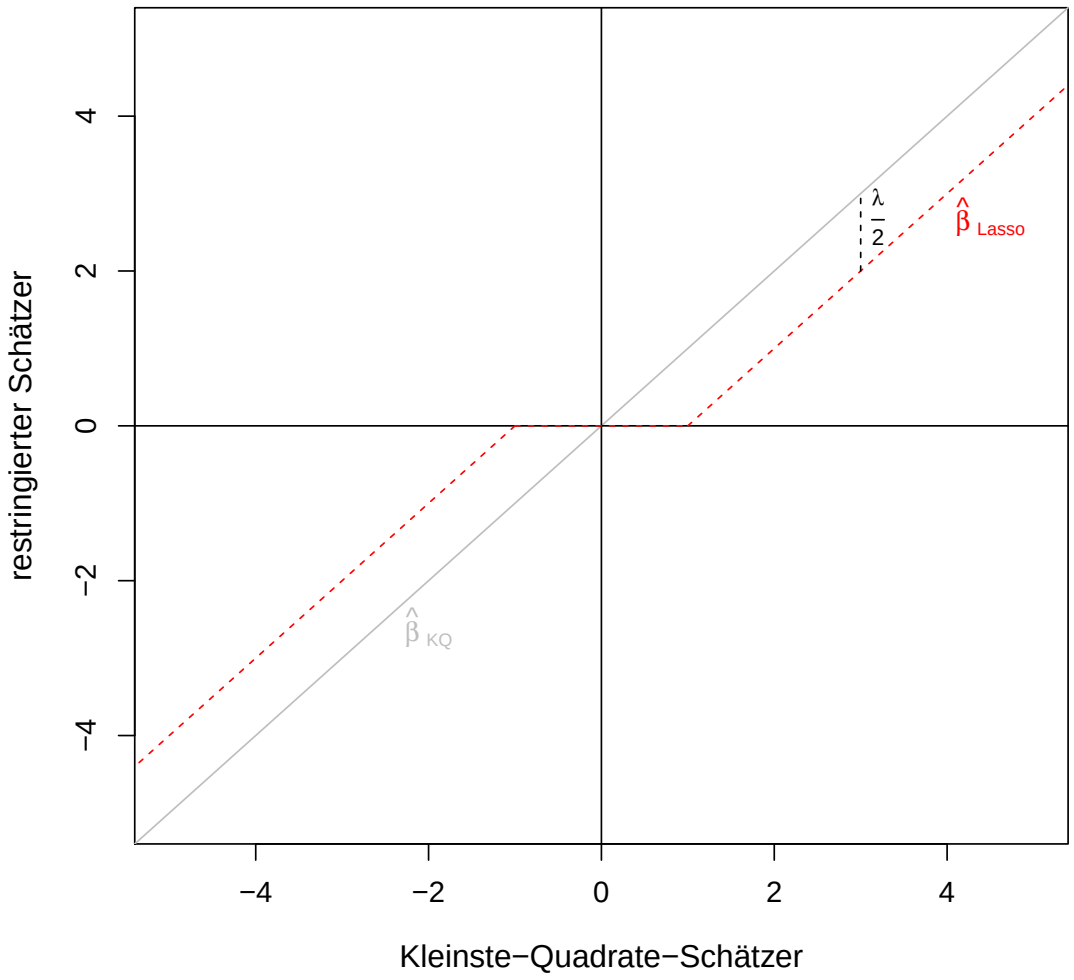

Abbildung 4: Zusammenhang von $\hat{\beta}_{KQ}$ mit $\hat{\beta}_{Lasso}$ im Orthonormalfall

2.3.6. Grenzen des Lasso-Verfahrens

Insgesamt weist das Lasso-Verfahren viele positive Eigenschaften auf. So kann es im Vergleich zum KQ-Schätzer die Varianz der Schätzer reduzieren und die Interpretierbarkeit der Modelle durch Variablenselektion steigern. Trotzdem sind dem Verfahren auch Grenzen gesetzt. Befinden sich mehr Kovariablen als Beobachtungen im Datensatz, das heißt $p > n$, so werden höchstens n davon in das geschätzte Modell aufgenommen. Dies stellt besonders für den Fall $p \gg n$ eine deutliche Einschränkung dar. Dementsprechend befinden sich bei einem Datensatz mit sehr wenigen Beobachtungen gegebenenfalls nicht alle relevanten Variablen im Regressionsmodell. Dies ist eine eher ungünstige Eigenschaft für ein Variablenselektionsverfahren.

Zudem stellen sich hohe paarweise Korrelationen zwischen mehreren Kovariablen als problematisch heraus. Hierbei tragen die einzelnen Variablen kaum zusätzlich zur Erklärung bei, erhöhen aber den Penalisierungsterm. Infolgedessen neigt das Lasso-Verfahren dazu, nur eine beliebige der korrelierten Variablen auszuwählen. Befinden sich im Extremfall zwei identische Kovariablen im Datensatz, so hat das Lasso-Verfahren keine eindeutige Lösung. Somit sollte bei starken Korrelationsstrukturen im Datensatz die Verwendung einer anderen Methode in Betracht gezogen werden. (Zou und Hastie, 2005)

2.3.7. Konsistenzbeschränkungen und Lösungsansätze

Eine weitere Einschränkung des Lasso-Verfahrens ist, dass eine konsistente Modellschätzung nur in bestimmten Fällen möglich ist. Konsistenz bedeutet in diesem Zusammenhang, dass die Wahrscheinlichkeit, die richtigen Variablen ins Modell aufzunehmen, für n gegen unendlich gegen 1 geht. Formal wird Konsistenz folgendermaßen definiert: Sei $\hat{S}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0, j = 1, ..., p\}$, so wird das Lasso-Verfahren genau dann als konsistent bezeichnet, wenn

$$\lim_{n\to\infty} P(\hat{S}(\lambda) = S) \to 1,$$

wobei $S = \{j : \beta_j \neq 0, j = 1, ..., p\}$. Damit diese Konsistenz gegeben ist, muss die sogenannte „Irrepresentable Condition“ nach Zou (2006) und Zhao und Yu (2006) erfüllt sein. Dafür wird die geschätzte Kovarianz-Matrix $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$ definiert. Ohne Beschränkung der Allgemeinheit wird angenommen, dass sich die relevanten Variablen aus den ersten s Variablen, das heißt $S = \{1, ..., s\}$, zusammensetzen. Unter dieser Voraussetzung wird $\hat{\Sigma}$ als

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{1,1} & \hat{\Sigma}_{1,2} \\ \hat{\Sigma}_{2,1} & \hat{\Sigma}_{2,2} \end{pmatrix}$$

dargestellt. Hierbei entspricht $\hat{\Sigma}_{1,1}$ der geschätzten $s \times s$ Kovarianzmatrix der relevanten Variablen, $\hat{\Sigma}_{1,2} = \hat{\Sigma}_{2,1}^T$ der $s \times (p - s)$ Kovarianzmatrix von relevanten und irrelevanten Variablen und $\hat{\Sigma}_{2,2}$ der $(p-s)\times(p-s)$ Kovarianzmatrix der irrelevanten Variablen. Die „Irrepresentable Condition“ ist unter diesen Annahmen als

$$\left\| \hat{\Sigma}_{2,1}\hat{\Sigma}_{1,1}^{-1} \operatorname{sign}(\beta_1, ..., \beta_s) \right\|_\infty \leq \theta, \quad 0 < \theta < 1 \tag{6}$$

definiert, wobei $\|x\|_\infty = \max_j |x_j|$ und $\operatorname{sign}(\beta_1, ..., \beta_p) = (\operatorname{sign}(\beta_1), ..., \operatorname{sign}(\beta_p))^T$. Um im Lasso-Verfahren eine konsistente Modellschätzung erreichen zu können ist

die „Irrepresentable Condition“ hinreichend und im Grunde genommen notwendig. Die Einschränkung „im Grunde genommen“ gilt deshalb, weil die notwendige Bedingung lediglich ein „ ≤ 1“, die hinreichende Bedingung aber ein „ $\leq \theta$“ mit $0 < \theta < 1$ fordert. Eine äquivalente Formulierung zur „Irrepresentable Condition“ ist die sogenannte „Neighbourhood Stability“, welche im Rahmen dieser Arbeit jedoch nicht weiter ausgeführt wird. (Bühlmann und van de Geer, 2011)

Bedingung (6) verdeutlicht, dass die Lasso-Schätzung in vielen Fällen nicht konsistent ist. So kann sich beispielsweise eine starke Korrelation zwischen relevanten und irrelevanten Variablen als problematisch erweisen. Um auch unter weniger strengen Annahmen eine konsistente Modellschätzung zu erhalten, können verschiedene Methoden verwendet werden. Eine mögliches Verfahren hierfür ist die Kombination von Lasso mit Resampling-Methoden. So stellen Meinshausen und Bühlmann (2010) mit ihrem Konzept der „Stability Selection“ ein Verfahren vor, das oftmals zu einer starken Verbesserung der Ergebnisse führt. Hierbei werden zunächst durch Resampling mehrere Pseudo-Datensätze gebildet und anschließend wird auf diese für verschiedene Werte von λ das Lasso-Verfahren angewandt. Somit resultieren für jeden Pseudo-Datensatz in Abhängigkeit von λ verschiedene Modelle mit einer unterschiedlichen Anzahl an Variablen. Dementsprechend kann für jede Variable x_j in Abhängigkeit von λ die empirische Wahrscheinlichkeit, in ein Modell aufgenommen zu werden, $\hat{\pi}_j^{\lambda}$ bestimmt werden. Um mit Hilfe dieser Wahrscheinlichkeiten einen Großteil der relevanten Variablen zu identifizieren, wird ein Grenzwert π festgelegt. Die Menge relevanter Variablen wird anschließend über $\hat{S} = \{j : \max_{\lambda \in \Lambda} \quad \hat{\pi}_j^{\lambda} > \pi\}$ geschätzt. Folglich hängen die Ergebnisse der Stability Selection wenig von der Wahl eines einzelnen Penalisierungsparameters ab. Zudem können deutlich stabilere Ergebnisse erzielt werden als bei einmaliger Anwendung des Lasso-Verfahrens auf den ursprünglichen Datensatz.

Als weiteres stabiles Verfahren wurde von Bach (2008) das sogenannte „Bolasso“ entwickelt. Dieses kombiniert das Lasso-Verfahren mit Bootstrap und liefert somit trotz nicht erfüllter „Irrepresentable Condition“ eine konsistente Variablenselektion. Als Motivation erläutert Bach, dass falls für den Penalisierungsparameter $\lambda_n = \lambda_0 \, n^{-0.5}$ mit $\lambda_0 \in (0, \infty)$ gilt, das Lasso-Verfahren für alle relevanten Variablen die richtigen Vorzeichen mit einer Wahrscheinlichkeit, die für n gegen unendlich gegen 1 geht, liefert. Folglich geht auch die Wahrscheinlichkeit dafür, den Effekt aller relevanten Variablen ungleich Null zu schätzen, gegen 1. Für alle nicht relevanten Variablen

besteht lediglich eine echt positive Wahrscheinlichkeit, in das Modell aufgenommen zu werden. Wird das Lasso-Verfahren für mehrere Datensätze aus der gleichen Verteilung wiederholt, so befindet sich jede relevante Variable mit einer Wahrscheinlichkeit, die gegen 1 geht, und zusätzlich noch weitere, irrelevanten Variablen im Modell. Dementsprechend liegen dem Anwender im Anschluss verschiedene Mengen $\hat{S}_i$ vor, von der jede mit hoher Wahrscheinlichkeit alle relevanten Variablen enthält. Wird nun der Schnitt aus allen Mengen $\hat{S}_i$ gebildet, so befinden sich in der resultierenden Schnittmenge $\hat{S} = \cap \hat{S}_i$ im Idealfall alle relevanten Variablen. Die irrelevanten Variablen, die sich zufällig in einzelnen Mengen $\hat{S}_i$ befinden, sind durch die Bildung des Schnitts darin nicht enthalten. Dies erlaubt, alle relevanten Variablen zu identifizieren. Da in der praktischen Anwendung im Normalfall nur ein Datensatz vorliegt, werden aus dem ursprünglichen Datensatz über Bootstrap Pseudo-Datensätze für die spätere Analyse gebildet. Mit Hilfe dieser Vorgehensweise kann auch ohne Bedingung (6) eine konsistente Modellschätzung erreicht werden (wenn log (Anzahl Pseudo-Datensätze) langsamer gegen unendlich geht als n).

2.4. Möglichkeiten des Resamplings

Die beiden oben erläuterten Beispiele verdeutlichen, dass auch unzureichende Ergebnisse des Lasso-Verfahrens leicht verbessert werden können, indem Resampling-Methoden verwendet werden. Hierbei ist es möglich auf Basis eines Original-Datensatzes beliebig viele Pseudo-Datensätze zu generieren. Dabei kann auf verschiedene Weise vorgegangen werden.

2.4.1. Bootstrap

Das wohl bekannteste Verfahren des Resamplings, welches inzwischen sehr vielfältig verwendet wird, ist das von Efron (1979) vorgestellte Bootstrap-Verfahren. Bei dieser Methode kann die Generierung der Pseudo-Datensätze sowohl über den nonparametrischen als auch über den parametrischen Ansatz erfolgen. Im Rahmen dieser Arbeit wird jedoch nur auf das nonparametrische Bootstrap-Verfahren eingegangen. (Henderson, 2005)

Im nonparametrischen Verfahren werden aus dem Original-Datensatz zufällig n Beobachtungen mit Zurücklegen gezogen, um einen Pseudo-Datensatz zu erhalten. Somit werden manche Beobachtungen mehrmals, andere überhaupt nicht in den Pseudo-Datensatz gezogen. Insgesamt können $\binom{2n-1}{n}$ verschiedene Pseudo-Datensätze

resultieren. Diese Vorgehensweise wird b - mal wiederholt und somit werden b Pseudo-Datensätze generiert. Die darauffolgenden, statistischen Analysen werden anschließend auf Basis aller b Pseudo-Datensätze durchgeführt. (Henderson, 2005)

Die Vorzüge des Bootstrap-Verfahrens im Allgemeinen können folgendermaßen erklärt werden: Da die Verteilung der gesamten Population im Normalfall unbekannt ist, wird die der zufälligen Stichprobe aus der Gesamtpopulation als Hinweis auf die wahre Verteilung gesehen. Werden nun aus dieser Stichprobe weitere Bootstrap-Stichproben gezogen, so kann die eigentliche Verteilung mit Hilfe dieser approximiert werden. Dementsprechend bringen Sprent und Smeeton (2007) zufolge Bootstrap-Verfahren einen besonders großen Mehrgewinn, wenn wenig über die wahre Verteilung in der Gesamtpopulation bekannt ist. (Henderson, 2005)

In der Praxis wird Bootstrap häufig dazu genutzt, Standardfehler, Konfidenzintervalle oder auch den Bias eines Schätzers zu quantifizieren. Besonders für die Analyse kleiner Datensätze kann dies sehr hilfreich sein. Als Vorteil sehen Efron und Tibshirani (1998), dass bei Verwendung des Bootstrap-Verfahrens für die statistische Analyse notwendige Annahmen reduziert werden können. Dies gilt, wie in Kapitel 2.3.7 erläutert, auch für das Lasso-Verfahren. (Henderson, 2005)

Eine Modifizierung des Bootstrap-Verfahrens ist der m-out-of-n Bootstrap. Hierbei werden nicht wie beim ursprünglichen Bootstrap n sondern $m < n$ Beobachtungen aus dem Original-Datensatz mit Zurücklegen gezogen. Somit kann das ursprüngliche Bootstrap-Verfahren im Fall von Inkonsistenz oftmals verbessert werden. (Davison et al., 2003)

2.4.2. Subsampling

Aktuellen Studien zufolge neigen Modellselektionsverfahren angewandt auf Bootstrap-Stichproben dazu, tendenziell zu viele Variablen auszuwählen (Janitza et al., in Druck; Binder und Schumacher, 2008). Ein alternatives, dem m-out-of-n Bootstrap angelehntes Verfahren ist das sogenannte Subsampling. Auch dieses weist in Fällen, in denen das ursprüngliche Bootstrap-Verfahren keine zufriedenstellenden Ergebnisse mehr liefert, asymptotische Konsistenz auf. Prinzipiell wird beim Subsampling wie im Bootstrap-Verfahren vorgegangen, diesmal werden jedoch die m

Beobachtungen aus dem Original-Datensatz ohne Zurücklegen gezogen. (Davison et al., 2003)

Für die Wahl von m gibt es verschiedene Möglichkeiten, die das Ergebnis stark beeinflussen können. Wird m zu groß gewählt, sind sich die einzelnen Stichproben sehr ähnlich. Zwar könnte somit der Lasso-Schätzer für die einzelnen Stichproben bessere Ergebnisse erzielen, aber eine potentielle Instabilität des Selektionsverfahrens durch Ausreißer würde durch diese Ähnlichkeit nicht erkannt werden. Wird m jedoch zu klein gewählt, so liefert die Stichprobe zu wenig Information und relevante Variablen werden möglicherweise nicht als solche identifiziert. Zum Vergleich von Bootstrap und Subsampling wird oftmals der Wert $m = 0.632\,n$ verwendet. Dieser setzt sich folgendermaßen zusammen:
Wird aus einem Datensatz der Größe n eine Bootstrap-Stichprobe der Größe n gezogen, so kann die Wahrscheinlichkeit, dass die Beobachtung i mindestens einmal in der Stichprobe B enthalten ist, über

$$P(i \in B) = 1 - (1 - \frac{1}{n})^n$$

berechnet werden. Geht der Stichprobenumfang n gegen unendlich, so nimmt der Grenzwert dieser Wahrscheinlichkeit den Wert

$$\lim_{n\to\infty} 1 - (1 - \frac{1}{n})^n = 1 - e^{-1} \approx 0.632$$

an. Somit beträgt für eine Bootstrap-Stichprobe der Größe n die erwartete Anzahl verschiedener Beobachtungen $0.632\,n$. Damit für spätere Vergleiche durchschnittlich die gleiche Anzahl verschiedener Beobachtungen in Bootstrap- und Subsampling-Stichproben enthalten sind, wird für die weitere Analyse der Wert $m = 0.632\,n$ verwendet. (Davison et al., 2003; De Bin et al., in Druck)

3. Anwendung auf simulierte Daten

Da bekannt ist, dass Variablenselektionsverfahren für Bootstrap-Stichproben oftmals sehr komplexe Modelle liefern, gilt es, dies auch für das Lasso-Verfahren zu untersuchen. Zudem werden die resultierenden Modelle bezüglich weiterer Gütekriterien geprüft. Die dabei erzielten Ergebnisse werden sowohl mit dem alternativen Subsampling verglichen als auch denen auf Basis der Original-Datensätze gegenübergestellt. Dementsprechend können mögliche Vorzüge und Nachteile der Kombination aus Lasso-Verfahren und verschiedenen Resampling-Methoden herausgearbeitet werden. Diese sollten für die Entwicklung neuer, kombinierter Methoden wie Bolasso berücksichtigt werden.

3.1. Datensimulation

Die statistischen Auswertungen zum Vergleich der verschiedenen Resampling-Methoden werden anhand eines simulierten Datensatzes durchgeführt. Somit ist bekannt, welche der potenziellen Prädiktoren tatsächlich einen Einfluss auf den Response haben. Dies ermöglicht nicht nur einen Vergleich zwischen den verschiedenen Methoden, sondern auch eine Gegenüberstellung mit dem wahren Modell. Um die spätere Auswertung zu erleichtern, werden im Rahmen dieser Arbeit Daten ohne Korrelationsstruktur simuliert. Dafür wird die statistische Software R (Version 3.0.2) verwendet. Für die Simulation wird folgendermaßen vorgegangen:
Der zu analysierende Datensatz soll $n = 100$ Beobachtungen und $p = 200$ Kovariablen beinhalten. Zu diesem Zweck werden unabhängig voneinander, zufällig 20 000 Werte aus der Standardnormalverteilung $\mathcal{N}(0, 1)$ gezogen. Mit diesen wird die Designmatrix X befüllt. Der Response wird anschließend durch die Gleichung

$$y_i = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + ... + \beta_{200} \, x_{i200} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$$

generiert. Dazu werden die Werte von ϵ_i ebenso zufällig aus der Standardnormalverteilung $\mathcal{N}(0, 1)$ gezogen. Die Werte für β_j werden so festgelegt, dass die ersten 30 Variablen einen Effekt auf den metrischen Response haben, die übrigen 170 keinen. Diese Effekte sind unterschiedlich stark ausgeprägt, wobei jeweils fünf Variablen den

gleichen Einfluss auf den Response haben. Demnach haben die β_j folgende Werte:

$$
\begin{aligned}
\beta_0 &= 0 \\
\beta_1 = \beta_2 = \ldots = \beta_5 &= 0.25 \\
\beta_6 = \beta_7 = \ldots = \beta_{10} &= 0.5 \\
\beta_{11} = \beta_{12} = \ldots = \beta_{15} &= 0.75 \\
\beta_{16} = \beta_{17} = \ldots = \beta_{20} &= 1.0 \\
\beta_{21} = \beta_{22} = \ldots = \beta_{25} &= 1.25 \\
\beta_{26} = \beta_{27} = \ldots = \beta_{30} &= 1.5 \\
\beta_{31} = \beta_{32} = \ldots = \beta_{200} &= 0
\end{aligned}
$$

Da sich in einer einzelnen Designmatrix zufällige Strukturen befinden könnten, die die spätere Analyse möglicherweise beeinträchtigen, wird das Verfahren 1 000 mal wiederholt. Demnach werden insgesamt 1 000 Datensätze nach obiger Vorgehensweise generiert. Anschließend werden aus jedem der 1 000 Datensätze Pseudo-Datensätze, mit Hilfe der in Kapitel 2.4 erläuterten Verfahren, erstellt. Hierbei wird $b = 1$ gesetzt, das heißt für jeden Original-Datensatz eine Bootstrap-Stichprobe der Größe n, eine Bootstrap-Stichprobe der Größe m und eine Subsampling-Stichprobe der Größe m gezogen. Diese werden zur vereinfachten Darstellung im weiteren Verlauf als „Bootstrap(n)“, „Bootstrap(m)“ und „Subsample(m)“ bezeichnet, die Original-Datensätze als „Daten“. Um die Ergebnisse von Subsampling und Bootstrap basierend auf gleicher Stichprobengröße vergleichen zu können, wird für beide Verfahren $m = 0.632\,n$ gesetzt. Dementsprechend liegen zur Analyse folgende Datensätze vor:

- 1 000 Original-Datensätze mit jeweils 100 Beobachtungen
- 1 000 Pseudo-Datensätze mit jeweils 100 Beobachtungen, generiert durch Bootstrap
- 1 000 Pseudo-Datensätze mit jeweils 63 Beobachtungen, generiert durch m-out-of-n Bootstrap
- 1 000 Pseudo-Datensätze mit jeweils 63 Beobachtungen, generiert durch Subsampling

Für jeden dieser 4 000 Datensätze wird anschließend das Lasso-Verfahren durchgeführt. Dabei wird, wie in Kapitel 2.3.4 erläutert, das entsprechende λ jeweils so gewählt, dass der MSE minimal wird.

3.2. Auswertung

Da sich die (Pseudo-)Datensätze in der Regel alle unterscheiden, werden für jeden Datensatz nach Anwendung des Lasso-Verfahrens verschiedene Ergebnisse erwartet. Es liegt jedoch die Vermutung nahe, dass die relevanten Variablen in den meisten Modellen enthalten sind, während die irrelevanten je nach Pseudo-Stichprobe unterschiedlich oft aufgenommen werden. Um die verschiedenen Resampling-Methoden vergleichen und bewerten zu können, werden unterschiedliche Kriterien untersucht. Im Fokus der Analyse stehen hierbei der detaillierte Vergleich der Prädiktionsgüte, der Inclusion Frequencies und der Modellkomplexität für Bootstrap(n), Bootstrap(m) und Subsample(m). Zur Einordnung der Güte der Charakteristika für die berechneten Modelle erfolgt zusätzlich eine Gegenüberstellung mit den Ergebnissen für 1 000 Original-Datensätze. Zudem sind durch die Simulation der Daten die wahren Modelle bekannt, das heißt auch ein Vergleich mit diesen ist möglich. Alle im Folgenden ausgewerteten Modelle wurden mit Hilfe des R-Pakets „glmnet“ berechnet. (Friedman et al., 2010)

3.2.1. Modellkomplexität

Zunächst wird die Komplexität der resultierenden Modelle betrachtet. Diese beschreibt die Anzahl an Variablen die nach Anwendung des Lasso-Verfahrens noch im Modell enthalten sind. Ein möglichst gutes Modell sollte sparsam sein, das heißt so wenig Variablen wie möglich beinhalten. Somit wird die Gefahr eines Overfittings vermieden und die Interpretierbarkeit des Modells steigt. Trotzdem sollten keine relevanten Variablen unnötig aus dem Modell entfernt werden, da dies zu einem Underfitting führen könnte. Bei der weiteren Betrachtung der Modellkomplexität muss berücksichtigt werden, dass auch hohe Werte noch keine Aussage darüber geben, wie viele relevante Variablen in das Modell aufgenommen wurden. (De Bin et al., in Druck)

Abbildung 5 stellt die Modellkomplexität in Form von Boxplots dar. Hierbei beschreibt die y-Achse, wie viele Variablen nach Anwendung des Lasso-Verfahrens in den Modellen enthalten sind. Jeder der vier Boxplots verkörpert die auftretenden Modellkomplexitäten innerhalb eines Resampling-Verfahrens und wird somit auf Basis von 1 000 berechneten Modellen erstellt. Die fetten, schwarzen Linien innerhalb der einzelnen Boxen kennzeichnen den Median der jeweiligen Gruppe, die Boxen selbst das 25%- und das 75%-Quantil. Dementsprechend stellen sie den Interquartilsabstand dar. An den Boxen befinden sich sogenannte Whiskers. Diese markieren

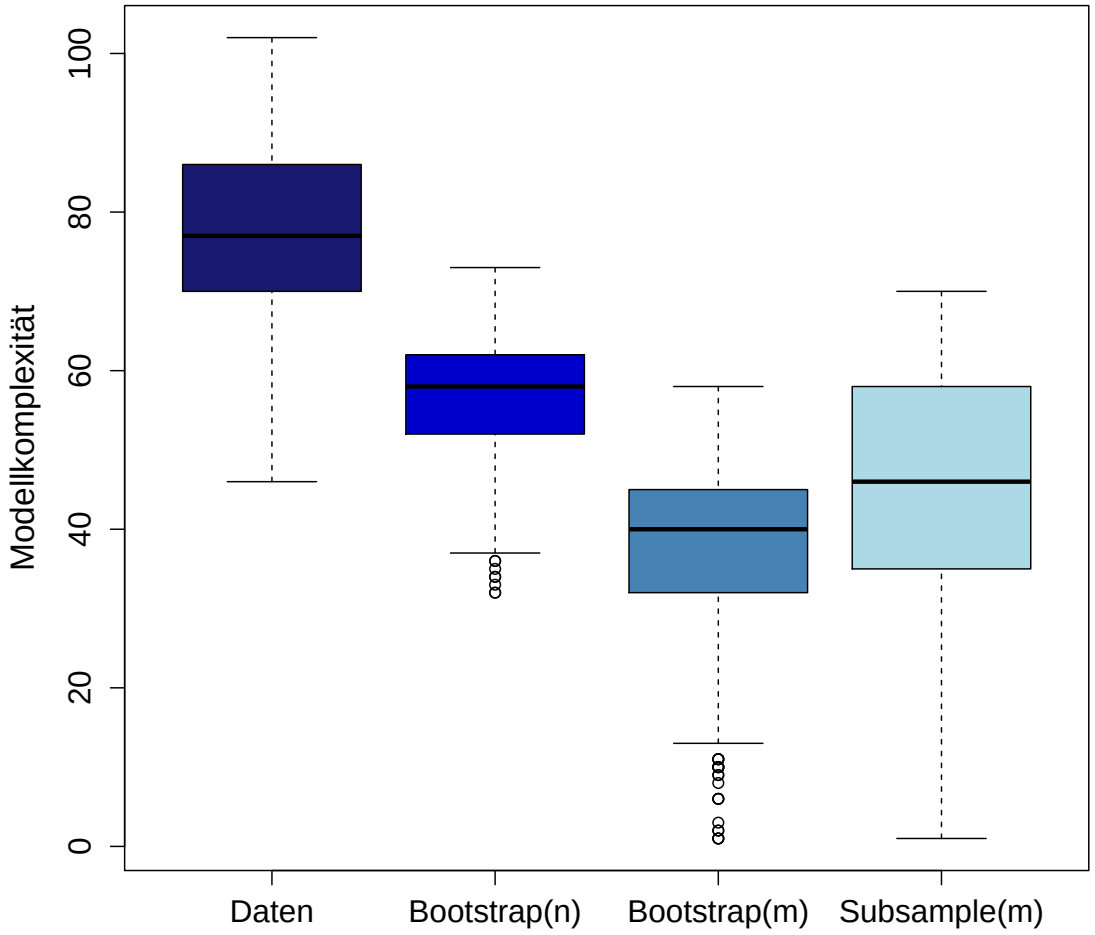

Abbildung 5: Vergleich der Modellkomplexität für Original-Datensätze und verschiedene Resampling-Methoden

die Werte der Modellkomplexität, die nicht weiter als der 1.5 - fache Interquartilsabstands von den Rändern der Box entfernt sind. Alle Werte die außerhalb der Whiskers liegen, werden als Ausreißer bezeichnet.

Vergleicht man die Mediane der einzelnen Methoden, so ist der für die Original-Datensätze am größten. Dies bedeutet, dass das Lasso-Verfahren, angewandt auf die Original-Datensätze, die komplexesten Modelle liefert. Der Median liegt in diesem Fall bei 77. Die maximale Modellkomplexität beträgt 102, die minimale 46. Da in der Literatur für das Lasso-Verfahren als maximale Anzahl an aufgenommen Variablen $\min(p{=}200, n{=}100)$ angegeben ist (siehe Kapitel 2.3.6), sollte diese höchstens 100 betragen. Der Wert von 102 ist verwunderlich. Insgesamt sind die aus dem Original-Datensatz resultierenden Modelle deutlich zu komplex. Die wahre Anzahl

an relevanten Variablen beträgt lediglich 30.

Weniger komplexe Modelle liefert das Lasso-Verfahren für Bootstrap(n). Der Median der Modellkomplexität liegt hier mit 58 deutlich niedriger als bei den Original-Datensätzen. Sogar die maximale Modellkomplexität von 73 liegt unter dem Median für die Original-Datensätze. Das kleinste resultierende Modell weist 32 Variablen auf. Insgesamt liefert das Lasso-Verfahren also auch für Bootstrap(n) zu komplexe Ergebnisse.

Wird das Lasso-Verfahren auf Bootstrap(m) angewandt, so erhält man die sparsamsten Modelle. Der Median der Modellkomplexität beträgt lediglich 40. Maximal werden 58 Variablen vom Lasso-Verfahren ausgewählt. Dies legt die Schlussfolgerung nahe, dass Bootstrap(m) bezüglich der Modellkomplexität für das Lasso-Verfahren gut geeignet ist. Hierbei müssen jedoch auch die minimalen Werte der Modellkomplexität betrachtet werden. Das kleinste resultierende Modell beinhaltet nur noch eine Variable. Somit wäre es deutlich zu sparsam. Auch wenn es sich hierbei um Ausreißer handelt, ist zu berücksichtigen, dass insgesamt 189 der 1 000 Modelle weniger als 30 Variablen beinhalten. Dementsprechend befinden sich in mindestens 18.9% der Modelle nicht alle relevanten Variablen. Hierbei wird von „mindestens“ gesprochen, da auch in Modellen mit 30 oder mehr Variablen, nicht zwangsläufig alle relevanten enthalten sind.

Die größten Unterschiede der Modellkomplexität innerhalb einer Resampling-Methode existieren für Subsample(m). Hierbei reicht die Anzahl aufgenommener Variablen von 1 bis 70. Wie bei Bootstrap(m) werden für viele Modelle (mindestens 180 von 1 000) nicht alle relevanten Variablen aufgenommen. Der Median liegt mit 46 über dem von Bootstrap(m). Insgesamt erweist sich für Subsample(m) eine Einordnung im Vergleich zu den anderen Methoden durch die große Spanne der Ergebnisse als schwierig. Bei Betrachtung des unteren und oberen Quartils wären jedoch Subsample(m) und Bootstrap(m) den anderen Verfahren bezüglich ihrer Komplexität vorzuziehen, da diese tendenziell nur wenig mehr als die relevante Anzahl an Variablen aufnehmen. Dabei ist jedoch zu beachten, dass, obwohl in manchen Modellen die Modellkomplexität der des wahren Modells entsprach, in keinem Fall das wahre Modell identifiziert werden konnte.

3.2.2. Inclusion Frequencies

Um nicht nur die Anzahl aufgenommener Variablen, sondern auch die Aufnahme relevanter Variablen zu quantifizieren, werden zusätzlich Inclusion Frequencies betrachtet. Diese geben für jede Variable den Anteil der Modelle an, in denen ihr Effekt ungleich Null geschätzt wird. Somit stellen Inclusion Frequencies einen Indikator für die geschätzte Relevanz einzelner Variablen dar. Dabei ist zu beachten, dass die hier betrachteten Inclusion Frequencies nicht wie in der Literatur üblich anhand von 1 000 Pseudo-Datensätzen aus einem Original-Datensatz ermittelt werden. Stattdessen erfolgt die Berechnung basierend auf 1 000 Pseudo-Datensätzen aus 1 000 verschiedenen Original-Datensätzen. Nach Anwendung des Lasso-Verfahrens wird erwartet, dass die relevanten Variablen fast immer im resultierenden Modell enthalten sind. Irrelevanten Variablen hingegen sollte nur in wenigen Fällen ein Effekt zugesprochen werden. Dementsprechend haben im Idealfall relevante Variablen eine Inclusion Frequency von 1, irrelevante Variablen eine von 0. Besonders die Aufnahme von Variablen mit schwachem Effekt ist in der praktischen Anwendung jedoch eher zufällig. So liegt unter Berücksichtigung der zuvor beschriebenen Modellkomplexität die Vermutung nahe, dass oftmals irrelevante Variablen aufgenommen werden. Trotzdem kann davon ausgegangen werden, dass Variablen mit starkem Effekt eine Inclusion Frequency nahe 1 und Variablen ohne Effekte eine Inclusion Frequency nahe 0 aufweisen. Für Variablen mit schwächeren Effekten wird eine mittlere Inclusion Frequency zwischen 0 und 1 erwartet. Durch die Kenntnis des wahren Modells können nicht nur die beobachteten Inclusion Frequencies der verschiedenen Resampling-Methoden miteinander verglichen, sondern diese auch den erwarteten Inclusion Frequencies gegenübergestellt werden. (De Bin et al., in Druck)

Insgesamt liegen, der Anzahl an Variablen entsprechend, 200 Inclusion Frequencies für jede Resampling-Methode zur Analyse vor. Um diese grafisch übersichtlicher darstellen zu können, werden die Inclusion Frequencies von Variablen mit gleichem Effekt zusammengefasst. So werden beispielsweise die Inclusion Frequencies für Variablen mit Effekt 0.25 als eine Gruppe betrachtet. Dementsprechend kann eine Reduktion auf sieben verschiedene Gruppen für jede Resampling-Methode erreicht werden. Abbildung 6 zeigt die mittleren Inclusion Frequencies für jeweils eine Gruppe von Variablen mit gleichem Effekt. Zum direkten Vergleich wurden die mittleren Inclusion Frequencies der verschiedenen Resampling-Methoden jeweils für den gleichen Effekt nebeneinander zu einem Block angeordnet. Diese Blöcke wurden so sortiert, dass der Effekt der Variablen von links nach rechts schwächer wird. Erwar-

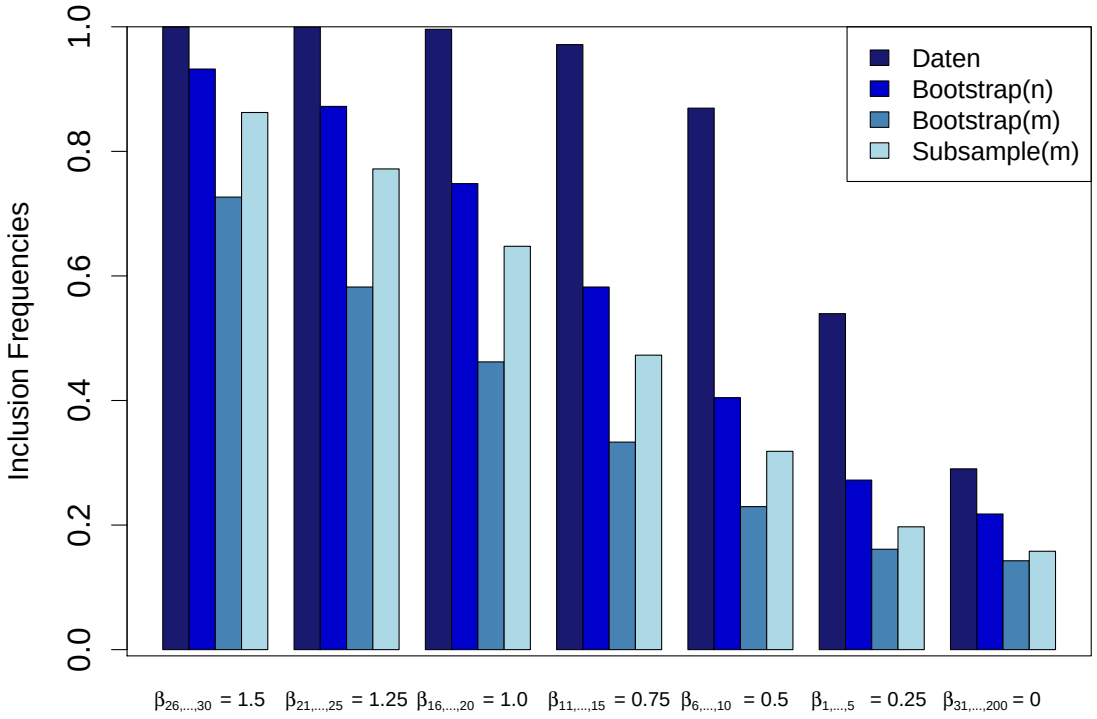

Abbildung 6: Vergleich der mittleren Inclusion Frequencies anhand verschieden generierter Datensätze für unterschiedlich starke Effekte

tungsgemäß weisen alle Methoden für den größten Effekt auch die höchsten Inclusion Frequencies auf. Dabei werden für die Original-Datensätze alle Variablen mit einem Effekt von 1.5 immer in das resultierende Modell aufgenommen. Doch auch für die Resampling-Methoden können diese Variablen in den meisten Fällen als relevant identifiziert werden. Nur Bootstrap(m) fällt mit einer mittleren Inclusion Frequency von 0.73 deutlich im Vergleich zu den anderen ab. Trotzdem ist für alle Methoden ein stufenförmiger Verlauf erkennbar. Dieser verdeutlicht, dass bei allen Methoden Variablen mit höherem Effekt öfter aufgenommen werden, als Variablen mit niedrigem oder keinem Effekt. Dabei ist für alle Effekte eine klare Abstufung zwischen den unterschiedlichen Methoden erkennbar. Die mittleren Inclusion Frequencies für die Original-Datensätze sind für alle Effekte am höchsten. Darauf folgen Bootstrap(n) und Subsample(m). Für Bootstrap(m) resultieren in jedem Fall die kleinsten mittleren Inclusion Frequencies. Somit werden für Bootstrap(n) relevante Variablen häufiger identifiziert als für Subsample(m) und Bootstrap(m), dafür aber auch irrelevante Variablen vermehrt in die Modelle aufgenommen. Auch für die Original-Datensätze identifiziert das Lasso-Verfahren zwar sehr gut die relevanten Variablen, nimmt jedoch eine irrelevante Variable im Schnitt in jedes dritte Modell auf. Da im Rahmen dieser Auswertung Mediane und Mittelwerte der Inclusion Frequencies zu nahezu

identischen Ergebnissen führten, wurde in obiger Beschreibung nur auf den Mittelwert eingegangen. Die entsprechende Darstellung der Mediane befindet sich im Anhang (Abbildung 9).

Um genauer zu untersuchen, welche Resampling-Methoden sich dazu eigenen, hinsichtlich relevanter und irrelevanter Variablen zu differenzieren, werden die Inclusion Frequencies einzelner Variablen ähnlich wie in De Bin et al. (in Druck) miteinander verglichen. Im Idealfall sollten die Inclusion Frequencies einer Variable mit Effekt immer größer sein als die einer Variable ohne Effekt. Auf Basis der hier verwendeten Datensätze ist dies für Variablen mit einem stärkerem Effekt als 0.25 für alle Resampling-Methoden gegeben. Für Variablen mit dem Effekt 0.25 hingegen ist eine eindeutige Abgrenzung zu den irrelevanten Variablen anhand der resultierenden Inclusion Frequencies nicht in allen Fällen möglich. Dementsprechend werden zur weiteren Analyse ausschließlich die Inclusion Frequencies der Variablen mit einem Effekt von 0.25 mit denen der irrelevanten Variablen verglichen. Hierbei ist zu beachten, dass Variablen mit gleichem Effekt nicht mehr wie zuvor gruppenweise sondern nun einzeln betrachtet werden. Um die verschiedenen Resampling-Methoden gegenüberzustellen, werden jeweils die Inclusion Frequencies aller Variablen mit Effekt 0.25 mit denen aller Variablen ohne Effekt paarweise verglichen. Dies bedeutet es werden insgesamt $5 * 170 = 850$ Paare untersucht. Anschließend wird die relative Häufigkeit der Paare berechnet, für die eine Variable mit Effekt 0.25 eine größere Inclusion Frequency als eine Variable ohne Effekt hat. Beträgt die relative Häufigkeit 1, so kann problemlos zwischen relevanten und irrelevanten Variablen abgegrenzt werden. Nimmt sie einen Wert um 0.5 an, so kann davon ausgegangen werden, dass das Lasso-Verfahren nicht zwischen Variablen mit Effekt 0.25 und irrelevanten Variablen unterscheiden konnte. (De Bin et al., in Druck)

Die resultierenden relativen Häufigkeiten werden in Abbildung 7 in Form von Balkendiagrammen dargestellt. Dabei ist darauf zu achten, dass zur detaillierteren Ansicht die y-Achse erst bei 0.9 beginnt. Die Grafik verdeutlicht, dass ausschließlich für die Original-Datensätze die Inclusion Frequencies der relevanten Variablen in allen Fällen größer sind, als die der irrelevanten. Doch auch auf Basis von Bootstrap(n) ist eine Abgrenzung zwischen relevanten und irrelevanten Variablen gut möglich. Lediglich eine der Variablen mit Effekt 0.25 hat eine kleinere Inclusion Frequency als eine der irrelevanten Variablen. Die relative Häufigkeit beträgt hier 0.999. Etwas schlechter in der Abgrenzung zeigt sich Subsample(n). Hierbei beträgt

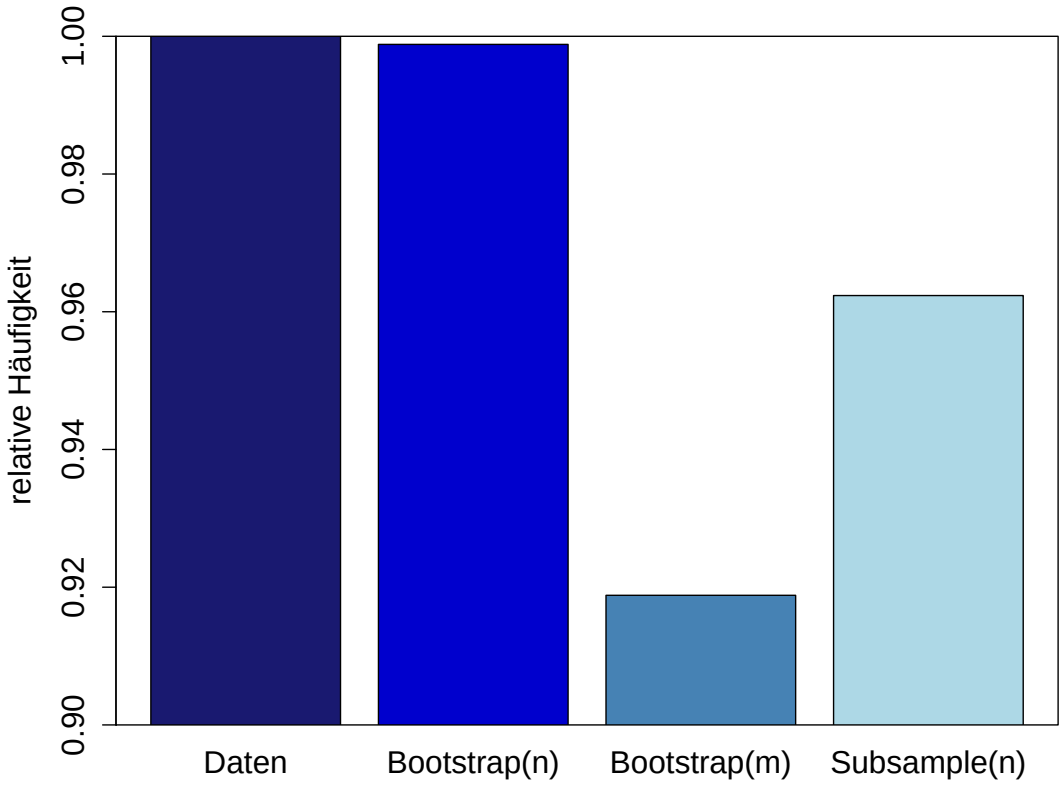

Abbildung 7: Anteil der Variablen mit Effekt 0.25 mit größerer Inclusion Frequency als Variablen ohne Effekt

die relative Häufigkeit 0.962. Insgesamt war somit in 32 Fällen die Inclusion Frequency einer irrelevanten Variable größer als die einer relevanten. Im Vergleich zu den Original-Datensätzen fällt Bootstrap(m) am stärksten ab. Hierbei beträgt die relative Häufigkeit nur noch 0.919, das heißt in 69 der 850 paarweisen Vergleiche war die Inclusion Frequency einer irrelevanten Variable größer als die einer Variable mit Effekt. Somit lässt sich schlussfolgern, dass das Lasso-Verfahren auf Basis von Bootstrap(m) am schlechtesten zwischen relevanten Variablen mit niedrigem Effekt und irrelevanten Variablen unterscheiden kann. Unter den untersuchten Resampling-Methoden kann für Bootstrap(n) die beste Abgrenzung erzielt werden. Hierbei sind die Ergebnisse mit denen der Original-Datensätze vergleichbar.

3.2.3. Prädiktionsgüte

Während die obigen Analysen dazu dienen, die verschiedenen Resampling-Methoden bezüglich der aufgenommenen Variablen zu vergleichen, werden in diesem Kapitel

deren Auswirkungen auf die prädiktiven Eigenschaften des Lasso-Verfahrens untersucht. Dementsprechend wird geprüft, für welche Resampling-Methode die jeweils geschätzten Modelle die besten Vorhersagen für neue Datensätze liefern. De Bin et al. (in Druck) zufolge erlaubt die Analyse der Prädiktionsgüte es nicht nur, Rückschlüsse auf die Leistung der Prädiktion zu ziehen, sondern auch indirekt auf die Eignung der ausgewählten Variablen. Als Maß für die Prädiktionsgüte wird hier die quadratische Abweichung zwischen wahren und gefitteten Werten betrachtet. Dazu wird in der Regel, wie in Kapitel 2.3.4 erläutert, eine Kreuzvalidierung durchgeführt und somit der mittlere, quadratische Prädiktionsfehler (MSE) bestimmt. Da in diesem speziellen Fall das wahre Modell bekannt ist, werden hier alle Modelle auf Basis der vollständigen Pseudo-Datensätze berechnet, das heißt der komplette Pseudo-Datensatz als Trainings-Datensatz verwendet. Anschließend wird für jeden Pseudo-Datensatz zur Evaluation ein eigener Test-Datensatz mit 100 000 Beobachtungen generiert. Dies garantiert, dass die berechneten Prädiktionsfehler nicht alle von dem selben Test-Datensatz abhängen, wodurch das Ergebnis verfälscht werden könnte. Zur Berechnung des Prädiktionsfehlers wird für jeden Test-Datensatz der Response mit Hilfe des für die Trainings-Daten gefitteten Modells geschätzt. Dieser wird mit den zuvor simulierten, wahren Werten verglichen. Dementsprechend wird der MSE bei n Beobachtungen über

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

gebildet. Durch die quadratische Form werden große Abweichungen noch weiter vergrößert, während sehr kleine Fehler (<1) verkleinert werden. Dementsprechend werden bei Verwendung des quadrierten Fehlers große Abweichungen stärker bestraft. Eine weitere Möglichkeit wäre die Verwendung des absoluten Prädiktionsfehlers. Dadurch würden alle Abweichungen gleich stark in den mittleren Fehler eingehen. Da im Rahmen dieser Auswertung die quadratischen und die absoluten Fehler zu vergleichbaren Ergebnissen führten, wird im Folgenden nur auf den MSE eingegangen. Die Darstellung der absoluten Prädiktionsfehler befindet sich im Anhang (Abbildung 10). (Fahrmeir et al., 2013)

Abbildung 8 stellt die berechneten MSE für die verschiedenen Resampling-Methoden in Form von Boxplots dar. Jeder Boxplot wird dementsprechend durch die MSE von 1 000 Test-Datensätzen gebildet. Der obigen Definition entsprechend eignet sich ein geschätztes Modell umso besser zur Prädiktion, desto niedriger der MSE ist. Die

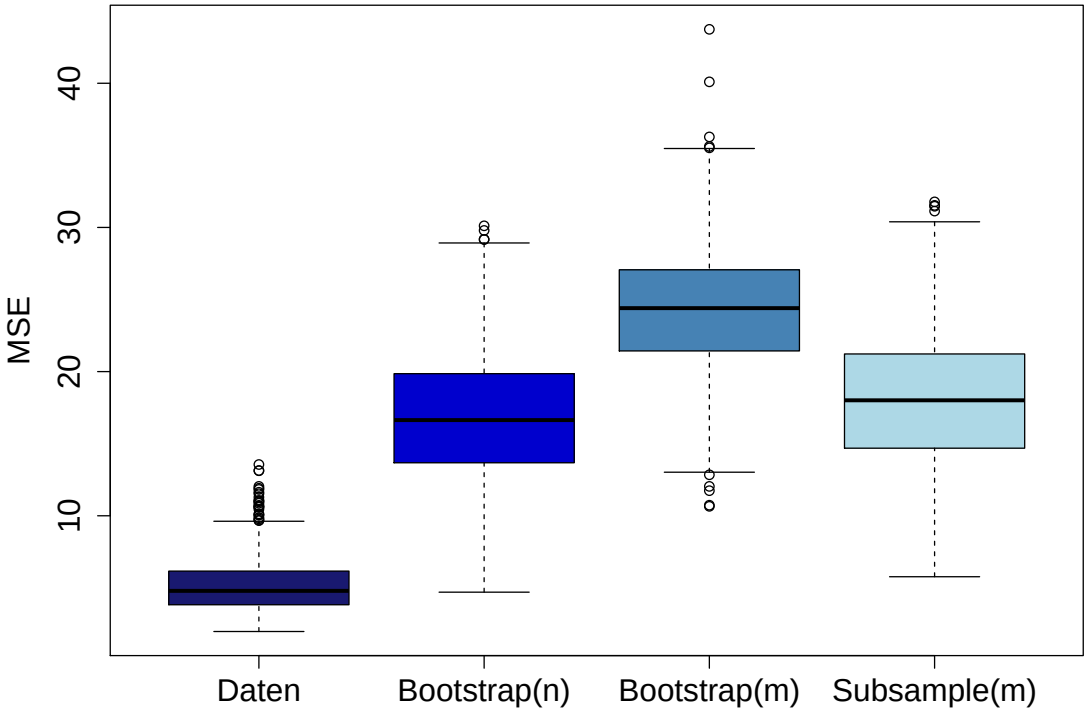

Abbildung 8: Vergleich des MSE für Original-Datensätze und verschiedene Resampling-Methoden

Modelle auf Basis der Original-Datensätze liefern deutlich erkennbar die kleinsten MSE. Da Variablen mit großem Effekt in alle Modelle und Variablen mit mittlerem Effekt in fast alle Modelle aufgenommen wurden (vergleiche Abbildung 6), liefern die Modelle sehr gute Ergebnisse für die Prädiktion. Auch die vielen irrelevanten Variablen, die fälschlicherweise im Modell enthalten sind (vergleiche Abbildung 5), scheinen keine Überanpassung an den jeweiligen Test-Datensatz zur Folge zu haben. Für Bootstrap(n) steigen die MSE deutlich an. Zudem streuen die Werte stärker als für die Original-Datensätze. Während für Bootstrap(n) Variablen mit sehr hohem Effekt in fast alle Modelle aufgenommen werden, sinken die Inclusion Frequencies für Variablen mit mittleren Effekten stark ab (vergleiche Abbildung 6). Somit werden für mittlere Effekte Variablen häufig nicht als relevant identifiziert. Dies erklärt die Verschlechterung der Prädiktionsgüte. Vergleichbare Ergebnisse werden für Subsample(m) erzielt. Obwohl für Subsample(m) Variablen mit starkem Effekt seltener erkannt werden als bei Bootstrap(n), werden hier ähnliche MSE erzielt, die im gleichen Ausmaß streuen. Dies könnte bedeuten, dass die tendenziell komplexeren

Modelle für Bootstrap(n) zu keiner Verbesserung der Prädiktion führen. Eine weitere Ursache hierfür könnte sein, dass Subsample(m) tendenziell weniger irrelevante Variablen aufnimmt als Bootstrap(n). Somit werden die geschätzten Modelle an weniger irrelevante Variablen angepasst. Die mit Abstand schlechteste Prädiktionsgüte wird für Bootstrap(m) erzielt. Dies war insofern zu erwarten, dass hier relevante Variablen in den wenigsten Fällen identifiziert werden, irrelevante jedoch fast genauso oft wie für Subsample(m).

4. Fazit und Ausblick

Alle Ergebnisse dieser Arbeit basieren auf Daten, die nach dem in Kapitel 3.1 beschriebenen Setting simuliert wurden. Im ersten Schritt wurde die Modellkomplexität für verschiedene Resampling-Methoden untersucht. Grundlage war die Annahme, dass auf Bootstrap-Stichproben basierende Modelle deutlich mehr Variablen beinhalten als auf Original-Datensätzen basierende. Diese Annahmen konnten für das Lasso-Verfahren im Rahmen der vorliegenden Analysen nicht bestätigt werden. So lieferten die Original-Datensätze mit Abstand die komplexesten Modelle. Für Bootstrap(m) und Subsample(m) konnten die sparsamsten Modelle erzielt werden. Hierbei muss jedoch berücksichtigt werden, dass diese Modelle teilweise zu sparsam waren. So enthielten einige weitaus weniger Variablen als für die Erklärung des Response relevant gewesen wären. Zudem variierte besonders für Subsample(m) die Modellkomplexität stark. Da mit Hilfe der Modellkomplexität nur die Anzahl aufgenommener Variablen, aber nicht der Anteil davon relevanter Variablen bestimmt werden kann, ist es nicht möglich auf Grundlage der Modellkomplexität eine Resampling-Methode klar zu favorisieren.

Im weiteren Verlauf wurden die Inclusion Frequencies für alle Variablen verglichen. Erwartungsgemäßg konnten sowohl für die Original-Datensätze als auch für alle Resampling-Methoden mit sinkendem Effekt sinkende Inclusion Frequencies beobachtet werden. Für die Original-Datensätze wurden jeweils deutlich höhere Inclusion Frequencies erzielt als für die Resampling-Methoden. Den Ergebnissen der Original-Datensätze am ähnlichsten sind die von Bootstrap(n). Hier werden Variablen mit starkem Effekt im Vergleich zu den anderen Resampling-Methoden am häufigsten aufgenommen. Dafür werden jedoch, wie bereits die Modellkomplexität zeigen konnte, vermehrt irrelevante Variablen aufgenommen. Dementsprechend zeigen die Ergebnisse, dass vor Verwendung einer Resampling-Methode genaue Überlegungen notwendig sind, für welchen Zweck die Modelle benötigt werden. Da für Bootstrap(m) Variablen mit starkem Effekt vergleichsweise selten aufgenommen werden, ist diese Methode nur mit Vorsicht zu verwenden. Insgesamt konnte jedoch gezeigt werden, dass in dem betrachteten Simulationssetting im Schnitt alle Verfahren sehr gut zwischen relevanten und irrelevanten Variablen differenzieren können. Als problematisch könnte es sich in der Anwendung jedoch erweisen, den Wert der Inclusion Frequencies zu finden, ab dem eine Variable nicht mehr als relevant angesehen werden sollte.

Zuletzt wurde die Prädiktionsgüte für die verschiedenen Datensätze betrachtet. Hierbei wurden für die Resampling-Methoden deutlich schlechtere Ergebnisse als für die Original-Datensätze erzielt. Die MSE für Bootstrap(n) und Subsample(m) hatten in etwa die gleiche Größe, betrugen aber ein Vielfaches der MSE für die Original-Datensätze. Dies verdeutlicht, dass die für Bootstrap(n) zusätzlich aufgenommenen Variablen im Vergleich zu Subsample(m) zu keiner Verbesserung der Prädiktion führen. Die schlechtesten Ergebnisse wurden für Bootstrap(m) erzielt. Somit sollte Modellen, die zur Prädiktion dienen, besser auf Grundlage von Bootstrap(n) oder Subsample(m) geschätzt werden.

Zusammenfassend lässt sich feststellen, dass keine Resampling-Methode für das Lasso-Verfahren klar bevorzugt werden kann. Vor Verwendung einer dieser Methoden sollte sich der Anwender in jedem Fall genau überlegen, wozu die Modelle später verwendet werden beziehungsweise auf welche Art das eigentliche Modell aus den Resampling-Ergebnissen gebildet wird. Sollen die Modelle möglichst alle relevanten Variablen beinhalten und gute Vorhersagen treffen, so wäre Bootstrap(n) das geeignete Verfahren. Einen guten Kompromiss zwischen Interpretierbarkeit und Prädiktion bietet Subsample(m).

Bei Betrachtung der vorliegenden Ergebnisse muss berücksichtigt werden, dass diese nicht im Allgemeinen gültig sind, sondern sich auf den hier betrachteten, speziellen Fall beziehen. So wurden alle Daten unabhängig voneinander aus einer Standardnormalverteilung gezogen. Dementsprechend lagen zur Auswertung nur unkorrelierte, metrische Variablen vor. In der praktischen Anwendung ist es ist äußerst unwahrscheinlich, dass diese vereinfachte Datenkonstellation auftritt. Folglich sollten die beschriebenen Resultate nur als Grundlage für weitere Untersuchungen angesehen werden. Ein interessanter Ansatz wäre beispielsweise die Betrachtung von paarweise hoch korrelierten Variablen. Diese werden im Lasso-Verfahren häufig nur abwechselnd aufgenommen, das heißt das resultierende Modell beinhaltet immer nur eine der Variablen. Somit würden die Inclusion Frequencies deutlich geringere Werte annehmen als im unkorrelierten Fall. Da in der Praxis oftmals komplexe, höherdimensionale Beziehungen vorkommen, wäre es zudem interessant zu wissen, wie sich die Ergebnisse der einzelnen Resampling-Methoden in diesem Fall verändern. Als weitere Modifikation der hier analysierten Daten wäre die zusätzlich Aufnahme bi-

närer Variablen möglich. Auch diese finden in der Praxis häufig Anwendung. Da das Lasso-Verfahren zudem auf generalisierte lineare Modelle erweiterbar ist, wäre neben der Analyse für einen normalverteilten Response auch die Betrachtung anderer Verteilungsstrukturen denkbar.

Doch nicht nur eine Variation in der Datenstruktur, sondern auch eine Abwandlung der Vorgehensweise kann zu stark veränderten Ergebnissen führen. So wurde in den vorliegenden Analysen die Stichprobengröße $m = 0.632\,n$ fest gewählt und deren Auswirkung auf die Ergebnisse nicht weiter betrachtet. Ein größerer Wert von m könnte jedoch gegebenenfalls stabilere Ergebnisse erzielen. Würden die Werte von m allerdings sehr groß gewählt, so wären die Pseudo-Datensätze vermutlich zu ähnlich und es könnte durch das Resampling kaum eine Verbesserung der Ergebnisse im Vergleich zur Anwendung des Lasso-Verfahrens auf einen einzigen Original-Datensatz erzielt werden. Auch für zu kleine m sind unzureichende Ergebnisse zu erwarten, da die einzelnen Pseudo-Datensätze nur sehr wenig Information enthalten würden. (De Bin et al., in Druck)

Literatur

Bach, F. R. (2008). Bolasso: Model Consistent Lasso Estimation Through the Bootstrap, *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, pp. 33–40.

Bühlmann, P. und van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, Berlin Heidelberg.

Binder, H. und Schumacher, M. (2008). Adapting Prediction Error Estimates for Biased Selection in High-Dimensional Bootstrap Samples, *Statistical Applications in Genetics and Molecular Biology* **7**: 1–28.

Davison, A. C., Hinkley, D. V. und Young, G. A. (2003). Recent Developments in Bootstrap Methodology, *Statistical Science* **18**: 141–157.

De Bin, R., Janitza, S., Sauerbrei, W. und Boulesteix, A.-L. (in Druck). Subsampling versus Bootstrapping in Resampling-Based Model-Selection for Multivariable Regression, *Biometrics* .

Efron, B. und Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*, CRC Press, Boca Raton.

Fahrmeir, L., Hamerle, A. und Tutz, G. (1996). *Multivariate Statistische Verfahren*, de Gruyter, Berlin.

Fahrmeir, L., Kneib, T. und Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen*, Springer, Berlin Heidelberg.

Fahrmeir, L., Kneib, T., Lang, S. und Marx, B. (2013). *Regression: Models, Methods and Applications*, Springer, Berlin Heidelberg.

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

Hastie, T., Tibshirani, R. und Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer, New York.

Henderson, A. R. (2005). The Bootstrap: A Technique for Data-Driven Statistics. Using Computer-Intensive Analyses to Explore Experimental Data, *Clinica Chimica Acta* **359**: 1–26.

Hoerl, A. und Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* **12**: 55–67.

Härdle, W. H. und Simar, L. (2015). *Applied Multivariate Statistical Analysis, Forth Edition*, Springer, Berlin Heidelberg.

Janitza, S., Binder, H. und Boulesteix, A.-L. (in Druck). Pitfalls of Hypothesis Tests and Model Selection on Bootstrap Samples: Causes and Consequences in Biometrical Applications, *Biometrical Journal* .

Leng, C., Lin, Y. und Wahba, G. (2006). A Note on the Lasso and Related Procedures in Model Selection, *Statistica Sinica* **16**: 1273–1284.

Meinshausen, N. und Bühlmann, P. (2006). High-Dimensional Graphs and Variable Selection with the Lasso, *The Annals of Statistics* **34**: 1434–1462.

Meinshausen, N. und Bühlmann, P. (2010). Stability Selection, *Journal of the Royal Statistical Society, Series B* **72**: 417–473.

Sprent, P. und Smeeton, N. (2007). *Applied Nonparametric Statistical Methods, Fourth Edition*, CRC Press, Boca Raton.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B* **58**: 267–288.

Zou, H. und Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society, Series B* **67**: 301–320.

A. Anhang zusätzlicher Grafiken

Abbildung 9 zeigt den Median der Inclusion Frequencies für jeweils eine Gruppe von Variablen mit gleichem Effekt. Es besteht kein merklicher Unterschied zwischen dem Median und dem Mittelwert der Inclusion Frequencies innerhalb einer Gruppe (vergleiche Abbildung 6).

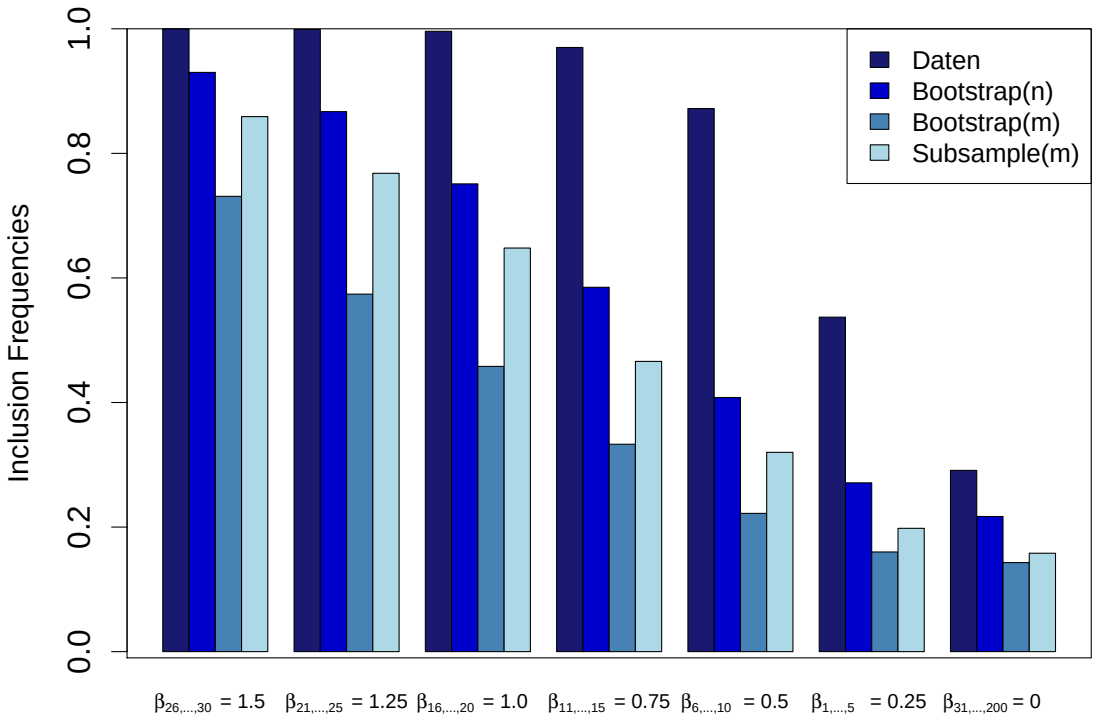

Abbildung 9: Vergleich Median der Inclusion Frequencies anhand verschieden generierter Datensätze für unterschiedlich starke Effekte

Abbildung 10 stellt den absoluten Prädiktionsfehler dar. Dieser unterscheidet sich zwar in der Größe vom quadratischen Prädiktionsfehler, die Verhältnisse der verschiedenen Methoden zueinander stimmen jedoch überein.

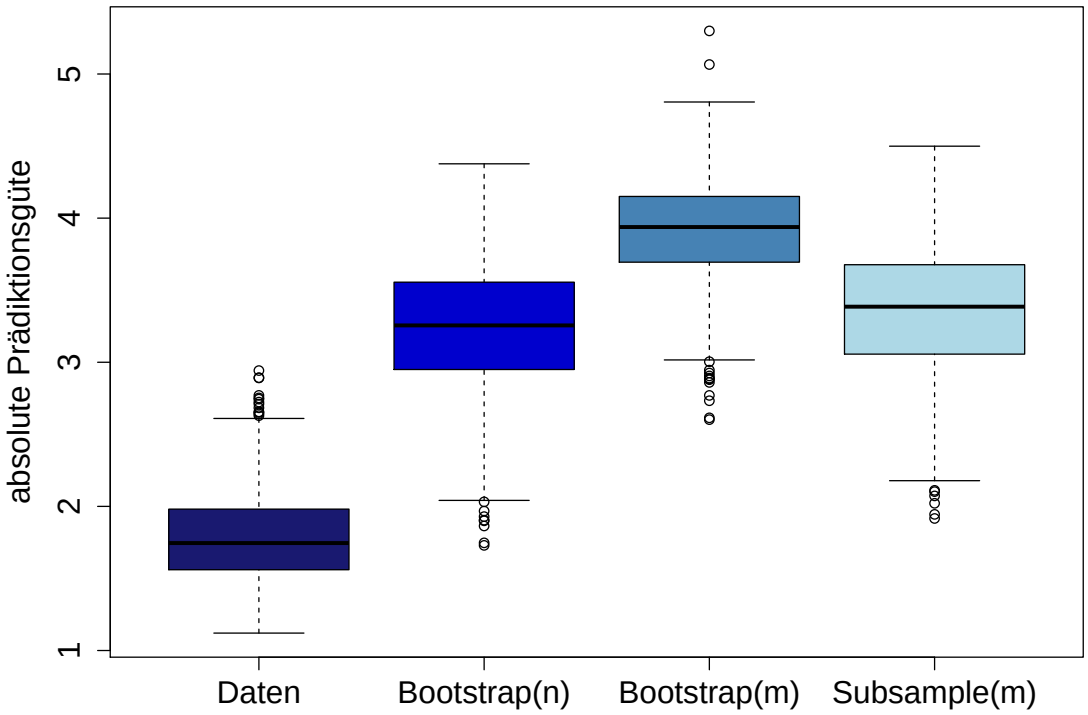

Abbildung 10: Vergleich des absoluten Prädiktionsfehlers für Original-Datensätze und verschiedene Resampling-Methoden

B. Elektronischer Anhang

Der elektronische Anhang besteht aus 3 Ordnern und einer Datei.

Der Ordner „Daten"beinhaltet die simulierten Original-Datensätze „data.RData“ sowie die daraus gezogenen Pseudo-Datensätze „bootstrap_n.RData“, „bootstrap_m.RData“ und „subsample.RData“

Der Ordner „Programme“ enthält den RCode „Simulation“ zur Simulation der Original-Datensätze und zur Ziehung der Pseudo-Datensätze. Die Anwendung des Lasso-Verfahrens und die Auswertung der resultierenden Modelle befinden sich in dem RCode „Lasso_Verfahren“. Der RCode zur anschließenden, grafischen Auswertung ist unter dem Namen „Grafische_Auswertungen“ gespeichert. Alle anderen Grafiken wurden mit Hilfe des RCodes „Weitere Grafiken“ erzeugt.

Die ausgewerteten Modelle nach Anwendung des Lasso-Verfahrens liegen unter den Dateinamen „ergebnisse_data.RData", „ergebnisse_bootstrap_n.RData“, „ergebnisse_bootstrap_m.RData“ und „ergebnisse_subsample.RData“ in dem Ordner „Ergebnisse“.

Zusätzlich zu den drei Ordnern befindet sich die vollständige vorliegende Arbeit unter dem Namen „Bachelorarbeit_Völkl.pdf“ im elektronischen Anhang.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich meine Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

.....	.....
Ort, Datum	*Johanna Völkl*