Niklas Pfister December 15, 2021

Causal inference for complex data structures

In statistics we attempt to connect observed data with a statistical model to draw conclusions about the underlying system giving rise to the data. A statistical model is always a simplification of reality and only captures the true underlying mechanism if certain assumptions hold. The model specification and the required assumptions depend on the problem at hand. It is useful to distinguish two important general types of statistical tasks.

(i) *Observational inference:* These are settings in which one observes a system in a fixed state and is interested in questions about the system in this specific state. For example, estimating the average number of smokers in a fixed population.

(ii) *Causal inference:* These are settings in which one observes a system (possibly in multiple states) and is interested in questions related to the effects of certain changes in the system. For example, estimating the effect a ban on tobacco advertisement has on the average number of smokers. This requires estimating the average number of smokers in a population where the ban is implemented but everything else remains as in the observed population.

A key difference between these tasks is that observational inference only requires interpolation while causal inference requires extrapolation. More specifically, the statistical model for observational problems is only assumed to be correct for a fixed data generating model. In contrast, for causal problems, additional assumptions are required to guarantee that the model also correctly captures specific changes (or interventions) to the data generating model.

To answer causal questions, statisticians therefore need to explicitly model parts of the mechanism underlying the data generating model. Given such a model it is then often possible to reduce the causal question to an observational one and estimate the quantity of interest with standard approaches. The main difficulty, however, is that the statistical conclusions are highly dependent on the extrapolation assumptions (that is, the assumptions guaranteeing that he model is correctly specified). It is therefore crucial to understand the limitations of these assumptions. In recent years, the sub-field of causality has successfully constructed a mathematical framework that allows to describe changes to a system and understand under which assumptions which part of a system can be inferred from the observed data. However, applying this framework to more complex data structures is unfortunately often still infeasible because we are lacking methods that are able to incorporate non-standard structural constraints. In my research, I am working on developing new ways to incorporate as much domain-specific knowledge as possible into a statistical model. In particular, I focus on problems in which there are data structures which induce non-standard statistical and causal constraints onto the model. This is best illustrated with two examples: (i) Microbiome data, in which the measuring procedure adds useful structure and (ii) offline sequential decision data, where the structure stems from the data generating process in which an agent is interacting with a system over time.

Microbiome data The microbiome refers to the collective genomes of microorganisms residing within multi-cellular organisms, such as plants, animals and humans. An important subclass is the human microbiome, which consists of microorganisms that reside on or within human tissue and biofluids, e.g., saliva, skin or the gastrointestinal tract. These ecosystems of bacteria, viruses, and fungi have been associated with various human health outcomes, such as cardiovascular and metabolic diseases, obesity, mental illness, and autoimmune disorders. To investigate these types of links between health outcomes and the microbiome, researchers are using next-generation sequencing technology to measure the microbiome. The resulting sequencing data can be processed to derive abundance measurements of the microbes in the sample, which is then used in subsequent analysis steps. Deriving statistical models for this type of data is challenging because the data is high-dimensional (there are mainly more microbes than samples), compositional (the abundances of individual microbes are correlated since they need to some to one), zero-inflated (many microbes only exist in few samples) and contains a phylogenetic tree-structure (since genetically similar microbes have similar functions). Microbiome researchers attempt to use this data, for example, to answer the following questions:

- Does an increased abundance of a certain microbe affect a health outcome?
- How does the abundance of some microbes affect the abundance of others?

These questions are causal in nature and answering them requires making assumptions that allow to extrapolate. Currently, we are working on ways to use the structure to constrain the feasible models. This allows for more reliable inference about the underlying mechanism compared to what would be possible if the structure is neglected.

Offline sequential decision data Many real-world systems can be described as sequential decision processes. That is, systems in which there is an agent – not necessarily human – that continuously interacts with the system by making decisions. Examples, include the federal reserve updating their monetary policy or a doctor deciding on a treatment for a patient. The agent is therefore part of the model and needs to be incorporated into the statistical model. Markov-decision processes are a well-established way of modeling such systems by focusing on optimizing the policy of an agent (i.e., the function determining which action to take given the current state) to maximize a pre-specified reward. One commonly separates two cases, either the policy of the agent can be actively controlled during data collection (the online case) or one only observes the agent interacting with the system without being able to engage with the system during data collection (the offline case). In the offline setting one can generally not directly observe the outcome of a specific policy of interest. Therefore, many problems in the offline setting fall into the category of causal inference. Any inference in these settings will depend on the assumptions we are willing to make. Fortunately, there is again a rich underlying structure that constraints the feasible models and helps the statistical analysis. For example, in many cases there is a clear time structure that separates actions from the state of the system and in some cases the system is observed under several partially randomized policies adding further constraints. There are many open questions that are interesting in these settings.

- Under what conditions on the observed policies and the underlying system is it possible to estimate the effect of a specific action?
- How can offline data be used to initialize an online policy learning algorithm to optimize a reward?

Our approach to answering these questions is to incorporate as many of the induced constraints as possible and investigate whether this can help us better understand the assumptions required to infer the causal quantities of interest.