Restricted structural equation models for causal inference

Author(s):
Peters, Jonas Martin

Publication Date:
2012

Permanent Link:

https://doi.org/10.3929/ethz-a-007597940 →

Diss. ETH No. 20756

Restricted Structural Equation Models for Causal Inference

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
JONAS MARTIN PETERS

Dipl.-Math. University of Heidelberg
born May 28, 1984
citizen of Germany

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Bernhard Schölkopf, co-examiner
PD Dr. Dominik Janzing, co-examiner

2012

Abstract

Causal inference tries to solve the following problem: given i.i.d. data from a joint distribution, one tries to infer the underlying causal DAG (directed acyclic graph), in which each node represents one of the observed variables.

For approaching this problem, we have to make assumptions that connect the causal graph with the joint distribution. Independence-based methods like the PC algorithm assume the causal Markov condition and faithfulness. These two conditions relate conditional independences and the graph structure; this allows to infer properties of the graph by testing for conditional independences in the joint distribution. Independence-based methods encounter the following difficulties: (1) One can discover causal structures only up to Markov equivalence classes, in particular one cannot distinguish between $X \to Y$ and $Y \to X$. (2) In practice, conditional independence testing is difficult. Especially when the conditioning set is large, their power is often relatively low. (3) When the data come fron a non-faithful distribution, the results may be wrong, but the user does not realize it. Also, when the set of variables is causally insufficient, i.e. some important variables have not been observed, those methods may draw wrong conclusions.

In structural equation models (SEMs) each variable X_j is a function of a set of nodes $\mathbf{PA}_j$ and some noise variable N_j:

$$X_j = f_j(\mathbf{PA}_j, N_j)\,, \qquad j = 1, \ldots, p$$

where the N_j are jointly independent. The corresponding graph is obtained by drawing directed arrows from each variable in $\mathbf{PA}_j$ to X_j (the $\mathbf{PA}_j$ become parents of X_j). In this form, SEMs are too general to be used for structure learning. Given a distribution, we can find for each DAG with respect to which the distribution is Markov to a

corresponding SEM. This changes, however, if we consider *restricted* SEMs, in which some combinations of function and the distribution of noise and parents are excluded. In Gaussian SEMs with linear functions and additive noise, for example, the graph can be identified from the joint distribution again up to Markov equivalence classes (assuming faithfulness). This, however, constitutes a somewhat exceptional case. If the functions are linear and the noise is non-Gaussian, the DAG becomes fully identifiable. In this thesis we present alternative directions of deviating from the linear Gaussian case: (i) apart from few exceptions, identifiability also holds for non-linear functions and arbitrarily distributed additive noise. And (ii), if we require all noise variables to have the same variances, again, the DAG can be recovered from the joint distribution. We also present restricted SEMs for discrete variables with similar identifiability results. Moreover, we apply restricted SEMs to time series data. We further investigate whether it is possible to distinguish between the cases "X is causing Y", "Y is causing X" and "both variables are caused by a third unobserved variable" (throughout this work we call a common cause a *confounder*).

From our point of view, SEM-based causal inference and the restriction of the function class leads to the following advantages: (1) We can identify causal relationships even within an equivalence class. (2) Fitting a model with additive noise is easier than general conditional independence testing. (3) We do not require faithfulness. (4) If the model assumptions are violated (e.g. the data do not follow an additive noise model or there are hidden common causes), the method is able to output "I do not know" instead of giving wrong answers.

For all of the proposed identifiability results we present practical methods and apply them to simulated and real data sets.

Zusammenfassung

Kausale Inferenz beschäftigt sich mit dem folgenden Problem: Seien unabhängig und identisch verteilte Daten einer gemeinsamen Verteilung gegeben. Das Ziel der kausalen Inferenz liegt darin, den zugrundeliegende kausalen Graphen zu schätzen, dessen Knoten die Zufallsvariablen repräsentieren. Wir nehmen dabei an, dass der Graph gerichtete Kanten, aber keine Zykel enthält (directed acyclic graph).

Um dieses Problem anzugehen, müssen wir Annahmen treffen, die den kausalen Graphen mit der gemeinsamen Verteilung in Verbindung bringen. Sogenannte unabhängigkeitsbasierte Methoden wie der PC Algorithmus nehmen an, dass die Verteilung bzgl. des Graphen Markov und treu ist. Diese beiden Bedingungen verbinden die (bedingten) Unabhängigkeiten in der Verteilung mit der Graphstruktur und ermöglichen mit Hilfe von Unabhängigkeitstests Teile des Graphen zu identifizieren. Unserer Meinung nach treten hierbei jedoch folgende Schwierigkeiten auf: (1) Man kann die Graphstruktur nur bis auf Markoväquivalenzklassen bestimmen. Insbesondere sind wir so nicht in der Lage zwischen $X \to Y$ und $Y \to X$ zu unterscheiden. (2) In der Praxis ist es schwierig, bedingte Unabhängigkeitstests durchzuführen. Vor allem, wenn die Menge an Variablen, auf die man bedingt, größer wird, besitzen die Tests oft nur eine relativ kleine Macht. (3) Wenn die Treuebedingung verletzt ist, liefern diese Methoden falsche Ergebnisse, ohne dass man dies erkennen kann. Gleiches gilt bei kausal unvollständigen Strukturen.

In sogenannten Structural Equation Models (SEMs) wird jede Variable X_j als Funktion einer Menge von Knoten $\mathbf{PA}_j$ und einer Rauschvariable N_j geschrieben:

$$X_j = f_j(\mathbf{PA}_j, N_j)\,, \qquad j = 1, \ldots, p$$

wobei alle Variablen N_j gemeinsam unabhängig sind. Den entsprech-

enden Graphen erhält man, indem man gerichtete Kanten von jeder Variablen auf der rechten Seite (den Eltern $\mathbf{PA}_j$) zu der entsprechenden Variable X_j auf der linken Seite zeichnet. In dieser Form sind SEMs jedoch zu allgemein, als dass man sie zum Strukturlernen verwenden könnte. Gegeben einer Verteilung kann man zu jedem Graphen, zu dem die Verteilung Markov ist, ein SEM mit genau dieser Struktur finden. Dies ändert sich, falls wir *eingeschränkte* oder *restricted* SEMs betrachten. In diesen schränkt man die Klasse der möglichen Kombinationen von Rauschen und Funktion ein. Betrachtet man beispielsweise nur SEMs mit linearen Funktionen und normalverteilten Rauschvariablen, kann man den Graphen bis auf Markoväquivalenzklasse aus der gemeinsamen Verteilung bestimmen (unter der zusätzlichen Annahme von Treue). Dieser Fall stellt allerdings eine Art Ausnahme da. Im Falle von linearen Funktionen und nicht-normalverteiltem Rauschen wird der Graph identifizierbar. In dieser Dissertation präsentieren wir alternative Modelle, die ebenfalls zur Identifizierbarkeit führen. (i) Mit Ausnahme weniger Beispiele erhalten wir Identifzierbarkeit ebenfalls bei nicht-linearen Funktionen mit beliebig verteiltem additiven Rauschen. Und (ii), im linearen normalverteilten Fall beweisen wir Identifzierbarkeit des Graphens aus der Verteilung unter der Annahme, dass alle Rauschvariablen die gleiche Varianz besitzen. Wir führen eingeschränkte SEMs für diskrete Variablen ein und erhalten analoge Ergebnisse. Ebenfalls zeigen wir, dass unsere Methoden auch auf Zeitreihen anwendbar sind. Ferner untersuchen wir, in wie weit man bei zwei beobachteten Variablen die Fälle "X verursacht Y", "Y verursacht X" und "beide Variablen werden durch eine dritte Variable verursacht" unterscheiden können (Detektion einer unbeobachteten gemeinsamen Ursache).

Unserer Meinung nach bringt die SEM-basierte kausale Inferenz und die Einschränkung der Funktionenklasse folgende Vorteile mit sich: (1) Wir können kausale Relationen auch innerhalb einer Markoväquivalenzklasse bestimmen. (2) Ein Model mit additivem Rauschen zu fitten (für jede Variable müssen wir eine multivariate Regression durchführen) ist ein einfacheres Problem als das Testen von bedingten Unabhängigkeiten. (3) Die Methoden basieren nicht auf der Treuebedingung. (4) Falls die (z.T. starken) Modellannahmen verletzt sind (beispielsweise sind die Daten nicht durch ein Modell mit additivem

Rauschen erzeugt oder es gibt unbeobachtete gemeinsame Ursachen), ist die Methode in der Lage, unentschlossen zu bleiben anstatt eine falsche Antwort zu geben.

Für alle vorgestellten Identifizierbarkeitsergebnisse stellen wir praktische Methoden vor, die wir auf simulierte und reale Datensätze anwenden.