Statistical aspects of inferring Bayesian networks from marginal observations

Kai von Prillwitz

supervised by Prof. David Gross

handed in October 23, 2015

The scope of the thesis is causal inference, the mathematical theory of 'what causes what'. Causal inference, or in general the concept of causation, is basic to human thinking, and philosophical theories about causation date back at least to Aristotle. Yet, a solid mathematical theory has long been missing. One major problem is to distinguish an actual causal dependence of two variables from mere correlation. To be precise, correlations are not necessarily the result of direct causal influences but they could also be due to a common cause, represented by a third (not observed) variable. Today, this falls under the concept of 'spurious correlation' and is related to the statement 'correlation does not imply causation'.

It has become popular to represent causal relations by networks of variables, so called *directed acyclic graphs* (DAGs). Vertices represent variables and directed edges can be interpreted as causal influence. In the first place, a DAG encodes (conditional) independence relations between its variables (leading to the term *Bayesian network*). The independence relations constrain the set of probability distributions that are compatible with the DAG. If any independence relation implied by the DAG is not found in a given probability distribution, the DAG and the distribution are deemed incompatible. If the distribution was obtained as an empirical distribution associated with a real data set, the DAG (and thus also its causal interpretation) can be rejected as an explanation for generating the data.

Note the similarity of this approach and bell inequalities. Bell inequalities are derived under certain assumptions of classical physics. Observing violations

of these inequalities in real data means that classical physics cannot explain the data generating mechanisms. Instead, quantum mechanics or yet another theory is required.

The application to real data is also where statistics enter the thesis. The smaller a data set, the less reliable is the estimation of the corresponding empirical distribution. A statistical hypothesis test becomes necessary to decide the compatibility of the data and the DAG. The whole task becomes dramatically more complicated if some of the variables in the model are not observable. Independence relations including such hidden variables cannot be evaluated and the remaining independence relations, if any exist, might carry only little information. Thus, less obvious constraints on the marginal distributions of the observed variables have to be found.

The goal of the thesis is to elaborate on such constraints and to use them to construct proper hypothesis tests that allow to decide the compatibility of some data and a proposed causal model. We start by working with inequality constraints in a recently developed entropic framework and implement likewise recent techniques of entropy estimation. In a second step we derive and implement analogous constraints on the level of certain generalized covariance matrices (which are contrary to actual covariances independent of the alphabets of the variables). While this is motivated by the search for a more powerful hypothesis test, deriving the new type of inequalities is interesting on its own as such constraints based on the structure of the DAG alone are rare to this day. Furthermore, we distinguish two different approaches to hypothesis testing.

By the above means we improve a previously proposed hypothesis test both in terms of its power (correctly identifying incompatibility between data and model) as well as its type-I-error control (minimizing falsely identified incompatibilities). For illustrative purposes we apply our methods to real empirical data, the so-called 'iris (flower) data set'. Roughly said, we show that several size attributes of the blossoms of iris flowers are most likely influenced by one common cause (a genetic factor or environmental influences) rather than by a more complex (and restrictive) causal structure. Note that the data set consists of three different types of iris flowers so that this result could be expected. A natural next step would be the application of our methods to more recent research data.