Algorithmik der Identifikation von Kausalen Effekten in Graphischen Modellen[1]

Benito van der Zander[2]

Abstract: Graphische, kausale Modelle repräsentieren Zufallsvariablen mitsamt ihren gegenseitigen Einflüssen als Graphen, und können die Ergebnisse von Experimenten aus rein beobachteten Daten vorhersagen. Diese Modelle haben große Bedeutung in Forschungsbereichen wie Epidemiologie, der Wirtschaftswissenschaft und der Sozialwissenschaft, in denen randomisierte kontrollierte Studien unmöglich sind oder unethisch wären, jedoch große Datenmengen zur Verfügung stehen. Obwohl graphische, kausale Modelle schon intensiv erforscht wurden, sind die meisten Ergebnisse theoretischer Natur und es fehlen effiziente Algorithmen, um die Modelle in künstlicher Intelligenz oder zur Analyse von Big Data anzuwenden.

In meiner Dissertation habe ich zwei Methoden untersucht, um die kausalen Effekte von Experimenten aus gegebenen beobachteten Daten und dem dazugehörigen graphischen kausalen Model zu berechnen: das Adjustieren für Störfaktoren in nicht-parametrisierten Systemen und die Instrumentvariablenmethode in linearen Systemen. Für beide Ansätze habe ich innovative Polynomialzeitalgorithmen entwickelt; abgesehen von einigen Situationen, für die ich gezeigt habe, dass das Problem der Berechnung NP-vollständig ist. Die vorgeschlagenen algorithmischen Methoden haben die bisher bekannten Verfahren wesentlich verbessert. Sie wurden in der weitverbreiteten Open-Source-Software DAGitty implementieren.

1 Einführung

Die Standardmethode, um die Auswirkungen von etwas wie einer Intervention oder einem Medikament zu untersuchen, ist es ein Experiment durchzuführen, bei dem Subjekte zufällig in eine Behandlungsgruppe und eine Kontrollgruppe aufgeteilt werden. Die Intervention wird nur in der ersten Gruppe durchgeführt, so dass die Effekte in beiden Gruppen miteinander verglichen werden können. Jedoch lassen sich viele wichtige Fragen nicht mit solchen randomisierten kontrollierten Studien beantworten, zum Beispiel „Verursacht Rauchen Lungenkrebs?“, „Verlangsamen nächtliche Ausgangssperren die Ausbreitung von Coronaviren?“, „Stammt die Klimaerwärmung von menschgemachten CO2-Emissionen?“, oder „Führen niedrigere Steuern zu mehr Wirtschaftswachstum?“. Für solche Fragen wären die nötigen experimentellen Studien unethisch, zu teuer, oder geradezu unmöglich. So wäre es ethisch nicht akzeptable Nichtraucher einer Rauchergruppe zuzuweisen; während einer Pandemie kann man nicht abwarten, bis die Effektivität der Maßnahmen ermittelt wurde; und die meisten Wissenschaftler haben nicht mehrere Länder und Planeten zur Verfügung, die sie für ihre Studien kontrollieren könnten.

Daher müssen solche Fragen aus rein beobachteten Daten beantwortet werden. Graphische, kausale Modelle sind eine wichtige Methode um beobachtete Daten entsprechend

[1] Englischer Titel der Dissertation: ”Algorithmics of Identifying Causal Effects in Graphical Models” [vdZ]

[2] Institut für Theoretische Informatik, Universität zu Lübeck, benito@tcs.uni-luebeck.de

zu analysieren, in Gebieten wie der Wirtschaftswissenschaft [AP08, Im14], Sozialwissenschaft [El13] und Epidemiologie [RGL08]. Solche Modelle repräsentieren Zufallsvariablen als Knoten in einem Graph und die Abhängigkeiten zwischen Variablen als Kanten.

Das wichtigste Modell sind azyklische gerichtete Graphen (DAGs), bei denen die Variablen von ihren Eltern abhängen [Pe09]. Ändert sich der Wert einer Variable, so beeinflusst sie direkt die Werte ihrer Kinder, welche wiederum ihre Kinder beeinflussen und so fort. Wir betrachten DAGs mit n Knoten (Zufallsvariablen) $\mathbf{V} = \{V_1, \ldots, V_n\}$. Für jede Variable V_i gibt es die bedingte Wahrscheinlichkeit $P_i(V_i = v_i \mid pa_i)$, dass Variable V_i den Wert v_i in Abhängigkeit der Werte pa_i aller Elternknoten $Pa(V_i)$ annimmt. Diese Wahrscheinlichkeitsverteilung $P_i(V_i = v_i \mid pa_i)$ kann beliebig gewählt werden, und auch der Wertebereich der Zufallsvariablen spielt keine Rolle. Zu beachten ist nur, dass P_i lokal ist, also von anderen Variablen als $V_i \cup Pa(V_i)$ nicht beeinflusst wird. Wir wollen Eigenschaften ermitteln, die für alle solche Wahrscheinlichkeitsverteilungen gelten, und nur von der Struktur des Graphens abhängen[3].

Aus diesen lokalen Verteilungen einzelner Variablen folgt eine Wahrscheinlichkeitsverteilung aller Variablen $P(V_1 = v_1, \ldots, V_n = v_n)$, kurz $P(\mathbf{v})$, als Produkt aller Verteilungen: $P(\mathbf{v}) = \prod_{i=1}^{n} P_i(v_i \mid pa_i)$. Üblicherweise wird der Index i der Faktoren weggelassen und $P_i(v_i \mid pa_i)$ als $P(v_i \mid pa_i)$ geschrieben, denn man kann P_i als Einschränkung der Verteilung P auf die Variable V_i und ihre Eltern betrachten.

Als ein Beispiel kann man mit dem Graph in Abbildung 1(a) untersuchen, ob das Risiko an Diabetes (Variable D) zu erkranken durch verbesserte Schulbildung (Variable *LE* „low education“) gesenkt werden kann [TL11, vdZLT19, RGL08, Kapitel 12]. In diesem Modell gibt es genau drei Ursachen des Diabetesrisikos: Das genetische Diabetesrisikos der Mutter (*MR*), ob sie Diabetes entwickelt hat (*MD*), und die Bildungsstufe (*LE*). Weder *MR* noch *MD* beeinflussen *LE*. Eine vierte Variable, Familieneinkommen (*FI*), beeinflusst *MD* und *LE*, hat jedoch keinen direkten Effekt auf *MR* oder *D*. Sie beeinflusst jedoch *D* über *MD* indirekt. Die zum Graph gehörende Wahrscheinlichkeitsverteilung ist dann $P(fi, mr, md, le, d) = P(fi)P(mr)P(md \mid fi, mr)P(le \mid fi)P(d \mid md, mr)$.

Eine bekannte Anwendung dieser graphischen Modelle, ohne jeglichen Kausalitätsbezug, ist es, dass man bedingte Unabhängigkeiten aus dem Graphen ablesen kann, welche für alle solche faktorisierte Wahrscheinlichkeitsverteilungen gelten. Mengen von Variablen $\mathbf{X}$ und $\mathbf{Y}$ sind nämlich unabhängig, also $P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y})$, wenn es keinen nicht-blockierten Pfad von einem Knoten $X \in \mathbf{X}$ zu einem Knoten $Y \in \mathbf{Y}$ gibt. Ein *Pfad* ist eine Sequenz von Knoten $V_{i_1}, V_{i_2}, \ldots, V_{i_k}$, so dass V_{i_j} und $V_{i_{j+1}}$ durch eine Kante verbunden sind. Der Pfad ist blockiert (von der leeren Menge), wenn er $V_{i_{j-1}} \to V_{i_j} \leftarrow V_{i_{j+1}}$ enthält[4]. In dem Fall wird der Knoten V_{i_j} *Kollider* (auf diesem Pfad) genannt. Im Beispiel 1 (a) sind *FI* und *MR* sowie *LR* und *MR* unabhängig, jedoch keine anderen Paare von Variablen[5].

[3] präziser gesagt, betrachtet der erste Teil der Dissertation allgemeine Verteilungen und der zweite Teil einen Spezialfall der Verteilungen in linearen Modellen.

[4] Die in einem nicht-blockierten Pfad erlaubten Kantenfolgen sind also $V_{i_{j-1}} \to V_{i_j} \to V_{i_{j+1}}$, $V_{i_{j-1}} \leftarrow V_{i_j} \leftarrow V_{i_{j+1}}$ und $V_{i_{j-1}} \leftarrow V_{i_j} \to V_{i_{j+1}}$.

[5] Der Zusammenhang zwischen Pfaden und Unabhängigkeit gilt nun erstmal nur in eine Richtung. Für alle entsprechend faktorisierbaren Wahrscheinlichkeitsverteilung sind *FI,MR* sowie *LE,MR* unabhängig. Für alle an-

Für die *bedingte Unabhängigkeit* ist zusätzlich eine Menge von Variablen $\mathbf{Z}$ gegeben. Die Variablenmengen $\mathbf{X}$ und $\mathbf{Y}$ sind bedingt unabhängig gegeben $\mathbf{Z}$, also $P(\mathbf{X},\mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Y} \mid \mathbf{Z})$, wenn $\mathbf{Z}$ alle Pfade zwischen $X \in \mathbf{X}$ und $Y \in \mathbf{Y}$ blockiert. Die Menge $\mathbf{Z}$ invertiert nun die Definition der Blockiertheit. Der Pfad ist blockiert, wenn er für ein $V_{i_j} \in \mathbf{Z}$ die Kanten $V_{i_{j-1}} \to V_{i_j} \to V_{i_{j+1}}$, $V_{i_{j-1}} \leftarrow V_{i_j} \leftarrow V_{i_{j+1}}$ oder $V_{i_{j-1}} \leftarrow V_{i_j} \to V_{i_{j+1}}$ enthält. Ebenso ist er blockiert, wenn er $V_{i_{j-1}} \to V_{i_j} \leftarrow V_{i_{j+1}}$ enthält und *kein* Nachfahre von V_{i_j} in $\mathbf{Z}$ ist.

Im Beispiel 1 (e) kann also mit $\mathbf{Z} = \{MD\}$ der Kollider *MD geöffnet* werden, so dass der Pfad $FI \to MD \leftarrow MR$ von *MD* nicht blockiert ist. Somit sind *FI* und *MR* nicht unabhängig gegeben *MD*. Es ergibt auch intuitiv Sinn, am Anfang können sich *FI* und *MR* über *MD* nicht beeinflussen, da *MD* nur eine gemeinsame Folge von beiden ist. Ist der Wert von *MD* jedoch fest, so kann die Kenntnis von *FI* zu Wissen über *MR* führen, oder umgekehrt. Zum Beispiel ist das Einkommen hoch und die Mutter hat Diabetes, so ist es wahrscheinlicher, dass sie ein hohes genetisches Risiko besitzt. Ist das Einkommen dagegen niedrig und die Mutter hat nicht Diabetes, hat sie vermutlich auch kein besonderes hohes genetisches Risiko[6].

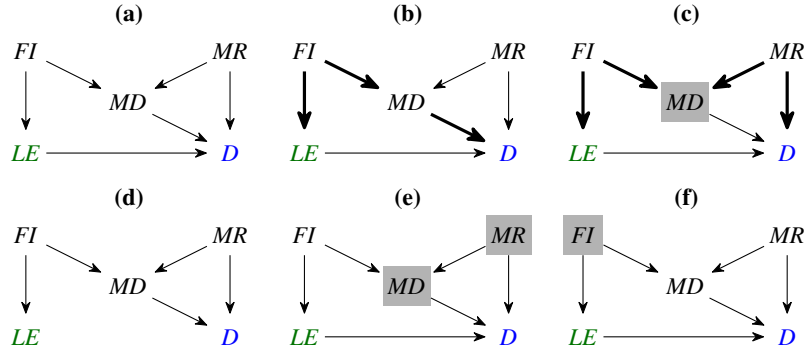

Abb. 1: Graphisches Modell, welches den Einfluss von Bildung (*LE*) auf das Diabetesrisiko (*D*) modelliert [TL11, vdZLT19, vdZ, RGL08, Kapitel 12]. Als Kovariablen gibt es das Familieneinkommen (*FI*), das genetische Diabetesrisikos der Mutter (*MR*), und ob die Mutter Diabetes hat (*MD*). Fall (a) zeigt den Graphen selbst. Fälle (b) und (c) zeigen jeweils einen nicht blockierten Pfad zwischen *LE* und *D*. Fall (d) zeigt den Graphen ohne gerichtete Pfade von *LE* zu *D*. Fälle (e) und (f) zeigen, wie man alle nicht-kausalen Pfade zwischen (*LE*) und (*D*) blockieren kann.

deren Paare gibt es Wahrscheinlichkeitsverteilungen, so dass diese nicht unabhängig sind. Die anderen Paaren sind also nicht *für alle* Wahrscheinlichkeitsverteilungen unabhängig. Es kann jedoch auch Wahrscheinlichkeitsverteilungen geben, für die alle Variablen unabhängig sind. Da die einzelnen Faktoren P_i von $P(\mathbf{v})$ beliebig gewählt werden können, können sie auch so gewählt werden, dass die Verteilung P_i nicht von den Eltern pa_i abhängt. Solche Verteilungen werden „non faithful“ genannt, spielen aber im Folgenden keine Rolle.

[6] Eine solche Korrelation von Ursachen nach Festhalten einer Folge kennen Statistiker auch als Berkson-Paradox.

2 Identifizieren des kausalen Effektes mittels Adjustierung

2.1 Der kausale Effekt

Die grundlegende Idee der kausalen Modellierung ist nun, dass aus der Faktorisierung der Verteilung $P(\mathbf{v})$ der Effekt eines Experimentes berechnet werden kann, ohne das Experiment tatsächlich durchzuführen[7]. Würde man ein Experiment durchführen, welches Variablen $\mathbf{X} \subseteq \mathbf{V}$ auf Werte $\mathbf{x}$ setzt, wären die Werte von $\mathbf{X}$ konstant $\mathbf{x}$, somit würden sie nicht mehr von ihren Eltern abhängen. Dies entspricht einfach dem Entfernen der Faktoren $P_{X_i}(x_i \mid pa_{X_i})$ für alle $X_i \in \mathbf{X}$ aus der Gesamtverteilung. Der Gesamteffekt der Variablen $\mathbf{X}$ ist also eine neue Wahrscheinlichkeitsverteilung $P(\mathbf{v} \mid do(\mathbf{x}))$, die Wahrscheinlichkeit, dass alle Variablen Werte $\mathbf{v}$ haben, nachdem ein Experiment Variablen $\mathbf{X}$ auf Werte $\mathbf{x}$ gesetzt hat:

$$P(\mathbf{v} \mid do(\mathbf{x})) = \begin{cases} \prod\limits_{V_i \in \mathbf{V} \setminus \mathbf{X}} P(v_i \mid pa_i) & \text{für } \mathbf{v} \text{ konsistent zu } \mathbf{x}, \\ 0 & \text{sonst,} \end{cases} \tag{1}$$

wobei konsistent bedeutet, dass die Werte $\mathbf{v}$ dieselben Werte für die Variablen $\mathbf{X}$ enthalten wie $\mathbf{x}$. Man kann auch den kausalen Effekt $P(\mathbf{y} \mid do(\mathbf{x}))$ auf eine Teilmenge $\mathbf{Y} \subseteq \mathbf{V}$ betrachten, der sich aus der Verteilung aller Variablen durch Marginalisieren berechnen lässt: $P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{v} \setminus \mathbf{y}} P(\mathbf{v} \mid do(\mathbf{x}))$. Dabei summiert $\sum_{\mathbf{v} \setminus \mathbf{y}}$ über die möglichen Werte aller Variablen $\mathbf{V} \setminus \mathbf{Y}$. Diese Verteilung entspricht ungefähr der Verteilung eines DAGs in dem alle in $\mathbf{X}$ eingehenden Kanten entfernt wurden.

Sind alle Faktoren $P(v_i \mid pa_i)$ bekannt, so lässt sich der kausale Effekt $P(\mathbf{v} \mid do(\mathbf{x}))$ nach Formel (1) leicht berechnen. In der Praxis, kennt man nicht alle Faktoren, und die Variablen sind partitioniert in eine Menge von beobachteten, „observed“ Variablen $\mathbf{O}$ und eine Menge von unbeobachteten, latenten Variablen $\mathbf{V} \setminus \mathbf{O}$. Die Faktoren der unbeobachteten Variablen dürfen bei der Berechnung nicht verwendet werden. Im Beispiel würde man auf jeden Fall *LE* und *D* kennen, da diese Variablen untersuchten werden, und vermutlich wurde auch *MD* abgefragt. *FI* könnte Datenschutzrechtlich problematisch zu ermitteln sein, und daher zu den unbeobachteten Variablen zählen. *MR* ist schwer zu bestimmen, selbst wenn die Gene der Mutter sequenziert wurden, ist es erstmal unbekannt, welche Gene für Diabetes relevant sind.

2.2 Adjustierung

Das Ziel der Dissertation ist es nun den kausalen Effekt $P(\mathbf{y} \mid do(\mathbf{x}))$ zu identifizieren. Dabei ist der Graph und Variablenmengen $\mathbf{X}, \mathbf{Y}, \mathbf{O}$ gegeben, und wir suchen eine Formel, die äquivalent zu $P(\mathbf{y} \mid do(\mathbf{x}))$ ist und nur Wahrscheinlichkeitsverteilungen über den beobachteten Variablen und keinen *do*-Operator enthält. Wir betrachten nicht beliebige Formeln,

[7] Woraus Judea Pearl seine Kausalitätstheorie entwickelt hat, für die er einen Turing-Award gewonnen hat.

sondern im ersten Teil ausschließlich Formeln, die den Effekt mittels Adjustierung identifizieren. Hierbei wird eine Menge von (Stör)variablen $\mathbf{Z} \subseteq \mathbf{O} \subseteq \mathbf{V}$ gesucht, ein sogenanntes *Adjustment Set* $\mathbf{Z}$, für die gilt:

$$P(\mathbf{Y} = \mathbf{y} \mid do(\mathbf{X} = \mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z}). \tag{2}$$

Diese Formel berechnet den Erwartungswert von der bedingten Wahrscheinlichkeit $P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ über alle möglichen Werte der Variablen $\mathbf{Z}$. Adjustment Sets sind die am häufigsten verwendete Methode zur Bestimmung des kausalen Effekts, weil sie gut erforschte statistische Eigenschaften haben [Va09, GK17]. Im Wesentlichen werden die möglichen Werte in Teilpopulationen aufgeteilt, so dass innerhalb jeder Teilpopulationen die rein beobachtete Wahrscheinlichkeit $P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ der durch ein Experiment entstehenden Wahrscheinlichkeit $P(\mathbf{Y} = \mathbf{y} \mid do(\mathbf{X} = \mathbf{x}), \mathbf{Z} = \mathbf{z})$ entspricht.

Das Standardverfahren zum Wählen der Menge $\mathbf{Z}$ ist Pearls *Back-Door-Kriterium* [Pe93, Pe09]. Das Back-Door Kriterium sagt, $\mathbf{Z}$ darf keinen von Nachfahren $\mathbf{X}$ enthalten und muss alle Pfade, deren erste Kante auf einen Knoten in $\mathbf{X}$ zeigt, blockieren[8]. In den Graphen von Abbildung 1 und Abbildung 2 (a) können mit dem Back-Door Kriterium alle Adjustment Sets gefunden werden. Im Graphen 2 (b) funktioniert das Kriterium jedoch nicht, denn es gibt nur ein Adjustment Set bestehend aus Nachfahren von X_1.

Ein vollständiges Adjustment-Kriterium, welches von einer Menge $\mathbf{Z}$ genau dann erfüllt ist, wenn $\mathbf{Z}$ ein Adjustment Set ist, wurde erst von [SVR10] entdeckt. Dieses Kriterium sagt, $\mathbf{Z}$ darf keinen Nachfahren eines Knotens auf einem echten, kausalen Pfad von $\mathbf{X}$ zu $\mathbf{Y}$ enthalten[9], und muss alle echten, nicht-kausalen Pfade zwischen $\mathbf{X}$ und $\mathbf{Y}$ blockieren. Ein *echter* Pfad ist ein Pfad, der nur einen Knoten von $\mathbf{X}$ enthält (d.h. als Startknoten); ein *kausaler* Pfad ist ein gerichteter Pfad, der nur von $\mathbf{X}$ weg gerichtete Kanten enthält; und ein *nicht-kausaler* Pfad ist ein Pfad, der nicht kausal ist.

Die zu untersuchende, offene Frage ist, wie man effizient überprüft, ob das Kriterium von einer Menge erfüllt ist, und wie man eine solche Menge findet. Viele Implementierungen von Algorithmen für kausalen Graphen [Ky98, Ka12] enumerieren alle möglichen Mengen oder Pfade, und testen dann auf triviale Weise, ob ein Kriterium erfüllt ist. Dies ergibt jedoch eine exponentielle Laufzeit und ist für große Graphen nicht praktikabel.

Dafür haben wir ein neues vollständiges Kriterium eingeführt, welches wir *Konstruktives Back-Door Kriterium* nennen und welches das Adjustment-Kriterium ändert, damit der Begriff des nicht-kausalen Pfades nicht mehr vorkommt. Die erste Bedingung verbietet weiterhin Nachfahren von echt kausalen Pfaden. Die zweite Bedingung ist, dass $\mathbf{Z}$ alle Pfade im Echten Back-Door-Graphen blockieren muss. Der *Echte Back-Door-Graphen* wird konstruiert, indem jeweils die erste Kante aller kausalen Pfade von $\mathbf{X}$ zu $\mathbf{Y}$ gelöscht wird. Eine Menge $\mathbf{Z}$ blockiert dann alle nicht-kausalen Pfade zwischen $\mathbf{X}$ und $\mathbf{Y}$, genau dann, wenn sie alle Pfade zwischen $\mathbf{X}$ und $\mathbf{Y}$ im echten Back-Door-Graphen blockiert.

[8] Es gibt auch eine äquivalente Formulierung mittels einem Back-Door-Graphen, in dem alle aus $\mathbf{X}$ ausgehenden Kanten gelöscht werden.

[9] Nachfahren des Startknotens sind erlaubt, Nachfahren des Endknoten eines Pfades nicht.

Im Beispiel von Abbildung 1 betrachten wir $\mathbf{X} = \{LE\}$ und $\mathbf{Y} = \{D\}$. Dann liegt nur $\{D\}$ auf kausalen Pfaden und der (echte) Back-Door-Graph ist der Graph von Fall (d). Blockiert werden muss der in (b) markierte Pfad. Dies kann entweder mit dem Adjustment Set $\{FI\}$ wie in Fall (f) erfolgen, oder mittels Knoten *MD*. Da *MD* wiederum den Pfad aus Fall (c) öffnet, ist ein zweites mögliche Adjustment Set nur $\{MD, MR\}$. Im Beispiel von Abbildung 2 (b), ist der (echte) Back-Door-Graph in Fall (b') gezeigt. Hier müssen alle Pfade von Y_2 zu $\{X_1, X_2\}$ blockiert werden, so dass das einzige Adjustment Set $\{Z_1, Z_2\}$ ist.

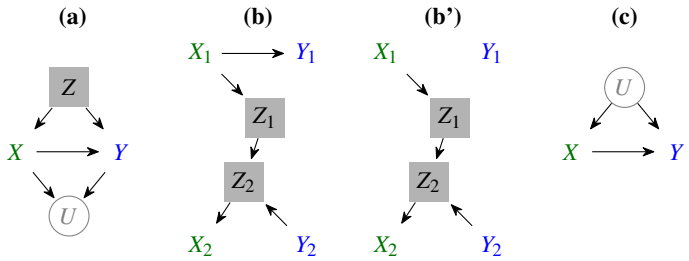

Abb. 2: Im Fall (a) kann der kausale Effekt von $\mathbf{X} = \{X\}$ auf $\mathbf{Y} = \{Y\}$ leicht mit den Eltern $\{Z\}$ von X als Adjustment Set identifiziert werden. Im Fall (b) kann der kausale Effekt von $\mathbf{X} = \{X_1, X_2\}$ auf $\mathbf{Y} = \{Y_1, Y_2\}$ mit dem Adjustment Set $\mathbf{Z} = \{Z_1, Z_2\}$ identifiziert werden. Graph (b') zeigt den dazugehörigen echten Back-Door-Graphen. Z_2 muss im Adjustment Set sein, um den Pfad zwischen X_2, Y_2 zu blockieren. Z_1 muss dann hinzugefügt werden, weil der Kollider Z_2 auf dem Pfad zu X_1 geöffnet wurde. Im Fall (c) ist es unmöglich den kausalen Effekt zu identifizieren. Das Adjustment Set müsste den Pfad über U blockieren, U ist jedoch als nicht beobachtet markiert und darf daher nicht im Adjustment Set sein.

Übrig bleibt das Problem eine Menge zu finden, welche alle Pfade in einem kausalen (Back-Door)-Graphen blockiert. Standardgraphtheoretische Algorithmen lassen sich hierfür nicht verwenden, weil die Definition mit Kollidern, welche erst blockiert sind, und dann geöffnet werden können, ungewöhnlich ist. Auch für kausale Graphen gab es zwar schon bekannte Algorithmen, welche überprüfen, ob eine Menge alle Pfade blockiert [Sh98], eine solche blockierende Menge finden, eine blockierende, minimale Menge finden [TPP98], eine Minimum-Menge finden [ADC96] oder alle diese Mengen aufzählen [TL11]. Wir mussten jedoch die Algorithmen erweitern, so dass sie Mengen $\mathbf{Z}$ unter der Einschränkung $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ für beliebige Mengen $\mathbf{I} \subseteq \mathbf{R}$ finden können. Zudem ist es uns gelungen, die Laufzeit für das Finden einer minimalen Menge von $\mathcal{O}(n^2)$ zu $\mathcal{O}(n+m)$ bezüglich der Zahl der Knoten n und Kanten m zu verbessern.

2.3 Adjustierung in generalisierten kausalen Modellen

DAGs sind die grundlegenden kausalen Modelle, aber sie setzen voraus, dass man alle relevanten Variablen und die Richtung aller Kanten kennt. In vielen Situationen fehlen diese Informationen jedoch. Daher haben wir alle unsere Ergebnisse und Algorithmen zu zwei weiteren Klassen von Graphen verallgemeinert: RCGs und MAGs.

Chain Graphen (CGs) enthalten zusätzlich ungerichtete Kanten, welche besagen, dass zwei Variablen einander direkt beeinflussen, jedoch unbekannt ist, welche der Variablen der Elternknoten und welche der Kindknoten ist [LW89, Fr90]. Ein Chain Graph repräsentiert eine Klasse von DAGs, und zwar alle DAGs, die durch das Ersetzen aller ungerichteten Kanten im CG durch gerichtete Kanten entstehen. Zudem müssen alle von einem CG repräsentierten DAGs dieselben (bedingten) Unabhängigkeiten haben, so dass sie statistisch ununterscheidbar sind. Nicht jede Ersetzung von ungerichteten Kanten ist erlaubt, zum Beispiel darf aus einem ungerichteten Zyklus kein gerichteter Zyklus entstehen, da sonst kein DAG entstünde. Auch darf bei der Ersetzung auf keinem Pfad zwischen zwei Variablen ein Kollider entstehen, außer es gibt es anderen Pfad der diese Variablen verbindet. Restricted Chain Graphs (RCGs) nennen wir diejenigen Chain Graphen, die mindestens einen DAG repräsentieren. Eine Unterklasse von RCGs sind CPDAGs, welche oft verwendet werden, da sie gut aus beobachteten Daten gelernt werden können.

Modelle mit unbekannten Variablen können durch einen *Ancestral Graph* dargestellt werden [RS02], der alle DAGs repräsentiert, welche dieselben Vorfahrrelationen wie der Ancestral Graph haben. Zum Beispiel, besagen die Vorfahrrelationen in Abbildung 1, dass *FI* ein Vorfahr von *D* und kein Vorfahr von *MR* ist. Wird der Graph als Ancestralgraph betrachtet, so repräsentiert er auch jeden DAG, der beispielsweise den Graph mit einer Variable erweitert, die zwei der bisherigen Variablen als Kinder hat. Es darf jedoch keine Variable eingefügt werden, die ein Kind von *FI* und ein Elternteil von *MR* ist, da diese Variable die Vorfahrrelationen ändern würde, indem *FI* ein Vorfahr von *MR* würde.

Maximale Ancestral Graphen (MAGs) sind Ancestral Graphen, bei denen alle Paare von Variablen, zwischen denen sich nicht alle Pfade blockieren lassen, direkt durch eine Kante verbunden sind. Wird der Graph aus Abbildung 1 als MAG betrachtet, dürfte man eine Variable mit Kindern *MR* und *D* einfügen. Eine Variable *U* mit Kindern *LE* und *D* darf jedoch nicht eingefügt werden. Dann gäbe es nämlich Pfade $FI \rightarrow LE \leftarrow U \rightarrow D$ und $FI \rightarrow LE \rightarrow D$. Der zweite Pfad kann nur an Knoten *LE* blockiert werden, wodurch sich der erste Pfad öffnen würde, der dann nicht mehr mit den bisherigen Variablen blockiert werden kann. Diese Variable *U* ließe sich jedoch einfügen, gäbe es eine Kante $FI \rightarrow D$. Anders als DAGs können MAGs auch Marginalisierung von Verteilung repräsentieren.

3 Identifizierung in linearen Modellen (SEMs)

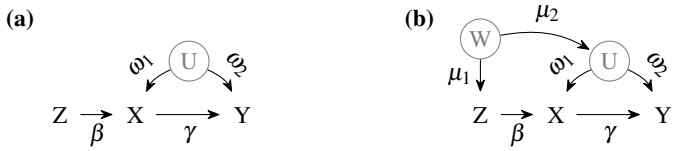

Abb. 3: (a): Das klassische Instrumentvariablenmodell. (b): *Z* ist eine bedingte Instrumentvariable. *U* and *W* sind unbeobachtete Variablen.

Sehr häufig werden kausale Modelle unter der Annahme verwendet, dass alle Variablen reelle Zahlen sind und alle Abhängigkeiten zwischen den Variablen linear sind [Bo89, Du75], so genannte Structural Equation Models (SEMs).

Zum Beispiel, repräsentiert der Graph in Abbildung 3 (a), vier Zufallszahlen mit je einer linearen Gleichung: $Z = \varepsilon_Z$; $U = \varepsilon_U$; $X = \beta Z + \omega_1 U + \varepsilon_X$; $Y = \gamma X + \omega_2 U + \varepsilon_Y$.
Jede Variable hängt von ihren Eltern und unabhängig gleichverteilten, zufälligen Errortermen $\varepsilon_\cdot$ ab[10]. Dies bedeutet beispielsweise, ändert man den Wert von Z um 1, dann ändert sich der (Mittel-)Wert von X um β. Das Modell ist kausal, ändert man den Wert von X, dann ändert sich der Wert von Z nicht. Die Abhängigkeiten der Variablen sind deterministisch, durch Addition der Zufallszahlen $\varepsilon_\cdot$ entstehen jedoch Zufallsvariablen. Geht man davon aus, alle Errorterms sind Gaußverteilt, dann sind die Faktoren $P_i(V_i \mid pa_i)$ in der Wahrscheinlichkeitsverteilung ebenfalls Gaußsche Normalverteilungen.

Aus diesem Modell lassen sich die Korrelation/Kovarianzen zwischen den Variablen quantitativ bestimmen, indem man die Parameter entlang der Pfade multipliziert und alle Pfade addiert. So ist die Kovarianz $\mathrm{Cov}(Z,X) = \beta$; $\mathrm{Cov}(Z,Y) = \beta\gamma$; und $\mathrm{Cov}(X,Y) = \gamma + \omega_1\omega_2$.

Die Kovarianzen zwischen allen Paaren von Variablen, sind dabei die beobachteten Daten. Das Ziel der Identifizierung in SEMs ist es die Parameter der Gleichungen aus den Kovarianzen zu ermitteln. Zum Beispiel folgt hier: $\gamma = \frac{\beta\gamma}{\beta} = \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(X,Z)}$. Man sagt, Z ist eine *Instrumentvariable* zur Identifizierung des direkten kausalen Effekts von X auf Y.

Auf dieselbe Weise kann man den Effekt in Abbildung 3 (b) als $\gamma = \frac{\beta\gamma}{\beta} = \frac{\mathrm{Cov}(Y,Z|W)}{\mathrm{Cov}(X,Z|W)}$ bestimmen. Es müssen jedoch die mit W bedingten Kovarianzen verwendet werden, um den Pfad über W zu blockieren, damit das Ergebnis nicht von $\mu_1\mu_2$ verfälscht wird.

Es gibt ein (nicht vollständiges) Kriterium, um zu entscheiden, ob eine Variable Z eine (bedingte) Instrumentvariable zur Identifizierung des Effekts von X auf Y ist [Pe01]: und zwar muss es eine Menge $\mathbf{W}$ geben, die keine Nachfahren von Y enthält, so dass es einen nicht von $\mathbf{W}$ blockierten Pfad von Z zu X gibt und alle Pfade von Z zu Y ohne die Kante $X \to Y$ von $\mathbf{W}$ blockiert sind.

Wir haben dieses Kriterium untersucht, und gezeigt, dass es ein NP-vollständiges Problem ist $\mathbf{W}$ zu finden. Daraus folgt, gegeben X, Y und Z, so ist es auch NP-vollständig zu entscheiden, ob Z eine bedingte Instrumentvariable für den Effekt von X auf Y ist.

Trotzdem ist es, gegeben X und Y, in $\mathcal{O}(n(n+m))$ Polynomialzeit möglich, eine bedingte Instrumentvariable Z für den Effekt von X auf Y und eine Menge $\mathbf{W}$ zu finden. Hierzu haben wir gezeigt: (1) Existiert eine bedingte Instrumentvariable Y für X auf Y, so existiert auch eine (möglicherweise andere) bedingte Instrumentvariable Y, deren Menge $\mathbf{W}$ ausschließlich aus Vorfahren von Y und Z besteht. (2) Besteht $\mathbf{W}$ aus Vorfahren von Y und Z, so kann $\mathbf{W}$ in Zeit $\mathcal{O}(n+m)$ gefunden werden. Obwohl $\mathbf{W}$ keine minimale Menge sein muss, kann unser Algorithmus zum Finden von minimalen Mengen aus der vorherigen Sektion dazu wiederverwendet werden.

Schließlich haben wir weitere Arten von Instrumentvariablen untersucht, einzelne Instrumentvariablen, die sich ähnlich verhalten, und Instrumental Sets [Br10], bei denen mehrere Instrumentvariablen kombiniert werden. Instrumental Sets mit leerer Menge $\mathbf{W} = \emptyset$ lassen

[10] Alternativ wird häufig die unbeobachtete Variable U in den Gleichungen weggelassen. D.h. $Z = \varepsilon_Z$; $X = \beta Z + \varepsilon_X$; $Y = \gamma X + \varepsilon_Y$, mit der zusätzlichen Bedingung, dass die Kovarianz zwischen ε_X und ε_Y gleich $\omega_1\omega_2$ ist.

sich in Polynomialzeit finden, aber im Allgemeinen ist das Finden von Instrumental Sets NP-vollständig.

4 Fazit

Die Forschung der Dissertation ergab die ersten effizienten Algorithmen, die ein Adjustment Set oder ein minimales Adjustment Set finden können, genau dann wenn ein solches existiert. Es ergab auch das erste vollständige Kriterium zur Charakterisierung von Adjustment Sets in MAGs und RCGs. Zudem war es die erste genauere Untersuchung der Komplexität von Instrumentvariablen mit einem Algorithmus zum effizienten Finden von bedingten Instrumentvariablen. Die Algorithmen ermöglichen die automatische Analyse kausaler Zusammenhänge. Sie sind praktisch implementierbar, und wurden in der weitverbreiteten Open-Source-Software DAGitty veröffentlicht.

Literaturverzeichnis

[ADC96] Acid, Silvia; De Campos, Luis M: An algorithm for finding minimum d-separating sets in belief networks. In: Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., S. 3–10, 1996.

[AP08] Angrist, Joshua D.; Pischke, Jörn-Steffen: Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, 2008.

[Bo89] Bollen, Kenneth A: Structural equations with latent variables. John Wiley & Sons, 1989.

[Br10] Brito, Carlos: Instrumental sets. Heuristics, Probability and Causality. A Tribute to Judea Pearl, S. 295–307, 2010.

[Du75] Duncan, Otis Dudley: Introduction to structural equation models. Academic Press, 1975.

[El13] Elwert, Felix: Graphical Causal Models. In: Handbook of Causal Analysis for Social Research, Handbooks of Sociology and Social Research, S. 245–273. Springer, 2013.

[Fr90] Frydenberg, Morten: The chain graph Markov property. Scandinavian Journal of Statistics, 17:333–353, 1990.

[GK17] Glynn, Adam; Kashin, Konstantin: Front-door Versus Back-door Adjustment with Unmeasured Confounding: Bias Formulas for Front-door and Hybrid Adjustments with Application to a Job Training Program. Journal of the American Statistical Association, 2017.

[Im14] Imbens, Guido: Instrumental Variables: An Econometrician's Perspective. Statistical Science, 29(3):323–358, 2014.

[Ka12] Kalisch, Markus; Mächler, Martin; Colombo, Diego; Maathuis, Marloes; Bühlmann, Peter: Causal Inference Using Graphical Models with the R Package pcalg. Journal of Statistical Software, 47(11):, 2012.

[Ky98] Kyono, Trent Mamoru: Commentator: A Front-End User-Interface Module for Graphical and Structural Equation Modeling. Bericht R-364, University of California, Los Angeles, 1998.

[LW89] Lauritzen, Steffen; Wermuth, Nanny: Graphical models for associations between variables, some of which are qualitative and some quantitative. The Annals of Statistics, 17:31–57, 1989.

[Pe93] Pearl, Judea: Comment: Graphical models, causality and intervention. Statistical Science, 8:266–269, 1993.

[Pe01] Pearl, Judea: Parameter identification: A new perspective. Bericht R-276, UCLA, 2001.

[Pe09] Pearl, Judea: Causality. Cambridge University Press, 2009.

[RGL08] Rothman, Kenneth J.; Greenland, Sander; Lash, Timothy L.: Modern Epidemiology. Wolters Kluwer, 2008.

[RS02] Richardson, Thomas; Spirtes, Peter: Ancestral Graph Markov Models. Annals of Statistics, 30:927–1223, 2002.

[Sh98] Shachter, Ross D.: Bayes-Ball: The Rational Pastime. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, S. 480–487, 1998.

[SVR10] Shpitser, Ilya; VanderWeele, Tyler; Robins, James: On the Validity of Covariate Adjustment for Estimating Causal Effects. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. AUAI Press, S. 527–536, 2010.

[TL11] Textor, Johannes; Liśkiewicz, Maciej: Adjustment Criteria in Causal Diagrams: An Algorithmic Perspective. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence. AUAI Press, S. 681–688, 2011.

[TPP98] Tian, Jin; Paz, Azaria; Pearl, Judea: Finding Minimal D-separators. Bericht R-254, University of California, Los Angeles, 1998.

[Va09] VanderWeele, Tyler J.: On the relative nature of overadjustment and unnecessary adjustment. Epidemiology, 20(4):496–499, 7 2009.

[vdZ] van der Zander, Benito: Algorithmics of Identifying Causal Effects in Graphical Models.

[vdZLT19] van der Zander, Benito; Liśkiewicz, Maciej; Textor, Johannes: Separators and Adjustment Sets in Causal Graphs: Complete Criteria and an Algorithmic Framework. Artificial Intelligence, 270:1–40, 2019.

Benito van der Zander hat seinen Bachelor-Abschluss über Ereignisprotokolle in Düsseldorf gemacht, seinen Master-Abschluss über Computer Vision in Aachen, und zuletzt in Lübeck über Kausalität promoviert. Während der Zeit der Promotion forschte er für ein DFG-Projekt und betreute Übungsgruppen am Institut für Theoretische Informatik. Bereits als Schüler hatte er Vorlesungen an der Universität besucht, eigene Software verkauft, wurde Bundessieger beim Bundeswettbewerb Informatik und errang ein paar Medaillen bei europäischen Informatikolympiaden. Er programmiert aktiv an Open-Source-Projekten: DAGitty mit den Algorithmen der Dissertation, Bibliothek-App VideLibri zur Literaturverwaltung von ausgeliehenen Büchern, LaTeX-Editor TeXstudio, sowie das Webseiten-Automatisierungstool und XQuery-Interpreter Xidel.