University of California, Irvine
Statistics Seminar

Approximate Data Deletion and Replication with the Bayesian Influence Function

Ryan Giordano
Asst. Professor, Statistics
UC Berkeley

4 p.m., Tuesday, Oct. 29, 2024
6011 Donald Bren Hall

Many model-agnostic statistical diagnostics are based on repeatedly re-fitting a model with some observations deleted or replicated. Cross-validation, the non-parametric bootstrap, and outlier detection via case deletion are examples of this technique. However, for Bayesian statistical procedures based on Markov Chain Monte Carlo (MCMC), re-computing posteriors for many slightly different datasets can be computationally prohibitive. Instead of exactly re-fitting, one might use the entire dataset and a single MCMC run to form a linear approximation to the effect of re-weighting observations. In the robust statistics literature, the leading term of this linear approximation is known as the influence function. We show that, for Bayesian posteriors, the influence function takes the form of a set of easily-estimated posterior covariances, and that the error of the linear approximation vanishes asymptotically for finite-dimensional posteriors under standard regularity conditions. However, in models for which the number of parameters grows with the size of the data, N, we show that the error of the linear approximation based on the influence function does not vanish, even for finite-dimensional subsets of the parameters whose posterior does concentrate at a $\sqrt{N}$ rate. We discuss the implications for infinitesimal jackknife covariances, the bootstrap, and approximate cross-validation, as well what is implicitly meant by ``exchangeability" when using the influence function in this way.