

LUDWIG-MAXIMILIANS-UNIVERSITÄT

INSTITUT FÜR STATISTIK



---

Vergleich von Methoden zur Schätzung kausaler  
Effekte unter Einbeziehung von Shrinkage-Verfahren

---

MASTERARBEIT

Alexander Neumaier

Betreuung: Prof. Dr. Christian Heumann

München, 13. Juni 2018

# Abstract

Während Assoziationen mit Hilfe statistischer Methoden meist gut geschätzt werden können, ist die Untersuchung kausaler Effekte eine komplexere Angelegenheit. Oft sorgen Confoundervariablen dafür, dass eine kausale Interpretation eines gemessenen Zusammenhangs nicht möglich ist. Deshalb ist es sinnvoll, solche Variablen zu identifizieren und bei der Modellierung zu berücksichtigen. In dieser Arbeit werden verschiedene Möglichkeiten zur Schätzung kausaler Effekte dargestellt. Dabei werden sowohl die gängigen Verfahren, IP-Weighting und Standardisierung, als auch eine Kombination der beiden Ansätze, Targeted Maximum Likelihood Estimation, betrachtet. Letzteres Verfahren kann durch Variablenselektions- und Shrinkage-Algorithmen ausgebaut werden. Dadurch ergeben sich erweiterte Schätzverfahren, welche ebenfalls herangezogen werden. Um die Güte der Methoden zu untersuchen und zu vergleichen, werden Simulationen unter verschiedenen datengenerierenden Prozessen durchgeführt. Dabei werden mehrere Variablen erzeugt, wobei für die Schätzung unklar ist, welche davon herangezogen werden sollten, um das Confounding-Problem zu beheben. Eine Evaluation der Verfahren erfolgt durch den Mean Squared Error (MSE).

Die Ergebnisse zeigen: Werden bei der Modellierung alle notwendigen Terme spezifiziert, wird der kausale Effekt durch Standardisierung am besten geschätzt. Die erweiterten Schätzverfahren weisen in den meisten Fällen höhere MSE auf als Standardisierung oder Targeted Maximum Likelihood Estimation. Bei Fehlspezifikation schneiden sie teilweise etwas besser ab. Dies lässt darauf schließen, dass die Einbeziehung von Variablenselektion und Shrinkage unter Umständen sinnvoll ist, sich jedoch auch negativ auf die Schätzung auswirken kann. Eine Verbesserung gegenüber der gewöhnlichen Targeted Maximum Likelihood Estimation kann jedoch im Großen und Ganzen nicht bestätigt werden.

# Notation

Bezeichnung	Bedeutung
$i = 1, \dots, n$	Index für die Beobachtungseinheiten
$A_i$	Binäre Treatmentvariable
$Y_i$	Outcome
$Y_i^{a=0}$	Potential Outcome ohne Treatment
$Y_i^{a=1}$	Potential Outcome bei Treatment
$\psi_{ATE}$	Kausaler Effekt von A auf Y (Average Treatment Effect)
$L_i = (L_{i1}, \dots, L_{ip})$	Weitere Variablen (Potenzielle Confounder)
$L_i^0 = (L_{i1}^0, \dots, L_{ip0}^0)$	Variablen mit Effekt auf A und auf Y (Confounder)
$L_i^A = (L_{i1}^A, \dots, L_{ipA}^A)$	Variablen mit Effekt auf A, aber nicht auf Y
$L_i^Y = (L_{i1}^Y, \dots, L_{ipY}^Y)$	Variablen mit Effekt auf Y, aber nicht auf A
$L_i^N = (L_{i1}^N, \dots, L_{ipN}^N)$	Variablen ohne Effekt auf A oder auf Y
$r, r_0, r_A, r_Y, r_N$	Indizes für die Variablenvektoren $L_i, L_i^0, L_i^A, L_i^Y, L_i^0$
$Z_i$	Verborgene Variable bzw. verborgener Variablenvektor
$\kappa$	Koeffizient für den Einfluss der Variable(n) $Z$ auf $L$
$\mu$	Interceptvektor zur Generierung der Variablen $L$
$\nu$	Standardabweichungsvektor zur Generierung der Variablen $L$
$\tau_0$	Intercept zur Generierung/Modellierung der Treatmentvariable
$\tau$	Koeffizientenvektor der einfachen Terme zur Generierung/ Modellierung der Treatmentvariable
$\nu$	Koeffizientenvektor der quadratischen Terme zur Generierung/ Modellierung der Treatmentvariable
$\hat{g}_i$	Vorhersage des Exposure-Modells
$\hat{H}_i$	„Cleverer Kovariable“ (zur Anpassung der Outcome-Vorhersagen)
$\phi$	Parameter für die „Cleverer Kovariable“

$\theta_0$	Intercept zur Generierung/Modellierung des Outcomes
$\theta_1$	Koeffizient der Variable $A$ zur Generierung/Modellierung des Outcomes
$\beta$	Koeffizientenvektor der (sonstigen) einfachen Terme zur Generierung/Modellierung des Outcomes
$\gamma$	Koeffizientenvektor der Interaktionsterme zur Generierung/Modellierung des Outcomes
$\delta$	Koeffizientenvektor der quadratischen Terme zur Generierung/Modellierung des Outcomes
$\epsilon_i$	Störterm zur Generierung/Modellierung des Outcomes
$\sigma^2$	Varianz der Störterme
$\hat{Q}_i$	Vorhersage des geschätzten Outcome-Modells
$\hat{Q}_i^*$	Angepasste Vorhersage des geschätzten Outcome-Modells
$l_s(Y, \hat{Q}^*)$	Verlust nach der negativen Log-Likelihood für bernoulliverteilte Zufallsvariablen
$m = 1, \dots, M$	Index für die Teildatensätze bei $M$ -facher Kreuzvalidierung
$\lambda$	Parameter für die Stärke der Penalisierung (Shrinkage)
$\alpha$	Parameter zwischen 0 und 1, der das Verhältnis zwischen Lasso- und Ridge-Penalisierung steuert
$k = 0, \dots, K$	Index über $K + 1$ verschiedene Penalisierungsstärken
$j = 1, \dots, R$	Index für die Wiederholungen einer Simulation
$\hat{\psi}_{ATE}(T)_j$	Schätzung des ATE in Wiederholung $j$ unter Schätzverfahren $T$
$MSE(\psi_{ATE}, T)$	Mean Squared Error von Schätzverfahren $T$ zur Schätzung von $\psi_{ATE}$
$MSE_{emp}(\psi_{ATE}, \hat{\psi}_{ATE}(T))$	Empirischer Mean Squared Error von Schätzungen unter Schätzverfahren $T$ zur Schätzung von $\psi_{ATE}$

# Inhaltsverzeichnis

<b>1</b>	<b>Kausalität - Einführung in die Problematik</b>	<b>1</b>
<b>2</b>	<b>Kausaler Effekt - Grundlagen</b>	<b>3</b>
2.1	Definition . . . . .	3
2.2	Confounding und Exchangeability . . . . .	5
<b>3</b>	<b>Methoden zur Schätzung kausaler Effekte</b>	<b>7</b>
3.1	„Einfache“ Methoden . . . . .	7
3.1.1	„Naive Schätzung“ . . . . .	7
3.1.2	Standardisierung . . . . .	8
3.1.3	IP-Weighting . . . . .	9
3.2	Targeted Maximum Likelihood Estimation . . . . .	11
3.3	Schätzung mit Variablenselektion . . . . .	13
3.3.1	Warum Variablenselektion? . . . . .	13
3.3.2	CTMLE-Variablenselektionsalgorithmus . . . . .	14
3.3.3	Einschub: Kreuzvalidierung . . . . .	15
3.4	Schätzung mit Shrinkage . . . . .	17
3.4.1	Einschub: Shrinkage-Schätzung . . . . .	17
3.4.2	CTMLE-Algorithmus mit Shrinkage-Schätzung . . . . .	19
3.4.3	CTMLE-Algorithmus mit komponentenweisem Shrinkage . . . . .	20
<b>4</b>	<b>Simulationsstudie zur Untersuchung der Schätzverfahren</b>	<b>23</b>
4.1	Aufbau der Simulationen . . . . .	23
4.1.1	Allgemeine Struktur . . . . .	23
4.1.2	Generierung der Daten . . . . .	24
4.1.3	Schätzung des kausalen Effekts . . . . .	28
4.1.4	MSE-Berechnung . . . . .	28
4.2	Durchführung der Simulationen . . . . .	30
4.2.1	Betrachtete datengenerierende Prozesse . . . . .	30
4.2.2	Verwendete Schätzmethoden . . . . .	32
4.3	Ergebnisse der Simulationen . . . . .	34
4.3.1	Erste Ergebnisse . . . . .	34

4.3.2	Verschiedene Assoziationsstrukturen . . . . .	38
4.3.3	Fehlspezifikation des Outcome-Modells . . . . .	40
4.3.4	Fehlspezifikation von Outcome-Modell und Exposure-Modell . . .	48
<b>5</b>	<b>Zusammenfassung und Abschluss</b>	<b>52</b>
5.1	Zusammenfassung . . . . .	52
5.2	Grenzen der Analysen und Ausblick . . . . .	54

# 1 Kausalität - Einführung in die Problematik

Den Begriff „Kausalität“ von Begriffen wie „Assoziation“, „Zusammenhang“ oder „Korrelation“ zu trennen, ist in der statistischen Datenanalyse von großer Bedeutung. Man betrachte dazu folgendes Beispiel:

Für jeden Tag seien die Gesamtumsatzzahlen von Eisdielen gegeben. Des Weiteren liegt pro Tag die Anzahl der Personen vor, die aufgrund von Kreislaufproblemen ins Krankenhaus eingeliefert werden. Von Interesse sei die Beziehung dieser beiden Variablen. Ein einfacher Ansatz, um diese zu untersuchen, ist die Anwendung eines linearen Regressionsmodells, wobei man den Eisdielenumsatz als Einflussgröße und die Anzahl der Kreislaufkrankungen als Zielgröße festlegt. Man gehe von dem (nicht ungewöhnlichen) Fall aus, dass man für den Steigungsparameter einen (bedeutenden) positiven Maximum-Likelihood-Schätzer erhält. Die folgenden zwei Interpretationen des geschätzten Koeffizienten werden betrachtet:

- „An Tagen, an denen viel Eis verkauft wird, ereignen sich tendenziell mehr Kreislaufkrankungen, als an Tagen, an denen wenig Eis verkauft wird.“
- „Eine höhere Umsatzzahl von Eisdielen sorgt für eine höhere Anzahl von Kreislaufkrankungen. Eis verursacht Kreislaufprobleme.“

Bei der ersten Aussage wird der Koeffizient lediglich „assoziativ“ interpretiert. Im zweiten Fall hingegen wird auf eine Kausalität, also eine Ursache-Wirkungsbeziehung, geschlossen. Letzteres ist nicht unproblematisch: Das aufgestellte Modell schätzt die Assoziation, ist jedoch nicht geeignet, um eine kausale Wirkungsbeziehung zu untersuchen. Der Grund dafür ist, dass eine sehr wichtige Variable nicht berücksichtigt wird: Die Temperatur an den betrachteten Tagen. An Tagen, an denen viel Eis verkauft wird, ist es tendenziell wärmer als an Tagen, an denen wenig Eis verkauft wird. Höhere Temperaturen sorgen außerdem für eine höhere Anzahl an Kreislaufproblemen. Somit gibt es noch einen zweiten „Weg“, der durch das aufgestellte Modell mitgemessen wird, ohne dass dies die Absicht ist. Dieser sorgt für einen Zusammenhang zwischen Eisdielenumsatz und Kreislaufproblemen, obwohl eine kausale Beziehung zwischen diesen Variablen eher

fragwürdig ist. Wird die Temperatur als Kovariable ins Modell aufgenommen und somit „kontrolliert“, ist zu erwarten, dass sich der Schätzer für den Koeffizienten ändert.

Das Beispiel zeigt bereits, dass sich ein kausaler Effekt und eine Assoziation unter Umständen deutlich unterscheiden können. Dies sollte sowohl bei der Wahl als auch bei der Interpretation des Modells berücksichtigt werden. Man sollte sich überlegen: Welche Absicht steht hinter der Modellierung? Geht es vorrangig um die Prognose einer Zielvariable mit Hilfe anderer vorliegender Variablen? Oder geht es hauptsächlich um die Kontrolle einer Zielvariable durch andere Variablen? Steht Letzteres im Fokus, so ist die Quantifizierung der Kausalität von zentraler Bedeutung.

Um eine geeignete Schätzung eines kausalen Effekts zu erhalten, ist es notwendig, Variablen, die weitere „Wege“ (wie oben beschrieben) öffnen, zu identifizieren und bei der Modellierung miteinzubeziehen. Solche Variablen werden als Confounder bezeichnet. Es gibt hierfür verschiedene Verfahren. Im Kapitel 2 der Arbeit wird der kausale Effekt und die Problematik statistisch formuliert, ehe in Kapitel 3 die Verfahren zur Schätzung kausaler Effekte dargestellt werden. Insbesondere werden hier auch Variablenselektions- und Shrinkage-Verfahren näher untersucht, da es auch Nachteile auf die Schätzung haben kann, wenn überflüssige Variablen ins Modell integriert werden. Anschließend werden in Kapitel 4 die erläuterten Verfahren mit Hilfe von Simulationen bezüglich ihrer Güte untersucht und verglichen. Hier werden verschiedene datengenerierende Prozesse herangezogen. Eine Zusammenfassung und ein Ausblick in Kapitel 5 schließen die Arbeit ab.



# 2 Kausaler Effekt - Grundlagen

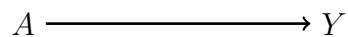
## 2.1 Definition

Man betrachte nun einen metrischen Outcome  $Y$  und der Einfachheit halber eine binäre Treatmentvariable  $A$ :

$$A = \begin{cases} 1, & \text{wenn Treatment stattfindet} \\ 0, & \text{wenn kein Treatment stattfindet} \end{cases}$$

„Treatment“ ist dabei im weiteren Sinne zu verstehen. Es wird angenommen, dass die statistischen Einheiten, an denen die Merkmale erhoben werden, unabhängig voneinander sind.

Von Interesse sei der kausale Effekt der Treatmentvariable auf den Outcome. Grafisch können kausale Effekte durch DAGs (directed acyclic graphs) dargestellt werden:



**Abbildung 2.1** : DAG - Kausaler Effekt

Die Ursache-Wirkungsbeziehungen werden hier durch gerichtete Wirkungspfeile veranschaulicht. Durch die Annahme, dass die Wirkungsstruktur azyklisch ist, wird bereits ausgeschlossen, dass ein Pfad wieder von  $Y$  zu  $A$  führt. Also insbesondere wird angenommen, dass  $Y$  keinen (direkten oder indirekten) kausalen Effekt auf  $A$  hat. Der kausale Effekte von  $A$  auf  $Y$  soll nun so definiert werden, dass er alle (direkten und indirekten) Pfade erfasst, welche von  $A$  wegführen und bei  $Y$  enden.

Es wird festgelegt:  $Y^{a=1}$  gibt den Potential Outcome für den Fall an, dass Treatment stattfindet, und  $Y^{a=0}$  für den Fall, dass kein Treatment stattfindet. Dies führt zur Annahme der Konsistenz. Im Kontext kausaler Effekte bedeutet das, dass der Potential Outcome  $Y^a$  dann dem beobachteten Outcome  $Y$  entspricht, wenn  $A = a$  als Treatmentvariable beobachtet wird. Oder anders formuliert:

$$Y = AY^{a=1} + (1 - A)Y^{a=0}$$

Somit ist pro Beobachtung immer nur einer der beiden Potential Outcomes bekannt, da für eine Beobachtung entweder  $A = 1$  oder  $A = 0$  ist, aber nicht beides (Die Beobachtung hat Treatment bekommen oder nicht). Eine Definition des kausalen Effekts für eine bestimmte Einheit ist über die Differenz der Potential Outcomes möglich:  $Y^{a=1} - Y^{a=0}$ . Dieser Wert ist unbekannt, da nur eine der beiden Größen vorliegen kann. Für die Grundgesamtheit lässt sich der kausale Effekt über den Erwartungswert definieren. Man spricht vom Average Treatment Effect (ATE):

$$\psi_{ATE} = E[Y^{a=1} - Y^{a=0}] = E[Y^{a=1}] - E[Y^{a=0}]$$

Der kausale Effekt bei binärer Treatmentvariable kann also als erwartete Differenz zwischen dem Potential Outcome bei Treatment  $a = 1$  und dem Potential Outcome bei Treatment  $a = 0$  einer Einheit definiert werden. Für jede Einheit sind zwei Y-Werte entscheidend - die zwei möglichen Outcomes, von denen maximal einer beobachtet werden kann. Der andere ist dann kontrafaktisch.

An dieser Stelle seien die bisher getroffenen Annahmen im Kontext kausaler Effekte noch einmal zusammengefasst. Diese werden auch für den Rest dieser Arbeit getroffen:

- Unabhängigkeit der Beobachtungseinheiten
- Azyklische Struktur der Ursache-Wirkungsbeziehungen
- Konsistenz:  $Y = AY^{a=1} + (1 - A)Y^{a=0}$

Wie unterscheidet sich der kausale Effekt nun von einer Assoziation?

Eine analoge Definition der Assoziation bei binärer Treatmentvariable ist durch die „assoziative Differenz“ möglich:

$$E[Y|A = 1] - E[Y|A = 0]$$

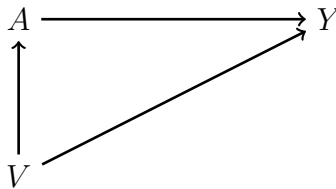
Diese entspricht der Differenz zwischen dem erwarteten Outcome, falls  $A = 1$  gegeben ist und dem erwarteten Outcome, falls  $A = 0$  gegeben ist. Lässt sich der Zusammenhang durch ein lineares Regressionsmodell mit  $A$  als Einflussgröße und  $Y$  als Zielgröße erklären, also in der Form  $E[Y|A] = \theta_0 + \theta_1 A$  (dummykodiert), so entspricht  $\theta_1$  der auf diese Weise definierten Assoziation. Im Gegensatz zum kausalen Effekt ist bei der Assoziation für jede Einheit nur ein Y-Wert entscheidend - der tatsächliche. Hier werden keine kontrafaktischen Outcomes betrachtet. Folglich kann die assoziative Differenz mit Hilfe eines geeigneten Regressionsmodells gut geschätzt werden.

(Vgl. Hernan; 2018, Part I S. 1 - 12 und Part II S. 11/12 ) (Vgl. Schomacker; 2017, Kapitel 1/2)

## 2.2 Confounding und Exchangeability

Wann unterscheiden sich kausaler Effekt und Assoziation?

Dies ist der Fall, wenn eine oder mehrere Variablen existieren, die einen sogenannten „Backdoor-Path“ öffnen. Ein „Backdoor-Path“ verbindet Treatmentvariable und Outcome durch eine gemeinsame Ursache. Werden die entsprechenden Variablen (sogenannte Confounder) nicht berücksichtigt, unterscheiden sich kausaler Effekt und Assoziation. Der „Backdoor-Path“ ändert nichts am ATE, da es sich nicht um einen Pfad handelt, der bei  $A$  beginnt und bei  $Y$  endet (Wirkungsrichtung ist entscheidend!). Er beeinflusst jedoch die Assoziation, da für diese die Wirkungsrichtung zwischen  $A$  und der Drittvariable nicht von Bedeutung ist. Man spricht dann vom sogenannten Confounding-Problem. Im einfachsten Fall (ein Confounder  $V$ ) kann das Problem folgendermaßen dargestellt werden:



**Abbildung 2.2 :** DAG - Confounding

Die Confoundervariable  $V$  stellt eine gemeinsame Ursache der Treatmentvariable  $A$  und des Outcomes  $Y$  dar und sorgt dadurch für einen Unterschied zwischen dem kausalen Effekt von  $A$  auf  $Y$  und der Assoziation dieser beiden Variablen.

Ein zentraler Begriff in diesem Zusammenhang ist die Eigenschaft (Mean) Exchangeability:

$$E[Y^a|A = a] = E[Y^a] \quad \forall a$$

Diese ist erfüllt, wenn der Erwartungswert von  $Y$  für die Treated-Einheiten derselbe ist wie für die Untreated-Einheiten, falls sie Treatment bekommen hätten, und umgekehrt. Exchangeability würde dazuführen, dass  $E[Y|A = 1] = E[Y^{a=1}]$  und  $E[Y|A = 0] = E[Y^{a=0}]$  gilt. Somit wären der Average Treatment Effect und die assoziative Differenz identisch, was zur Folge hätte, dass der kausale Effekt durch eine geeignete Regression gut geschätzt werden kann. Exchangeability ist jedoch nicht erfüllt, wenn Confounding vorliegt.

Um einen kausalen Effekt geeignet zu schätzen, ist es notwendig, die Confoundervariablen  $L^0$  zu identifizieren und bei der Modellierung zu berücksichtigen. Durch Bedingen auf diese Variablen wird Exchangeability erreicht. Man spricht von auf  $L^0$  bedingter

(Mean) Exchangeability:

$$E[Y^a|A = 1, L^0 = l^0] = E[Y^a|A = 0, L^0 = l^0] \quad \forall a, l^0$$

Beseitigt man das Confounding-Problem und erreicht dadurch Exchangeability, kann der ATE mit Hilfe von entsprechenden Regressionsmodellen geschätzt werden.

In der Praxis sind die Confounder oft nur teilweise oder gar nicht bekannt. Dennoch ist es wichtig, möglichst viele von ihnen zu erfassen, um eine geeignete Schätzung des kausalen Effekts zu erhalten. Nun könnte man schlussfolgern, dass jede denkbare Variable ins Modell aufgenommen werden sollte, um die Chance auf möglichst viele Confounder zu erhöhen. Dies ist jedoch problematisch:

- Ein Collider ist eine gemeinsame Wirkung von Treatmentvariable und Outcome (praktisch das Gegenteil eines Confounders). Auf eine Collidervariable sollte nicht bedingt werden, da hier eine Differenz zwischen dem Average Treatment Effect und der assoziativen Differenz erst durch das Bedingen entsteht. Bedingte Exchangeability liegt nicht vor, wenn in den entsprechenden Variablen Collider enthalten sind.
- Werden überflüssige Variablen in das Modell integriert, kann sich die Varianz der Schätzung erhöhen.

Es kann also sinnvoll sein, Variablenselektionsverfahren in die Schätzung zu integrieren. Im nächsten Kapitel werden die möglichen Schätzverfahren dargestellt. Diese bauen auf den Überlegungen in diesem Kapitel auf.

(Vgl. Hernan; 2018, Part I S. 83 - 95) (Vgl. Schomacker; 2017, Kapitel 1/2)

# 3 Methoden zur Schätzung kausaler Effekte

Es wird davon ausgegangen, dass die Daten bei der Schätzung in der folgenden Form vorliegen: Für  $n$  Beobachtungseinheiten ( $i = 1, \dots, n$ ) sind die binäre Treatmentvariable  $A_i$ , der (tatsächliche) Outcome  $Y_i$  und  $p$  potenzielle Confoundervariablen  $L_i = (L_{i1}, \dots, L_{ip})$  gegeben. Dabei sei nicht bekannt, welche der Variablen aus  $L$  einen Effekt auf  $A$  und/oder auf  $Y$  haben. Die Verfahren schätzen den Average Treatment Effect aus den Daten. Hyperparameter bzw. Modelle müssen für einzelne Verfahren zusätzlich festgelegt werden.

## 3.1 „Einfache“ Methoden

### 3.1.1 „Naive Schätzung“

Unter der „Naiven Schätzung“ versteht man die Anwendung eines einfachen Regressionsmodells mit  $Y$  als Zielgröße und  $A$  als Einflussgröße (dummykodiert):

$$Y_i = \theta_0 + A_i\theta_1 + \epsilon_i \quad \text{mit} \quad \epsilon_i \sim N(0, \sigma^2)$$

Der ATE wird in diesem Fall durch den zum Regressor  $A$  gehörigen ML-Schätzer ermittelt:  $\hat{\psi}_{ATE} = \hat{\theta}_1$ . Denn unter Exchangeability gilt hier:

$$\psi_{ATE} = E[Y^{a=1}] - E[Y^{a=0}] = E[Y|A = 1] - E[Y|A = 0] = (\theta_0 + \theta_1) - \theta_0 = \theta_1$$

Hier gehen zum einen die Annahmen des verwendeten Regressionsmodells ein. Zum anderen wird angenommen, dass kein Confounding vorliegt und Exchangeability bereits erfüllt ist, ohne dass auf weitere Variablen bedingt wird. Ist dies nicht der Fall, entsteht bei der Schätzung eine Verzerrung (wie in Kapitel 2.2 erläutert). Diese Methode kann somit als „naiv“ betrachtet werden, da Assoziation und Kausalität gleichgesetzt werden.

Generell lassen sich für nicht normalverteilte Outcomes auch generalisierte lineare Modelle verwenden (z.B. ein logistisches Regressionsmodell für binäres  $Y$ ). In diesem Fall entspricht der Schätzer des ATE nicht einfach dem Regressionschätzer, da die Parame-

ter dann unter Umständen anders interpretiert werden. Entsprechende Transformationen müssen vorgenommen werden.

### 3.1.2 Standardisierung

Bei der Standardisierung werden die Variablen  $L$  als Kovariablen ins Modell aufgenommen, um ein mögliches Confounding-Problem zu beseitigen und Exchangeability zu erreichen. Dadurch werden diese Variablen kontrolliert und die entsprechenden „Backdoor-Paths“ gegebenenfalls blockiert. Im einfachsten Fall kann dafür ein lineares Modell verwendet werden, welches lediglich einfache Terme enthält:

$$Y_i = \theta_0 + A_i\theta_1 + L_i\beta + \epsilon_i \quad \text{mit } \epsilon_i \sim N(0, \sigma^2)$$

Dabei ist  $\beta = (\beta_1, \dots, \beta_p)$  ein Vektor der Länge  $p$ .

Aber auch komplexere Modelle sind möglich. Zum Beispiel solche, die Interaktionsterme zwischen  $A$  und  $L$  oder quadratische Terme  $L^2$  enthalten:

$$Y_i = \theta_0 + A_i\theta_1 + L_i\beta + L_iA_i\gamma + \epsilon_i \quad \text{mit } \epsilon_i \sim N(0, \sigma^2)$$

$$Y_i = \theta_0 + A_i\theta_1 + L_i\beta + L_i^2\delta + \epsilon_i \quad \text{mit } \epsilon_i \sim N(0, \sigma^2)$$

$\gamma = (\gamma_1, \dots, \gamma_p)$  und  $\delta = (\delta_1, \dots, \delta_p)$  sind ebenfalls Koeffizientenvektoren der Länge  $p$ . Es gilt (unter auf  $L$  bedingter Exchangeability): Eine Schätzung des kausalen Effekts ist möglich, indem über die Variablen  $L$  in der folgenden Form „gemittelt“ wird:

$$E[Y^a] = \sum_l E[Y|A = a, L = l] \times P[L = l]$$

Da zumeist sehr viele Variablen  $L$  vorliegen, von denen ein Teil auch stetig ist, muss die Summe durch ein Integral und  $P[L = l]$  durch eine Wahrscheinlichkeitsdichte ersetzt werden. Folgende Vorgehensweise wird bei der Standardisierung herangezogen:

1. Es wird ein Modell mit  $Y$  als Zielgröße und  $A$  und  $L$  als Einflussgrößen geschätzt.
2. Für jede Beobachtung wird  $A_i = 1$  gesetzt und auf Basis des Modells die Vorhersage  $\hat{Q}_i(A_i = 1, L_i)$  der Zielvariable bestimmt. Anschließend bildet man den Mittelwert über alle Beobachtungen:  $\frac{1}{n} \sum_{i=1}^n \hat{Q}_i(A_i = 1, L_i)$ . Der errechnete Wert ist der Schätzer für  $E[Y^{a=1}]$ .
3. Analog werden die Vorhersagen  $\hat{Q}(A = 0, L)$  gebildet und damit  $E[Y^{a=0}]$  geschätzt.
4. Die Schätzung des ATE erfolgt gemäß der Definition:  $E[Y^{a=1}] - E[Y^{a=0}]$ .

Dieses Verfahren lässt sich auch für generalisierte lineare Modelle anwenden.

Verwendet man die Methode für den einfachen Fall (Lineares Modell, nur einfache Terme), so ergibt sich  $\hat{\theta}_1$  als Schätzer für den kausalen Effekt.

Die Annahmen des Standardisierungsverfahrens werden am Ende des nächsten Abschnitts zusammengefasst.

(Vgl. Hernan; 2018, Part II S. 23 - 31) (Vgl. Schomacker; 2017, Kapitel 5)

### 3.1.3 IP-Weighting

Ziel beim IP-Weighting ist es, den Wirkungspfeil von den Confoundern auf die Treatmentvariable zu entfernen. Damit wären die „Backdoor-Paths“ blockiert und der kausale Effekt würde der Assoziation entsprechen. Jede Beobachtung erhält dazu ein Gewicht:

$$W_i^A = 1/f(A_i|L_i)$$

Dies entspricht  $1/P[A_i = 1|L_i]$  für die Treated-Einheiten und  $1/P[A_i = 0|L_i] = 1/(1 - P[A_i = 1|L_i])$  für die Untreated-Einheiten. Es sollen also sozusagen Beobachtungen stärker gewichtet werden, bei denen die Treatmentvariable A weniger durch die Confounder beeinflusst wird.

$P[A = 1|L]$  ist unbekannt und kann mit Hilfe eines logistischen Regressionsmodells geschätzt werden. Dabei werden die Variablen  $L$  als Einflussgrößen und  $A$  als binäre Zielgröße modelliert. Werden nur einfache Terme aufgenommen, hat das Modell die folgende Form:

$$P[A_i = 1|L_i] = \frac{1}{1 + \exp(-(\tau_0 + L_i\tau))} \quad \text{mit } \tau = (\tau_1, \dots, \tau_p)$$

Dadurch ergibt sich ein Schätzer  $\hat{P}[A_i = 1|L_i]$  für die Treatmentwahrscheinlichkeit jeder Beobachtung. Man spricht auch vom Exposure-Modell und den Vorhersagen  $\hat{g}_i$ .

Die beobachtungsspezifischen Gewichte errechnen sich wie oben beschrieben:  $W_i = 1/\hat{f}(A_i|L_i)$ .

Anschließend wird eine gewichtete KQ-Schätzung des Modells  $E[Y|A] = \theta_0 + \theta_1 A$  mit den Gewichten  $W_i$  durchgeführt. Der folgende Term wird dazu bezüglich  $(\theta_0, \theta_1)$  minimiert:

$$\sum_{i=1}^n W_i [Y_i - (\theta_0 + \theta_1 A_i)]^2$$

Der ATE wird dann durch  $\hat{\theta}_1$  geschätzt.

Im Prinzip passiert dabei Folgendes: Es wird eine Pseudo-Population erzeugt, in der jedes Individuum zweimal berücksichtigt wird (einmal mit  $a = 1$  und einmal mit  $a = 0$ ).

Somit ist in dieser Pseudopopulation die Wirkung von  $L$  auf  $A$  eliminiert und die Assoziation von  $A$  und  $Y$  kann kausal interpretiert werden.

Oft werden auch stabilisierte Gewichte benutzt. Diese haben folgende Form:

$$W_i^A = f(A_i)/f(A_i|L_i)$$

$f(A_i)$  entspricht dabei  $P[A_i = 1]$  für die Treated-Einheiten und  $P[A_i = 0]$  für die Untreated-Einheiten.

Folgende zentrale Annahmen werden beim IP-Weighting bzw. bei der Standardisierung getroffen:

- Zunächst einmal wird auf Modelle zurückgegriffen. Beim IP-Weighting wird  $P[A = 1|L]$  modelliert, bei der Standardisierung modelliert man  $E[Y|A, L]$ . Dabei werden bereits verschiedene Modellannahmen getroffen. Sind diese nicht erfüllt, so ist das Modell fehlspezifiziert und erzeugt eine Verzerrung oder eine höhere Varianz bei der Schätzung. Eine gewisse Fehlspezifikation ist in der Praxis oft unumgänglich, es geht zumeist nur darum, diese möglichst gering zu halten.
- Ein zweite wichtige Annahme ist, dass durch Bedingen auf  $L$  Exchangeability erreicht wird. Dies ist beispielsweise nicht möglich, wenn Confounder existieren, die nicht in  $L$  enthalten sind. Ist diese Annahme nicht erfüllt, entsteht wie bereits erläutert eine Verzerrung.
- Außerdem wird angenommen, dass Positivität gilt: Das bedeutet, dass die Treatmentwahrscheinlichkeit für jede mögliche Ausprägung  $l$  der möglichen Confounder nicht 0 sein darf. (Ebenso darf die Wahrscheinlichkeit für  $A = 0$  auch nicht 0 sein.) Eine Verletzung der Annahme entsteht somit, wenn gewisse Umstände  $l$  vorliegen, die für unmögliches oder sicheres Treatment sorgen. Aber auch sehr starke Zusammenhänge zwischen  $A$  und  $L$  können bereits dafür sorgen, dass die Positivität in den Daten praktisch verletzt ist.

Besonders das IP-Weighting ist sehr anfällig für Positivitätsverletzungen.

Bei den weiteren Verfahren werden die bisher verwendeten Schätzansätze wieder aufgegriffen und erweitert.

(Vgl. Hernan; 2018, Part II S. 11 - 22, 28) (Vgl. Schomacker; 2017, Kapitel 6)



## 3.2 Targeted Maximum Likelihood Estimation

Die Targeted Maximum Likelihood Estimation verbindet die Ansätze aus der Standardisierung und dem IP-Weighting. Für diese Methode ist es sinnvoll, den Outcome  $Y$  auf den Wertebereich  $[0, 1]$  zu beschränken. Dies geschieht über folgende Transformation:

$$Y^* = \frac{Y - \min}{\max - \min}$$

Dabei stellen  $\max$  und  $\min$  die Grenzen des Wertebereichs dar. Man geht anschließend folgendermaßen vor:

1. Zunächst wird, wie bei der Standardisierung, das Outcome-Modell  $E[Y|A, L]$  geschätzt. Dafür kann (aufgrund des beschränkten Outcomes) zum Beispiel ein logistisches Regressionsmodell bzw. ein Quasi-Binomialmodell verwendet werden. Mit Hilfe des Modells kann auch hier für jede Beobachtung eine Vorhersage  $\hat{Q}_i(A_i = 1, L_i)$  und eine Vorhersage  $\hat{Q}_i(A_i = 0, L_i)$  berechnet werden.
2. Als Nächstes schätzt man das Exposure-Modell  $P[A = 1|L]$  wie beim IP-Weighting. Es ergeben sich die Schätzungen  $\hat{g}_i(1|L_i) = \hat{P}(A_i = 1|L_i)$  und  $\hat{g}_i(0|L_i) = \hat{P}(A_i = 0|L_i) = 1 - \hat{P}(A_i = 1|L_i)$  für jede Beobachtung. Damit lässt sich eine „Cleverer Kovariable“ berechnen:

$$\hat{H}_i(A_i, L_i) = \frac{I(A_i = 1)}{\hat{g}_i(1|L_i)} - \frac{I(A_i = 0)}{\hat{g}_i(0|L_i)}$$

$I$  stellt dabei die Indikatorfunktion dar.

3. Es folgt der Update-Schritt. Man modelliert dazu:

$$\log \left( \frac{P(Y_i = 1|A_i, L_i)}{1 - P(Y_i = 1|A_i, L_i)} \right) = \log \left( \frac{\hat{Q}_i(A_i, L_i)}{1 - \hat{Q}_i(A_i, L_i)} \right) + \phi \hat{H}_i(A_i, L_i)$$

Dies entspricht einer logistischen Regression mit Zielvariable  $Y$ , „Cleverer Kovariable“  $\hat{H}(A, L)$ , ohne Intercept und mit Offset (Variable, für die der zugehörige Koeffizient mit 1 festgelegt wird):  $\log \left( \frac{\hat{Q}(A, L)}{1 - \hat{Q}(A, L)} \right)$ . Dadurch erhält man einen Schätzer für den Koeffizienten der „Cleveren Kovariable“:  $\hat{\phi}$

Nun können die Vorhersagen für das Outcome-Modell  $\hat{Q}(A = 1, L)$  und  $\hat{Q}(A = 0, L)$  angepasst werden:

$$\begin{aligned} \hat{Q}_i^*(A_i = 1, L_i) &= \text{expit}(\text{logit}(\hat{Q}_i(A_i = 1, L_i)) + \hat{\phi} \hat{H}_i(A_i = 1, L_i)) \\ \hat{Q}_i^*(A_i = 0, L_i) &= \text{expit}(\text{logit}(\hat{Q}_i(A_i = 0, L_i)) + \hat{\phi} \hat{H}_i(A_i = 0, L_i)) \end{aligned}$$

4. Zuletzt wird der ATE über die angepassten Vorhersagen geschätzt:

$$\hat{\psi}_{ATE}^* = \frac{1}{n} \sum_{i=1}^n \hat{Q}_i^*(A_i = 1, L_i) - \hat{Q}_i^*(A_i = 0, L_i)$$

Wurde zu Beginn eine Transformation von  $Y$  durchgeführt, muss  $\hat{\psi}_{ATE}^*$  wieder entsprechend zurücktransformiert werden:  $\hat{\psi}_{ATE} = \hat{\psi}_{ATE}^* \times (max - min)$ , ansonsten wird  $\hat{\psi}_{ATE} = \hat{\psi}_{ATE}^*$  gesetzt.

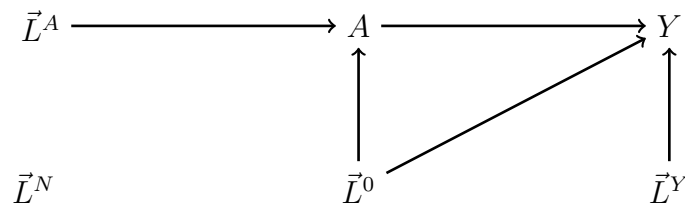
Durch die Verwendung des Outcome-Modells  $E[Y|A, L]$  und des Exposure-Modells  $P[A = 1|L]$  bekommt das TMLE-Verfahren eine zentrale Eigenschaft: Es ist doppelt robust. Dies bedeutet, dass der Schätzer für den kausalen Effekt konsistent ist, wenn eines der beiden Modelle fehlspezifiziert ist, das andere jedoch korrekt ist. Ist das Outcome-Modell bereits richtig spezifiziert, so ist kein Update der Vorhersagen notwendig und  $\phi \approx 0$ . Falls es fehlspezifiziert ist, das Exposure-Modell jedoch richtig, so erfolgt eine Korrektur der Vorhersagen. Konsistenz ist nur dann nicht gegeben, wenn beide Modelle fehlspezifiziert sind.

(Vgl. Schuler; 2016, S. 65 - 68) (Vgl. Schomacker; 2017, Kapitel 9)

## 3.3 Schätzung mit Variablenselektion

### 3.3.1 Warum Variablenselektion?

Bei allen bisher dargestellten Verfahren werden alle Variablen  $L$  benutzt, um einem möglichen Confounding-Problem entgegen zu wirken. Wie jedoch bereits in Kapitel 2 erläutert, sollte nicht jede beliebige Variable herangezogen werden. Zum einen können Collider-Variablen einen Bias erzeugen, wenn auf diese bedingt wird (Diese Problematik wird in der Arbeit nicht genauer untersucht). Zum anderen können überflüssige Variablen die Varianz der Schätzung erhöhen. Man betrachte folgende kausale Beziehungen:



**Abbildung 3.1** : DAG - Confounder und weitere Variablen

Dabei sind  $L^0$ ,  $L^A$ ,  $L^Y$  und  $L^N$  jeweils vektoriell zu sehen. D.h. es handelt sich jeweils um Variablenvektoren, deren Komponenten alle die entsprechenden kausalen Beziehungen erfüllen.

- $L^0$  enthält die wahren Confounder
- $L^A$  enthält Variablen, die einen kausalen Effekt auf  $A$ , aber nicht auf  $Y$  haben
- $L^Y$  enthält Variablen, die einen kausalen Effekt auf  $Y$ , aber nicht auf  $A$  haben
- $L^N$  enthält Variablen, die weder auf  $A$  noch auf  $Y$  einen kausalen Effekt haben

In dieser Situation wäre es wünschenswert, die Gruppenzugehörigkeit jeder Variable aus  $L$  zu kennen. Liegen keine Vorinformationen vor, die Aufschluss darüber geben, welche Variablen aus  $L$  zu welcher Gruppe gehören (was per Annahme der Fall ist), so ist unter Umständen ein Schätzverfahren sinnvoll, das Variablenselektion integriert. Variablenselektion „per Hand“ oder im Vorfeld sollte aufgrund des multiplen Testproblems vermieden werden.

Das dargestellte kausale Wirkungsmuster wird auch bei den in Kapitel 4 durchgeführten Simulationen herangezogen.

### 3.3.2 CTMLE-Variablenselektionsalgorithmus

Die dargestellten Überlegungen motivieren den iterativen CTMLE-Variablenselektionsalgorithmus:

Bei diesem gibt  $r$  den Index für die aktuelle Iteration an. Die Vorhersagen des Outcome-Modells  $\hat{Q}_{(r)}$  bzw.  $\hat{Q}_{(r)}^*$  und des Exposure-Modells  $\hat{g}_{(r)}$  werden bei diesem Verfahren in jeder Iteration  $r$  neu berechnet. Dies geschieht über die Aufnahme zusätzlicher Variablen ins Exposure-Modell. Für jede Beobachtung erfolgt die Bestimmung der Vorhersagen jeweils unter  $A_i = 1$  und unter  $A_i = 0$ .  $Y$  sollte auch hier wieder auf den Bereich  $[0, 1]$  transformiert werden.

Wie beim TMLE-Verfahren wird zunächst das Outcome-Modell  $E[Y|A, L]$  geschätzt. Dadurch erhält man wie gewohnt initiale Vorhersagen  $\hat{Q}_{(0)}$ . Für das Exposure-Modell  $P[A = 1|L]$  (von hier an mit  $G$  bezeichnet) wird zunächst ein reines Intercept-Modell  $G_{(0)}$  ohne Kovariablen verwendet, um die Vorhersagen  $\hat{g}_{(0)}$  zu bestimmen.

Für  $r = 0, \dots, p - 1$  (D.h. eine Iteration pro Variable in  $L$ ) wird Folgendes wiederholt:

1. Zunächst wird mit Hilfe von  $\hat{Q}_{(r)}$  und  $\hat{g}_{(r)}$  ein Update-Schritt durchgeführt, um dadurch die angepassten Vorhersagen  $\hat{Q}_{(r)}^*$  zu erhalten (genau wie beim TMLE-Verfahren Schritt 3). Diese werden verwendet, um einen Verlust  $Ls_{(r)}$  nach der negativen Log-Likelihood für bernoulliverteilte Zufallsvariablen zu bestimmen. Dafür werden die Vorhersagen unter der beobachteten Treatmentvariable herangezogen:

$$Ls(Y, \hat{Q}^*) = - \sum_{i=1}^n [Y_i \log(\hat{Q}_i^*(A_i, L_i)) + (1 - Y_i) \log(1 - \hat{Q}_i^*(A_i, L_i))]$$

2. Für jede Variable aus  $L$ , die noch nicht im aktuellen Exposure-Modell  $G_{(r)}$  enthalten ist, wird ein Exposure-Modell geschätzt, das zusätzlich zu den Kovariablen von  $G_{(r)}$  noch die jeweilige Variable als Kovariable enthält. Dadurch entstehen  $p - r$  neue Modelle.
3. Für jedes dieser neuen Modelle bildet man die Vorhersagen  $\hat{g}$  und führt mit diesen und  $\hat{Q}_{(r)}$  einen Update-Schritt durch. Dadurch erhält man korrigierte Vorhersagen  $\hat{Q}^*$  für jedes Modell. Mit diesen wiederum wird pro Modell ein Verlust gebildet, wofür die Verlustfunktion aus Schritt 1 verwendet wird. Das Exposure-Modell, das zum geringsten Verlust  $\tilde{L}s_{(r)}$  führt, wird mit  $\tilde{G}_{(r)}$  bezeichnet.
4. Folgendes ist zu überprüfen:
  - Ist  $\tilde{L}s_{(r)} < Ls_{(r)}$ , bedeutet das, dass die Vorhersagen durch die „optimale aufgenommene Variable“ im Exposure-Modell verbessert wurden. Man setzt  $G_{(r+1)} = \tilde{G}_{(r)}$  als Exposure-Modell für die nächste Iteration. Daraus erge-

ben sich neue Vorhersagen  $\hat{g}_{(r+1)}$ . Die nicht angepassten Vorhersagen des Outcome-Modells bleiben unverändert:  $\hat{Q}_{(r+1)} = \hat{Q}_{(r)}$

- Falls  $\tilde{L}S_{(r)} \geq LS_{(r)}$ , werden die Vorhersagen durch eine Erweiterung des Exposure-Modells nicht verbessert. Es wird nun ein „tatsächliches Update“ durchgeführt. D.h. man setzt:  $\hat{Q}_{(r)} = \hat{Q}_{(r)}^*$  und wiederholt die Schritte 3 und 4 dieser Iteration.

Zum Schluss ist noch ein letzter Update-Schritt notwendig, um  $\hat{Q}_{(p)}^*$  zu erhalten. Die Prozedur erzeugt eine Sequenz  $(\hat{Q}_{(r)}^*, \hat{g}_{(r)}; r = 0, \dots, p)$  von Vorhersagen. In jeder Iteration wird dem Exposure-Modell eine Variable aus  $L$  als Kovariable hinzugefügt: Jeweils diejenige, welche für die größte Verbesserung bezüglich der Verlustfunktion sorgt. Das letzte Exposure-Modell enthält somit alle Variablen  $L$  als Kovariablen. Die Aufgabe ist es nun, den optimalen Iterationsschritt  $r_{opt}$  zu bestimmen. Dies geschieht über Kreuzvalidierung, um Overfitting zu vermeiden. Der ATE bestimmt sich dann schließlich über:

$$\hat{\psi}_{ATE} = \frac{1}{n} \sum_{i=1}^n \hat{Q}_{i(r_{opt})}^*(A_i = 1, L_i) - \hat{Q}_{i(r_{opt})}^*(A_i = 0, L_i)$$

Eventuell ist dann wieder eine Rücktransformation (wie beim TMLE-Verfahren) notwendig.

(Vgl. Schnitzer; 2017, S. 4/5)

### 3.3.3 Einschub: Kreuzvalidierung

Würde man den Algorithmus alleine nach den Verlusten der Daten, auf denen der Algorithmus durchgeführt wird, bewerten, käme man zu dem Ergebnis, dass alle Variablen berücksichtigt werden sollten und  $r_{opt} = p$  ist. Jedoch sollte Überanpassung vermieden werden, da diese die Varianz der Schätzung erhöht. Somit ist eine Aufteilung in Trainingsdaten, auf denen der Algorithmus durchgeführt wird, und in Testdaten, auf denen evaluiert wird, sinnvoll. Kreuzvalidierung ist eine bewährte Methode, um einen Bias-Varianz Trade-Off zu erreichen:

- Der Datensatz wird dazu in  $M$  gleich große Teile geteilt.
- Für  $m = 1, \dots, M$  wird Folgendes durchgeführt:
  - Der  $m$ -te Teil des geteilten Datensatzes  $Val(m)$  wird als Testdatensatz benutzt. Dieser enthält  $1/M \times n$  Beobachtungen. Die restlichen  $\frac{M-1}{M} \times n$  Beobachtungen werden als Trainingsdaten verwendet.
  - Die Trainingsdaten werden benutzt, um den Algorithmus auszuführen.

- Durch das zu Beginn geschätzte Outcome-Modell (auf den Trainingsdaten) werden die initialen Vorhersagen  $\hat{Q}_0$  für die Testdaten bestimmt. In jeder Iteration  $r$  werden auf Basis der aktuellen (auf den Trainingsdaten geschätzten) Koeffizienten korrigierte Vorhersagen  $\hat{Q}^*$  für die Testdaten berechnet. Mit diesen wird nach der bereits verwendeten Funktion ein Verlust berechnet. Wurde bei den Trainingsdaten in Schritt 4 einer Iteration ein tatsächliches Update durchgeführt, ist zu beachten, dass die Vorhersagen der Testdaten ebenfalls angepasst werden müssen.

Insgesamt ergibt sich der Verlust für eine Iteration  $r$  zu:

$$cvError_r = - \sum_{m=1}^M \sum_{i \in Val(m)} [Y_i \log(\hat{Q}_{i(r)}^*(A_i, L_i)) + (1 - Y_i) \log(1 - \hat{Q}_{i(r)}^*(A_i, L_i))]$$

Pro Iteration  $r$  wird jede Beobachtung genau einmal evaluiert.  $r_{opt}$  ist die Iteration, zu welchem der korrespondierende Verlust ( $cvError_{r_{opt}}$ ) am geringsten ist.

(Vgl. Schnitzer; 2017, S. 6 / 7) (Vgl. Hastie; 2009, S. 241 - 255)

## 3.4 Schätzung mit Shrinkage

### 3.4.1 Einschub: Shrinkage-Schätzung

Shrinkage-Verfahren kommen insbesondere bei hochdimensionalen Daten und/oder Multikollinearitätsproblemen zum Einsatz. Sie penalisieren betragsmäßig hohe Schätzer für den Koeffizientenvektor und schrumpfen diese gegen bzw. komponentenweise auf 0. Dies erzeugt einen Bias bei der Schätzung, da der Log-Likelihood ein Strafterm hinzugefügt wird. Jedoch kann dadurch die Varianz der Schätzung unter Umständen verringert werden.

Bei der Schätzung eines kausalen Effekts können die Koeffizienten auch jeweils über Shrinkage-Verfahren geschätzt werden. Wie auch bei der Variablenselektion liegt der Fokus hier beim Exposure-Modell. Man verwendet den folgenden Schätzer für den Koeffizientenvektor  $\tau^*$  des Exposure-Modells ( $\tau^*$  mit Länge  $p+1$  stellt den Koeffizientenvektor mit Intercept  $\tau_0$  dar,  $\tau$  den Koeffizientenvektor ohne Intercept).

$$\hat{\tau}^*(\lambda) = \operatorname{argmax}_{\tau^*} \left[ \sum_{i=1}^n \{A_i \log(g_i(\tau^*)) + (1 - A_i) \log(1 - g_i(\tau^*))\} - \lambda \operatorname{pen}(\tau) \right]$$

Dies entspricht der penalisierten Log-Likelihood für bernoulliverteilte Zufallsvariablen.  $\operatorname{pen}(\tau)$  stellt den Penalisierungsterm dar. Mit  $\lambda$  wird die Stärke der Penalisierung kontrolliert. Es gilt, dass  $\lambda \geq 0$  zu wählen ist, wobei  $\lambda = 0$  bedeutet, dass keine Penalisierung stattfindet und die Shrinkage-Schätzung der gewöhnlichen Maximum-Likelihood-Schätzung entspricht. Eine Herausforderung ist es somit,  $\lambda$  geeignet zu wählen bzw. Verfahren zu konstruieren, die ein geeignetes  $\lambda$  bestimmen. Auf eine Penalisierung des Intercepts wird in der Regel verzichtet.

Alternativ kann die Shrinkage-Schätzung des Exposure-Modells auch folgendermaßen dargestellt werden:

$$\hat{\tau}(\lambda) = \operatorname{argmax}_{\tau^*} \left[ \sum_{i=1}^n \{A_i \log(g_i(\tau^*)) + (1 - A_i) \log(1 - g_i(\tau^*))\} \right]$$

unter der Bedingung:  $\operatorname{pen}(\tau) \leq c$

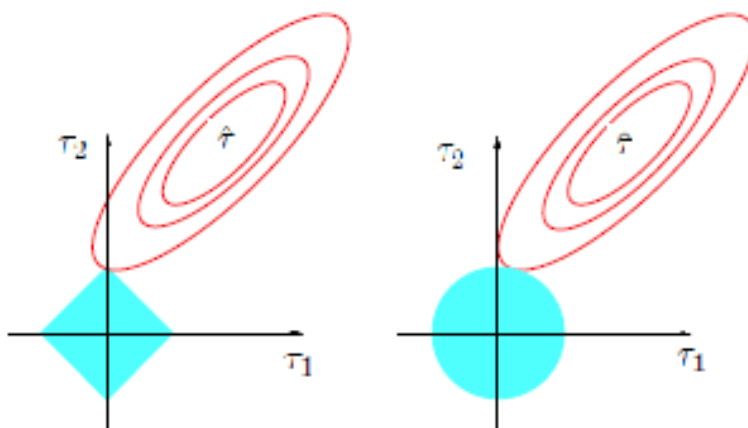
Es wird also eine ML-Schätzung unter einer Restriktion durchgeführt, welche die Größe der Komponenten von  $\hat{\tau}$  beschränkt. Der Parameter  $c$  ergibt sich dabei eindeutig aus  $\lambda$ .

Von Bedeutung ist die Wahl des Penalisierungsterms. Besonders folgende Terme werden häufig benutzt:

- Ridge-Penalisierung:  $\operatorname{pen}(\tau) = \sum_{r=1}^p \frac{1}{2} \tau_r^2$

- Lasso-Penalisierung:  $\text{pen}(\tau) = \sum_{r=1}^p |\tau_r|$
- Elastic-Net-Penalisierung:  $\text{pen}(\tau) = \sum_{r=1}^p (1 - \alpha) \frac{1}{2} \tau_r^2 + \alpha |\tau_r|$

Bei der Ridge-Penalisierung werden die Schätzer gegen 0 geschrumpft, bei der Lasso-Penalisierung werden manche Komponenten sogar auf 0 geschrumpft. Das Elastic-Net kombiniert die beiden Penalisierungsterme, wobei die Gewichte der Terme durch den Parameter  $\alpha \in [0, 1]$  kontrolliert werden.  $\alpha = 0$  entspricht der Ridge-Penalisierung und  $\alpha = 1$  der Lasso-Penalisierung.



**Abbildung 3.2 :** Lasso-Penalisierung (links) und Ridge-Penalisierung (rechts)

In der Grafik ist abgebildet, wie der Schätzer  $\hat{\tau}$  gegen 0 geschrumpft wird. Die blaue Raute bzw. der Kreis stellen die Beschränkung des Schätzers dar. Der Maximum-Likelihood-Schätzer wird solange gegen den 0-Vektor geschrumpft (dargestellt durch die roten Ellipsen), bis die Begrenzung eingehalten wird. Sobald Ellipse und Raute bzw. Kreis zusammentreffen, ist der penalisierte ML-Schätzer gefunden. Aus der Grafik wird auch ersichtlich, warum bei der Ridge-Penalisierung die Komponenten des Schätzers nur gegen 0 geschrumpft werden, während sie bei der Lasso-Penalisierung komponentenweise auf 0 geschrumpft werden. Dies liegt an der Form (Raute), welche durch die Betragsfunktion zustande kommt.

Shrinkage kann somit als Alternative zur klassischen Variablenselektion gesehen werden. Während die Variablen bei letzterem Verfahren entweder aufgenommen oder nicht aufgenommen werden, werden die zugehörigen Koeffizienten beim Shrinkage gegen bzw. komponentenweise auf 0 geschrumpft. Dies richtet sich nach Stärke der Penalisierung und Wahl des Penalisierungsterms.

(Vgl. Hastie; 2009, S. 61 - 79)



### 3.4.2 CTMLE-Algorithmus mit Shrinkage-Schätzung

Der CTMLE-Algorithmus mit Shrinkage-Schätzung konstruiert eine mehrstufige Shrinkage-Schätzung des Exposure-Modells. Man beginnt mit einer hohen Penalisierung und verringert diese schrittweise. Dies kann durch eine absteigende Sequenz von  $\lambda$ -Werten bewerkstelligt werden:  $\{\lambda_k : k = 0, \dots, K\}$ , wobei  $\lambda_K = 0$  (keine Penalisierung) festgelegt wird.  $\hat{g}^\lambda$  stellt die Vorhersagen des Exposure-Modells dar, wenn eine Shrinkage-Schätzung des Modells mit Parameter  $\lambda$  durchgeführt wird (mit allen Variablen  $L$  als Kovariablen). Zunächst werden wie beim CTMLE-Variablenselektionsalgorithmus initiale Vorhersagen  $\hat{Q}_{(0)}$  und ein auf  $[0, 1]$  transformierter Outcome benötigt. Man wiederholt die folgenden Schritte für  $k = 0, \dots, K - 1$ :

1. Wie gewohnt wird mit  $\hat{Q}_{(k)}$  und  $\hat{g}^{\lambda_k}$  ein TMLE-Update durchgeführt, um die angepassten Vorhersagen  $\hat{Q}_{(k)}^*$  zu erhalten. Mit diesen wird nach der negativen Log-Likelihood für bernoulliverteilte Zufallsvariablen ein Verlust  $L_{S(k)}$  berechnet.
2. Dasselbe wird für  $\hat{Q}_{(k)}$  und  $\hat{g}^{\lambda_{k+1}}$  wiederholt. Dadurch erhält man einen Verlust  $\tilde{L}_{S(k)}$ .
3. Anschließend werden die beiden Verluste verglichen:
  - Ist  $\tilde{L}_{S(k)} < L_{S(k)}$ , so bleiben die Vorhersagen des Outcome-Modells unverändert:  $\hat{Q}_{(k+1)} = \hat{Q}_{(k)}$
  - Ist  $\tilde{L}_{S(k)} \geq L_{S(k)}$ , so wird ein tatsächliches Update durchgeführt und  $\hat{Q}_{(k+1)} = \hat{Q}_{(k)}^*$  gesetzt.

Zum Schluss ist ein letzter Update-Schritt notwendig, um  $\hat{Q}_{(K)}^*$  zu erhalten. Dadurch erhält man eine Sequenz von angepassten Vorhersagen:  $(\hat{Q}_k^*; k = 1, \dots, K)$ . Auch hier wird per Kreuzvalidierung (analog zum CTMLE-Variablenselektionsalgorithmus) die optimale Iteration  $k_{opt}$  ermittelt. Der Average Treatment Effect kann dann folgendermaßen geschätzt werden (ggf. wieder mit Rücktransformation):

$$\hat{\psi}_{ATE} = \frac{1}{n} \sum_{i=1}^n \hat{Q}_{i(k_{opt})}^*(A_i = 1, L_i) - \hat{Q}_{i(k_{opt})}^*(A_i = 0, L_i)$$

(Vgl. Schnitzer; 2017, S. 5/6)

### 3.4.3 CTMLE-Algorithmus mit komponentenweisem Shrinkage

Die Shrinkage-Schätzung kann noch flexibler gestaltet werden. Beim Shrinkage-Algorithmus aus dem vorherigen Abschnitt werden alle Kovariablen gleich stark penalisiert. Es gibt jedoch auch die Möglichkeit, kovariablenspezifische Werte  $\lambda_r$  zu benutzen und die Regressoren dadurch unterschiedlich stark zu penalisieren. Im Fall des Exposure-Modells kann der Schätzer dann folgendermaßen definiert werden:

$$\hat{\tau}^*(\lambda) = \operatorname{argmax}_{\tau^*} \left[ \sum_{i=1}^n \{A_i \log(g_i(\tau^*)) + (1 - A_i) \log(1 - g_i(\tau^*))\} - \sum_{r=1}^p \lambda_r \operatorname{pen}(\tau_r) \right]$$

Die üblichen Penalisierungsterme können wie folgt dargestellt werden:

- Ridge-Penalisierung:  $\operatorname{pen}(\tau_r) = \frac{1}{2} \tau_r^2$
- Lasso-Penalisierung:  $\operatorname{pen}(\tau_r) = |\tau_r|$
- Elastic-Net-Penalisierung:  $\operatorname{pen}(\tau_r) = (1 - \alpha) \frac{1}{2} \tau_r^2 + \alpha |\tau_r|$

Diese entsprechen den Termen aus den vorherigen Abschnitten - mit dem Unterschied, dass die Summe nicht mehr auftaucht und der Strafterm kovariablenspezifisch ist.

Der CTMLE-Algorithmus mit komponentenweisem Shrinkage reduziert die Penalisierung schrittweise, wählt dazu aber in jeder Iteration eine Kovariable aus, deren  $\lambda$ -Wert verringert werden soll. Sei  $\Lambda = (\lambda_{k,r}; k = 1, \dots, K; r = 1, \dots, p)$  eine Matrix von Penalisierungsparametern. Jede Spalte enthält eine Sequenz von absteigenden  $\lambda$ -Werten für eine bestimmte Kovariable. Des Weiteren gilt:

- $t$  sei der Index für die Iterationen des Algorithmus.
- $\vec{\lambda}_{(t)}$  gibt den Penalisierungsvektor der Länge  $p$  in der Iteration  $t$  an.
- $k_{(t)}^r$  entspricht dem Index des Penalisierungsparameters für den Regressor  $r$  in Iteration  $t$ .
- Somit stellt  $\lambda_{k_{(t)}^r, r}$  den Penalisierungsparameter für die  $r$ -te Kovariable in der  $t$ -ten Iteration dar.

Zu Beginn ist  $k$  für jede Variable gleich 1:  $k_{(1)}^r = 1$ . Wie bei den anderen Algorithmen werden zunächst mit Hilfe eines auf  $[0, 1]$  transformierten Outcomes  $Y$  und eines geeigneten Outcome-Modells Vorhersagen bestimmt:  $\hat{Q}_{(1)}$ . Die initialen Vorhersagen  $\hat{g}_{(1)}$  des Exposure-Modells erhält man durch eine Shrinkage-Schätzung mit dem Startvektor für die Penalisierung:  $\vec{\lambda}_{(1)} = \{\lambda_{k_{(1)}^r, r}; r = 1, \dots, R\}$ . Anschließend werden für  $t = 1, \dots, p \times (K - 1)$  die folgenden Schritte ausgeführt:

1. Mit Hilfe der Vorhersagen  $\hat{g}_{(t)}$  und  $\hat{Q}_{(t)}$  führt man den TMLE-Update Schritt durch und erhält dadurch  $\hat{Q}_{(t)}^*$ . Der Verlust  $Ls_{(t)}$  errechnet sich dann wieder mit der negativen Log-Likelihood für bernoulliverteilte Zufallsvariablen.
2. Für jede Variable  $r = 1, \dots, p$ , für die  $k_{(t)}^r < K$  gilt (Das heißt: Für jede Variable, bei welcher man noch nicht bei der minimalen Penalisierung angelangt ist), wird ein Exposure-Modell per Shrinkage geschätzt. Dafür wird der Penalisierungsvektor  $\tilde{\lambda}_{(t)}^r = (\lambda_{k_{(t)}^1, 1}, \dots, \lambda_{k_{(t)}^{r-1}, r-1}, \lambda_{k_{(t)}^r + 1, r}, \lambda_{k_{(t)}^{r+1}, r+1}, \dots, \lambda_{k_{(t)}^p, p})$  verwendet. Dies entspricht einem Penalisierungsvektor, der sich von  $\vec{\lambda}_{(t)}$  nur dadurch unterscheidet, dass die Variable  $r$  mit dem nächst kleineren  $\lambda$ -Wert penalisiert wird.
3. Für jedes dieser neuen Exposure-Modelle werden Vorhersagen  $\hat{g}$  gebildet. Diese und  $\hat{Q}_{(t)}$  werden verwendet, um jeweils angepasste Vorhersagen  $\hat{Q}^*$  zu bekommen. Mit diesen wird nach der gewohnten Verlustfunktion pro Modell/pro Variable ein Verlust gebildet, wofür die Verlustfunktion aus Schritt 1 verwendet wird. Die Variable, die nach diesem Prozess zum kleinsten Verlust führt, sei mit  $r^*$  bezeichnet. Der zugehörige Verlust sei  $\tilde{L}s_{(t)}^{r^*}$ .
4. Es wird überprüft:
  - Ist  $\tilde{L}s_{(t)}^{r^*} < Ls_{(t)}$ , so wurde durch eine weniger starke Penalisierung der Variable  $r^*$  eine Verbesserung der Vorhersagen bezüglich der Verlustfunktion erzielt. Man setzt  $\vec{\lambda}_{(t+1)} = \tilde{\lambda}_{(t)}^{r^*}$  and erhält dadurch auch neue Vorhersagen des Exposure-Modells:  $\hat{g}_{(t+1)}$ . Die Vorhersagen des Outcome-Modells bleiben unverändert:  $\hat{Q}_{(t+1)} = \hat{Q}_{(t)}$ .
  - Ansonsten wird ein „tatsächliches Update“ durchgeführt und  $\hat{Q}_{(t)} = \hat{Q}_{(t)}^*$  gesetzt. Mit diesen werden die Schritte 3 und 4 des Algorithmus wiederholt.

Auch hier ist noch ein letzter Update-Schritt notwendig, um  $\hat{Q}_{(p \times (K-1))}^*$  zu erhalten. Um die optimale Iteration  $t_{opt}$  zu bestimmen, wird wieder auf Kreuzvalidierung zurückgegriffen (Funktionsweise analog zu den anderen Algorithmen). Anschließend kann der ATE geschätzt werden (ggf. wieder mit Rücktransformation):

$$\hat{\psi}_{ATE} = \frac{1}{n} \sum_{i=1}^n \hat{Q}_{i(t_{opt})}^*(A_i = 1, L_i) - \hat{Q}_{i(t_{opt})}^*(A_i = 0, L_i)$$

Wählt man für diesen Algorithmus  $K = 2$  und  $\lambda_{1,r} = \infty$  für alle  $r$  sowie  $\lambda_{2,r} = 0$  für alle  $r$ , so entspricht das Verfahren dem CTMLE-Variablenselektionsalgorithmus. Bei Letzterem kann entweder keine Penalisierung (Variable wird aufgenommen) oder maximale Penalisierung (Variable wird nicht aufgenommen) gewählt werden. Dadurch wird auch

die höhere Flexibilität des CTMLE-Algorithmus mit komponentenweisem Shrinkage ersichtlich.

(Vgl. Schnitzer; 2017, S. 6)

# 4 Simulationsstudie zur Untersuchung der Schätzverfahren

In diesem Kapitel werden die in Kapitel 3 dargestellten Schätzverfahren anhand ihrer Güte untersucht und verglichen. Als Kennzahl wird der Mean Squared Error (MSE) herangezogen. Da eine analytische Ermittlung des MSE für die Verfahren nicht möglich ist, werden Simulationen durchgeführt. Mit deren Hilfe kann der MSE durch den empirischen MSE geschätzt werden.

## 4.1 Aufbau der Simulationen

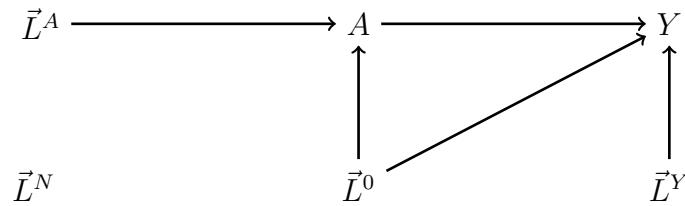
### 4.1.1 Allgemeine Struktur

Für eine Simulation legt man einen datengenerierenden Prozess inklusive des wahren Average Treatment Effects  $\psi_{ATE}$ , verschiedene Schätzverfahren und eine bestimmte Anzahl  $R$  an Wiederholungen fest. Der Prozess ist dann folgender:

- Für  $j = 1, \dots, R$ :
  - Die Daten werden nach dem datengenerierenden Prozess (Zufallsprozess) gebildet
  - Für jede Schätzmethode wird der Average Treatment Effect aus den Daten geschätzt.
- Dadurch ergibt sich eine Matrix von Schätzungen. Die Zeilen repräsentieren die Wiederholungen und die Spalten die Schätzmethoden. Aus der Matrix und dem wahren ATE lässt sich der empirische MSE-Vektor (ein MSE pro Schätzmethode) bestimmen.

## 4.1.2 Generierung der Daten

Die Daten werden nach folgenden kausalen Beziehungen erzeugt:



**Abbildung 4.1** : DAG - Kausale Beziehungen für die Simulationen

### Generierung der möglichen Confounder $L$

Für den Variablenvektor  $L$  ( $r = 1, \dots, p$ ) werden pro Variable  $n$  Beobachtungen ( $i = 1, \dots, n$ ) aus einer variablenspezifischen Normalverteilung gezogen. Im einfachsten Fall wird Unabhängigkeit zwischen all diesen Variablen festgelegt. Dann können die Daten für jede Variable  $L_{.r}$  aus  $L$  folgendermaßen generiert werden:

$$L_{ir} \sim N(\mu_r, \nu_r^2)$$

Folgende Parameter müssen festgelegt werden:

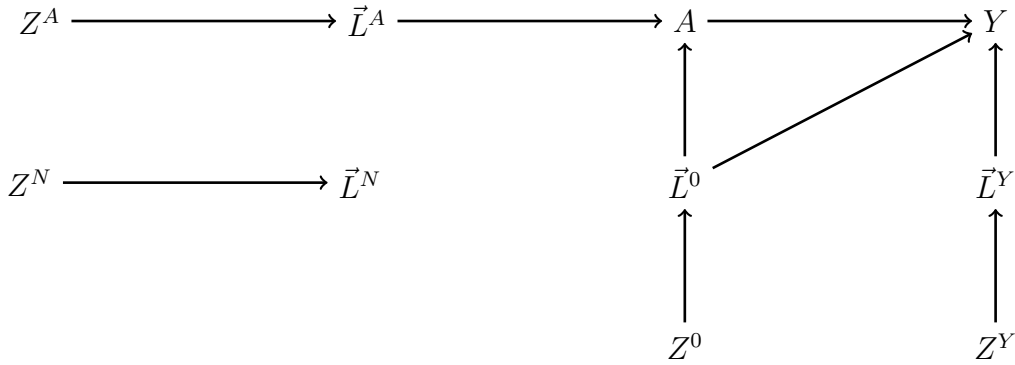
- $n$ : Anzahl der Beobachtungen
- $p$ : Anzahl der Variablen in  $L$
- $p_0, p_A, p_Y, p_N$ : Anzahl der Variablen in jeder der 4  $L$ -Gruppen
- $\mu = (\mu_1, \dots, \mu_p)$ : Interceptvektor (unter diesen Umständen: Erwartungswertvektor) für die Normalverteilung der Variablen  $L$
- $\nu = (\nu_1, \dots, \nu_p)$ : Standardabweichungsvektor für die Normalverteilung der Variablen  $L$

Optional kann ein Teil der erzeugten Variablen mit Hilfe von Cutpoints (Jeweils der Median) in binäre Größen umgewandelt werden:

$$\tilde{L}_{ir} = \begin{cases} 1, & \text{falls } L_{ir} > \text{Med}(L_{.r}) \\ 0, & \text{sonst} \end{cases}$$

Optional können auch Assoziationen zwischen den Variablen  $L$  simuliert werden. Dies geschieht über verborgene Hilfsvariablen  $Z$ . Sollen die Variablen innerhalb einer  $L$ -Gruppe

Zusammenhänge aufweisen, lässt sich dies durch folgendes kausale Wirkungsmuster realisieren:



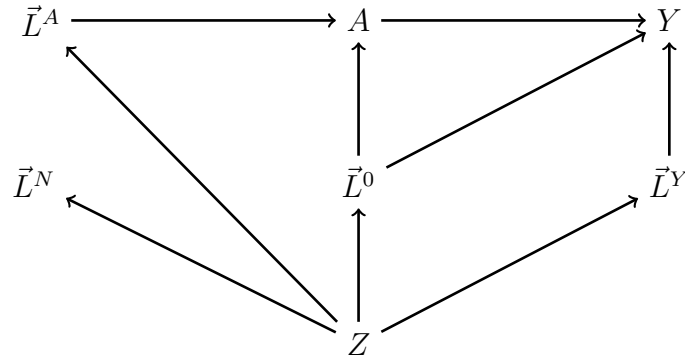
**Abbildung 4.2 :** DAG - Kausale Beziehungen für die Simulationen bei Assoziationen zwischen den Variablen  $L$  innerhalb einer L-Gruppe

Eine Komponente aus dem Variablenvektor  $Z = (Z^0, Z^A, Z^Y, Z^N)$  hat jeweils einen kausalen Effekt auf alle Variablen  $L$  der zugehörigen L-Gruppe. Dies führt dazu, dass die Variablen  $L$  einer Gruppe Assoziationen aufweisen, ohne in einer kausalen Beziehung zueinander zu stehen. Die 4 Variablen aus  $Z = (Z^0, Z^A, Z^Y, Z^N)$  werden alle aus derselben Normalverteilung (gleicher Erwartungswert und gleiche Varianz) gezogen. Anschließend können die Beobachtungen der Variablen  $L$  folgendermaßen erzeugt werden:

$$L_{ir} | Z_i \begin{cases} \sim N(\mu_r + Z_i^0 \kappa, \nu_r^2), & \text{falls } L_{.r} \text{ Variable aus } L^0 \\ \sim N(\mu_r + Z_i^A \kappa, \nu_r^2), & \text{falls } L_{.r} \text{ Variable aus } L^A \\ \sim N(\mu_r + Z_i^Y \kappa, \nu_r^2), & \text{falls } L_{.r} \text{ Variable aus } L^Y \\ \sim N(\mu_r + Z_i^N \kappa, \nu_r^2), & \text{falls } L_{.r} \text{ Variable aus } L^N \end{cases}$$

Der Einfachheit halber wird angenommen, dass der Koeffizient, der den Einfluss von  $Z$  widerspiegelt, für jede Variable aus  $L$  derselbe ist:  $\kappa$ . Die Erwartungswerte sind also jetzt beobachtungsspezifisch (nicht nur:  $\mu_r$ ).

Auch Assoziationen zwischen allen Variablen aus  $L$  lassen sich simulieren. Dafür wird lediglich eine Variable  $Z$  benötigt, welche einen Einfluss auf alle Variablen aus  $L$  hat:



**Abbildung 4.3 :** DAG - Kausale Beziehungen für die Simulationen bei Assoziationen zwischen allen Variablen  $L$

Die Normalverteilung, aus welcher die Beobachtungen der Variablen  $L$  erzeugt werden, ist hier beobachtungsspezifisch, jedoch nicht mehr gruppenspezifisch:

$$L_{ir}|Z_i \sim N(\mu_r + Z_i\kappa, \nu_r^2)$$

Zu beachten ist: Simuliert man die Variablen nach diesem Schema (Assoziationen zwischen allen Variablen aus  $L$ ), so entsteht ein neuer „Backdoor-Path“! Somit ist es in so einem Fall nicht mehr ausreichend, die Variablen  $L^0$  in die Schätzung zu integrieren. Zusätzlich müssen entweder die Variablen  $L^A$  oder  $L^Y$  berücksichtigt werden. (Oder  $Z$ , jedoch ist diese Variable per Annahme verborgen)

Möchte man Assoziationen zwischen den Variablen aus  $L$  erzeugen, so muss zusätzlich zu den oben erwähnten Parametern festgelegt werden:

- Erwartungswert und Varianz der Normalverteilung, aus der die Variablen  $Z$  gezogen werden
- Der Koeffizient  $\kappa$  für den Einfluss der Variablen  $Z$

Die Variablen  $Z$  sind in dem Sinne verborgen, dass sie in den Daten nicht auftauchen und für die Schätzung somit nicht zur Verfügung stehen.

### Generierung der Treatmentvariable $A$

Um die Treatmentvariable zu generieren, müssen zunächst die Variablen aus  $L$  bestimmt werden, die einen kausalen Effekt auf die Treatmentvariable haben. Es wird festgelegt:  $L^{A0} = [L^A|L^0]$ . Das heißt  $L^{A0}$  verknüpft die Vektoren  $L^0$  und  $L^A$  und enthält somit alle Variablen aus  $L$ , welche einen Effekt auf  $A$  haben. Die  $n$  Beobachtungen der



Treatmentvariable  $A$  können dann folgendermaßen generiert werden:

$$A_i | L_i^{A0} \sim B \left( 1, \frac{1}{1 + \exp(-(\tau_0 + L_i^{A0}\tau + (L_i^{A0})^2\nu))} \right)$$

$A$  ist also bernoulliverteilt, wobei die Wahrscheinlichkeit von den Variablen aus  $L^{A0}$  abhängt. Dazu werden einfache und quadratische Terme verwendet. Es muss festgelegt werden:

- $\tau_0$ : Intercept für die Treatmentwahrscheinlichkeit
- $\tau$ : Vektor der Länge  $p_A + p_0$  für die Koeffizienten der einfachen Terme  $L^{A0}$
- $\nu$ : Vektor der Länge  $p_A + p_0$  für die Koeffizienten der quadratischen Terme  $(L^{A0})^2$

### Generierung des Outcomes $Y$

Analog zur Generierung von  $A$  müssen auch hier die Variablen aus  $L$  festgelegt werden, die einen kausalen Effekt auf  $Y$  haben:  $L^{Y0} = [L^Y | L^0]$ . Der Outcome wird dann nach folgendem Schema gebildet:

$$Y_i = \theta_0 + A_i\theta_1 + L_i^{Y0}\beta + L_i^{Y0}A_i\gamma + (L_i^{Y0})^2\delta + \epsilon_i \quad \text{mit } \epsilon_i \sim N(0, \sigma^2)$$

Einfache Terme  $A$  und  $L^{Y0}$ , Interaktionsterme  $A : L^{Y0}$  und quadratische Terme  $(L^{Y0})^2$  werden zur Generierung von  $Y$  verwendet. Werden die Daten auf diese Weise gebildet, so gilt für den Treatment Effect einer Beobachtung  $i$ :

$$Y_i^{a=1} - Y_i^{a=0} = (\theta_0 + \theta_1 + L_i^{Y0}\beta + L_i^{Y0}\gamma + (L_i^{Y0})^2\delta + \epsilon_i) - (\theta_0 + L_i^{Y0}\beta + (L_i^{Y0})^2\delta + \epsilon_i) = \theta_1 + L_i^{Y0}\gamma$$

Somit gilt für den Average Treatment Effect:

$$\psi_{ATE} = E[Y^{a=1} - Y^{a=0}] = E[\theta_1 + L^{Y0}\gamma] = \theta_1 + E[L^{Y0}]\gamma$$

Der Erwartungswertvektor  $E[L^{Y0}]$  wiederum lässt sich anhand von Parametern berechnen - je nachdem, wie die Variablen  $L$  im ersten Schritt erzeugt wurden.

Folgende Parameter werden benötigt:

- $\theta_0$ : Intercept für den Outcome
- $\beta$ : Vektor der Länge  $p_Y + p_0$  für die Koeffizienten der einfachen Terme  $L^{Y0}$
- $\gamma$ : Vektor der Länge  $p_Y + p_0$  für die Koeffizienten der Interaktionsterme zwischen  $L^{Y0}$  und  $A$
- $\delta$ : Vektor der Länge  $p_Y + p_0$  für die Koeffizienten der quadratischen Terme  $(L^{Y0})^2$

- $\sigma$ : Standardabweichung für die Störterme
- $\psi_{ATE}$ : Der Average Treatment Effect. Ist dieser festgelegt, so ergibt sich der Parameter  $\theta_1$  auf eindeutige Weise aus den anderen Parametern.

### 4.1.3 Schätzung des kausalen Effekts

Es folgt die Schätzung des Average Treatment Effects. Alle Schätzmethoden verwenden dazu die erzeugten Daten. Es muss dabei nicht nur das Verfahren selbst, sondern auch gegebenenfalls Hyperparameter und Modelle festgelegt werden. Diese sind (im Wesentlichen) die folgenden:

- Naive Schätzung: Keine
- Standardisierung: Outcome-Modell
- IP-Weighting: Exposure-Modell
- TMLE: Outcome-Modell, Exposure-Modell
- CTMLE-Variablenselektionsalgorithmus: Outcome-Modell, (Volles) Exposure-Modell,  $M$
- CTMLE-Algorithmus mit Shrinkage-Schätzung: Outcome-Modell, Exposure-Modell,  $M$ ,  $\lambda$ -Vektor,  $\alpha$
- CTMLE-Algorithmus mit komponentenweisem Shrinkage: Outcome-Modell, Exposure-Modell,  $M$ ,  $\lambda$ -Vektor (pro Variable),  $\alpha$

### 4.1.4 MSE-Berechnung

Nachdem die Daten  $R$  mal erzeugt wurden und der ATE mit jedem Verfahren  $R$  mal geschätzt wurde, kann der MSE geschätzt werden. Der MSE für den zu schätzenden Parameter  $\psi_{ATE}$  und ein Schätzverfahren  $T$  ist folgendermaßen definiert:

$$MSE(\psi_{ATE}, T) = E[(T - \psi_{ATE})^2] = Var(T) + E[T - \psi_{ATE}]^2$$

Dies entspricht der mittleren quadratischen Abweichung des Schätzers vom zu schätzenden Parameter. Der MSE kann in Varianz des Schätzers und mittlere Abweichung des Schätzers vom Parameter (Bias) zerlegt werden. Da der MSE (für die meisten Verfahren) analytisch nicht zugänglich ist, muss er jeweils durch den empirischen MSE geschätzt werden, welcher sich aus den Simulationen ergibt.  $\hat{\psi}_{ATE}(T)_j$  ( $j = 1, \dots, R$ ) sei der unter

Schätzverfahren  $T$  geschätzte Parameter in Wiederholung  $j$ . Dann ist der empirische MSE für die Schätzungen eines Verfahrens  $T$  zur Schätzung von  $\psi_{ATE}$ :

$$MSE_{emp}(\psi_{ATE}, \hat{\psi}_{ATE}(T)) = \frac{1}{R} \sum_{j=1}^R (\hat{\psi}_{ATE}(T)_j - \psi_{ATE})^2$$

## 4.2 Durchführung der Simulationen

In diesem Abschnitt wird erläutert, mit welchen Inputs die Simulationen tatsächlich durchgeführt werden.

### 4.2.1 Betrachtete datengenerierende Prozesse

Zur Erzeugung der Daten werden folgende Inputs variiert:

- Die Anzahl der Beobachtungen. Gewählt werden die Stichprobenumfänge  $n = 200$  und  $n = 800$ .
- Die Assoziationsstruktur der Variablen  $L$ . Die folgenden drei werden betrachtet:
  - Keine Assoziationen
  - Assoziationen innerhalb einer  $L$ -Gruppe
  - Assoziationen zwischen allen Variablen aus  $L$

Diese Strukturen wurden bereits in Kapitel 4.1.2 genauer erläutert.

- Bildung von Outcome und Treatmentvariable. Vier verschiedene Szenarien werden herangezogen:
  1.  $Y$  und  $A$  werden jeweils nur durch einfache Terme gebildet. Das heißt  $\gamma$ ,  $\delta$  und  $v$  entsprechen jeweils dem 0-Vektor.
  2.  $Y$  wird durch einfache und Interaktionsterme gebildet. In diesem Szenario sind die Komponenten von  $\gamma$  von 0 verschieden.  $A$  wird nach wie vor nur durch einfache Terme gebildet.
  3.  $Y$  wird durch einfache und quadratische Terme gebildet. Somit enthält  $\delta$  keine 0-en mehr.  $A$  wird nach wie vor nur durch einfache Terme gebildet.
  4.  $Y$  wird durch einfache, quadratische und Interaktionsterme gebildet.  $A$  wird durch einfache und quadratische Terme gebildet. Da die Wirkungsstruktur hier am komplexesten ist, wird für diesen Fall zusätzlich eine Simulation mit Stichprobenumfang  $n = 3200$  herangezogen.

Die übrigen Inputs werden folgendermaßen gewählt:

- Um den MSE möglichst gut durch den empirischen MSE zu schätzen und dadurch ein aussagekräftiges Ergebnis zu erhalten, ist die Wahl einer hohen Anzahl an Wiederholungen sinnvoll. Je höher die Anzahl, desto höher ist allerdings auch die Laufzeit. Als „Kompromiss“ wird  $R = 500$  gewählt.

- Die Anzahl der Variablen  $L$  und die Anzahl jeweils in den 4  $L$ -Gruppen sollte im Verhältnis zum Stichprobenumfang nicht zu hoch sein, jedoch hoch genug, dass Variablenselektion und Shrinkage sinnvoll erscheint:

$$p = 10, p_0 = 4, p_Y = 2, p_A = 2, p_N = 2$$

- Der Anteil der Variablen  $L$ , die jeweils in eine binäre Größe umgewandelt werden: 0.2
- Der zu schätzende Average Treatment Effect:  $\psi_{ATE} = 1$
- Weitere Parameter:

- Gegebenenfalls Erwartungswert und Standardabweichung der Variablen  $Z$ : 0 und 1

- Interceptvektor von  $L$ :  $\mu = (1, 2, 3, 4, 2, 3, 2, 3, 2, 3)$

- Standardabweichungsvektor von  $L$ :  $\nu = (0.8, 1, 1.2, 1.4, 1, 1.2, 1, 1.2, 1, 1.2)$

- Gegebenenfalls Koeffizient für den Einfluss von  $Z$ :  $\kappa = 1$

- Intercept zur Generierung der Treatmentvariable:

$$\tau_0 = \begin{cases} -7.5, & \text{in den Szenarien 1, 2 und 3 der Treatment- und Outcomebildung,} \\ -14.5, & \text{in Szenario 4 der Treatment- und Outcomebildung} \end{cases}$$

- Koeffizienten für die einfachen Terme  $L^{A0}$  zur Generierung der Treatmentvariable:

$$\tau = (0.3, 0.4, 0.5, 0.6, 0.45, 0.55)$$

- Koeffizienten für die quadratischen Terme  $(L^{A0})^2$  zur Generierung der Treatmentvariable:

$$v = \begin{cases} (0, 0, 0, 0, 0, 0), & \text{in den Szenarien 1, 2 und 3,} \\ (0.5, 0.25, 0.15, 0.1, 0.18, 0.18), & \text{in Szenario 4} \end{cases}$$

- Intercept für die Generierung des Outcomes:  $\theta_0 = -15$

- Koeffizienten für die einfachen Terme zur Generierung des Outcomes:

$$\beta = (0.5, 1, 1.5, 2, 0.75, 1.25)$$

- Koeffizienten für die Interaktionsterme zur Generierung des Outcomes:

$$\gamma = \begin{cases} (0, 0, 0, 0, 0, 0), & \text{in den Szenarien 1 und 3,} \\ (0.6, 0.9, 1.3, 1.7, 0.75, 1.25), & \text{in Szenario 2,} \\ (0.9, 0.95, 0.85, 1.05, 1, 0.8), & \text{in Szenario 4} \end{cases}$$

- Koeffizienten für die quadratischen Terme zur Generierung des Outcomes:

$$\delta = \begin{cases} (0, 0, 0, 0, 0, 0), & \text{in den Szenarien 1 und 2,} \\ (0.5, 0.25, 0.15, 0.1, 0.18, 0.18), & \text{in den Szenarien 3 und 4} \end{cases}$$

- Standardabweichung der Störterme:  $\sigma = 1$
- Der Koeffizient  $\theta_1$  für die Variable  $A$  zur Generierung des Outcomes ergibt sich (wie bereits erläutert)

Ziel bei dieser Parameterwahl ist es, eine Confounding Situation in ausreichendem Maß zu simulieren. Um zu verhindern, dass sich Effekte gegenseitig „neutralisieren“, werden nur positive Werte verwendet. Diese werden durch die negativen Intercepts einigermaßen ausgeglichen. Das heißt: Die Prädiktoren sind im Mittel ca. 0. Dies ist insbesondere bei der Bildung der Treatmentvariable von Bedeutung. Des Weiteren sollten sich die Effekte unterscheiden, jedoch nicht zu stark. Ein gewisser Abstand von der 0 ist wichtig. Auch die Erwartungswerte der Variablen  $L$  sollten nicht zu nah an der 0 sein, damit man insbesondere bei den Interaktionstermen und quadratischen Termen nicht auf Spezialfälle trifft. Diese könnten die Terme sonst nahezu bedeutungslos machen.

## 4.2.2 Verwendete Schätzmethoden

Die folgenden Methoden werden zur Schätzung des ATE jeweils benutzt. Die später in den Grafiken verwendeten Abkürzungen sind jeweils in Klammern dahinter.

- Naive Schätzung (Naiv)
- IP-Weighting (IPW)
- Standardisierung (Std)
- Targeted Maximum Likelihood Estimation (TMLE)
- CTMLE-Variablenselektionsalgorithmus (VS)
- CTMLE-Algorithmus mit Shrinkage-Schätzung unter Ridge-Penalisierung (Rid)

- CTMLE-Algorithmus mit Shrinkage-Schätzung unter Elastic-Net-Penalisation mit  $\alpha = 0.5$  (EN)
- CTMLE-Algorithmus mit Shrinkage-Schätzung unter Lasso-Penalisation (Las)
- CTMLE-Algorithmus mit komponentenweisem Shrinkage unter Ridge-Penalisation (cRid)
- CTMLE-Algorithmus mit komponentenweisem Shrinkage unter Elastic-Net-Penalisation mit  $\alpha = 0.5$  (cEN)
- CTMLE-Algorithmus mit komponentenweisem Shrinkage unter Lasso-Penalisation (cLas)

Die folgenden Modellierungen des Outcome-Modells werden jeweils verwendet:

- Nur mit  $A$  als Kovariable
- Mit einfachen Termen  $A$  und  $L$

Je nach datengenerierendem Prozess noch zusätzlich:

- Mit einfachen Termen  $A$  und  $L$  sowie Interaktionstermen  $A : L$
- Mit einfachen Termen  $A$  und  $L$  sowie quadratischen Termen  $L^2$
- Mit einfachen Termen  $A$  und  $L$  sowie Interaktionstermen  $A : L$  und quadratischen Termen  $L^2$

Da beim IP-Weighting und bei der naiven Schätzung kein Outcome-Modell in dem Sinne spezifiziert wird, gibt es für diese Methoden keine verschiedenen „Varianten“ zur Modellierung des Outcomes. Die Methode Standardisierung, bei welcher man nur  $A$  als Kovariable modelliert, entspricht der naiven Schätzung. Zu beachten ist, dass für die Modellierung (ggf.) alle Variablen aus  $L$  verwendet werden, da per Annahme keine Information darüber vorliegt, welche Variablen aus  $L$  einen Effekt auf  $A$  oder  $Y$  haben. Dies gilt auch für die Interaktionsterme und quadratischen Terme.

Des Weiteren wird gewählt:

- Eine Modellierung des Exposure-Modells nur mit einfachen Termen  $L$  (bei allen Verfahren)
- $M = 10$  Teildatensätze für die Kreuzvalidierung. 10 ist eine klassische Zahl, die hierfür häufig verwendet wird.
- Die  $\lambda$ -Sequenz sollte auf einer Log-Skala gewählt werden. Hier wird auf die 10er-Log-Skala zurückgegriffen:  $\{100, 10, 1, 0.1, 0.01, 0.001, 0.0001, 0\}$

## 4.3 Ergebnisse der Simulationen

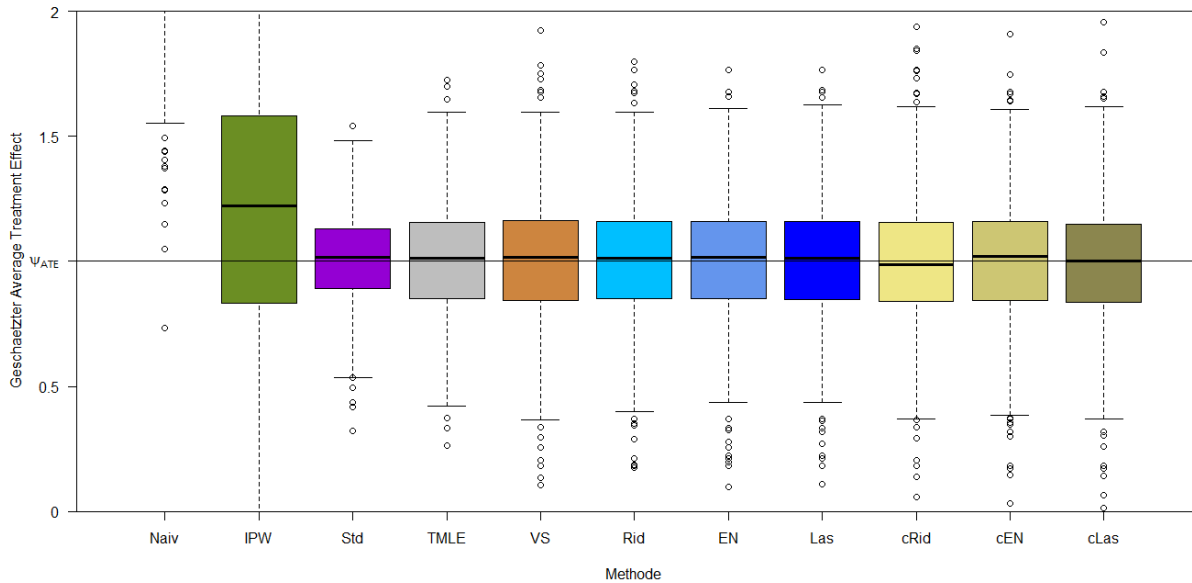
In diesem Abschnitt werden die wesentlichen Ergebnisse der Simulationen dargestellt. Vollständige Tabellen mit den Ergebnissen befinden sich im elektronischen Anhang der Arbeit.

Zu beachten ist, dass aufgrund der Anschaulichkeit teilweise unterschiedliche Skalen für den MSE verwendet werden. Zudem sind die Schätzungen der Verfahren, besonders bei der naiven Schätzung und dem IP-Weighting, nicht immer vollständig abgebildet. Die Extremwerte sind teilweise nicht alle sichtbar. Auch bei den MSE-Grafiken sind die Balken von der naiven Schätzung und dem IP-Weighting zumeist nicht ganz sichtbar. Ist in diesem Kapitel vom MSE die Rede, ist der empirische MSE gemeint. Dasselbe gilt für Bias und Varianz.

### 4.3.1 Erste Ergebnisse

Hier werden die Simulationen dargestellt, bei welchen der Outcome nur durch einfache Terme  $A$  und  $L^{Y^0}$  und die Treatmentvariable nur durch einfache Terme  $L^{A^0}$  gebildet wird und keine Assoziationen zwischen den Variablen aus  $L$  vorliegen. Es werden die Schätzungen betrachtet, bei welchen Outcome-Modell und Exposure-Modell richtig spezifiziert werden. Das heißt: Letzteres beinhaltet die Kovariablen  $L$  und Ersteres die Kovariablen  $A$  und  $L$ . Richtige Modellspezifikation bedeutet in diesem Fall, dass alle richtigen Terme im spezifizierten Modell enthalten sind, aber auch überflüssige Terme enthalten sein können (wobei sich keine Collider im Modell befinden). Man erhält die folgenden Schätzungen für den Stichprobenumfang  $n = 200$ :

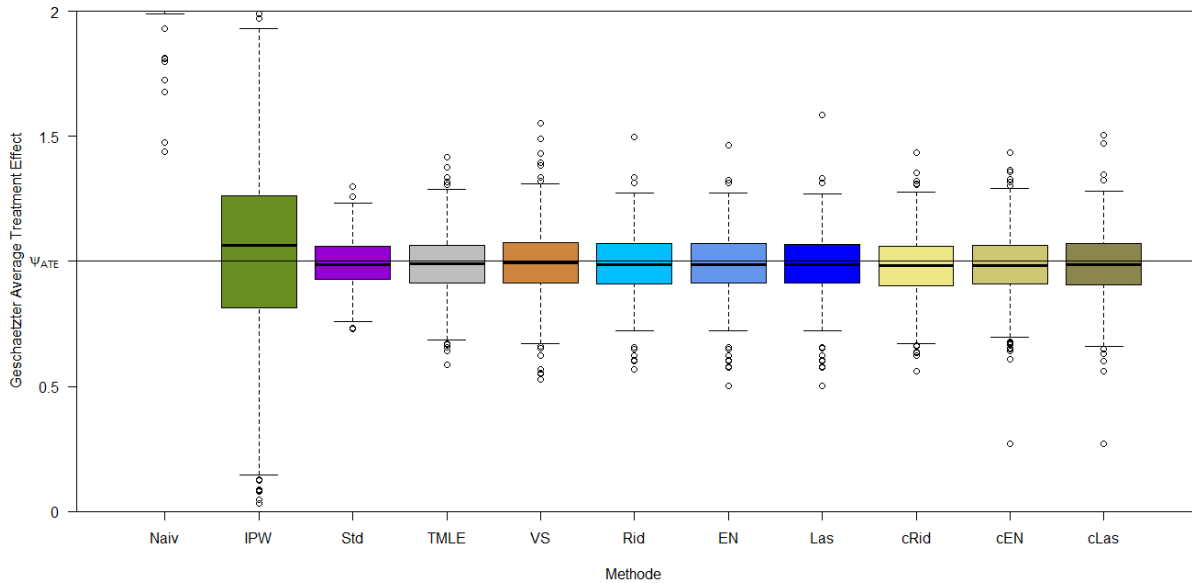




**Abbildung 4.4 :** Schätzungen des ATE für  $n = 200$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird und das Outcome-Modell mit den Kovariablen  $A$  und  $L$  spezifiziert wird

Es lässt sich feststellen, dass die naive Schätzung stark verzerrt ist. Dies ist auf das erläuterte Confounding-Problem zurückzuführen. Das IP-Weighting weist eine sehr hohe Varianz auf, während der ATE bei der Standardisierung sehr gut geschätzt wird: Ein Bias ist kaum vorhanden und die Varianz ist geringer als bei den anderen Methoden. TMLE und die erweiterten Verfahren zeigen alle ein ähnliches Bild: Die Verzerrung ist ebenfalls sehr gering, die Varianz etwas höher als bei der Standardisierung, jedoch deutlich geringer als beim IP-Weighting. Bei TMLE sind Spannweite und Varianz etwas geringer als bei den Variablenselektions- und Shrinkage-Verfahren. Eine Verbesserung der Güte durch diese Verfahren ist hier somit nicht ersichtlich.

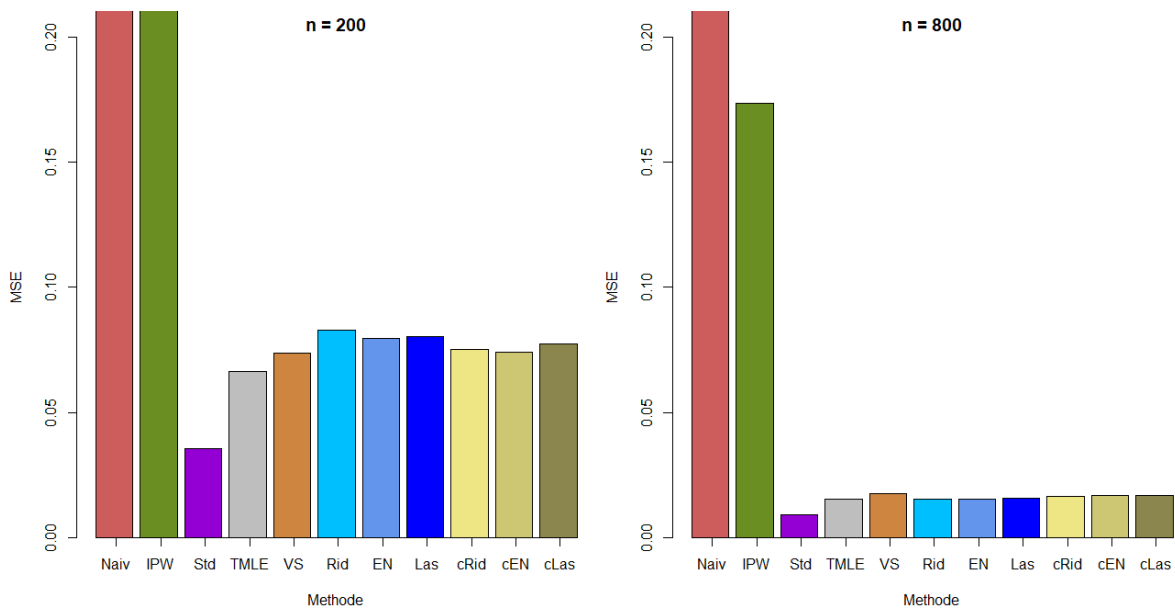
Als Nächstes wird dasselbe für den Stichprobenumfang  $n = 800$  dargestellt.



**Abbildung 4.5 :** Schätzungen des ATE für  $n = 800$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird und das Outcome-Modell mit den Kovariablen  $A$  und  $L$  spezifiziert wird

Die Schätzungen weisen hier ein sehr ähnliches Muster auf. Die Varianz wird im Vergleich zur Simulation mit  $n = 200$  deutlich geringer. Dies ist sowohl an der Spannweite als auch am Interquartilsabstand ersichtlich.

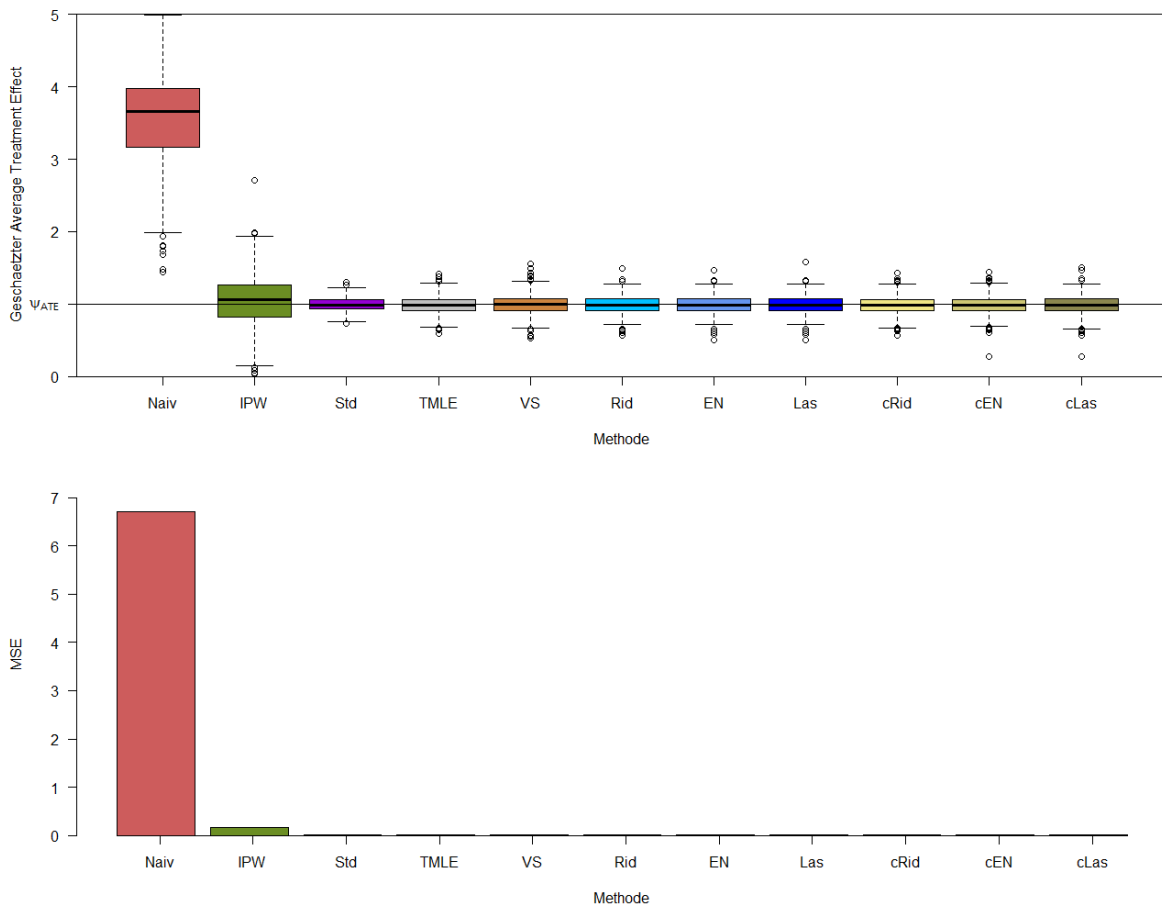
Man betrachte nun die MSE zu den Verfahren für die zwei Simulationen:



**Abbildung 4.6 :** MSE-Vergleich der Schätzungen des ATE für die beiden Stichprobenumfänge  $n = 200$  und  $n = 800$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird und das Outcome-Modell mit den Kovariablen  $A$  und  $L$  spezifiziert wird

Auffällig ist, dass die MSE beim Übergang von  $n = 200$  auf  $n = 800$  deutlich geringer werden. Dies liegt an der geringeren Varianz. Konsistenz lässt sich somit bei allen Verfahren (außer bei der naiven Schätzung) vermuten, wenn Outcome-Modell und Exposure-Modell richtig spezifiziert sind. Die hier erwähnte Konsistenz bedeutet, dass der MSE mit steigendem  $n$  gegen 0 geht und ist nicht zu verwechseln mit der Konsistenzannahme im Kontext kausaler Effekte, welche zu Beginn der Arbeit erwähnt wurde. Bei richtiger Modellspezifikation weist Standardisierung die besten Ergebnisse bezüglich des MSE auf. Bei den erweiterten Verfahren sind die MSE sehr ähnlich und etwas höher als bei TMLE.

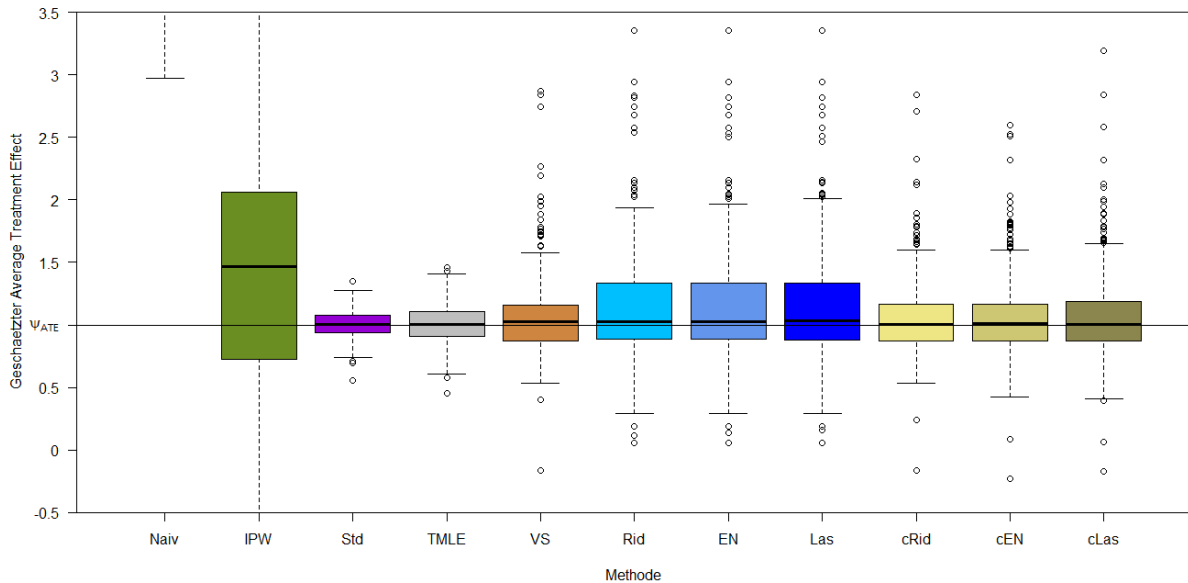
Confounding wurde hier in großem Maß simuliert, weswegen die naive Schätzung einen hohen Bias aufweist. Folgende Grafik zeigt das Ausmaß. Es handelt sich dabei um denselben Plot wie Abbildung 4.4 bzw. Abbildung 4.6 rechts, jedoch wird die Skala hier anders gewählt, um den Bias der naiven Schätzung aufzuzeigen.



**Abbildung 4.7 :** Schätzungen des ATE und MSE für den Stichprobenumfang  $n = 800$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird und das Outcome-Modell mit den Kovariablen  $A$  und  $L$  spezifiziert wird, mit breiterer Skala

### 4.3.2 Verschiedene Assoziationsstrukturen

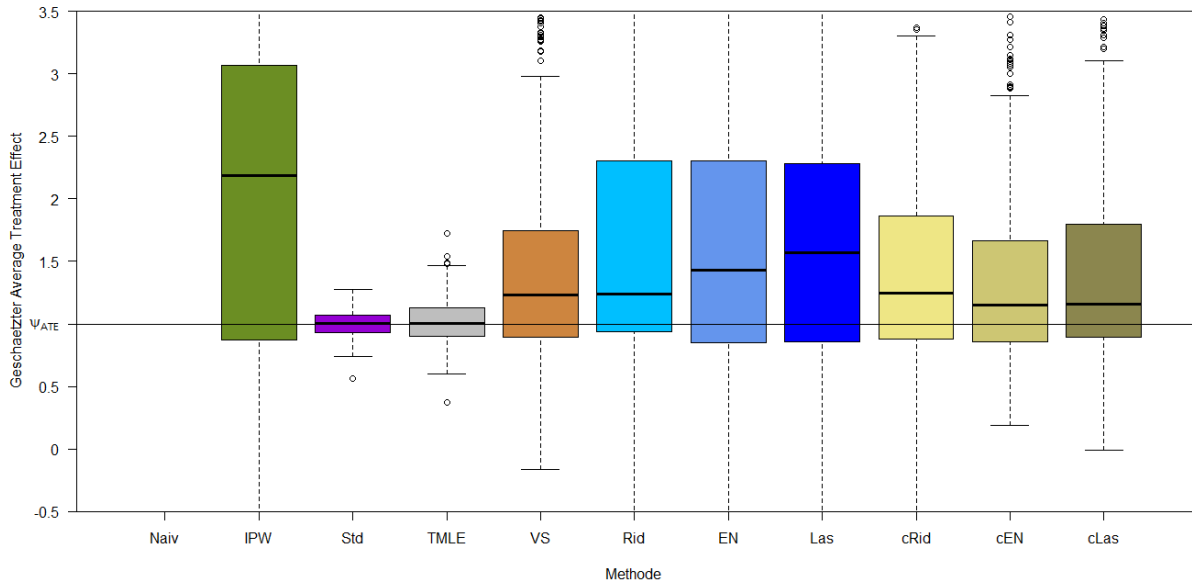
Es folgt die Darstellung der Simulationsergebnisse, bei welchen Assoziationen zwischen den Variablen aus  $L$  bestehen. Diese kommen, wie in Abschnitt 4.1.2 erläutert, durch entsprechende verborgene Variablen  $Z$  zustande. Weisen die Variablen  $L$  einer Gruppe Assoziationen auf, erhält man folgende Schätzungen für  $n = 800$  (Die Bildung von Treatmentvariable und Outcome sowie die Modellspezifikation bleiben wie in Abschnitt 4.3.1):



**Abbildung 4.8** : Schätzungen des ATE für den Stichprobenumfang  $n = 800$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird und das Outcome-Modell mit den Kovariablen  $A$  und  $L$  spezifiziert wird und zwischen den Variablen  $L$  innerhalb einer Gruppe Assoziationen bestehen

Zu beachten ist, dass die Skala nun von  $-0.5$  bis  $3.5$  geht und damit breiter ist als in den Grafiken von Abschnitt 4.3.1 (Außer Abbildung 4.7).

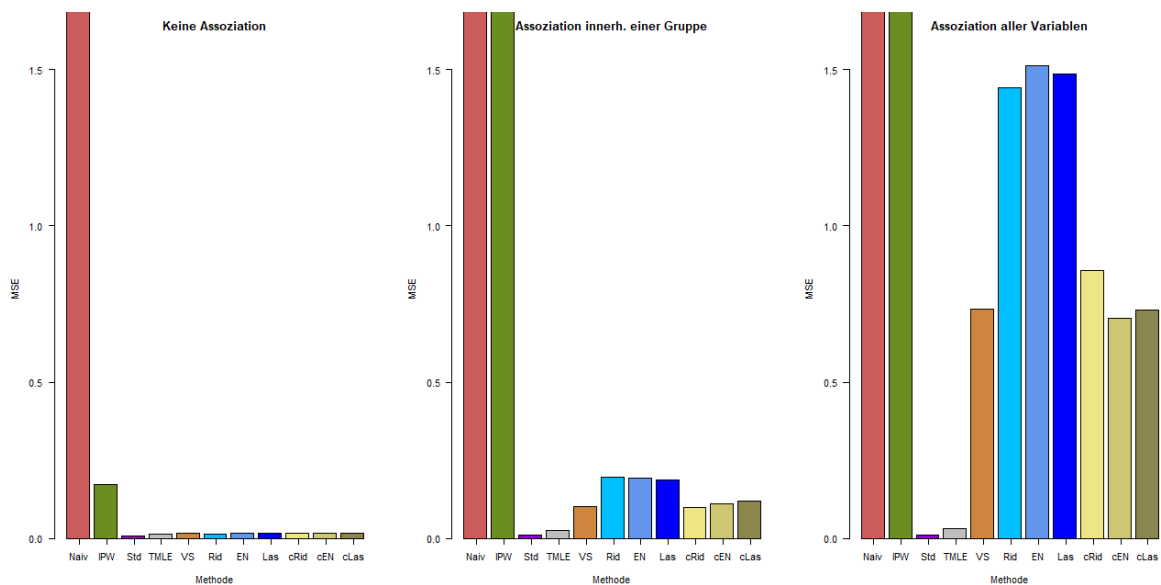
Es zeigt sich, dass der kausale Effekt hier durch Standardisierung und TMLE im Vergleich zu den anderen Methoden sehr gut geschätzt wird. Die Variablenselektions- und Shrinkage-Verfahren schneiden schlechter ab. Die Varianz steigt im Vergleich zum Fall „Keine Assoziationen zwischen den Variablen aus  $L$ “. Insbesondere sind bei diesen Methoden auch einige Extremwerte nach oben zu sehen. Variablenselektion und komponentenweises Shrinkage zeigen etwas bessere Ergebnisse als nicht-komponentenweises Shrinkage. Die Verzerrung der naiven Schätzung sowie die Varianz des IP-Weighting nehmen im Vergleich zum Fall ohne Assoziationen zwischen den Variablen aus  $L$  zu. Weisen alle Variablen aus  $L$  einen Zusammenhang auf, ergeben sich die folgenden Ergebnisse:



**Abbildung 4.9 :** Schätzungen des ATE für den Stichprobenumfang  $n = 800$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird und das Outcome-Modell mit den Kovariablen  $A$  und  $L$  spezifiziert wird und zwischen allen Variablen  $L$  Assoziationen bestehen

Beim Übergang von der Struktur „Assoziationen zwischen den Variablen  $L$  einer Gruppe“ zum Fall „Assoziationen zwischen allen Variablen aus  $L$ “ zeigt sich, dass sich die Güte der Schätzverfahren weiter verschlechtert. Lediglich bei der Standardisierung und bei TMLE bleiben die Schätzungen sehr ähnlich, weswegen diese auch hier deutlich besser schätzen als die anderen Verfahren.

Von Interesse ist nun ein MSE-Vergleich unter den drei Assoziationsstrukturen:



**Abbildung 4.10 :** MSE-Vergleich der Schätzungen des ATE für die Fälle keine Assoziationen, Assoziationen innerhalb einer  $L$ -Gruppe und Assoziationen zwischen allen Variablen aus  $L$  beim Stichprobenumfang  $n = 800$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird und das Outcome-Modell mit den Kovariablen  $A$  und  $L$  spezifiziert wird

Man sieht: Der MSE nimmt mit der Anzahl der Assoziationen unter den Variablen aus  $L$  zu. Bei Standardisierung und TMLE ist dies nur in geringem Maße zu beobachten, bei den anderen Verfahren ist der Anstieg sehr deutlich.

### 4.3.3 Fehlspezifikation des Outcome-Modells

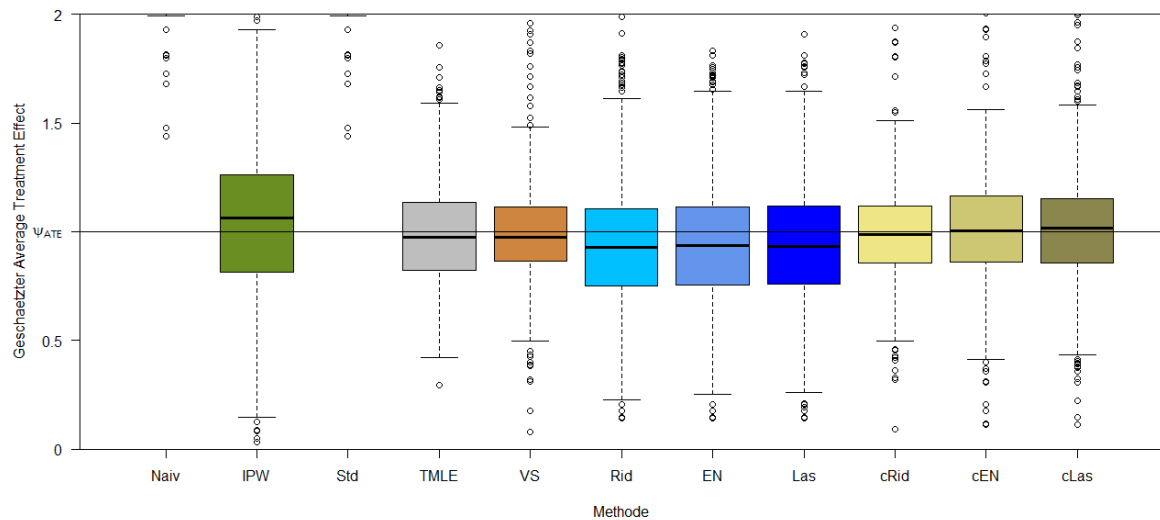
Bisher wurden nur die Fälle betrachtet, in denen sowohl Outcome-Modell als auch Exposure-Modell richtig spezifiziert wurden (Richtig in dem Sinne, dass alle richtigen Terme im spezifizierten Modell enthalten sind). In diesem Abschnitt steht die Fehlspezifikation des Outcome-Modells im Vordergrund. Es werden die Fälle betrachtet, dass

- einfache Terme fehlen
- Interaktionsterme fehlen
- quadratische Terme fehlen

Somit ist die Korrektur durch das Exposure-Modell in diesem Fall von großer Bedeutung. Das Exposure-Modell wird nach wie vor richtig spezifiziert: Bildung der Treatmentvariable und Modellspezifikation erfolgen jeweils durch einfache Terme.

## Fehlende einfache Terme

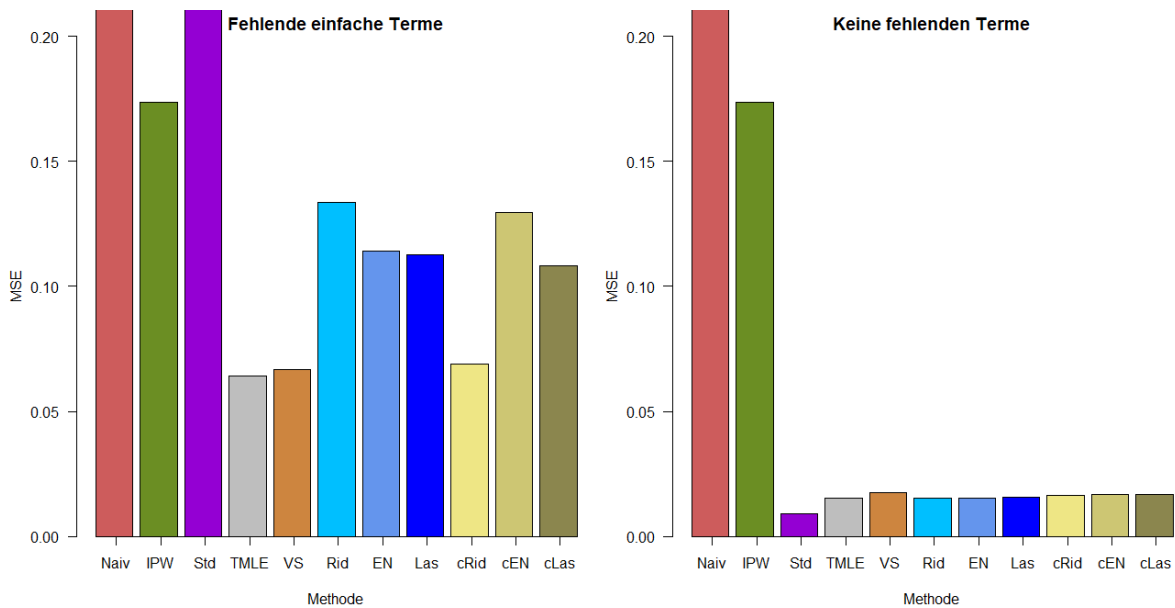
Der Outcome wird generiert wie in Abschnitt 4.3.1. Spezifiziert man das Modell nur mit  $A$  als Kovariable, so fehlen alle einfachen Terme der Variablen  $L^{Y^0}$ . Standardisierung entspricht in diesem Fall der naiven Schätzung, weshalb die Schätzungen dieser beiden Verfahren hier identisch sind. Man erhält folgende Schätzungen:



**Abbildung 4.11 :** Schätzungen des ATE für den Stichprobenumfang  $n = 800$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird und das Outcome-Modell nur mit Kovariable  $A$  spezifiziert wird

Die naive Schätzung/Standardisierung weist eine starke Verzerrung auf. Bei den anderen Verfahren ist der Bias sehr gering. Die Korrektur durch das Exposure-Modell funktioniert also. Von den erweiterten Verfahren schneiden der CTMLE-Variablenselektionsalgorithmus und der CTMLE-Algorithmus mit komponentenweisem Shrinkage unter Ridge-Penalisierung am besten ab. Die Interquartilsabstände der beiden Verfahren sind geringer als bei der gewöhnlichen TMLE-Schätzung, die Varianzen sind dennoch etwas höher.

Es folgt ein Vergleich der MSE unter dieser Fehspezifikation und unter richtiger Modellspezifikation wie in Abschnitt 4.3.1:



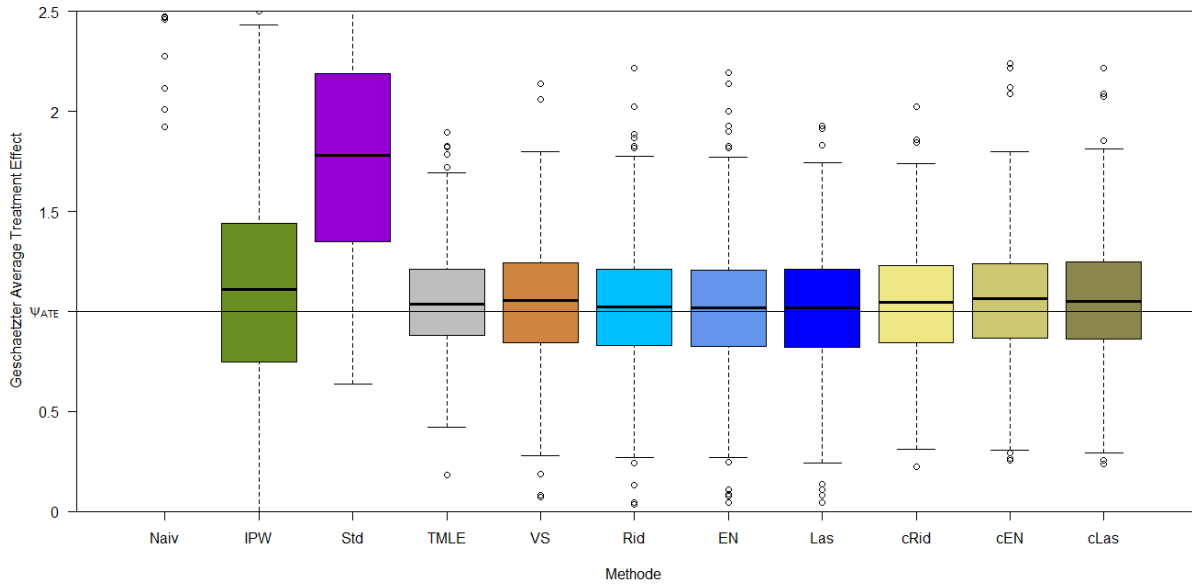
**Abbildung 4.12 :** MSE-Vergleich der Schätzungen des ATE für die Spezifikation des Outcome-Modells nur mit Kovariable  $A$  (Fehlende Terme) und die Spezifikation des Outcome-Modells mit den Kovariablen  $A$  und  $L$  (Keine fehlenden Terme) beim Stichprobenumfang  $n = 800$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  generiert wird

Bei der naiven Schätzung und beim IP-Weighting ändert sich der MSE nicht, da bei diesen Verfahren kein Outcome-Modell spezifiziert wird. Bei den anderen Verfahren wird der MSE deutlich geringer, wenn die einfachen Terme  $L$  aufgenommen werden. Dieses Verhalten zeigt sich bei der Standardisierung am stärksten. Wie bereits bei Abbildung 4.11 festgestellt, weisen TMLE, VS und cRid von den Schätzverfahren die besten Schätzungen auf, wenn einfache Terme fehlen.

### Fehlende Interaktionsterme

In diesem Abschnitt wird der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  sowie Interaktionsterme  $A : L^{Y^0}$  gebildet. Zunächst werden die Ergebnisse dargestellt, falls das Modell fehlspezifiziert wird, indem nur einfache Terme  $A$  und  $L$  angenommen werden:

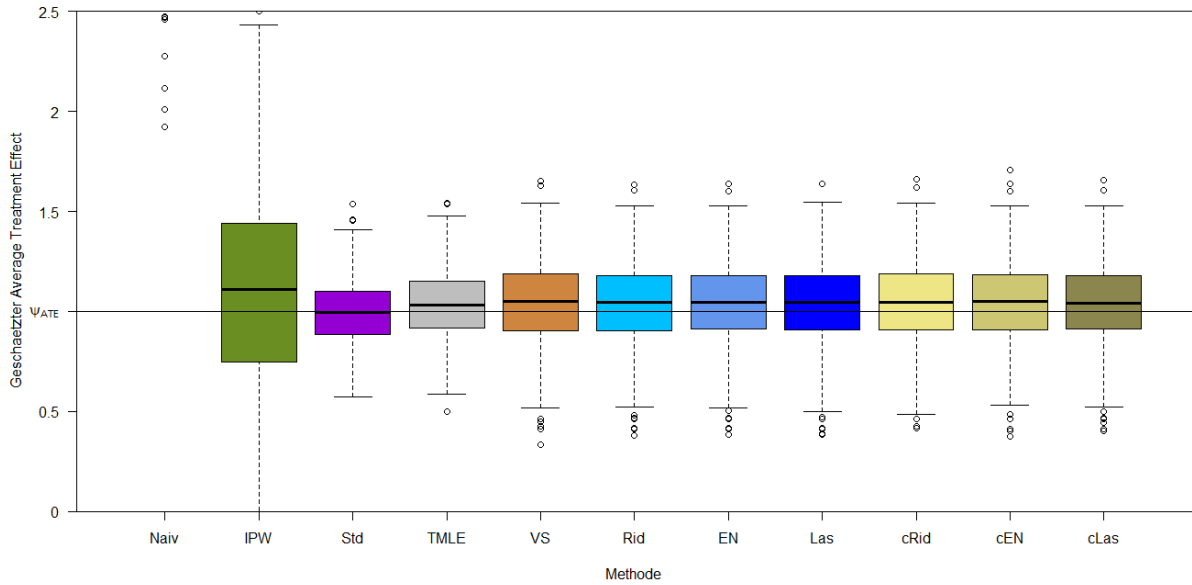




**Abbildung 4.13 :** Schätzungen des ATE für den Stichprobenumfang  $n = 800$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  sowie Interaktionsterme  $A : L^{Y^0}$  generiert wird und das Outcome-Modell nur mit einfachen Termen  $A$  und  $L$  spezifiziert wird

Man sieht: Die Schätzungen der Standardisierung sind wieder verzerrt. Bei den anderen Verfahren (außer der naiven Schätzung) ist der Bias dank der Korrektur des richtig spezifizierten Exposure-Modells gering. TMLE weist auch hier wieder eine etwas geringere Streuung auf als die Variablenselektions- und Shrinkage-Verfahren.

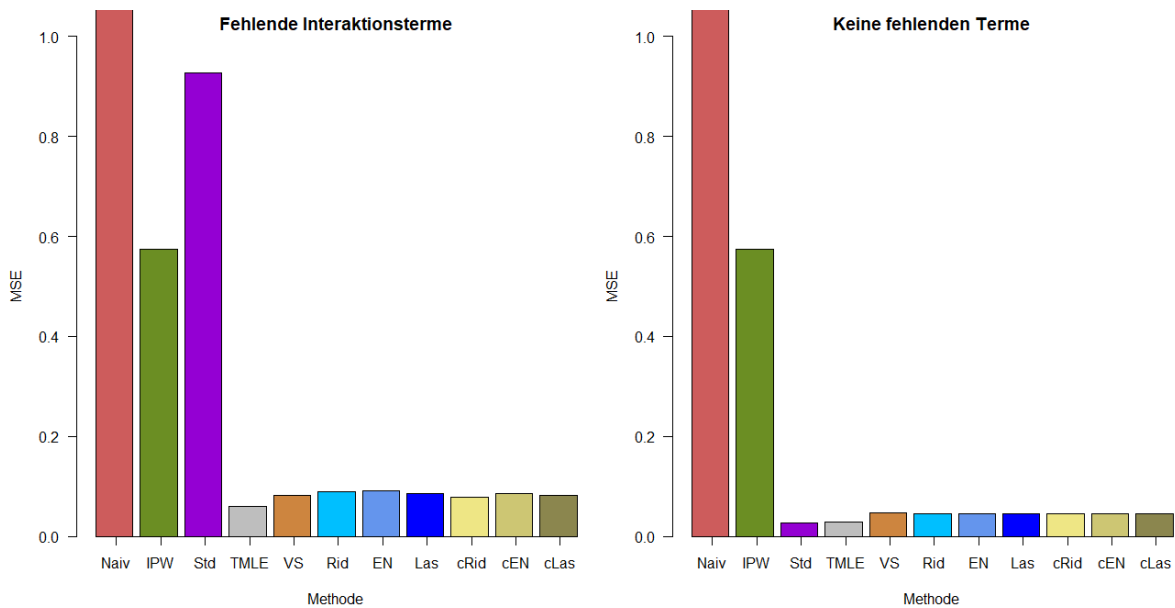
Werden Interaktionsterme  $A : L$  bei der Modellspezifikation berücksichtigt, ergeben sich folgende Schätzungen:



**Abbildung 4.14 :** Schätzungen des ATE für den Stichprobenumfang  $n = 800$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y^0}$  sowie Interaktionsterme  $A : L^{Y^0}$  generiert wird und das Outcome-Modell mit einfachen Termen  $A$  und  $L$  sowie Interaktionstermen  $A : L$  spezifiziert wird

Standardisierung schneidet nun wieder am besten ab: Verzerrung und Varianz werden deutlich geringer, wenn Interaktionsterme bei der Modellierung berücksichtigt werden. Bei TMLE und den erweiterten Schätzverfahren wird die Streuung auch etwas geringer, jedoch nicht in so großem Maße wie bei der Standardisierung.

Auch hier werden die MSE unter den beiden Modellspezifikationen verglichen:

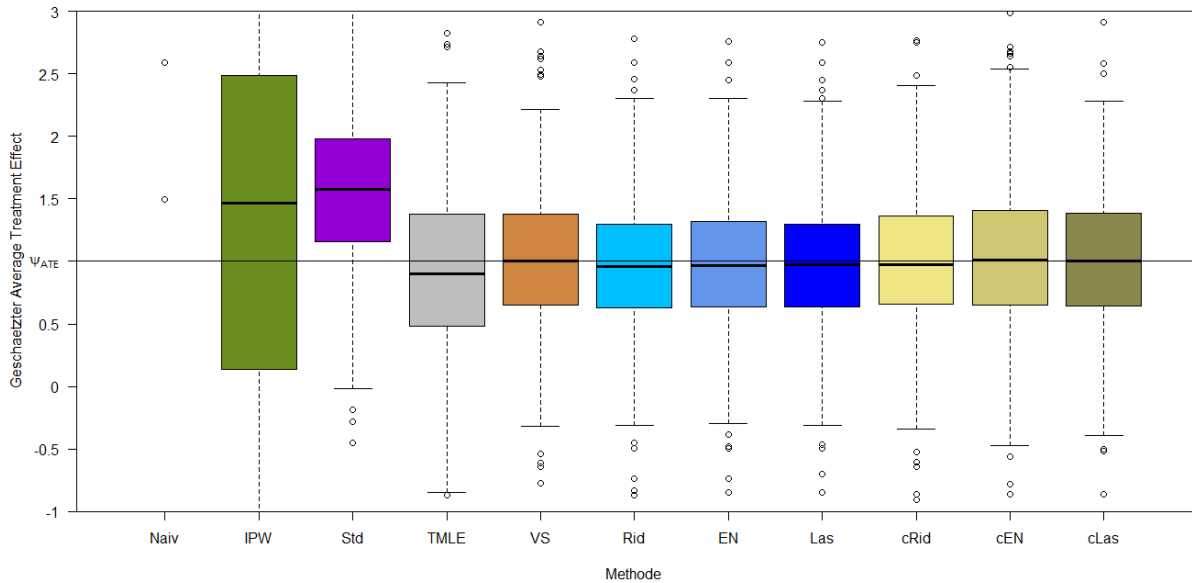


**Abbildung 4.15 :** MSE-Vergleich der Schätzungen des ATE für die Spezifikation des Outcome-Modells nur mit einfachen Termen  $A$  und  $L$  (Fehlende Interaktionsterme) und die Spezifikation des Outcome-Modells mit einfachen Termen  $A$  und  $L$  sowie Interaktionstermen  $A : L$  (Keine fehlenden Terme) beim Stichprobenumfang  $n = 800$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y_0}$  sowie Interaktionsterme  $A : L^{Y_0}$  generiert wird

Der MSE nimmt bei der Spezifikation von Interaktionstermen ab. Diese Abnahme zeigt sich bei der Standardisierung stärker als bei den anderen Verfahren. Nach wie vor weist der gewöhnliche TMLE-Ansatz etwas geringere MSE auf als die Variablenselektions- und Shrinkage-Verfahren.

### Fehlende quadratische Terme

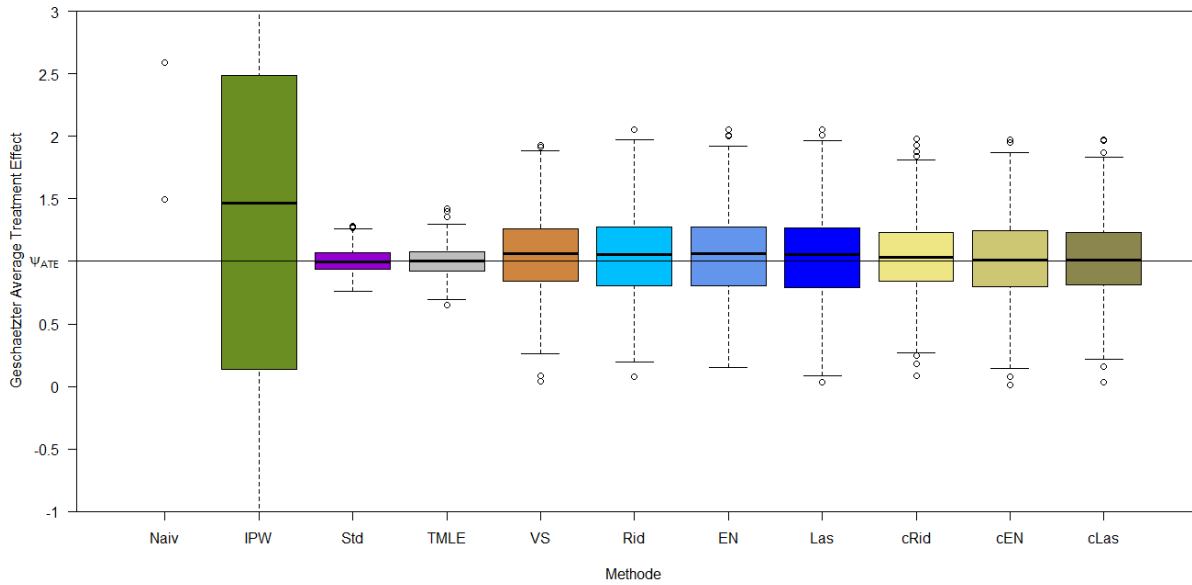
Analog zum Fall „Fehlende Interaktionsterme“ wird nun die Bildung des Outcomes durch einfache Terme  $A$  und  $L^{Y_0}$  und quadratische Terme  $(L^{Y_0})^2$  herangezogen. Falls das Modell durch fehlende quadratische Terme fehlspezifiziert wird, erhält man folgende Schätzungen:



**Abbildung 4.16 :** Schätzungen des ATE für den Stichprobenumfang  $n = 800$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y_0}$  sowie quadratische Terme  $(L^{Y_0})^2$  generiert wird und das Outcome-Modell nur mit einfachen Termen  $A$  und  $L$  spezifiziert wird

Wie bei den anderen Fehlspezifikationen sind die Schätzungen der Standardisierung verzerrt. Zum ersten Mal weisen die Variablenselektions- und Shrinkage-Verfahren bessere Ergebnisse auf als der gewöhnliche TMLE-Ansatz. Dies zeigt sich sowohl an einem etwas geringeren Bias als auch an einer etwas geringeren Varianz. Die klassische Variablenselektion schneidet etwas schlechter ab als die Shrinkage-Algorithmen. Letztere zeigen bei Ridge-Penalisation etwas bessere Ergebnisse als bei den beiden anderen Penalisierungstermen.

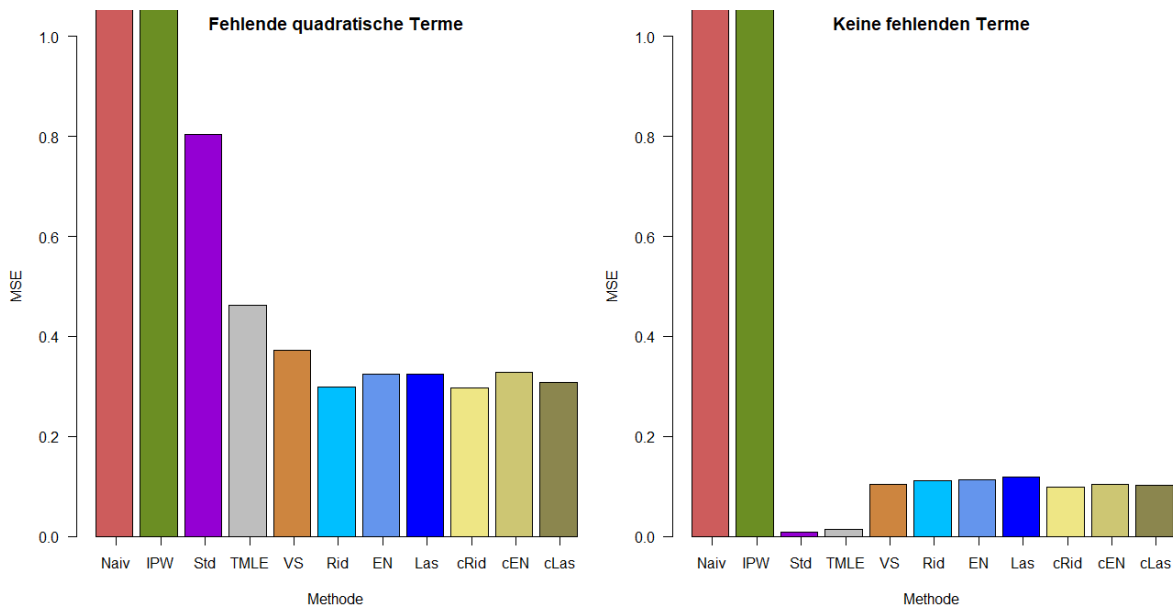
Eine Berücksichtigung der quadratischen Terme bei der Modellspezifikation führt zu folgenden Schätzungen:



**Abbildung 4.17** : Schätzungen des ATE für den Stichprobenumfang  $n = 800$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y_0}$  sowie quadratische Terme  $(L^{Y_0})^2$  generiert wird und das Outcome-Modell mit einfachen Termen  $A$  und  $L$  sowie quadratischen Termen  $L^2$  spezifiziert wird

Standardisierung und TMLE schneiden nun wieder deutlich besser ab als die anderen Verfahren.

Ein MSE-Vergleich bestätigt die Ergebnisse:

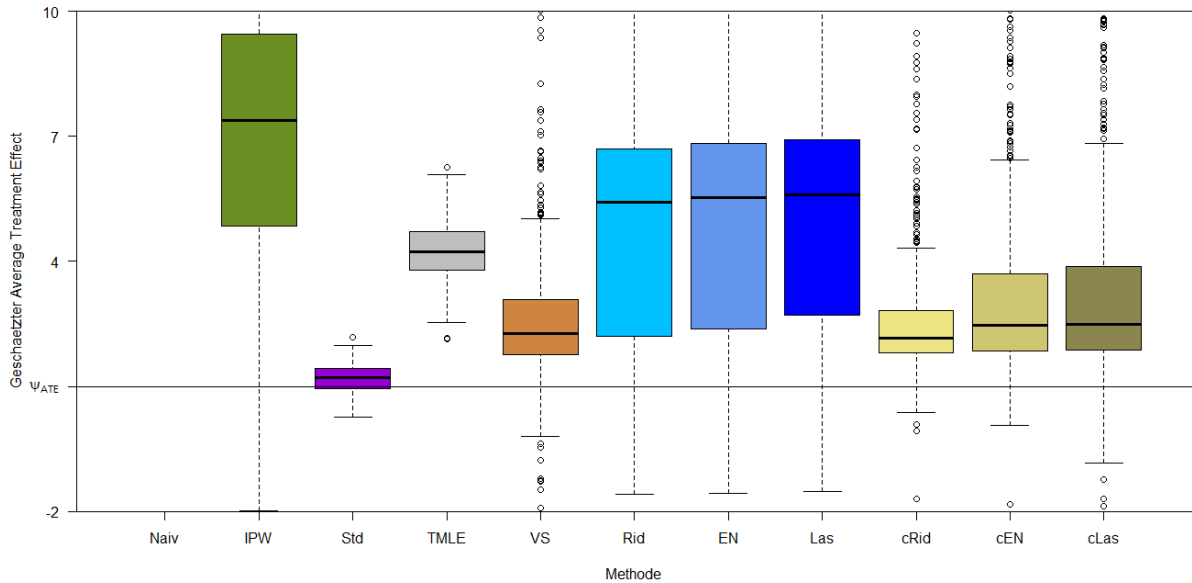


**Abbildung 4.18** : MSE-Vergleich der Schätzungen des ATE für die Spezifikation des Outcome-Modells nur mit einfachen Termen  $A$  und  $L$  (Fehlende quadratische Terme) und die Spezifikation des Outcome-Modells mit einfachen Termen  $A$  und  $L$  sowie quadratischen Termen  $L^2$  (Keine fehlenden Terme) beim Stichprobenumfang  $n = 800$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y_0}$  sowie quadratische Terme  $(L^{Y_0})^2$  generiert wird

Bei fehlenden quadratischen Termen weisen die erweiterten Methoden geringere MSE auf als TMLE. Der MSE für die Standardisierung ist sehr hoch. Wird das Outcome-Modell durch die Annahme quadratischer Terme richtig spezifiziert, werden die MSE bei allen Verfahren geringer, besonders jedoch bei Standardisierung und TMLE, so dass diese dann eine höhere Güte bezüglich des MSE aufweisen als die erweiterten Verfahren.

#### 4.3.4 Fehlspezifikation von Outcome-Modell und Exposure-Modell

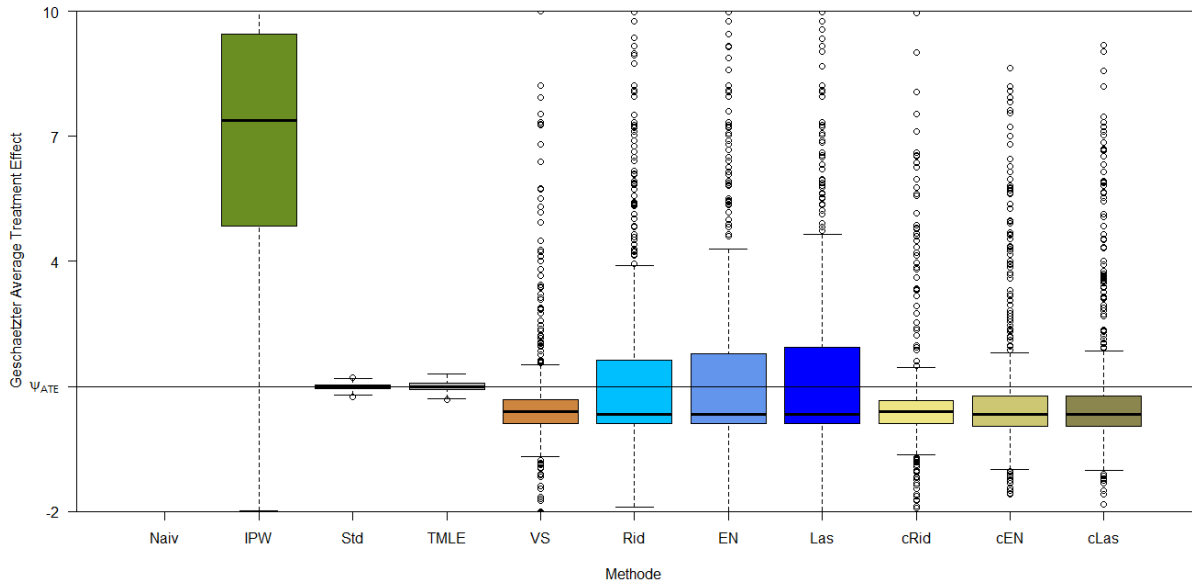
Die bisherigen Analysen bezogen sich auf ein richtig spezifiziertes Exposure-Modell. Es stellt sich die Frage, wie gut die Schätzungen sind, wenn sowohl Outcome-Modell als auch Exposure-Modell fehlspezifiziert sind. Dafür wird  $Y$  durch einfache Terme  $L^{Y_0}$ , Interaktionsterme  $A : L^{Y_0}$  und quadratische Terme  $(L^{Y_0})^2$  gebildet. Bei der Modellierung werden nur einfache Terme  $A$  und  $L$  spezifiziert.  $A$  wird durch einfache Terme  $L^{A_0}$  und quadratische Terme  $(L^{A_0})^2$  gebildet. Das Exposure-Modell wird jedoch nur mit einfachen Termen  $L$  spezifiziert. Da die Wirkungsstruktur hier komplexer ist, wird ein höherer Stichprobenumfang herangezogen:  $n = 3200$ . Es ergeben sich folgende Schätzungen:



**Abbildung 4.19** : Schätzungen des ATE für den Stichprobenumfang  $n = 3200$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y0}$ , Interaktionsterme  $A : L^{Y0}$  sowie quadratische Terme  $(L^{Y0})^2$  generiert wird und das Outcome-Modell nur mit einfachen Termen  $A$  und  $L$  spezifiziert wird, und wobei die Treatmentvariable durch einfache Terme  $L^{A0}$  und quadratische Terme  $(L^{A0})^2$  generiert wird und das Exposure-Modell nur mit einfachen Termen  $L$  spezifiziert wird

Alle Schätzverfahren sind hier verzerrt. Trotz fehlspezifiziertem Outcome-Modell zeigt Standardisierung allerdings nur einen sehr geringen Bias. Die Verzerrung des CTMLE-Variablenselektionsalgorithmus und der komponentenweisen CTMLE-Shrinkage-Algorithmen ist geringer als bei der Targeted Maximum Likelihood Estimation, jedoch höher als bei der Standardisierung. IP-Weighting und die „normalen“ CTMLE-Shrinkage-Algorithmen zeigen noch stärkere Verzerrungen. Die Varianz ist bei der Standardisierung und bei TMLE sehr gering, während die erweiterten Schätzverfahren hohe Streuungen zeigen. Dies ist am Interquartilsabstand und an einer hohen Anzahl von Extremwerten ersichtlich.

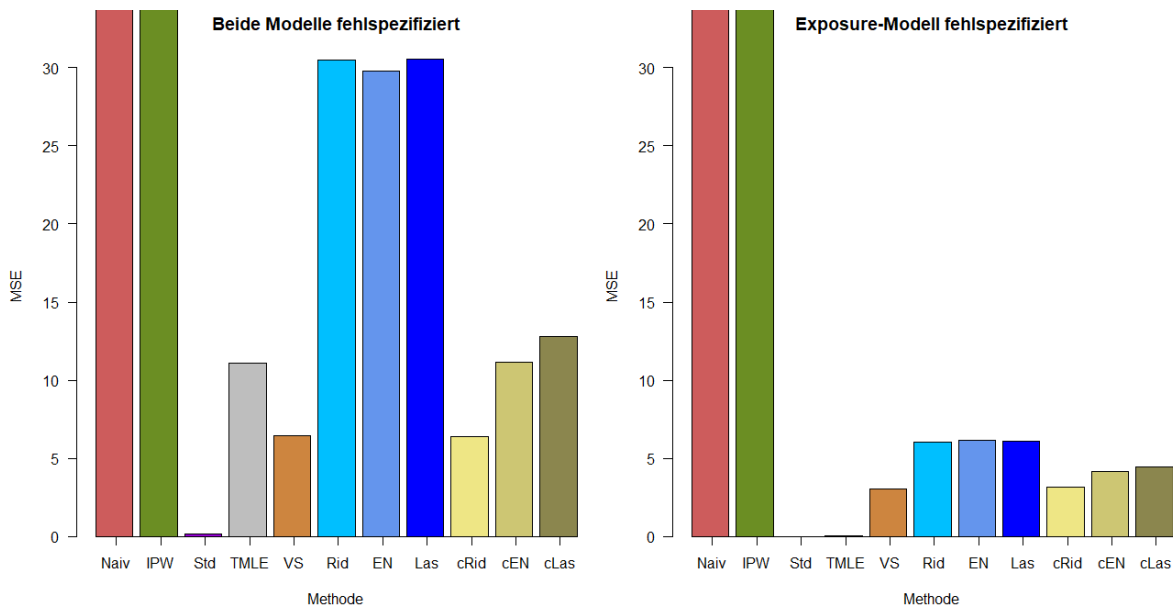
Wird das Outcome-Modell richtig spezifiziert (Das Exposure-Modell bleibt nach wie vor fehlspezifiziert), ergeben sich folgende Schätzungen:



**Abbildung 4.20 :** Schätzungen des ATE für den Stichprobenumfang  $n = 3200$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y_0}$ , Interaktionsterme  $A : L^{Y_0}$  sowie quadratische Terme  $(L^{Y_0})^2$  generiert wird und das Outcome-Modell mit einfachen Termen  $A$  und  $L$ , Interaktionstermen  $A : L$  sowie quadratischen Termen  $L^2$  spezifiziert wird, und wobei die Treatmentvariable durch einfache Terme  $L^{A_0}$  und quadratische Terme  $(L^{A_0})^2$  generiert wird und das Exposure-Modell nur mit einfachen Termen  $L$  spezifiziert wird

Die Verzerrung wird bei allen Schätzverfahren (außer naiv und IP-Weighting) deutlich geringer. Der Bias ist bei TMLE nun kleiner als bei den erweiterten Schätzverfahren. Auch die Varianz nimmt ab. Jedoch zeigen sich nach wie vor viele Extremwerte bei den erweiterten Schätzverfahren. Diese schneiden hier nun wieder schlechter ab als TMLE. Die MSE sehen folgendermaßen aus:





**Abbildung 4.21** : MSE-Vergleich der Schätzungen des ATE für die Spezifikation des Outcome-Modells nur mit einfachen Termen  $A$  und  $L$  (Fehlende Interaktionsterme und quadratische Terme) und die Spezifikation des Outcome-Modells mit einfachen Termen  $A$  und  $L$ , Interaktionstermen  $A : L$  sowie quadratischen Termen  $L^2$  (Keine fehlenden Terme) beim Stichprobenumfang  $n = 3200$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y_0}$ , Interaktionsterme  $A : L^{Y_0}$  sowie quadratische Terme  $(L^{Y_0})^2$  generiert wird, und wobei die Treatmentvariable durch einfache Terme  $L^{A_0}$  und quadratische Terme  $(L^{A_0})^2$  generiert wird und das Exposure-Modell nur mit einfachen Termen  $L$  spezifiziert wird

Standardisierung weist in beiden Fällen den geringsten MSE auf. VS und cRid zeigen einen kleineren MSE als TMLE, wenn beide Modelle fehlspezifiziert sind. Dies ändert sich, wenn das Outcome-Modell richtig spezifiziert wird. Beim Übergang wird der MSE bei allen Verfahren, bei welchen ein Outcome-Modell spezifiziert wird, geringer.

# 5 Zusammenfassung und Abschluss

## 5.1 Zusammenfassung

In dieser Arbeit wurde die Problematik der Schätzung kausaler Effekte aufgezeigt. Ist man am kausalen Effekt einer Treatmentvariable  $A$  auf einen Outcome  $Y$  interessiert, reicht es in der Regel nicht, eine einfache Regression mit  $Y$  als Zielvariable und  $A$  als Einflussvariable durchzuführen. Dadurch wird lediglich eine Assoziation geschätzt. Existieren Variablen  $L^0$ , die einen Effekt auf  $A$  und  $Y$  haben, liegt Confounding vor, was dazu führt, dass sich Assoziation und Kausalität unterscheiden. Somit ist es notwendig, solche Variablen (Confounder) bei der Schätzung eines kausalen Effekts zu berücksichtigen.

Ein gängiges Verfahren zur Schätzung kausaler Effekte ist Standardisierung. Hier werden die möglichen Confounder  $L$  in einem Outcome-Modell  $E[Y|A, L]$  als Kovariablen modelliert. Anschließend wird über diese Variablen gemittelt. Beim IP-Weighting wird die Wirkung von den Confoundern auf die Treatmentvariable eliminiert, indem  $P[A = 1|L]$ , das sogenannte Exposure-Modell, modelliert wird. Entsprechend der Vorhersagen dieses Modells werden die Beobachtungen bei der Schätzung des Modells  $E[Y|A]$  gewichtet. Targeted Maximum Likelihood Estimation kombiniert die beiden Ansätze, wodurch es robuster gegenüber Modellfehlspezifikationen ist als die beiden Standardverfahren. TMLE wiederum kann erweitert werden, indem das Exposure-Modell durch Variablen-selektion und/oder Shrinkage schrittweise aufgebaut wird. Die Idee dahinter ist, dass in der Regel unklar ist, welche Variablen aus den Daten einen Effekt auf  $A$  und/oder auf  $Y$  haben und demzufolge verwendet werden sollten, um das Confounding-Problem zu beheben. Shrinkage und Variablenselektion kann auch verbunden werden: In diesem Fall können Variablen unterschiedlich stark penalisiert werden (Komponentenweises Shrinkage).

Um die Güte der Schätzverfahren zu untersuchen, wurden Simulationen durchgeführt, in welchen die Daten 500 mal erzeugt wurden und der kausale Effekte mit jeder Methode 500 mal geschätzt wurde. Dadurch konnten Mean Squared Errors unter verschiedenen datengenerierenden Prozessen geschätzt werden. 10 Variablen mit 4 enthaltenen Confoundern wurden jeweils generiert. Bei den 6 anderen Variablen war nur ein Effekt auf  $A$ , nur ein Effekt auf  $Y$  oder kein Effekt auf  $A$  oder auf  $Y$  vorhanden. Da für

die Schätzung angenommen wurde, dass die kausalen Beziehungen nicht bekannt sind, wurden immer alle 10 Variablen in die Schätzmethoden miteinbezogen. Es wurden die Stichprobenumfänge  $n = 200$  und  $n = 800$  (in einem Fall:  $n = 3200$ ) und verschiedene Assoziationsstrukturen innerhalb der möglichen Confounder  $L$  betrachtet. Zudem wurden sowohl richtige Spezifikationen von Outcome-Modell und Exposure-Modell als auch Fehlspezifikationen der beiden Modelle herangezogen (wobei richtige Modellspezifikation in diesem Fall bedeutet, dass alle richtigen Terme spezifiziert werden).

Die Ergebnisse bestätigen, dass eine naive Schätzung ohne Berücksichtigung des Confounding-Problems stark verzerrt sein kann. IP-Weighting weist überwiegend hohe Streuungen der Schätzungen auf. Bei richtiger Spezifikation des Exposure-Modells ist der Bias hier gering, ansonsten kann auch dieser sehr hoch sein. Standardisierung zeigt von den betrachteten Verfahren die besten Ergebnisse, wenn das Outcome-Modell richtig spezifiziert ist: Sowohl Bias als auch Varianz sind sehr gering. Bei Fehlspezifikation des Outcome-Modells nimmt die Verzerrung zu. Diese Zunahme kann gering, aber auch sehr hoch sein. Bei TMLE ist der Bias nur dann hoch, wenn Outcome-Modell und Exposure-Modell fehlspezifiziert sind. Die Methode zeigt aufgrund höherer Varianz etwas höhere MSE als Standardisierung, wenn beide Modelle richtig spezifiziert sind. Ist nur das Outcome-Modell fehlspezifiziert, sind die Schätzungen besser als die in diesem Fall verzerrten Schätzungen der Standardisierung.

Die Variablenselektions- und Shrinkageverfahren schneiden bei richtiger Modellspezifikation aufgrund höherer Varianz etwas schlechter ab als Standardisierung und TMLE. Der Bias ist wie bei den anderen Verfahren sehr gering. Bei Fehlspezifikation zeigen sie teilweise bessere Ergebnisse als die anderen Verfahren. Wird nur das Outcome-Modell fehlspezifiziert, bleibt die Unverzerrtheit bestehen, weshalb diese Verfahren hier geringere MSE aufweisen als Standardisierung. Gegenüber TMLE zeigen sich bei Fehlspezifikation nur in manchen Szenarien bessere Schätzergebnisse. Beispielsweise ist die Verzerrung des CTMLE-Variablenselektionsalgorithmus und der CTMLE-Schätzung mit komponentenweisem Shrinkage bei Fehlspezifikation beider Modelle geringer als bei TMLE, die Varianz wiederum jedoch höher. Von den Variablenselektions- und Shrinkage-Verfahren zeigt komponentenweise Shrinkage-Schätzung unter Ridge-Penalisation zumeist den geringsten MSE. Eine Verbesserung dieser Verfahren gegenüber TMLE kann jedoch im Großen und Ganzen nicht bestätigt werden.

Höhere Stichprobenumfänge sorgen für geringere Streuungen (Konsistenzeigenschaft bei richtiger Modellspezifikation außer bei naiver Schätzung). Mit zunehmenden Assoziationen unter den Variablen aus  $L$  nimmt der MSE zu. Dies ist bei Standardisierung und TMLE nur in geringem Maße zu beobachten, bei den Variablenselektions- und Shrinkagealgorithmen ist es sehr deutlich zu sehen.

## 5.2 Grenzen der Analysen und Ausblick

Man sieht: Der datengenerierende Prozess sowie die Modellspezifikation haben einen erheblichen Einfluss auf die Güte der betrachteten Methoden. Dieser Einfluss unterscheidet sich von Schätzverfahren zu Schätzverfahren. Die Struktur ist dabei teilweise noch etwas unklar. In der Praxis sind Modellfehlspezifikationen wohl die Regel, das Ausmaß einer Fehlspezifikation ist jedoch von Fall zu Fall unterschiedlich. Somit ist es zumeist schwierig, für bestimmte Daten festzustellen, welches Verfahren und welche Modellspezifikation verwendet werden sollte.

Zudem wurde nur ein Teil von möglichen datengenerierenden Prozessen herangezogen. Weiterführende Simulationen zu den Schätzverfahren können durchgeführt werden:

- Collider sind, wie schon angesprochen, Variablen, die eine Verzerrung verursachen, wenn auf diese bedingt wird. Eine Erweiterung der Simulationen kann erfolgen, indem Collider erzeugt werden, welche dann ebenfalls in den Daten auftauchen. Ist nicht bekannt, welche Variablen Collider sind, ist Variablenselektion von Bedeutung.
- Es wurde angenommen, dass in den Variablen  $L$  alle Confounder enthalten sind. Es ist jedoch möglich, dass Confounder unbekannt und/oder schwer zu erheben sind. Demzufolge kann es sinnvoll sein, einen Teil der simulierten Confounder nicht in die Daten aufzunehmen, die für die Schätzung zur Verfügung stehen. In diesem Fall kann Exchangeability nicht erreicht werden, da auf  $L$  bedingte Exchangeability nicht erfüllt ist.
- In den durchgeführten Simulationen wurden 10 Variablen  $L$  mit 4 enthaltenen Confoundervariablen generiert. Die Anzahl der Variablen, und auch der Anteil der Confounder kann anders gewählt werden. Die Simulationen können beispielsweise mit einem höheren Anteil an Variablen durchgeführt werden, die einen (starken) Effekt auf  $A$  haben, aber keinen Effekt auf  $Y$  (Positivitätsverletzungen).
- Zur Erzeugung der Daten können auch andere Verteilungen als die gewählten verwendet werden. Hier wurde bei den Variablen  $L$  und beim Outcome  $Y$  auf die Normalverteilung zurückgegriffen. Für die binäre Treatmentvariable  $A$  wurde die Bernoulliverteilung herangezogen. Interessant wäre eine Durchführung der Simulationen mit anderen Verteilungen: Z.B. Poisson-Verteilung, Exponentialverteilung, Weibullverteilung etc. Auch kann  $A$  metrisch und damit nicht mehr bernoulliverteilt sein. In diesem Fall muss der kausale Effekt jedoch anders definiert werden.

- Auch wenn bereits Interaktionsterme und quadratische Terme herangezogen wurden, können die Funktionen zur Bildung der Variablen noch deutlich komplexer sein: Z.B. höhere Polynome, abschnittsweise Funktionen etc. In diesem Zusammenhang wäre bei der Schätzung auch eine Anwendung von Machine-Learning-Verfahren zur Bildung der Vorhersagen von Interesse. Oft werden sogenannte Super Learner verwendet, welche verschiedene Vorhersagemethoden kombinieren. Das heißt sowohl komplexere Wirkungsstrukturen zwischen den Variablen als auch komplexere Modellspezifikationen können für die betrachteten Schätzverfahren untersucht werden.

In der Arbeit wurden einige Analyseergebnisse bezüglich der Güte verschiedener Methoden zur Schätzung kausaler Effekte gewonnen. Eine Struktur ist dabei teilweise zu sehen, teilweise ist diese jedoch noch unklar. Insbesondere wann die Variablenselektions- und Shrinkage-Verfahren der gewöhnlichen TMLE vorzuziehen sind (und gegebenenfalls welche), ist noch nicht genau ersichtlich. Es bieten sich weiterführende Analysen an. Die oben genannten Punkte bieten einige Ideen.

# Literaturverzeichnis

- Hernan MA, Robins JM (2018). *Causal Inference*, Boca Raton: Chapman Hall/CRC, forthcoming.  
URL: <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>
- Mireille E. Schnitzer, Matthew Cefalu (2017). *Collaborative targeted learning using regression shrinkage*, Statistics in Medicine, Faculty of Pharmacy, Université de Montréal
- Megan S. Schuler, Sherri Rose (2016). *Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies*, American Journal of Epidemiology.
- Judea Pearl (2009). *Causal inference in statistics: An overview*, Computer Science Department, University of California
- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009). *The Elements of Statistical Learning*, Second Edition, Springer
- Michael Schomacker (2017). *Kausale Inferenz* (Lehrveranstaltung), Institut für Statistik, Ludwig-Maximilians-Universität München  
URL: <https://moodle.lmu.de/course/view.php?id=1958>
- Jerome Friedman, Trevor Hastie, Rob Tibshirani, Noah Simon, Balasubramanian Narasimhan, Junyang Qian (2018). Package 'glmnet'.  
URL: <https://cran.r-project.org/web/packages/glmnet/glmnet.pdf>
- Susan Gruber, Mark van der Laan (2018). Package 'tmle'.  
URL: <https://cran.r-project.org/web/packages/tmle/tmle.pdf>

# Abbildungsverzeichnis

2.1	DAG - Kausaler Effekt . . . . .	3
2.2	DAG - Confounding . . . . .	5
3.1	DAG - Confounder und weitere Variablen . . . . .	13
3.2	Lasso-Penalisierung (links) und Ridge-Penalisierung (rechts) . . . . .	18
4.1	DAG - Kausale Beziehungen für die Simulationen . . . . .	24
4.2	DAG - Kausale Beziehungen für die Simulationen bei Assoziationen zwischen den Variablen $L$ innerhalb einer L-Gruppe . . . . .	25
4.3	DAG - Kausale Beziehungen für die Simulationen bei Assoziationen zwischen allen Variablen $L$ . . . . .	26
4.4	Schätzungen des ATE für $n = 200$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird und das Outcome-Modell mit den Kovariablen $A$ und $L$ spezifiziert wird . . . . .	35
4.5	Schätzungen des ATE für $n = 800$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird und das Outcome-Modell mit den Kovariablen $A$ und $L$ spezifiziert wird . . . . .	36
4.6	MSE-Vergleich der Schätzungen des ATE für die beiden Stichprobenumfänge $n = 200$ und $n = 800$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird und das Outcome-Modell mit den Kovariablen $A$ und $L$ spezifiziert wird . . . . .	36
4.7	Schätzungen des ATE und MSE für den Stichprobenumfang $n = 800$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird und das Outcome-Modell mit den Kovariablen $A$ und $L$ spezifiziert wird, mit breiterer Skala . . . . .	37
4.8	Schätzungen des ATE für den Stichprobenumfang $n = 800$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird und das Outcome-Modell mit den Kovariablen $A$ und $L$ spezifiziert wird und zwischen den Variablen $L$ innerhalb einer Gruppe Assoziationen bestehen . . . . .	38

4.9	Schätzungen des ATE für den Stichprobenumfang $n = 800$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird und das Outcome-Modell mit den Kovariablen $A$ und $L$ spezifiziert wird und zwischen allen Variablen $L$ Assoziationen bestehen . . . . .	39
4.10	MSE-Vergleich der Schätzungen des ATE für die Fälle keine Assoziationen, Assoziationen innerhalb einer $L$ -Gruppe und Assoziationen zwischen allen Variablen aus $L$ beim Stichprobenumfang $n = 800$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird und das Outcome-Modell mit den Kovariablen $A$ und $L$ spezifiziert wird . . . . .	40
4.11	Schätzungen des ATE für den Stichprobenumfang $n = 800$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird und das Outcome-Modell nur mit Kovariable $A$ spezifiziert wird . . . . .	41
4.12	MSE-Vergleich der Schätzungen des ATE für die Spezifikation des Outcome-Modells nur mit Kovariable $A$ (Fehlende Terme) und die Spezifikation des Outcome-Modells mit den Kovariablen $A$ und $L$ (Keine fehlenden Terme) beim Stichprobenumfang $n = 800$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ generiert wird . . . . .	42
4.13	Schätzungen des ATE für den Stichprobenumfang $n = 800$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ sowie Interaktionsterme $A : L^{Y^0}$ generiert wird und das Outcome-Modell nur mit einfachen Termen $A$ und $L$ spezifiziert wird . . . . .	43
4.14	Schätzungen des ATE für den Stichprobenumfang $n = 800$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ sowie Interaktionsterme $A : L^{Y^0}$ generiert wird und das Outcome-Modell mit einfachen Termen $A$ und $L$ sowie Interaktionstermen $A : L$ spezifiziert wird . . . . .	44
4.15	MSE-Vergleich der Schätzungen des ATE für die Spezifikation des Outcome-Modells nur mit einfachen Termen $A$ und $L$ (Fehlende Interaktionsterme) und die Spezifikation des Outcome-Modells mit einfachen Termen $A$ und $L$ sowie Interaktionstermen $A : L$ (Keine fehlenden Terme) beim Stichprobenumfang $n = 800$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ sowie Interaktionsterme $A : L^{Y^0}$ generiert wird . . . . .	45



4.16	Schätzungen des ATE für den Stichprobenumfang $n = 800$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ sowie quadratische Terme $(L^{Y^0})^2$ generiert wird und das Outcome-Modell nur mit einfachen Termen $A$ und $L$ spezifiziert wird	46
4.17	Schätzungen des ATE für den Stichprobenumfang $n = 800$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ sowie quadratische Terme $(L^{Y^0})^2$ generiert wird und das Outcome-Modell mit einfachen Termen $A$ und $L$ sowie quadratischen Termen $L^2$ spezifiziert wird	47
4.18	MSE-Vergleich der Schätzungen des ATE für die Spezifikation des Outcome-Modells nur mit einfachen Termen $A$ und $L$ (Fehlende quadratische Terme) und die Spezifikation des Outcome-Modells mit einfachen Termen $A$ und $L$ sowie quadratischen Termen $L^2$ (Keine fehlenden Terme) beim Stichprobenumfang $n = 800$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ sowie quadratische Terme $(L^{Y^0})^2$ generiert wird	48
4.19	Schätzungen des ATE für den Stichprobenumfang $n = 3200$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ , Interaktionsterme $A : L^{Y^0}$ sowie quadratische Terme $(L^{Y^0})^2$ generiert wird und das Outcome-Modell nur mit einfachen Termen $A$ und $L$ spezifiziert wird, und wobei die Treatmentvariable durch einfache Terme $L^{A^0}$ und quadratische Terme $(L^{A^0})^2$ generiert wird und das Exposure-Modell nur mit einfachen Termen $L$ spezifiziert wird	49
4.20	Schätzungen des ATE für den Stichprobenumfang $n = 3200$ ohne Assoziationen zwischen den Variablen aus $L$ , wobei der Outcome durch einfache Terme $A$ und $L^{Y^0}$ , Interaktionsterme $A : L^{Y^0}$ sowie quadratische Terme $(L^{Y^0})^2$ generiert wird und das Outcome-Modell mit einfachen Termen $A$ und $L$ , Interaktionstermen $A : L$ sowie quadratischen Termen $L^2$ spezifiziert wird, und wobei die Treatmentvariable durch einfache Terme $L^{A^0}$ und quadratische Terme $(L^{A^0})^2$ generiert wird und das Exposure-Modell nur mit einfachen Termen $L$ spezifiziert wird	50

4.21 MSE-Vergleich der Schätzungen des ATE für die Spezifikation des Outcome-Modells nur mit einfachen Termen  $A$  und  $L$  (Fehlende Interaktionsterme und quadratische Terme) und die Spezifikation des Outcome-Modells mit einfachen Termen  $A$  und  $L$ , Interaktionstermen  $A : L$  sowie quadratischen Termen  $L^2$  (Keine fehlenden Terme) beim Stichprobenumfang  $n = 3200$  ohne Assoziationen zwischen den Variablen aus  $L$ , wobei der Outcome durch einfache Terme  $A$  und  $L^{Y_0}$ , Interaktionsterme  $A : L^{Y_0}$  sowie quadratische Terme  $(L^{Y_0})^2$  generiert wird, und wobei die Treatmentvariable durch einfache Terme  $L^{A_0}$  und quadratische Terme  $(L^{A_0})^2$  generiert wird und das Exposure-Modell nur mit einfachen Termen  $L$  spezifiziert wird . 51

# Elektronischer Anhang

Der beiliegende USB-Stick enthält die folgenden Dateien:

- Im Ordner 'R-Code' befinden sich die folgenden R-Dateien:
  - Die Datei 'simulate\_CausalEffect.R' enthält die „Hauptfunktion“ (simulate\_CausalEffect()), welche die Simulationen auf die beschriebene Art durchführt.
  - Die Datei 'generate\_data.R' enthält die Funktion, die die Daten nach einem festgelegten Schema und übergebenen Parametern erzeugt. (Unterfunktion zu 'simulate\_CausalEffect()')
  - Die Datei 'estimate\_Effect.R' enthält die Funktion, die aus bestimmten Daten und nach einer bestimmten Methode den Average Treatment Effect schätzt. (Unterfunktion zu 'simulate\_CausalEffect()')
  - Die Datei 'estimate\_simple.R' enthält Funktionen, die den Average Treatment Effect mittels naiver Schätzung, Standardisierung oder IPW aus den Daten schätzen. (Unterfunktion zu 'estimate\_Effect()')
  - Die Datei 'estimate\_tmle.R' enthält die Funktion, die den Average Treatment Effect per TMLE schätzt. (Unterfunktion zu 'estimate\_Effect()')
  - Die Datei 'estimate\_variableselection.R' enthält die Funktion, die den Average Treatment Effect per CTMLE-Variablenselektionsalgorithmus schätzt. (Unterfunktion zu 'estimate\_Effect()')
  - Die Datei 'estimate\_shrinkage.R' enthält die Funktion, die den Average Treatment Effect per CTMLE-Algorithmus mit Shrinkage schätzt. (Unterfunktion zu 'estimate\_Effect()')
  - Die Datei 'estimate\_comp\_shrinkage.R' enthält die Funktion, die den Average Treatment Effect per CTMLE-Algorithmus mit komponentenweisem Shrinkage schätzt. (Unterfunktion zu 'estimate\_Effect()')
  - Die Datei 'compute\_MSE.R' enthält die Funktion, die anhand der Schätzungen und des wahren Average Treatment Effects für jede Methode einen empirischen MSE berechnet (Unterfunktion zu 'simulate\_CausalEffect()')

- Die Datei 'perform\_simulation.R' ruft die Funktion 'simulate\_CausalEffect()' für verschiedene Inputs auf und speichert die Outputs (MSE, Schätzungen)
- Die Datei 'plot\_simulation.R' plottet die Outputs aus 'perform\_simulation.R' auf geeignete Weise
- Im Ordner 'R-Objekte' befinden sich die Outputs von 'perform\_simulation.R', welche die Kennzahlen der durchgeführten Simulationen beinhalten. Anhand des Namens der R-Objekte lässt sich erkennen, welche Simulation jeweils enthalten ist:
  - „n200“, „n800“ oder „n3200“ gibt den Stichprobenumfang an
  - „no“, „insideGr“ oder „all“ gibt die Assoziationsstruktur der Variablen  $L$  an
  - Bildung von Outcome und Treatmentvariable:
    - \* „einf“: nur einfache Terme zur Erzeugung von Treatmentvariable und Outcome
    - \* „inter“: einfache und Interaktionsterme zur Erzeugung des Outcomes sowie einfache Terme zur Erzeugung der Treatmentvariable
    - \* „quadr“: einfache und quadratische Terme zur Erzeugung des Outcomes sowie einfache Terme zur Erzeugung der Treatmentvariable
    - \* „complex“: einfache, Interaktionsterme und quadratische Terme zur Erzeugung des Outcomes sowie einfache und quadratische Terme zur Erzeugung der Treatmentvariable

Jedes R-Objekt ist eine Liste mit zwei Elementen:

- Estimated\_Effects: Matrix der geschätzten Effekte mit  $R = 500$  Zeilen (Anzahl der Wiederholungen) und einer Spalte für jede Schätzmethode
- MSE: Matrix von MSE, Bias und Varianz (3 Spalten) und einer Zeile für jede Schätzmethode

Die Namen der Schätzmethode geben Auskunft darüber, um welche allgemeine Methode es sich handelt und wie das Outcome-Modell spezifiziert wurde:

- „emp“ / „empty“: Nur  $A$  als Kovariable
- „norm“:  $A$  und  $L$  als Kovariablen
- „inter“:  $A$  und  $L$  als Kovariablen sowie die Interaktionsterme  $A : L$
- „quadr“:  $A$  und  $L$  als Kovariablen sowie die quadratischen Terme  $L^2$
- „inter\_quadr“:  $A$  und  $L$  als Kovariablen sowie die Interaktionsterme  $A : L$  und die quadratischen Terme  $L^2$
- Die Masterarbeit als pdf-Datei

# Erklärung der Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

---

Ort, Datum

---

Unterschrift