Backfitting for large scale crossed random effects regressions

Swarnadip Ghosh, 4th-Year PhD Student

Stanford University

October 29, 2020

Abstract

Large-scale genomic and electronic commerce data sets often have a crossed random effects structure, arising from genotypes × environments or customers × products. Naive methods of handling such data will produce inferences that do not generalize. Regression models that properly account for crossed random effects can be very expensive to compute. The cost of both generalized least squares and Gibbs sampling can easily grow as $N^{(3/2)}$ (or worse) for N observations. Papaspiliopoulos, Roberts and Zanella (2020) present a collapsed Gibbs sampler that costs $O(N)$, but under an extremely stringent sampling model. We propose a backfitting algorithm to compute a generalized least squares estimate and prove that it costs $O(N)$ under greatly relaxed though still strict sampling assumptions. Empirically, the backfitting algorithm costs $O(N)$ under further relaxed assumptions. We illustrate the new algorithm on a ratings data set from Stitch Fix.

Making reliable predictions even when distributions shift

Suyash Gupta, 4th-Year PhD Student

Stanford University
October 29, 2020

Abstract

While the traditional viewpoint in machine learning and statistics assumes training and testing samples come from the same population, practice belies this fiction. One strategy—coming from robust statistics and optimization—is thus to build a model robust to distributional perturbations. In this talk, we present a different approach to describe procedures for robust predictive inference, where a model provides uncertainty estimates on its predictions rather than point predictions. We present a method that produces prediction sets (almost exactly) giving the right coverage level for any test distribution in an f-divergence ball around the training population. The method, based on conformal inference, achieves (nearly) valid coverage in finite samples, under only the condition that the training data be exchangeable. An essential component of our methodology is to estimate the amount of expected future data shift and build robustness to it; we develop estimators and prove their consistency for protection and validity of uncertainty estimates under shifts. By experimenting on several large-scale benchmark datasets, including Recht et al.'s CIFAR-v4 and ImageNet-V2 datasets, we provide complementary empirical results that highlight the importance of robust predictive validity.

This is joint work with Maxime Cauchois, Alnur Ali and John Duchi.

Market Impact in the Latent Order Book

Ismael Lemhadri, 4th-Year PhD Student

Stanford University
October 29, 2020

Abstract

In all modern financial markets, the concept of price impact is fundamental to the design of execution strategies. Price impact can represent a large fraction of transaction costs. Too much trading volume can degrade execution performance, or turn a winning strategy into a losing one. In this talk, I will give an overview of price impact models, focusing on the latent order book model of Donier et al. I derive some justification as well as theoretical results about the model. If time permits, I will also present evidence of market impact on Bitcoin data.

glmnet v4.0: Extending glmnet to all generalized linear models

Kenneth Tay, 5th-Year PhD Student

Stanford University

October 28, 2020

Abstract

The elastic net (Zou & Hastie 2005) is a popular penalized regression method which has a fast R implementation in the glmnet package (Friedman et al. 2010). The elastic net penalty can be applied to all generalized linear models (GLMs); however, before v4.0 glmnet only supported special GLM families. In this talk, we give some background on the algorithm for this generalization and explain how we developed glmnet v4.0 to support all GLM families.

Canonical Correlation Analysis in high dimensions with structured regularization

Elena Tuzhilina, 4th-Year PhD Student

Stanford University
October 28, 2020

Abstract

Canonical Correlation Analysis is a technique for measuring the association between two multivariate sets of variables. The Regularized modification of Canonical Correlation Analysis imposes L2 penalty on the CCA coefficients and is widely applied to the analysis of high dimensional data. The first part of the talk will cover the challenges that we have encountered while conducting Regularized CCA in the context of brain data analysis problems. These challenges arise from both dimensionality and particular structure of the data under consideration. The second part of this talk will be devoted to solutions that we have elaborated to address these challenges from both computational and theoretical aspects.

Generating random contingency tables via the Burnside process

Chenyang Zhong, 4th-Year PhD Student

Stanford University

October 28, 2020

Abstract

Contingency tables display frequency counts of categorical variables and are widely used in the analysis of experimental and survey data. Significance tests for such tables motivate the problem of generating contingency tables with fixed row and column sums uniformly at random. Burnside process is a class of Markov chains that sample from unlabelled structures. In this talk, I will present a fast algorithm for generating random contingency tables via the Burnside process. Mixing time analysis and practical performance of the algorithm will also be discussed.