technische universität
dortmund

Effiziente Variablenselektion aus großen SNP Daten mittels approximierter Cross-Leverage Scores

Bachelorarbeit

von Rieke Deborah Möller-Ehmcke

27. April 2022

Gutachter 1: Dr. Alexander Munteanu
Gutachterin 2: Prof. Dr. Katja Ickstadt

Inhaltsverzeichnis

1 Einleitung

Bei Untersuchungen in der Genetik werden oft sehr viele Variablen erhoben, während die Zahl der beobachteten Individuen, beispielsweise aus Kostengründen, vergleichsweise klein ist. Gleichzeitig ist man nicht nur daran interessiert, wie sich einzelne Variablen auf ein Merkmal auswirken, sondern insbesondere auch daran, welchen Einfluss bestimmte Variablenkombinationen haben. Bei Ruczinski, Kooperberg und LeBlanc (2003) wird die logische Regression als Methode vorgestellt, mit der solche Variablenkombinationen bei binären Einflussvariablen untersucht werden können. Diese Methode ist jedoch abhängig von der Variablenzahl sehr zeitaufwändig. Um die Zahl der Variablen vor der Durchführung der logischen Regression zu reduzieren, wird bei Parry u. a. (2021) vorgeschlagen, sogenannte Leverage und Cross-Leverage Scores zur Selektion relevanter Variablen zu verwenden. Dieser Ansatz wird hier aufgegriffen und an Datenbeispielen angewandt. Die Cross-Leverage Scores, welche jeweils zu einer Einflussvariablen und der Zielvariablen gehören, werden genauer betrachtet, da diese bei den Daten bei Parry u. a. (2021) als Selektionskriterien gut zu funktionieren scheinen. Für Datensätze, die so groß sind, dass auch die exakte Berechnung der (Cross-)Leverage Scores schwierig wird, gibt es Möglichkeiten, diese effizient zu approximieren. Auch dies wird an Datenbeispielen betrachtet. Es werden in dieser Arbeit zwei Fragestellungen untersucht. Zuerst geht es darum, inwiefern Leverage und Cross-Leverage Scores dazu geeignet sind, in bestimmten Datensätzen als Variablenselektionskriterium verwendet zu werden. Anschließend werden Approximationen der Cross-Leverage Scores in diesem Zusammenhang studiert. Dazu werden Daten aus dem internationalen HapMap Projekt, sowie simulierte binäre Daten betrachtet.

1.1 Übersicht der Kapitel

Nachdem die Fragestellungen der Arbeit noch einmal formuliert werden, werden kurz die biologischen Grundlagen bezüglich SNP Daten erläutert. Anschließend wird ein Überblick über die verwendeten Methoden gegeben. Unter anderem wird dort erläutert, wie Leverage und Cross-Leverage Scores hier berechnet und zur Variablenselektion eingesetzt werden. Außerdem wird dargelegt, wie die-

se (Cross-)Leverage Scores mit weniger Rechenaufwand approximativ bestimmt werden können. Im nächsten Abschnitt werden sowohl die betrachteten echten Daten aus dem HapMap Projekt, als auch die simulierten Daten beschrieben. In der Auswertung wird erst die Variablenselektion mit den exakt bestimmten (Cross-)Leverage Scores analysiert. Anschließend wird untersucht, wie gut die Ergebnisse mit den Approximationen angenähert werden können. Eine kurze Zusammenfassung der gewonnenen Erkenntnisse und ihrer Bedeutung für weitergehende Untersuchungen schließt diese Arbeit ab.

1.2 SNPs

Der folgende Abschnitt basiert auf Tomiuk und Loeschcke (2017, Kap. 2). Ein großer Teil der Eigenschaften von Lebewesen lässt sich durch ihr Erbgut erklären. Träger dieses Erbguts ist die Desoxyribonukleinsäure, kurz DNS oder englisch DNA. Die DNA ist ein Molekül, dessen Grundbausteine die Basen Adenin, Cytosin, Guanin und Thymin sind. Die Verbindung aus einer Base, einem Zucker und Phosphatresten heißt Nukleotid. Die DNA hat die Struktur einer Doppelhelix. Sie besteht aus zwei Nukleotidketten, bei der jedes Nukleotid aus einer Kette mit einem bestimmten Nukleotid in der anderen Kette verbunden ist. Eine solche Verbindung zweier Nukleotide nennt man auch Basenpaar. Ein Basenpaar kann entweder die Basen Adenin und Thymin oder Cytosin und Guanin enthalten. Ein fester DNA-Abschnitt heißt Locus. Die Erbinformation an einem bestimmten Locus heißt Allel (Tomiuk und Loeschcke 2017, Kap.1). Wie bei vielen anderen Organismen, gibt es die meisten Gene der Menschen zweifach, da sie von Mutter und Vater vererbt werden (Graw 2015, Kap. 1). Liegt in beiden Fällen das gleiche Allel vor, so ist das entsprechende Merkmal homozygot. Unterscheiden die Allele von Vater und Mutter sich, so spricht man von einem heterozygoten Merkmal (Tomiuk und Loeschcke 2017, Kap. 1). Die menschliche DNA enthält mehr als drei Milliarden Basenpaare (Kruglyak und Nickerson 2001). Um nicht alle dieser Basenpaare untersuchen zu müssen, können sogenannte SNPs betrachtet werden. Diese werden in Tomiuk und Loeschcke (2017, Kap. 3.6.4) beschrieben. Ein Single Nucleotide Polymorphism, kurz SNP, gesprochen Snip ist ein Locus, an dem in einer Bevölkerung mindestens zwei verschiedene

Basenpaare vorkommen. Dabei darf der Anteil der häufigen Variante in einer Population nicht größer als 99% sein. Dies unterscheidet SNPs von Mutationen. Ein großer Teil der Variation menschlicher Erbinformation lässt sich durch SNPs erklären (Consortium u. a. 2015). In dieser Arbeit werden ausschließlich Merkmale betrachtet, deren Ausprägungen binär kodiert werden können. Dies ist beispielsweise der Fall, wenn untersucht wird, ob ein Individuum an einer bestimmten Krankheit erkrankt ist oder nicht. Untersuchungen mit SNPs sind im Allgemeinen jedoch nicht auf derartige Merkmale beschränkt.

2 Notation und Methoden

Wir betrachten im Folgenden eine Matrix $\boldsymbol{X} \sim n \times p$, die aus n Beobachtungen mit jeweils p Variablen $X_1, ..., X_p$ besteht. Zum Teil wird eine Variable X_i auch als Variable i bezeichnet, wenn dies nicht zu Verwechslungen führt. Der Vektor der Zielvariablen ist $\boldsymbol{y} \in \mathbb{R}^n$. Die Matrix $\begin{pmatrix} \boldsymbol{X} & \boldsymbol{y} \end{pmatrix}^T$ wird mit $\tilde{\boldsymbol{X}}$ bezeichnet. Die i-te Zeile einer Matrix $\boldsymbol{A}$ wird mit $\boldsymbol{A}_{i\cdot}$ und die j-te Spalte entsprechend mit $\boldsymbol{A}_{\cdot j}$ notiert.

2.1 Logische Regression

Eine Methode, die bei binären Daten angewandt wird, um den Einfluss von Variableninteraktionen auf die Zielvariable zu untersuchen, ist die logische Regression. Die folgende Beschreibung basiert auf Ruczinski, Kooperberg und LeBlanc (2003). Das hier betrachtete Modell ist eine Vereinfachung des dort aufgestellten Modells für den Spezialfall einer binären Zielvariable (Parry u. a. 2021). Es hat die Form

$$g(E(\boldsymbol{y})) = \beta_0 + \beta_1 L.$$

L ist dabei eine boolesche Verbindung einiger der erhobenen Einflussvariablen. $X_1, ..., X_p$. $L = (X_1 \wedge X_2) \vee X_3^c$ ist demnach beispielsweise genau dann 1, wenn $X_1 = 1$ und $X_2 = 1$ oder $X_3 = 0$ ist. Zur Unterscheidung zwischen den erhobenen Einflussvariablen $X_1, ..., X_p$ und der booleschen Verbindung L wird L in dieser Arbeit als logische Einflussvariable bezeichnet. Es soll die logische Kombination L gefunden werden, welche die für die Ausprägung der Zielvariablen

relevanten Variableninteraktionen beschreibt. Die Zahl der logischen Einflussvariablen wächst als doppelte Exponentialfunktion von p, wird also sehr schnell sehr groß (Parry u. a. 2021). Aus diesem Grund ist es gerade bei Daten mit sehr vielen Variablen notwendig, vor der Anwendung der logischen Regression die Variablenzahl zu reduzieren.

2.2 Selektionskriterien

2.2.1 Cross-Leverage und Leverage Scores

Cross-Leverage Scores, kurz CLS, und Leverage Scores, kurz LS, werden über die sogenannte Hat-Matrix definiert. Bei einem linearen Modell der Form $y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ wird sie bei $n > p$ und $Rang(X) = p$ mit folgender Formel berechnet:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

(Hoaglin und Welsch 1978). Die Leverage Scores sind dann die Diagonalelemente dieser Matrix, die Cross-Leverage Scores sind die Nebendiagonalelemente. Die Berechnung von $\boldsymbol{H}$ kann für binäre Datenmatrizen auf die gleiche Weise durchgeführt werden. Allein die Interpretation der (Cross-)Leverage Scores ist nicht exakt übertragbar (Parry u. a. 2021). In dieser Arbeit liegt der Fokus auf Matrizen mit $n \ll p$. Außerdem sollen insbesondere auch die CLS der Einflussvariablen mit der Zielvariable betrachtet werden. Anstatt der Matrix $\boldsymbol{X}$ wird daher die oben eingeführte Matrix $\tilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{y} \end{pmatrix}^T$ zur Berechnung von $\boldsymbol{H}$ verwendet. Hier gilt also

$$\boldsymbol{H} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T$$

(Parry u. a. 2021). Bei Parry u. a. (2021) wird gezeigt, dass für eine andere, numerisch stabilere Berechnung der Hat-Matrix eine QR-Zerlegung bestimmt werden kann. Dazu wird $\tilde{\boldsymbol{X}}$ in das Produkt $\boldsymbol{QR}$ einer orthogonalen Matrix $\boldsymbol{Q}$ und einer oberen Dreiecksmatrix $\boldsymbol{R}$ zerlegt (Karpfinger 2014). Es gibt verschiedene Methoden, eine QR-Zerlegung zu erhalten, in R werden dazu Householder Matrizen verwendet (Schreiber und Van Loan 1989). Für die QR-Zerlegung wird die Funktion `qr()` in R verwendet (R Core Team 2021). Die Cross-Leverage Scores, die hier von Interesse sind, sind die zwischen jeweils einer Einflussvariablen

und der Zielvariablen, also $c_{i(p+1)}, i = 1, ..., p$. Als Kurzform von LS und CLS wird hier die Schreibweise (C)LS verwendet. Die (C)LS werden mit R bestimmt (R Core Team 2021).

2.2.2 Korrelation

Ein weiteres Mittel, um nach relevanten Variablen zu suchen, ist die Betrachtung der Korrelationen mit der Zielvariablen (Parry u. a. 2021). Dabei wird der Korrelationskoeffizient von Pearson der jeweiligen Einflussvariable X_i mit der Zielvariablen Y bestimmt. In Parry u. a. (2021) wird gezeigt, dass die Selektion über die Korrelation im Allgemeinen nicht zu derselben Variablenauswahl wie die Selektion mit CLS führt. Ein Vergleich beider Kriterien ist also sinnvoll.

2.3 Auswahl der Variablen

In jedem Datensatz werden jeweils $n \log(n)$ der Variablen ausgewählt. Dies entspricht der als Mindestwert angegebenen Zahl in Parry u. a. (2021). Bei Verwendung der Cross-Leverage Scores zur Selektion werden die Variablen mit den betragsmäßig größten CLS ausgewählt. Auch dies entspricht dem Vorgehen bei Parry u. a. (2021) und wird durch die Erkenntnisse bei Wollenberg (2016) motiviert. Die Auswahl über die Leverage Scores erfolgt über die kleinsten LS. Analog werden in den entsprechenden Verfahren die Variablen mit den betragsmäßig größten Korrelationen ausgewählt.

2.4 Approximative Berechnung von Leverage und Cross-Leverage Scores

Ein menschliches Genom enthält etwa 11 Millionen SNPs (Kruglyak und Nickerson 2001). Auch die QR-Zerlegung wird bei einer so großen Zahl an Variablen aufwändig. Eine exakte Berechnung benötigt $O(pn^2)$ Zeit (Drineas u. a. 2012). Es ist jedoch möglich, diese Zeit zu reduzieren, wenn man bei der Bestimmung der (C)LS eine sinnvolle Approximation der Werte zulässt. Wichtig ist, dass die approximative Berechnung einerseits weniger Zeit als die exakte benötigt, sich andererseits aber im Ergebnis nicht zu sehr von ihr unterscheidet. Bei Drineas

u. a. (2012) wird eine Methode vorgestellt, mit der eine Approximation der Matrix $\boldsymbol{Q}$ aus der QR-Zerlegung von $\tilde{\boldsymbol{X}}$ mit einer bestimmten Wahrscheinlichkeit nur geringfügig von der exakten Matrix $\boldsymbol{Q}$ abweicht. Diese Methode wird bei Clarkson und Woodruff (2013), sowie darauf aufbauend bei Nelson und Nguyen (2013) und Cohen (2016), verbessert. In einem ersten Schritt wird die Matrix $\boldsymbol{R}$ aus der QR-Zerlegung von $\tilde{\boldsymbol{X}}$ approximativ bestimmt. Dabei wird eine sogenannte Sparse Embedding Matrix verwendet. Die folgende Definition für eine solche Matrix basiert auf Cohen (2016) und Clarkson und Woodruff (2013). Ein Eintrag $\boldsymbol{\Pi}_{i,j}$ einer Sparse Embedding Matrix $\boldsymbol{\Pi} \sim r \times p$ mit Sparsity s wird zufällig gewählt. Dafür wird zunächst für jede Spalte j $\boldsymbol{f}(j)$ als s-dimensionaler Vektor gewählt, dessen Einträge ohne Zurücklegen und mit gleicher Wahrscheinlichkeit aus der Menge $\{1, ..., r\}$ gezogen werden. Für verschiedene Spalten ist dieser Vorgang unabhängig. Zusätzlich wird für jeden Eintrag ein zufälliges Vorzeichen $\sigma_{i,j}$ ausgewählt. Auch hier geschieht die Auswahl unabhängig und mit gleicher Wahrscheinlichkeit. Die Einträge $\boldsymbol{\Pi}_{f(j),j}$ werden dann als $\sigma_{f(j),j}\frac{1}{\sqrt{s}}$ definiert, alle übrigen Einträge werden auf 0 gesetzt. Nach Drineas u. a. (2012) reicht es hier, bei einer Sparsity von $s = 1$, $r = n^2$ zu wählen. Laut (Cohen 2016) kann bei einer Sparsity von $s = \log(p)$ r auf $n\log(n)$ gesetzt werden, um die Abweichung von den exakten Werten mit hoher Wahrscheinlichkeit klein zu halten. $\boldsymbol{R}^{-1}$ wird als Inverse, beziehungsweise Moore-Penrose Inverse, der so approximierten Matrix $\boldsymbol{R}$ bestimmt. In einem zweiten Approximationsschritt wird, wie bei Drineas u. a. (2012) beschrieben, eine ϵ-Johnson-Lindenstrauss Transformation (ϵ-JLT) verwendet, um die Berechnung von $\tilde{\boldsymbol{X}}\boldsymbol{R}^{-1}$ effizient anzunähern. Eine ϵ-JLT ist wie folgt definiert: Seien $\epsilon > 0$ und eine Menge $\boldsymbol{x}_1, ..., \boldsymbol{x}_p \in \mathbb{R}^n$ gegeben. Dann ist $\boldsymbol{\Pi} \in \mathbb{R}^{r\times n}$ eine ϵ-JLT, wenn gilt

$$(1-\epsilon)\|\boldsymbol{x}_i\|_2^2 \leq \|\boldsymbol{\Pi}\boldsymbol{x}_i\|_2^2 \leq (1+\epsilon)\|\boldsymbol{x}_i\|_2^2$$

(Drineas u. a. 2012). Für ein $\epsilon \in (0, 0.5]$ wird eine ϵ-JLT hier wie folgt konstruiert (Peters 2021). Zunächst wird $k = \frac{\log(p)}{\epsilon^2}$ gesetzt. Anschließend wird eine zufällige Matrix $\boldsymbol{P} \in \mathbb{R}^{n\times k}$ bestimmt, deren Einträge unabhängig voneinander aus einer Normalverteilung mit Erwartungswert 0 und Varianz $1/k$ gezogen werden. Mit dieser ϵ-JLT kann daraufhin das Matrixprodukt $\boldsymbol{\Omega} = \tilde{\boldsymbol{X}}\boldsymbol{R}^{-1}\boldsymbol{P}$ bestimmt wer-

den. Die approximativen Leverage Scores $\tilde{l}_i$ ergeben sich aus den quadrierten Zeilennormen $\|\boldsymbol{\Omega}_{i.}\|_2^2$, die approximativen Cross-Leverage Scores $\tilde{c}_{i(p+1)}$ für jeweils eine Einflussvariable und die Zielvariable werden mit dem Skalarprodukt $\langle\boldsymbol{\Omega}_{i.}\rangle\langle\boldsymbol{\Omega}_{(p+1).}\rangle$ ermittelt. Der Aufbau der für diese Approximationen verwendeten Funktionen basiert in Teilen auf dem bei Peters (2021) beschriebenen Vorgehen. In *R* (R Core Team 2021) wird jeweils eine Funktion zur Berechnung der CLS $c_{i(p+1)}$ für die Parameter $s = 1, r = n^2$, welche im Folgenden Algorithmus 1 genannt wird, und die Parameter $s = \log(p), r = n\log(n)$, im Folgenden Algorithmus 2, aufgestellt, wobei eingestellt werden kann, dass r um den Faktor c vergrößert wird. Die Standardeinstellung für ϵ ist 0.5, kann aber verändert werden. Außerdem kann eingestellt werden, dass die exakte Berechnung durchgeführt wird, wenn die entsprechende Sketchingmatrix eine höhere Dimension als die ursprüngliche Matrix hat. Sofern die Matrix $\boldsymbol{R}$ nicht invertierbar ist, wird die Moore-Penrose Inverse mit der Funktion `ginv()` aus dem *R*-Paket `MASS` (Venables und Ripley 2002) berechnet. Die Approximation der LS, sowie anderer CLS kann auf analoge Weise bestimmt werden, wird hier jedoch nicht weiter betrachtet. Es werden außerdem Funktionen aufgestellt, welche die Algorithmen so anwenden, dass die Daten zeilen- oder blockweise eingelesen werden, um auch mit sehr großen Daten umgehen zu können. Diese werden jedoch nicht in der Auswertung verwendet.

2.5 Kerndichteschätzer

Damit (C)LS als Selektionskriterien funktionieren, müssen sie sich abhängig davon, ob die entsprechende Einflussvariable relevant ist oder nicht, unterscheiden. Ein graphisches Mittel, um nach Verteilungsunterschieden zu suchen, ist die Abbildung eines Kerndichteschätzers. Ein Kern-Dichteschätzer für eine Dichte $f(x)$ ist gegeben durch

$$\hat{f}(x) = \frac{1}{p_1 h}\sum_{i=1}^{p_1} K(\frac{x - x_i}{h}),$$

wobei $x_1, ..., x_{p_1}$ die Realisationen, h eine gewählte Bandbreite und $K(u)$ ein Kern ist. Der hier verwendete Gauß-Kern ist definiert als

$$K(u) = \tfrac{1}{\sqrt{2\pi}} exp(-0.5u^2) \text{ für } u \in \mathbf{R} \text{ (Fahrmeir u. a. 2016, S.93).}$$

2.6 Bewertungskriterien für die Ergebnisse

Sowohl für die Variablenselektion, als auch für die Approximation der CLS werden Kriterien zur Bewertung beschrieben. Diese Kriterien sind zum Großteil sehr einfach und ihre Aussagekraft begrenzt sich auf die spezifischen Daten.

2.6.1 Güte der Variablenselektion

Um beurteilen, wie gut die (C)LS zur Variablenselektion in den betrachteten Simulationen funktionieren, werden verschiedene Kriterien angesehen. Dazu werden Simulationen mehrfach mit den gleichen Parametern durchgeführt. Dann wird gezählt, wie oft jeweils wie viele der relevanten Variablen selektiert werden. Das arithmetische Mittel der gefundenen relevanten Variablen wird als einfacher Indikator für die Performance der Selektion betrachtet. Führt die Interaktion zwischen verschiedenen Variablen zu einem Einfluss auf die Zielvariable, so wird in dieser Arbeit die Menge der entsprechenden Variablen als relevante Interaktion bezeichnet. Diese gilt als vollständig gefunden, wenn in einem Datensatz alle enthaltenen Variablen selektiert werden. Die Visualisierung davon erfolgt über Histogramme. Offensichtlich funktioniert eine Selektion besser, wenn mehr relevante Variablen gefunden werden. Für die logische Regression ist es wichtig, dass Interaktionen vollständig gefunden werden.

2.6.2 Güte der Approximation

Um zu beurteilen, wie gut die Approximation mit den in Abschnitt 2.4 beschriebenen Algorithmen funktioniert, werden zwei Ansätze verfolgt. Zum einen wird untersucht, wie groß die Distanz zwischen den approximierten und den exakten Werten ist. Dafür wird die absolute Differenz betrachtet, welche jeweils mit $\|\boldsymbol{Q}_{i.}\|_2\|\boldsymbol{Q}_{(p+1).}\|_2$ für den CLS der i-ten Einflussvariable mit der Zielvariable normiert wird (Drineas u. a. 2012). Zum anderen wird die Variablenselektion mit den approximierten CLS untersucht. Dazu wird bei den HapMap Daten der Anteil der Variablen, der sowohl mit den exakten, als auch mit den approximierten Daten selektiert wird, bestimmt. Für einige der simulierten Daten wird verglichen, wie viele vollständige Interaktionen jeweils bei Verwendung der approximierten

oder der exakten CLS gefunden werden. Als weiterer Vergleich wird die Zahl gefundener irrelevanter Interaktionen ähnlicher Form betrachtet.

3 Daten

In diesem Abschnitt werden die hier verwendeten Daten beschrieben. Neben echten SNP Daten aus dem internationalen HapMap Projekt werden dabei vor allem simulierte binäre Daten betrachtet. Neben der Tatsache, dass die Zahl der Variablen stets größer als die Zahl der Beobachtungen ist, haben die in dieser Arbeit betrachteten Datensätze außerdem gemein, dass die Zielvariable stets binäre Ausprägungen hat. Im Falle von genetischen Daten kann dies beispielsweise als Unterscheidung in Fall- und Kontrollgruppe oder, wie bei den HapMap Daten, zwischen zwei ethnischen Gruppen verstanden werden.

3.1 HapMap

Wie auch bei Parry u. a. (2021) werden die im *R*-Paket `SNPassoc` (Moreno und Gonzalez 2021) verfügbaren Daten verwendet, welche aus den Erhebungen des internationalen HapMap Projektes (Hapmap 2003) stammen. Diese Daten enthalten 9307 SNPs von 120 Menschen. Die betrachtete Zielvariable ist hier die ethnische Gruppe. Ein Teil der Individuen stammt aus Europa (CEU), bei ihnen nimmt die Zielvariable den Wert 1 an. Die anderen gehören dem westafrikanischen Yorùbá Volk (YRI) an. Bei ihnen hat die Zielvariable den Wert 0. Unter Verwendung der Funktion `additive` (Moreno und Gonzalez 2021) werden die SNP Variablen so kodiert, dass sie den Wert 0 annehmen, wenn beide entsprechenden Allele die häufige Version an diesem Locus sind, 1, wenn das Merkmal heterozygot ist, und 2, wenn nur das seltene Merkmal vorliegt (*PLINK: Whole genome data analysis toolset - Harvard University* 2017). Außerdem werden die fehlenden Werte mittels Imputation ergänzt. Dazu wird das *R*-Paket `missMethods` verwendet (Rockel 2020). Anschließend werden diejenigen Variablen, die bei jeder Beobachtung den Wert 0 annehmen, entfernt, da sie keine zusätzlichen Informationen liefern. Die Zahl der übrigen Variablen ist ohne die Zielvariable 7648. Dieses Vorgehen entspricht im Wesentlichen dem bei Parry

u. a. (2021).

3.2 Simulation

Es wird im Folgenden davon ausgegangen, dass die betrachteten Simulationen sinnvoll für die Analyse sind. Allerdings liegen hier, anders als bei tatsächlichen SNP Daten, binäre Daten vor, welche nicht aus einer Dummy-Kodierung hervorgehen. Außerdem ist die Wahrscheinlichkeit dafür, dass eine Variable den Wert 1 annimmt bei relevanten Einflussvariablen anders als bei irrelevanten. Des weiteren werden die Einflussvariablen hier unabhängig voneinander simuliert, während davon ausgegangen wird, dass zwischen verschiedenen Allelen durchaus Abhängigkeiten bestehen können. Um zu untersuchen, inwiefern die (C)LS bei der Variablenselektion helfen, werden binäre Datensätze simuliert. Die Funktion `simulatebinarydata`, die dazu verwendet wird, entspricht der Simulationsfuntkion in Parry u. a. (2021). Zunächst lassen sich die Zahl der Beobachtungen n und die Zahl der Einflussvariablen p angeben. Mit dem Argument $vars$ können die Variableninteraktionen, die einen tatsächlichen Einfluss auf die Zielvariable haben, festgelegt werden. Um die Anzahl der Beobachtungen, in der die Zielvariable den Wert 0 oder 1 annimmt, etwa gleich zu halten, kann die Wahrscheinlichkeit, dass bestimmte Variableninteraktionen vorkommen, eingestellt werden. Des weiteren gibt es eine Möglichkeit, die Wahrscheinlichkeit, dass die Zielvariable 1 wird, obwohl kein der relevanten Interaktionen vorliegen, sowie die Wahrscheinlichkeit, dass die Zielvariable 0 wird, obwohl eine relevante Interaktion vorliegt, zu verändern. In den betrachteten Simulationen wird die Zielvariable jedoch genau dann 1, wenn mindestens eine der relevanten Interaktionen vorliegt. Da die Anordnung der relevanten und irrelevanten Variablen keinen Einfluss auf die Selektion haben (Parry u. a. 2021), sind die relevanten Variablen stets die ersten Einflussvariablen im Datensatz. Betrachtet man also beispielsweise die folgende Simulation

$$\texttt{simulatebinarydata}(n = 10, p = 100, vars = list(c(1, 2), -3),$$
$$prob = \texttt{profun}(vars = \texttt{list}(c(1, 2), -3), p = 100), yprobpos = 1, yprobneg = 0),$$

so simuliert man 10 Beobachtungen mit 100 Einflussvariablen. Die Zielvariable wird genau dann 1, wenn die ersten beiden Einflussvariablen den Wert 1 annehmen oder die dritte den Wert 0, anders ausgedrückt wenn $L = (X_1 \wedge X_2) \vee X_3^c = 1$. Die Funktion `profun`, die ebenfalls von Parry u. a. (2021) stammt, sorgt mit den entsprechenden Parametern dafür, dass die Realisationen der Zielvariable in etwa gleich oft 0 und 1 sind. Im Folgenden werden die Elemente der Liste *vars* gleichbedeutend mit den entsprechenden booleschen Kombinationen verwendet. Als nächstes werden die für die Analyse verwendeten Parameter beschrieben. Bei Parry u. a. (2021) werden drei verschiedene Szenarien betrachtet, welche sich darin unterscheiden, welche Variablenkombinationen die Ausprägung der Zielvariablen beeinflussen. Auch hier werden diese Szenarien mit einer Variablenzahl von 1000 und 60 Beobachtungen simuliert. Zusätzlich werden Simulationen mit anders gewählten Parametern betrachtet. Zur Abgrenzung werden diese mit Simulation 1 bis 20 bezeichnet. Bei den Simulationen 1 bis 11 werden Daten mit verschiedenen Parametern simuliert, um einen Eindruck davon zu erhalten, wie sich die (C)LS in unterschiedlichen Situationen verhalten, und wann mit ihrer Hilfe die relevanten Einflussvariablen gut gefunden werden können. In Tabelle 1 Simulation 12 soll zusätzlich zu den HapMap Daten weitere Datensätze liefern, bei denen die Differenz zwischen exakten und approximierten Cross-Leverage Scores untersucht werden kann. Dazu wird 100 mal ein Datensatz mit $n = 60$ und $p = 100000$ simuliert. Die relevante logische Einflussvariable ist von der gleichen Form wie bei Simulation 11, also $L = (X_1 \wedge X_2 \wedge X_3^c) \vee (X_4 \wedge X_5) \vee (X_6 \wedge X_7) \vee (X_8 \wedge X_9^c) \vee (X_{10}^c \wedge X_{11}^c) \vee X_{12} \vee X_{13}$. Bei den Simulationen 13 bis 20 liegt der Fokus darauf, wie sich die (C)LS relevanter Variablen von denen der irrelevanten Variablen unterscheiden. Dafür werden Daten mit weniger Variablen und Beobachtungen simuliert. Für diese Simulationen gilt $n = 10$ und $p = 100$. Verschiedene boolesche Kombinationen der ersten 6 Variablen entscheiden jeweils über die Ausprägung der Zielvariablen.

Tabelle 1: Simulationen 1 bis 11

Simulation	n	p	relevante logische Einflussvariable
1	40	10000	$X_1 \vee X_2 \vee ... \vee X_{100}$
2	40	10000	$(X_1 \wedge X_2 \wedge X_3 \wedge X_4 \wedge X_5) \vee (X_6 \wedge X_7 \wedge X_8 \wedge X_9) \vee (X_{10} \wedge X_{11} \wedge X_{12} \wedge X_{13}) \vee (X_{14} \wedge X_{15} \wedge X_{16}) \vee (X_{17} \wedge X_{18} \wedge X_{19}) \vee (X_{20} \wedge X_{21} \wedge X_{22}) \vee (X_{23} \wedge X_{24}) \vee (X_{25} \wedge X_{26}) \vee (X_{27} \wedge X_{28}) \vee (X_{29} \wedge X_{30}) \vee X_{31} \vee X_{32} \vee X_{33} \vee X_{34} \vee X_{35}$
3	50	5000	$(X_1 \wedge X_2 \wedge X_3 \wedge X_4 \wedge X_5) \vee (X_6 \wedge X_7 \wedge X_8 \wedge X_9) \vee (X_{10} \wedge X_{11} \wedge X_{12} \wedge X_{13}) \vee (X_{14} \wedge X_{15} \wedge X_{16}) \vee (X_{17} \wedge X_{18} \wedge X_{19}) \vee (X_{20} \wedge X_{21} \wedge X_{22}) \vee (X_{23} \wedge X_{24}) \vee (X_{25} \wedge X_{26}) \vee (X_{27} \wedge X_{28}) \vee (X_{29} \wedge X_{30}) \vee X_{31} \vee X_{32} \vee X_{33} \vee X_{34} \vee X_{35}$
4	60	10000	$(X_1 \wedge X_2 \wedge X_3 \wedge X_4^c \wedge X_5^c) \vee (X_6 \wedge X_7 \wedge X_8 \wedge X_9) \vee (X_{10} \wedge X_{11} \wedge X_{12}^c \wedge X_{13}^c) \vee (X_{14} \wedge X_{15} \wedge X_{16}) \vee (X_{17} \wedge X_{18} \wedge X_{19}^c) \vee (X_{20}^c \wedge X_{21}^c \wedge X_{22}^c) \vee (X_{23} \wedge X_{24}) \vee (X_{25} \wedge X_{26}^c) \vee (X_{27}^c \wedge X_{28}^c) \vee (X_{29}^c \wedge X_{30}^c) \vee X_{31} \vee X_{32} \vee X_{33} \vee X_{34}^c \vee X_{35}^c$
5	40	1000	$X_1 \vee X_2 \vee X_3 \vee X_4 \vee X_5 \vee X_6^c \vee X_7^c \vee X_8^c \vee X_9^c \vee X_{10}^c$
6	50	5000	$(X_1 \wedge X_2 \wedge X_3^c) \vee (X_4 \wedge X_5) \vee (X_6 \wedge X_7) \vee (X_8 \wedge X_9^c) \vee (X_{10}^c \wedge X_{11}^c) \vee X_{12} \vee X_{13}^c$
7	50	5000	$(X_1 \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5 \wedge X_6) \vee (X_7 \wedge X_8) \vee X_9 \vee X_{10} \vee (X_{11}^c \wedge X_{12}^c \wedge X_{13}^c) \vee (X_{14}^c \wedge X_{15}^c \wedge X_{16}^c) \vee (X_{17}^c \wedge X_{18}^c) \vee X_{19}^c \vee X_{20}^c$
8	60	1000	$(X_1 \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5 \wedge X_6) \vee (X_7 \wedge X_8) \vee X_9 \vee X_{10} \vee (X_{11}^c \wedge X_{12}^c \wedge X_{13}^c) \vee (X_{14}^c \wedge X_{15}^c \wedge X_{16}^c) \vee (X_{17}^c \wedge X_{18}^c) \vee X_{19}^c \vee X_{20}^c$
9	100	10000	$(X_1 \wedge X_2 \wedge X_3^c) \vee (X_4 \wedge X_5) \vee (X_6 \wedge X_7) \vee (X_8 \wedge X_9^c) \vee (X_{10}^c \wedge X_{11}^c) \vee X_{12} \vee X_{13}^c$
10	100	10000	$(X_1 \wedge X_2^c \wedge X_3^c) \vee (X_4 \wedge X_5) \vee (X_6 \wedge X_7^c) \vee (X_8^c \wedge X_9^c) \vee (X_{10}^c \wedge X_{11}^c) \vee X_{12} \vee X_{13}^c$
11	100	10000	$(X_1 \wedge X_2 \wedge X_3^c) \vee (X_4 \wedge X_5) \vee (X_6 \wedge X_7) \vee (X_8 \wedge X_9^c) \vee (X_{10}^c \wedge X_{11}^c) \vee X_{12} \vee X_{13}$

Tabelle 2: Simulationen 13 bis 20

Simulation	n	p	relevante logische Einflussvariable
13	10	100	$(X_1 \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5) \vee X_6$
14	10	100	$(X_1^c \wedge X_2^c \wedge X_3^c) \vee (X_4^c \wedge X_5^c) \vee X_6^c$
15	10	100	$(X_1^c \wedge X_2 \wedge X_3) \vee (X_4^c \wedge X_5) \vee X_6$
16	10	100	$(X_1^c \wedge X_2^c \wedge X_3) \vee (X_4^c \wedge X_5) \vee X_6^c$
17	10	100	$X_1 \vee X_2 \vee X_3 \vee X_4 \vee X_5 \vee X_6$
18	10	100	$X_1^c \vee X_2^c \vee X_3^c \vee X_4^c \vee X_5^c \vee X_6^c$
19	10	100	$(X_1^c \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5 \wedge X_6)$
20	10	100	$(X_1 \wedge X_2) \vee (X_3 \wedge X_4^c) \vee (X_5^c \wedge X_6^c)$

4 Auswertung

Für die Analyse der Daten mit den oben beschriebenen Methoden wird *R* benutzt (R Core Team 2021). Die Grafiken werden mit `ggplot2` erstellt (Wickham 2016), wobei die Farbgebung zum Teil mit `viridis` (Garnier u. a. 2021) gewählt wird. Für Grafiken, die aus mehreren Abbildungen bestehen, wird außerdem das Paket `gridExtra` verwendet (Auguie 2017). Zunächst werden die (C)LS bei verschiedenen Datensätzen betrachtet. Es wird untersucht, wie gut sie als Selektionskriterium geeignet sind. Für den zweiten Teil der Auswertung wird angenommen, dass sie bei der Variablenauswahl sinnvoll sind. Hier werden die vorgestellten Algorithmen zur Approximation der (C)LS angewandt. Die approximativ bestimmten Werte werden dann mit den exakten Leverage und Cross-Leverage Scores verglichen.

4.1 Auswahl der relevanten Einflussvariablen

Da die Variablenselektion der HapMap Daten vermutlich der Selektion bei (Parry u. a. 2021) entspricht, wird sie hier nicht erneut durchgeführt. Für die drei Szenarien aus Parry u. a. (2021) werden bei 1000 Simulationen $60 \log(60) = 256$ Variablen mit CLS, LS und Korrelation ausgewählt. Das arithmetische Mittel der jeweils gefundenen Variablen ist in Tabelle 3 zu finden. Wie bei auch bei Parry u. a. (2021), werden bei Szenario 1 die meisten Variablen mit den Leverage Scores gefunden, während bei den Szenarien, in denen auch einflussreiche Variableninteraktionen vorliegen, die Selektion mit den CLS am besten abschneidet. Es wird nun untersucht, ob bei Datensätzen mit anderen Dimensionen und relevanten Einflussvariablen ähnliche Erkenntnisse gezogen werden können. Mit den in Tabelle 1 aufgelisteten Parametern werden jeweils 1000 Datensätze simuliert.

Tabelle 3: Mittlere Anzahl der mit verschiedenen Kriterien gefundenen relevanten Variablen in den Szenarien 1 bis 3

Szenario	CLS	LS	Korrelation
1	3.983	10	9.287
2	6.257	4.94	6.24
3	6.48	5.339	6.281

Jedes Mal werden die $n \log(n)$ Variablen mit den betragsmäßig höchsten CLS oder den niedrigsten LS, beziehungsweise den betragsmäßig höchsten Korrelationen, ausgewählt. Wie in Abschnitt 2.6 beschrieben, wird untersucht, wie oft relevante Variablen, beziehungsweise Interaktionen gefunden werden.

Tabelle 4: Mittlere Anzahl der mit verschiedenen Kriterien gefundenen relevanten Variablen in den Simulationen 1 bis 11

Simulation	CLS	LS	Korrelation
1	0	100	0.007
2	0.559	15.582	0.75
3	1.532	18.394	2.513
4	0.416	18.225	0.055
5	0	10	0
6	0.347	8.489	2.027
7	0.033	10.687	0.242
8	1.117	13.762	0.448
9	0.903	10.166	2.221
10	0.469	10.164	2.724
11	3.244	10.222	4.088

Wie oft die Variablen gefunden werden, unterscheidet sich abhängig von den jeweiligen Parametern stark. Auffällig ist jedoch, dass bei der Selektion mit den Leverage Scores in jeder der Simulationen 1 bis 11 die relevanten Variablen deutlich häufiger gefunden werden, als bei der Verwendung von CLS oder Korrelationen. Wie in Tabelle 4 zu erkennen ist, werden beispielsweise in Simulation 1 mit den LS jedes Mal alle relevanten Variablen gefunden, während mit den CLS keine und mit den Korrelationen fast keine gefunden werden. In diesem Fall entspricht das Ergebnis den Aussagen in (Parry u. a. 2021), da hier ausschließlich Haupteffekte betrachtet werden. Allerdings werden auch bei den Datensätzen, in denen hauptsächlich Interaktionen, die aus mehr als einer Variablen bestehen, relevant für die Ausprägungen der Zielvariablen sind, mehr relevante Variablen mit den LS als mit den CLS gefunden. Dies ist beispielsweise bei den Simulationen 9 bis 11 zu sehen. Weiterhin fällt auf, dass eine Variable, die die Zielvariable negativ beeinflusst gelegentlich deutlich häufiger gefunden wird als eine Variable mit positivem Einfluss, auch wenn beide aus Interaktionen gleicher Größe stammen. In Abbildung 1 ist beispielsweise zu erkennen, dass

sogar beide Variablen 10 und 11 der Interaktion $X_{10}^c \wedge X_{11}^c$ gemeinsam sehr viel häufiger gefunden werden als die einzelnen Variablen 12 und 13. Dafür dass in

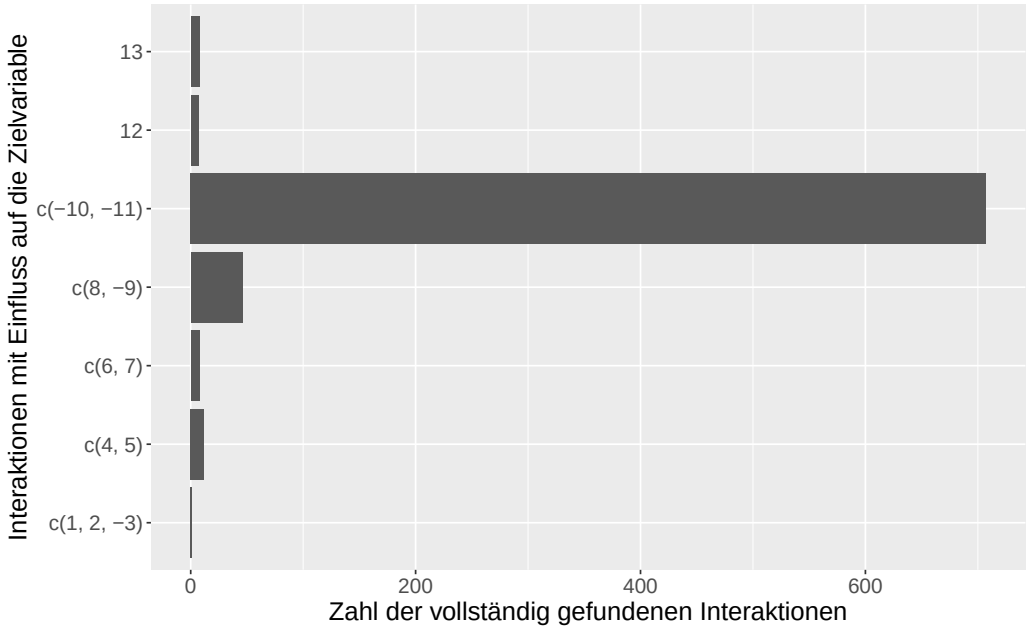

Abbildung 1: Zahl der mit den höchsten Beträgen der CLS gefundenen Interaktionen in Simulation 11

einigen Fällen über die betragsmäßig größten CLS kaum relevante Variablen gefunden werden, gibt es neben dem Zufall zwei denkbare Erklärungen. Entweder unterscheiden sich die CLS der relevanten Variablen bei diesen Daten nicht wesentlich von den CLS der irrelevanten Variablen oder sie unterscheiden sich zwar, liegen aber im Mittel nicht weiter von 0 entfernt. Um einen Eindruck davon zu erhalten, wie die CLS der relevanten Variablen im Vergleich zu den CLS der relevanten Variablen verteilt sind, können entsprechende Kerndichteschätzer betrachtet werden. In Abbildung 2 sind die Kerndichteschätzer für die Beträge der CLS in Simulation 2 dargestellt. Variablen, die zu Interaktionen gleicher Größe gehören, werden gemeinsam betrachtet. Zumindest die absoluten CLS der Variablen 31 bis 35 unterscheiden sich in ihrer Verteilung von denen der irrelevanten Variablen 36 bis 1000. Bei den anderen relevanten Variablen ist der Unterschied zu den irrelevanten weniger deutlich. Bei den Variablen 31 bis 35 gibt es weniger betragsmäßig große Variablen als bei den Variablen 36 bis 1000.

Eine Selektion über betragsmäßig große CLS ergibt hier also wenig Sinn.

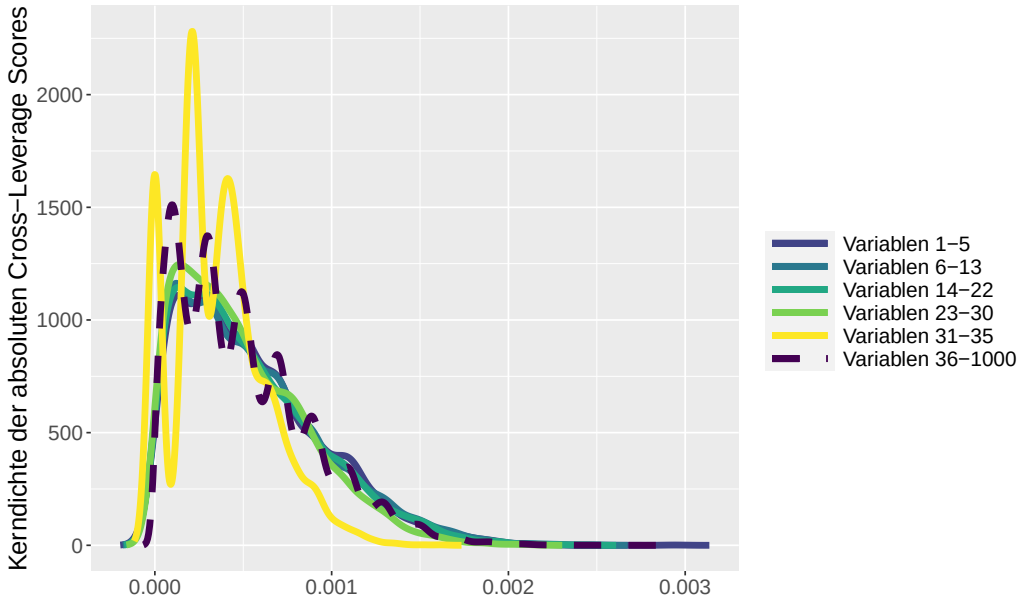

Abbildung 2: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 2. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1 \wedge X_2 \wedge X_3 \wedge X_4 \wedge X_5) \vee (X_6 \wedge X_7 \wedge X_8 \wedge X_9) \vee (X_{10} \wedge X_{11} \wedge X_{12} \wedge X_{13}) \vee (X_{14} \wedge X_{15} \wedge X_{16}) \vee (X_{17} \wedge X_{18} \wedge X_{19}) \vee (X_{20} \wedge X_{21} \wedge X_{22}) \vee (X_{23} \wedge X_{24}) \vee (X_{25} \wedge X_{26}) \vee (X_{27} \wedge X_{28}) \vee (X_{29} \wedge X_{30}) \vee X_{31} \vee X_{32} \vee X_{33} \vee X_{34} \vee X_{35}$

Um einen besseren Überblick über die CLS und die LS relevanter Variablen verschiedener Form zu erhalten, werden bei den Simulationen 13 bis 20 n, p, sowie die Zahl der relevanten Einflussvariablen nicht verändert. Aus Zeitgründen sind n und p mit 10 und 100 vergleichsweise klein. Verschiedene Interaktionen der ersten sechs Variablen beeinflussen die Zielvariable. Diese werden in Abschnitt 3 beschrieben. Die Kerndichteschätzer der CLS, der absoluten CLS, sowie der LS werden bestimmt. Dabei werden die Variablen gemeinsam betrachtet, die in gleich großen Interaktionen auftreten und in der Schreibweise von `simulatebinarydata` das gleiche Vorzeichen haben. Exemplarisch werden hier die absoluten Cross-Leverage Scores und die Leverage Scores bei Simulation 13 abgebildet. Die übrigen Kerndichteschätzer befinden sich im Anhang in Abschnitt A. In Abbildung 3 ist zu erkennen, dass die absoluten CLS sich hier bei

allen relevanten Variablen von denen der irrelevanten unterscheiden, während bei den LS in Abbildung 4 allein die Variable 6, die die Zielvariable als Haupteffekt beeinflusst zu einem deutlich anderen Kerndichteschätzer führt. Betrachtet man alle Kerndichteschätzer im Anhang, so sieht man, dass relevante Variablen sich zum Teil in den CLS, aber nicht in den LS voneinander unterscheiden. In diesen Fällen ist davon auszugehen, dass die CLS als Selektionskriterium besser funktionieren. Allerdings scheinen die betragsmäßig höchsten CLS nicht immer die der relevanten zu sein. In einigen Fällen scheinen es eher die betragsmäßig kleinsten zu sein.

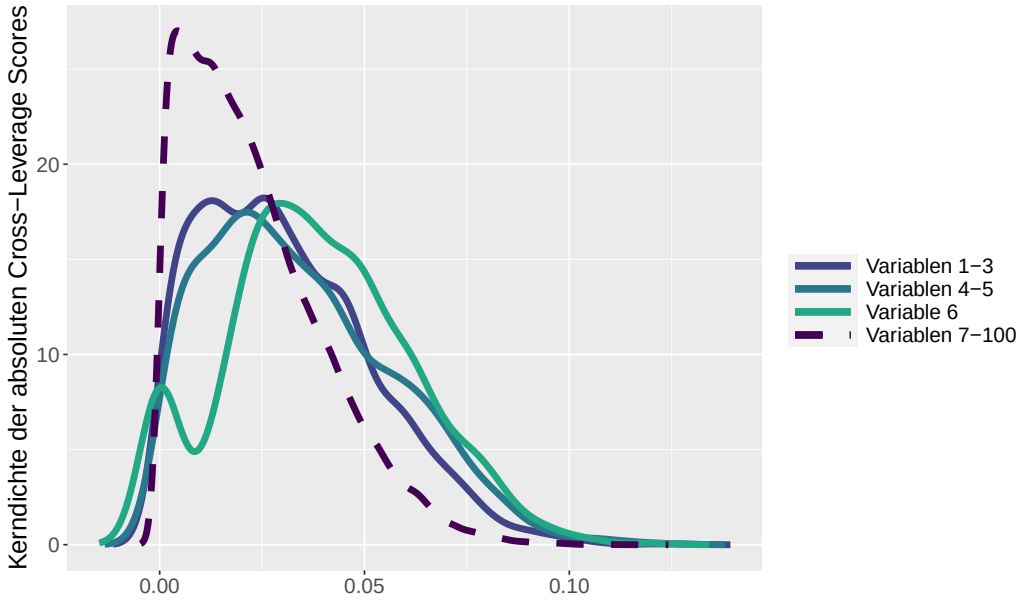

Abbildung 3: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 13. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1 \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5) \vee X_6$

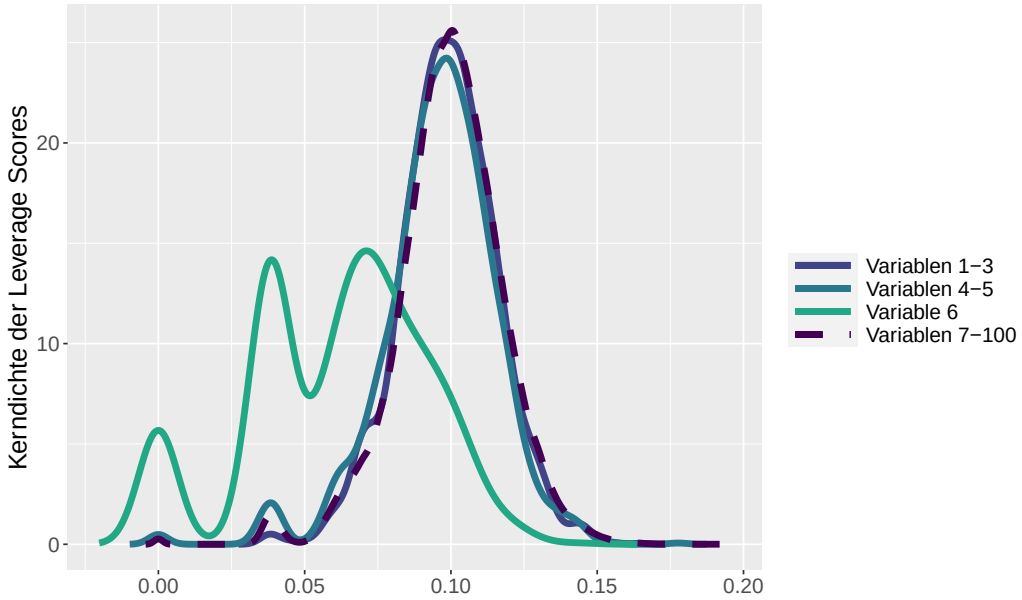

Abbildung 4: Kerndichteschätzer der LS der Variablen für 1000 Wiederholungen von Simulation 13. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1 \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5) \vee X_6$

4.2 Approximation der CLS und LS

Ist die Zahl der erhobenen Variablen zu groß, um Cross-Leverage und Leverage Scores exakt zu berechnen, so kann eine approximative Berechnung in Betracht gezogen werden, wie sie in Abschnitt 2.4 vorgestellt wird. In diesem Abschnitt wird dies am Beispiel der HapMap Daten, sowie einiger simulierter Daten, durchgeführt. Hier werden stets die CLS, die jeweils zu einer Einflussvariablen und der Zielvariablen gehören, betrachtet. Zunächst werden diese CLS im HapMap Datensatz approximiert. Dabei werden die Funktionen, die hier Algorithmus 1 und Algorithmus 2 heißen, angewendet. Eine Veränderung der Parameter c und ϵ wird hier beispielhaft für die Werte $c = 2$ und $\epsilon = 0.3$ untersucht. Zunächst werden die CLS auf die verschiedenen Arten jeweils 100 mal approximiert. Diese Werte werden dann mit den exakt berechneten verglichen, indem die absoluten Differenzen mit der in Abschnitt 2.6.2 beschriebenen Normierung bestimmt werden. Dann wird untersucht, wie hoch der Anteil der mit den approximierten Werten ausgewählten Variablen an den Variablen ist, die mit den exakten Cross-Leverage Scores selektiert werden. Sowohl die Differenzen, als auch dieser Anteil sind in Tabelle 5 festgehalten. Da eine hohe Übereinstimmung mit der Selektion durch die exakten CLS und eine niedrige absolute Differenz für eine gute Approximation sprechen, ist es bemerkenswert, dass der Anteil gleicher selektierter Variablen hier bei Algorithmus 1 stets etwas größer ist als bei Algorithmus 2 mit gleichen Einstellungen, während die absoluten Differenzen bei Algorithmus 2 tendenziell kleiner sind. Auch für Simulation 12 werden die

Tabelle 5: Mittlerer Anteil der über die Approximationen selektierten Variablen im HapMap Datensatz, die auch mit den exakten CLS selektiert werden. In den Klammern sind die entsprechenden normierten mittleren Abstände zu den exakt berechneten CLS zu finden.

	Standardeinstellung	$c = 2$	$\epsilon = 0.3$
Algorithmus 1	23.75% (0.1337)	36.79% (0.13079)	52.31% (0.08075)
Algorithmus 2	23.22% (0.06468)	36.07% (0.06768)	49.21% (0.06272)

Approximationen betrachtet. Hier ist $n = 60$ und $p = 100000$, der Unterschied zwischen Variablen- und Beobachtungsanzahl ist also deutlich höher als bei den HapMap Daten. Wenn man für einen Durchlauf der Simulation die CLS 100

mal mit den Algorithmen 1 und 2 approximiert, so ist die mittlere normierte absolute Differenz zu den exakten Werten, wie auch bei den HapMap Daten, bei Algorithmus 1 tendenziell höher. Sie liegt dort im Mittel bei etwa 0.12, bei Algorithmus 2 nur bei etwa 0.09. Auch die höchsten mittleren absoluten Differenzen bei einer Approximation sind bei Algorithmus 1 zu finden. Exemplarisch wird auch hier untersucht, wie eine Veränderung der in Abschnitt 2.4 beschriebenen Parameter c und ϵ sich auf die mittleren absoluten Differenzen zu den exakten Werten auswirkt. Eine Verdopplung von c führt bei beiden Algorithmen nur zu einer geringfügigen Verkleinerung des Mittelwerts der Differenzen. Anders als bei Algorithmus 2, wird bei Algorithmus 1 dieser Wert deutlich kleiner, wenn ϵ nicht auf 0.5, sondern auf 0.3 gesetzt wird. Dann ist der mittlere absolute Mittelwert der Differenzen zu den exakten CLS mit etwa 0.07 unter allen betrachteten Approximationen am kleinsten. Inwiefern die Variablenauswahl über approximierte CLS zum finden relevanter Interaktionen führt, wird für die Szenarien 1 bis 3 aus Parry u. a. (2021) analysiert. Die Abbildungen 5 bis 7 zeigen die Zahl der bei 1000 Simulationen vollständig gefundenen relevanten Interaktionen bei Verwendung der exakten CLS, sowie der mit Algorithmus 1 und 2 approximierten CLS. In allen Szenarien ist die Zahl der mit Algorithmus 1 und der mit Algorithmus 2 gefundenen Interaktionen ähnlich groß und deutlich niedriger als bei der Selektion mit exakten CLS. Wie in Abbildung 5 zu erkennen ist, werden in Szenario 1 die irrelevanten Einflussvariablen 11 bis 20 sogar häufiger selektiert als die relevanten. Bei Szenario 2 und 3 werden hingegen auch mit den approximativen Werten mehr relevante Interaktionen gefunden. Wie zu erwarten ist, werden Interaktionen mit weniger Variablen häufiger vollständig gefunden.

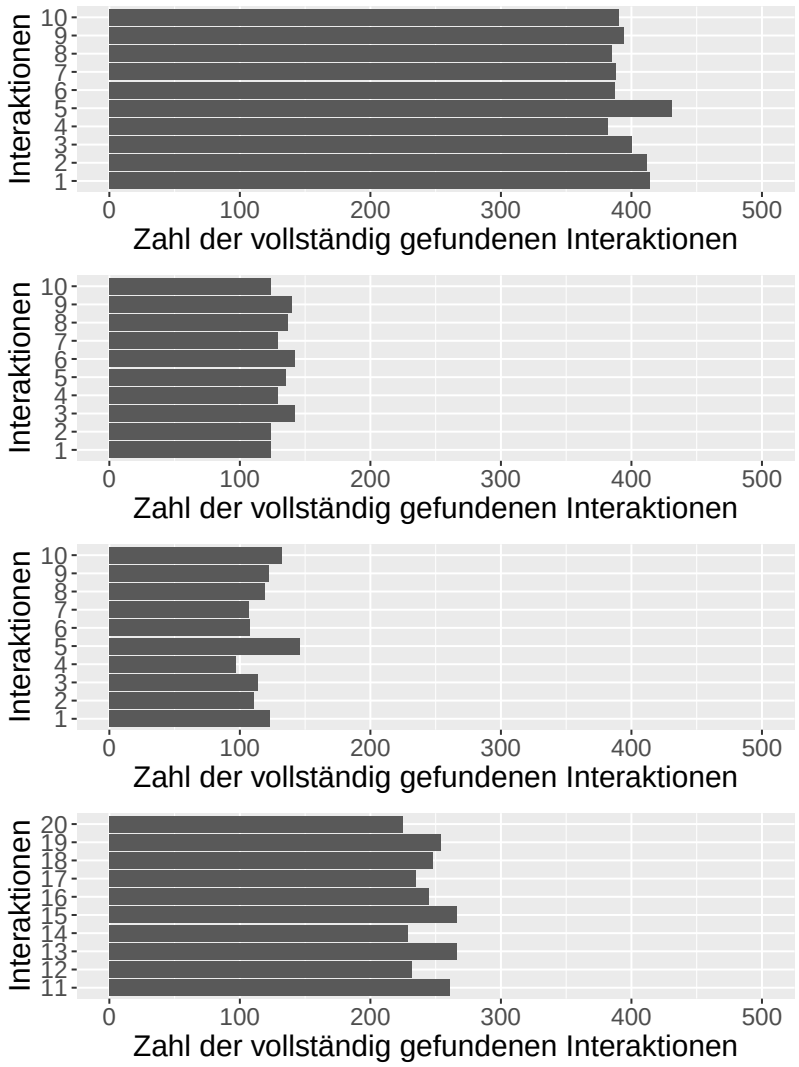

Abbildung 5: Selektion der Variablen mit (von oben nach unten) exakten CLS, mit Algorithmus 1 und mit Algorithmus 2 approximierten CLS bei Szenario 1. Ganz unten ist die Zahl vergleichbarer irrelevanter Interaktionen, die bei Verwendung der exakten CLS gefunden werden, abgebildet.

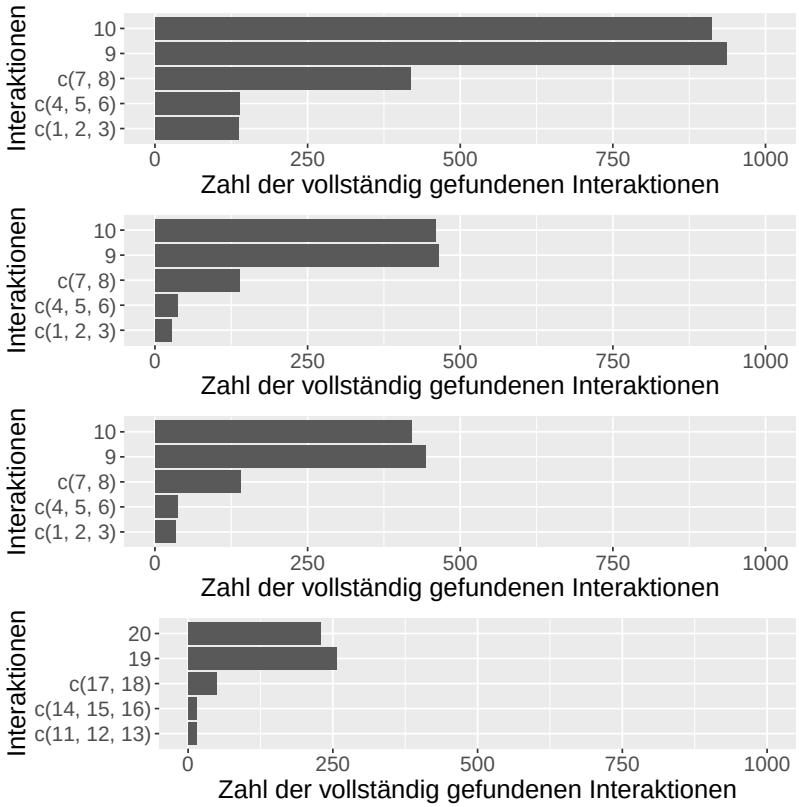

Abbildung 6: Selektion der Variablen mit (von oben nach unten) exakten CLS, mit Algorithmus 1 und mit Algorithmus 2 approximierten CLS bei Szenario 2. Ganz unten ist die Zahl vergleichbarer irrelevanter Interaktionen, die bei Verwendung der exakten CLS gefunden werden, abgebildet.

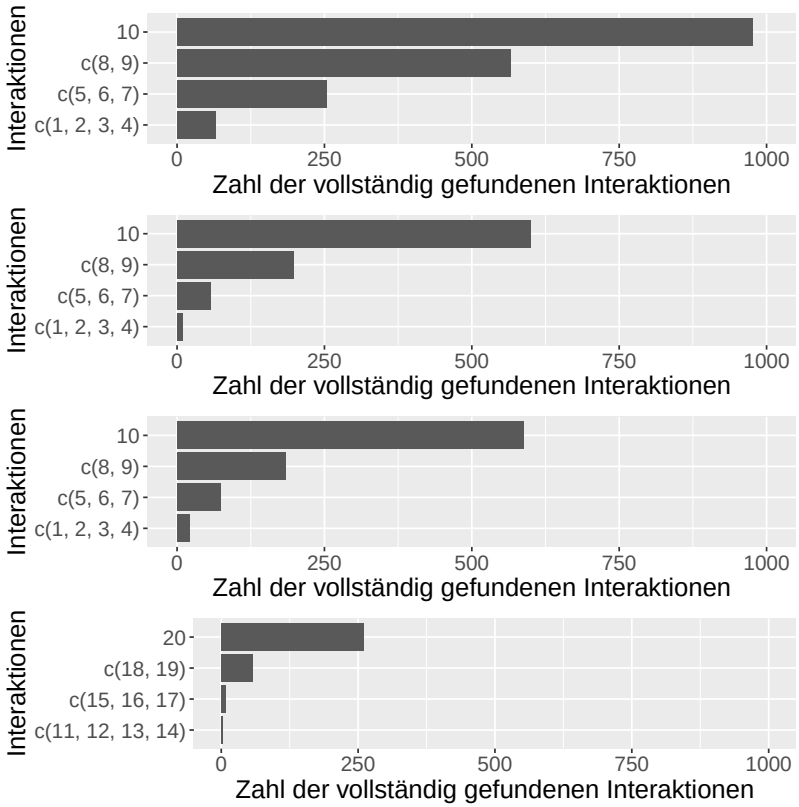

Abbildung 7: Selektion der Variablen mit (von oben nach unten) exakten CLS, mit Algorithmus 1 und mit Algorithmus 2 approximierten CLS bei Szenario 3. Ganz unten ist die Zahl vergleichbarer irrelevanter Interaktionen, die bei Verwendung der exakten CLS gefunden werden, abgebildet.

5 Zusammenfassung

Bei Datensätzen mit sehr vielen Variablen, wie sie beispielsweise in der Genetik häufig vorkommen, kann es sinnvoll sein, zunächst eine Auswahl potentiell wichtiger Variablen zu treffen, um die effizient analysieren zu können. In dieser Arbeit wurde untersucht, inwieweit Leverage Scores und insbesondere Cross-Leverage Scores für eine solche Vorauswahl geeignet sein können. Außerdem wurden Möglichkeiten, Cross-Leverage Scores mit weniger Aufwand approximativ zu berechnen, betrachtet. Die Datensätze, die dazu verwendet wurden, enthalten dabei alle mehr Variablen als Beobachtungen. Außerdem liegt jeweils eine Zielvariable mit binär kodierten Ausprägungen vor. Neben sogenannten SNP Daten aus dem internationalen HapMap Projekt (Hapmap 2003), werden simulierte Daten, in denen auch die Einflussvariablen binär sind, analysiert. Die Simulationen sind von Parry u. a. (2021) übernommen. Nicht nur einzelne Variablen, sondern auch Variablenkombinationen können dabei als einflussreich festgelegt werden. Auch das Vorgehen bei der Variablenselektion basiert darauf. Für die Approximation der Cross-Leverage Scores wird ein Algorithmus verwendet, der auf Drineas u. a. (2012), Clarkson und Woodruff (2013), Nelson und Nguyen (2013) und Cohen (2016) beruht. Die Ergebnisse dieser Arbeit beziehen sich allein auf die verwendeten Daten. Es können keine allgemeingültigen Schlüsse gezogen werden. Haupterkenntnis im Bezug auf die Variablenselektion mit Leverage Scores und Cross-Leverage Scores ist, dass die Selektion anhand betragsmäßig großer Cross-Leverage Scores, die bei Parry u. a. (2021) gut funktioniert, hier in vielen Fällen nicht sinnvoll ist. Da sowohl die Cross-Leverage, als auch Leverage Scores der relevanten und irrelevanten Variablen sich bei vielen der betrachteten Datensätze jedoch unterscheiden, kann durchaus davon ausgegangen werden, dass diese Werte als Kriterium zur Variablenselektion nützlich sein können. Bei Parry u. a. (2021) wird eine Variablenauswahl vorgeschlagen, die zu einem Teil über Cross-Leverage und zum anderen Teil über Leverage Scores geschieht. Dies ist für weitergehende Untersuchungen sicherlich interessant. Bei den Approximationen der Cross-Leverage Scores scheint die Variablenselektion zum Teil noch zu funktionieren, allerdings in den betrachteten Beispielen deutlich schlechter als bei Verwendung der exakten Werte. Es ist jedoch möglich,

dass auch bei den betrachteten Daten Approximationen mit anderen als den gewählten Parametern zu besseren Ergebnissen führen.

Literatur

Auguie, Baptiste (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. URL: https://CRAN.R-project.org/package=gridExtra.

Clarkson, Kenneth L. und David P. Woodruff (2013). „Low rank approximation and regression in input sparsity time". In: *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*. Hrsg. von Dan Boneh, Tim Roughgarden und Joan Feigenbaum. ACM, S. 81–90. DOI: 10.1145/2488608.2488620. URL: https://doi.org/10.1145/2488608.2488620.

Cohen, Michael B. (2016). „Nearly Tight Oblivious Subspace Embeddings by Trace Inequalities". In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*. Hrsg. von Robert Krauthgamer. SIAM, S. 278–287. DOI: 10.1137/1.9781611974331.ch21. URL: https://doi.org/10.1137/1.9781611974331.ch21.

Consortium, 1000 Genomes Project u. a. (2015). „A global reference for human genetic variation". In: *Nature* 526.7571, S. 68.

Drineas, Petros u. a. (2012). „Fast approximation of matrix coherence and statistical leverage". In: *The Journal of Machine Learning Research* 13.1, S. 3475–3506.

Fahrmeir, Ludwig u. a. (2016). *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag.

Garnier u. a. (2021). *viridis - Colorblind-Friendly Color Maps for R*. R package version 0.6.2. DOI: 10.5281/zenodo.4679424. URL: https://sjmgarnier.github.io/viridis/.

Graw, Jochen (2015). *Genetik*. Springer.

Hapmap, CA (2003). „The international HapMap project: The international HapMap consortium". In: *Nature* 426, S. 789–96.

Hoaglin, David C und Roy E Welsch (1978). „The hat matrix in regression and ANOVA". In: *The American Statistician* 32.1, S. 17–22.

Karpfinger, Christian (2014). „Die QR-Zerlegung einer Matrix“. In: *Arbeitsbuch Höhere Mathematik in Rezepten*. Springer, S. 87–88.

Kruglyak, Leonid und Deborah A Nickerson (2001). „Variation is the spice of life“. In: *Nature genetics* 27.3, S. 234–236.

Moreno, Victor und Juan R Gonzalez (2021). *SNPassoc: SNPs-Based Whole Genome Association Studies*. R package version 2.0-11. URL: https://CRAN.R-project.org/package=SNPassoc.

Nelson, Jelani und Huy L. Nguyen (2013). „OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings“. In: *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*. IEEE Computer Society, S. 117–126. DOI: 10.1109/FOCS.2013.21. URL: https://doi.org/10.1109/FOCS.2013.21.

Parry, Katharina u. a. (2021). „Cross-Leverage Scores for Selecting Subsets of Explanatory Variables“. In: *arXiv preprint arXiv:2109.08399*.

Peters, Christian (2021). *l2s_sampling.py*. https://github.com/cxan96/oblivious-sketching-logreg/blob/main/sketching/l2s_sampling.py. [besucht am 26.04.2022].

PLINK: Whole genome data analysis toolset - Harvard University (2017). https://zzz.bwh.harvard.edu/plink/dataman.shtml. [besucht am 27.04.2022].

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rockel, Tobias (2020). *missMethods: Methods for Missing Data*. R package version 0.2.0. URL: https://CRAN.R-project.org/package=missMethods.

Ruczinski, Ingo, Charles Kooperberg und Michael LeBlanc (2003). „Logic regression“. In: *Journal of Computational and graphical Statistics* 12.3, S. 475–511.

Schreiber, Robert und Charles Van Loan (1989). „A storage-efficient WY representation for products of Householder transformations“. In: *SIAM Journal on Scientific and Statistical Computing* 10.1, S. 53–57.

Tomiuk, Jürgen und Volker Loeschcke (2017). *Grundlagen der evolutionsbiologie und formalen genetik*. Springer.

Venables, W. N. und B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wollenberg, Alexander (2016). „Reduktion hochdimensionaler Datensätze für die logische Regression unter der Verwendung von Leverage Scores mit besonderer Berücksichtigung von SNP-Daten“. Magisterarb. Technische Universität Dortmund.

A Kerndichteschätzer für die Simulationen 13 bis 20

A.1 Kerndichteschätzer der CLS der Simulationen 13 bis 20

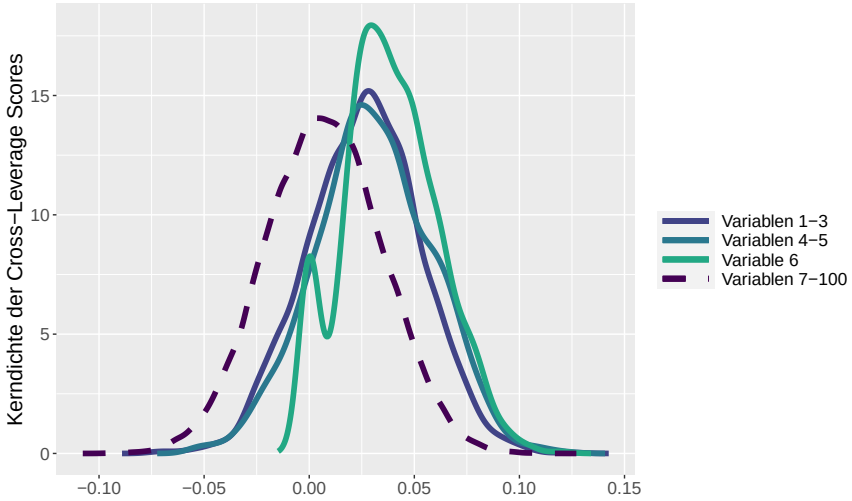

Abbildung 8: Kerndichteschätzer der CLS der Variablen für 1000 Wiederholungen von Simulation 13. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1 \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5) \vee X_6$

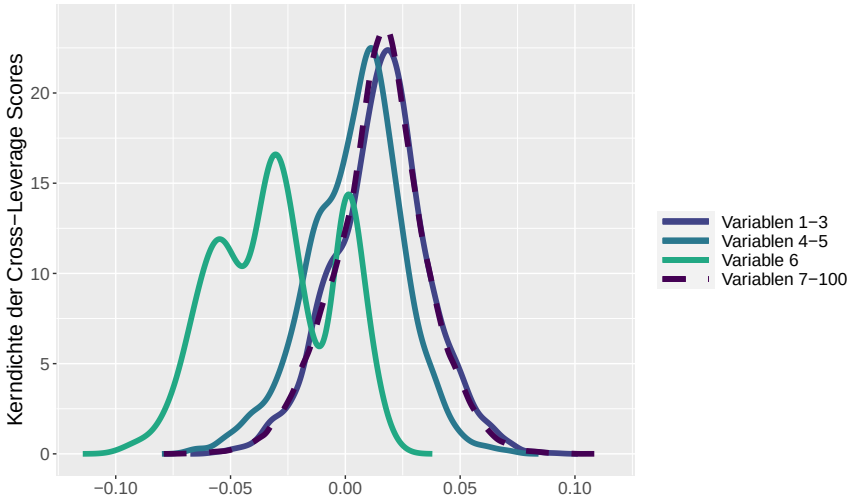

Abbildung 9: Kerndichteschätzer der CLS der Variablen für 1000 Wiederholungen von Simulation 14. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2^c \wedge X_3^c) \vee (X_4^c \wedge X_5^c) \vee X_6^c$

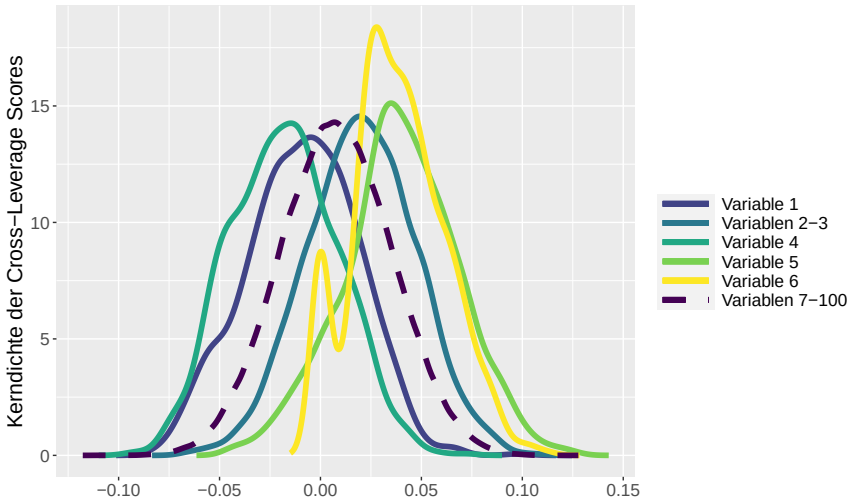

Abbildung 10: Kerndichteschätzer der CLS der Variablen für 1000 Wiederholungen von Simulation 15. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2 \wedge X_3) \vee (X_4^c \wedge X_5) \vee X_6$

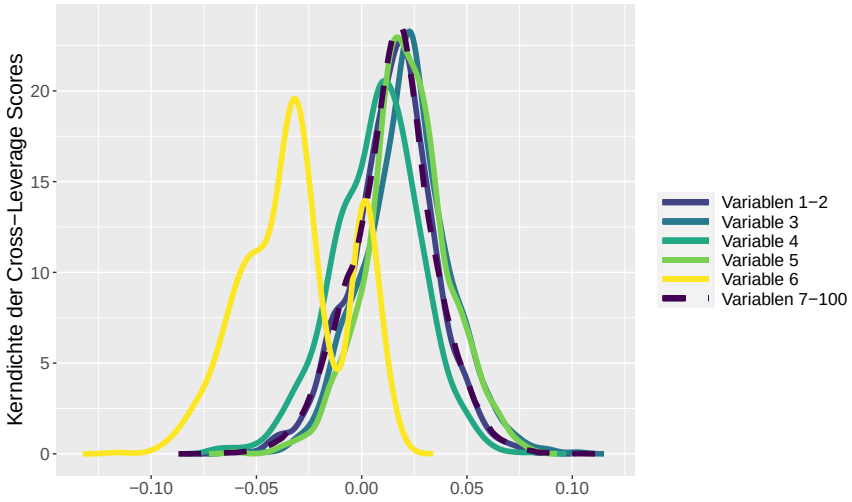

Abbildung 11: Kerndichteschätzer der CLS der Variablen für 1000 Wiederholungen von Simulation 16. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2^c \wedge X_3) \vee (X_4^c \wedge X_5) \vee X_6^c$

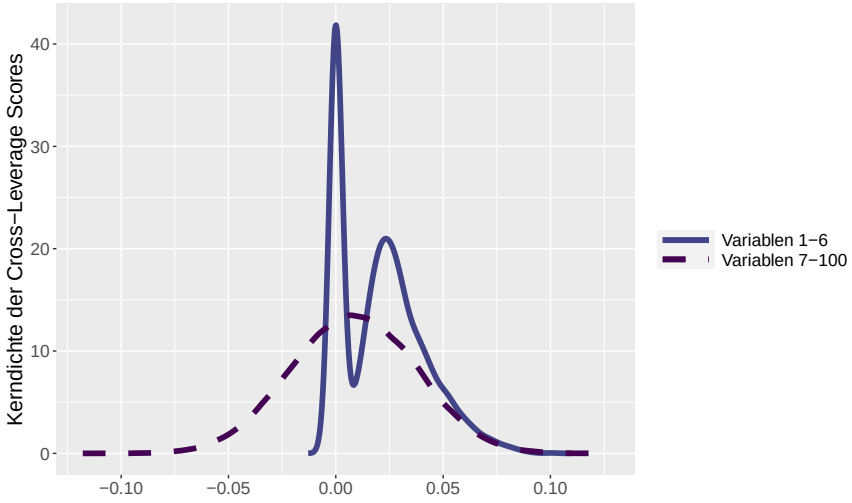

Abbildung 12: Kerndichteschätzer der CLS der Variablen für 1000 Wiederholungen von Simulation 17. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = X_1 \vee X_2 \vee X_3 \vee X_4 \vee X_5 \vee X_6$

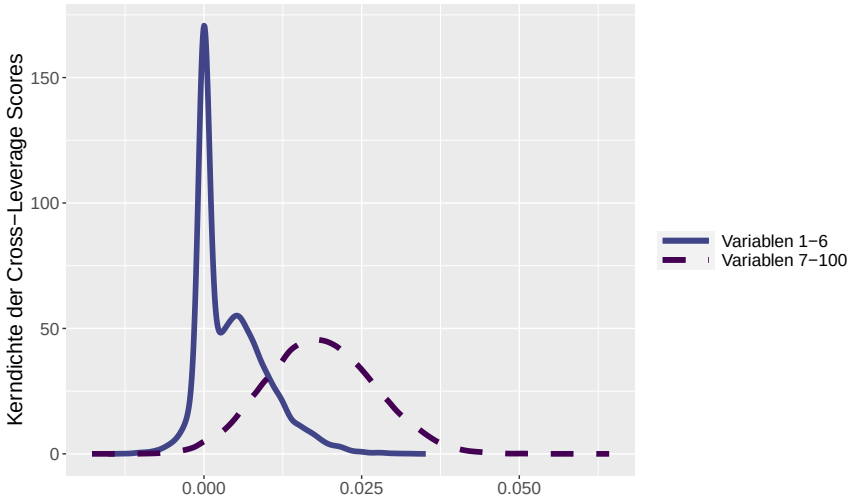

Abbildung 13: Kerndichteschätzer der CLS der Variablen für 1000 Wiederholungen von Simulation 18. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = X_1^c \vee X_2^c \vee X_3^c \vee X_4^c \vee X_5^c \vee X_6^c$

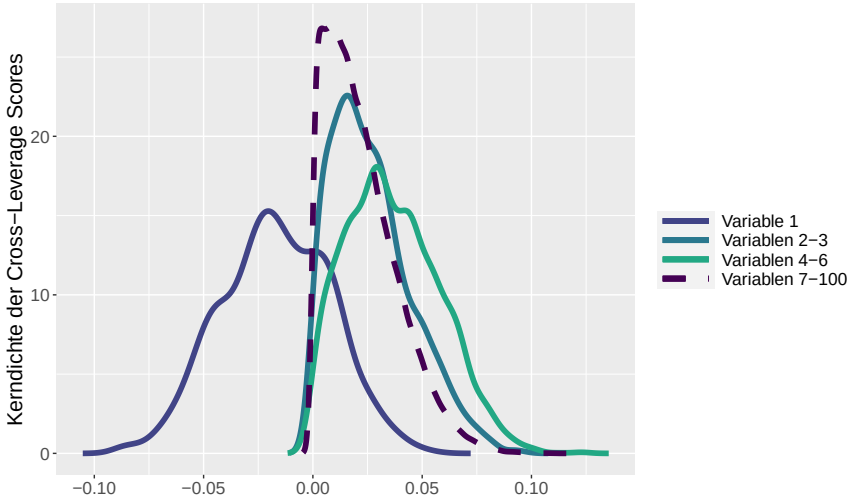

Abbildung 14: Kerndichteschätzer der CLS der Variablen für 1000 Wiederholungen von Simulation 19. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5 \wedge X_6)$

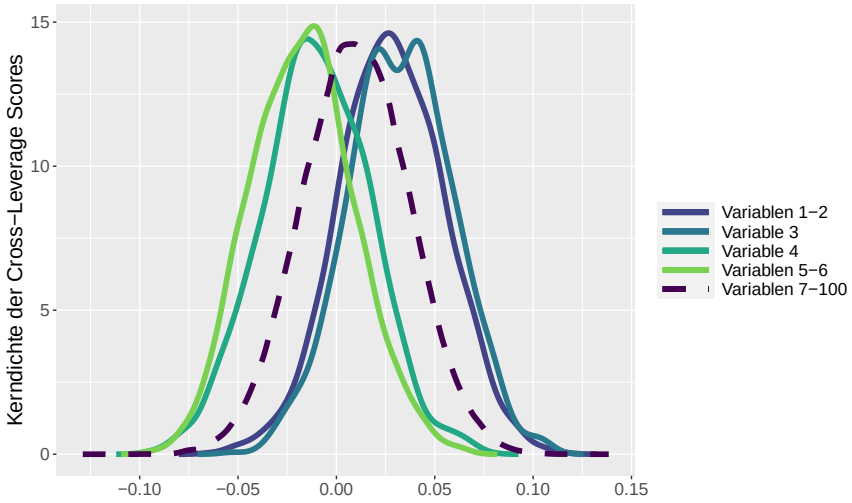

Abbildung 15: Kerndichteschätzer der CLS der Variablen für 1000 Wiederholungen von Simulation 20. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1 \wedge X_2) \vee (X_3 \wedge X_4^c) \vee (X_5^c \wedge X_6^c)$

A.2 Kerndichteschätzer der absoluten CLS der Simulationen 14 bis 20

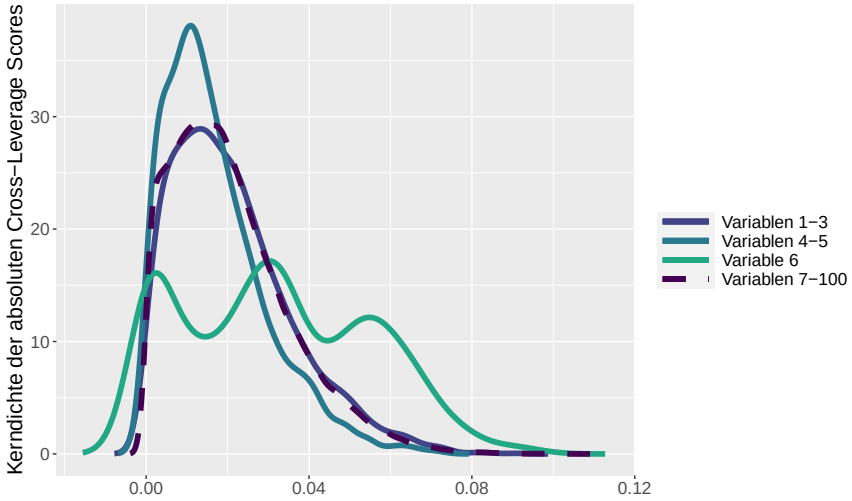

Abbildung 16: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 14. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2^c \wedge X_3^c) \vee (X_4^c \wedge X_5^c) \vee X_6^c$

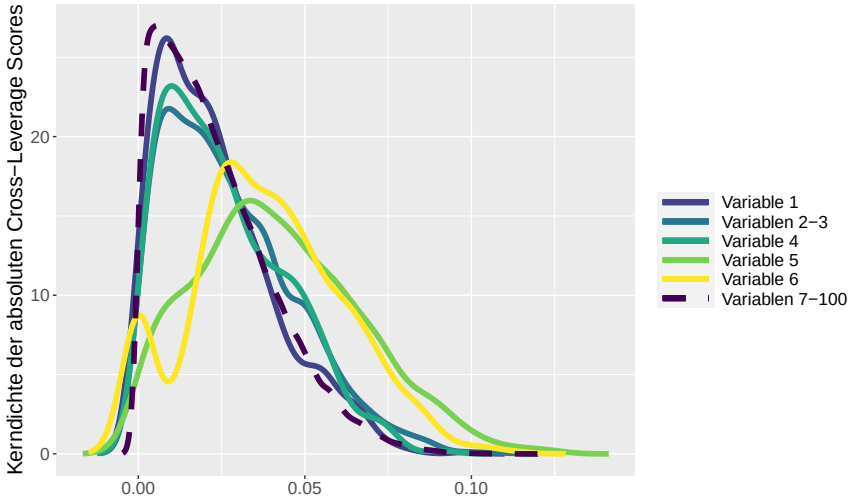

Abbildung 17: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 15. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2 \wedge X_3) \vee (X_4^c \wedge X_5) \vee X_6$

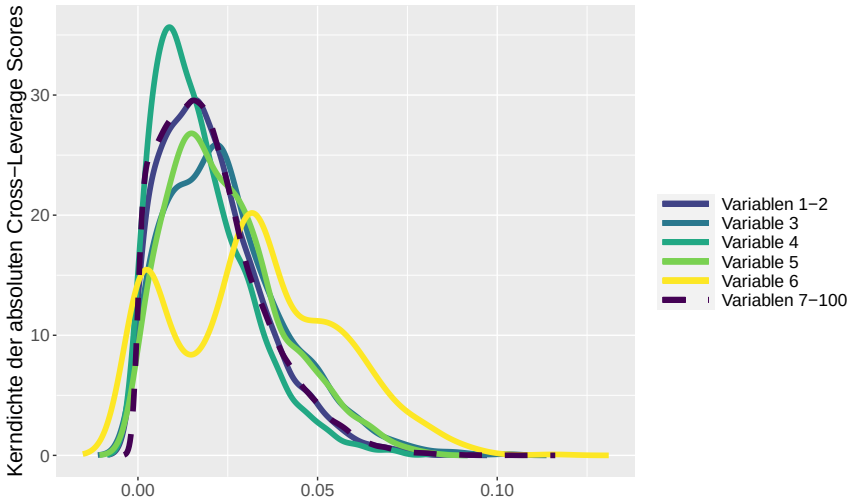

Abbildung 18: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 16. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2^c \wedge X_3) \vee (X_4^c \wedge X_5) \vee X_6^c$

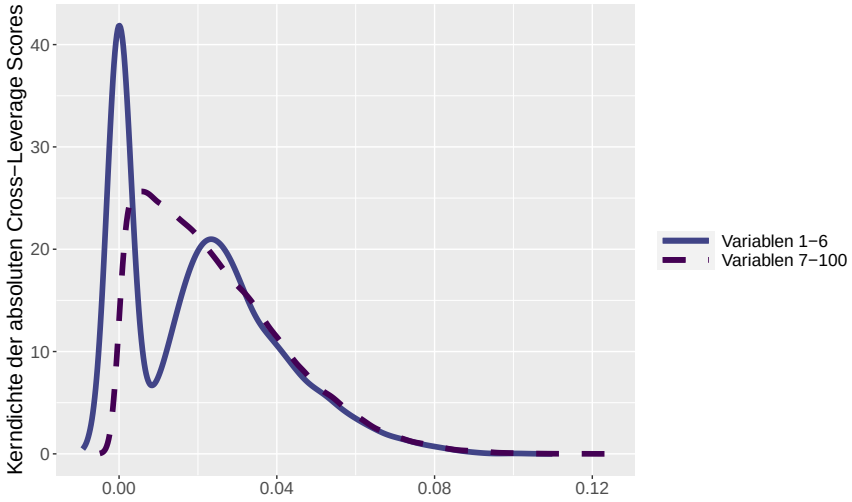

Abbildung 19: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 17. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = X_1 \vee X_2 \vee X_3 \vee X_4 \vee X_5 \vee X_6$

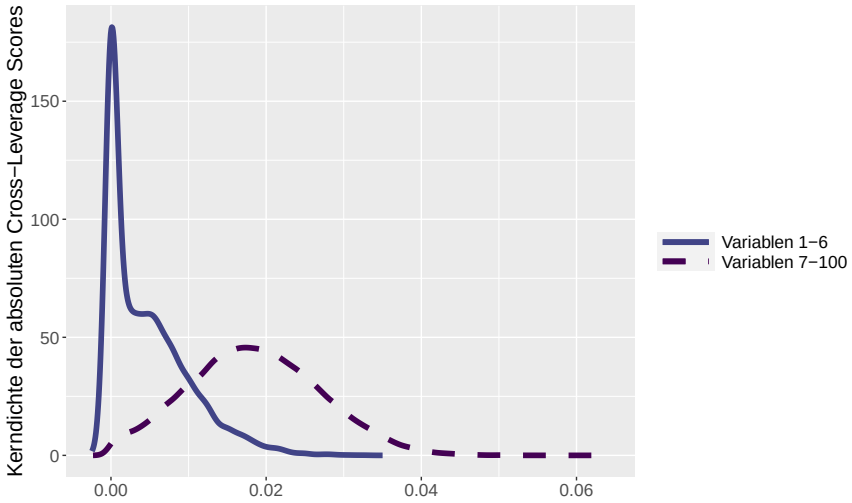

Abbildung 20: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 18. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = X_1^c \vee X_2^c \vee X_3^c \vee X_4^c \vee X_5^c \vee X_6^c$

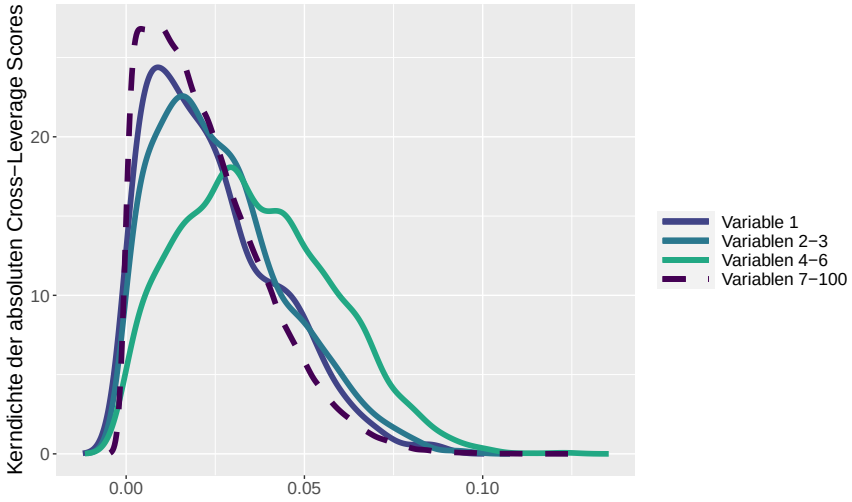

Abbildung 21: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 19. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5 \wedge X_6)$

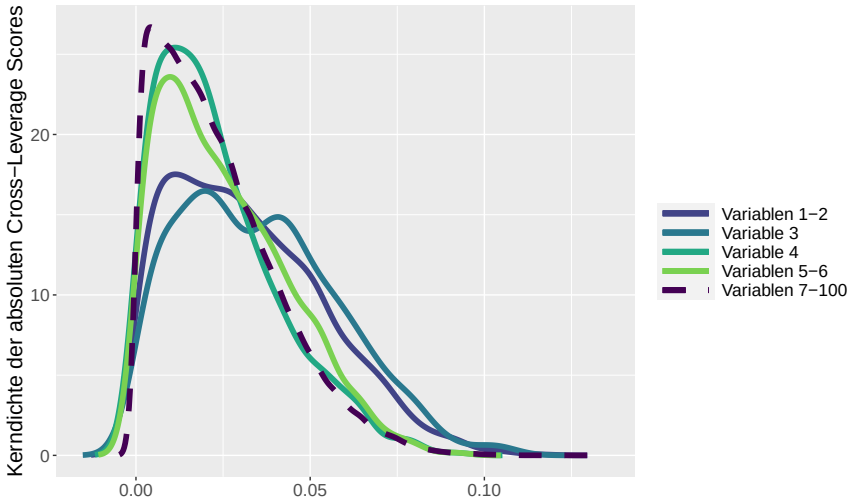

Abbildung 22: Kerndichteschätzer der betragsmäßigen CLS der Variablen für 1000 Wiederholungen von Simulation 20. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1 \wedge X_2) \vee (X_3 \wedge X_4^c) \vee (X_5^c \wedge X_6^c)$

A.3 Kerndichteschätzer der LS der Simulationen 14 bis 20

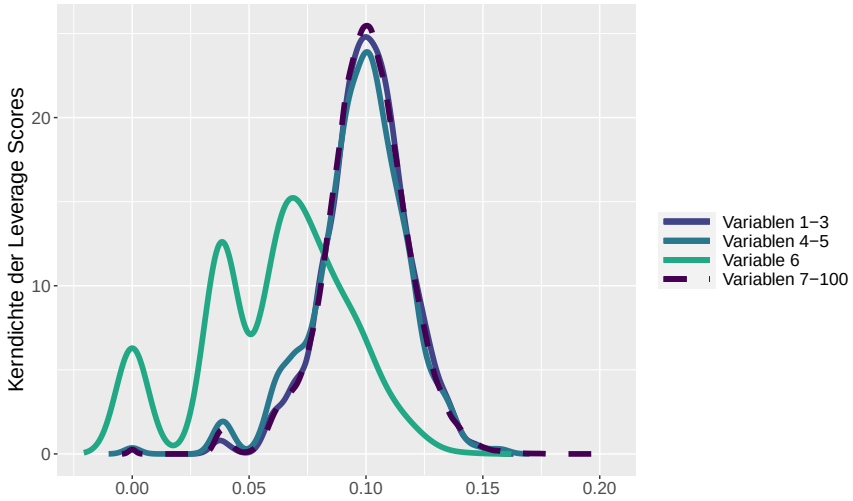

Abbildung 23: Kerndichteschätzer der LS der Variablen für 1000 Wiederholungen von Simulation 14. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2^c \wedge X_3^c) \vee (X_4^c \wedge X_5^c) \vee X_6^c$

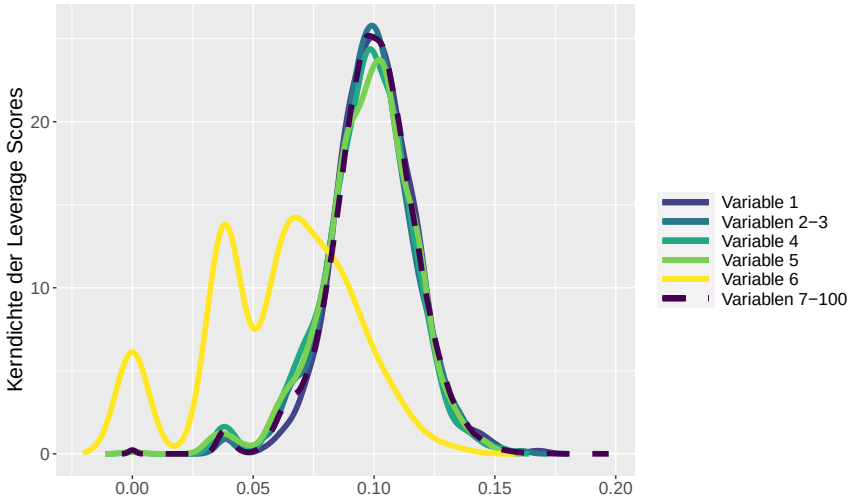

Abbildung 24: Kerndichteschätzer der LS der Variablen für 1000 Wiederholungen von Simulation 15. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2 \wedge X_3) \vee (X_4^c \wedge X_5) \vee X_6$

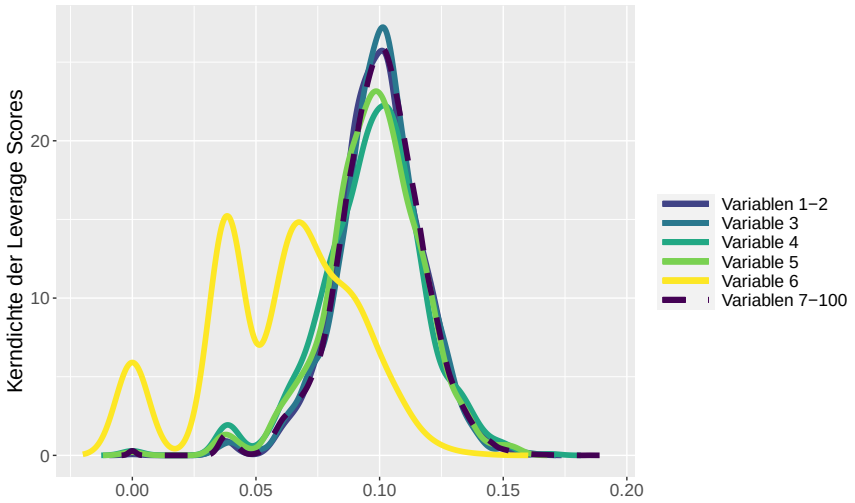

Abbildung 25: Kerndichteschätzer der LS der Variablen für 1000 Wiederholungen von Simulation 16. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2^c \wedge X_3) \vee (X_4^c \wedge X_5) \vee X_6^c$

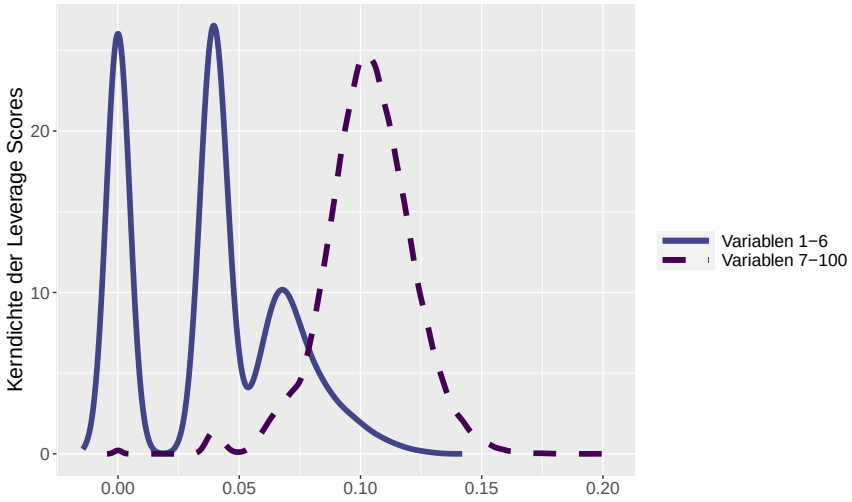

Abbildung 26: Kerndichteschätzer der LS der Variablen für 1000 Wiederholungen von Simulation 17. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = X_1 \vee X_2 \vee X_3 \vee X_4 \vee X_5 \vee X_6$

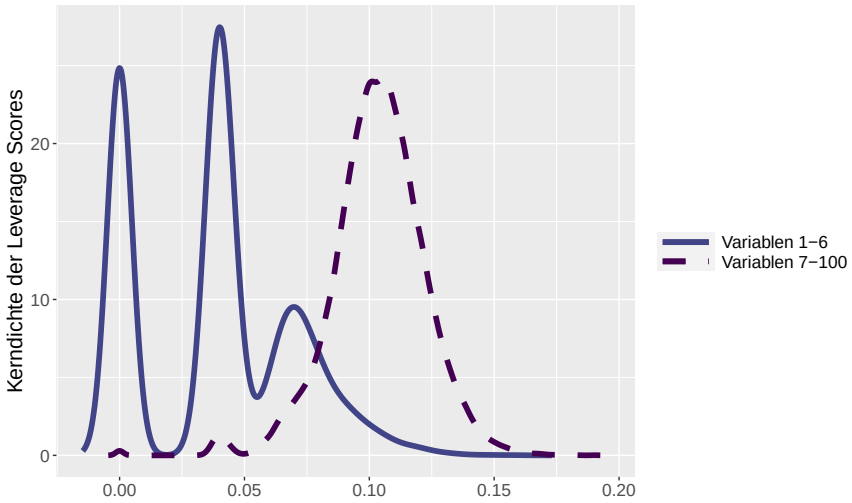

Abbildung 27: Kerndichteschätzer der LS der Variablen für 1000 Wiederholungen von Simulation 18. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = X_1^c \vee X_2^c \vee X_3^c \vee X_4^c \vee X_5^c \vee X_6^c$

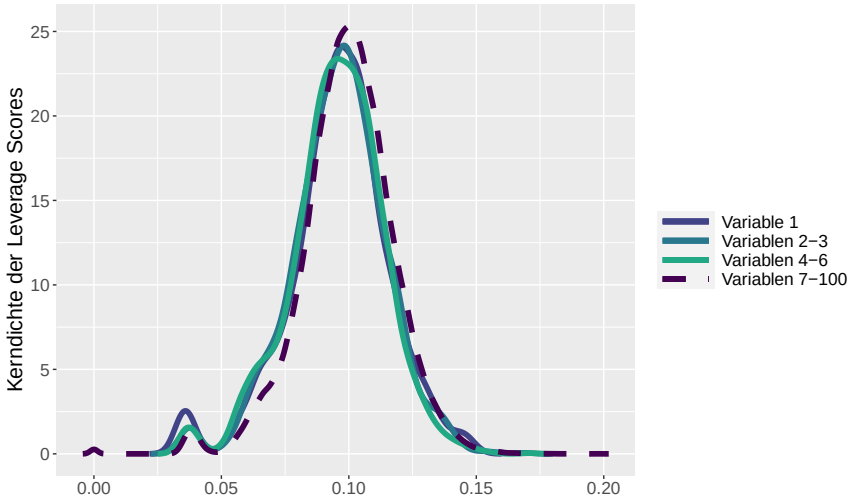

Abbildung 28: Kerndichteschätzer der LS der Variablen für 1000 Wiederholungen von Simulation 19. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1^c \wedge X_2 \wedge X_3) \vee (X_4 \wedge X_5 \wedge X_6)$

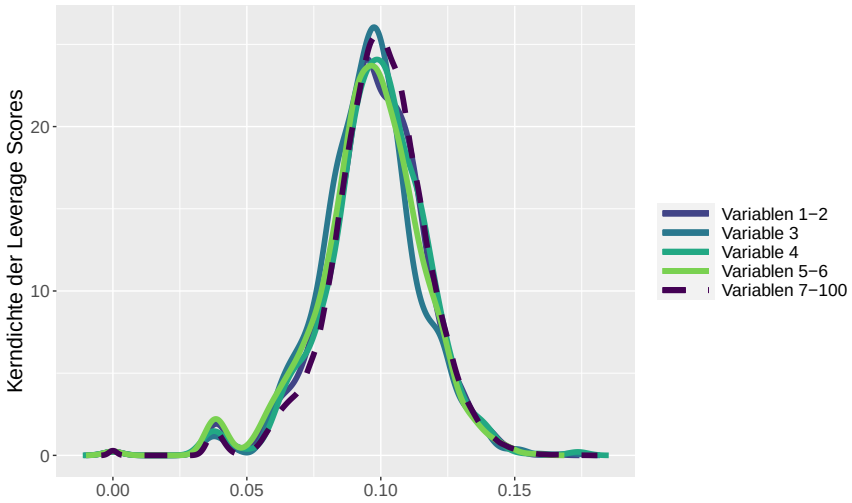

Abbildung 29: Kerndichteschätzer der LS der Variablen für 1000 Wiederholungen von Simulation 20. Der Kerndichteschätzer, der zu den irrelevanten Variablen gehört, ist mit einer gestrichelten Linie abgebildet. Die logische Einflussvariable ist $L = (X_1 \wedge X_2) \vee (X_3 \wedge X_4^c) \vee (X_5^c \wedge X_6^c)$