Causal learning from high-dimensional observational data

Diss. ETH No. 23751

CAUSAL LEARNING FROM HIGH-DIMENSIONAL OBSERVATIONAL DATA

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
PREETAM NANDY
Master of Statistics, Indian Statistical Institute
born August 15, 1988
citizen of India

accepted on the recommendation of
Prof. Dr. Marloes Maathuis, examiner
Prof. Dr. Nicolai Meinshausen, co-examiner
Prof. Dr. Thomas Richardson, co-examiner

2016

Abstract

Many scientific questions are causal in nature. Observational studies are sometimes the only source of information for answering such questions, especially when intervention experiments are impractical or unethical. Moreover, even if intervention experiments are possible, readily available observational data can be used, for example, in prioritizing much more expensive and time-consuming experiments.

In the first part of this thesis, we consider the estimation of joint intervention effects from high-dimensional observational data. For example, one can think of predicting the effects of double or triple gene-knockouts on other genes, based on observational gene expression profiles. There is an existing method, called IDA, to predict (bounds on) single intervention effects (e.g., single gene knock-outs) from observational data. Since the space of possible intervention experiments grows exponentially in the number of simultaneous interventions, having an IDA-like tool to predict the effect of multiple simultaneous interventions is highly desirable in order to plan and prioritize such experiments. We therefore develop a new method, called joint-IDA, to estimate (bounds on) the effect of multiple simultaneous interventions. This is a challenging extension, because the estimation of causal effects with covariate adjustments no longer works for joint interventions. Moreover, we relax the linearity and the Gaussianity assumptions of IDA and extend its high-dimensional consistency results to joint-IDA. We evaluate the estimators in simulation studies and also illustrate them on data from the DREAM4 challenge.

The second part of the thesis focuses on causal structure learning from high-dimensional observational data. This is, in fact, the first step of both IDA and joint-IDA, as these methods do not assume the causal structure to be known. The main approaches for causal structure learning are so-called

constraint-based, score-based or hybrid methods. Hybrid methods borrow ideas from both constraint-based and score-based methods, and often use a greedy search on a restricted search space in order to achieve computational efficiency. Such methods tend to work well in practice, but very little was known about their theoretical properties. We show that a naive hybrid version of the score-based greedy equivalence search (GES) is inconsistent, meaning that the algorithm cannot learn the correct causal structure even with infinite samples. We also show that we can achieve consistency with an adaptive modification of the search space. This leads to the Adaptively Restricted GES (ARGES) algorithm. Further, we show that both GES and ARGES are consistent in certain sparse high-dimensional settings. To our knowledge, these are the first high-dimensional consistency results for score-based and hybrid algorithms. In simulation studies, we found that ARGES combines the best aspects of the constraint-based PC algorithm and the score-based GES algorithm: the fast computation of PC and the good estimation performance of GES.

Zusammenfassung

Viele wissenschaftliche Fragestellungen sind von kausaler Natur. Beobachtungsstudien sind manchmal die einzige verfügbare Informationsquelle, um solche Fragen zu beantworten, insbesondere dann, wenn Interventionsexperimente nicht durchführbar sind. Aber auch wenn Interventionsexperimente möglich sind, können bereits vorhandene Beobachtungsdaten zum Beispiel zusätzlich benutzt werden, um teurere und zeitaufwändige Experimente zu priorisieren.

Im ersten Teil dieser Doktorarbeit befassen wir uns mit dem Schätzen von gemeinsamen Interventionseffekten basierend auf hochdimensionalen Beobachtungsdaten. Zum Beispiel möchten wir den Effekt eines doppelten oder dreifachen Gen-Knockouts auf andere Gene, ausschliesslich basierend auf beobachteten Genexpressionsmustern, vorhersagen. Es existiert bereits eine Methode namens IDA, um basierend auf Beobachtungsdaten (Schranken für) einzelne Interventionseffekte (zum Beispiel, von einzelnen Gen-Knockouts) vorherzusagen. Da aber der Raum möglicher Interventionsexperimente exponentiell wächst in der Anzahl gleichzeitig zugelassener Interventionen, wäre es äusserst wünschenswert, eine IDA-ähnliche Methode zu haben, die den Effekt mehrerer gleichzeitig stattfindender Interventionen vorhersagt und damit die Planung und Priorisierung solcher Interventionsexperimente ermöglicht. Zu diesem Zweck entwickeln wir die neue Methode joint-IDA, um (Schranken für) den Effekt mehrerer gleichzeitig stattfindender Interventionen zu schätzen. Dies ist eine anspruchsvolle Erweiterung, da die Schätzung kausaler Effekte via covariate adjustments nicht mehr funktioniert für gleichzeitig stattfindende Interventionen. Wir schwächen zudem die Linearitäts- und Gaussianitätsannahmen von der IDA-Methode ab und erweitern deren Konsistenzresultate auf joint-IDA. Wir evaluieren die Schätzer in Simulationsexperimenten und wenden sie zusätzlich auf einen realen Datensatz der “DREAM4 challenge” an.

Der zweite Teil der Arbeit beschäftigt sich mit dem Lernen kausaler Strukturen basierend auf hoch-dimensionalen Beobachtungsdaten. Dieses Lernen der kausalen Struktur ist der erste Schritt von sowohl IDA als auch joint-IDA, da beide diese Methoden nicht auf der Annahme beruhen, dass die kausale Struktur im Vorraus bekannt ist. Dazu gibt es folgende drei Hauptansätze: Constraint-basierte Methoden, score-basierte Methoden und Hybridmethoden. Die Hybridmethoden borgen dabei Ideen von constraint- und score-basierten Methoden und beruhen häufig auf einer Greedy-Suche auf einem eingeschränkten Suchraum, um eine effiziente Rechenleistung zu gewährleisten. Diese Hybridmethoden tendieren dazu, in praktischen Anwendungen gut zu funktionieren, es ist jedoch nur sehr wenig über ihre theoretischen Eigenschaften bekannt. Wir zeigen, dass eine naive Hybridversion des score-basierten greedy equivalence search (GES) Algorithmus inkonsistent ist, was bedeutet, dass sie auch mit unendlicher Anzahl an Daten nicht in der Lage ist, die korrekte kausale Struktur zu lernen. Wir zeigen dann, dass wir mittels einer adaptiven Anpassung des Suchraums eine konsistente Version erreichen können. Diese Anpassung führt zum Adaptively Restricted GES (ARGES) Algorithmus. Wir beweisen, dass sowohl GES als auch ARGES konsistent sind in gewissen hoch-dimensionalen Situationen. Soweit uns bekannt ist, sind dies die ersten hoch-dimensionalen Konsistenzresultate für score-basierte und Hybrid-Algorithmen. In Simulationsstudien haben wir gesehen, dass ARGES die jeweils besten Aspekte des constraint-basierten PC-Algorithmus und des score-basierten GES-Algorithmus vereinigt: Die schnelle Rechenleistung von PC und die guten Schätzeigenschaften von GES.