Research Statement

Michael Schweinberger

My research is concerned with methodological, computational, and theoretical aspects of learning from

- discrete and dependent data without independent replications: e.g., network, spatial, and temporal data;
- data with additional structure: e.g., block, multilevel, spatial, and temporal structure;
- social and health science data: e.g., educational, network, and epidemiological data.

My research has been made possible by NWO award Rubicon-44606029 (sole PI), NSF award DMS-1513644 (sole PI), NSF award DMS-1812119 (sole PI), and ARO award W911NF-21-1-0335 (lead PI).

Overview

My research is motivated by discrete and dependent data without independent replications, such as network, spatial, and temporal data. My ideas of how to learn from discrete and dependent data without independent replications are elaborated in one of the simplest possible settings: statistical exponential families (Wainwright and Jordan, Foundations and Trends in Machine Learning, 2008). Exponential families are widely used throughout data science, as stand-alone models or building blocks of more complex models. The fundamental role of exponential families in data science is exemplified by the prominent role of multivariate Gaussians, but there are numerous other examples, including generalized linear models, undirected graphical models and random graph models with exponential-family parameterizations, Markov random fields in machine learning, and Boltzmann machines in artificial intelligence. For example, random graph models with exponential-family parameterizations are widely used for specifying models that capture dependencies in network data. Such models represent an alternative to latent variable models that induce dependence through latent variables, but do not provide data scientists with a simple and flexible approach to specifying and comparing a wide range of models with competing forms of dependence and other features of interest, despite the fact that latent variable models can be viewed as natural models of exchangeable random graphs by an appeal to the Aldous-Hoover theorem.

Selected highlight

An example of discrete and dependent data without independent replications is network data. Since the 1950s, scientists have argued that connections depend on other connections: e.g., the frequent observation that "a friend of a friend is a friend" suggests that social contacts are dependent, which has implications in terms of understanding and predicting pandemics and other network-mediated phenomena of interest. In applications, population probability models are learned from a single observation of a population network or subnetworks sampled from a population network. That raises an important question:

How can we construct models of network-mediated phenomena that respect the fact that connections depend on other connections and learn them from data, without having the benefit of independent observations from the same source?

I have attempted to contribute constructive answers to these questions in a decade-long sequence of first- and single-authored papers (e.g., Annals of Statistics, 2020; Bernoulli, 2020; Statistical Science, 2020; Journal of the Royal Statistical Society, Series B, 2015; Journal of the American Statistical Association, 2011) and more recent papers (e.g., Stewart and Schweinberger, 2021):

1. I demonstrated how models should not be constructed, by studying ill-posed statistical exponential-family models of discrete and dependent network data. My single-authored JASA (2011) paper was among the earliest papers on the topic and preceeded Chatterjee and Diaconis (AOS, 2013).
2. I demonstrated how well-posed models can be constructed, by introducing next-generation models of discrete and dependent network data that combine salient features of latent structure models (capturing who is close to whom) and statistical exponential-family models (capturing local dependence).

3. I showed that statistical learning of an unbounded number of parameters based on a single observation of dependent data from a statistical exponential family is possible, with theoretical guarantees.
4. I showed that scalable statistical learning of an unbounded number of parameters based on a single observation of dependent data from a statistical exponential family is possible, with theoretical guarantees.

There is a common thread that connects these advances: **the importance of additional structure.** Models that lack mathematical structure to control the dependence among connections can be ill-posed, but endowing models with additional structure can help control dependence and result in well-posed models with desirable properties. In addition, weak dependence facilitates concentration-of-measure results, which in turn facilitate theoretical guarantees. In practice, there are many forms of additional structure (e.g., block, multi-level, spatial, and temporal structure) and statistical algorithms can take advantage of additional structure to scale up computing. To conclude, additional structure has at least two advantages:

1. It facilitates the construction of well-posed models with desirable properties.
2. It facilitates scalable statistical learning with theoretical guarantees.

Last, but not least, additional structure helps answer fundamental questions about the statistical analysis of dependent network data raised by leading probabilists (e.g., Chatterjee and Diaconis, AOS, 2013) and statisticians (e.g., Fienberg, JCGS, 2012). We have provided tentative answers in Statistical Science (2020).

Selected directions of future research

Online educational assessment data: In collaboration with Minjeong Jeon (Graduate School of Education & Information Studies, University of California, Los Angeles), I am working on educational assessment data, including online educational assessment data. Among other things, we are developing statistical interaction and learning progression maps based on latent space models, with a view to providing educators with visual tools for monitoring student progress and detecting disadvantaged groups of students who need more support than other students, with applications to traditional and online educational assessments.

Stochastic processes involving networks, space, and time: Many real-world processes involve networks, space, and time: e.g., infectious diseases spread by way of contact, contacts depend on geographical space, and contacts change over time. While there are existing stochastic processes indexed by networks, space, time or combinations of them, many of them make either simplifying assumptions or have unknown probabilistic and statistical properties. One of my directions of future research is to design stochastic processes indexed by networks, space, and time that do justice to the complexity of network-mediated phenomena and develop scalable statistical methods for learning them from data, leveraging my decade-long research on the building blocks for learning from discrete and dependent data without independent replications.

Scalable selection of models of dependent data without independent replications and intractable likelihood functions: Developing scalable model selection procedures with theoretical guarantees is non-trivial when the likelihood function is intractable, the number of parameters is large, and the data consists of a single observation of dependent random variables. Such scenarios arise in the statistical analysis of discrete and dependent data, including network, spatial, and temporal data. For example, there are many models of dependent network data, but no scalable model selection procedures with theoretical guarantees are known. I am working on a scalable approach to model selection in dependent data problems with intractable likelihood functions based on regularized pseudo- and composite-likelihood methods, with theoretical guarantees.

Quantifying uncertainty of statistical learning based on dependent data without independent replications: In applications of data science, it is important to provide a disclaimer, acknowledging that statistical conclusions based on data are subject to error. In scenarios when the number of parameters is unbounded and a single observation of discrete and dependent data is available, it is not obvious how to quantify the uncertainty about statistical conclusions based on data, because the distributions of many statistical quantities are unknown. A natural approach to capturing uncertainty is a Bayesian approach. While Bayesian approaches to learning from discrete and dependent data without independent replications have long flourished, many of them are either not scalable or theoretical guarantees are unknown. I intend to elaborate on scalable Bayesian approaches to uncertainty quantification for discrete and dependent data without independent replications based on factorized objective functions (e.g., pseudo- and composite-likelihood functions), with theoretical guarantees.