Research Statement

Michael Schweinberger

My research is concerned with methodological and theoretical aspects of statistical learning from discrete and dependent data without independent replications, including network, spatial, and temporal data arising in the social sciences, the health sciences, and other fields. My research would have been impossible without the generous support of the Netherlands Organisation for Scientific Research (the Dutch equivalent of the NSF), in the form of NWO award Rubicon-44606029 (sole PI); the U.S. National Science Foundation, in the form of NSF awards DMS-1513644 (sole PI) and DMS-1812119 (sole PI); and the U.S. Department of Defense, in the form of 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 arising in the social sciences, health sciences, and other fields. My main 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). Statistical exponential families are widely used throughout data science, as stand-alone models or building blocks of more complex models. The fundamental role of statistical 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 parameterizations, Markov random fields in machine learning, and Boltzmann machines in artificial intelligence.

Selected highlight

Consider network data. Since the 1950s, scientists have argued that connections depend on other connections: e.g., the observation that "a friend of a friend is a friend" suggests that friendships are dependent, which has implications in terms of understanding and predicting network-mediated phenomena of interest (e.g., pandemics). In many 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?

In a decade-long sequence of first-authored publications (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., arXiv:2012.07167), I have attempted to contribute constructive answers to these questions by leveraging the statistical exponential-family platform:

1. I have demonstrated how models should not be constructed, by studying ill-posed statistical exponential-family models of discrete and dependent network data. My JASA (2011) paper was among the earliest papers on the topic and preceded the widely read paper of Chatterjee and Diaconis (AOS, 2013).
2. I have shown how well-posed models can be constructed, by developing models that combine the advantages of latent structure models (capturing who is close to whom) and statistical exponential-family models (capturing local dependence).
3. I have demonstrated that scalable statistical learning of statistical exponential-family models of discrete and dependent network data with an unbounded number of parameters and an intractable likelihood function is possible, with theoretical guarantees.

There is a common thread that connects these advances: **the importance of additional structure.** Statistical exponential-family 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 for statistical learning. In practice, there are many forms of additional structure (e.g., block, multilevel, spatial, and temporal structure), and statistical algorithms can take advantage of additional structure for the purpose of large-scale computing. As a result, 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 network data on the statistical exponential-family platform, raised by many probabilists and statisticians in the field. We have provided tentative answers to these questions in Statistical Science (2020).

Selected directions of future research

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 distance, and contacts change over time. While there are existing stochastic processes indexed by networks, space, and time, 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 basics of learning from discrete and dependent data without independent replications.

Scalable selection of models of discrete and dependent data without independent replications: 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. As a case in point, there are many models of dependent network data, but model selection procedures are scarce and lack either computational scalability or theoretical guarantees or both. 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 based on discrete and dependent data without independent replications: In applications of statistics, 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 random variables is available, it is not obvious how to quantify the uncertainty about statistical conclusions, because the small- and large-sample distributions of many statistical quantities are unknown. A natural approach to capturing uncertainty is a Bayesian approach. I intend to elaborate on scalable Bayesian approaches to uncertainty quantification for discrete and dependent data without independent replications and with intractable likelihood functions based on factorized objective functions (e.g., pseudo- and composite-likelihood functions), with theoretical guarantees.

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.