Improve Survey Inference Using Bayesian Machine Learning

We consider survey inference from nonrandom samples in data-rich settings where high-dimensional auxiliary information is available both in the sample and the target population. When we have access to the individual-level data of the auxiliary variables in the population, we propose a regularized predictive inference approach that predicts the outcomes in the population based on the large number of auxiliary variables using Bayesian additive regression trees (BARTs) and its extensions. Our simulation studies reveal that the regularized predictions using BARTs yield valid inferences for the population means with coverage rates close to the nominal levels. We extend the method to accommodate two-phase designs, scenarios involving population data with confidentiality constraints, and cases where only the population margins of the auxiliary variables are available. We demonstrate the application of the proposed methods using health surveys.

Dr. Qixuan Chen is Associate Professor of Biostatistics at Columbia University. She obtained her PhD in Biostatistics from the University of Michigan in 2009. Her research focuses on survey sampling, missing data, measurement error, data integration, and Bayesian modeling. She collaborates extensively with interdisciplinary researchers on the design and analysis of longitudinal and cross-sectional health surveys at local, national, and international levels. Since 2018, Dr. Chen has served as Associate Editor for Biometrics.