Editorial
Big Data? More Challenges!

Peihua Qiu

Editor, *Technometrics*

Recent advances in data acquisition technologies have led to massive amount of data being collected routinely in the physical, chemical, and engineering sciences as well as information sciences and technology. In addition to volume, the data often have complicated structure. Examples of such Big Data include the data streams obtained from complex engineering systems, image sequences, climate data, website transaction logs, credit card records, and so forth. Because of their big volume and complicated structure, big data are difficult to handle using traditional database management and statistical analysis tools. They create many new challenges for statisticians to describe and analyze them properly. To face the challenges and promote new statistical methods in handling big data, *Technometrics* decided to have a special issue on that topic in late 2013, and a guest editorial board was established soon after the decision. The board includes Drs. Ming-Hui Chen, Radu V. Craiu, Robert B. Gramacy, Willis A. Jensen, Faming Liang, Chuanhai Liu, and William Q. Meeker as associate editors, and me as editor. The Call for Papers was published in the journal and some other media in early 2014. We received 23 high-quality submissions before the deadline. All these papers went through the regular review procedure of the journal. Besides the people in the guest editorial board, some associate editors in the regular editorial board of the journal also help handle some submissions. Finally, 11 papers were selected to publish in the special issue, which cover a wide range of topics in describing, analyzing and computing big data. These papers are briefly discussed below.

The first five papers proposed numerical algorithms that can analyze big data fast. In the paper titled "Orthogonalizing EM: A design-based least squares algorithm" by Shifeng Xiong, Bin Dai, Jared Huling, and Peter Z. G. Qian, an efficient iterative algorithm intended for various least squares problems, based on a design of experiments perspective, was proposed. The algorithm, called orthogonalizing EM (OEM), works for ordinary least squares and can be extended easily 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. In the paper titled "Speeding up neighborhood search in local Gaussian process prediction" by Robert B. Gramacy and Benjamin Haaland, the authors suggested an algorithm for speeding up neighborhood search in local Gaussian process prediction that is commonly used in various non-linear and non-parametric prediction problems, particularly when deployed as emulators for computer experiments. The third paper titled "A bootstrap Metropolis-Hastings algorithm for Bayesian analysis of big data" by Faming Liang, Jinsu Kim and Qifan Song proposed a so-called bootstrap Metropolis-Hastings (BMH) algorithm that provided a general framework to tame powerful MCMC methods for big data analysis. The major idea of the algorithm 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 fourth paper titled "Compressing an ensemble with statistical models: an algorithm for global 3D spatio-temporal temperature" by Stefano Castruccio and Marc G. Genton suggested an algorithm for compressing 3D spatio-temporal temperature using statistics-based approach that explicitly accounted for the space-time dependence of the data. The fifth paper titled "Partitioning a Large Simulation as It Runs" by Kary Myers, Earl Lawrence, Michael Fugate, Claire McKay Bowen, Lawrence Ticknor, Jon Woodring, Joanne Wendelberger, and Jim Ahrens was about analysis of data streams, in which data were generated sequentially and data storage, transferring and analysis were all challenging. The authors suggested a so-called online *in situ* method for identifying a reduced set of time steps of the data and data analysis results to save in the storage facility, in order to significantly reduce the data transfer and storage requirements.

The next two papers were about machine learning methods for handling big data. The first paper titled "High-performance kernel machines with implicit distributed optimization and randomization" by Vikas Sindhwani and Haim Avron proposed a framework for massive-scale training of kernel-based statistical models, based on combining distributed convex optimization with randomization techniques. The second paper titled "Statistical learning of neuronal functional connectivity" by Chunming Zhang, Yi Chai, Xiao Guo, Muhong Gao, David Devilbiss, and Zhengjun Zhang was on identifying the network structure of a neuron ensemble beyond the standard measure of pairwise correlations, which was critical for understanding how information was transferred within such a neural population. The spike train data posed a significant challenge to conventional statistical methods due to not only the complexity, massive size and large scale, but also the high dimensionality. In this paper, the authors proposed a novel "*Structural Information Enhanced*" (SIE) regularization method for estimating the conditional intensities under the generalized linear model (GLM)

TECHNOMETRICS, AUGUST 2016, VOL. 58, NO. 3
DOI: 10.1080/00401706.2016.1196946

framework to better capture the functional connectivity among neurons.

The last four papers are on some specific big data problems. The paper titled "Measuring influence of users in Twitter ecosystems using a counting process modeling framework" by Donggeng Xia, Shawn Mankad, and George Michailidis was about analyzing data extracted from social media platforms, such as Twitter, which were both large in scale and complex in nature, since they contained both unstructured text, as well as structured data, such as time stamps and interactions between users. In this paper, the authors developed a modeling framework using multivariate interacting counting processes to capture the detailed actions that users undertook on such platforms, namely posting original content, reposting and/or mentioning other users' postings. Profile monitoring is an important problem in manufacturing industries. The paper titled "Discovering the Nature of Variation in Nonlinear Profile Data" by Zhenyu Shi, Daniel W. Apley, and George C. Runger proposed a method for exploratory analysis of a sample of profiles for the purpose to discover the nature of any profile-to-profile variation that was present over the sample. The next paper titled "Variable selection in a log-linear Birnbaum-Saunders regression model for high-dimensional survival data via the elastic-net and stochastic EM" by Yukun Zhang, Xuewen Lu, and Anthony F. Desmond proposed a simultaneous parameter estimation and variable selection procedure in a log-linear Birnbaum-Saunders regression model for analyzing high-dimensional survival data. The last paper of the special issue titled "Online updating of statistical inference in the big data setting" by Elizabeth D. Schifano, Jing Wu, Chun Wang, Jun Yan, Ming-Hui Chen developed iterative estimating algorithms and statistical inferences for linear models and estimating equations for analyzing big data arising from online analytical processing, where large amounts of data arrived in streams and required a fast analysis without storage/access to the historical data.

The 11 papers described above will definitely play an important role in the big data analysis literature. I would like to thank the contributing authors, the reviewers, and the associate editors who handled the submitted manuscripts to the special issue for their help and support to make this special issue possible. I also want to thank the *Technometrics* Management Committee, the ASA Journal Manager Eric Sampson, and the journal editorial assistant Janet Wallace for their support, help and assistance. Finally, I want to thank all readers for your reading of the papers in this issue which was prepared specifically for your research and applications.