Technometrics, ISSN 0040-1706
Volume 58, number 3 (august 2016): "Special Issue on Big Data"

Abstract
We introduce an efficient iterative algorithm, intended for various least squares problems, based on a design of experiments perspective. The algorithm, called orthogonalizing EM (OEM), works for ordinary least squares (OLS) and can be easily extended to penalized least squares. The main idea of the procedure is to orthogonalize a design matrix by adding new rows and then solve the original problem by embedding the augmented design in a missing data framework. We establish several attractive theoretical properties concerning OEM. For the OLS with a singular regression matrix, an OEM sequence converges to the Moore-Penrose generalized inverse-based least squares estimator. For ordinary and penalized least squares with various penalties, it converges to a point having grouping coherence for fully aliased regression matrices. Convergence and the convergence rate of the algorithm are examined. Finally, we demonstrate that OEM is highly efficient for large-scale least squares and penalized least squares problems, and is considerably faster than competing methods when n is much larger than p. Supplementary materials for this article are available online.

Abstract
Recent implementations of local approximate Gaussian process models have pushed computational boundaries for nonlinear, nonparametric prediction problems, particularly when deployed as emulators for computer experiments. Their flavor of spatially independent computation accommodates massive parallelization, meaning that they can handle designs two or more orders of magnitude larger than previously. However, accomplishing that feat can still require massive computational horsepower. Here we aim to ease that burden. We study how predictive variance is reduced as local designs are built up for prediction. We then observe how the exhaustive and discrete nature of an important search subroutine involved in building such local designs may be overly conservative. Rather, we suggest that searching the space radially, that is, continuously along rays emanating from the predictive location of interest, is a far thriftier alternative. Our empirical work demonstrates that ray-based search yields predictors with accuracy comparable to exhaustive search, but in a fraction of the time—for many problems bringing a supercomputer implementation back onto the desktop. Supplementary materials for this article are available online.

Abstract
RMarkov chain Monte Carlo (MCMC) methods have proven to be a very powerful tool for analyzing data of complex structures. However, their computer-intensive nature, which typically require a large number of iterations and a complete scan of the full dataset for each iteration, precludes their use for big data analysis. In this article, we propose the so-called bootstrap Metropolis–Hastings (BMH) algorithm that provides a general framework for how to tame

powerful MCMC methods to be used for big data analysis, that is, to replace the full data log-likelihood by a Monte Carlo average of the log-likelihoods that are calculated in parallel from multiple bootstrap samples. The BMH algorithm possesses an embarrassingly parallel structure and avoids repeated scans of the full dataset in iterations, and is thus feasible for big data problems. Compared to the popular divide-and-combine method, BMH can be generally more efficient as it can asymptotically integrate the whole data information into a single simulation run. The BMH algorithm is very flexible. Like the Metropolis–Hastings algorithm, it can serve as a basic building block for developing advanced MCMC algorithms that are feasible for big data problems. This is illustrated in the article by the tempering BMH algorithm, which can be viewed as a combination of parallel tempering and the BMH algorithm. BMH can also be used for model selection and optimization by combining with reversible jump MCMC and simulated annealing, respectively. Supplementary materials for this article are available online.

Compressing an Ensemble With Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature P. 319-328

Stefano Castruccio & Marc G. Genton

Abstract

One of the main challenges when working with modern climate model ensembles is the increasingly larger size of the data produced, and the consequent difficulty in storing large amounts of spatio-temporally resolved information. Many compression algorithms can be used to mitigate this problem, but since they are designed to compress generic scientific datasets, they do not account for the nature of climate model output and they compress only individual simulations. In this work, we propose a different, statistics-based approach that explicitly accounts for the space-time dependence of the data for annual global three-dimensional temperature fields in an initial condition ensemble. The set of estimated parameters is small (compared to the data size) and can be regarded as a summary of the essential structure of the ensemble output; therefore, it can be used to instantaneously reproduce the temperature fields in an ensemble with a substantial saving in storage and time. The statistical model exploits the gridded geometry of the data and parallelization across processors. It is therefore computationally convenient and allows to fit a nontrivial model to a dataset of 1 billion data points with a covariance matrix comprising of 1018entries. Supplementary materials for this article are available online.

Partitioning a Large Simulation as It Runs P. 329-340

Kary Myers, Earl Lawrence, Michael Fugate, Claire McKay Bowen, Lawrence Ticknor, Jon Woodring, Joanne Wendelberger & Jim Ahrens

Abstract

As computer simulations continue to grow in size and complexity, they present a particularly challenging class of big data problems. Many application areas are moving toward *exascale* computing systems, systems that perform 1018 FLOPS (FLoating-point Operations Per Second)—a billion billion calculations per second. Simulations at this scale can generate output that exceeds both the storage capacity and the bandwidth available for transfer to storage, making post-processing and analysis challenging. One approach is to embed some analyses in the simulation while the simulation is running—a strategy often called *in situ analysis*—to reduce the need for transfer to storage. Another strategy is to save only a reduced set of time steps rather than the full simulation. Typically the selected time steps are evenly spaced, where the spacing can be defined by the budget for storage and transfer. This article combines these two ideas to introduce an online in situ method for identifying a reduced set of time steps of the simulation to save. Our approach significantly reduces the data transfer and storage requirements, and it provides improved fidelity to the simulation to facilitate post-processing and reconstruction. We illustrate the method using a computer simulation that supported NASA's 2009 Lunar Crater Observation and Sensing Satellite mission.

High-Performance Kernel Machines With Implicit Distributed Optimization and Randomization P. 341-349

Haim Avron & Vikas Sindhwani

Abstract

We propose a framework for massive-scale training of kernel-based statistical models, based on combining distributed convex optimization with randomization techniques. Our approach is based on a block-splitting variant of the alternating directions method of multipliers, carefully reconfigured to handle very large random feature matrices under memory constraints, while exploiting hybrid parallelism typically found in modern clusters of multicore machines. Our high-performance implementation supports a variety of statistical learning tasks by enabling several loss functions, regularization schemes, kernels, and layers of randomized approximations for both dense and sparse datasets, in an extensible framework. We evaluate our implementation on large-scale model construction tasks and provide a comparison against existing sequential and parallel libraries. Supplementary materials for this article are available online.

Statistical Learning of Neuronal Functional Connectivity P. 350-359

Chunming Zhang, Yi Chai, Xiao Guo, Muhong Gao, David Devilbiss & Zhengjun Zhang

Abstract

Identifying the network structure of a neuron ensemble beyond the standard measure of pairwise correlations is critical for understanding how information is transferred within such a neural population. However, the spike train data pose significant challenges to conventional statistical methods due to not only the complexity, massive size, and large scale, but also high dimensionality. In this article, we propose a novel "structural information enhanced" (SIE) regularization method for estimating the conditional intensities under the generalized linear model (GLM) framework to better capture the functional connectivity among neurons. We study the consistency of parameter estimation of the proposed method. A new "accelerated full gradient update" algorithm is developed to efficiently handle the complex penalty in the SIE-GLM for large sparse datasets applicable to spike train data. Simulation results indicate that our proposed method outperforms existing approaches. An application of the proposed method to a real spike train dataset, obtained from the prelimbic region of the prefrontal cortex of adult male rats when performing a T-maze based delayed-alternation task of working memory, provides some insight into the neuronal network in that region.

Measuring Influence of Users in Twitter Ecosystems Using a Counting Process Modeling Framework P. 360-370

Donggeng Xia, Shawn Mankad & George Michailidis

Abstract

Data extracted from social media platforms are both large in scale and complex in nature, since they contain both unstructured text, as well as structured data, such as time stamps and interactions between users. A key question for such platforms is to determine influential users, in the sense that they generate interactions between members of the platform. Common measures used both in the academic literature and by companies that provide analytics services are variants of the popular web-search PageRank algorithm applied to networks that capture connections between users. In this work, we develop a modeling framework using multivariate interacting counting processes to capture the detailed actions that users undertake on such platforms, namely posting original content, reposting and/or mentioning other users' postings. Based on the proposed model, we also derive a novel influence measure. We discuss estimation of the model parameters through maximum likelihood and establish their asymptotic properties. The proposed model and the accompanying influence measure are illustrated on a dataset covering a five-year period of the Twitter actions of the members of the U.S. Senate, as well as mainstream news organizations and media personalities. Supplementary material is available online including computer code, data, and derivation details.

Discovering the Nature of Variation in Nonlinear Profile Data P. 371-382

Zhenyu Shi, Daniel W. Apley & George C. Runger

Abstract

Profile data have received substantial attention in the quality control literature. Most of the prior work has focused on the profile monitoring problem of detecting sudden changes in the characteristics of the profiles, relative to an in-control sample set of profiles. In this article, we present an approach for exploratory analysis of a sample of profiles, the purpose of which is to discover the nature of any profile-to-profile variation that is present over the sample. This is especially challenging in big data environments in which the sample consists of a stream of high-dimensional profiles, such as image or point cloud data. We use manifold learning methods to find a low-dimensional representation of the variation, followed by a supervised learning step to map the low-dimensional representation back into the profile space. The mapping can be used for graphical animation and visualization of the nature of the variation, to facilitate root cause diagnosis. Although this mapping is related to a nonlinear mixed model sometimes used in profile monitoring, our focus is on discovering an appropriate characterization of the profile-to-profile variation, rather than assuming some prespecified parametric profile model and monitoring for variation in those specific parameters. We illustrate with two examples and include an additional example in the online supplement to this article on the *Technometrics* website.

Variable Selection in a Log–Linear Birnbaum–Saunders Regression Model for High-Dimensional Survival Data via the Elastic-Net and Stochastic EM P. 383-392

Yukun Zhang, Xuewen Lu & Anthony F. Desmond

Abstract

The Birnbaum–Saunders (BS) distribution is broadly used to model failure times in reliability and survival analysis. In this article, we propose a simultaneous parameter estimation and variable selection procedure in a log–linear BS regression model for high-dimensional survival data. To deal with censored survival data, we iteratively run a combination of the stochastic EM algorithm (SEM) and variable selection procedure to generate pseudo-complete data and select variables until convergence. Treating pseudo-complete data as uncensored data via SEM makes it possible to incorporate iterative penalized least squares and simplify computation. We demonstrate the efficacy of our method using simulated and real datasets.

Online Updating of Statistical Inference in the Big Data Setting P. 393-403

Elizabeth D. Schifano, Jing Wu, Chun Wang, Jun Yan & Ming-Hui Chen

Abstract

We present statistical methods for big data arising from online analytical processing, where large amounts of data arrive in streams and require fast analysis without storage/access to the historical data. In particular, we develop iterative estimating algorithms and statistical inferences for linear models and estimating equations that update as new data arrive. These algorithms are computationally efficient, minimally storage-intensive, and allow for possible rank deficiencies in the subset design matrices due to rare-event covariates. Within the linear model setting, the proposed online-updating framework leads to predictive residual tests that can be used to assess the goodness of fit of the hypothesized model. We also propose a new online-updating estimator under the estimating equation setting. Theoretical properties of the goodness-of-fit tests and proposed estimators are examined in detail. In simulation studies and real data applications, our estimator compares favorably with competing approaches under the estimating equation setting. Supplementary materials for this article are available online.