Diss. ETH No. 23806

CAUSAL INFERENCE IN SEMIPARAMETRIC AND NONPARAMETRIC STRUCTURAL EQUATION MODELS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

JAN ERNEST

MSc ETH Mathematics
born on 09.04.1987
citizen of Zürich ZH

accepted on the recommendation of

Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Marloes M. Maathuis, co-examiner

2016

Abstract

The goals of causal inference are inherently different from the ones of classical statistics. Instead of measuring statistical associations between variables, the main focus is on the characterization of the underlying causal mechanisms. This is typically achieved via the estimation of causal graphs (*structure learning*) or the prediction of causal effects under interventions. Both problems are well-understood and elaborated for linear structural equation models (SEMs). This thesis addresses them for specific classes of semiparametric and nonparametric SEMs.

First, we study structure learning for causal additive models (CAMs). CAMs constitute a natural semiparametric extension of linear Gaussian SEMs: while still relying on the additivity of the functions and Gaussianity of the noise, all functions are assumed to be exclusively nonlinear. We present a score-based structure learning methodology based on (restricted) maximum likelihood estimation that is consistent in low- and high-dimensional settings. The key idea of our approach is to decouple order search among the variables from subsequent edge selection in the graph. We provide an efficient implementation of our proposed methodology in the `R`-package `CAM` and evaluate its performance in extensive simulations.

In the second part of the thesis, we study the identifiability and estimation of partially linear additive SEMs with Gaussian noise (PLSEMs). Thus, we drop the assumption of exclusivity of the functional type and with that we address one of the major limitations of both, linear SEMs and CAMs. We precisely specify how linear and nonlinear additive functions impose restrictions on the underlying causal model and derive a systematic characterization of the identifiability of PLSEMs. Thereby, we close a relevant gap, as the identifiability theory of additive models with Gaussian noise was only elaborated for linear SEMs and CAMs. We complement

the theoretical findings with an efficient score-based estimation procedure that, given one PLSEM, finds all equivalent PLSEMs. We prove low- and high-dimensional consistency results for our algorithm and evaluate its performance on simulated datasets.

In the last part, we additionally relax the additivity and Gaussianity assumptions. Structure learning for unstructured nonparametric SEMs is a highly ambitious task as it is plagued by the curse of dimensionality. Interestingly, the situation is different for the estimation of (total) causal effects. We show that a specific marginal integration regression technique (S-mint) theoretically achieves the optimal univariate convergence rate of nonparametric regression for a very general class of nonparametric SEMs with known (or approximately known) structure (assuming sufficient smoothness). Specifically, S-mint does not suffer from the curse of dimensionality. We propose an implementation based on an additive regression approximation with subsequent L_2-boosting. In extensive simulations, our method demonstrates a more pronounced robustness with respect to model misspecification than other methods that rely more heavily on the correct estimation of the causal structure.

Zusammenfassung

Das Gebiet der kausalen Inferenz hat eine grundsätzlich andere Zielsetzung als die klassische Statistik. Statt statistische Assoziationen zwischen Variablen zu messen, besteht der Hauptfokus der kausalen Inferenz darin, die zugrundeliegenden kausalen Zusammenhänge zu charakterisieren. Dies kann auf verschiedene Arten angegangen werden. Zum Beispiel, indem man einen Graphen schätzt, der die kausalen Mechanismen abbildet. Alternativ kann man versuchen, direkt die kausalen Effekte vorherzusagen, welche durch Interventionen verursacht werden. Beide Ansätze sind seit längerem bekannt und ausgereift für lineare Strukturgleichungsmodelle. Die vorliegende Doktorarbeit untersucht diese Ansätze in spezifischen Klassen von semiparametrischen und nichtparametrischen Strukturgleichungsmodellen.

Eine naheliegende semiparametrische Erweiterung der linearen Gauss'schen Modelle sind sogenannte kausale additive Modelle. Sie sind immer noch additiv mit Gauss'schen Fehlertermen, bestehen jedoch ausschliesslich aus nichtlinearen Funktionen. Wir entwickeln eine score-basierte Maximum-Likelihood Methode, um für diese Modellklasse die zugrundeliegenden kausalen Graphen zu schätzen und zeigen deren Konsistenz für den tief- und hoch-dimensionalen Fall. Die entscheidende Idee der Methode besteht darin, die Suche nach einer korrekten kausalen Ordnung der Variablen von der Suche nach individuellen Kanten im kausalen Graphen zu entkoppeln. Wir stellen eine effiziente Implementierung der Methode im `R`-Paket `CAM` zur Verfügung und untersuchen deren Leistungsfähigkeit in diversen numerischen Experimenten.

Sowohl die linearen Gauss'schen Modelle als auch die kausalen additiven Modelle besitzen den grossen Nachteil, dass alle additiven Komponenten vom selben Typ sein müssen, das heisst, entweder alle linear oder alle nichtlinear. Diese restriktive Annahme kann umgangen werden, indem man

beide Funktionstypen im gleichen Modell zulässt, das heisst, durch Betrachten von partiell linearen additiven Strukturgleichungsmodellen mit Gauss'schen Fehlertermen. Wir untersuchen, wie lineare und nichtlineare Funktionen das zugrundeliegende kausale Modell einschränken, und leiten daraus eine systematische Charakterisierung der Identifizierbarkeit der gesamten Modellklasse her. Dadurch schliessen wir eine grosse Lücke in der Identifizierbarkeitstheorie additiver Modelle, welche bisher nur für lineare Gauss'sche und kausale additive Modelle ausgearbeitet wurde. Wir ergänzen die Theorie durch einen effizienten Algorithmus, der für ein gegebenes partiell lineares Modell alle dazu äquivalenten Modelle auflistet. Wir beweisen dessen Konsistenz im tief- und hoch-dimensionalen Fall und untersuchen die Leistungsfähigkeit auf simulierten Datensätzen.

Zuguterletzt stellen wir uns die Frage, welche Aussagen ohne die Annahme von additiven Funktionen und Gauss'schen Fehlertermen getroffen werden können. Unglücklicherweise ist das Schätzen von unstrukturierten nichtparametrischen Modellen geprägt vom sogenannten Fluch der Dimensionalität. Dies ist interessanterweise nicht der Fall, wenn wir versuchen, (totale) kausale Effekte zu schätzen. Wir zeigen, dass eine spezifische Regressionsmethode, die auf marginaler Integration beruht, für eine allgemeine Klasse von nichtparametrischen Strukturgleichungsmodellen mit bekannter (oder ungefähr bekannter) Struktur die optimale univariate Konvergenzrate für nichtparametrische Regression erreicht (unter Annahme genügender Differenzierbarkeit). Insbesondere umgeht diese Methode den Fluch der Dimensionalität. Als Ergänzung zur Theorie schlagen wir eine Implementierung der Methode vor, welche auf einer additiven Approximation mit anschliessendem L_2-boosting beruht. In ausgiebigen Simulationen erweist sich diese Methode als robuster gegenüber Abweichungen vom Modell als andere Schätzverfahren, welche stärker von der korrekten Schätzung der kausalen Struktur abhängig sind.