Ludwig-Maximilians-Universität

Institut für Statistik

Masterarbeit

# Lasso-Inferenz bei multipel imputierten Daten


*Betreuer:* Prof. Dr. Christian Heumann

*Autor:* Armin Reichert

14. März 2018

## Abstract

In the last years different models based on the lasso for constructing confidence intervals or p-values even in the high-dimensional setting have been proposed. However in most cases the authors don't consider the common problem of missing data. Moreover it is not apparent how these inference methods can be combined with multiple imputation as most of the time the parameter distribution is non-normal and/or there are no easily accesable standard errors, which are essential for using Rubin's pooling rules. In this master thesis a combination of bootstrap and multiple imputation, which only require the coefficient estimates, is used to construct confidence intervals. The validity of this approach is shown in a simulation study under various settings. Especially the used Lasso-Partial-Ridge Regression and the Lasso-Projection achieve overall good results.

# Inhaltsverzeichnis

**1 Problemstellung und Aufbau der Arbeit** . . . . . 1

**2 Das Lasso und darauf aufbauende Inferenz Methoden** . . . . . 4

2.1 Einführende Notation und lineare Regression . . . . . 4

2.2 Das Lasso . . . . . 5

2.3 Das adaptive Lasso . . . . . 6

2.4 Multi-Sample-Splitting . . . . . 7

2.5 Lasso-Partial-Ridge . . . . . 10

2.6 Die Lasso Projektion . . . . . 11

**3 Multiple Imputation** . . . . . 13

3.1 Kombination der multipel imputierten Datensätze . . . . . 13

3.2 Bootstrap-Inferenz unter Berücksichtigung fehlender Werte . . . . . 16

**4 Aufbau der Simulationsstudie** . . . . . 19

4.1 Konstruktion der Designmatrix . . . . . 19

4.2 Festlegung der Inferenzmethode, des Beta-Vektors und der Varianz des Störterms . . . . . 20

4.3 Erzeugung fehlender Werte . . . . . 22

4.4 Gütekriterien und Parameterspezifikationen für die statistischen Modelle 23

4.5 Abschließende Motivation . . . . . 25

**5 Ergebnisse der Simulationsstudie** . . . . . 27

5.1 Vorstellung der Ergebnisse für ein ausgewähltes Setting . . . . . 27

5.2 Coverageraten: Mittlere quadratische Abweichung und Vergleich zwischen den Inferenzmethoden . . . . . 32

5.3 Median Konfidenzintervallbreite: Lineare Regression . . . . . 33

5.4 Zusammenfassung der Ergebnisse . . . . . 35

5.5 Rechenzeit der statistischen Modelle . . . . . 35

**6 P-Werte und Ausblick** . . . . . 36

6.1 Inferenz mit p-Werten . . . . . 36

6.2 Weiterführende Forschung . . . . . 36

**Tabellenverzeichnis** . . . . . 38

**Abbildungsverzeichnis** . . . . . 39

**Literatur** . . . . . 41

**Anhang** . . . . . 44

A Tabellen . . . . . 44

B Grafiken . . . . . 52

C Elektronischer Anhang . . . . . 80

# 1 Problemstellung und Aufbau der Arbeit

Das Lasso hat sich in den letzten Jahren aufgrund seiner Simplizität und Flexibilität als eine der Standardmethoden für die Variablenselektion etabliert. Vor allem für hochdimensionale Daten, welche auch immer mehr von wirtschaftlichen Unternehmen gesammelt und analysiert werden, stellt das Lasso eine attraktive Analysemethode dar. Einfache Modelle, wie zum Beispiel die Ordinary-Least-Squares Regression (OLS), können in solchen Situationen nicht eindeutig gelöst werden, da die Designmatrix keinen vollen Rang hat. Aber auch in niedrigen Dimensionen kann das Lasso einer OLS-Regression vorzuziehen sein. Nähert sich die Zahl der Variablen den Beobachtungen an, werden die OLS-Schätzer instabil (James et al., 2014, Kapitel 6). Das Lasso umgeht diese Probleme durch den Einbezug einer $l_1$-Penalisierung. Hierdurch werden die Koeffizienten zwar verzerrt geschätzt, die Varianz der Schätzer hingegen reduziert (Hastie et al., 2001, Kapitel 3). Zusätzlich ermöglicht diese Penalisierung es dem Modell bestimmte Koeffizienten des Regressionsmodells auf exakt null zu setzen, was einer Variablenselektion entspricht.

Die Schätzung von Regressionskoeffizienten ist meist eng mit der zugehörigen Inferenz dieser Parameter verknüpft. Konfidenzintervalle, Hypothesentests und p-Werte sind wichtige Werkzeuge der Datenanalyse. Für hochdimensionale Daten ist Inferenz jedoch schwierig, da die asymptotische Verteilung vieler Schätzer kompliziert und schwer zu berechnen ist (Liu et al., 2017a). In den letzten Jahren wurde viel Forschung betrieben, dieses Problem, vor allem für das Lasso, zu lösen. Hierbei wird häufig ein zweistufiges Verfahren aus Lasso und regulärer OLS angewandt. Eine naive Methode wäre es zum Beispiel zuerst eine Lasso Regression zu berechnen, die ausgewählten Variablen als Prädiktoren einer OLS zu nutzen und mit diesem Modell Inferenz zu betreiben. Dieses Verfahren wird im Folgenden als Lasso+OLS bezeichnet. Zhao et al. (2017) raten allerdings von dieser Vorgehensweise ab, da Variablenselektion und Inferenz auf dem gleichen Datensatz erfolgen. Dezeure et al. (2015) schlagen eine Splitting Prozedur vor. Der Datensatz wird halbiert und auf der einen Hälfte wird das Lasso zur Identifizierung der Variablen berechnet. Diese ausgewählten Variablen werden daraufhin als Prädiktoren einer einfachen OLS-Regression, welche auf der zweiten Hälfte durchgeführt wird, verwendet. Da beide Hälften statistisch von-

einander unabhängig sind, ist in diesem Fall klassische Inferenz über das OLS-Modell möglich. Liu et al. (2017a) beschreiben einen Partial-Least-Square Ansatz, bei der die Variablen, welche durch das Lasso ausgewählt wurden, unpenalisiert in ein OLS Modell eingehen, die Variablen hingegen, die nicht ausgewählt wurden, mit einer $l_2$-Penalisierung versehen werden, um zusätzliche Variabilität in diesen Koeffizienten zu erzeugen. Mithilfe des Bootstraps können dann Konfidenzintervalle berechnet werden. In Zhang und Zhang (2014) wird eine Lasso-Projektionsmethode vorgestellt, dessen Koeffizienten asymptotisch normalverteilt sind und somit analytische Konfidenzintervalle konstruiert werden können.

Die Autoren gehen allerdings in ihrer Beschreibung der Verfahren und durchgeführten Simulationsstudien stets von vollständigen Datensätzen aus. Das in der Realität jedoch oftmals vorkommende Problem fehlender Werte wird ignoriert. Bei sehr geringen Ausfallquoten kann im Zweifel noch argumentiert werden, einen Listenweisen Fallausschluss durchzuführen, die state of the art Prozedur zur Handhabung unvollständiger Daten ist jedoch die multiple Imputation (Schafer und Graham, 2002). Hierbei ist die Verknüpfung der statistischen Modelle mit der multiplen Imputation jedoch nicht offensichtlich, da die pooling Regeln von Rubin normalverteilte Parameter voraussetzen (van Buuren, 2012, Kapitel 6). Diese Annahme wird allerdings für manche Lasso-Inferenzmethoden verletzt. Zudem sind nicht für alle Verfahren analytische Standardfehler vorhanden, wodurch diese auch nicht gepooled werden können. Ohne diese können jedoch weder Konfidenzintervalle, Hypothesentests oder p-Werte analytisch berechnet werden.

Ziel dieser Arbeit ist es die bereits exisiterenden Inferenzmethoden für Lasso Regressionen auf multipel imputierte Datensätze zu übertragen, um hieraus Konfidenzintervalle zu berechnen. Hierbei wird vor allem darauf Wert gelegt, dass eine Anwendung sowohl im niedrig- als auch hochdimensionalen Raum möglich ist. Von zentraler Bedeutung ist das Paper von Schomaker und Heumann (2016), in dem verschiedene Bootstrapverfahren für Situationen vorgeschlagen werden, in denen die pooling Regeln scheitern. Wir werden sehen, dass diese nur den geschätzten Parametervektor benötigen, um Inferenz zu betreiben.

In Kapitel 2 werden sowohl die Lasso Regression als auch weitere Verfahren, die auf dem Lasso aufbauen und in dieser Arbeit Anwendung finden, vorgestellt. Kapitel 3

gibt eine kurze Einführung in die Theorie fehlender Werte mit Fokus auf die multiple Imputation und zeigt auf, wie man für die in Kapitel 2 vorgestellten Verfahren über eine Kombination aus Bootstrap und multipel imputierten Datensätzen selbst dann Konfidenzintervalle erhalten kann, wenn die pooling Regeln scheitern. In einer Simulationsstudie wird untersucht, ob dieses Vorgehen valide Konfidenzintervalle liefert und welche Verfahren unter welchen Situationen besser oder schlechter abschneiden. Siehe hierfür Kapitel 4 und 5. Kapitel 6 verweist den Leser auf die entsprechende Literatur, wenn man sich für p-Werte anstelle von Konfidenzintervallen interessiert und beendet die Arbeit mit einem Ausblick möglicher Forschungsfragen, die in der Zukunft verfolgt werden könnten. Im Anhang finden sich alle Tabellen und Abbildungen der Simulationsstudie, die nicht im Ergebnissteil von Kapitel 5 eingebunden sind.

# 2 Das Lasso und darauf aufbauende Inferenz Methoden

## 2.1 Einführende Notation und lineare Regression

Im Folgenden bezeichnen wir $y = (y_1, \ldots, y_n)^T$ als die zu interessierende Zielvariable und $x_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$ als die $j$-te Kovariable, welche die Ausprägungen von $n$ Beobachtungen beinhaltet. Diese $p$ Vektoren werden in Matrixnotation als $X = (x_1, \ldots, x_p)$ definiert und stellen die Designmatrix der Dimension $n \times p$ dar. $Y$ hingegen entspricht der Matrixnotation der Zielgröße $y$. Zusätzlich definieren wir $D = (x_1, \ldots, x_p, y)$ als eine $n \times (p+1)$ Datenmatrix, welche sowohl $X$ als auch $Y$ enthält und somit alle vorhandenen Variablen und Beobachtungen beinhaltet.
Wir motivieren die in diesem Kapitel beschriebenen Verfahren über das einfache lineare Regressionsmodell (OLS):

$$Y = X\beta + \epsilon,\ \epsilon \overset{iid}{\sim} N(0, \sigma^2 I) \tag{2.1}$$

bei dem $\beta = (\beta_1, \ldots, \beta_p)^T$ den Vektor der unbekannten Regressionskoeffizienten darstellt. Für den Störterm $\epsilon$ wird eine Normalverteilung mit der Varianz $\sigma^2$ angenommen. Wir gehen zudem davon aus, dass die Daten zentriert sind. Der Interzept kann somit ohne Verlust von Generalität in den Modellgleichungen ignoriert werden.
Die Koeffizienten $\beta$ ergeben sich, indem die Residuenquadratsumme (RSS) minimiert wird (Hastie et al., 2001, Kapitel 2):

$$\mathrm{RSS}(\beta) = (Y - X\beta)^T (Y - X\beta) \tag{2.2}$$

Unter der Annahme, dass $X$ vollen Rang hat, ist $X^T X$ invertierbar und es kann die analytische Lösung für den Beta-Vektor gefunden werden:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \tag{2.3}$$

Allerdings kommt es in vielen statistischen Fragestellungen immer häufiger vor, dass sich die Anzahl der erhobenen Merkmale an die der Beobachtungen annähert oder diese sogar übersteigt ($p > n$). Im letzteren Fall hat $X$ keinen vollen Rang und eine eindeutige Lösung von Gleichung 2.3 existiert nicht. Aber auch für niedrig-

dimensionale Datensätze ($p < n$) kann es zu Problemen kommen, wenn das Verhältnis zwischen Beobachtungen und Merkmalen zu gering ist. Eine ad-hoc Regel lautet, dass jeder Variable mindestens 20 Beobachtungen gegenüberstehen sollten (Department of Biostatistics, Vanderbilt University, o. J.). Für zu kleine Verhältnisse sind instabile Schätzer, Overfitting und schlechte Prognosen für zukünftige Daten die Folge. In solchen Fällen können penalisierte Regressionen, wie die Ridge- oder Lasso-Regression, genutzt werden, welche die Varianz der Schätzer stabilisieren (James et al., 2014, Kapitel 6). In der folgenden Arbeit werden wir uns auf die Lasso-Regression und deren Erweiterungen beschränken.

## 2.2 Das Lasso

Das Lasso ist eine Abwandlung der einfachen linearen Regression (Gleichung 2.1) und wird hauptsächlich zur Variablenselektion, Schätzung der Koeffizienten und Prognose genutzt (James et al., 2014, Kapitel 6):

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \left( \text{RSS}(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{2.4}$$

Der erste Teil des Minimierungsproblems ist die Residuenquadratsumme aus Gleichung 2.2. Der zweite Summand stellt einen $l_1$ Penalisierungsterm dar, dessen Einfluss zusätzlich durch $\lambda$ gelenkt wird. Es ist leicht zu erkennen, dass für $\lambda = 0$ die Lasso Koeffizienten identisch mit denen einer OLS-Regression sind. Ist $\lambda$ hingegen hinreichend groß, können Koeffizienten exakt auf null gedrückt werden und das Lasso führt somit neben der Schätzung der Koeffizienten gleichzeitig eine Variablenselektion durch. Da die Lösung dieses Minimierungsproblems von $\lambda$ abhängt, muss dieses sinnvoll gewählt werden. Dies geschieht meist mithilfe einer Kreuzvalidierung. Je größer hierbei das $\lambda$ gewählt wird, desto stärker werden die einzelnen Koeffizienten gegen null gedrückt und desto größer ist die Verzerrung. Auf der anderen Seite nimmt mit steigendem $\lambda$ die Flexibilität des Lasso-Modells ab, wodurch die Varianz der Schätzer, wie bereits angesprochen, stabilisiert wird (James et al., 2014, Kapitel 6). Diese Eigenschaft ist typisch für penalisierte Regressionen und wird als Bias-Varianz-Tradeoff bezeichnet. Zusätzlich werden vor der Lösung des Minimierungsproblems die Prädiktoren standardisiert, da dieses nicht invariant gegenüber

Transformationen ist (Hastie et al., 2001, Kapitel 3). Diese Standardisierung wird ebenfalls für alle weiteren Verfahren dieses Kapitels angewandt, allerdings nicht jedes Mal explizit erwähnt.

## 2.3 Das adaptive Lasso

Das adaptive Lasso wurde von Zou (2006) entwickelt, da dieses Verfahren im Vergleich zu dem einfachen Lasso die sogenannten Orakel-Eigenschaften erfüllt. Ein statistisches Modell zur Variablenselektion besitzt diese Orakel-Eigenschaften, wenn es konsistent die richtigen Variablen auswählt und dabei optimale Prognosegüte besitzt. Zou (2006) zeigt, dass es Situationen für das einfache Lasso gibt, bei denen ein festes $\lambda$, welches optimal für die Variablenselektion ist, allerdings zu verzerrten Schätzern selbst von großen Regressionskoeffizienten führt, wodurch die Prognosegüte des Modells abnimmt. Umgekehrt gibt es ebenfalls Fälle, bei denen ein festes $\lambda$, welches zu optimalen Prognosen führt, inkonsistent in der Variablenselektion ist (Meinshausen und Bühlmann, 2006). Als Lösung wird von Zou (2006) das adaptive Lasso vorgeschlagen:

$$\hat{\beta}_{Ad.Lasso} = \arg\min_{\beta} \left( \mathrm{RSS}(\beta) + \lambda \sum_{j=1}^{p} \hat{\omega}_j \left|\beta_j\right| \right) \tag{2.5}$$

Dieses Minimierungsproblem ähnelt dem aus Gleichung 2.4 stark und unterscheidet sich einzig durch den zusätzlichen Gewichtungsfaktor $\hat{\omega}_j$ im Penalisierungsteil mit:

$$\hat{\omega}_j = \frac{1}{\left|\hat{\beta}_j^*\right|^{\gamma}} \tag{2.6}$$

Für jeden Koeffizienten kann somit eine individuelle Regularisierung durchgeführt werden, wobei $\hat{\beta}^*$ meist durch eine vorangehende OLS- oder Ridge-Regression geschätzt wird und $\gamma$ eine positive Konstante darstellt, für die von dem Autor die Werte 0.5, 1 oder 2 vorgeschlagen werden (Zou, 2006).

Das adaptive Lasso kann für die gleichen Fragestellungen wie das einfache Lasso benutzt werden. Auch bei dem erweiterten Verfahren werden Variablen mit wachsendem $\lambda$ gegen null gedrückt und eine Varianzreduktion der Koeffizienten geht auf Kosten einer eigens eingeführten Verzerrung einher. Wie für das Lasso sind auch

bei dem adaptive Lasso die Koeffizienten jedoch nicht normalverteilt (Pötscher und Schneider, 2007).

## 2.4 Multi-Sample-Splitting

Statistische Fragestellungen lassen sich meist in zwei grobe Gruppen einteilen: Prognose und Inferenz. Die beiden bisher vorgestellten Lasso-Modelle werden hauptsächlich zur Schätzung der Regressionskoeffizienten und für die Erstellung zukünftiger Prognosen genutzt. Die Inferenz der geschätzten Parameter bleibt hingegen meist außen vor. Die gängigen R-Pakete zur Berechnung penalisierter Regressionen wie 'glmnet' (Friedman et al., 2010) oder 'penalized' (Goeman et al., 2017) geben in den entsprechenden Funktionen beispielsweise keine Standardfehler der Lasso-Koeffizienten heraus, welche für analytische Inferenzmethoden jedoch ein fundamentaler Bestandteil sind. Goeman et al. (2016) argumentiert, dass diese zwar durch Bootstrapverfahren berechnet werden könnten, allerdings nicht aussagekräftig für stark verzerrte Schätzer sind. Penalisierte Regressionen benutzen gerade den Bias-Varianz-Tradeoff, um die Variabilität der Koeffizienten zu reduzieren. Somit entsteht allerdings bei Betrachtung der Standardfehler ein zu optimistisches Bild der Präzision des Modells, da zwar die Standardfehler klein sind, diese sich jedoch auf verzerrte Koeffizienten beziehen. Vor allem die Konstruktion valider Konfidenzintervalle erschwert sich durch diese Merkmale und der Verletzung der Normalverteilungs-Annahme der Koeffizienten für das Lasso und das adaptive Lasso ungemein. Um dem entgegenzuwirken, wurden verschiedene auf dem Lasso aufbauende Verfahren entwickelt, die versuchen diese Verzerrung aufzuheben und in der Lage sind für vollständige Datensätze valide Inferenz zu betreiben.

Wir beginnen mit einer Datensatz-Splitting Prozedur, welche im Folgenden als Multi-Sample-Splitting (MSS) bezeichnet wird (Dezeure et al., 2015) und definieren das aktive Set an Variablen

$$S = \{j; \beta_j \neq 0,\, j = 1, \ldots, p\}, \tag{2.7}$$

dessen Mächtigkeit $|S|$ und die Anzahl der Beobachtungen eines Datensatzes $D$ mit $N(D)$. Die grundlegende Idee dieses Verfahrens ist es, $D$ in zwei Hälften zu teilen, auf der ersten eine Variablenselektion durchzuführen und mit dem hieraus erhaltenen

Set $S$ an Variablen mithilfe der zweiten Datensatzhälfte valide Inferenz zu betreiben (Dezeure et al., 2015). Diese Teilung des Datensatzes erinnert stark an Maschinelles Lernen, bei dem ebenfalls eine Aufteilung des ursprünglichen Datensatzes in einen Trainings- und Testdatensatz stattfindet. Hierdurch wird vor allem Overfitting vermieden, was die Prognosegüte massiv verschlechtern kann. Selbiges Prinzip gilt auch für die MSS-Prozedur. Würden Variablenselektion und Inferenz, wie bei dem Lasso+OLS Verfahren, auf ein und dem selben Datensatz stattfinden, ist intuitiv leicht nachvollziehbar, dass dies zu verzerrten Inferenzaussagen führen kann. Denn in dem ersten Schritt werden gerade die Variablen ausgewählt, bei denen sich das Variablenselektionsmodell sicher ist, dass sich die zugehörigen Koeffizienten von null unterscheiden. Ansonsten würde zum Beispiel eine Lasso-Regression diese auf null setzen. Rechnet man nun in einem zweiten Schritt eine einfache OLS-Regression auf den gleichen Beobachtungen, wobei als Prädiktoren das Set $S$ genutzt wird, kann nicht ohne weiteres angenommen werden, dass dieses Vorgehen zu valider Inferenz führt (Zhao et al., 2017). Durch die Teilung des Datensatzes wird dieses Problem jedoch vermieden, da die beiden Hälften zwar aus der gleichen Population stammen, statistisch gesehen jedoch unabhängig voneinander sind.

Zurück zu dem MSS-Verfahren. Der ursprüngliche Datensatz $D$ mit den Indizes der Beobachtungen $\{1,\ldots,n\}$ wird in zwei möglichst gleich große Hälften $D_1$ und $D_2$ geteilt. Somit gilt $N(D_1) = \lfloor n/2 \rfloor$, $N(D_2) = n - \lfloor n/2 \rfloor$ und $D_1 \cap D_2 = \emptyset$. $D_2$ ist damit genauso groß wie $D_1$ oder wenn $n$ ungerade ist eine Beobachtung reicher. Wie wir sehen werden, ist dies eine wichtige Eigenschaft, damit die MSS-Prozedur auch auf hochdimensionale Datensätze mit ungerader Anzahl an Beobachtungen angewendet werden kann.

Auf $D_1$ kann nun beispielsweise das Lasso zur Variablenselektion genutzt werden und wir erhalten das Set an Koeffizienten, die ungleich null sind:

$$\hat{S}(D_1) \subseteq \{1,\ldots,p\} \tag{2.8}$$

Eine bisher nicht genannte Limitation des Lasso's ist, dass es im Fall $p > n$ maximal $n$ Variablen Koeffizienten ungleich null zuordnen kann (Zou und Hastie, 2005). Für $D_1$ können somit maximal $\lfloor n/2 \rfloor$ Variablen ausgewählt werden und es gilt $|\hat{S}(D_1)| \leq n/2 \leq N(D_2)$. Dies ermöglicht es mit dem Set $\hat{S}(D_1)$ auf $D_2$ eine OLS-

Regression durchzuführen, bei der mindestens gleich viele Beobachtungen Variablen gegenüberstehen und somit eine eindeutige Lösung für die Regressionskoeffizienten, welche unverzerrt geschätzt werden, vorhanden ist. Neben den Standardfehlern der Koeffizienten können unter den klassischen Gauß-Markov-Annahmen die t-Tests $H_0 : \beta_j = 0$ v.s. $H_1 : \beta_j \neq 0$ konstruiert werden, wodurch man für die in $\hat{S}(D_1)$ ausgewählten Variablen die zugehörigen p-Werte erhält (Dezeure et al., 2015). Für alle $j \notin \hat{S}(D_1)$ können wir den p-Wert $P_j = 1$ zuordnen. Somit ist für jeden Prädiktor in $D$ nun ein p-Wert berechnet worden. Das Set $\hat{S}(D_1)$ und die zugehörigen p-Werte hängen allerdings stark davon ab, welche Beobachtungen durch das zufällige Aufteilen in $D_1$ und $D_2$ fallen. Die Autoren nennen dieses Phänomen eine p-Werte Lotterie (Dezeure et al., 2015). Als Lösung schlagen sie vor, das hier vorgestellte Verfahren $R$-mal durchzuführen. Somit erhält man für jede Variable eine Menge an p-Werten:

$$P_j^{[1]}, \ldots, P_j^{[R]} \qquad (j = 1, \ldots, p) \tag{2.9}$$

Ziel ist es nun diese durch ein Aggregationsverfahren zu einem einzelnen p-Wert für jede Variable zusammenzufassen. Hierbei muss vor allem berücksichtigt werden, dass die p-Werte einer Variable miteinander korreliert sind. Schließlich stammen die Datensatzhälften $D_1$ und $D_2$ stets von $D$ ab. Hierfür kann beispielsweise eine empirische Quantilsfunktion mit $0 < \gamma < 1$ verwendet werden (Dezeure et al., 2015):

$$Q_j(\gamma) = min\left(emp.\gamma - Quantil\left\{P_j^{[r]}/\gamma;\ r = 1, \ldots, R\right\}, 1\right) \tag{2.10}$$

Für $\gamma = 1/2$ entspricht dies dem beispielsweise dem Median, der dann mit dem Faktor 2 multipliziert wird. Zudem können Konfidenzintervalle über die Dualität mit den p-Werten konstruiert werden, falls man sich für diese interessiert. Das detaillierte Vorgehen kann in Dezeure et al. (2015) nachgelesen werden.

Der Vorteil, Inferenz zu betreiben, geht allerdings mit dem Nachteil einher, dass für die MSS-Prozedur die Eigenschaft der Variablenselektion verloren geht. Denn für jede der $R$ Teilungen des Datensatzes können unterschiedliche Variablen in dem Lasso-Schritt ausgewählt werden, dessen Koeffizienten anschließend über OLS geschätzt werden. Wenn man nun beispielsweise die $R$ Koeffizienten einer jeden Variable mittelt, um einen einzigen Vektor der Länge $p$ zu erhalten, ist es wahrscheinlich, dass

sich alle Koeffizienten von null unterscheiden. Nur wenn in jeder der $R$ Teilungen eine Variable stets auf null gesetzt wird, bleibt dies auch nach der Aggregation erhalten.

## 2.5 Lasso-Partial-Ridge

Ein weiteres Inferenzverfahren stellt die Lasso-Partial-Ridge (LPR) Regression dar. Die Idee ist simpel. In einem ersten Schritt wird das Lasso auf den kompletten Datensatz $D$ angewandt, wodurch einem Set an Variablen $\hat{S}(D) \subseteq \{1, \ldots, p\}$ Koeffizienten ungleich null zugewiesen werden. In einem zweiten Schritt wird eine Partial-Ridge-Regression ebenfalls auf $D$ durchgeführt, bei der die empirische $l_2$ Verlustfunktion mit keiner Penalisierung für die ausgewählten Variablen, aber einem $l_2$ Penalisierungsterm für die nicht ausgewählten Variablen minimiert wird (Liu et al., 2017a). Der LPR-Schätzer hat somit folgende Form:

$$\hat{\beta}_{LPR} = \arg\min_{\beta} \left( \frac{1}{2n} \|Y - X\beta\|_2^2 + \frac{\lambda_2}{2} \sum_{j \notin \hat{S}} \beta_j^2 \right) \tag{2.11}$$

wobei $\lambda_2$ ein zusätzlicher Regularisierungs-Parameter ist. Liu et al. (2017a) konnten in ihrer Studie zeigen, dass $\lambda_2 = 1/n$ für verschiedene Varianzen des Störterms gut funktioniert.

Durch den LPR-Schätzer wird die Verzerrung in den Regressionskoeffizienten der ausgewählten Variablen vermindert, während die Variation für nicht ausgewählte Variablen erhöht wird. Warum dieses Vorgehen sinnvoll, ist lässt sich leicht an einem Beispiel in Anlehnung an Liu et al. (2017a) vorführen. Nehmen wir an, der wahre Beta-Vektor besteht aus einigen großen und kleinen Koeffizienten. Würde man versuchen aus einer Kombination aus Bootstrap und Lasso+OLS Konfidenzintervalle zu konstruieren, werden vor allem in dem Lasso-Schritt die größeren Koeffizienten ausgewählt. Die kleineren werden somit in dem OLS-Fit nicht mehr betrachtet und implizit auf null gesetzt. Als Folge kann es passieren, dass für einige Variablen mit kleinen Koeffizienten sehr schmale Konfidenzintervalle entstehen (im Extremfall [0,0], wenn eine Variable im Lasso-Schritt nie ausgewählt wird), was zu schlechten Coverageraten führt. Der LPR-Schätzer versucht dem entgegen zu wirken, indem er zusätzliche Variation in den nicht ausgewählten Variablen durch die $l_2$ Penalisierung erzeugt, um auch für kleine Koeffizienten die Coverageraten zu verbessern. Konfi-

denzintervalle können für den LPR-Schätzer über Bootstrapverfahren, wie in Liu et al. (2017a) beschrieben, konstruiert werden. Hierbei stellt sich jedoch die Frage, ab welcher Größenordnung Koeffizienten als klein oder sogar als zu klein gelten, um konsistent von verschiedenen Methoden entdeckt zu werden. Ein hierfür oft genutztes Kriterium stellt die Beta-Min Kondition dar. Diese verlangt, dass der Betrag des kleinsten Regressionskoeffizienten (mit der null ausgeschlossen) viel größer als $1/\sqrt{n}$ ist. Die oben genannte Bootstrap Lasso+OLS Prozedur benötigt zum Beispiel die in der Realität oft verletzte Beta-Min Kondition (Liu und Yu, 2013), während dies für den LPR-Schätzer nicht der Fall ist (Liu et al., 2017a).
Allerdings geht auch hier die praktische Funktionalität des Lasso's Variablenselektion zu betreiben, aufgrund der $l_2$ Penalisierung für nicht ausgewählte Variablen verloren. Die Idee der LPR-Regression, ist gerade für diese Variablen zusätzliche Variation zu erzeugen. Hierdurch werden die Schätzer jedoch wieder von der null weggedrückt. Als Folge kann das LPR-Verfahren nicht so einfach zur Variablenselektion genutzt werden. Man könnte zwar argumentieren, die Variablen auszuwählen, für die das konstruierte Konfidenzintervall die null nicht enthält, allerdings sind dann die Regressionskoeffizienten nicht mehr valide, da diese unter dem vollen Modell geschätzt wurden.

## 2.6 Die Lasso Projektion

Die letzte Methode, die in diesem Kapitel vorgestellt wird, ist die Lasso-Projektion (LP), welche laut den Autoren Zhang und Zhang (2014) wie der LPR-Schätzer in der Lage ist, auch für kleine Regressionskoeffizienten valide Konfidenzintervalle zu kreieren. Das Verfahren wird wie bei Dezeure et al. (2015) zuerst an dem einfachen OLS-Fall motiviert. Neben der Schätzung des Beta-Vektors über die Normalengleichung $X^T X\beta = X^T Y$ lässt sich der $j$-te Regressionskoeffizient mit $j = 1, \ldots, p$ auch über folgende Form darstellen:

$$\hat{\beta}_{OLS;j} = Y^T Z^{(j)} / (X^{(j)})^T Z^{(j)} \tag{2.12}$$

wobei $Z^{(j)}$ die Residuen der OLS-Regression $X^{(j)}$ gegen die restlichen Prädiktoren $X^{(-j)}$ sind. Wie bei der Methode der kleinsten Quadrate ist diese Vorgehensweise

nur in dem $p < n$ Fall anwendbar, da ansonsten die Residuen $Z^{(j)}$ null sind und Gleichung 2.12 nicht eindeutig lösbar ist.

Für hochdimensionale Daten kann jedoch stattdessen eine regularisierte Projektion verwendet werden. Wir benutzen eine Lasso Regression $X^{(j)}$ gegen $X^{(-j)}$, um die entsprechenden Residuen $Z^{(j)}_{Lasso}(\lambda_j) = Z^{(j)}_{Lasso}$ zu erhalten, welche von dem gewählten Regularisierungsparameter $\lambda_j$ abhängen.

Da im Gegensatz zu $Z^{(j)}$ die Residuen $Z^{(j)}_{Lasso}$ nicht orthogonal zueinander stehen, impliziert dies einen Bias in den Regressionskoeffizienten, welcher schon in den vorherigen Kapiteln zu Penaliserungsmethoden erwähnt, allerdings anders motiviert wurde (Dezeure et al., 2015). Somit ist es notwendig, eine Korrektur der Verzerrung vorzunehmen, für die der Lasso-Schätzer $\hat{\beta}$ aus $Y$ gegen $X$ genutzt wird (Dezeure et al., 2015):

$$\hat{\beta}_{LP;j} = \frac{Y^T Z^{(j)}_{Lasso}}{(X^{(j)})^T Z^{(j)}_{Lasso}} - \sum_{k \neq j} P_{jk} \hat{\beta}_k \tag{2.13}$$

mit

$$P_{jk} = (X^{(k)})^T Z^{(j)}_{Lasso} / (X^{(j)})^T Z^{(j)}_{Lasso} \tag{2.14}$$

Es lässt sich zeigen, dass Koeffizienten mindestens asymptotisch normalverteilt sind und man für diese analytische Konfidenzintervalle berechnen kann. Da diese Eigenschaft im späteren Verlauf nicht genutzt wird, sei der interessierte Leser auf Dezeure et al. (2015) und Zhang und Zhang (2014) verwiesen.

Wie die MSS-und die LPR-Methode verliert die Lasso-Projektion jedoch die Eigenschaft der Variablenselektion, da die Regressionskoeffizienten nicht gegen null gedrückt werden.

# 3 Multiple Imputation

In den bisherigen Ausführungen wurde stets angenommen, dass der zugrunde liegende Datensatz vollständig ist. Dies ist in der Realität jedoch oftmals nicht der Fall, fehlende Werte sind eher die Normalität als die Ausnahme.
Die Standardprozedur, mit diesen umzugehen, ist die multiple Imputation. Ein Listenweiser Fallausschluss verschwendet unnötig Informationen und kann zu einer starken Verzerrung der geschätzten Mittelwerte, Regressionskoeffizienten und Korrelationen führen (van Buuren, 2012, Kapitel 1). Auch singuläre Imputationsmethoden sind in den meisten Fällen nicht zu rechtfertigen, da die Unsicherheit, die mit den fehlenden Werten einhergeht, nicht berücksichtigt wird. Singulär imputierten Werten wird das gleiche Gewicht zugeordnet wie erfassten Daten. Vor allem Standardfehler geschätzter Parameter werden somit unterschätzt (Enders, 2010, Kapitel 2). Die Multiple Imputation behebt dieses Problem, indem sie jeden fehlenden Wert mehrmals imputiert und somit $M$ vollständige Datensätze erzeugt, die sich alleine in den imputierten Werten unterscheiden. Wir werden allerdings sehen, dass die pooling Regeln, welche zur Aggregation der geschätzten Parameter und Standardfehler der $M$ Datensätze genutzt werden, für einige Verfahren aus Kapitel 2 nicht angewandt werden können. Eine mögliche Lösung, wie dennoch Inferenz für multipel imputierte Datensätze betrieben werden kann, wird in Kapitel 3.2 vorgestellt.

## 3.1 Kombination der multipel imputierten Datensätze

Wir interessieren uns für einen Parametervektor $\theta$ eines beliebigen Analysemodells (in Bezug zu dieser Arbeit wären dies zum Beispiel die Regressionskoeffizienten) und definieren $\hat{\theta}_m$ als die geschätzten Parameter und $\hat{V}_m$ als deren Varianz-Kovarianzmatrix für den $m$-ten imputierten Datensatz mit $m = 1, \ldots, M$ (Carpenter und Kenward, 2012, Kapitel 2). Der gepoolte Parametervektor lässt sich dann als einfacher Mittelwert darstellen:

$$\hat{\theta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{3.1}$$

Die gepoolte Kovarianzmatrix $\hat{V}_{MI}$ hingegen wird aus einer Kombination aus der mittleren Kovarianzmatrix innerhalb der Gruppen:

$$\hat{W} = \frac{1}{M} \sum_{m=1}^{M} \hat{V}_m \tag{3.2}$$

und der Kovarianzmatrix zwischen den Gruppen:

$$\hat{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta}_{MI})(\hat{\theta}_m - \hat{\theta}_{MI})^T \tag{3.3}$$

berechnet (van Buuren, 2012, Kapitel 2):

$$\hat{V}_{MI} = \hat{W} + \left(1 + \frac{1}{M}\right) \hat{B} \tag{3.4}$$

Die gepoolte Varianz ist somit nicht die Summe aus der Varianz zwischen und innerhalb der Gruppen, sondern beinhaltet zusätzlich noch $\hat{B}/M$. Dieser Term ist notwendig, da $\hat{\theta}_{MI}$ selbst nur aus einer finiten Anzahl ($M$) von Datenpunkten berechnet wurde. Steigt die Anzahl der imputierten Datensätze gegen unendlich, verschwindet diese Streuung nach und nach (Enders, 2010, Kapitel 8). Standardfehler erhält man wie üblich über die Wurzel der Varianzen aus $\hat{V}_{MI}$. Konfidenzintervalle für $\hat{\theta}_{MI}$ können dann in dem skalaren Fall mithilfe der Quantile einer $t_\nu$ Verteilung mit approximativen Freiheitsgraden

$$\nu = (M-1)\left(1 + \frac{\hat{W}}{\hat{B} + \hat{B}/M}\right)^2 \tag{3.5}$$

berechnet werden (Carpenter und Kenward, 2012, Kapitel 2). Diese pooling Regeln gehen jedoch davon aus, dass die geschätzten Parameter $\hat{\theta}$ einer Normalverteilung um den wahren Populationswert $\theta$ mit Varianz $W$ folgen (van Buuren, 2012, Kapitel 6). Diese Annahme wird beispielsweise bei dem Lasso, adaptive Lasso und dem LPR-Schätzer verletzt, da deren Parameter weder t- noch normalverteilt sind. Zusätzlich sind für diese Verfahren auch keine analytischen Standardfehler vorhanden, so dass diese auch nicht gepooled werden können.

Aber auch für das MSS-Verfahren stößt man auf Probleme. Wenn man sich zurückerinnert, wurde die Teilung des Datensatzes $R$-mal vorgenommen, um die p-Lotterie zu vermeiden. Hierbei entstehen für jede Variable mehrere geschätzte Koeffizienten und Standardfehler, welche vor Anwendung der pooling Regeln ebenfalls aggregiert werden müssen. Eine auf dem ersten Blick intuitive Möglichkeit, wäre auch die poo-

ling Regeln für die $R$-Iterationen anzuwenden. Die $R$ Koeffizienten könnten über Gleichung 3.1 gemittelt werden. Für die Standardfehler hingegen ist eine solche Aggregation schwieriger, da nicht für jede Variable $R$ Standardfehler zur Verfügung stehen. Nur wenn eine Variable in jeder der $R$ Lasso-Regressionen auf $D_1$ ausgewählt wurde, ist dies der Fall. Im Gegensatz zu den Koeffizienten ist hierbei nicht sofort klar, wie mit dieser Problematik umzugehen ist. Eine Möglichkeit wäre, auch die Standardfehler für nicht ausgewählte Variablen auf null zu setzen (Schomaker und Heumann, 2014). Somit wären nun für jede Variable $R$ Standardfehler vorhanden, die über Gleichung 3.4 kombiniert werden könnten. Denn wie bei der multiplen Imputation liegen für das MSS-Verfahren mehrere geschätzte Koeffizienten mit zugehörigen Standardfehlern für eine einzelne Variable vor. Somit gibt es auch hier eine Streuung innerhalb und zwischen den $R$ Datensätzen. Allerdings ist für diese Vorgehensweise ebenfalls die Annahme der pooling Regeln nötig, dass die geschätzten Parameter einer Normalverteilung um den wahren Populationswert folgen, welche jedoch verletzt wird. Die Koeffizienten der durch das Lasso ausgewählten Variablen sind zwar für $D_2$ normalverteilt, betrachtet man jedoch die Verteilung der $R$ Koeffizienten einer Variable ist dies nicht mehr der Fall. Denn nur wenn eine Variable ausgewählt wurde, wird anschließend der Koeffizient auf $D_2$ über eine OLS-Regression geschätzt. Ansonsten wird der Variable ein Wert von null zugeordnet. Hierbei kann man intuitiv leicht nachvollziehen, dass die Koeffizienten über die $R$ Iterationen hinweg keiner Normalverteilung, um den wahren Populationswert folgen. Stattdessen liegt aufgrund der vorangehenden Variablenselektion über das Lasso eine erhöhte Wahrscheinlichkeitsmasse, als unter einer Normalverteilung zu erwarten wäre, an der Stelle null vor. Zusammenfassend kann man somit sagen, dass auch für das MSS-Verfahren eine Aggregation der Parameter alles andere als offensichtlich ist.

Eine Ausnahme hingegen bildet die Lasso-Projektion. Für dieses Verfahren liegen analytische Standardfehler vor und die Parameter sind asymptotisch normalverteilt (Zhang und Zhang, 2014). In diesem Fall können die Koeffizienten mit Gleichung 3.1 und die analytischen Standardfehler über Gleichung 3.4 gepooled werden, um hieraus Konfidenzintervalle basierend auf einer $t_\nu$ Verteilung zu berechnen. Allerdings muss man sich stets im Klaren sein, dass asymptotische Eigenschaften für endliche Stichproben ein falsches Bild der tatsächlichen Verteilung geben können (Zou, 2006), so

dass auch die Anwendung der pooling Regeln für die Lasso-Projektion mit Vorsicht zu betrachten ist.

Es konnte somit gezeigt werden, dass die meisten der in Kapitel 2 beschriebenen Verfahren die klassischen pooling Regeln nicht nutzen können. Es stellt sich daher die Frage, wie mit dieser Problematik umzugehen ist, denn schließlich ist die multiple Imputation das Standardverfahren für die Handhabung fehlender Werte.
Ziel dieser Arbeit ist es, eine Inferenzmethode vorzustellen und zu validieren, welche sowohl auf alle bisher betrachteten als auch in der Zukunft entwickelten statistischen Modelle, welche die benötigten Annahmen der pooling Regeln verletzen, angewandt werden kann. Ein solcher Lösungsansatz wird im Folgenden Kapitel vorgestellt.

## 3.2 Bootstrap-Inferenz unter Berücksichtigung fehlender Werte

Schomaker und Heumann (2016) beschäftigten sich allgemein mit der Frage, wie Inferenz für multipel imputierte Datensätze erfolgen kann, wenn es keine analytische Lösung zur Schätzung von Standardfehlern gibt, oder die Verteilungsannahme der Parameter verletzt ist und die allgemeinen pooling Regeln somit nicht angewendet werden können. Hierbei fokussieren sie sich hauptsächlich auf die Konstruktion valider Konfidenzintervalle und schlagen vier verschiedene Kombinationen aus Bootstrapping und multipel imputierten Datensätzen vor, welche in den folgenden Absätzen vorgestellt werden:

- **Methode 1, MI-Boot gepoolte Stichprobe (PS[1])**: Der unvollständige Datensatz wird $M$-mal multipel imputiert und für alle $M$ Datensätze werden jeweils $K$ Bootstrapstichproben gezogen, so dass man $M \times K$ vollständige Datensätze erhält. Auf all diesen Datensätzen wird der Schätzer $\hat{\theta}_{m,k}$ mit $m = 1,\ldots,M$ und $k = 1,\ldots,K$ des statistischen Modells berechnet. Diese $M \times K$ Parameter werden jeweils geordnet, um hieraus die zugehörigen $1-\alpha\%$ Konfidenzintervalle für $\theta$ zu erzeugen, bei dem die untere ($\hat{\theta}_u$) und obere ($\hat{\theta}_o$) Intervallgrenze durch die entsprechenden Quantile $\hat{\theta}^{\frac{\alpha}{2}}$ und $\hat{\theta}^{1-\frac{\alpha}{2}}$ definiert sind:

$$\left[\hat{\theta}_u; \hat{\theta}_o\right] = \left[\hat{\theta}^{\frac{\alpha}{2}}; \hat{\theta}^{1-\frac{\alpha}{2}}\right] \tag{3.6}$$

[1] Für den englischen Begriff: pooled sample

- **Methode 2, MI-Boot**: Wie in Methode 1 werden jeweils $K$ Bootstrapstichproben für alle $M$ Datensätze gezogen. Diese $K$ Bootstrapziehungen werden jeweils benutzt, um $M$ Standardfehler für jeden skalaren Parameter zu schätzen. Diese können nun nach den allgemeinen pooling Regeln aus Kapitel 3 kombiniert werden, um somit Konfidenzintervalle basierend auf einer $t_\nu$ Verteilung zu konstruieren [2].

- **Methode 3, Boot-MI gepoolte Stichprobe (PS)**: Dieses Setting ähnelt Methode 1 sehr stark. Allein die Reihenfolge von Imputation und Bootstrap wird getauscht. Für den unvollständigen Datensatz werden $K$ Bootstrapstichproben erzeugt und diese jeweils multipel imputiert. Hierbei entstehen $K \times M$ vollständige Datensätze, bei der die geschätzten Parameter wie in Methode 1 geordnet und aus den Quantilen die Intervallgrenzen erzeugt werden.

- **Methode 4, Boot-MI**: Auch hier werden zuerst $K$ Bootstrapziehungen auf dem unvollständigen Datensatz durchgeführt und diese anschließend $M$-mal imputiert. Für jede Bootstrapstichprobe kann nun Gleichung 3.1 angewandt werden, um die jeweils $M$ Parameter zu kombinieren. Somit entstehen $K$ gepoolte Parameter $\hat{\theta}_{MI,k}$ mit $k = 1,\ldots,K$, aus dessen geordneter Menge die Quantile wie in Methode 1 und 3 die Grenzen der Konfidenzintervalle bilden.

Die Attraktivität dieser Methoden liegt vor allem darin, dass sie leicht zu implementieren sind und von dem statistischen Analysemodell nur die geschätzten Parameter benötigt werden. Für alle vier Methoden werden keine analytisch berechneten Standardfehler genutzt und in drei von vier Fällen werden die Konfidenzintervalle über schlichte Quantilsfunktionen bestimmt. Somit ist für diese Fälle keine Annahme über die Verteilung der geschätzten Parameter nötig. Die Qualität der in Kapitel 2 diskutierten Verfahren hängt alleine davon ab, das richtige Set an Variablen auszuwählen und für dieses die Parameter so unverzerrt wie möglich zu schätzen. Hierbei ist theoretisch zu erwarten, dass das Lasso und adaptive Lasso eher schlecht abschneiden, da für beide Verfahren die Koeffizienten aufgrund der $l_1$ Penalisierung verzerrt sind. Das MSS, LPR oder LP Verfahren sollten hier einen Vorteil besitzen, da diese entwe-

[2] Diese Methode ist aufgrund der Anwendung der pooling Regeln nur nutzbar, wenn die Parameter normalverteilt sind und ist daher für Modelle gedacht, für die keine leicht zugänglichen Standardfehler vorhanden sind.

der eine OLS-Regression zur Schätzung nutzen oder eine Korrektur der Verzerrung vornehmen.

Für die in Kapitel 2 beschriebenen Verfahren können daher die Methoden 1, 3 und 4 angewandt werden. Methode 2 kann hingegen aufgrund der Nutzung der pooling Regeln nur für Modelle mit normalverteilten Koeffizienten genutzt werden.

# 4 Aufbau der Simulationsstudie

Wir betrachten für die folgende Simulationsstudie einen Datensatz aus 200 Beobachtungen ($n$) und 39 Prädiktoren ($p$), bei der die abhängige Variable mithilfe eines linearen Modells gebildet wird:

$$Y = X\beta + \epsilon,\ \epsilon \overset{iid}{\sim} N(0,\sigma^2 I) \tag{4.1}$$

Die hierfür benötigten Parameter $X$, $\beta$ und $\sigma$ werden in den nachfolgenden Kapiteln definiert.

## 4.1 Konstruktion der Designmatrix

Die Kovarianzmatrix der Designmatrix $X$ wird folgendermaßen konstruiert: Es werden in einem ersten Schritt zuerst drei kleinere Kovarianzmatrizen $\Sigma^A$, $\Sigma^B$ und $\Sigma^C$ der Dimension $p/3 \times p/3 = 13 \times 13$ mit jeweils gleicher Kovarianz und einer Varianz von eins erzeugt. In Matrix $\Sigma^A$ liegt eine Kovarianz zwischen allen Variablen von 0.8 in $\Sigma^B$ von 0.5 und in $\Sigma^C$ von 0.2 vor. Somit hat beispielsweise Matrix $\Sigma^A$ folgende Form:

$$\Sigma^A_{ij} = 0.8\ \forall\ i \neq j \wedge \Sigma^A_{ij} = 1\ \forall\ i = j \tag{4.2}$$

Diese Submatrizen werden nun in eine Blockmatrix eingefügt:

$$\Sigma = \begin{pmatrix} \Sigma^A & 0 & 0 \\ 0 & \Sigma^B & 0 \\ 0 & 0 & \Sigma^C \end{pmatrix} \tag{4.3}$$

wobei, die 0-er auf den Nebendiagonalen ebenfalls Matrizen der Dimension $13 \times 13$ sind. Da alle Varianzen der Variablen auf eins gesetzt wurden, stellt diese Kovarianzmatrix gleichzeitig die Korrelationsmatrix dar. Es gibt somit drei Blöcke an Variablen, welche im Folgenden als $A$, $B$ und $C$ bezeichnet werden, die miteinander unterschiedlich stark korreliert sind, zu den Variablen der anderen Blöcke allerdings keine Korrelation aufweisen. Der Vorteil einer solchen Aufteilung liegt darin, dass untersucht werden kann, wie sich unterschiedliche Korrelationen zum Beispiel auf die Breite der Konfidenzintervalle auswirken. Die Prädiktoren können nun mithilfe der

Funktion 'mvrnorm' des R-Paketes 'MASS' (Venables und Ripley, 2017) von einer multivariaten Normalverteilung mit $N_p(0,\Sigma)$ erzeugt werden.

## 4.2 Festlegung der Inferenzmethode, des Beta-Vektors und der Varianz des Störterms

Insgesamt werden acht verschiedene Settings kreiert, die sich hinsichtlich der angewandten Inferenzmethode, des wahren Beta-Vektors und der Varianz des Störterms unterscheiden.
In dieser Simulationsstudie werden die Methoden MI-Boot (PS) und Boot-MI aus Kapitel 3.2 verwendet, um die entsprechenden 95%-Konfidenzintervalle zu kreieren, da diese auf alle Verfahren aus Kapitel 2 angewandt werden können. Hierdurch wird eine direkte Vergleichbarkeit der statistischen Modelle möglich, da Unterschiede in den Ergebnissen alleine auf diese zurückzuführen sind. Theoretisch hätte auch die Methode Boot-MI (PS) genutzt werden können, laut Schomaker und Heumann (2016) ist Boot-MI jedoch typischerweise effizienter und somit vorzuziehen.
Für den wahren Beta-Vektor werden ebenfalls zwei unterschiedliche Settings in Betracht gezogen:

- Der Beta-Vektor hat für alle Blöcke $A$, $B$ und $C$ die folgende Form:

  $$\beta_A = \beta_B = \beta_C = (20,\ 15,\ 10,\ 5,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)^T$$

  Somit gilt $\beta = \left(\beta_A^T,\ \beta_B^T,\ \beta_C^T\right)^T$. Dies entspricht einer einfachen Beta-Situation. Die meisten Variablen (27) haben keinen Einfluss auf die Zielgröße und können somit von dem Lasso im Sinne einer Variablenselektion auf null gesetzt werden, während eine übersichtliche Auswahl von zwölf Variablen recht hohe Koeffizienten aufweist, welche es zu schätzen gilt. Für die weitere Simulationsstudie wird dieser Beta-Vektor als $\beta_{\text{spärlich}}$ bezeichnet. In diesem Fall hält die Beta-Min Annahme, da $5 \gg 1/\sqrt{n} = 0.071$.

- Der Beta-Vektor hat für alle Blöcke $A$, $B$ und $C$ die folgende Form:

  $$\beta_A = \beta_B = \beta_C = (20,\ 15,\ 10,\ 1,\ 0.5,\ 0.1,\ 0.08,\ 0.06,\ 0.04,\ 0.02,\ 0.01,\ 0.005,\ 0)^T$$

  Somit gilt $\beta = \left(\beta_A^T,\ \beta_B^T,\ \beta_C^T\right)^T$. Dies entspricht einem schwierigen Setting für Lasso-Regressionen. Fast alle Regressionskoeffizienten sind somit von null ver-

schieden und teilen sich zudem in zwei Gruppen aus relativ großen und aus kleinen Werten, die gegen null gehen, auf. Somit entsteht eine Art Kliff zwischen beiden Gruppen. Für die weitere Simulationsstudie wird dieser Beta-Vektor als $\beta_{\text{Kliff}}$ bezeichnet. In diesem Fall ist die Beta-Min Annahme verletzt, da $0.005 < 1/\sqrt{n} = 0.071$. Dieses Setting wurde gewählt, um zu untersuchen, ob die angewandten Verfahren auch kleine Koeffizienten entdecken können, oder ob ab einer bestimmten Schwelle die Coverageraten sinken.

Zusätzlich wird die Varianz des Störterms in Gleichung 4.1 variiert, sodass das Verhältnis aus Signal und Störung (eng: Signal-Noise-Ratio (SNR)) den Wert 5 oder 10 ergibt (Liu et al., 2017a):

$$\text{SNR} = \|X\beta\|_2^2/n\sigma^2 \tag{4.4}$$

Somit entstehen insgesamt $2 \times 2 \times 2 = 8$ unterschiedliche Settings für jedes Modell aus Kapitel 2. Eine kompakte Darstellung dieser ist in Tabelle 1 gegeben.
Wie bereits beschrieben, kann nun die abhängige Variable über Gleichung 4.1 gebildet werden. Es ist leicht zu sehen, dass die Werte von $y$ unter den verschiedenen Settings variieren, da diese sowohl abhängig von der SNR als auch von dem gewählten Beta-Vektor sind. Die vollständige Designmatrix $X$ bleibt hingegen konstant.

| Setting | Inferenzmethode | Beta | SNR |
|---|---|---|---|
| 1 | MI-Boot (PS) | $\beta_{\text{spärlich}}$ | 5 |
| 2 | MI-Boot (PS) | $\beta_{\text{spärlich}}$ | 10 |
| 3 | MI-Boot (PS) | $\beta_{\text{Kliff}}$ | 5 |
| 4 | MI-Boot (PS) | $\beta_{\text{Kliff}}$ | 10 |
| 5 | Boot-MI | $\beta_{\text{spärlich}}$ | 5 |
| 6 | Boot-MI | $\beta_{\text{spärlich}}$ | 10 |
| 7 | Boot-MI | $\beta_{\text{Kliff}}$ | 5 |
| 8 | Boot-MI | $\beta_{\text{Kliff}}$ | 10 |

Tabelle 1: Auflistung der acht Settings, welche in der Simulationsstudie berücksichtigt werden in Bezug auf die Inferenzmethode, den wahren Beta-Vektor und der Signal-Noise-Ratio. Jedes statistische Verfahren aus Kapitel 2 wird auf jedes Setting angewandt.

## 4.3 Erzeugung fehlender Werte

Nachdem für ein beliebiges Setting die $y$-Werte berechnet wurden, werden die Wahrscheinlichkeiten für fehlende Werte in ausgewählten Prädiktoren über folgende Funktion erzeugt:

$$\pi_{\text{fehlend}}(y) = \left(1 - \frac{1}{(\delta \cdot y)^2}\right) \kappa \qquad \text{mit } \delta \neq 0 \text{ und } \kappa \leq 1 \tag{4.5}$$

Somit liegt ein missing at random Mechanismus vor, da die Wahrscheinlichkeiten für fehlende Werte in einer unabhängigen Variable von den beobachtbaren $y$-Werten abhängt. Die Parameter $\delta$ und $\kappa$ steuern dabei den Anteil der fehlenden Werte. Während $\delta$ die Struktur der Kurve ändern kann, ist leicht erkennbar, dass $\kappa$ die entstehende Kurve nur staucht ($\kappa < 1$) oder beibehält ($\kappa = 1$). Für $\delta = 0.01$ und $\kappa = 1$ sind die Wahrscheinlichkeiten für fehlende Werte eines beliebigen Prädiktors in Abbildung 1 dargestellt.

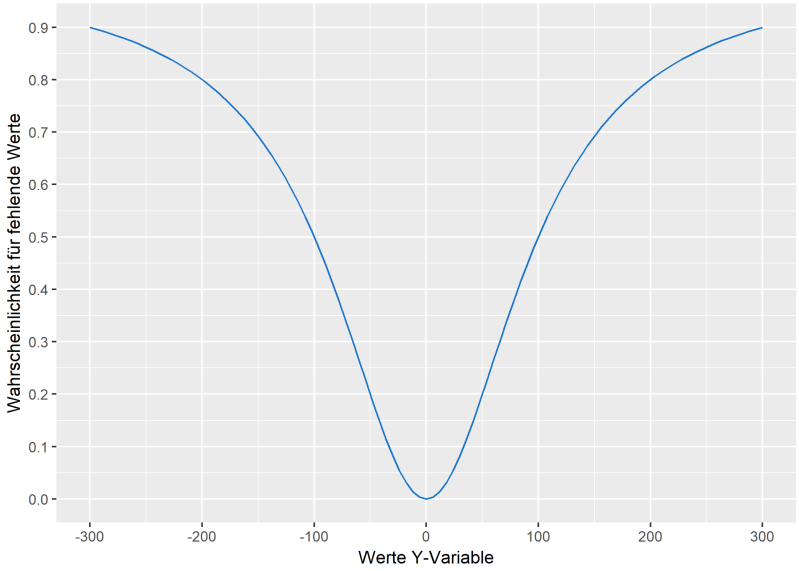


Abbildung 1: Beispielhafte Darstellung der Funktion aus Gleichung (4.5) zur Erzeugung von fehlenden Werten in einer unabhängigen Variable für $\delta = 0.01$ und $\kappa = 1$

Sowohl für hohe als auch niedrige Werte in $y$ steigen somit die Wahrscheinlichkeiten für fehlende Werte. Die Designmatrix enthält für $\beta = \beta_{\text{spärlich}}$ neun Variablen mit unvollständigen Einträgen und für $\beta = \beta_{\text{Kliff}}$ zwölf. Hierbei liegt eine gewisse Symmetrie zwischen den Variablen der Blöcke $A$, $B$ und $C$ vor, um auch hier eine bessere

Vergleichbarkeit zu gewährleisten. Enthält beispielsweise die erste Variable in Block $A$ fehlende Werte, gilt dies auch für die erste Variable in Block $B$ und $C$.
Der Parameter $\delta$ liegt hierbei über alle Settings hinweg zwischen 0.008 und 0.013 und $\kappa$ zwischen 0.8 und 1. Hierdurch entsteht etwas Variation in der Ausfallrate. Die Parameter $\delta$ und $\kappa$ wurden dabei so gewählt, dass ungefähr 15% bis 35% der Beobachtungen einer Variable fehlen. Welche Variablen unvollständig sind, als auch die exakten Werte für $\delta$ und $\kappa$ können Tabelle 2 entnommen werden.

| Beta-spärlich | Variablen mit fehlenden Werten | $\delta$ | $\kappa$ |
|---|---|---|---|
| | $X_1$, $X_{14}$, $X_{27}$ | 0.013 | 0.8 |
| | $X_3$, $X_{16}$, $X_{29}$ | 0.009 | 0.9 |
| | $X_5$, $X_{18}$, $X_{31}$ | 0.011 | 0.85 |
| Beta-Kliff | | | |
| | $X_2$, $X_{15}$, $X_{28}$ | 0.01 | 1 |
| | $X_5$, $X_{18}$, $X_{31}$ | 0.008 | 0.85 |
| | $X_8$, $X_{21}$, $X_{34}$ | 0.009 | 0.9 |
| | $X_{11}$, $X_{24}$, $X_{37}$ | 0.009 | 0.8 |

Tabelle 2: Variablen mit fehlenden Werten für $\beta = \beta_{\text{spärlich}}$ und $\beta = \beta_{\text{Kliff}}$. Die Parameter $\delta$ und $\kappa$ beziehen sich auf Gleichung 4.5, mit der fehlende Werte erzeugt werden.

Die multiple Imputation erfolgt stets mit dem R-Paket 'mice' (van Buuren und Groothuis-Oudshoorn, 2011).
An dieser Stelle soll erwähnt werden, dass die gängigen R-Pakete zur multiplen Imputation (darunter auch 'mice') hauptsächlich für niedrig dimensionale Datensätze geeignet sind (Zhao und Long, 2016). Für hochdimensionale Settings kann beispielsweise die Erweiterung 'miceadds' (Robitzsch et al., 2017) verwendet werden, bei der eine Partial-Least-Squares (PLS) Regression zur multiplen Imputation genutzt wird. Weitere Verfahren für hochdimensionale Datensätze, welche auf 'mice' aufbauen, werden in Deng et al. (2016) vorgestellt.

## 4.4 Gütekriterien und Parameterspezifikationen für die statistischen Modelle

Um für die Prädiktoren sowohl Coverageraten als auch die Median Breite der 95%-Konfidenzintervalle zu berechnen, wird $Y$ für $S = 500$ Monte-Carlo-Iterationen durch die Simulation unabhängiger Fehlerterme $\epsilon$ über Gleichung 4.1 neu erzeugt.

Es wurde sich für den Median entschieden, da dieser robust gegenüber Ausreißern ist. Zudem werden die mittlere quadratische Abweichung der Coverageraten von dem gewünschten 95%-Niveau, als auch die Verteilung der Koeffizientenschätzer über alle $S$ Iterationen als Gütekriterien herangezogen. Die Ergebnisse der Simulationsstudie werden in Kapitel 5 vorgestellt. In Bezug auf die in Kapitel 3.2 beschriebenen Inferenzmethoden wird die Anzahl der multiplen Imputationen auf $M = 10$ und die Bootstrapziehungen auf $K = 200$ gesetzt.

Für das adaptive Lasso wird der zusätzliche Gewichtungsfaktor $\hat{\omega}_j$ aus Gleichung 2.6 mit den Koeffizienten einer Ridge-Regression bestimmt. Im Falle dieser Simulation mit $p < n$ hätte man ebenfalls eine OLS-Regression nutzen können, da die Verfahren aber in den hochdimensionalen Raum übertragbar sein sollen, wurde sich für die Ridge-Regression entschieden. Zudem wird $\gamma$ auf eins gesetzt. Die Schätzung erfolgt mit dem R-Paket 'glmnet' (Friedman et al., 2010).

Das in Kapitel 2 vorgestellte MSS-Verfahren ist zwar in dem R-Paket 'hdi' (Dezeure et al., 2015) implementiert, die entsprechende Funktion erzeugt allerdings nur Konfidenzintervalle und die entsprechenden p-Werte. Daher wurde eine eigene Funktion geschrieben, welche das Vorgehen der MSS-Prozedur nachahmt, aber Regressionskoeffizienten ausgibt. Der ursprüngliche Datensatz $D$ wird $R$-mal geteilt und auf der jeweiligen Hälfte eine Lasso- und daraufhin eine OLS-Regression mit dem Set der ausgewählten Variablen berechnet. Dies erzeugt für jeden Prädiktor $R$ Regressionskoeffizienten[3], welche gemittelt werden können, um einen Parametervektor der Länge $p$ zu erhalten. Da dieses Verfahren sehr rechenintensiv ist, wurde sich für $R = 25$ entschieden. Dies lieferte in verschiedenen Settings ähnliche Ergebnisse zu dem Vorschlag der Autoren, $R$ auf 50 zu setzen (Dezeure et al., 2015).

Für den LPR-Schätzer wird $\lambda_2$ aus Gleichung 2.11, wie empfohlen, auf $1/n$ gesetzt (Liu et al., 2017a). Das $\lambda$ der vorangehenden Lasso-Regressionen zur Variablenselektion wird hingegen wie üblich mit Kreuzvalidierung bestimmt. Zur Berechnung dieses Modells wird das R-Paket 'HDCI' (Liu et al., 2017b) genutzt.

Für die einfache Lasso-Regression wird das R-Paket 'glmnet' (Friedman et al., 2010) und für den LP-Schätzer das Paket 'hdi' (Dezeure et al., 2015) verwendet. Ansonsten

[3] Nicht ausgewählten Variablen im Lasso-Schritt wird wie bisher ein Koeffizient von null zugewiesen.

gibt es zu beiden Modellen keine Besonderheiten der Parameter oder der Vorgehensweise, die erwähnt werden müssten.

Alle aufgeführten Pakete sind unter der R-Version 3.4.3 verfügbar.

## 4.5 Abschließende Motivation

In manchen Aspekten wurde die Simulationsstudie bewusst recht einfach gehalten. Der Anteil der fehlenden Werte ist moderat und da $p < n$ gilt, befinden wir uns in einem eher niedrig dimensionalen Setting. Alle in Kapitel 2 vorgestellten Methoden können jedoch auch mit hochdimensionalen Daten umgehen und wurden teilweise im Hinblick auf solche Situationen konstruiert. Warum wird also ein Simulationssetting betrachtet, welches niedrigdimensional ist? Wie bereits beschrieben, gehen viele Autoren nicht auf die Problematik fehlender Werte ein. Für einfache Variablenselektionsmodelle wurden inzwischen zwar einige Lösungsansätze für multipel imputierte Datensätze vorgeschlagen (siehe zum Beispiel Schomaker und Heumann (2014), Chen und Wang (2013) oder Wood et al. (2008) ), bei einer zusätzlichen Betrachtung der Inferenz sieht die Lage jedoch eher mager aus. Daher soll in dieser Arbeit vor allem ein Grundstein gelegt werden, diese Lücke zu schließen. Denn liefern die in Kapitel 2 vorgestellten statistischen Methoden in Kombination mit den Bootstrapverfahren von Schomaker und Heumann (2016) schlechte Ergebnisse, ist es wahrscheinlich, dass dies auch für komplexere hochdimensionale Settings gilt. Schneiden die Verfahren hingegen gut ab, können die in dieser Arbeit vorgestellten Methoden für weitere Simulationsstudien genutzt werden. Das Verhältnis zwischen Beobachtungen und Variablen ist hingegen recht klein (ca. 5:1) und liegt hierdurch weit unter der geforderten 20:1 Regel. In solchen Situationen sind aus statistischer Sicht penalisierte Verfahren einer einfachen OLS-Regression vorzuziehen. Möchte man für die geschätzten Parameter allerdings neben der Schätzung der Koeffizienten auch Inferenz betreiben und es liegen zusätzlich fehlende Werte vor, stößt das einfache Lasso an seine Grenzen. Als Lösung können die in dieser Arbeit vorgestellten Inferenzmethoden in Kombination mit den statistischen Modellen verwendet werden. Somit haben diese nicht nur in hochdimensionalen Settings einen Nutzen und es lohnt sich die Güte dieser Verfahren für den Fall $p < n$ zu untersuchen. Zudem ist die Interpretierbarkeit und Darstellung der Coverageraten für die einzelnen

Variablen mit $p = 39$ noch überschaubar und daher für einen ersten Einblick gut geeignet. Für hochdimensionale Simulationsstudien bei denen die Designmatrix $X$ jedoch aus hunderten Variablen besteht, gelingt dies nicht mehr. Vor allem in Bezug auf die Variablen der Blöcke $A$, $B$ und $C$ können so Unterschiede noch anschaulich hervorgehoben werden.

Eine mögliche Schwierigkeit stellt hingegen die Struktur der Kovarianzmatrix $\Sigma$ dar. Hier lassen sich zwei gegenläufige theoretische Effekte vermuten, die sich auf die Größe der Konfidenzintervalle auswirken. Die Qualität imputierter Werte erhöht sich, wenn in dem Imputationsmodell weitere Variablen vorhanden sind, die hoch mit der zu imputierenden Variable korreliert sind (Enders, 2010, Kapitel 5). Für den Extremfall, dass zwei Variablen eine wahre Korrelation von eins aufweisen, aber nur eine Variable fehlende Werte enthält, lassen sich beispielsweise perfekte Imputationen erzeugen. Hiervon profitieren vor allem die Variablen aus Block $A$ und $B$. Denn je besser die imputierten Werte, desto höher sollte auch die Güte der angewandten statistischen Methoden sein. Auf der anderen Seite liegt besonders für die Variablen in Block $A$ eine hohe Multikollinearität vor, mit der die statistischen Modelle umgehen müssen. Dies könnte dazu führen, dass nicht-relevante Variablen in den Lasso-Schritten der Verfahren ausgewählt werden (Zou und Hastie, 2005).

# 5 Ergebnisse der Simulationsstudie

Im Folgenden werden beispielhaft die Ergebnisse für das Setting 7 aus Tabelle 1 mithilfe der entsprechenden Abbildungen vorgestellt. Anschließend findet eine Aggregation der Ergebnisse für die Coverageraten und die Median Konfidenzintervallbreite über alle Settings hinweg statt.

## 5.1 Vorstellung der Ergebnisse für ein ausgewähltes Setting

Wir beginnen mit den Coverage Wahrscheinlichkeiten für das Setting 7, welche in Abbildung 2 dargestellt sind. Für jedes statistische Modell aus Kapitel 2 wurden die Anteile der 500 Monte-Carlo-Iterationen berechnet, bei denen das Konfidenzintervall den wahren Beta-Wert enthält. Die Variablen der Gruppen $A$, $B$ und $C$ sind durch verschiedene Farben gekennzeichnet (siehe Legende). Variablen mit fehlenden Werten sind als Rauten dargestellt. Zusätzlich ist in jedem Plot das Maximum und das Minimum der Coverage Wahrscheinlichkeiten eingetragen. Optimal wäre es, wenn sich alle Punkte auf dem 0.95 Level befinden würden. Man darf allerdings nicht vergessen, dass die Coverage Wahrscheinlichkeiten stochastische Elemente und somit dem Zufall unterworfen sind. Nehmen wir an, ein Verfahren besitzt eine wahre Coverage Wahrscheinlichkeit von 0.95. Da diese nichts weiter als ein Anteil ($\pi$) ist, können wir leicht ein Konfidenzintervall mit $\alpha = 0.05$ kreieren, welches die Abweichung angibt, die von dem Monte-Carlo-Fehler erklärt werden kann:

$$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{S}} \approx [0.93, 0.97] \tag{5.1}$$

mit $\hat{\pi} = 0.95$, $S = 500$ und dem Quantil der Standardnormalverteilung $z_{0.975} \approx 1.96$. Liegen die Anteile in diesem Intervall, kann dies nicht als Fehler der Verfahren angesehen werden. In Abbildung 2 sieht man beispielsweise, dass für die meisten Variablen die Coverage Wahrscheinlichkeiten recht nah an der gewünschten 0.95 liegen.
Über das Minimum sehen wir, dass kein Verfahren für auch nur eine einzige Variable Werte unter der 0.93 Marke aufweist. Allerdings erkennt man recht deutlich, dass für alle Modelle die Coverage Wahrscheinlichkeiten meist etwas über dem 0.95 Niveau liegen. Und obwohl, wie beschrieben, geringe Abweichungen auf den Monte-Carlo-Fehler zurückzuführen sind, können trotzdem allgemeine Trends der Inferenzmetho-

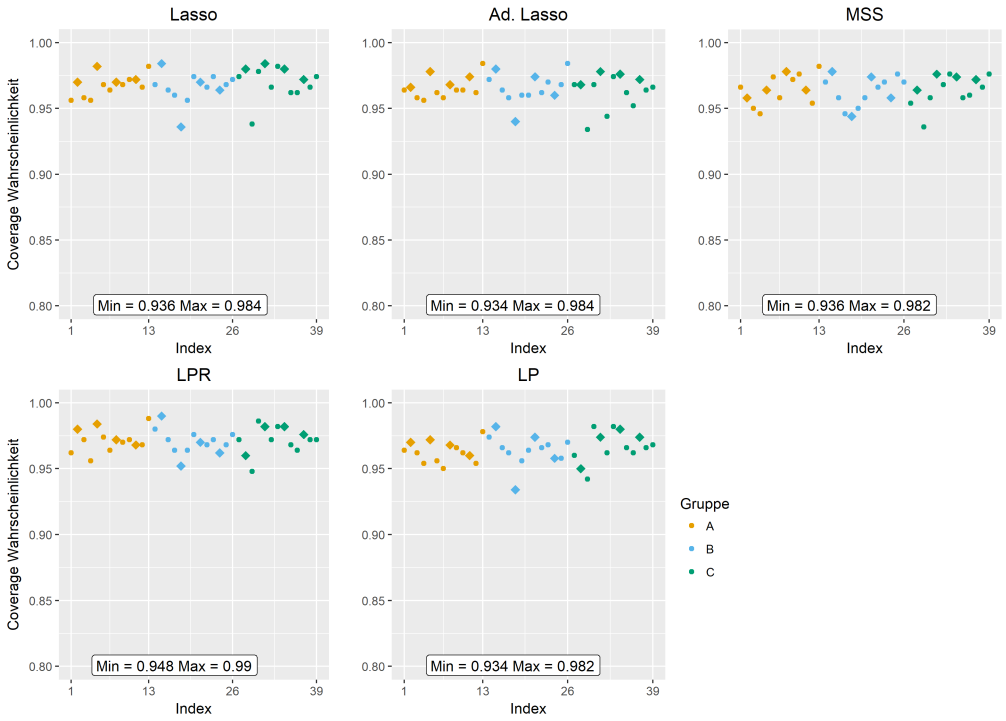


Abbildung 2: Coverage Wahrscheinlichkeiten der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

den und Modelle herausgelesen werden, wenn die Wahrscheinlichkeiten konsistent über- oder unterschätzt werden. Am besten schneiden in dieser Kategorie der MSS und der LP-Schätzer mit einem Maximum von 0.982 ab. Vergleicht man die Werte hinsichtlich der Gruppen $A$, $B$ und $C$, lassen sich kaum Unterschiede feststellen. Die Korrelation scheint somit keinen Einfluss auf die Coverage Wahrscheinlichkeiten zu haben. Selbiges gilt für die Variablen mit fehlenden Werten. Auch im Hinblick auf die wahren Koeffizienten gibt es keinen offensichtlichen Zusammenhang. Variablen mit großen als auch extrem kleinen Koeffizienten besitzen ähnliche Coverageraten.

Die Median Breite der $S = 500$ Konfidenzintervalle ist in Abbildung 3 dargestellt. Im Gegensatz zu den Coverage Wahrscheinlichkeiten sieht man sofort, dass sowohl zwischen den Gruppen $A$, $B$ und $C$ als auch in Bezug, ob eine Variable vollständig ist oder nicht, große Unterschiede bestehen. Für Variablen mit hoher Multikollinearität ergeben sich breitere Konfidenzintervalle. Dies deckt sich mit der Überlegung aus Kapitel 4, dass für hohe Korrelationen die Verfahren möglicherweise Probleme haben, die richtigen Variablen auszuwählen und sich diese zusätzliche Unsicherheit

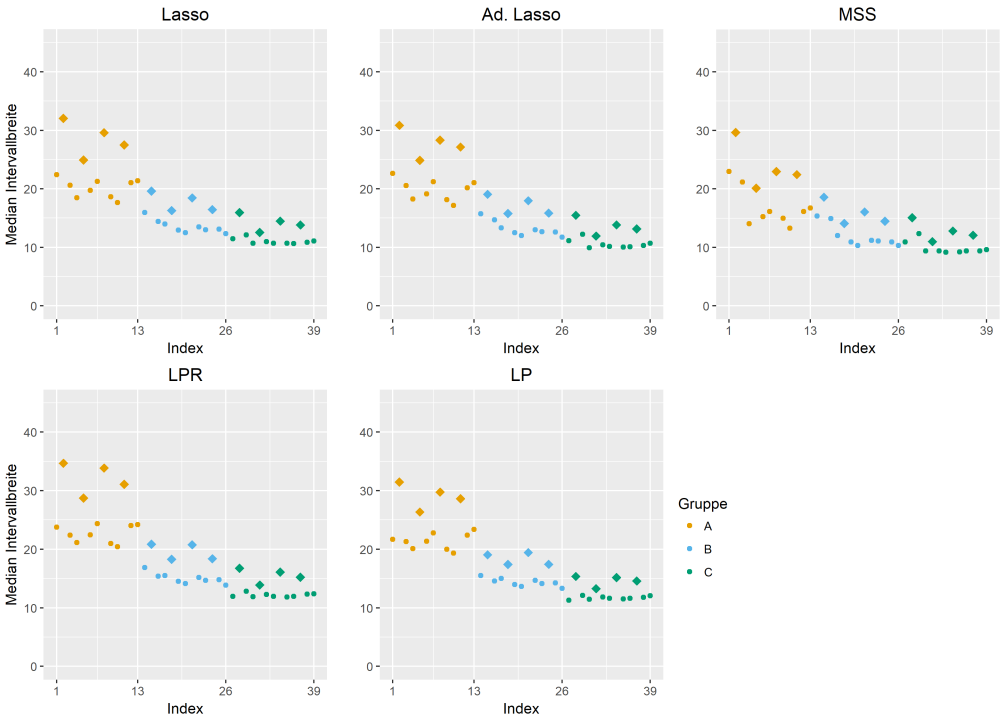


Abbildung 3: Median Konfidenzintervallbreite der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

auf die Größe der Intervalle auswirkt. Selbiges gilt für Variablen mit fehlenden Werten. Diese sind von Grund auf mit einer erhöhten Unsicherheit verbunden, welche zwar durch die multiple Imputation berücksichtigt, aber nicht komplett aufgehoben werden kann. Bei einem Vergleich der Verfahren untereinander besitzt das MSS in diesem Setting die kleinsten Konfidenzintervalle für die meisten Variablen. Wie bei den Coverageraten lassen sich keine großen Unterschiede für die Median Breite der Intervalle in Bezug auf die Größe des wahren Koeffizienten feststellen.

In Abbildung 4 ist die Median-Breite der Konfidenzintervalle gegen die Coverage Wahrscheinlichkeit abgetragen. Hier sieht man nochmals anschaulich, dass Variablen mit hohen paarweisen Korrelationen und fehlenden Werten die größten Konfidenzintervalle aufweisen, die Coverageraten hingegen recht stabil bleiben. Die exakten Werte für beide Maße sind in Anhang A Tabelle 11 zusammengefasst.

Neben den bisher vorgestellten Gütekriterien ist es auch möglich die Verteilung der geschätzten Koeffizienten mit Boxplots zu visualisieren (siehe hierfür Abbildung 5).

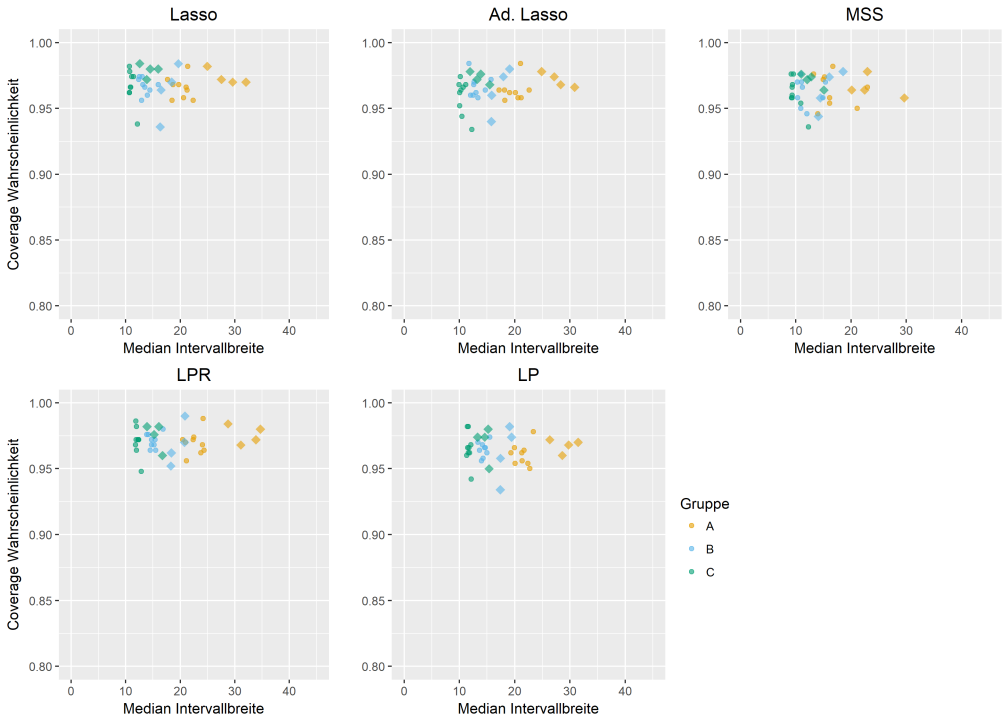


Abbildung 4: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt

Hierbei sei darauf hingewiesen, dass jeder Boxplot auf der Verteilung der Koeffizienten beruht, auf dessen Menge über die Quantile die Konfidenzintervalle konstruiert wurden. In dem MI-Boot (PS) Fall stellt somit ein Boxplot die Verteilung der $S \times K \times M = 1.000.000$ Punktschätzer einer Variable dar [4]. Erinnern wir uns hingegen an die Vorgehensweise der Boot-MI Methode zurück, so wird für jede Bootstrapziehung der Datensatz $M$-mal imputiert und diese $M$ Punktschätzer jeweils gemittelt. Hieraus entstehen über alle Monte-Carlo-Iterationen hinweg für jeden Prädiktor nur $S \times K = 100.000$ geschätzte Koeffizienten. Da in dem hier vorgestellten Setting 7 die Boot-MI Methode angewandt wurde, repräsentiert jeder einzelne Boxplot in Abbildung 5 die Verteilung der entsprechenden 100.000 Datenpunkte. Die vertikalen Linien trennen hierbei die Variablen der Blöcke $A$, $B$ und $C$, während die Punkte innerhalb der Boxplots den wahren Beta-Wert darstellen. Die Box entspricht, wie üblich, dem Bereich, in dem die mittleren 50% der Daten liegen. Die Whiskers

[4] mit $S = 500$, $K = 200$ und $M = 10$.

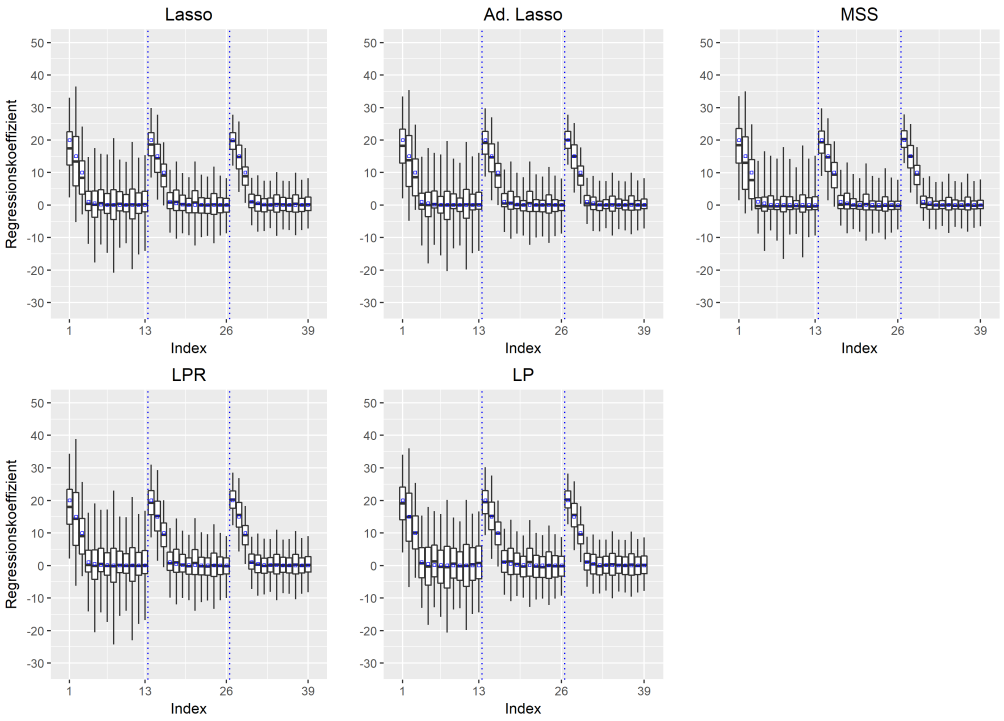


Abbildung 5: Boxplots für die Verteilung der Parameter-Schätzer der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-partial-Ridge und Lasso-Projektion. Whiskers repräsentieren das 2.5% und das 97.5% Quantil. Ausreißer, die außerhalb des Whiskers liegen, sind nicht abgetragen. Die vertikalen Linien trennen die Variablen der Blöcke $A$, $B$ und $C$. Die wahren Beta-Werte sind als blaue Punkte eingezeichnet.

hingegen stellen, statt des 1.5-fachen Interquartilsabstandes, das 2.5% und 97.5% Quantil dar. Ausreißer, welche kleiner oder größer als diese Quantile sind, wurden der Übersichtlichkeit halber nicht eingezeichnet. Es ist leicht erkennbar, dass für alle dargestellten Boxplots der wahre Beta-Wert von dem unteren und oberen Quartil eingeschlossen wird. In den meisten Fällen liegt dieser sogar sehr nahe an dem Median der entsprechenden Verteilung. Die größte Variabilität in den Verteilungen der geschätzten Koeffizienten findet sich für Variablen mit hoher Multikollinearität und fehlenden Werten. Diese Erkenntnisse decken sich mit den bisherigen Beobachtungen aus Abbildung 2, 3 und 4. Die Abbildungen der bisher nicht betrachteten Settings sind in Anhang B zu finden.

## 5.2 Coverageraten: Mittlere quadratische Abweichung und Vergleich zwischen den Inferenzmethoden

Als weiteres Gütekriterium kann die mittlere quadratische Abweichung (MSE) der Coverageraten von dem gewünschten 95%-Niveau herangezogen werden, welche für alle Settings und Modelle in Tabelle 3 dargestellt ist.

| Setting | Lasso | Ad. Lasso | MSS | LPR | LP |
|---|---|---|---|---|---|
| 1 | 10.44 | 13.91 | 9.26 | **2.85** | 5.20 |
| 2 | 8.84 | 14.53 | 5.86 | **5.00** | 6.74 |
| 3 | 8.34 | 45.40 | 8.54 | 3.29 | **1.76** |
| 4 | 7.66 | 48.81 | 7.69 | 2.95 | **2.47** |
| 5 | 6.08 | 5.70 | 5.87 | 7.08 | **4.32** |
| 6 | 6.64 | 6.50 | 6.83 | 7.29 | **5.39** |
| 7 | 4.41 | 3.38 | **3.18** | 5.38 | 3.19 |
| 8 | 3.82 | 2.59 | 2.93 | 4.32 | **2.49** |
| Mittelwert 1-4 | 8.82 | 30.66 | 7.84 | **3.52** | 4.04 |
| Mittelwert 5-8 | 5.24 | 4.54 | 4.70 | 6.02 | **3.85** |

Tabelle 3: Mittlerer quadratische Abweichung der Coverageraten zu dem 95%-Niveau. Der niedrigste MSE aller Verfahren für ein Setting ist hervorgehoben und Mittelwerte wurden für die MI-Boot (PS) Settings (1-4) und die Boot-MI (5-8) berechnet.

Wie bereits beschrieben können kleine Abweichungen durch den Monte-Carlo-Fehler erklärt werden. Bei einem direkten Vergleich der Verfahren ist allerdings zu erwarten, dass dieser konstant ist und große Unterschiede alleine auf die Modelle und Inferenzverfahren zurückzuführen sind. Dies ist zum Beispiel für das adaptive Lasso der Fall, welches für alle MI-Boot-Settings $(1, 2, 3, 4)$ die größten quadratischen Abweichungen von dem 95%-Niveau aufweist. Für $\beta = \beta_{\text{Kliff}}$ (Setting 3 und 4) explodiert der MSE mit Werten von 45.40 und 48.81 regelrecht. Für die Boot-MI-Settings $(5, 6, 7, 8)$ erzielt das adaptive Lasso hingegen recht gute Coverageraten, was sich in kleinen MSE-Werten äußert. Eine ähnliche Struktur lässt, sich auch für das Lasso feststellen, bei dem für die Boot-MI Settings konsistent kleinere MSE erzielt werden. Für die anderen Verfahren (MSS, LPR und LP) gelten diese Unterschiede in Bezug auf die Inferenzmethode jedoch nicht[5]. Über alle Settings hinweg erzielt vor allem die Lasso-Projektion und das Lasso-Partial-Ridge-Modell konsistent gute

[5] Setting 1 zu 5, 2 zu 6, 3 zu 7 und 4 zu 8 unterscheiden sich alleine durch die angewandte Inferenzmethode und können in Bezug auf diesen Parameter direkt verglichen werden.

Ergebnisse. Interessanterweise sind für die Lasso-Projektion bei dem schwierigeren $\beta = \beta_{\text{Kliff}}$ (Setting 3, 4, 7 und 8) bei Konstanthaltung der anderen Parameter die MSE konsistent geringer als bei $\beta = \beta_{\text{spärlich}}$. Das MSS-Verfahren hingegen schneidet unter MI-Boot (PS) mittelmäßig ab, da vor allem für die Settings 1, 3 und 4 der MSE im Vergleich zu dem der LPR und LP erhöht ist. Bei Anwendung der Boot-MI Inferenzmethode sind die MSE hingegen recht klein.

Für die Coverageraten aller Settings kann man zusammenfassen, dass diese für die Boot-MI Inferenzmethode recht robust gegenüber dem wahren Beta-Vektor, der SNR, der Korrelation und Variablen mit fehlenden Werten sind. Allerdings liegen über alle Modelle hinweg meist leicht erhöhte Coverageraten in Bezug auf das gewünschte 95% Niveau vor.
Für die MI-Boot (PS) Inferenzmethode werden diese für den LPR-Schätzer und die LP in der Regel etwas unterschätzt, liegen jedoch meist noch in einem akzeptablen Bereich. Das einfache Lasso und das adaptive Lasso hingegen haben für beide Beta-Vektoren Probleme. Die Coverageraten werden hier für Koeffizienten ungleich null unter- und für Koeffizienten gleich null überschätzt. Maxima von $> 99\%$ und Minima $< 90\%$ sind hierbei vor allem für das adaptive Lasso die Regel anstatt die Ausnahme. Besonders bei $\beta = \beta_{\text{Kliff}}$ erzielt das adaptive Lasso für Variablen mit kleinen Regressionskoeffizienten extrem niedrige Coverage Anteile von teilweise nur 81% (siehe beispielsweise Abbildung 18 in Anhang B). Für das MSS-Modell liegt eine ähnliche Problematik mit dem Unterschied vor, dass die Coverageraten der kleinen Koeffizienten von Beta-Kliff im Vergleich zu dem Lasso und dem adaptiven Lasso eher über- statt unterschätzt werden.

## 5.3 Median Konfidenzintervallbreite: Lineare Regression

Auch für die Median Breite der Konfidenzintervalle kann es sinnvoll sein, diese noch einmal aggregiert darzustellen. Schließlich sind über alle Settings 1560 von diesen vorhanden[6]. Hierbei bietet sich eine lineare Regression an. Als Prädiktoren werden die statistischen Modelle (Referenz: Lasso), die Signal-Noise-Ratio (Referenz: SNR = 10), der wahre Beta-Vektor (Referenz: $\beta_{\text{spärlich}}$), die Inferenzmethode (Referenz: Boot-

[6] 39 Variablen für jedes der fünf statistischen Modelle für jedes der acht Settings ergibt 1580 Variablen und somit genauso oft das Gütekriterium Median Konfidenzintervallbreite.

MI), als auch in welchem Korrelationsblock sich die entsprechende Variable befindet (Referenz: Korrelation 0.2) und ob sie fehlende Werte enthält (Referenz: vollständig) verwendet. Zusätzlich wird die Median Breite als abhängige Variable logarithmiert, da somit die geschätzten Koeffizienten multiplikativ interpretiert werden können und die Schiefe der Verteilung korrigiert wird. Die sich hieraus ergebenden exponierten Koeffizienten und p-Werte sind in Tabelle 4 dargestellt.

| Variablen | $\exp\left(\hat{\beta}\right)$ | p-Wert |
|---|---|---|
| Ad-Lasso | 0.99 | 0.13 |
| MSS | 0.91 | 0.00 |
| LPR | 1.12 | 0.00 |
| LP | 1.06 | 0.00 |
| SNR 5 | 1.33 | 0.00 |
| Beta-Kliff | 0.81 | 0.00 |
| MI-Boot (PS) | 0.77 | 0.00 |
| Korrelation 0.5 | 1.23 | 0.00 |
| Korrelation 0.8 | 1.86 | 0.00 |
| Fehlende Werte | 1.25 | 0.00 |
| | Adjustierte $R^2 = 0.92$ | |

Tabelle 4: Exponierte Koeffizienten und p-Werte der Prädiktoren für eine lineare Regression mit logarithmierter Median Konfidenzintervallbreite als Zielgröße.

Über alle Settings hinweg sind eine hohe Multikollinearität, fehlende Werte in Variablen und eine kleinere Signal-Noise-Ratio, was einer größeren Varianz des Störterms entspricht, Faktoren, die bei Konstanthaltung der anderen Variablen zu einem höheren Median der Konfidenzintervalle führen. Zwischen den einzelnen statistischen Modellen gibt es nur relativ geringe Unterschiede. Beispielsweise ist für die LPR-Regression die Median Breite der Konfidenzintervalle im Vergleich zu dem Lasso um 12% erhöht. Für die Inferenzmethode gilt, dass MI-Boot (PS) kürzere Intervalle erzeugt als Boot-MI. Zudem ist der Median der Konfidenzintervalle bei $\beta = \beta_{\text{Kliff}}$ um den Faktor 0.81 kleiner als bei $\beta = \beta_{\text{spärlich}}$

Obwohl das Modell der linearen Regression recht simpel ist, da zum Beispiel Interaktionseffekte oder Polynome nicht berücksichtigt wurden, erklärt es 92% der gesamten Streuung. Die Ergebnisse des Modells decken sich auch mit den Beobachtungen, die man bei einer Betrachtung der entsprechenden Abbildungen im Anhang B entdecken kann.

## 5.4 Zusammenfassung der Ergebnisse

Zusammenfassend kann man sagen, dass die Lasso-Projektion und der LPR-Schätzer über alle Settings hinweg die konsistentesten Ergebnisse im Hinblick auf die Coverage Wahrscheinlichkeiten liefern und somit für reale Situationen gut geeignet sind, bei denen der wahre Beta-Vektor und die Varianz des Störterms nicht von vornherein bekannt sind. Entscheidet man sich hingegen für das Lasso oder das adaptive Lasso, sollte auf jeden Fall die Boot-MI Inferenzmethode gewählt werden. Zusätzlich kann es sinnvoll sein, die Anzahl der Imputationen zu erhöhen, um stabilere Ergebnisse zu erhalten. Auch die MSS-Prozedur erzielt in dieser Studie bessere Coverageraten in den Boot-MI Settings und kann ohne Bedenken genutzt werden. Wie für die multiple Imputation kann es nicht schaden, die Anzahl der Iterationen $R$ zu erhöhen.
Die Tabellen mit den exakten Werten für die Coverage Anteile und die Median Breite der Konfidenzintervalle, sowie die zugehörigen Abbildungen finden sich in Anhang A und B.

## 5.5 Rechenzeit der statistischen Modelle

Betrachtet man die gemeinsame Rechenzeit aller statistischen Verfahren, benötigen die LP mit ungefähr 48% und das MSS-Verfahren (für $R = 25$) mit 31% bei weitem den größten Anteil. Die LPR-Regression (18%) und vor allem das adaptive Lasso (2%) als auch das Lasso (1%) sind hingegen recht schnell.
Für die Inferenzmethode ist Boot-MI rechenintensiver als MI-Boot (PS), da für erstere $K$-mal mehr Imputationen durchgeführt werden müssen, die in der Regel recht zeitaufwendig sind. Daher kann vor allem die LPR-Regression empfohlen werden, wenn Zeit ein limitierender Faktor ist. Dieses Modell hat von sich aus eine moderate Rechenzeit und zeichnet sich zudem für die weniger rechenaufwändige MI-Boot (PS) Inferenzmethode mit dem kleinsten mittleren MSE aus.

# 6 P-Werte und Ausblick

## 6.1 Inferenz mit p-Werten

In den bisherigen Ausführungen sind wir stets von der Annahme ausgegangen, dass das Ziel der Inferenz in der Konstruktion valider Konfidenzintervalle für unvollständige Daten liegt. Interessiert man sich hingegen beispielsweise für die p-Werte eines t-Testes in Bezug auf die Regressionskoeffizienten, ist es nicht offensichtlich, wie diese über die $M$ multipel imputierten Datensätze aggregiert werden können.
Das Hauptproblem liegt darin, dass p-Werte unter der Nullhypothese gleichverteilt sind, die pooling Regeln jedoch eine Normalverteilung voraussetzen. Als mögliche Lösung schlägt Licht (2011) eine z-Transformation über die Quantilsfunktion der Standardnormalverteilung vor. Hierdurch wird für die p-Werte eine Normalverteilung induziert, wodurch die pooling Regeln angewandt werden können. Die genauen Details dieser Vorgehensweise, als auch einige Ratschläge zur praktischen Anwendung dieses Verfahrens, können in Licht (2011) nachgelesen werden. Natürlich kann diese Inferenzmethode nur genutzt werden, wenn aus den statistischen Modellen p-Werte berechnet werden können. Dies gilt beispielsweise für das Multi-Sample-Splitting und die Lasso-Projektion. Für das einfache Lasso, das adaptive Lasso und die Lasso-Partial-Ridge-Regression liegen hingegen keine leicht zugänglichen p-Werte der Regressionskoeffizienten vor.

## 6.2 Weiterführende Forschung

In dieser Arbeit wurde eine Methodik aus Bootstrap und multipler Imputation vorgestellt, um für die Lasso-Regression und deren Erweiterungen Konfidenzintervalle für Datensätze mit unvollständigen Daten zu erzeugen. Es konnte gezeigt werden, dass vor allem die LPR-Regression und die Lasso-Projektion gute Ergebnisse liefern. Allerdings wurde sich in der durchgeführten Simulationsstudie auf die Analyse eines niedrig dimensionalen Settings beschränkt. Das Problem hochdimensionaler Datensätze besteht darin, dass jede unabhängige Variable als Linearkombination der restlichen Prädiktoren dargestellt werden kann. Es liegt somit perfekte Multikollinearität vor. Eine Konsequenz hieraus ist, dass Variablenselektionsmodelle nicht in der Lage sind zu erkennen, welche Variablen wirklich einen Einfluss auf die Zielgröße

ausüben (James et al., 2014, Kapitel 6). Für die Prognose ist diese Eigenschaft weniger gravierend, da selbst wenn die falschen Variablen ausgewählt werden, ein Modell entstehen kann, welches für zukünftige Daten gute Vorhersagen trifft. Für die Inferenz ist hingegen zu erwarten, dass das Problem der perfekten Multikollinearität negative Auswirkungen mit sich bringt. Schließlich konnte schon in der Simulationsstudie beobachtet werden, dass mit steigender Korrelation die Coverageraten zwar stabil bleiben, die Breite der Konfidenzintervalle jedoch zunimmt. Vor allem in Bezug auf die Coverageraten ist es unklar, ob diese Eigenschaft für $p > n$ erhalten bleibt. Neben der Performance im hochdimensionalem Raum, sollte in weiteren Studien ebenfalls untersucht werden, welche Auswirkungen eine Fehlspezifikation des Modells auf die Ergebnisse hat. In dieser Simulationsstudie sind wir zudem davon ausgegangen, dass alle Variablen einen linearen Einfluss auf die Zielgröße besitzen. Interaktionseffekte oder nichtlineare Abhängigkeiten, welche über Polynome modelliert werden können, wurden nicht berücksichtigt. Zudem sollte erforscht werden, ob die angewandte Inferenzmethode nicht nur für die lineare, sondern auch für generalisierte Regressionen, wie beispielsweise die logistische Regression, valide Ergebnisse liefert. Hierfür ist es natürlich notwendig die statistischen Modelle entsprechend anzupassen. Für das MSS-Verfahren gelingt dies am einfachsten, indem auf der zweiten Datensatzhälfte eine logistische statt einer linearen Regression berechnet wird. Und auch der LPR-Schätzer kann verwendet werden, wenn in dem Minimierungsproblem von Gleichung 2.11 der erste Summand durch die negative Log-Likelihood der Binomialverteilung ersetzt wird. Zusätzlich müssen alle Lasso-Regressionen, die beispielsweise zur Identifizierung des aktiven Sets an Variablen genutzt werden, an die binäre Zielgröße angepasst werden. Das weitere Vorgehen zur Berechnung der Konfidenzintervalle ist hingegen identisch zu der in dieser Arbeit betrachteten linearen Regression. Diese Flexibilität unterstreicht nochmals die Attraktivität der Inferenzmethode und der statistischen Modelle für den Anwender in der Praxis.

# Tabellenverzeichnis

1 Auflistung der acht Simulation-Settings . . . . . 21

2 Auflistung der Variablen mit fehlenden Werten und der zugehörigen Parameter $\delta$ und $\kappa$ . . . . . 23

3 Mittlere quadratischer Abweichung der Coverageraten . . . . . 32

4 Lineare Regression mit logarithmierter Median Konfidenzintervallbreite als Zielgröße . . . . . 34

5 Coverageraten und Median Konfidenzintervallbreiten für: MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR $= 5$ . . . . . 44

6 Coverageraten und Median Konfidenzintervallbreiten für: MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR $= 10$ . . . . . 45

7 Coverageraten und Median Konfidenzintervallbreiten für: MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR $= 5$ . . . . . 46

8 Coverageraten und Median Konfidenzintervallbreiten für: MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR $= 10$ . . . . . 47

9 Coverageraten und Median Konfidenzintervallbreiten für: Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR $= 5$ . . . . . 48

10 Coverageraten und Median Konfidenzintervallbreiten für: Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR $= 10$ . . . . . 49

11 Coverageraten und Median Konfidenzintervallbreiten für: Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR $= 5$ . . . . . 50

12 Coverageraten und Median Konfidenzintervallbreiten für: Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR $= 10$ . . . . . 51

# Abbildungsverzeichnis

1 Wahrscheinlichkeitsfunktion fehlender Werte für $\delta = 0.01$ und $\kappa = 1$ . 22

2 Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR = 5: Coverage Wahrscheinlichkeiten . . . 28

3 Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR = 5: Median Konfidenzintervallbreite . . 29

4 Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR = 5: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit . . . . . 30

5 Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR = 5: Boxplot Parameterverteilung . . . . 31

6 MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR = 5: Coverage Wahrscheinlichkeiten 52

7 MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR = 5: Median Konfidenzintervallbreite . . . . . 53

8 MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR = 5: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit . . . . . 54

9 MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR = 5: Boxplot Parameterverteilung 55

10 MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR = 10: Coverage Wahrscheinlichkeiten . . . . . 56

11 MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR = 10: Median Konfidenzintervallbreite . . . . . 57

12 MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR = 10: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit . . . . . 58

13 MI-Boot (PS), $\beta = \beta_{\text{spärlich}}$ und SNR = 10: Boxplot Parameterverteilung 59

14 MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR = 5: Coverage Wahrscheinlichkeiten 60

15 MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR = 5: Median Konfidenzintervallbreite 61

16 MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR = 5: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit . . . . . 62

17 MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR = 5: Boxplot Parameterverteilung . 63

18 MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR = 10: Coverage Wahrscheinlichkeiten 64

19 MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR = 10: Median Konfidenzintervallbreite 65

20 MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR = 10: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit . . . . . 66

21 MI-Boot (PS), $\beta = \beta_{\text{Kliff}}$ und SNR = 10: Boxplot Parameterverteilung 67

22 Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR = 5: Coverage Wahrscheinlichkeiten . 68

23 Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR = 5: Median Konfidenzintervallbreite . 69

24 Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR = 5: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit . . . . . 70

25 Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR = 5: Boxplot Parameterverteilung . . 71

26 Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR = 10: Coverage Wahrscheinlichkeiten . 72

27 Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR = 10: Median Konfidenzintervallbreite 73

28 Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR = 10: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit . . . . . 74

29 Boot-MI, $\beta = \beta_{\text{spärlich}}$ und SNR = 10: Boxplot Parameterverteilung . . 75

30 Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR = 10: Coverage Wahrscheinlichkeiten . . 76

31 Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR = 10: Median Konfidenzintervallbreite . 77

32 Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR = 10: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit . . . . . 78

33 Boot-MI, $\beta = \beta_{\text{Kliff}}$ und SNR = 10: Boxplot Parameterverteilung . . . 79

# Literatur

Carpenter, J. und Kenward, M. (2012). *Multiple Imputation and its Application*, Statistics in Practice, Wiley.

Chen, Q. und Wang, S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study, *Statistics in Medicine* **32**(21): 3646–3659.

Deng, Y., Chang, C., Ido, M. S. und Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data, *Scientific Reports* **6**(21689).

Department of Biostatistics, Vanderbilt University (o. J.). Statistical problems to document and to avoid. Letzter Zugriff 14. März 2018.
**URL:** *http://biostat.mc.vanderbilt.edu/wiki/Main/ManuscriptChecklist*

Dezeure, R., Bühlmann, P., Meier, L. und Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi, *Statistical Science* **30**(4): 533–558.

Enders, C. (2010). *Applied Missing Data Analysis*, Methodology in the social sciences, Guilford Publications.

Friedman, J., Hastie, T. und Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**(1): 1–22. Letzter Zugriff 14. März 2018.
**URL:** *https://www.jstatsoft.org/article/view/v033i01*

Goeman, J. J., Meijer, R. J. und Chaturvedi, N. (2016). L1 and l2 penalized regression models. Letzter Zugriff 14. März 2018.
**URL:** *https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf*

Goeman, J. J., Meijer, R. J. und Chaturvedi, N. (2017). *Penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model.* Letzter Zugriff 14. März 2018.
**URL:** *https://cran.r-project.org/web/packages/penalized/index.html*

Hastie, T., Tibshirani, R. und Friedman, J. (2001). *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc.

James, G., Witten, D., Hastie, T. und Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*, Springer Publishing Company, Incorporated.

Licht, C. (2011). New methods for generating significance levels from multiply-imputed data. Letzter Zugriff 14. März 2018.
**URL:** *https://d-nb.info/101104966X/34*

Liu, H., Xu, X. und Li, J. J. (2017a). A Bootstrap Lasso + Partial Ridge Method to Construct Confidence Intervals for Parameters in High-dimensional Sparse Linear Models, *ArXiv e-prints* . Letzter Zugriff 14. März 2018.
**URL:** *https://arxiv.org/abs/1706.02150*

Liu, H., Xu, X. und Li, J. J. (2017b). Hdci: High dimensional confidence interval based on lasso and bootstrap. Letzter Zugriff 14. März 2018.
**URL:** *https://cran.r-project.org/web/packages/HDCI/index.html*

Liu, H. und Yu, B. (2013). Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression, *Electron. J. Statist.* **7**: 3124–3169.

Meinshausen, N. und Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso, *Ann. Statist.* **34**(3): 1436–1462.

Pötscher, B. und Schneider, U. (2007). On the distribution of the adaptive lasso estimator. Letzter Zugriff 14. März 2018.
**URL:** *https://EconPapers.repec.org/RePEc:pra:mprapa:6913*

Robitzsch, A., Grund, S. und Henke, T. (2017). *miceadds: Some additional multiple imputation functions, especially for mice*. Letzter Zugriff 14. März 2018.
**URL:** *https://CRAN.R-project.org/package=miceadds*

Schafer, J. L. und Graham, J. W. (2002). Missing data: Our view of the state of the art, *Psychological Methods* **7**(2): 147–177.

Schomaker, M. und Heumann, C. (2014). Model selection and model averaging after multiple imputation, *Computational Statistics and Data Analysis* **71**: 758 – 770.

Schomaker, M. und Heumann, C. (2016). Bootstrap inference when using multiple imputation. Letzter Zugriff 14. März 2018.
**URL:** *https://arxiv.org/abs/1602.07933*

van Buuren, S. (2012). *Flexible Imputation of Missing Data*, Chapman & Hall/CRC Press, Boca Raton, FL.

van Buuren, S. und Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r, *Journal of Statistical Software* **45**(3): 1–67.

Venables, W. N. und Ripley, B. D. (2017). Mass: Support functions and datasets for venables and ripley's mass. Letzter Zugriff 14. März 2018.
**URL:** *https://cran.r-project.org/web/packages/MASS/index.html*

Wood, A. M., White, I. R. und Royston, P. (2008). How should variable selection be performed with multiply imputed data?, *Statistics in Medicine* **27**(17): 3227–3246.

Zhang, C.-H. und Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society Series B* **76**(1): 217–242.

Zhao, S., Shojaie, A. und Witten, D. (2017). In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. Letzter Zugriff 14. März 2018.
**URL:** *https://arxiv.org/abs/1705.05543*

Zhao, Y. und Long, Q. (2016). Multiple imputation in the presence of high-dimensional data, *Statistical Methods in Medical Research* **25**(5): 2021–2035.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* **101**: 1418–1429.

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.

# Anhang

## A Tabellen

| | Inferenzmethode = MI-Boot (PS) $\beta = \beta_{\text{spärlich}}$ SNR = 5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage in % | | | | | Median KI Breite | | | | |
| Variable | Lasso | Ad-Lasso | MSS | LPR | LP | Lasso | Ad-Lasso | MSS | LPR | LP |
| $X_1$ | 91.6 | 93.4 | 94.6 | 93.4 | 92.6 | 25.86 | 26.78 | 28.00 | 27.85 | 25.72 |
| $X_2$ | 95.4 | 95.0 | 96.0 | 95.6 | 95.0 | 23.59 | 25.42 | 26.99 | 25.63 | 23.92 |
| $X_3$ | 90.4 | 90.2 | 89.0 | 91.4 | 90.0 | 21.65 | 22.92 | 23.65 | 24.31 | 23.62 |
| $X_4$ | 92.4 | 92.2 | 86.8 | 94.0 | 93.6 | 17.36 | 18.30 | 16.35 | 20.39 | 20.36 |
| $X_5$ | 99.2 | 99.2 | 96.6 | 91.6 | 91.6 | 22.45 | 23.93 | 19.50 | 26.77 | 25.55 |
| $X_6$ | 99.2 | 99.6 | 98.6 | 95.0 | 95.0 | 17.73 | 18.70 | 14.39 | 21.64 | 20.98 |
| $X_7$ | 99.4 | 99.4 | 98.8 | 96.0 | 94.0 | 18.38 | 19.64 | 14.42 | 22.39 | 21.77 |
| $X_8$ | 100.0 | 100.0 | 99.6 | 94.8 | 95.4 | 18.31 | 19.46 | 13.56 | 23.01 | 21.39 |
| $X_9$ | 98.0 | 99.2 | 97.8 | 95.0 | 91.6 | 16.58 | 17.39 | 14.76 | 19.96 | 19.63 |
| $X_{10}$ | 99.4 | 99.8 | 99.6 | 93.2 | 93.4 | 15.88 | 16.79 | 11.96 | 19.78 | 18.69 |
| $X_{11}$ | 99.0 | 99.8 | 99.0 | 94.8 | 94.4 | 18.72 | 20.07 | 14.77 | 22.76 | 22.22 |
| $X_{12}$ | 99.2 | 99.2 | 99.0 | 94.6 | 94.8 | 17.56 | 18.43 | 14.72 | 21.25 | 20.83 |
| $X_{13}$ | 98.2 | 99.2 | 98.6 | 93.8 | 93.0 | 18.72 | 19.99 | 15.02 | 23.07 | 21.90 |
| $X_{14}$ | 90.0 | 91.8 | 91.2 | 91.4 | 91.2 | 18.07 | 18.42 | 18.95 | 19.00 | 17.59 |
| $X_{15}$ | 91.4 | 93.4 | 94.4 | 93.8 | 93.4 | 14.32 | 15.02 | 15.38 | 15.10 | 14.09 |
| $X_{16}$ | 92.6 | 92.8 | 92.6 | 93.2 | 91.8 | 15.64 | 16.45 | 17.12 | 16.77 | 16.11 |
| $X_{17}$ | 95.4 | 96.0 | 95.2 | 95.6 | 93.6 | 13.71 | 14.12 | 15.16 | 15.25 | 14.98 |
| $X_{18}$ | 95.6 | 97.8 | 94.6 | 91.0 | 89.8 | 14.86 | 15.14 | 13.97 | 17.19 | 16.42 |
| $X_{19}$ | 98.6 | 99.2 | 98.0 | 95.0 | 95.2 | 12.04 | 12.19 | 10.84 | 14.20 | 14.04 |
| $X_{20}$ | 99.0 | 99.6 | 98.0 | 95.6 | 93.8 | 11.57 | 11.60 | 10.15 | 13.74 | 13.46 |
| $X_{21}$ | 97.2 | 98.8 | 98.2 | 94.2 | 92.2 | 12.79 | 13.10 | 11.74 | 15.10 | 14.46 |
| $X_{22}$ | 98.0 | 98.2 | 95.8 | 94.2 | 94.0 | 12.29 | 12.43 | 10.86 | 14.62 | 14.27 |
| $X_{23}$ | 99.0 | 99.4 | 98.4 | 95.6 | 95.8 | 11.80 | 11.95 | 10.23 | 14.07 | 13.61 |
| $X_{24}$ | 97.6 | 99.0 | 97.6 | 94.0 | 93.2 | 11.96 | 12.01 | 10.55 | 14.17 | 13.59 |
| $X_{25}$ | 98.0 | 98.8 | 97.2 | 92.6 | 92.4 | 12.22 | 12.49 | 10.58 | 14.64 | 14.07 |
| $X_{26}$ | 99.0 | 100.0 | 98.2 | 95.0 | 93.8 | 11.38 | 11.29 | 9.99 | 13.46 | 13.02 |
| $X_{27}$ | 94.0 | 93.6 | 93.8 | 92.8 | 92.6 | 12.74 | 12.53 | 12.47 | 13.06 | 12.21 |
| $X_{28}$ | 93.2 | 95.0 | 95.6 | 94.8 | 94.4 | 12.17 | 12.57 | 12.83 | 12.70 | 12.01 |
| $X_{29}$ | 92.6 | 94.4 | 94.8 | 94.4 | 94.0 | 13.43 | 14.09 | 14.65 | 14.48 | 13.68 |
| $X_{30}$ | 94.6 | 95.0 | 95.4 | 95.0 | 93.0 | 10.91 | 11.20 | 11.92 | 11.93 | 11.37 |
| $X_{31}$ | 95.4 | 98.4 | 94.2 | 91.4 | 91.0 | 11.37 | 11.37 | 11.00 | 12.93 | 12.44 |
| $X_{32}$ | 98.8 | 99.4 | 98.6 | 95.0 | 94.0 | 10.15 | 10.06 | 9.21 | 12.01 | 11.57 |
| $X_{33}$ | 98.0 | 99.6 | 97.0 | 93.6 | 94.4 | 10.00 | 9.90 | 9.16 | 11.71 | 11.33 |
| $X_{34}$ | 98.2 | 99.0 | 96.4 | 94.2 | 94.2 | 10.31 | 10.20 | 9.41 | 12.01 | 11.57 |
| $X_{35}$ | 96.8 | 99.4 | 96.8 | 94.0 | 92.8 | 10.06 | 9.79 | 9.30 | 11.54 | 11.28 |
| $X_{36}$ | 98.6 | 99.6 | 98.0 | 94.8 | 94.0 | 10.04 | 9.94 | 9.26 | 11.63 | 11.36 |
| $X_{37}$ | 98.2 | 99.8 | 97.6 | 96.2 | 96.4 | 10.12 | 10.00 | 9.24 | 11.87 | 11.34 |
| $X_{38}$ | 98.0 | 99.4 | 97.6 | 94.4 | 92.6 | 10.18 | 10.10 | 9.25 | 11.92 | 11.60 |
| $X_{39}$ | 96.2 | 98.0 | 96.2 | 91.8 | 90.8 | 10.65 | 10.70 | 10.14 | 12.25 | 12.03 |

Tabelle 5: Coverageraten und Median Konfidenzintervallbreiten der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion.

| | Inferenzmethode = MI-Boot (PS) $\beta = \beta_{\text{spärlich}}$ SNR = 10 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage in % | | | | | Median KI Breite | | | | |
| Variable | Lasso | Ad-Lasso | MSS | LPR | LP | Lasso | Ad-Lasso | MSS | LPR | LP |
| $X_1$ | 87.6 | 88.8 | 90.0 | 90.0 | 88.8 | 19.47 | 19.76 | 20.33 | 20.89 | 18.76 |
| $X_2$ | 92.0 | 92.8 | 92.8 | 93.0 | 93.0 | 17.66 | 18.10 | 19.11 | 19.04 | 16.89 |
| $X_3$ | 92.2 | 92.4 | 92.4 | 92.2 | 92.6 | 19.11 | 19.84 | 20.90 | 20.57 | 19.77 |
| $X_4$ | 92.6 | 92.0 | 91.2 | 94.4 | 94.6 | 14.77 | 14.90 | 14.77 | 16.64 | 16.45 |
| $X_5$ | 97.8 | 98.2 | 95.8 | 91.4 | 91.8 | 17.29 | 17.74 | 14.95 | 20.45 | 18.67 |
| $X_6$ | 99.6 | 100.0 | 98.4 | 93.4 | 93.4 | 14.58 | 15.06 | 11.26 | 17.70 | 16.35 |
| $X_7$ | 99.0 | 99.6 | 97.8 | 94.4 | 92.2 | 14.12 | 14.22 | 11.86 | 16.72 | 15.82 |
| $X_8$ | 98.6 | 99.4 | 97.6 | 93.8 | 95.0 | 13.04 | 13.34 | 10.55 | 15.44 | 14.59 |
| $X_9$ | 99.0 | 99.6 | 98.0 | 93.6 | 94.6 | 14.76 | 15.20 | 12.00 | 17.71 | 16.75 |
| $X_{10}$ | 98.6 | 99.6 | 98.6 | 95.0 | 94.6 | 13.34 | 13.58 | 11.12 | 15.95 | 14.90 |
| $X_{11}$ | 97.2 | 99.2 | 97.8 | 95.0 | 92.8 | 13.04 | 13.15 | 11.73 | 15.14 | 14.63 |
| $X_{12}$ | 97.8 | 99.0 | 97.2 | 92.4 | 93.4 | 12.82 | 13.16 | 11.01 | 15.26 | 14.70 |
| $X_{13}$ | 98.8 | 99.4 | 98.4 | 95.6 | 96.0 | 14.18 | 14.62 | 11.40 | 17.08 | 15.93 |
| $X_{14}$ | 92.4 | 92.8 | 93.4 | 93.4 | 91.8 | 12.61 | 12.49 | 12.67 | 13.18 | 12.14 |
| $X_{15}$ | 92.8 | 94.4 | 94.2 | 95.0 | 93.8 | 10.76 | 10.89 | 11.07 | 11.23 | 10.47 |
| $X_{16}$ | 92.8 | 90.6 | 91.2 | 92.0 | 90.6 | 11.61 | 11.93 | 12.25 | 12.22 | 11.51 |
| $X_{17}$ | 94.4 | 94.4 | 93.8 | 95.8 | 95.8 | 9.58 | 9.88 | 10.56 | 10.47 | 10.12 |
| $X_{18}$ | 96.2 | 98.8 | 93.8 | 92.2 | 92.0 | 11.77 | 11.78 | 10.86 | 13.56 | 12.58 |
| $X_{19}$ | 98.2 | 99.2 | 96.4 | 93.6 | 94.0 | 8.90 | 8.83 | 7.90 | 10.40 | 9.81 |
| $X_{20}$ | 93.0 | 96.2 | 94.6 | 92.4 | 90.0 | 9.39 | 9.21 | 8.75 | 10.62 | 10.24 |
| $X_{21}$ | 97.4 | 99.2 | 96.6 | 95.4 | 95.0 | 9.64 | 9.62 | 8.60 | 11.13 | 10.70 |
| $X_{22}$ | 98.8 | 99.8 | 94.2 | 92.0 | 93.2 | 9.69 | 9.57 | 8.56 | 11.24 | 10.68 |
| $X_{23}$ | 97.8 | 99.2 | 97.4 | 95.0 | 94.2 | 9.75 | 9.60 | 8.76 | 11.16 | 10.74 |
| $X_{24}$ | 98.4 | 99.4 | 97.6 | 95.0 | 94.2 | 8.35 | 8.30 | 7.65 | 9.80 | 9.26 |
| $X_{25}$ | 96.2 | 98.6 | 96.8 | 94.2 | 93.4 | 9.28 | 9.19 | 8.43 | 10.64 | 10.21 |
| $X_{26}$ | 98.2 | 98.8 | 96.4 | 93.8 | 93.4 | 9.81 | 9.88 | 8.85 | 11.45 | 10.79 |
| $X_{27}$ | 93.6 | 94.0 | 94.8 | 94.8 | 93.6 | 8.92 | 8.74 | 8.79 | 9.22 | 8.58 |
| $X_{28}$ | 94.8 | 95.2 | 96.0 | 95.4 | 95.0 | 8.63 | 8.79 | 8.84 | 9.09 | 8.51 |
| $X_{29}$ | 93.0 | 93.2 | 92.8 | 91.8 | 91.0 | 9.64 | 9.97 | 10.10 | 10.02 | 9.43 |
| $X_{30}$ | 92.8 | 93.2 | 93.6 | 92.8 | 91.8 | 8.44 | 8.86 | 9.35 | 9.04 | 8.67 |
| $X_{31}$ | 95.8 | 98.8 | 95.6 | 92.6 | 92.4 | 8.89 | 8.72 | 8.56 | 10.05 | 9.50 |
| $X_{32}$ | 98.8 | 100.0 | 96.6 | 95.0 | 94.8 | 7.34 | 7.14 | 6.79 | 8.48 | 8.07 |
| $X_{33}$ | 95.2 | 98.2 | 90.6 | 86.8 | 87.4 | 7.89 | 7.72 | 7.59 | 8.94 | 8.53 |
| $X_{34}$ | 98.0 | 100.0 | 98.0 | 94.8 | 94.6 | 7.90 | 7.68 | 7.38 | 9.06 | 8.67 |
| $X_{35}$ | 96.0 | 98.8 | 96.6 | 93.8 | 92.0 | 8.10 | 7.94 | 7.73 | 9.22 | 8.82 |
| $X_{36}$ | 97.8 | 99.2 | 97.4 | 95.2 | 95.0 | 7.09 | 6.78 | 6.75 | 8.05 | 7.71 |
| $X_{37}$ | 98.6 | 99.4 | 96.6 | 95.2 | 96.0 | 7.88 | 7.63 | 7.32 | 9.04 | 8.58 |
| $X_{38}$ | 97.0 | 98.8 | 96.8 | 95.0 | 93.0 | 7.95 | 7.75 | 7.46 | 8.94 | 8.64 |
| $X_{39}$ | 98.0 | 100.0 | 97.8 | 96.2 | 96.4 | 7.43 | 7.20 | 6.92 | 8.65 | 8.21 |

Tabelle 6: Coverageraten und Median Konfidenzintervallbreiten der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion.

| | Inferenzmethode = MI-Boot (PS) $\beta = \beta_{\text{Kliff}}$ SNR = 5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage in % | | | | | Median KI Breite | | | | |
| Variable | Lasso | Ad-Lasso | MSS | LPR | LP | Lasso | Ad-Lasso | MSS | LPR | LP |
| $X_1$ | 90.6 | 92.6 | 92.6 | 92.4 | 92.8 | 19.89 | 20.98 | 21.81 | 21.18 | 19.54 |
| $X_2$ | 92.4 | 94.2 | 95.2 | 93.6 | 94.8 | 23.35 | 24.63 | 25.38 | 25.47 | 24.01 |
| $X_3$ | 93.4 | 95.0 | 94.0 | 96.2 | 95.4 | 18.15 | 19.73 | 21.21 | 19.92 | 19.39 |
| $X_4$ | 90.4 | 87.0 | 88.2 | 88.8 | 96.2 | 13.96 | 15.06 | 11.00 | 17.18 | 17.47 |
| $X_5$ | 92.8 | 85.8 | 95.4 | 90.2 | 94.6 | 17.48 | 19.01 | 13.89 | 21.44 | 21.28 |
| $X_6$ | 94.4 | 92.0 | 98.2 | 95.0 | 94.6 | 14.86 | 15.74 | 11.81 | 18.14 | 18.65 |
| $X_7$ | 91.4 | 87.8 | 100.0 | 93.8 | 93.4 | 15.66 | 17.26 | 11.75 | 19.28 | 19.39 |
| $X_8$ | 88.2 | 83.0 | 98.8 | 92.2 | 93.6 | 18.38 | 19.93 | 13.79 | 23.00 | 21.67 |
| $X_9$ | 93.6 | 89.6 | 98.2 | 93.8 | 93.4 | 14.32 | 15.22 | 12.42 | 17.27 | 17.55 |
| $X_{10}$ | 90.0 | 82.4 | 99.4 | 94.2 | 95.0 | 13.45 | 14.46 | 9.87 | 16.90 | 16.94 |
| $X_{11}$ | 91.0 | 88.0 | 98.4 | 94.0 | 95.4 | 18.53 | 20.15 | 15.11 | 22.51 | 22.19 |
| $X_{12}$ | 92.8 | 89.0 | 99.2 | 94.4 | 94.6 | 15.15 | 16.00 | 12.21 | 18.54 | 18.77 |
| $X_{13}$ | 99.4 | 99.8 | 99.4 | 95.8 | 95.8 | 15.95 | 17.32 | 13.00 | 19.30 | 19.79 |
| $X_{14}$ | 91.8 | 92.4 | 94.0 | 94.0 | 92.4 | 13.80 | 13.97 | 14.06 | 14.45 | 13.51 |
| $X_{15}$ | 94.4 | 94.8 | 95.6 | 95.8 | 95.2 | 14.80 | 15.42 | 15.48 | 15.81 | 14.37 |
| $X_{16}$ | 94.0 | 95.2 | 95.2 | 95.4 | 95.2 | 12.87 | 14.05 | 14.54 | 13.65 | 13.08 |
| $X_{17}$ | 96.0 | 94.4 | 95.6 | 96.6 | 95.8 | 11.01 | 11.03 | 10.30 | 12.79 | 13.05 |
| $X_{18}$ | 92.2 | 91.4 | 94.4 | 93.2 | 93.2 | 11.85 | 12.08 | 10.50 | 13.97 | 13.89 |
| $X_{19}$ | 96.0 | 91.0 | 98.0 | 95.8 | 95.2 | 10.24 | 10.44 | 8.73 | 12.14 | 12.30 |
| $X_{20}$ | 92.2 | 86.0 | 97.8 | 93.8 | 94.6 | 9.89 | 9.90 | 8.37 | 11.85 | 11.97 |
| $X_{21}$ | 92.2 | 90.2 | 95.8 | 93.0 | 91.8 | 12.72 | 13.19 | 11.86 | 14.89 | 14.69 |
| $X_{22}$ | 94.2 | 87.0 | 99.0 | 97.0 | 96.8 | 10.58 | 10.79 | 8.97 | 12.79 | 12.69 |
| $X_{23}$ | 92.6 | 83.2 | 98.6 | 95.6 | 95.6 | 10.32 | 10.41 | 8.81 | 12.34 | 12.19 |
| $X_{24}$ | 96.0 | 92.4 | 98.8 | 98.0 | 96.6 | 11.77 | 11.90 | 10.54 | 13.69 | 13.69 |
| $X_{25}$ | 91.2 | 81.4 | 98.6 | 95.8 | 95.2 | 10.41 | 10.56 | 8.94 | 12.45 | 12.43 |
| $X_{26}$ | 99.0 | 99.8 | 97.4 | 94.6 | 93.8 | 9.67 | 9.76 | 8.19 | 11.56 | 11.61 |
| $X_{27}$ | 94.0 | 95.6 | 95.8 | 96.0 | 94.4 | 10.00 | 10.09 | 10.04 | 10.27 | 9.88 |
| $X_{28}$ | 92.6 | 93.8 | 94.4 | 94.6 | 94.2 | 12.44 | 12.56 | 12.68 | 12.98 | 12.07 |
| $X_{29}$ | 93.0 | 94.6 | 95.6 | 94.8 | 94.4 | 10.77 | 11.61 | 12.06 | 11.52 | 10.81 |
| $X_{30}$ | 96.8 | 95.2 | 95.4 | 96.6 | 95.8 | 8.52 | 8.19 | 7.95 | 9.78 | 9.71 |
| $X_{31}$ | 94.2 | 90.0 | 94.4 | 94.4 | 95.0 | 9.16 | 8.94 | 8.51 | 10.59 | 10.41 |
| $X_{32}$ | 92.2 | 84.0 | 98.0 | 94.8 | 93.8 | 8.80 | 8.81 | 7.91 | 10.47 | 10.37 |
| $X_{33}$ | 94.4 | 88.4 | 98.0 | 95.4 | 94.6 | 8.62 | 8.57 | 7.66 | 10.20 | 10.17 |
| $X_{34}$ | 92.6 | 86.6 | 97.6 | 94.6 | 93.2 | 10.15 | 10.15 | 9.72 | 11.75 | 11.53 |
| $X_{35}$ | 93.2 | 88.4 | 96.8 | 93.4 | 92.6 | 8.58 | 8.45 | 7.86 | 10.01 | 9.99 |
| $X_{36}$ | 93.2 | 84.8 | 98.2 | 95.4 | 95.6 | 8.57 | 8.44 | 7.68 | 10.08 | 10.01 |
| $X_{37}$ | 92.0 | 90.8 | 96.6 | 93.2 | 91.8 | 9.94 | 9.88 | 9.26 | 11.47 | 11.14 |
| $X_{38}$ | 92.0 | 84.8 | 98.4 | 93.2 | 94.4 | 8.81 | 8.77 | 7.74 | 10.49 | 10.35 |
| $X_{39}$ | 99.0 | 100.0 | 99.0 | 96.4 | 96.2 | 9.03 | 8.97 | 8.16 | 10.63 | 10.50 |

Tabelle 7: Coverageraten und Median Konfidenzintervallbreiten der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion.

| | Inferenzmethode = MI-Boot (PS) $\beta = \beta_{\text{Kliff}}$ SNR = 10 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage in % | | | | | Median KI Breite | | | | |
| Variable | Lasso | Ad-Lasso | MSS | LPR | LP | Lasso | Ad-Lasso | MSS | LPR | LP |
| $X_1$ | 90.0 | 94.0 | 94.2 | 93.2 | 94.8 | 14.65 | 15.03 | 15.19 | 15.49 | 14.27 |
| $X_2$ | 93.2 | 94.0 | 94.4 | 93.4 | 92.2 | 17.54 | 18.03 | 18.91 | 18.85 | 17.20 |
| $X_3$ | 93.8 | 94.6 | 94.2 | 95.0 | 94.4 | 14.38 | 15.30 | 16.20 | 15.41 | 14.28 |
| $X_4$ | 91.6 | 88.6 | 88.2 | 91.4 | 94.6 | 10.59 | 11.15 | 8.41 | 12.80 | 12.86 |
| $X_5$ | 88.4 | 85.2 | 89.6 | 88.2 | 91.4 | 12.94 | 13.66 | 10.92 | 15.54 | 15.28 |
| $X_6$ | 95.4 | 93.4 | 99.4 | 96.2 | 93.8 | 11.25 | 11.64 | 9.34 | 13.53 | 13.66 |
| $X_7$ | 91.8 | 83.2 | 98.8 | 94.0 | 95.6 | 11.86 | 12.58 | 9.20 | 14.41 | 14.11 |
| $X_8$ | 89.8 | 85.0 | 98.4 | 92.6 | 94.0 | 13.39 | 14.07 | 10.26 | 16.79 | 15.23 |
| $X_9$ | 93.4 | 91.0 | 97.6 | 94.2 | 93.0 | 10.76 | 10.92 | 9.19 | 12.89 | 12.72 |
| $X_{10}$ | 91.4 | 84.0 | 98.8 | 94.8 | 94.8 | 10.07 | 10.51 | 7.79 | 12.37 | 12.22 |
| $X_{11}$ | 90.4 | 86.6 | 98.2 | 92.4 | 93.4 | 13.57 | 14.29 | 11.16 | 16.45 | 15.72 |
| $X_{12}$ | 94.4 | 88.6 | 99.2 | 95.4 | 95.8 | 11.41 | 11.52 | 8.95 | 13.85 | 13.63 |
| $X_{13}$ | 98.6 | 99.4 | 98.8 | 94.6 | 94.0 | 12.07 | 12.75 | 9.69 | 14.69 | 14.42 |
| $X_{14}$ | 93.4 | 94.4 | 94.2 | 94.6 | 93.2 | 10.17 | 10.13 | 10.23 | 10.57 | 9.93 |
| $X_{15}$ | 92.6 | 95.0 | 96.6 | 96.0 | 95.8 | 10.31 | 10.14 | 10.02 | 10.72 | 9.84 |
| $X_{16}$ | 93.2 | 93.2 | 94.6 | 94.2 | 94.4 | 9.66 | 10.03 | 10.42 | 10.07 | 9.59 |
| $X_{17}$ | 95.2 | 94.6 | 93.2 | 96.2 | 96.2 | 8.19 | 8.01 | 7.62 | 9.50 | 9.46 |
| $X_{18}$ | 94.8 | 92.2 | 93.2 | 94.6 | 93.4 | 8.69 | 8.65 | 7.95 | 10.06 | 9.96 |
| $X_{19}$ | 91.6 | 88.8 | 96.8 | 92.8 | 93.4 | 7.71 | 7.63 | 6.72 | 9.07 | 8.96 |
| $X_{20}$ | 92.6 | 83.0 | 97.0 | 94.4 | 94.2 | 7.46 | 7.37 | 6.27 | 8.88 | 8.74 |
| $X_{21}$ | 93.4 | 91.2 | 96.8 | 92.6 | 90.2 | 9.38 | 9.53 | 8.80 | 10.87 | 10.50 |
| $X_{22}$ | 93.4 | 86.0 | 98.4 | 95.8 | 96.0 | 7.91 | 7.96 | 6.72 | 9.43 | 9.23 |
| $X_{23}$ | 92.0 | 83.4 | 97.2 | 94.6 | 94.2 | 7.62 | 7.65 | 6.58 | 9.07 | 8.88 |
| $X_{24}$ | 93.0 | 90.6 | 95.8 | 93.8 | 92.8 | 8.49 | 8.39 | 7.70 | 9.87 | 9.61 |
| $X_{25}$ | 91.2 | 81.4 | 97.8 | 94.0 | 93.6 | 7.76 | 7.80 | 6.55 | 9.27 | 9.04 |
| $X_{26}$ | 98.8 | 99.4 | 97.2 | 95.4 | 95.0 | 7.24 | 7.21 | 6.15 | 8.65 | 8.48 |
| $X_{27}$ | 94.6 | 94.6 | 93.8 | 94.2 | 94.4 | 7.40 | 7.32 | 7.42 | 7.57 | 7.25 |
| $X_{28}$ | 92.8 | 93.6 | 94.2 | 94.4 | 93.4 | 8.55 | 8.36 | 8.31 | 8.78 | 8.20 |
| $X_{29}$ | 92.8 | 94.2 | 95.6 | 95.4 | 95.4 | 7.97 | 8.23 | 8.38 | 8.31 | 7.93 |
| $X_{30}$ | 95.0 | 93.6 | 93.2 | 96.8 | 94.8 | 6.35 | 5.89 | 6.00 | 7.16 | 7.06 |
| $X_{31}$ | 92.6 | 90.6 | 92.6 | 92.8 | 93.4 | 6.63 | 6.36 | 6.17 | 7.66 | 7.43 |
| $X_{32}$ | 93.4 | 82.6 | 98.2 | 95.0 | 94.6 | 6.61 | 6.45 | 5.93 | 7.73 | 7.49 |
| $X_{33}$ | 94.6 | 89.4 | 98.2 | 96.2 | 95.2 | 6.44 | 6.23 | 5.79 | 7.46 | 7.37 |
| $X_{34}$ | 91.6 | 90.0 | 96.4 | 93.2 | 92.8 | 7.49 | 7.30 | 6.99 | 8.53 | 8.25 |
| $X_{35}$ | 94.8 | 93.2 | 96.2 | 94.0 | 93.8 | 6.43 | 6.14 | 5.95 | 7.38 | 7.28 |
| $X_{36}$ | 92.4 | 83.4 | 97.4 | 95.0 | 94.2 | 6.43 | 6.30 | 5.90 | 7.50 | 7.36 |
| $X_{37}$ | 93.2 | 90.2 | 97.4 | 95.0 | 92.2 | 7.10 | 6.90 | 6.62 | 8.14 | 7.86 |
| $X_{38}$ | 93.8 | 81.0 | 98.2 | 95.6 | 95.6 | 6.51 | 6.35 | 5.79 | 7.63 | 7.46 |
| $X_{39}$ | 98.0 | 99.0 | 97.6 | 96.0 | 94.8 | 6.69 | 6.62 | 6.14 | 7.79 | 7.60 |

Tabelle 8: Coverageraten und Median Konfidenzintervallbreiten der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion.

| | Inferenzmethode = Boot-MI $\beta = \beta_{\text{spärlich}}$ SNR = 5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage in % | | | | | Median KI Breite | | | | |
| Variable | Lasso | Ad-Lasso | MSS | LPR | LP | Lasso | Ad-Lasso | MSS | LPR | LP |
| $X_1$ | 98.2 | 98.0 | 98.0 | 98.6 | 97.6 | 38.82 | 36.48 | 36.39 | 41.42 | 36.41 |
| $X_2$ | 98.0 | 97.8 | 97.8 | 98.4 | 98.0 | 29.37 | 28.56 | 29.33 | 31.40 | 28.69 |
| $X_3$ | 96.4 | 96.6 | 96.4 | 96.4 | 95.8 | 32.81 | 31.64 | 30.57 | 35.80 | 32.25 |
| $X_4$ | 96.4 | 95.6 | 91.6 | 97.0 | 96.8 | 24.82 | 24.07 | 21.74 | 27.42 | 25.39 |
| $X_5$ | 97.8 | 97.4 | 97.6 | 97.6 | 97.2 | 38.44 | 36.63 | 33.00 | 42.53 | 37.33 |
| $X_6$ | 98.6 | 99.2 | 98.0 | 98.4 | 97.6 | 26.18 | 25.03 | 22.03 | 29.27 | 26.40 |
| $X_7$ | 97.2 | 97.8 | 98.4 | 97.6 | 97.0 | 27.27 | 26.61 | 22.61 | 30.57 | 27.58 |
| $X_8$ | 98.4 | 98.6 | 97.8 | 98.4 | 98.0 | 27.87 | 26.90 | 22.99 | 31.54 | 27.71 |
| $X_9$ | 97.0 | 97.2 | 98.0 | 97.2 | 96.4 | 23.90 | 23.27 | 20.59 | 26.63 | 24.47 |
| $X_{10}$ | 97.4 | 98.0 | 97.6 | 97.6 | 96.8 | 22.65 | 21.72 | 18.96 | 25.57 | 23.05 |
| $X_{11}$ | 97.6 | 98.0 | 95.4 | 97.4 | 95.6 | 26.63 | 26.10 | 22.58 | 30.10 | 27.13 |
| $X_{12}$ | 97.0 | 96.2 | 96.4 | 96.4 | 95.8 | 26.28 | 25.29 | 22.13 | 29.58 | 26.33 |
| $X_{13}$ | 98.8 | 97.6 | 99.0 | 98.4 | 98.0 | 26.95 | 26.00 | 22.08 | 30.19 | 27.19 |
| $X_{14}$ | 98.2 | 97.6 | 97.6 | 98.2 | 97.0 | 26.81 | 25.45 | 25.48 | 28.22 | 25.32 |
| $X_{15}$ | 97.4 | 97.2 | 96.4 | 97.6 | 97.2 | 17.52 | 17.40 | 17.39 | 18.49 | 17.09 |
| $X_{16}$ | 97.0 | 96.8 | 96.8 | 97.6 | 96.6 | 21.96 | 21.32 | 21.18 | 23.46 | 21.85 |
| $X_{17}$ | 95.8 | 95.6 | 94.6 | 96.2 | 95.4 | 18.03 | 17.13 | 17.11 | 19.31 | 18.10 |
| $X_{18}$ | 98.2 | 97.2 | 98.0 | 99.0 | 97.8 | 24.49 | 23.34 | 21.97 | 26.87 | 24.35 |
| $X_{19}$ | 97.6 | 97.8 | 97.6 | 97.8 | 97.2 | 16.32 | 15.65 | 14.57 | 17.97 | 16.82 |
| $X_{20}$ | 98.0 | 98.2 | 97.4 | 98.0 | 97.2 | 16.13 | 15.42 | 14.07 | 17.86 | 16.60 |
| $X_{21}$ | 94.0 | 95.2 | 96.0 | 94.8 | 94.4 | 17.64 | 16.94 | 15.67 | 19.39 | 18.19 |
| $X_{22}$ | 97.2 | 97.0 | 96.8 | 97.0 | 97.2 | 17.11 | 16.61 | 15.17 | 18.96 | 17.70 |
| $X_{23}$ | 97.8 | 98.0 | 97.0 | 97.6 | 97.2 | 16.00 | 15.40 | 14.04 | 17.73 | 16.56 |
| $X_{24}$ | 98.6 | 98.4 | 98.6 | 99.0 | 97.8 | 16.09 | 15.28 | 14.22 | 17.82 | 16.60 |
| $X_{25}$ | 97.2 | 96.8 | 97.6 | 97.4 | 96.4 | 16.94 | 16.34 | 14.93 | 18.83 | 17.36 |
| $X_{26}$ | 98.0 | 97.8 | 98.6 | 98.2 | 97.4 | 15.37 | 14.75 | 13.55 | 17.05 | 15.94 |
| $X_{27}$ | 95.0 | 94.2 | 94.0 | 93.2 | 93.4 | 18.43 | 17.51 | 17.32 | 19.20 | 17.40 |
| $X_{28}$ | 97.8 | 97.6 | 97.6 | 98.2 | 98.0 | 14.72 | 14.46 | 14.44 | 15.42 | 14.30 |
| $X_{29}$ | 96.2 | 96.2 | 96.0 | 97.0 | 96.6 | 18.93 | 18.30 | 18.22 | 20.01 | 18.89 |
| $X_{30}$ | 97.0 | 96.4 | 96.2 | 97.4 | 97.0 | 14.36 | 13.63 | 13.67 | 15.39 | 14.49 |
| $X_{31}$ | 96.6 | 97.0 | 96.6 | 97.6 | 96.2 | 17.91 | 16.83 | 16.19 | 19.40 | 17.92 |
| $X_{32}$ | 97.6 | 98.2 | 97.8 | 98.0 | 97.8 | 13.53 | 12.91 | 12.20 | 14.97 | 14.05 |
| $X_{33}$ | 97.4 | 97.2 | 97.2 | 97.2 | 96.6 | 13.46 | 12.84 | 12.06 | 14.86 | 13.95 |
| $X_{34}$ | 96.4 | 96.8 | 96.4 | 96.8 | 96.2 | 14.02 | 13.40 | 12.71 | 15.43 | 14.54 |
| $X_{35}$ | 98.4 | 98.4 | 98.4 | 99.0 | 98.8 | 13.66 | 12.90 | 12.19 | 14.90 | 14.02 |
| $X_{36}$ | 95.4 | 95.4 | 96.0 | 95.4 | 95.4 | 13.47 | 12.85 | 12.11 | 14.66 | 13.94 |
| $X_{37}$ | 97.6 | 97.2 | 97.0 | 97.6 | 96.8 | 13.73 | 12.98 | 12.29 | 15.03 | 14.12 |
| $X_{38}$ | 96.6 | 96.8 | 97.8 | 97.0 | 96.6 | 13.85 | 13.28 | 12.59 | 15.22 | 14.43 |
| $X_{39}$ | 96.8 | 95.8 | 96.8 | 96.8 | 96.6 | 14.51 | 13.83 | 13.17 | 15.90 | 15.03 |

Tabelle 9: Coverageraten und Median Konfidenzintervallbreiten der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion.

| | Inferenzmethode = Boot-MI $\beta = \beta_{\text{spärlich}}$ SNR = 10 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage in % | | | | | Median KI Breite | | | | |
| Variable | Lasso | Ad-Lasso | MSS | LPR | LP | Lasso | Ad-Lasso | MSS | LPR | LP |
| $X_1$ | 98.2 | 98.0 | 98.0 | 98.0 | 97.6 | 28.75 | 27.14 | 28.15 | 30.12 | 26.44 |
| $X_2$ | 98.0 | 97.6 | 97.8 | 98.2 | 97.6 | 23.22 | 22.62 | 23.68 | 24.58 | 22.04 |
| $X_3$ | 99.0 | 98.8 | 98.8 | 99.0 | 98.4 | 25.52 | 24.10 | 24.30 | 27.21 | 24.16 |
| $X_4$ | 96.0 | 95.4 | 94.4 | 96.8 | 96.0 | 19.56 | 18.79 | 18.12 | 21.51 | 19.39 |
| $X_5$ | 97.6 | 97.2 | 97.4 | 98.2 | 97.0 | 28.95 | 27.77 | 26.27 | 32.22 | 27.71 |
| $X_6$ | 97.4 | 97.6 | 96.0 | 96.8 | 96.0 | 21.05 | 20.07 | 18.95 | 23.18 | 20.61 |
| $X_7$ | 97.6 | 97.2 | 98.4 | 97.8 | 97.2 | 21.47 | 20.81 | 19.21 | 23.74 | 21.27 |
| $X_8$ | 96.8 | 97.0 | 95.8 | 96.2 | 96.2 | 22.53 | 21.39 | 19.73 | 25.19 | 21.86 |
| $X_9$ | 97.6 | 97.0 | 97.8 | 97.4 | 97.4 | 19.03 | 18.13 | 17.36 | 21.01 | 18.87 |
| $X_{10}$ | 98.0 | 98.4 | 97.8 | 98.0 | 98.0 | 17.88 | 17.30 | 15.67 | 19.98 | 17.90 |
| $X_{11}$ | 97.2 | 97.2 | 96.8 | 97.8 | 96.8 | 21.47 | 20.81 | 19.25 | 23.89 | 21.29 |
| $X_{12}$ | 95.8 | 95.2 | 97.2 | 96.8 | 95.0 | 20.79 | 19.59 | 18.46 | 23.05 | 20.26 |
| $X_{13}$ | 97.2 | 97.2 | 98.4 | 97.4 | 97.0 | 21.85 | 20.88 | 19.26 | 24.23 | 21.14 |
| $X_{14}$ | 97.4 | 97.4 | 97.8 | 97.2 | 96.8 | 19.60 | 18.62 | 18.92 | 20.47 | 18.24 |
| $X_{15}$ | 97.4 | 97.2 | 97.4 | 98.4 | 97.4 | 13.72 | 13.44 | 13.49 | 14.48 | 13.17 |
| $X_{16}$ | 97.8 | 97.8 | 97.8 | 98.2 | 97.6 | 16.60 | 16.15 | 16.39 | 17.45 | 16.01 |
| $X_{17}$ | 98.4 | 98.2 | 98.4 | 98.6 | 98.0 | 14.38 | 13.74 | 14.10 | 15.18 | 14.19 |
| $X_{18}$ | 95.8 | 95.2 | 96.0 | 96.0 | 95.4 | 18.30 | 17.37 | 17.02 | 19.80 | 17.85 |
| $X_{19}$ | 97.8 | 97.8 | 98.2 | 98.2 | 97.4 | 12.83 | 12.29 | 12.04 | 14.06 | 13.02 |
| $X_{20}$ | 98.2 | 98.4 | 98.6 | 98.4 | 98.4 | 12.73 | 12.09 | 11.60 | 13.96 | 12.75 |
| $X_{21}$ | 95.2 | 94.4 | 97.6 | 95.6 | 94.8 | 13.87 | 13.27 | 12.96 | 15.07 | 13.92 |
| $X_{22}$ | 98.0 | 97.8 | 97.6 | 97.8 | 97.6 | 13.48 | 13.06 | 12.46 | 14.87 | 13.58 |
| $X_{23}$ | 96.4 | 97.0 | 96.6 | 96.6 | 96.2 | 12.74 | 12.20 | 11.79 | 13.89 | 12.80 |
| $X_{24}$ | 97.0 | 96.4 | 97.8 | 97.8 | 96.4 | 12.74 | 12.12 | 11.79 | 13.83 | 12.84 |
| $X_{25}$ | 98.4 | 98.0 | 97.8 | 98.2 | 97.6 | 13.17 | 12.58 | 12.14 | 14.55 | 13.30 |
| $X_{26}$ | 98.2 | 98.4 | 97.4 | 98.0 | 97.6 | 12.43 | 11.78 | 11.41 | 13.49 | 12.39 |
| $X_{27}$ | 95.6 | 94.0 | 94.6 | 93.6 | 93.6 | 12.97 | 12.31 | 12.35 | 13.40 | 12.29 |
| $X_{28}$ | 97.8 | 97.6 | 97.6 | 97.6 | 97.4 | 11.63 | 11.32 | 11.37 | 12.14 | 11.27 |
| $X_{29}$ | 98.6 | 99.0 | 98.6 | 98.6 | 98.8 | 14.29 | 13.87 | 14.15 | 15.02 | 13.91 |
| $X_{30}$ | 97.4 | 97.2 | 97.2 | 97.4 | 97.0 | 11.36 | 10.79 | 11.09 | 11.98 | 11.13 |
| $X_{31}$ | 98.4 | 99.0 | 98.6 | 98.4 | 98.6 | 13.49 | 12.78 | 12.75 | 14.59 | 13.30 |
| $X_{32}$ | 97.6 | 97.6 | 97.8 | 97.8 | 97.2 | 10.85 | 10.33 | 10.20 | 11.75 | 10.96 |
| $X_{33}$ | 97.2 | 97.4 | 98.0 | 97.8 | 97.2 | 10.71 | 10.08 | 9.88 | 11.56 | 10.77 |
| $X_{34}$ | 98.0 | 98.4 | 97.4 | 98.0 | 97.6 | 11.05 | 10.51 | 10.30 | 12.00 | 11.07 |
| $X_{35}$ | 97.4 | 97.4 | 96.8 | 97.4 | 97.6 | 10.66 | 10.10 | 10.13 | 11.50 | 10.77 |
| $X_{36}$ | 98.0 | 97.6 | 98.4 | 98.0 | 97.8 | 10.64 | 10.15 | 10.06 | 11.50 | 10.78 |
| $X_{37}$ | 98.4 | 98.2 | 97.2 | 97.8 | 97.8 | 10.74 | 10.16 | 9.98 | 11.69 | 10.85 |
| $X_{38}$ | 96.2 | 96.8 | 97.2 | 97.0 | 97.6 | 10.87 | 10.32 | 10.12 | 11.84 | 10.94 |
| $X_{39}$ | 96.0 | 94.8 | 96.0 | 96.0 | 95.0 | 11.29 | 10.72 | 10.62 | 12.21 | 11.38 |

Tabelle 10: Coverageraten und Median Konfidenzintervallbreiten der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion.

| | Inferenzmethode = Boot-MI $\beta = \beta_{\text{Kliff}}$ SNR = 5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage in % | | | | | Median KI Breite | | | | |
| Variable | Lasso | Ad-Lasso | MSS | LPR | LP | Lasso | Ad-Lasso | MSS | LPR | LP |
| $X_1$ | 95.6 | 96.4 | 96.6 | 96.2 | 96.4 | 22.40 | 22.63 | 23.00 | 23.77 | 21.69 |
| $X_2$ | 97.0 | 96.6 | 95.8 | 98.0 | 97.0 | 32.06 | 30.87 | 29.65 | 34.69 | 31.48 |
| $X_3$ | 95.8 | 95.8 | 95.0 | 97.2 | 96.2 | 20.63 | 20.58 | 21.14 | 22.38 | 21.31 |
| $X_4$ | 95.6 | 95.6 | 94.6 | 95.6 | 95.4 | 18.47 | 18.25 | 14.03 | 21.12 | 20.10 |
| $X_5$ | 98.2 | 97.8 | 96.4 | 98.4 | 97.2 | 24.95 | 24.88 | 20.12 | 28.75 | 26.37 |
| $X_6$ | 96.8 | 96.2 | 97.4 | 97.4 | 95.6 | 19.73 | 19.12 | 15.23 | 22.46 | 21.36 |
| $X_7$ | 96.4 | 95.8 | 95.8 | 96.4 | 95.0 | 21.25 | 21.21 | 16.15 | 24.37 | 22.77 |
| $X_8$ | 97.0 | 96.8 | 97.8 | 97.2 | 96.8 | 29.62 | 28.34 | 22.97 | 33.86 | 29.76 |
| $X_9$ | 96.8 | 96.4 | 97.2 | 97.0 | 96.6 | 18.66 | 18.16 | 15.00 | 20.96 | 19.98 |
| $X_{10}$ | 97.2 | 96.4 | 97.6 | 97.2 | 96.2 | 17.67 | 17.16 | 13.26 | 20.42 | 19.34 |
| $X_{11}$ | 97.2 | 97.4 | 96.4 | 96.8 | 96.0 | 27.53 | 27.15 | 22.45 | 31.09 | 28.63 |
| $X_{12}$ | 96.6 | 96.2 | 95.4 | 96.8 | 95.4 | 21.07 | 20.17 | 16.15 | 24.06 | 22.41 |
| $X_{13}$ | 98.2 | 98.4 | 98.2 | 98.8 | 97.8 | 21.40 | 21.07 | 16.72 | 24.21 | 23.41 |
| $X_{14}$ | 96.8 | 97.2 | 97.0 | 98.0 | 97.4 | 15.96 | 15.75 | 15.38 | 16.86 | 15.49 |
| $X_{15}$ | 98.4 | 98.0 | 97.8 | 99.0 | 98.2 | 19.64 | 19.08 | 18.59 | 20.87 | 19.07 |
| $X_{16}$ | 96.4 | 96.4 | 95.8 | 97.2 | 96.6 | 14.44 | 14.72 | 14.92 | 15.41 | 14.60 |
| $X_{17}$ | 96.0 | 95.8 | 94.6 | 96.4 | 96.2 | 14.00 | 13.34 | 12.01 | 15.48 | 14.99 |
| $X_{18}$ | 93.6 | 94.0 | 94.4 | 95.2 | 93.4 | 16.31 | 15.80 | 14.11 | 18.29 | 17.41 |
| $X_{19}$ | 95.6 | 96.0 | 95.0 | 96.4 | 95.6 | 12.92 | 12.50 | 10.95 | 14.51 | 13.99 |
| $X_{20}$ | 97.4 | 96.0 | 95.8 | 97.6 | 96.4 | 12.49 | 12.01 | 10.33 | 14.16 | 13.63 |
| $X_{21}$ | 97.0 | 97.4 | 97.4 | 97.0 | 97.4 | 18.45 | 17.96 | 16.08 | 20.77 | 19.46 |
| $X_{22}$ | 96.6 | 96.2 | 96.6 | 96.8 | 96.6 | 13.51 | 12.99 | 11.19 | 15.18 | 14.70 |
| $X_{23}$ | 97.4 | 97.0 | 97.0 | 97.2 | 96.8 | 13.01 | 12.67 | 11.06 | 14.71 | 14.16 |
| $X_{24}$ | 96.4 | 96.0 | 95.8 | 96.2 | 95.8 | 16.48 | 15.87 | 14.47 | 18.38 | 17.44 |
| $X_{25}$ | 96.8 | 96.8 | 97.6 | 96.8 | 95.8 | 13.14 | 12.60 | 10.94 | 14.77 | 14.27 |
| $X_{26}$ | 97.2 | 98.4 | 97.0 | 97.6 | 97.0 | 12.37 | 11.76 | 10.31 | 13.87 | 13.34 |
| $X_{27}$ | 97.4 | 96.8 | 95.4 | 97.2 | 96.0 | 11.44 | 11.16 | 10.90 | 11.93 | 11.30 |
| $X_{28}$ | 98.0 | 96.8 | 96.4 | 96.0 | 95.0 | 15.97 | 15.49 | 15.10 | 16.75 | 15.36 |
| $X_{29}$ | 93.8 | 93.4 | 93.6 | 94.8 | 94.2 | 12.14 | 12.25 | 12.32 | 12.84 | 12.13 |
| $X_{30}$ | 97.8 | 96.8 | 95.8 | 98.6 | 98.2 | 10.73 | 9.93 | 9.38 | 11.87 | 11.43 |
| $X_{31}$ | 98.4 | 97.8 | 97.6 | 98.2 | 97.4 | 12.57 | 11.94 | 11.00 | 13.90 | 13.28 |
| $X_{32}$ | 96.6 | 94.4 | 96.8 | 97.2 | 96.2 | 10.95 | 10.43 | 9.40 | 12.26 | 11.84 |
| $X_{33}$ | 98.2 | 97.4 | 97.6 | 98.2 | 98.2 | 10.68 | 10.16 | 9.17 | 11.96 | 11.65 |
| $X_{34}$ | 98.0 | 97.6 | 97.4 | 98.2 | 98.0 | 14.50 | 13.85 | 12.82 | 16.10 | 15.19 |
| $X_{35}$ | 96.2 | 96.2 | 95.8 | 96.8 | 96.6 | 10.69 | 10.07 | 9.25 | 11.82 | 11.51 |
| $X_{36}$ | 96.2 | 95.2 | 96.0 | 96.4 | 96.2 | 10.67 | 10.07 | 9.36 | 11.97 | 11.64 |
| $X_{37}$ | 97.2 | 97.2 | 97.2 | 97.6 | 97.4 | 13.85 | 13.18 | 12.10 | 15.22 | 14.60 |
| $X_{38}$ | 96.6 | 96.4 | 96.6 | 97.2 | 96.6 | 10.89 | 10.30 | 9.36 | 12.32 | 11.77 |
| $X_{39}$ | 97.4 | 96.6 | 97.6 | 97.2 | 96.8 | 11.09 | 10.68 | 9.59 | 12.39 | 12.08 |

Tabelle 11: Coverageraten und Median Konfidenzintervallbreiten der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion.

| | Inferenzmethode = Boot-MI $\beta = \beta_{\text{Kliff}}$ SNR = 10 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coverage in % | | | | | Median KI Breite | | | | |
| Variable | Lasso | Ad-Lasso | MSS | LPR | LP | Lasso | Ad-Lasso | MSS | LPR | LP |
| $X_1$ | 96.0 | 96.2 | 95.4 | 96.0 | 95.4 | 17.11 | 16.68 | 16.83 | 17.93 | 16.22 |
| $X_2$ | 97.0 | 96.8 | 96.6 | 97.6 | 96.6 | 23.46 | 22.56 | 22.89 | 25.02 | 21.92 |
| $X_3$ | 95.8 | 95.2 | 95.2 | 96.4 | 96.2 | 16.15 | 16.21 | 16.82 | 17.26 | 16.12 |
| $X_4$ | 96.6 | 95.2 | 94.4 | 96.6 | 96.0 | 14.23 | 13.81 | 11.58 | 15.98 | 14.93 |
| $X_5$ | 97.6 | 97.4 | 96.4 | 98.0 | 97.0 | 18.94 | 18.23 | 15.60 | 21.31 | 19.15 |
| $X_6$ | 97.6 | 97.0 | 96.8 | 97.6 | 96.6 | 15.22 | 14.50 | 12.34 | 17.22 | 15.83 |
| $X_7$ | 99.4 | 98.4 | 99.0 | 99.0 | 98.4 | 16.42 | 15.92 | 13.05 | 18.65 | 16.99 |
| $X_8$ | 97.8 | 97.2 | 97.4 | 97.6 | 96.2 | 22.13 | 20.80 | 17.83 | 25.09 | 21.38 |
| $X_9$ | 96.8 | 96.4 | 96.4 | 97.6 | 96.4 | 14.33 | 13.64 | 12.05 | 16.02 | 14.81 |
| $X_{10}$ | 96.4 | 95.6 | 97.0 | 97.0 | 95.4 | 13.69 | 13.18 | 10.79 | 15.67 | 14.25 |
| $X_{11}$ | 97.8 | 97.8 | 97.8 | 97.6 | 97.4 | 20.31 | 19.90 | 17.15 | 22.96 | 20.63 |
| $X_{12}$ | 96.6 | 96.4 | 93.6 | 96.4 | 96.2 | 15.86 | 14.87 | 12.51 | 18.05 | 16.40 |
| $X_{13}$ | 97.8 | 98.0 | 95.6 | 97.0 | 97.2 | 16.34 | 15.81 | 13.49 | 18.49 | 17.06 |
| $X_{14}$ | 96.6 | 95.8 | 97.2 | 97.0 | 96.6 | 12.10 | 11.64 | 11.55 | 12.72 | 11.67 |
| $X_{15}$ | 95.8 | 95.6 | 95.2 | 95.8 | 95.2 | 13.35 | 12.66 | 12.36 | 13.98 | 12.52 |
| $X_{16}$ | 93.6 | 93.8 | 95.2 | 95.8 | 94.6 | 11.03 | 11.06 | 11.18 | 11.58 | 10.76 |
| $X_{17}$ | 96.6 | 96.0 | 95.0 | 97.2 | 96.6 | 10.55 | 9.84 | 9.39 | 11.59 | 11.05 |
| $X_{18}$ | 97.2 | 97.2 | 97.6 | 97.6 | 97.0 | 12.15 | 11.50 | 10.74 | 13.31 | 12.60 |
| $X_{19}$ | 96.0 | 95.8 | 97.0 | 96.4 | 96.0 | 9.94 | 9.51 | 8.69 | 11.07 | 10.47 |
| $X_{20}$ | 97.2 | 96.6 | 96.4 | 97.2 | 97.2 | 9.68 | 9.20 | 8.41 | 10.79 | 10.21 |
| $X_{21}$ | 97.6 | 97.0 | 97.6 | 97.6 | 97.4 | 13.64 | 13.03 | 12.19 | 15.03 | 13.95 |
| $X_{22}$ | 97.0 | 95.8 | 95.2 | 96.6 | 95.8 | 10.22 | 9.72 | 8.93 | 11.41 | 10.74 |
| $X_{23}$ | 97.0 | 96.6 | 96.2 | 97.0 | 96.4 | 10.02 | 9.50 | 8.62 | 11.13 | 10.39 |
| $X_{24}$ | 97.0 | 97.0 | 97.4 | 97.2 | 96.8 | 12.01 | 11.48 | 10.61 | 13.19 | 12.26 |
| $X_{25}$ | 96.0 | 95.2 | 95.0 | 96.2 | 95.2 | 10.12 | 9.73 | 8.64 | 11.26 | 10.56 |
| $X_{26}$ | 96.8 | 97.0 | 96.2 | 96.6 | 96.4 | 9.45 | 8.97 | 8.18 | 10.58 | 9.97 |
| $X_{27}$ | 97.2 | 97.0 | 96.0 | 96.8 | 96.0 | 8.67 | 8.40 | 8.25 | 8.96 | 8.41 |
| $X_{28}$ | 95.6 | 94.6 | 94.4 | 95.0 | 94.2 | 10.71 | 10.19 | 9.91 | 11.07 | 10.20 |
| $X_{29}$ | 96.4 | 96.6 | 97.2 | 97.6 | 97.2 | 9.31 | 9.24 | 9.24 | 9.74 | 9.15 |
| $X_{30}$ | 96.6 | 96.4 | 96.4 | 97.2 | 96.0 | 8.12 | 7.48 | 7.39 | 8.83 | 8.43 |
| $X_{31}$ | 96.6 | 96.8 | 97.0 | 97.0 | 96.6 | 9.24 | 8.64 | 8.27 | 10.20 | 9.63 |
| $X_{32}$ | 97.2 | 96.0 | 97.2 | 97.2 | 97.2 | 8.31 | 7.87 | 7.50 | 9.20 | 8.79 |
| $X_{33}$ | 96.0 | 95.8 | 95.8 | 96.8 | 95.4 | 8.12 | 7.68 | 7.28 | 9.03 | 8.54 |
| $X_{34}$ | 97.0 | 96.6 | 96.6 | 97.2 | 97.0 | 10.54 | 9.86 | 9.36 | 11.59 | 10.81 |
| $X_{35}$ | 95.8 | 95.4 | 96.0 | 95.8 | 95.4 | 8.12 | 7.60 | 7.35 | 8.91 | 8.53 |
| $X_{36}$ | 97.2 | 96.4 | 97.0 | 97.0 | 96.2 | 8.11 | 7.68 | 7.29 | 8.98 | 8.54 |
| $X_{37}$ | 96.4 | 95.8 | 96.6 | 97.2 | 97.0 | 9.99 | 9.36 | 8.87 | 10.99 | 10.30 |
| $X_{38}$ | 96.6 | 95.8 | 95.8 | 96.8 | 96.2 | 8.26 | 7.74 | 7.25 | 9.15 | 8.67 |
| $X_{39}$ | 96.8 | 96.6 | 97.6 | 97.0 | 96.8 | 8.54 | 8.08 | 7.60 | 9.41 | 8.90 |

Tabelle 12: Coverageraten und Median Konfidenzintervallbreiten der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion.

# B Grafiken

## MI-Boot (PS) Beta spärlich SNR 5

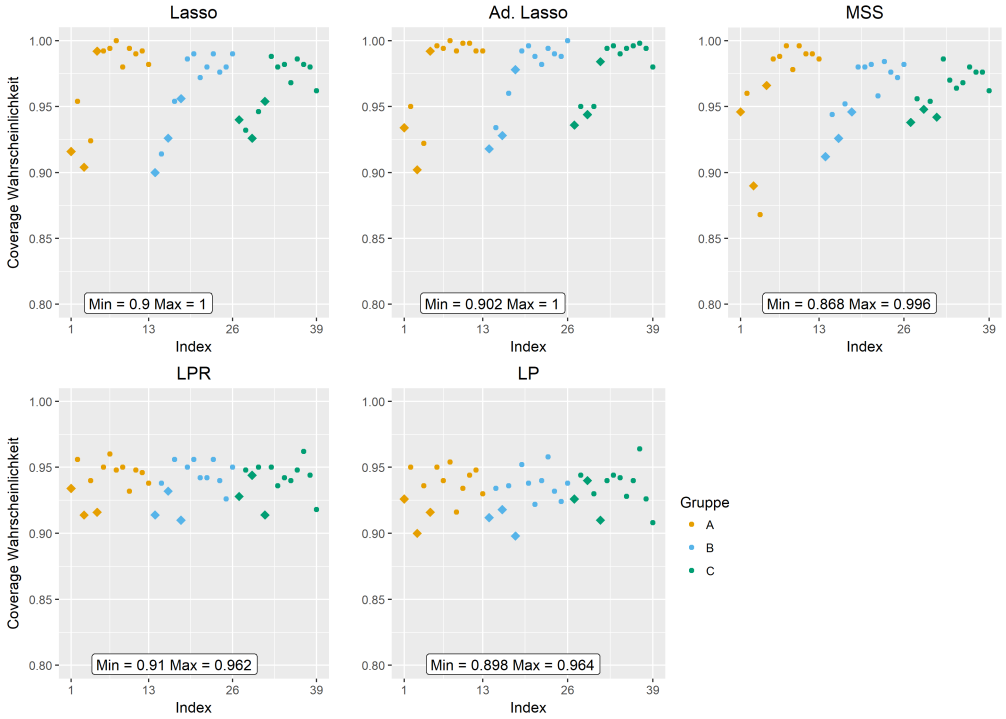


Abbildung 6: Coverage Wahrscheinlichkeiten der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

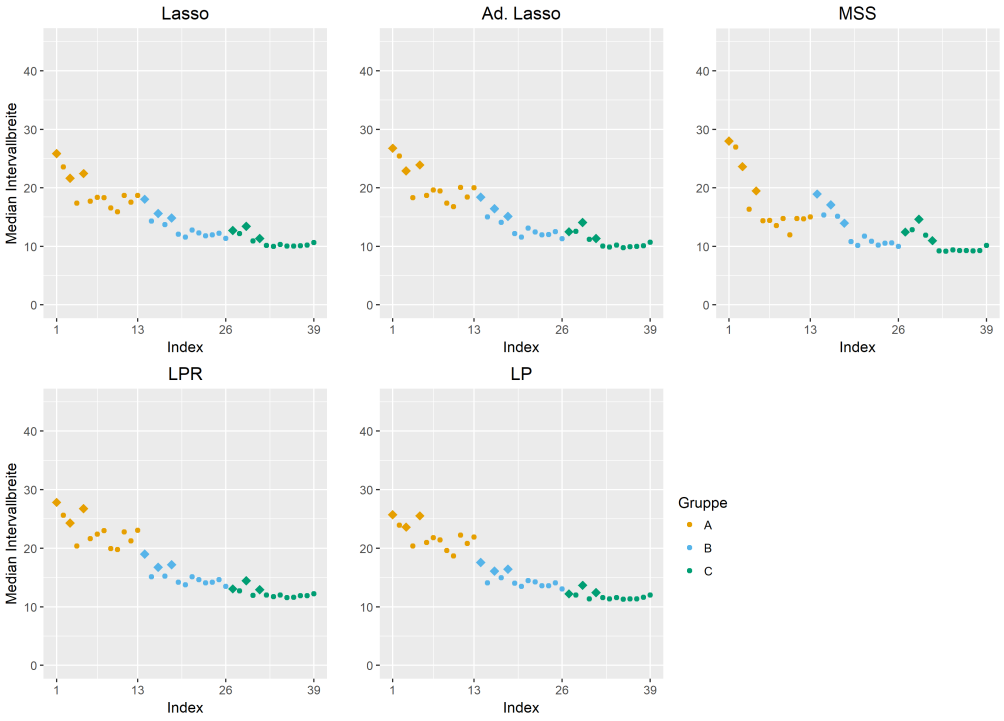


Abbildung 7: Median Konfidenzintervallbreite der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

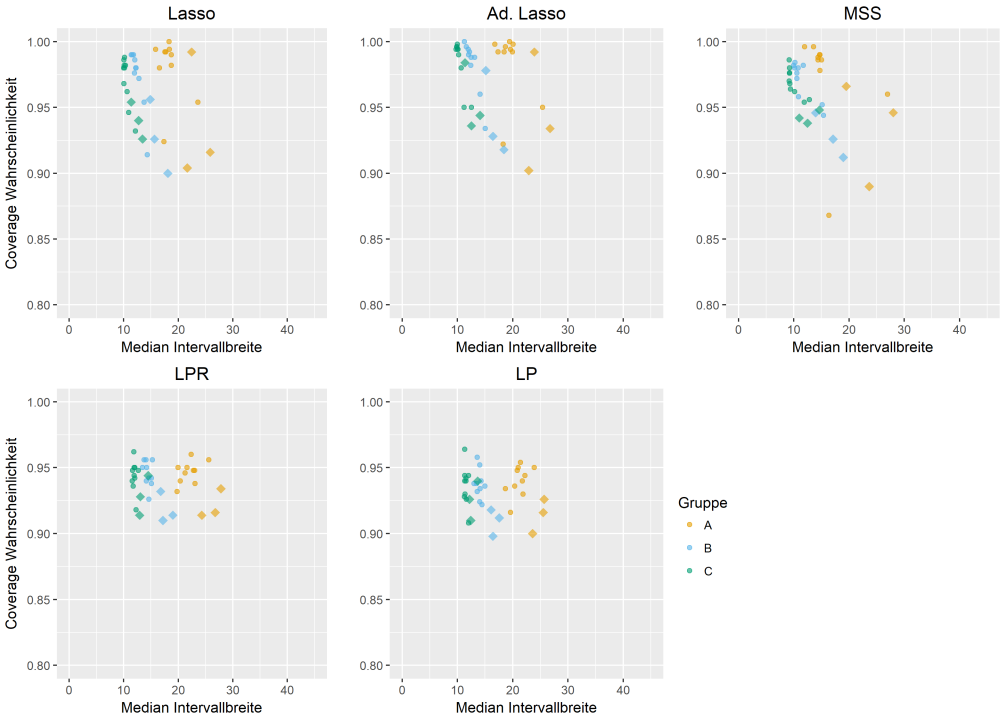


Abbildung 8: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

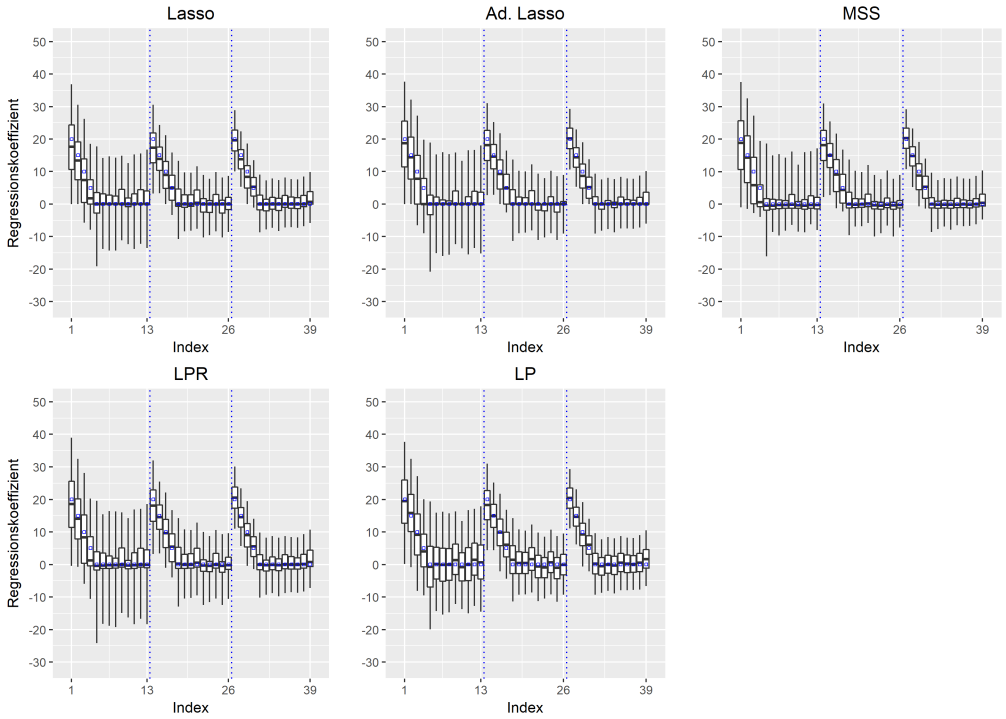


Abbildung 9: Boxplots für die Verteilung der Parameter-Schätzer der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Whiskers repräsentieren das 2.5% und das 97.5% Quantil. Ausreißer, die außerhalb des Whiskers liegen, sind nicht abgetragen. Die vertikalen Linien trennen die Variablen der Blöcke $A$, $B$ und $C$. Die wahren Beta-Werte sind als blaue Punkte eingezeichnet.

### MI-Boot (PS) Beta-spärlich SNR 10

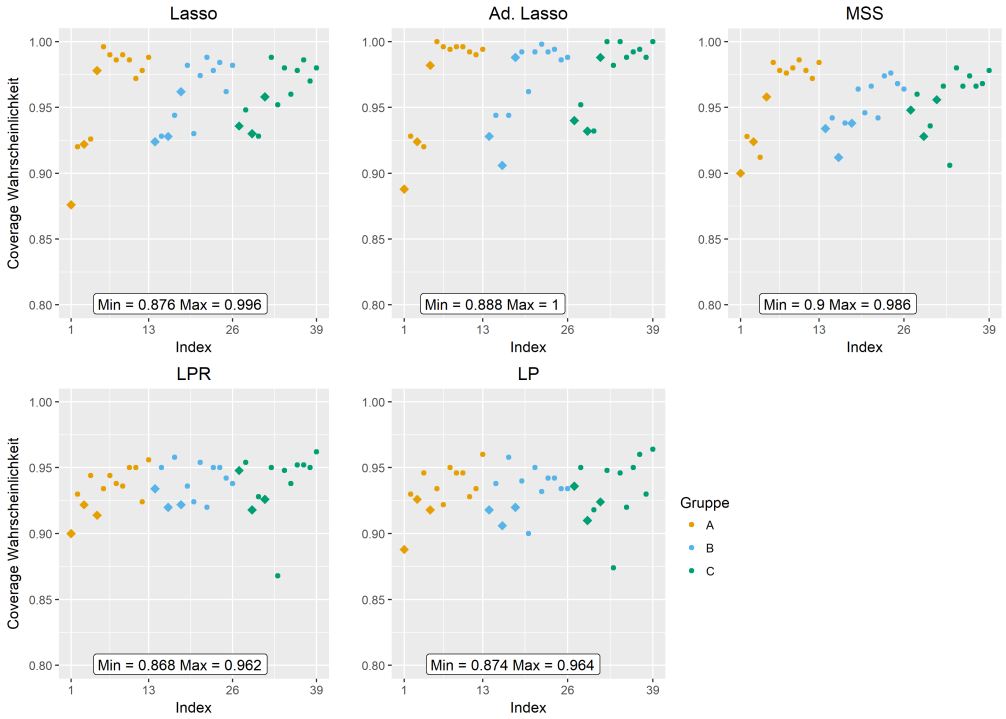


Abbildung 10: Coverage Wahrscheinlichkeiten der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.


Abbildung 11: Median Konfidenzintervallbreite der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

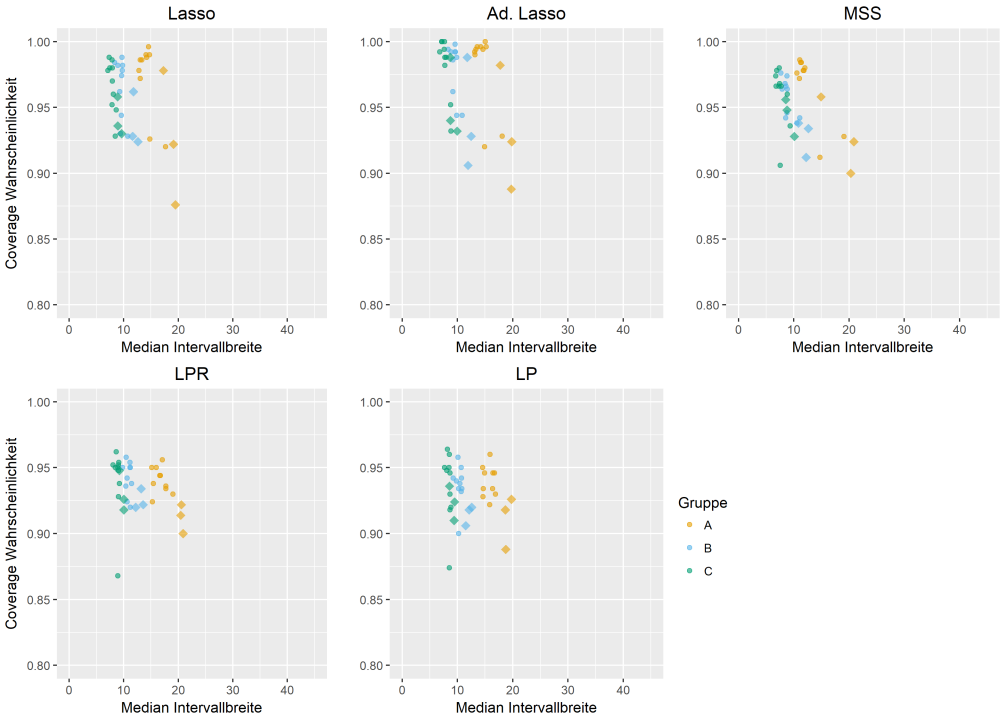


Abbildung 12: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

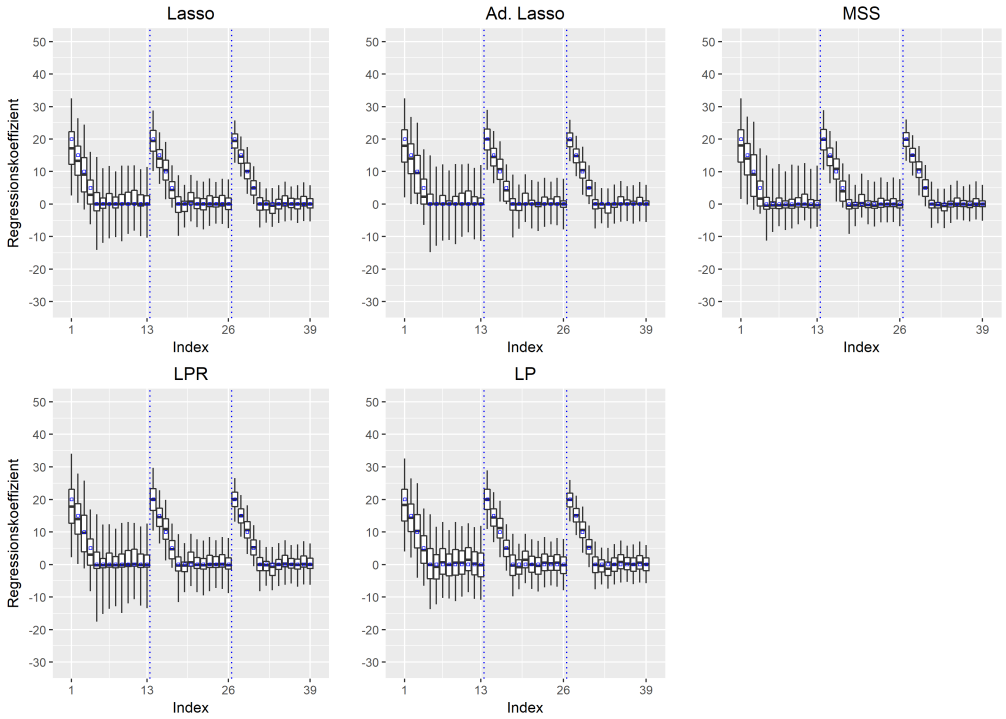


Abbildung 13: Boxplots für die Verteilung der Parameter-Schätzer der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Whiskers repräsentieren das 2.5% und das 97.5% Quantil. Ausreißer, die außerhalb des Whiskers liegen, sind nicht abgetragen. Die vertikalen Linien trennen die Variablen der Blöcke $A$, $B$ und $C$. Die wahren Beta-Werte sind als blaue Punkte eingezeichnet.

### MI-Boot (PS) Beta-Kliff SNR 5

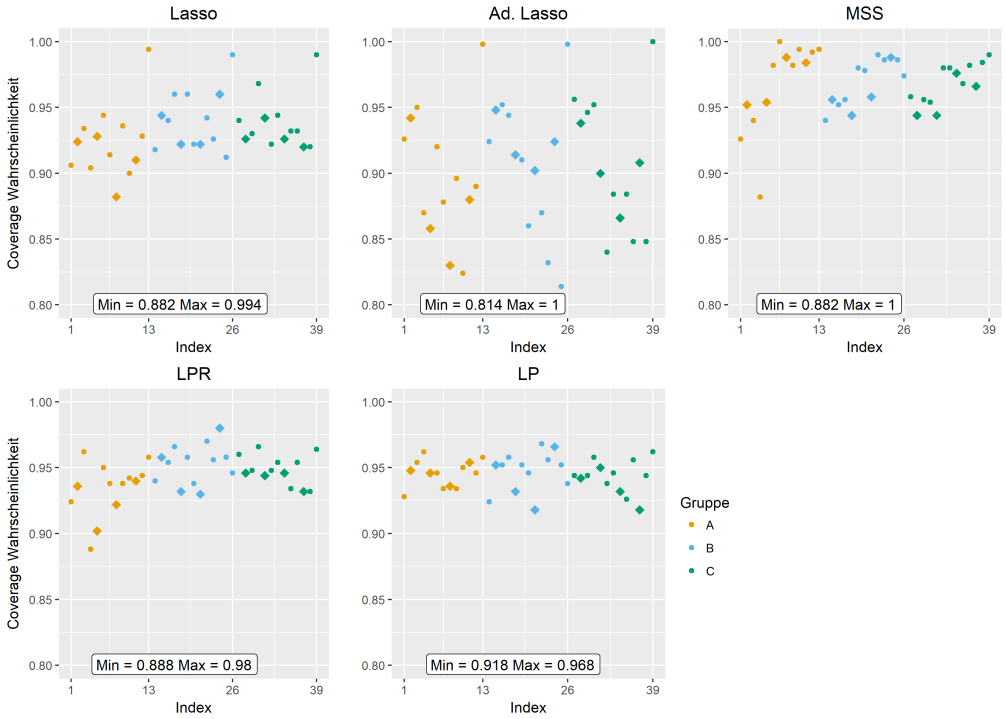


Abbildung 14: Coverage Wahrscheinlichkeiten der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

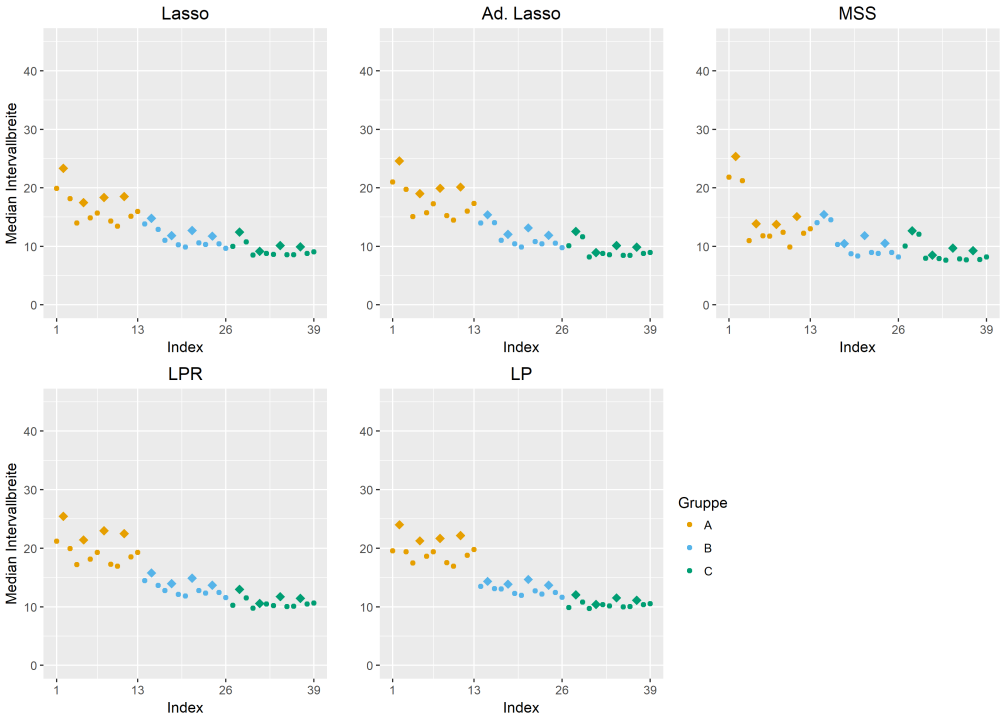


Abbildung 15: Median Konfidenzintervallbreite der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

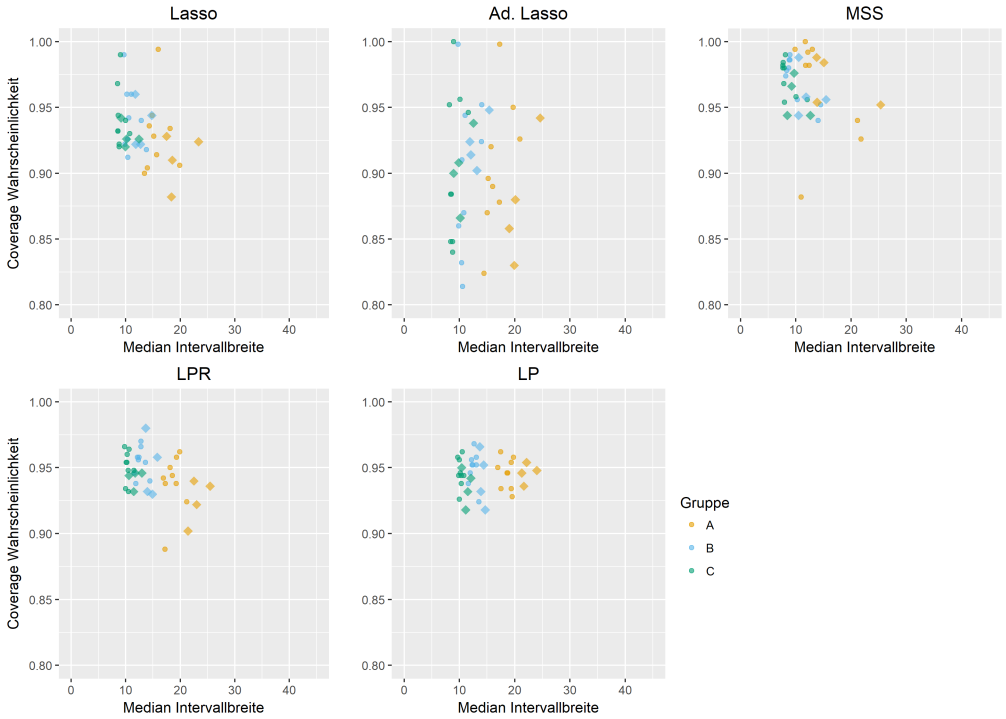


Abbildung 16: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

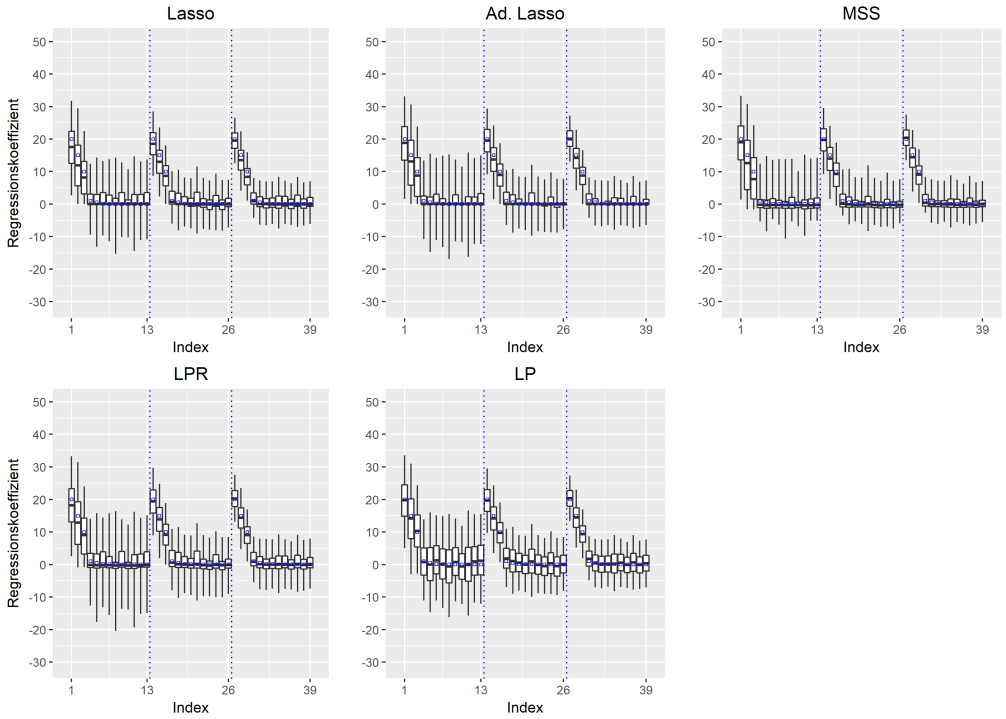


Abbildung 17: Boxplots für die Verteilung der Parameter-Schätzer der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Whiskers repräsentieren das 2.5% und das 97.5% Quantil. Ausreißer, die außerhalb des Whiskers liegen, sind nicht abgetragen. Die vertikalen Linien trennen die Variablen der Blöcke $A$, $B$ und $C$. Die wahren Beta-Werte sind als blaue Punkte eingezeichnet.

### MI-Boot (PS) Beta-Kliff SNR 10

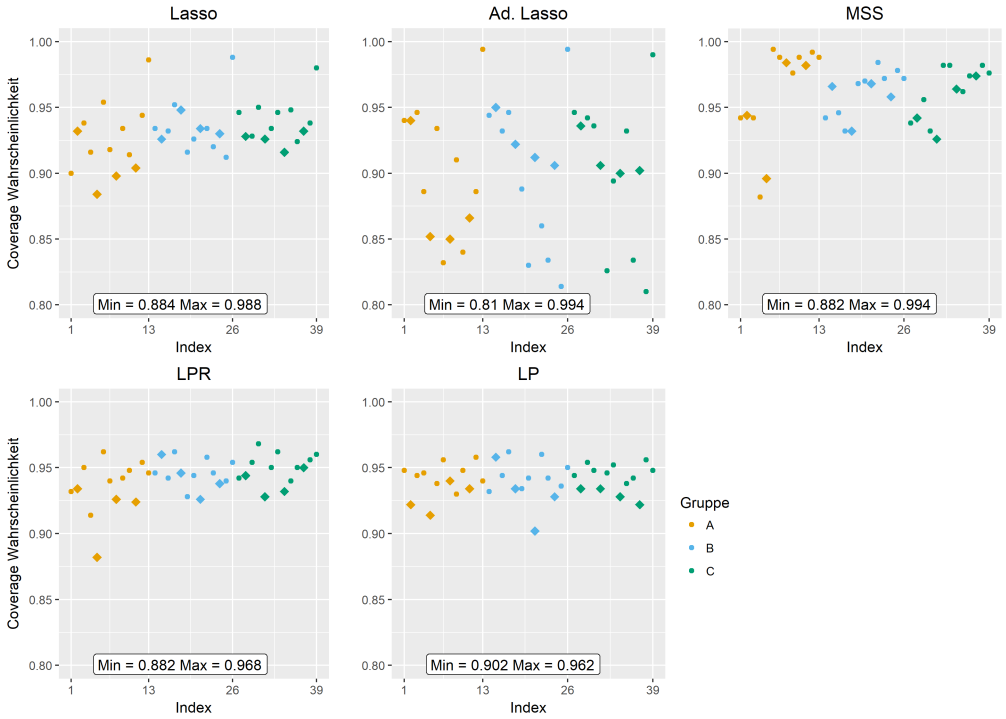


Abbildung 18: Coverage Wahrscheinlichkeiten der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

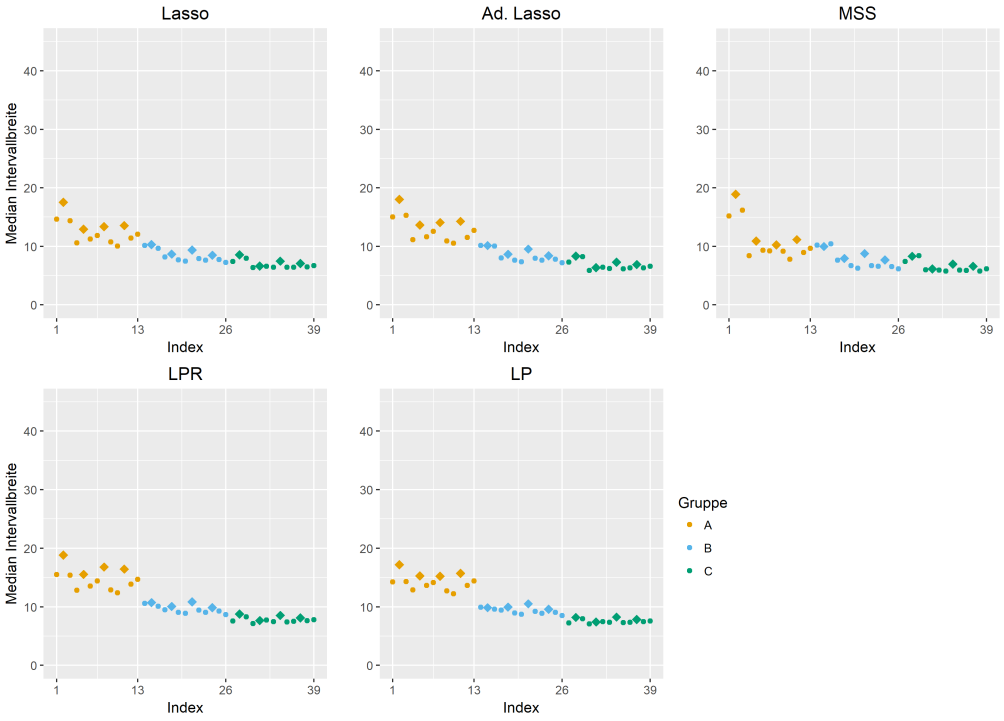


Abbildung 19: Median Konfidenzintervallbreite) der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

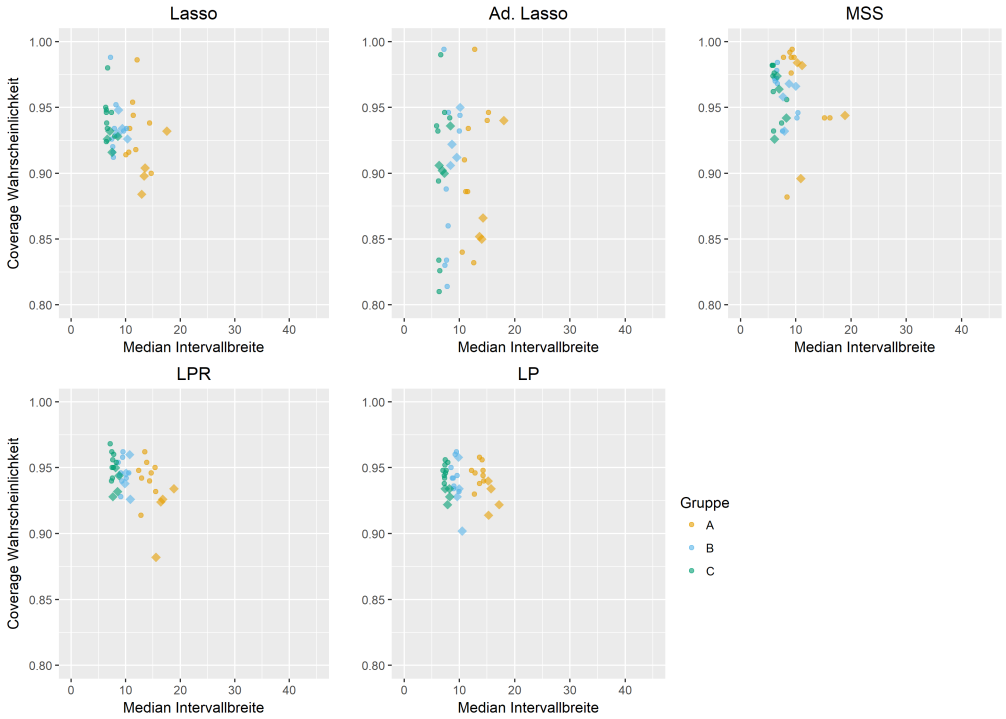


Abbildung 20: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

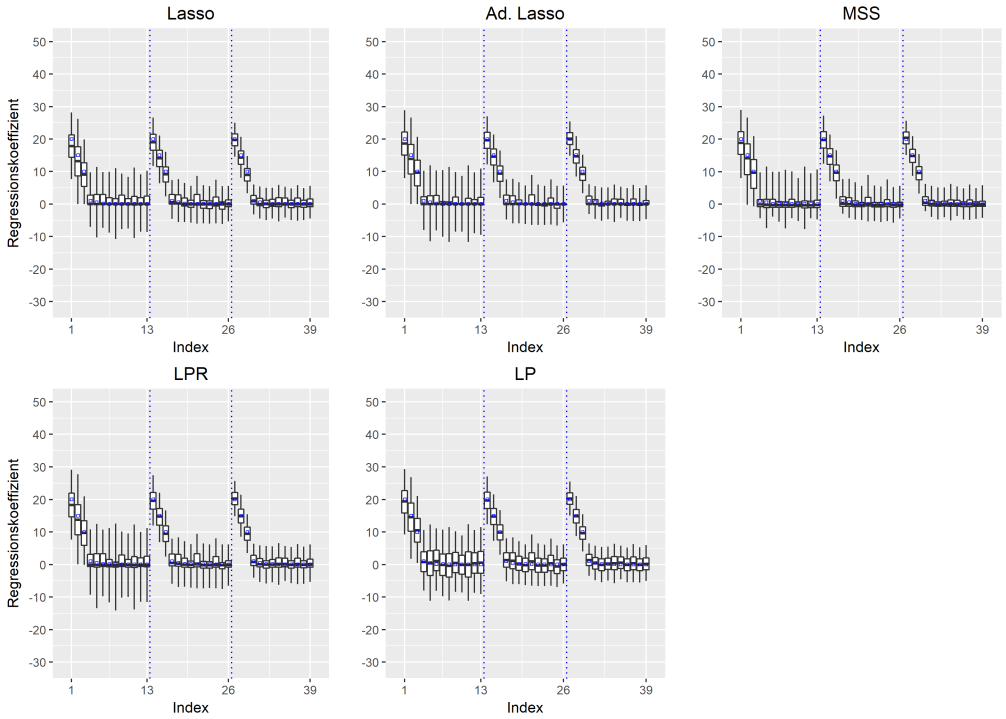


Abbildung 21: Boxplots für die Verteilung der Parameter-Schätzer der 500 Monte-Carlo-Iterationen für das MI-Boot (PS) Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Whiskers repräsentieren das 2.5% und das 97.5% Quantil. Ausreißer, die außerhalb des Whiskers liegen, sind nicht abgetragen. Die vertikalen Linien trennen die Variablen der Blöcke $A$, $B$ und $C$. Die wahren Beta-Werte sind als blaue Punkte eingezeichnet.

### Boot-MI Beta-spärlich SNR 5

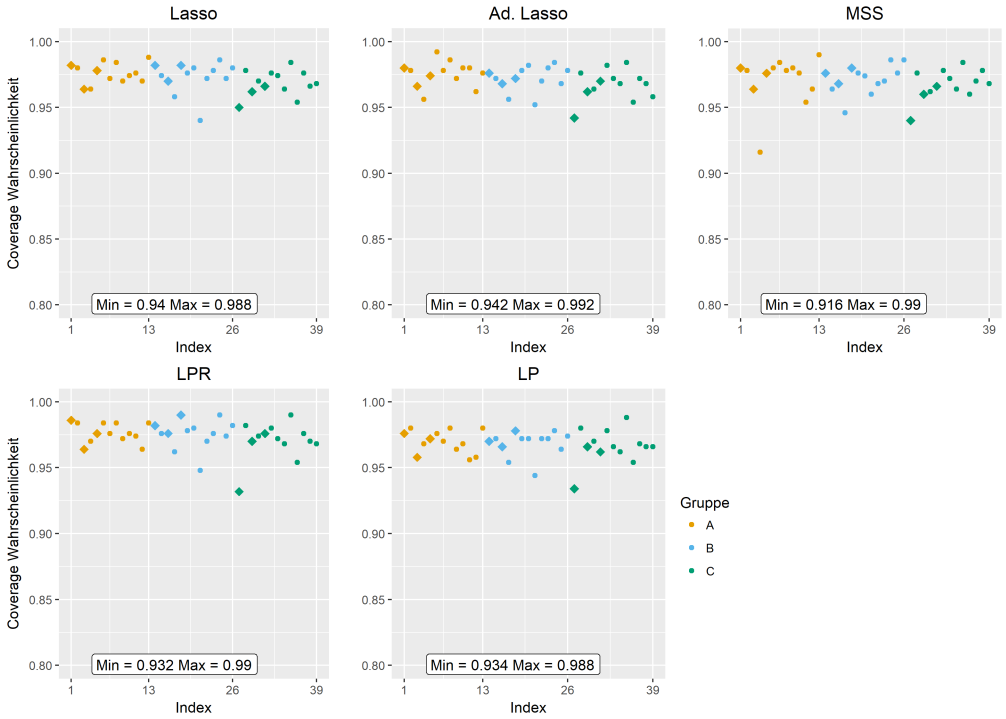


Abbildung 22: Coverage Wahrscheinlichkeiten der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.


Abbildung 23: Median Konfidenzintervallbreite der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 5. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

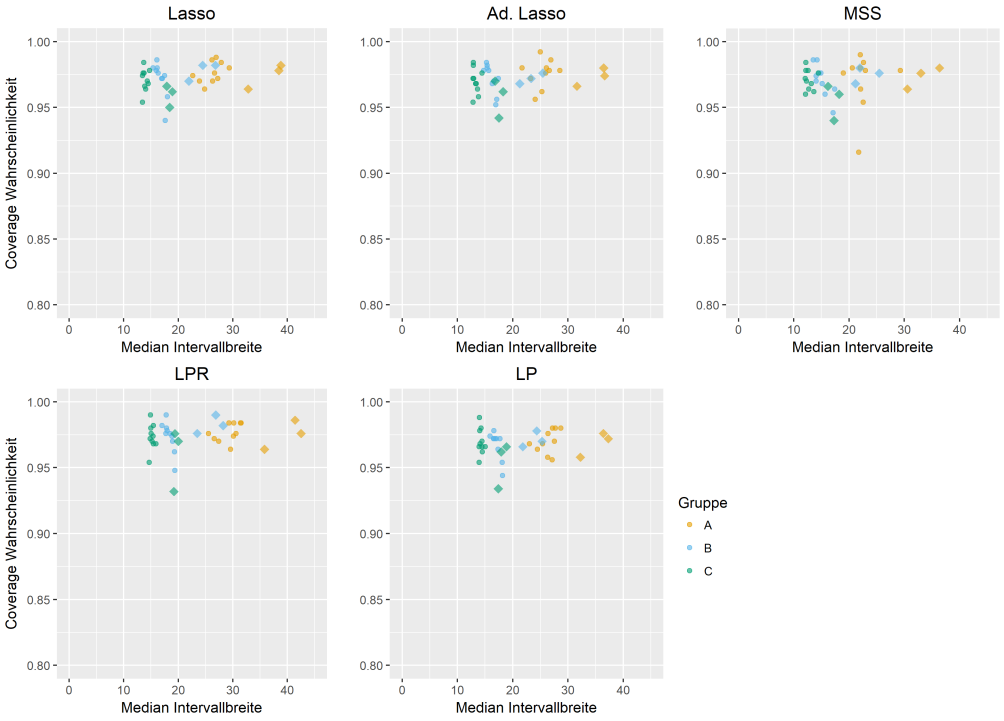


Abbildung 24: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

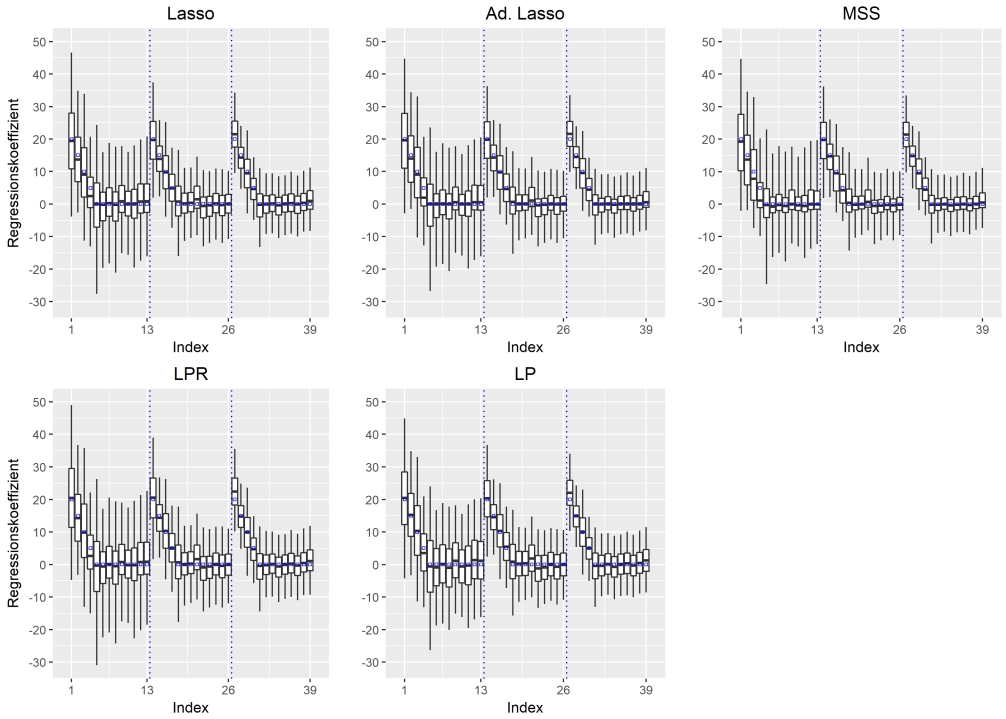


Abbildung 25: Boxplots für die Verteilung der Parameter-Schätzer der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 5$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Whiskers repräsentieren das 2.5% und das 97.5% Quantil. Ausreißer, die außerhalb des Whiskers liegen, sind nicht abgetragen. Die vertikalen Linien trennen die Variablen der Blöcke $A$, $B$ und $C$. Die wahren Beta-Werte sind als blaue Punkte eingezeichnet.

### Boot-MI Beta-spärlich SNR 10

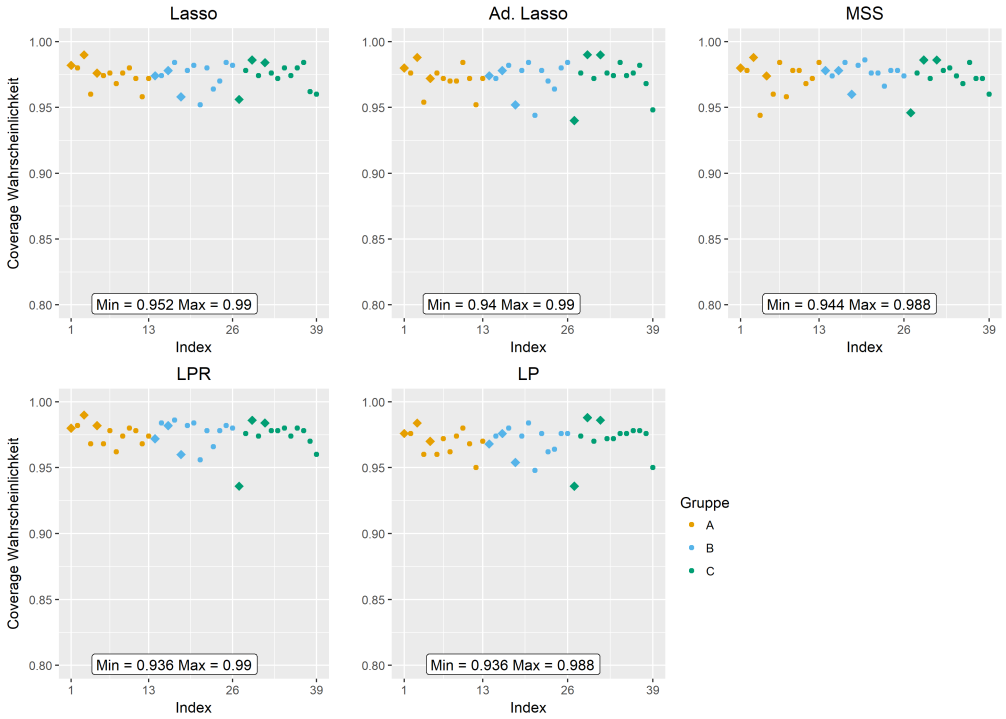


Abbildung 26: Coverage Wahrscheinlichkeiten der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

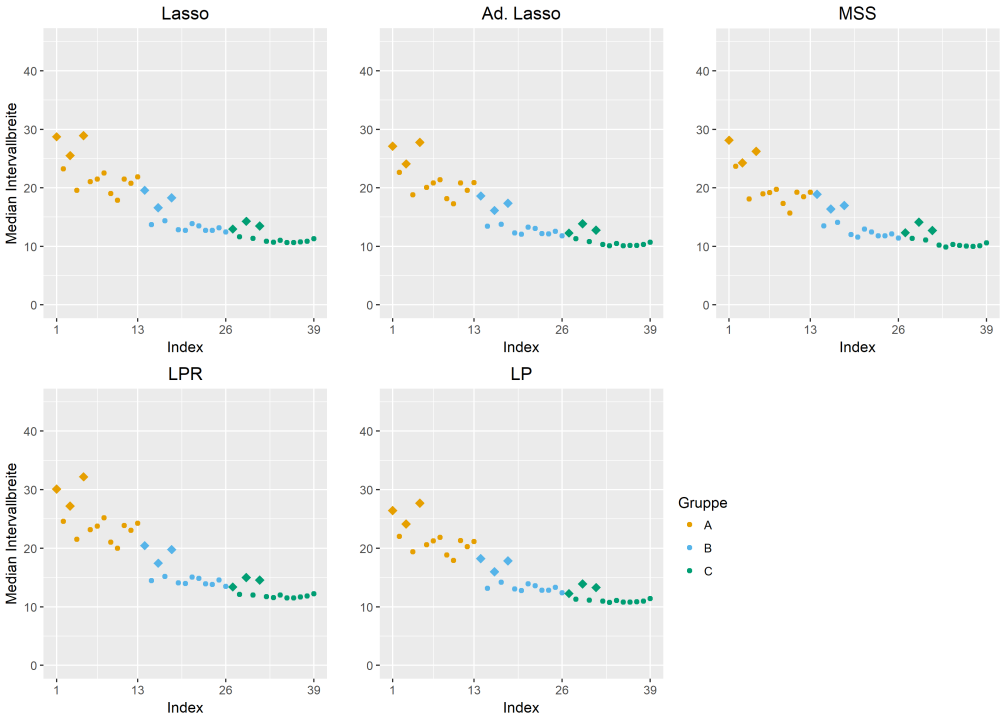


Abbildung 27: Median Konfidenzintervallbreite der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

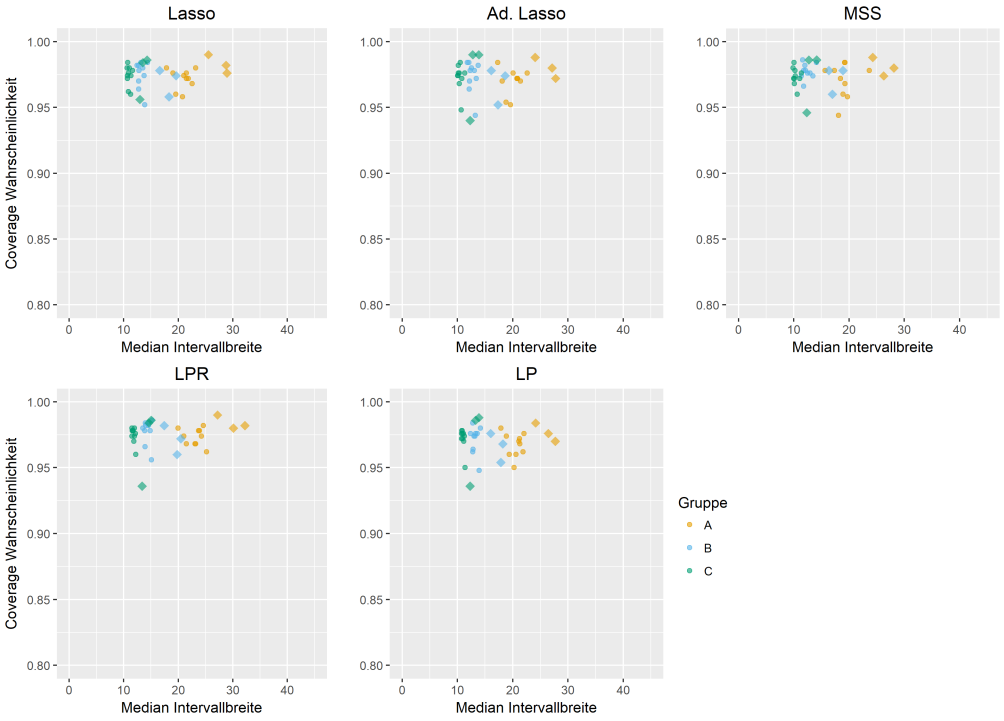


Abbildung 28: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

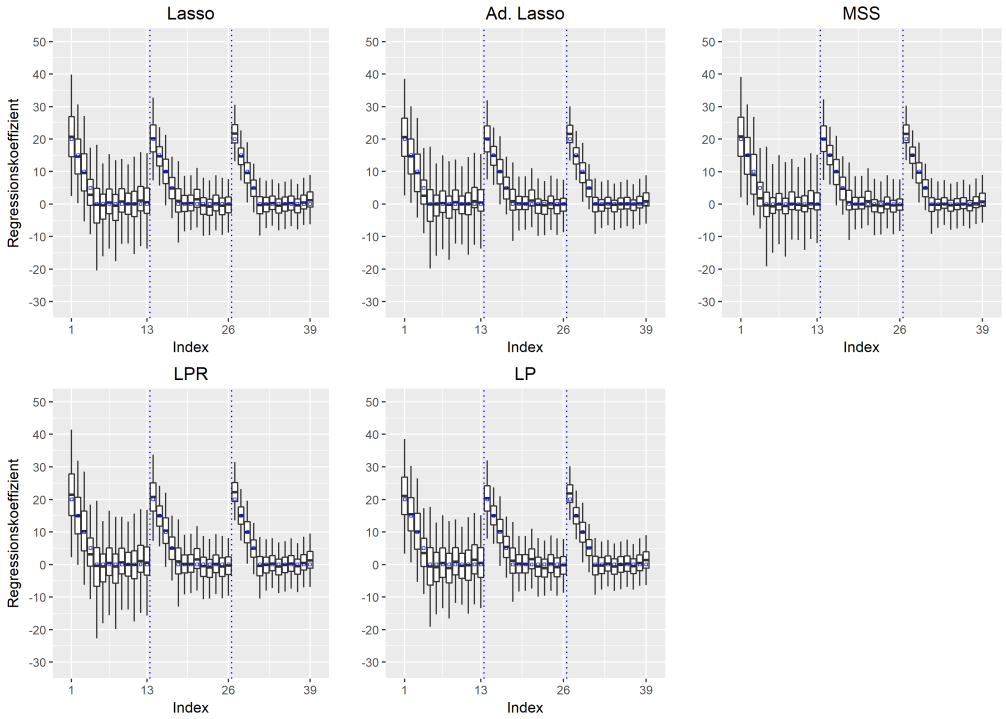


Abbildung 29: Boxplots für die Verteilung der Parameter-Schätzer der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{spärlich}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Whiskers repräsentieren das 2.5% und das 97.5% Quantil. Ausreißer, die außerhalb des Whiskers liegen, sind nicht abgetragen. Die vertikalen Linien trennen die Variablen der Blöcke $A$, $B$ und $C$. Die wahren Beta-Werte sind als blaue Punkte eingezeichnet.

### Boot-MI Beta-Kliff SNR 10

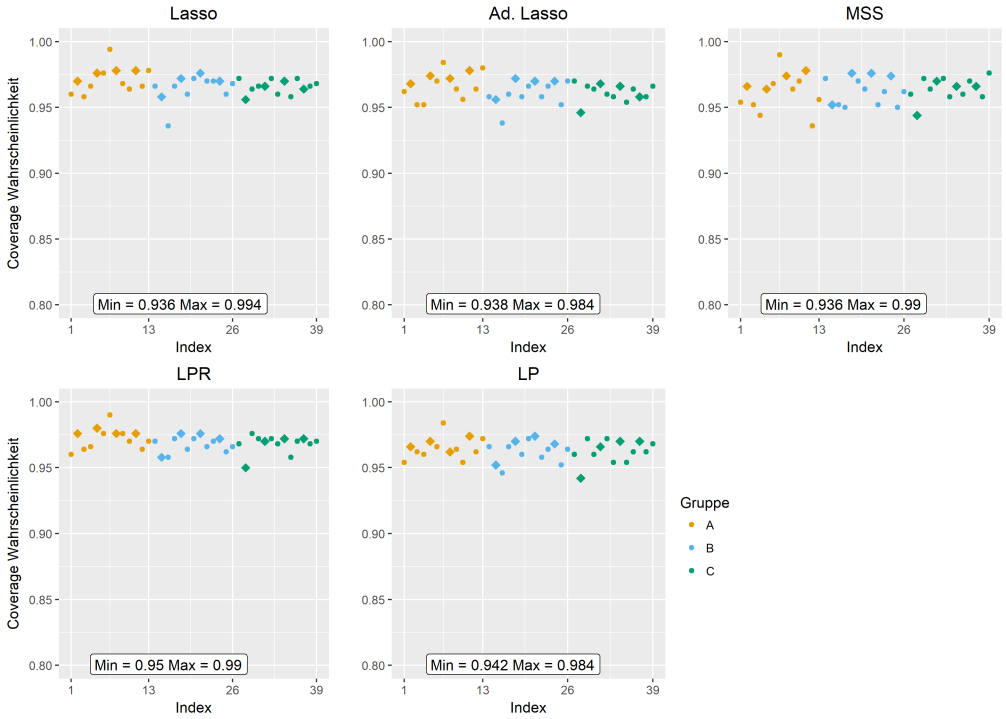


Abbildung 30: Coverage Wahrscheinlichkeiten der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

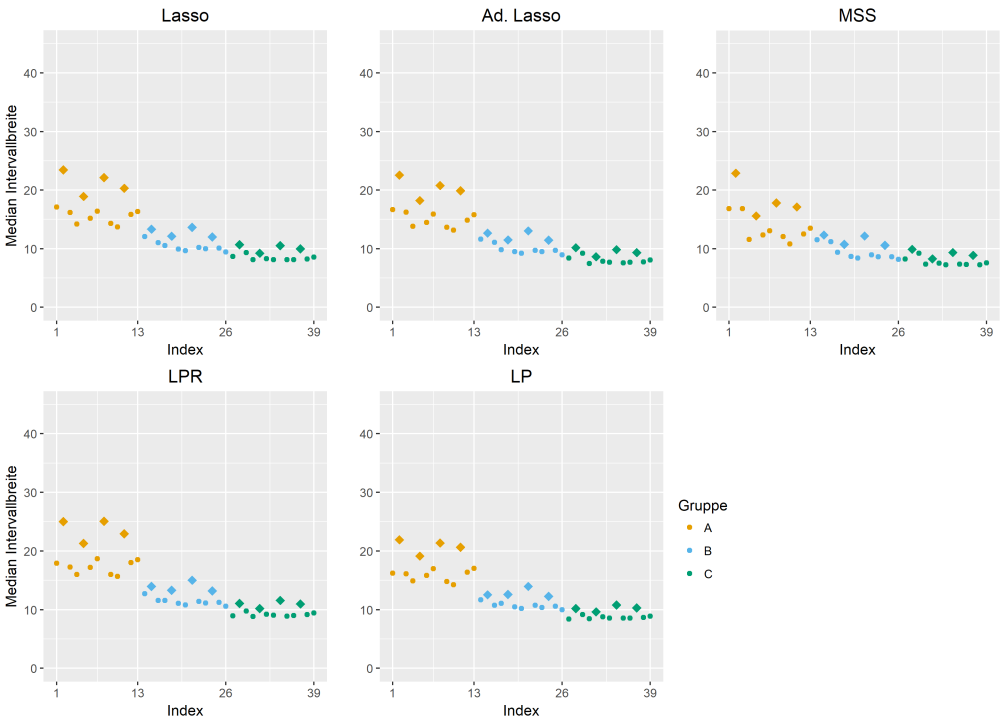


Abbildung 31: Median Konfidenzintervallbreite der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

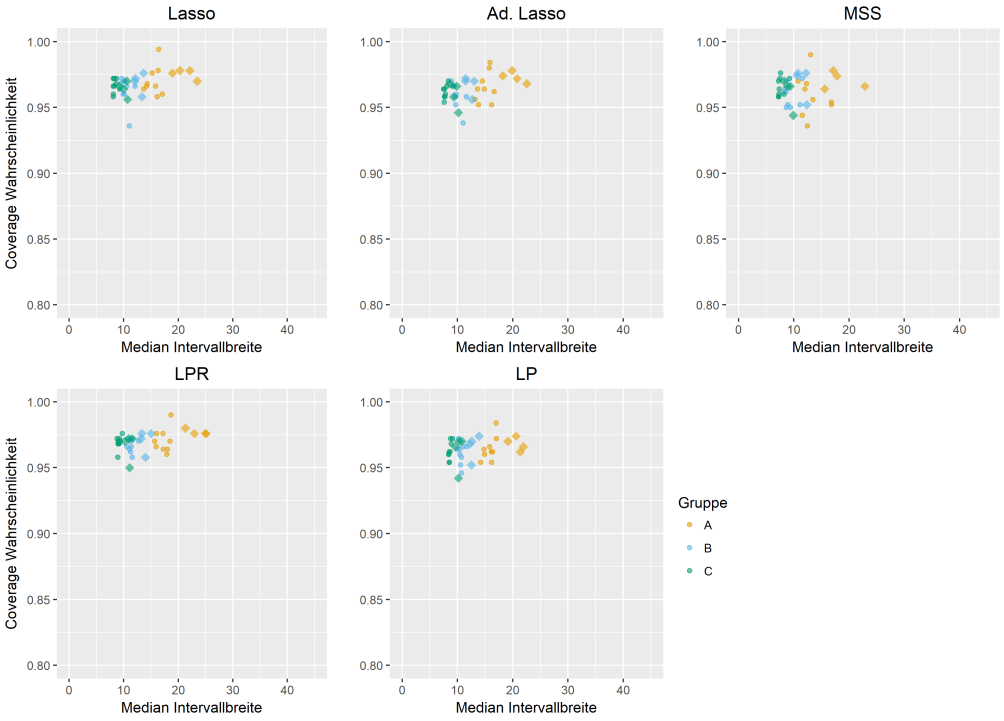


Abbildung 32: Median Konfidenzintervallbreite v.s. Coverage Wahrscheinlichkeit der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR = 10. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Variablen mit fehlenden Werten sind als Rauten dargestellt.

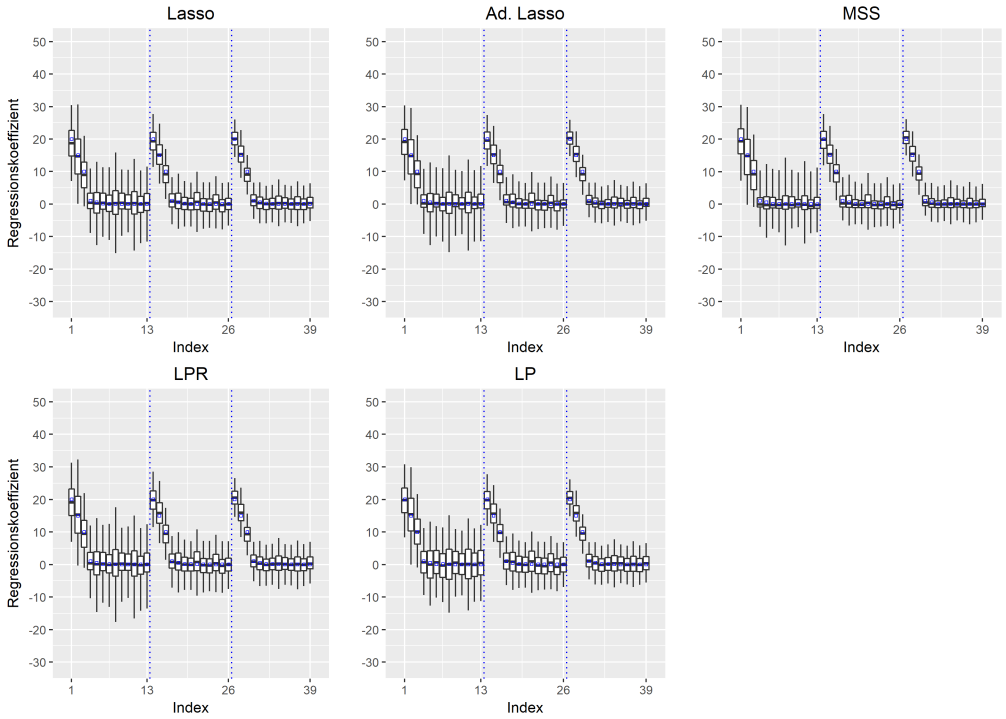


Abbildung 33: Boxplots für die Verteilung der Parameter-Schätzer der 500 Monte-Carlo-Iterationen für das Boot-MI Verfahren, $\beta = \beta_{\text{Kliff}}$ und SNR $= 10$. Die Abkürzungen Ad-Lasso, MSS, LPR und LP stehen für adaptive Lasso, Multi-Sample-Splitting, Lasso-Partial-Ridge und Lasso-Projektion. Whiskers repräsentieren das 2.5% und das 97.5% Quantil. Ausreißer, die außerhalb des Whiskers liegen, sind nicht abgetragen. Die vertikalen Linien trennen die Variablen der Blöcke $A$, $B$ und $C$. Die wahren Beta-Werte sind als blaue Punkte eingezeichnet.

## C Elektronischer Anhang

Der elektronische Anhang besteht aus:

- Bericht in PDF Format
- Kommentierter R Code für die Simulationsstudie
- Objekte der Simulationsstudie als RData Dateien
- Alle in dieser Arbeit genutzten Grafiken

## Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe. Die Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form vorgelegt. Sie wurde bisher nicht veröffentlicht.

München, den 14. März 2018