Bayesian methods for variable selection in high-dimensional regression models

Crispin M. Mutshinda, PhD

Marine Macroecology and Biogeochemistry Lab

Dalhousie University, Halifax, NS, Canada

Abstract

Regression is arguably the most popular method in applied statistics. The statistical theory for estimation and hypothesis testing is well established for situations where the sample size n is smaller than the number p of explanatory variables. The Gauss-Markov theorem states that ordinary least square estimators are the best linear unbiased estimators of the regression coefficients. Moreover, Wald confidence intervals are valid, providing a tool for hypothesis testing about variable importance. These desirable properties break down when the covariates are strongly correlated predictors or when the number of explanatory variables exceeds the sample size, a situation often referred to as the "large p, small n" setting. While the $p > n$ situation poses challenges that are both theoretical and practical, "short-fat" data are increasingly common in modern data analysis due to technological advances in data collection techniques. A prevalent feature of short-fat data is that most explanatory variables are either irrelevant or redundant meaning that the underlying model is typically sparse, and it turns out that sparsity assumptions are critical for learning from such data. Variable selection methods aim at identifying a small subset of explanatory variables that can better explain the variation in the response variable. The literature on statistical methods and computational algorithms for parameter estimation and variable selection in high-dimensional regression models is quite extensive. In this talk, I review Bayesian stochastic search and shrinkage methods for variable selection in high-dimensional regression models and briefly discuss the advantages and limitations of each approach, drawing illustrative examples from my own past and on-going research.

KeyWords: Bayesian inference, Hamiltonian Monte Carlo, Markov chain Monte Carlo, regularization, stochastic search