Diss. ETH No. 22930

Estimating Causal Networks from Multivariate Observational Data.

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
CHRISTOPHER NOWZOHOUR

Master of Science, University of Oxford
born July 2, 1986
citizen of Germany

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Marloes Maathuis, co-examiner

2015

Abstract

The field of statistical causal inference is concerned with estimating cause-effect relationships between some variables from i.i.d. observations. This is impossible in general, e.g. one cannot distinguish $X \to Y$ from $X \leftarrow Y$ without making further assumptions. However, when more variables are involved or certain structural or distributional assumptions are made, causal inference becomes possible. This is relevant for applications, where randomized experiments are not feasible to test causal hypotheses (econometrics) or the large number of hypotheses requires some kind of pre-screening (genomics).

This thesis is about structure learning, which means estimating the underlying causal graph from data. Specifically, the focus of interest is score-based methods, which assign every possible causal graph a numeric score (depending on the observed data) and then try to find the graph maximizing this score. The two main challenges are:

1. Defining a meaningful score, that is maximized by the true underlying graph only, and is easily computable at the same time.
2. Solving the combinatorial optimization problem of maximizing the score over all possible graphs.

An important class of causal models are directed acyclic graphs (DAGs), where there are no cyclic relations and no hidden variables. DAGs (also known as Bayesian networks) encode conditional independencies in the joint distribution, and are only identifiable up to their equivalence class in general (there is generally more than one DAG encoding the same set of conditional independencies). When the model is restricted to additive noise, the independence of the noise terms can be used to identify DAGs completely, unless the model is linear and Gaussian (in the continuous case). This thesis presents a score-based method for continuous identifiable additive noise models. Specifically, a penalized

pseudo-likelihood score is developed for this nonparametric setting and proved to be consistent. The method is also successfully tested on simulated and real datasets.

To also accomodate hidden variables, the class of DAGs needs to be extended. A useful way to do this are bow-free acyclic path diagrams (BAPs), which put some restrictions on the hidden structure, but are statistically viable. The parametrization is assumed to be linear and Gaussian to facilitate likelihood scoring. This means full identifiability is not possible anymore. In contrast to DAGs, no established theory exists about model equivalency for BAPs. This thesis presents a greedy search method for this case, that estimates the equivalence class of the underlying graph, as well as some theoretical results about model equivalency. The method is shown to work on simulated data and is applied to a well-known genomics dataset, where the statistical fit is shown to be much better than for DAG models.

Zusammenfassung

Das Gebiet der statistischen kausalen Inferenz beschäftigt sich mit dem Schätzen von Ursache-Wirkungs-Zusammenhängen zwischen einer Reihe von Variablen basierend auf i.i.d. Daten. Ganz generell ist das nicht möglich (z.B. $X \to Y$ von $X \leftarrow Y$ zu unterscheiden) ohne zusätzliche Annahmen zu treffen. Dies ändert sich, sobald mehr Variablen involviert sind oder bestimmte strukturelle oder verteilungstechnische Annahmen getroffen werden. Kausale Inferenz ist besonders relevant für Anwendungen, für die randomisierte Experimente nicht möglich sind (z.B. in der Ökonometrie) oder wo die grosse Anzahl der zu testenden Hypothesen eine Art Vorauswahl erfordert (z.B. in der Genomik).

In dieser Dissertation geht es darum, den zugrundeliegenden kausalen Graphen von Daten zu schätzen. Der Fokus liegt insbesondere auf Score-basierten Methoden, die jedem Graphen einen (von den Daten abhängigen) numerischen Vergleichswert—die Score—zuordnen und dann versuchen den wertmaximierenden Graphen zu finden. Die zwei Hauptherausforderungen sind:

1. Das Definieren einer sinnvollen Score-Funktion, die nur vom wahren kausalen Graphen maximiert wird und zugleich einfach zu berechnen ist.

2. Das Lösen des kombinatorischen Optimierungsproblems um die Score über alle Graphen zu maximieren.

Eine wichtige Klasse von kausalen Modellen sind DAGs (directed acyclic graphs), in denen es keine zyklischen Strukturen und keine verborgenen Variablen gibt. DAGs (auch als Bayes'sche Netze bekannt) kodieren konditionelle Unabhängigkeiten in der gemeinsamen Wahrscheinlichkeitsverteilung und sind im Allgemeinfall nur bis auf ihre Äquivalenzklasse identifizierbar (es gibt in der Regel mehrere DAGs die die gleichen konditionellen Unabhängigkeiten kodieren). Wenn man das

Modell auf additive Fehlerterme beschränkt, kann die Unabhängigkeit dieser Fehlerterme dazu genutzt werden den DAG komplett zu identifizieren, ausser das Modell ist linear und normalverteilt (im kontinuierlichen Fall). Diese Arbeit präsentiert eine Score-basierte Methode für kontinuierliche und identifizierbare Modelle mit additiven Fehlertermen. Da dieses Szenario nichtparametrisch ist, wurde eine penalisierte pseudo-Likelihood Score entwickelt und deren Konsistenz bewiesen. Die Methode wurde ausserdem erfolgreich an simulierten und reellen Daten getestet.

Um auch verborgene Variablen zu modellieren muss die Klasse der DAGs erweitert werden. Eine Möglichkeit dies zu tun sind BAPs (bow-free acyclic path diagrams). Die verborgenen Variablen sind hier bestimmten Restriktionen unterlegen, aber dafür ist das statistische Modell praktikabel. Die Parametrisierung ist linear und normalverteilt, sodass eine likelihood score eingesetzt werden kann. Das heisst aber auch, dass das Modell nicht mehr komplett identifizierbar ist. Im Gegensatz zu DAGs gibt es für BAPs keine vollständige Theorie der Äquivalenzklassen. In dieser Arbeit wird ein Greedy-Algorithmus für dieses Szenario präsentiert, der die Äquivalenzklasse des zugrundeliegenden Graphen schätzt. Ausserdem werden einige theoretische Resultate über die Äquivalenzstruktur von BAPs vorgestellt. Die Methode wurde ebenfalls erfolgreich an simulierten Daten getestet und wurde darüberhinaus auf ein bekanntes Genomik-Datenset angewendet, wo der statistische Fit erheblich besser war als für DAG Modelle.