Research Statement

Michael Schweinberger

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

- dependent and high-dimensional data without independent replications: e.g., network, spatial, and temporal data;
- structured data, that is, data with additional structure, either observed or unobserved: e.g., hierarchical, multilevel, and multiscale structure;
- social science data: e.g., educational data, epidemiological data, and network data.

Overview

My research is motivated by dependent and high-dimensional data without independent replications, such as network, spatial, and temporal data. My ideas of how to learn from dependent data without independent replications are elaborated in the simplest possible setting: statistical exponential families (Wainwright and Jordan, Foundations and Trends in Machine Learning, 2008). Statistical exponential families are widely used throughout data science, either as stand-alone models or as 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 applications of exponential families, including generalized linear models, Markov random fields in machine learning, and Boltzmann machines in artificial intelligence. In fact, some of my research (e.g., Schweinberger, JASA 2011) has contributed to the understanding of generative deep learning models in artificial intelligence: see Kaplan et al. (2020). *On the S-instability and degeneracy of discrete deep learning models.* Information and Inference.

Selected highlight

Consider network data, which are dependent data without independent replications. Since the 1950s, social scientists have pointed out that connections depend on other connections: e.g., the phenomenon that "a friend of a friend is a friend" suggests that friendships are dependent. 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:

What can we learn about a connected world where connections depend on other connections, without having the benefit of independent observations from the same source? In general, what can we learn in high-dimensional scenarios with $p \to \infty$ parameters and a single observation of dependent random variables?

In a decade-long sequence of papers published in the most prestigious journals in statistics (e.g., The Annals of Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society, Series B, Bernoulli, Statistical Science), I have:

1. Studied the properties of ill-behaved models of dependent random variables, with applications to models of dependent network data and generative deep learning models in artificial intelligence (e.g., restricted Boltzmann machines). My work (Schweinberger, JASA, 2011) preceeded the related work of Chatterjee and Diaconis (AOS, 2013).

2. Shown how well-posed models of dependent network data can be constructed, with desirable properties.

3. Demonstrated that statistical learning of $p \to \infty$ parameters based on a single observation of dependent random variables is possible, with theoretical guarantees.

4. Developed scalable methods for statistical learning of $p \to \infty$ parameters based on a single observation of dependent random variables, with theoretical guarantees.

There is a common thread that connects all of these advances: the importance of additional structure. Models that lack mathematical structure to control the dependence among random variables can be ill-behaved, 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 consistency results. In other words, endowing models with additional structure has two advantages:

1. It facilitates the construction of well-posed models with desirable properties.

2. It facilitates statistical learning with theoretical guarantees.

In practice, there are many forms of additional structure (e.g., spatial, temporal, and multilevel structure), and it makes sense to take advantage of additional structure when available. Besides, additional structure helps answer fundamental questions about the statistical analysis of non-standard, dependent, and high-dimensional network data raised raised by leading probabilists (e.g., Chatterjee and Diaconis, AOS, 2013) and statisticians (e.g., Fienberg, JCGS, 2012).

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. Among other things, we are developing statistical interaction and learning progression maps, with a view to providing educators with visual tools for monitoring student progress and detecting (underrepresented) groups of students who need more, and different support than other students. The basic idea is to embed both students and test items in a shared metric space. By embedding students and test items in the same metric space, teachers can assess how students interact with items, and how students progress over time. A practical advantage is that interaction and learning progression maps provide a simple and appealing visual summary of student learning in a metric space (e.g., a low-dimensional Euclidean space), helping detect students from underrepresented groups who need more, and different support than other students.
- **Stochastic processes involving networks, space, and time:** Many real-world processes involve networks, space, and time: e.g., infectious diseases spread through contacts among population members, contacts depend on geographical space, and contacts change over time. While there are existing stochastic processes indexed by networks, space, time or a combination of them, many of them are either simplistic or complex but 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 real-world phenomena, and develop scalable statistical and computational methods and theoretical guarantees for learning them from data.

- **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, such as discrete 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 plan to develop a scalable approach to model selection in dependent data problems with intractable likelihood functions based on pseudolikelihood-based Dantzig selectors. Pseudo-likelihood Dantzig selectors are a natural extension of the pseudolikelihood-based M-estimators of Stewart and Schweinberger (2021). On computational grounds, pseudolikelihood-based Dantzig selectors are attractive, because computing them does not require intractable normalizing constants. On theoretical grounds, pseudolikelihood-based Dantzig selectors are appealing as well, because theoretical guarantees for model selection can be obtained in dependent data problems without independent replications, provided the data have additional structure that helps control dependence.
- **Quantifying uncertainty of statistical learning based on dependent data without independent replications:** In applications of data science, it is important to provide a disclaimer, stating how uncertain we are about statistical conclusions based on data. In scenarios when the number of parameters is unbounded and a single observation of dependent random variables is available, it is not obvious how to quantify uncertainty, because the 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 based on pseudolikelihood and composite likelihood functions, with applications to network data and other discrete and dependent data without independent replications.