Statistics and Its Interface Volume 9 (2016) 397–398

Editorial: Special Issue on Statistical and Computational Theory and Methodology for Big Data

The integration of computer technology into science and daily life has enabled the collection of big data, such as high-throughput biological assay data, large-scale genomic sequencing data, climate data, website transaction logs, and credit card records. Big data are collected on a massive scale in terms of volume, intensity, and complexity that exceed the capacity of standard analytic tools. Big data require drastic improvements in the state of the art methods for storage, analysis and interpretation which led to a revolution in science and technology. They also present challenges to the existing statistical and computational theory and methodology. This special issue reflects a variety of current research on the development of statistical methods and computational algorithms at the frontier of this vital and rapidly developing area.

The Call for papers for this special issue was first announced at the Banff workshop on Statistical and Computational Theory and Methodology for Big Data Analysis, which was held February 9–14, 2014. All submissions went through regular review process and four co-guest editors handled the peer reviews. Finally, 12 articles were selected to publish in this special issue, which cover a broad range of topics in handling and analyzing big data.

Wang, Chen, Schifano, Wu, and Yan summarize recent methodological and software developments in statistics that address the big data challenges. Methodologies are grouped into three classes: subsampling based, divide and conquer, and online updating for stream data. They also extend the online updating approach to variable selection with commonly used criteria, carry out a simulation study to examine their performances with stream data. Software packages are summarized with focuses on the open source R and R packages, covering recent tools that help break the barriers of computer memory and computing power.

The solution path clustering (SPC) method is capable to recognize noise and it also provides a short path of clustering solutions. However, the SPC method is not sufficiently fast for big datasets. **Marchetti and Zhou** propose a method that iterates between clustering a small subsample of the full data and sequentially assigning the other data points to attain orders of magnitude of computational savings. Their proposed method preserves the ability to isolate noise, includes a solution selection mechanism that ultimately provides one clustering solution with an estimated number of clusters, and is able to extract small tight clusters from noisy data. With implementation simplicity and the orders of magnitude of computational savings, their proposed methodology is quite appealing for big data. Smoothing spline ANOVA is a promising approach for extracting information from noisy data. However, the heavy computational cost may prevent this approach to be used for analyzing big data. **Helwig and Ma** propose a new algorithm by introducing rounding parameters to make the computation scalable for fitting smoothing spline ANOVA models to large data. They demonstrate that the rounding parameters algorithm takes only a few seconds on a standard laptop or tablet computer to fit a nonparametric regression model to very large samples. The U-statistic-based functional regression is a nonparametric method that allows direct estimation of higher-order moments without imposing assumptions on the mean structure. However, the estimation relies on a U-statistics-based estimating equation, which is generally too computationally expensive for big data. To overcome this computational barrier, **Xi and Lin** construct a computationally more succinct surrogate estimating equation using the “divide-and-conquer” strategy. They show that their proposed method significantly reduces the computational time and meanwhile enjoys the same asymptotic behavior as the original estimation method.

The singular value decomposition (SVD) is a key linear algebraic operation at the heart of many statistical and data mining methods. **Liang** develops a split-and-merge SVD algorithm. This new algorithm is particularly suitable for big data problems. To tackle the problem of robust covariance estimation in the “large p small n” setting, **Huang and Lee** develop COVariance Eigenvalue-Regularized estimation (*Cover*) based on eigenvalue regularization and robust *Cover* by incorporating Huber’s loss function into the estimation procedure. Cover is computationally efficient and enjoys good theoretical properties when p is fixed and there are no outliers while robust cover is resistant to outliers. **K. Chen** proposes a set of tools, including leverage score, information score, and approximated Cook’s distance, to perform model diagnostics and outlier detection for high-dimensional reduced-rank estimation.

The convolutional neural networks (CNNs) have proven to be a powerful tool for discriminative learning. **Dai, Lu, and Wu** carry out an in-depth investigation of generative modeling of CNNs. They construct a generative model for the CNN in the form of exponential tilting of a reference distribution, propose a generative gradient for pre-training CNNs by a non-parametric importance sampling scheme, and develop a generative visualization method for the CNNs by sampling from an explicit parametric image distribution. In the case of massive data sets, running many MCMC samplers become prohibitive due to the large number of likelihood calculations. In order to carry out Bayesian inference for a large set of time series, **Casarin, Craiu, and Leisen** propose a new algorithm that combines the "divide and conquer" idea based on parallel MCMC runs with a sequential MCMC strategy. Their approach yields important reductions in the computation time when the number of available cores is large and does not require storing the data on a single computer which is crucial when data volume is massive. **Liquet and Saracco** propose a new Sliced Inverse Regression (SIR) estimator (called BIG-SIR) of the effective Dimension Reduction direction by following the "divide and conquer" strategy. BIG-SIR can handle the analysis of big data and can be implemented using a combination of parallel strategy and shared memory structures, providing a solution for analyzing massive data sets which exceed the size of available RAM.

Principal Component Analysis (PCA) produces inconsistent estimators when the dimensionality is moderate to high. **Zhang and She** generalize sparse PCA to the broad exponential family distributions under high-dimensional setup, with built-in treatment for missing values, and propose a family of iterative sparse generalized PCA (SG-PCA) algorithms such that despite the non-convexity and nonsmoothness of the optimization task, the loss function decreases in every iteration. In terms of ease and intuitive parameter tuning, their sparsity-inducing regularization is superior to the popular Lasso. **Stein, van Dyk, and Kashyap** present an image segmentation framework for preprocessing such images in order to reduce the data volume while preserving as much thermal information as possible for later downstream analyses. They employ a parametric class of dissimilarities that can be expressed as cosine dissimilarity functions or Hellinger distances between nonlinearly transformed vectors of multi-passband observations in each pixel and develop a decision theoretic framework for choosing the dissimilarity that minimizes the expected loss that arises when estimating identifiable thermal properties based on segmented images rather than on a pixel-by-pixel basis.

Big data is still a rapidly growing research field. We hope that this special issue helps further stimulate new breakthrough researches and promote developing powerful statistical and computational methods in the era of big data. We also hope that this special issue makes *Statistics and Its Interface* (SII) a friendly home to many more exciting developments and innovations on big data research.

Ming-Hui Chen (Co-Guest Editor), University of Connecticut
Radu V. Craiu (Co-Guest Editor), University of Toronto
Faming Liang (Co-Guest Editor), University of Florida
Chuanhai Liu (Co-Guest Editor), Purdue University
Heping Zhang (Editor-in-Chief), Yale University