Statistical and Computational Theory and Methodology for Big Data Analysis

Ming-Hui Chen (University of Connecticut)
Radu Craiu (University of Toronto)
Faming Liang (Texas A&M University)
Chuanhai Liu (Purdue University)

February 9, 2014 – February 14, 2014

1 Overview of the Field

The integration of computer technology into science and daily life has enabled the collection of massive volumes of data, such as high-throughput biological assay data, climate data, website transaction logs, and credit card records. However, such big data sets cannot be practically analyzed on a single commodity computer because their sizes are too large to fit in memory or it is too time consuming to process when the current statistical methods are used. To circumvent this obstacle, one may have to resort to parallel and distributed architectures, with multicore and cloud computing platforms providing access to hundreds or thousands of processors. While the parallel and distributed architectures present new capabilities for storage and manipulation of data, from an inferential point of view, it is unclear how the current statistical methodology can be transported to the paradigm of big data. Also, with growing size typically comes a growing complexity of data structures, of the patterns in the data, and of the models needed to account for the patterns. Big data has put a great challenge on the current statistical methodology.

2 Recent Developments and Open Problems

There are several algorithms that are recently developed and feasible for statistical inference of big data and workable on parallel machines, including the bag of little bootstraps [1], aggregated estimation equation [2, 3], split-and-conquer algorithms [4], and the subsampling-based stochastic approximation algorithm [5]. The bag of little bootstraps algorithm is designed to assess the quality of an estimator, which functions by averaging the results of bootstrapping multiple small subsets of the original data. The small subsets can be processed in parallel, each on an individual or a very small sets of computer nodes. The aggregated estimation equation and split-and-conquer algorithms are based on the same idea of divide-and-conquer, but focus on different types of problems; the former is for parameter estimation and the latter for variable selection of regression models. In [5], a general principle for big data analysis is proposed: i.e., using Monte Carlo averages, that are calculated from subsamples in parallel, to approximate the quantities that originally need to calculate from the full data. Under this principle, a general parameter estimation approach, maximum mean log-likelihood estimation, is developed in [5] for big data models based on the technique of stochastic approximation.

On the other hand, iterative algorithms have been widely used in the current society of scientific computing. Examples of such iterative algorithms include various Markov chain Monte Carlo (MCMC) algorithms [6, 7, 8], and the EM algorithm [9], which typically require a large number of iterations and a complete scan of the full dataset for each iteration. The MCMC algorithms are rooted in the work of physicists such as Metropolis and von Neumann during the period 1945-55 when they employed modern computers for simulations of some probabilistic problems in atomic bomb designs. After six decades of development, it has proven to be very powerful and typically unique computational tools for analyzing data of complex structures. The EM algorithm represents a hallmark achievement in the history of statistics, and has been widely used in scientific computing for parameter estimation in presence of missing data. Given the successes of the iterative algorithms in modern scientific computing, it would be of great interest to develop some innovative iterative algorithms that are feasible for big data.

There have been significant advances made by the statistical community on big data research. One of open problems is how to generalize and scale up such proposed techniques to the true big data settings. One of the key features of big data is that the statistical methods, which work well on small-scale datasets, usually perform poorly in big data settings. Thus, it is not easy to expect the performance of those statistical methods for the big data problems. The field of neuroimaging has also witnessed big progress made in integratively analyzing imaging data of multiple subjects and multiple modalities, together with genomic data. Other open problems include: (i) to have a better understand of big data and associated statistical issues; (ii) to think more carefully about how to solve big data issues; and (iii) to have a more concrete focus on big data problems. The workshop participant, Christophe Andrieu of the University of Bristol, suggested that an open problem is currently to go beyond "linear regression" and logit type strategies. Since it seems that most of the recent work on big data has been confined to this, is this enough? Should one bother with more complicated models? What would be the gains?

3 Presentation Highlights

3.1 Day 1: February 10, 2014

The workshop kicked off with the American Statistical Association (ASA) video from the Executive Director, Ron Wasserstein. Ron started with a warm greeting to the workshop participants and then presented the recent ASA activities and initiatives on big data. The morning Session II on February 10 featured two presentations delivered by Hongzhe Li of the University of Pennsylvania and Heping Zhang of the Yale University. Hongzhe's presentation was on "Microbiome, Metagenomics and High Dimensional Compositional Data Analysis". Human gut microbiome plays an important role in human health and disease. Next generation sequencing technologies have made it possible to study all microbes in human gut in an unbiased way. Analysis of such large volumes of reads data, usually in 100s of terabytes, raises many challenges in statistical analysis and computation. He presented several methods for analysis of such data, including model-based methods for quantifying the composition of all bacteria and methods for analysis of high dimensional compositional data. These methods have showed close associations between diets and microbiome composition and microbiome and obesity. Heping's presentation was on "Tree-based Rare Variants Analyses". Heping introduced a tree-based method that adopts a non-parametric disease model and is capable of exploring gene-gene interactions. Their method outperforms the sequence kernel association test (SKAT) in most of our simulation scenarios, and by notable margins in some cases. By applying the tree-based method to the Study of Addiction: Genetics and Environment (SAGE) data, they successfully detected gene CTNNA2 and its 44 specific variants that increase the risk of alcoholism in women. This gene has not been detected in the SAGE data. Post hoc literature search also supports the role of CTNNA2 as a likely risk gene for alcohol addiction. This finding suggests that their tree-based method can be effective in dissecting genetic variants for complex diseases using rare variants data.

Marc A. Suchard of the University of California Los Angeles and Minge Xie of Rutgers University were the presenters in the afternoon Session I on February 10. Marc's presentation was on "When Multi-Core Statistical Computing Fails for Massive Sample Sizes ...". Much of statistical computing is memory-bandwidth limited, not floating-pointing operation throughput limited as commonly assumed. This often restricts the utility of multi-core computing techniques to improve statistical estimation run-time. Marc explored this conundrum in inference tools for a massive Bayesian model of sea-surface temperatures across the global and further described approaches for computing the data likelihood that exploit fine-scale parallelization for potential scalability to real-time satellite surveillance data. These simple algorithmic changes open the door on using advancing computing technology involving many-core architectures. These architectures provide significantly higher memory-bandwidth and inexpensively afford order-of-magnitude run-time speed-ups. Minge presented "A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data" [4]. If there are extraordinarily large data, too large to fit into a single computer or too expensive to perform a computationally intensive data analysis, what should we do? To deal with this problem, Minge discussed a "split-and-conquer" approach and illustrated it using several computationally intensive penalized regression methods, along with a theoretical support. Specifically, consider a regression setting of generalized linear models with n observations and p covariates, in which n is extraordinarily large and p is either bounded or goes to ∞ at a certain rate of n. They proposed to randomly split the data of size n into K subsets of size $O(n/K)$. For each subset of data, they performed a penalized regression analysis and the results from each of the K subsets are then combined to obtain an overall result. They showed that under mild conditions the combined overall result still retains desired properties of many commonly used penalized estimators, such as the model selection consistency and asymptotic normality. When K is well controlled, they also showed that the combined result is asymptotically equivalent to the result of analyzing the entire data all at once (assuming that there is a super computer that could carry out such an analysis). In addition, when a computational intensive algorithm is used, they showed that the split-and-conquer approach can substantially reduce computing time and computer memory requirement. Furthermore, they demonstrated that the approach has an inherent advantage of being more resistant to false model selections caused by spurious correlations. Similar to what reported in the literature, they established an upper bound for the expected number of falsely selected variables and a lower bound for the expected number for truly selected variables. The proposed methodology was illustrated numerically using both simulation and real data examples.

In the afternoon Session II on February 10, there were three presentations delivered by Ping Li of Rutgers University, Christophe Andrieu of University of Bristol, and Lingsong Zhang of Purdue University. Ping presented "BigData: Efficient Search and Learning Using Sparse Random Projections and Probabilistic Hashing". Modern applications of search and learning have to deal with datasets with billions of examples in billion or even billion square dimensions (e.g., text documents represented by high-order n-grams). Ping first presented the use of very sparse random projections for learning with high-dimensional data. It is evident that the projection matrix can be extremely sparse (e.g., 0.1% or less nonzeros) without hurting the learning performance. For binary sparse data (which are common in practice), however, b-bit minwise hashing turns out to be much more efficient than random projections. In addition, the recent development of one-permutation hashing [10] substantially reduced the processing time of (b-bit) minwise hashing, from (e.g.,) 500 permutations to merely one. There are many other exciting new progresses in the basic research of random projections and hashing, for example, the new work on sign Cauchy random projections for approximating chi-square distances [11] and the work on using stable random projections for very fast and accurate compressed sensing [12]. Christophe's presentation was on "Uniform Ergodicity of the Iterated Conditional SMC and Geometric Ergodicity of Particle Gibbs Samplers" [13]. He discussed the quantitative bounds for rates of convergence and asymptotic variances for iterated conditional sequential Monte Carlo (i-cSMC) Markov chains and associated particle Gibbs samplers. Their main findings are that the essential boundedness of potential functions associated with the i-cSMC algorithm provide necessary and sufficient conditions for the uniform ergodicity of the i-cSMC Markov chain, as well as quantitative bounds on its (uniformly geometric) rate of convergence. Lingsong presented "Scale-Space Inference with Application on Spatial Clustering Detection". He proposed a novel multi-resolution cluster detection (MCD) method to identify irregularly shaped clusters in space and derived the multi-scale test statistic on a single cell based on likelihood ratio statistic for Bernoulli sequence, Poisson sequence and Normal sequence. A neighborhood variability measure is defined to select the optimal test threshold. In his presentation, the MCD method was compared with single scale testing methods controlling for false discovery rate and the spatial scan statistics using simulation and f-MRI data, and the MCD method was shown to be more effective for discovering irregularly shaped clusters. The implementation of his proposed method does not require heavy computation, making it suitable for cluster detection for large spatial data.

3.2 Day 2: February 11, 2014

In the morning Session I on February 11, Peihua Qiu of the University of Florida presented "On Nonparametric Profile Monitoring". Quality of a process is often characterized by the functional relationship between a response and one or more predictors. Profile monitoring is for checking the stability of this relationship over time. In the literature, most existing control charts are for monitoring parametric profiles, and they assume that within-profile observations are independent of each other, which is often invalid. In this presentation, he discussed some of their recent research on nonparametric profile monitoring when within-profile data are correlated. He also briefly described the problems of online image monitoring and dynamic disease screening that are closely related to profile monitoring. Hongtu Zhu of the University of North Carolina presented "Functional Analysis of Big Neuroimaging Data". Motivated by recent work on studying massive imaging data in various neuroimaging studies, Hongtu's group proposed several classes of spatial regression models including spatially varying coefficient models, spatial predictive Gaussian process models, tensor regression models, and Cox functional linear regression models for the joint analysis of large neuroimaging data and clinical and behavioral data. Their statistical models explicitly account for several stylized features of neuorimaging data: the presence of multiple piecewise smooth regions with unknown edges and jumps and substantial spatial correlations. They developed some fast estimation procedures to simultaneously estimate the varying coefficient functions and the spatial correlations. They also systematically investigated the asymptotic properties (e.g., consistency and asymptotic normality) of the multiscale adaptive parameter estimates. Their Monte Carlo simulation and real data analysis confirmed the excellent performance of their models in different applications.

In the morning Session II on February 11, Jian Zhang of the University of Kent presented "High-dimensional Inference in Magnetoencephalographic Neuroimaging". Reconstructing neural activities using non-invasive sensor arrays outside the brain is an ill-posed inverse problem since the observed (Magnetoencephalography) MEG sensor measurements could result from an infinite number of possible neuronal sources [14]. MEG data can be complex and of large scale, in particular, when multiple trials and multiple subjects are involved. Should we build a big model for such kinds of large scale data? In this presentation, he focused on a local approach which contains a series of small local models in the source space and is scalable to parallel computing. They proposed a family of procedures called beamformers by using covariance thresholding [15]. A general theory was developed on how their spatial and temporal dimensions determine their performance. Conditions were provided for the convergence rate of the associated beamformer estimation and the implications of the theory were illustrated by simulations and a real data analysis on face-perception MEG data [16]. Momiao Xiong of the University of Texas Health Science Center at Houston discussed "Classification Analysis of Big Image Data". Due to advances in sensors, growing large and complex medical images provides invaluable information for holistic discovery of the genetic and epigenetic structure of disease and has the potential to enhance diagnosis of disease, prediction of clinical outcomes, characterization of disease progression, management of health care and development of treatments, but also pose great methodological and computational challenges. An enormous amount of increasingly larger, more complex and more diverse demand developing unified frameworks and novel statistical methods for cluster and classification analysis of medical image data, which will provide low-cost and powerful tools for early detection and efficient management of complex diseases such as cancers, mental disorders, vascular diseases. The medical images have the ability to visualize the pathology change in the cellular or even the molecular level or anatomical changes in tissues and organs. However, the medical images for the same type of disease from different individuals might be quite similar. As a result, it is a big challenge to extract the key information from a large amount of medical images for early detection of the complex diseases and the prediction of the drug response. To address this issue, he presented an extension of one dimensional functional principal component analysis to the two dimensional functional principle component analysis (2DFPCA). To reduce high dimensional image data to low dimensional space, they developed novel space sufficient dimension reduction methods to select variables. The proposed methods were applied to 250 liver cancer histology image data (99 tumor tissues and 151 normal tissues) and 176 ovarian cancer histology images with the drug response status from TCGA database. For the liver cancer dataset, they obtained almost 84%, 79.8% and 86.8% classification accuracy, sensitivity and specificity, respectively. For the ovarian cancer drug response dataset, classification accuracy, sensitivity and specificity were 80.1%, 85.8% and 71.4%, respectively.

The afternoon on February 11 featured two tutorial and review sessions on the recent developments of statistical methods and computational methods for big data analysis. In the afternoon Session I on February 11, Faming Liang of the Taxes A&M University presented "Recent Developments of Iterative Monte Carlo Methods for Big Data Analysis". Iterative Monte Carlo methods, such as MCMC, stochastic approximation, and EM, have proven to be very powerful tools for statistical data analysis. 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. Faming provided an overview of the recent developments of iterative Monte Carlo methods for big data analysis. Chuanhai Liu of the Purdue University presented "Big Data Analysis and Beyond" Based on his current understanding of big data, which essentially says that there is no big data but only bigger data (i.e., data that cannot be analyzed efficiently using a single computer), Chuanhai listed three challenges: (1) modeling - how uncertainly data are related to scientific questions, (2) inference - how data can adequately be converted to knowledge, and (3) computing - how inferential output can be computed from big data. In addition to agreeing with the common recognition that approximate inference can be made by making use of MapReduce, he pointed it out that improved results can be obtained using an iteratveMapReduce-type computational framework, which he developed in R at Purdue for his topic course on massive data analysis. Due to both time limit and overwhelming interest from participants, in the remaining of his presentation he focused on a new inferential framework called inferential models [17, 18, 19, 20]. Unlike all other existing schools of thought, this new framework produces scientifically desirable prior-free and frequency-calibrated probabilistic inference. While it is apparently difficult to introduce such a new framework in 30 minutes to cover its philosophy, theory, computation, and application, Chuanhai delivered a clear message: due to the lack of a solid foundation in statistics, developing such inferential methods is critically important for scientific inference, especially from big data. This calls for attention on research on foundations of statistics, a topic well beyond but fundamentally important in big data analysis.

In the afternoon Session II on February 11, Jun Yan of the University of Connecticut delivered "A Partial Review of Software for Big Data Statistics". Big data brings challenges to even simple statistical analysis because of the barriers in computer memory and computing time. The computer memory barrier is usually handled by a database connection that extracts data in chunks for processing. The computing time barrier is handled by parallel computing, often accelerated by graphical processing units. In this partial review, Jun summarized the open source R packages that break the computer memory limit such as biglm and bigmemory, as well as the academic version of the commercial Revolution R, and R packages that support parallel computing. Products from commercial software will also be sketched for completeness. This review work was a joint effort of Jun Yan, Ming-Hui Chen, Elizabeth Schifano, Chun Wang, and Jing Wu of the University of Connecticut.

3.3 Day 3: February 12, 2014

In the morning Session I on February 12, Kun Chen of the University of Connecticut presented "Sparse and low-rank Regression in High Dimensions". The talk discussed the combination of distinct but interrelated dimension reduction techniques in fitting high-dimensional multivariate models. This results in models with some composite low-dimensional structures, which often enjoy enhanced interpretability and improved predictive accuracy. Kun also presented a general methodology and some theoretical results for recovering a sparse and low-rank matrix structure from noisy data, in both unsupervised and supervised learning problems. Linglong Kong of the Purdue University presented "Quantile Regression in Variable Screening". Linglong first introduced a quantile regression framework for linear and nonlinear variable screening with high-dimensional heterogeneous data. Motivated by success of various variable screening methods, especially the quantile-adaptive framework, He discussed how to combine the information from different quantile levels to provide more efficient variable screening procedure. In particular, there are two ways to do so: one is to simply take (weighted) average across different levels of quantile regression; the other one is to use (weighted) composite quantile regression. Asymptotically, these two approaches are equivalent in terms of efficiency. Numerical studies confirm the fine performance of the proposed method for various linear and nonlinear models.

In the morning Session II on February 12, Ying Nian Wu of the University of California at Los Angeles presented "What Is Beyond Sparse Coding?" Many types of data such as natural images admit sparse representations by redundant dictionaries of basis functions (or regressors), and these dictionaries can either be designed or learned from training data. However, it is still unclear how to go beyond sparsity and continue to learn structures behind the sparse representations. Ying Nian reviewed some recent progresses and the major issues and difficulties that need to be addressed. He also presented our own recent work that seeks to learn dictionaries of compositional patterns in the sparse representations. Xiao Wang of the Purdue University presented "Functional Regression and Image Regression" based on his recent papers [21, 22]. Xiao first discussed generalized functional linear models with scalar responses and image predictors. Predicting clinical outcomes on the basis of quantitative image data is quite promising in current psychiatric neuroimaging research. In his talk, he considered prediction with image predictors in the framework of functional linear model and bounded total variation space. The slope image is assumed to belong to the space of bounded total variation. He presented near-optimal guarantee for stable recovery of the slope image using total variation minimization and nonasymptotic error bounds on the excess risk by exploiting various techniques from compressive sensing, as well as the approximation theory. These techniques allows us to obtain finite-sample bounds that hold with high probability, and are specified explicitly in terms of the sample size and the image size.

The afternoon Session I on February 12 featured three presentations delivered by Xiaotong Shen of the University of Minnesota, Xiaojing Wang of the University of Connecticut, and Guanghua Xiao of the UT Southwestern Medical Center. Xiaotong's talk was on "Sentiment Analysis". Sentiment analysis identifies the relevant content as well as determines and understands opinions, from documents or texts, towards a specific event of interest. In this presentation, Xiaotong discussed large margin methods for ordinal classification involving word predictors, where imprecise information is available for prediction regarding linguistic relations among predictors, expressed in terms of a directed graph. Then the methods are used for sentiment analysis, where sentiment function representations of words are derived, on which the imprecise predictor relations are integrated as linear relational constraints over sentiment function coefficients. Computational and theoretical aspects were discussed, in addition to an application to opinion survey. Xiaojing presented "A Bayesian Approach to Subgroup Identification". Xiaojing discussed subgroup identification, the goal of which is to determine the heterogeneity of treatment effects across subpopulations. Searching for differences among subgroups is challenging because it is inherently a multiple testing problem with the complication that test statistics for subgroups are typically highly dependent, making simple multiplicity corrections such as the Bonferroni correction too conservative. In this talk, Xiaojing presented a Bayesian approach to identify subgroup effects, with a scheme for assigning prior probabilities to possible subgroup effects that accounts for multiplicity and yet allows for (pre-experimental) preference to specific subgroups. The analysis utilizes a new Bayesian model selection methodology and, as a byproduct, produces individual probabilities of treatment effect that could be of use in personalized medicine. Xiaojing illustrated the analysis using an example involving subgroup analysis of biomarker effects on treatments. Guanghua presented "Detection of Tumor Driver Genes Using a Fully Integrated Bayesian Approach". DNA copy number alterations (CNAs), including amplifications and deletions, can result in significant changes in gene expression, and are closely related to the development and progression of many diseases, especially cancer. For example, CNA-associated expression changes in certain genes (called tumor driver genes) can alter the expression levels of many downstream genes through transcription regulation, and cause cancer. Identification of such tumor driver genes leads to discovery of novel therapeutic targets for personalized treatment of cancers. Several approaches have been developed for this purpose by using both copy number and gene expression data. In this talk, Guanghua discussed a Bayesian approach to identify tumor driver genes, in which the copy number and gene expression data are modeled together, and the dependency between the two data types is modeled through conditional probabilities. The joint modeling approach can identify CNA and differentially expressed (DE) genes simultaneously, leading to improved detection of tumor driver genes and comprehensive understanding of underlying biological processes. Guanghua also presented simulation studies to evaluate their proposed method and then applied their method to a head and neck squamous cell carcinoma (HNSCC) dataset. Both simulation studies and data application showed that the joint modeling approach can significantly improve the performance in identifying tumor driver genes, when compared to other existing approaches.

The third day (February 12) of the workshop concluded with two featured presentations by Elizabeth D. Schifano of the University of Connecticut on "Online Updating of Statistical Inference in the Big Data Setting" and Nan Lin of the Washington University in St. Louis on "Statistical Aggregation in Massive Data Environment". Elizabeth presented statistical regression 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, Elizabeth and her collaborators (Ming-Hui Chen, Jun Yan, Chun Wang, Jing Wu of the University of Connecticut) developed iterative estimating algorithms and statistical inferences for linear models and estimating equations that update as new data arrive. Elizabeth introduced predictive residuals in the online-updated linear model setting that can be used to test the goodness-of-fit of the hypothesized model, as well as a new online-updated estimator under the estimating equation framework that has less bias in finite samples as compared to other online-updated estimators. In simulation studies, their approaches compared favorably with competing approaches in terms of timing and accuracy. Due to their size and complexity, massive data sets bring many computational challenges for statistical analysis, such as overcoming the memory limitation and improving computational efficiency of traditional statistical methods. In Nan's talk, he discussed the statistical aggregation strategy to conquer such challenges posed by massive data sets. Statistical aggregation partitions the entire data set into smaller subsets, compresses each subset into certain low-dimensional summary statistics and aggregates the summary statistics to approximate the desired computation based on the entire data. Results from statistical aggregation are required to be asymptotically equivalent. Statistical aggregation is particularly useful to support sophisticated statistical analyses for online analytical processing in data cubes. Nan detailed its application to two large families of statistical methods, estimating equation estimation and U-statistics.

3.4 Day 4: February 13, 2014

In the morning Session I on February 13, Hongyu Zhao of the Yale University could not attend the workshop but sent his presentation slides to all the workshop participants. Hongyu's slides were on "Detecting Genetic Association Signals Leveraging Network Information". Although Genome Wide Association Studies (GWAS) have identified many sceptibility loci for common diseases, these loci only explain a small portion of heritability. It is challenging to identify the remaining disease loci because their association signals are likely weak and difficult to identify among millions of candidates. One potentially useful direction to increase statistical power is to incorporate pathway and functional genomics information to prioritize GWAS signals. In his slides, he first described a method to utilize network information to prioritize disease genes based on the "guilt by association" principle, in which networks are treated as static, and disease associated genes are assumed to locate closer with each other than random pairs in the network. Hongyu then introduced a novel "guilt by rewiring" principle that postulates that disease genes more likely undergo rewiring in disease patients, whereas most of the network is unaffected in disease condition. A Markov random field framework was used for both methods to integrate network information to prioritize genes. Applications in Crohn's disease and Parkinson's disease show that these methods lead to more replicable and biologically meaningful results. Zhang Zhang of the Beijing Institute of Genomics at Chinese Academy of Sciences, China presented "Biocuration in the Era of Big Data". Biology enters the era of big data. More than a dozen biological wikis (bio-wiki) have been constructed to call on community intelligence in big biological data curation. However, one of the major limitations in bio-wikis is insufficient participation from the scientific community, which is intrinsically because of lack of explicit authorship and thus no credit for community-curated contributions. To increase community curation in bio-wikis, Zhang and his collaborators developed AuthorReward [23] to reward community-curated efforts by contribution quantification and explicit authorship. It quantifies researchers' contributions by properly factoring both edit quantity and quality and yields automated explicit authorship according to their quantitative contributions. They also constructed RiceWiki[24] (http://ricewiki.big.ac.cn), a wiki-based, publicly editable, and open-content platform for community curation of rice genes. To test the functionality of AuthorReward, they installed it in RiceWiki. As testified in RiceWiki, AuthorReward is capable of yielding sensible quantitative contributions and providing automated explicit authorship, consistent well with perceptions of all participated contributors. Based on collective intelligence, RiceWiki bears the potential to deal with big data and make it possible to build a rice encyclopedia by and for the scientific community.

In the morning Session II on February 13, Xin Gao of the King Abdullah University of Science and Technology (KAUST) presented "Poly(A) motif prediction using spectral latent features from human DNA sequences" [25]. They propose a novel machine learning method for poly(A) motif prediction by marrying generative learning (hidden Markov models) and discriminative learning (support vector machines). Generative learning provides a rich palette on which the uncertainty and diversity of sequence information can be handled, while discriminative learning allows the performance of the classification task to be directly optimized. They employed hidden Markov models for fitting the DNA sequence dynamics, and developed an efficient spectral algorithm for extracting latent variable information from these models. These spectral latent features were then fed into support vector machines to fine tune the classification performance. The proposed method was evaluated on a comprehensive human poly(A) dataset that consists of 14,740 samples from 12 of the most abundant variants of human poly(A) motifs. Compared with one of previous state-of-art methods in the literature (the random forest model with expert-crafted features), our method reduces the average error rate, false negative rate and false positive rate by 26%, 15% and 35%, respectively. Matthias Katzfuss of the Texas A&M University presented "Statistical Inference for Massive Distributed Spatial Data Using Low-Rank Model" Matthias's talk focused on computationally feasible spatial inference and prediction approaches using spatial low-rank models, for analyzing massive distributed spatial data. The proposed methods adopt the divide-and-conquer strategy to dealing with massive spatial data, which allow local spatial modeling and computations at separated data servers with minimal communication among them. The proposed methods can be very useful in dealing with contemporaneous large-scale spatial-temporal applications.

In the afternoon Session I on February 13, Ruslan Salakhutdinov of the University of Toronto presented "Annealing Between Distributions by Averaging Moments" [26]. Many powerful Monte Carlo techniques for estimating partition functions, such as annealed importance sampling (AIS), are based on sampling from a sequence of intermediate distributions which interpolate between a tractable initial distribution and the intractable target distribution. The near-universal practice is to use geometric averages of the initial and target distributions, but alternative paths can perform substantially better. Ruslan presented a novel sequence of intermediate distributions for exponential families defined by averaging the moments of the initial and target distributions. He discussed and analyzed the asymptotic performance of both the geometric and moment averages paths and derive an asymptotically optimal piecewise linear schedule. AIS with moment averaging performs well empirically at estimating partition functions of restricted Boltzmann machines (RBMs), which form the building blocks of many deep learning models, including Deep Belief Networks and Deep Boltzmann Machines. Alexander Y. Shestopaloff of the University of Toronto presented "MCMC for Non-Linear State Space Models Using Ensembles of Latent Sequences" [27]. Alexander introduced a new MCMC method for non-linear, non-Gaussian state space models using Embedded HMM MCMC and Ensemble MCMC. In contrast to existing methods, which only consider a single state sequence during parameter sampling, their new method considers an enormously large ensemble of latent state sequences at once. This allows making larger proposals while keeping a high acceptance rate, leading to more efficient sampling. He showed that when applied to the problem of Bayesian inference in the Ricker model of population dynamics, their new MCMC method improves performance relative to methods that only look at a single state sequence during sampling. Xiaoyi Min of the Yale School of Public Health presented "Detection of Chromosome Copy Number Variations in Multiple Sequences". Xiaoyi introduced the concept of DNA copy number variation (CNV) and its potential role in human complex diseases. To help identify inherited CNVs which are generally short and common in the population, he introduced an extension to the Screening and Ranking Algorithm [28] which integrates information from multiple samples. In particular, Xiaoyi and his collaborators proposed an adaptive Fisher's method for combining screening statistics across samples, which has a high power regardless of the carrier proportion of the CNV. Furthermore, Xiaoyi gave both theoretical and numerical results to demonstrate that this method performs better than other current methods. Profs. Peihua Qiu, Min-ge Xie, and Hongtu Zhu gave many insightful comments and suggestions from different aspects after his presentation.

The afternoon Session II on February 13 was the last session of this workshop. Lee H. Dicker of the Rutgers University presented "Variance estimation in high-dimensional linear models" and Philip Gautier of the Purdue University presented "Divide & Recombine for Large Complex Data: Likelihood Modelling for Logistic Regression". Lee's talk focused on the role of variance in model fitting procedures for high-dimensional data. Lee provided an overview of popular methods for variance estimation, emphasizing the role of efficiency. The residual variance and the proportion of explained variation are important quantities in many statistical models and model fitting procedures. They play an important role in regression diagnostics, model selection procedures, and in determining the performance limits in many problems. Recently, methods for estimating these and other related summary statistics in high-dimensional linear models have received significant attention. In this talk, Lee discussed some of the various approaches to estimating these quantities (e.g., residual sum-of-squares-based estimators, the method-of-moments) and the conditions required to ensure reliable performance (sparsity, conditions on the predictor covariance matrix). Efficiency was be discussed, along with new estimators that are closely related to ridge regression. Lee also presented an application related to estimating heritability, an important concept in genetics. There were three thought-provoking points of discussion that came up during and after Lee's talk:

(a) Efficiency and M-estimators. The MLE/ridge estimators described in Lee's talk were derived under a "random β" assumption, but they're interested in their performance in a "fixed β" model. A detailed analysis of the estimators in the random β model is straightforward, by standard likelihood theory. Converting consistency and asymptotic normality results for the random β model into corresponding results for the fixed β model is also fairly straightforward. However, efficiency for the fixed β model seems to be a bit more difficult. One of the comments/suggestions made during Lee's talk was that one could consider efficiency within an appropriate class of M-estimators, and that this might be more tractable than efficiency within the broader likelihood framework. It seems like this may be a very fruitful approach to this problem.

(b) Estimating heritability using "sparse" estimators for σ^2. The main application discussed in Lee's talk was estimating heritability in genetics, i.e. estimating r^2. The methods Lee used were based on "non-sparse" estimators for σ^2 and τ^2 discussed in the talk. However, it also seems feasible to estimate r^2 using "sparse" estimators for σ^2 that have been proposed elsewhere (these estimators require β to be sparse in order to be effective). This approach to estimating r^2 has not been considered in the literature (thus, we would have to derive the sparse estimators for r^2 ourselves, along with their asymptotic distribution), and it would be interesting to compare the performance of the "sparse" and "non-sparse" methods in a real data example.

(c) Towards the end of the talk, someone asked (approximately): "Can you apply these methods in a genomic dataset where $p = 1,000,000$ and $n = 1,000$?" His response during the talk was pretty brief, and the Session Chair, Paul Kvam, basically suggested that this would probably be difficult. However, challenges like this can also be very interesting. To address problems with $p = 1,000,000$ and $n = 1,000$, one might initially seek methods for reducing the ratio p/n (perhaps to around 10-100), before applying methods similar to those proposed in the talk. Reducing p/n could be achieved by either (i) increasing n (this is often not possible) or (ii) decreasing p. Screening-out predictors could be used to decrease p; this could be conducted using only the "X" data and not peeking at the "y" data (e.g. screen out highly correlated predictors, as in the heritability example from the talk). More broadly, Lee was very excited to explore how methods like those proposed in the talk can be used in challenging applications; the comments/feedback he received during and after the talk have provided some great leads for this.

Philip Gautier, a Ph.D. student at Purdue who is working with Professors William Cleveland and Chuanhai Liu, talked about using Divide and Recombine (D&R) as a statistical framework for the analysis of big data. Philip provided context to relate this likelihood-based analysis to previous talks that featured "Split and Conquer" approaches in similar large-data problems. Examples for logistic regression are especially effective in comparing the D&R approach to the simpler "all-data" MLE. Matthias Katzfuss commented that it would be necessary to see how the method I presented (divide & recombine likelihood modelling) would perform on real data sets with many explanatory variables After the talk finished, Nan Lin discussed asymptotics for divide & recombine (a.k.a divide & conquer, split & conquer) methods with the author. The asymptotic results presented earlier by Nan Lin, Minge Xie, and Elizabeth Schifano have the subset size going to infinity. Under these conditions, their estimators were consistent. In real applications, we might expect to be able to collect more data, but the subset size will probably have some upper limit based on our computing environment. The simulations presented showed that, with a fixed subset size and growing number of subsets, bias persists. Philip commented that the focus should be on reducing the size of that bias by pursuing better division and recombination methods, rather than focusing on consistency.

4 Scientific Progress Made

Much progress has been made in this workshop. We summarize the comments from some of the workshop participants on this regard.

Christophe Andrieu, School of Mathematics, University of Bristol, UK: It has really helped raise awareness of the multiple facets of what "big data" means. The issues are multiple and not necessarily faced simultaneously (size of the data, computational burden). In Bristol, in collaboration with Warwick, Oxford and Lancaster we have started a reading group to discuss the practical issues faced when using parallel architectures (GPUs and others), and my interest has been raised by the Banff workshop. In terms of computing, the workshop has helped me realise that straightforward parallelisation is not a panacea and there is going to be a two way interaction between computation and statistical modelling and inference (in contrast to what has happened since the advent of cheap and powerful computers, since then computing was ancillary to the inferential task). And this is where statisticians are likely to have the edge, but they need to understand "computation" better. The mix of people with different perspectives on all these issues was really a big + for me.

Minge Xie, Department of Statistics, Rutgers University, USA: First, let me take this opportunity to thank you [the organizers] all for organizing this wonderful workshop. The workshop was a success in every aspect. I had a good time. We all appreciate your handwork and effort for making this an impactful event.

Peihua Qiu, Department of Biostatistics, University of Florida, USA. As I told you [the organizer] during the workshop, this big data workshop is one of the best ones I ever attended. You did an excellent job in organizing the sessions well.

Kun Chen, Department of Statistics, University of Connecticut, USA: I would like to thank the organizers for organizing such a timely and successful workshop on big data. This workshop has greatly facilitated interactions among the statisticians who are currently working on the frontiers of statistical big data research, and also provided great learning opportunities for young researchers who are eager to dedicate to big data research. I believe the workshop will continue to stimulate new ideas on how to conduct statistical big data research and how to better prepare younger generation statisticians to tackle the many challenges we are facing in the big data era.

Guanghua (Andy) Xiao, Quantitative Medical Research Center, Department of Clinical Sciences, University of Texas, Southwestern Medical Center, USA: Thank you [the organizers] very much for giving me the opportunity to attend the workshop. It was a great experience for me at Banff and I have learned a lot from the workshop.

Hongzhe Li, Department of Biostatistics, University of Pennsylvania, USA: Thank you [the organizers] very much for putting together this great workshop and for inviting me to participate. It was a great workshop and all the presentations are very interesting and address many important statistical issues related to big data analysis. The talks and discussions will definitely lead to further works in this very important area of statistical research.

Heping Zhang, Department of Biostatistics, Yale University, USA: It was a very successful and informative meeting. I have learned various important topics related to big data analyses.

Hongtu Zhu, Department of Biostatistics, University of North Carolina, USA: (i) Focus on developing novel statistical and computational methods on integrating imaging-genetic data; and (ii) solve several deep theories associated with our new methods.

Paul Kvam, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Tech, USA: There were a lot of important concepts that I was probably not able to take in during the workshop, but I learned so much about the current state and future direction of the nonparametric methods that are most helpful for big data analysis, my future research is sure to be affected in a great way. I was surely inspired and happy to be there. I am so grateful to you and the hosts for making my experience there possible!

Jian Zhang, School of Mathematics, Statistics and Actuarial Science, University of Kent, UK: It has been one of greatest workshops I attended so far. I have learnt a lot about Big Data Analysis via this workshop.

Xiaojing Wang, Department of Statistics, University of Connecticut, USA: Thank you very much for organizing such a wonderful workshop at Banff! It is really a very interesting workshop and I learned a lot.

Alex Shestopaloff, Department of Statistical Sciences, University of Toronto, CA: As a PhD student attending my first conference outside of Toronto, I must say that it was a very enjoyable and informative experience. I look forward to meeting you (and the others) again! Many thanks for your efforts.

Xiaoyi Min, Yale School of Public Health, USA: From the talks presented at the workshop, I learned a lot about the challenges in "big data" as well as the current developments addressing these challenges. From my personal understanding, the workshop covered recent progresses in three directions. First, many studies deal with the big volume of "big data". The "Divide and Conquer" strategy, for example, is employed in handling

data that cannot be stored or computed on a single computer. Second, several talks tackled the computation burden coming from "big data". They use, for example, specific modeling or sampling techniques to allow parallel computing on multiple nodes or even GPUs. Last but not least, many researches focus on the specific data structures arising from different forms of big and complex data such as image data, spatio-temporal data, and genetic and genomic data. To me, furthering and integrating the progresses in all these directions and providing solutions for real data analyses with "big data" is an important next step. Real data may be large in size, complex in structure, and difficult to compute at the same time. Therefore, an ideal solution should take all these challenges into account. This requires the collaboration of experts in all three aspects. Researchers also need a comprehensive understanding of the problems and the available tools. To this aspect, the workshop provided a great opportunity for researchers to learn from each other, exchange ideas, and start collaboration, which is very beneficial for the future development in the field of big data.
Philip Gautier, Department of Statistics, Purdue University, USA: Thank you for all of your work organizing the conference. The discussions I had with other presenters were very productive.

5 Outcome of the Meeting

There are many outcomes of the meeting. The workshop participants learned several new approaches and software for big data analysis. The workshop has initiated potential collaborations. After the workshop, Xiao Wang has obtained image datasets from both Professors Peihua Qiu and Hongtu Zhu. Peihua Qiu and Xiao Wang (two workshop participants) started their collaborative research on statistical process control of images, which is a new research area and involves big data processing and analysis. Hongtu Zhu and Xiao Wang are working on a paper on image classification after workshop. This workshop provided them a great opportunity to communicate with each other. Also, certain methods presented in the workshop, such as the divide-and-conquer method, will be probably used in their research. Minge Xie and Ming-Hui Chen discussed how to combine interval estimates in the steam data setting and they will continue to collaborate after the workshop.

There will be two special issues in Technometrics and Statistics and Its Inference on big data after the workshop. The special issue of Technometrics will publish original high-quality papers that deal with all aspects of the statistical analysis of big data, including but not limited to data visualization and exploratory data analysis, statistical computation, statistical modeling and inferences, and innovative applications. Papers in any application domain that fits within the broad scope of Technometrics will be considered. This special issue is expected to be published in February 2016. The Guest Editorial Board consists of Ming-Hui Chen (University of Connecticut), Radu V. Craiu (University of Toronto), Robert B. Gramacy (University of Chicago), Willis A. Jensen (W. L. Gore and Associates), Faming Liang (Texas A&M University), Chuanhai Liu (Purdue University), William Q. Meeker (Iowa State University), and Peihua Qiu (Editor) (University of Florida).

The special issue of Statistics and Its Inference (SII) will be on "Statistical and Computational Theory and Methodology for Big Data". SII strongly encourages substantive applications and computational developments for analyzing big data in all areas of sciences. High-quality review articles in this emerging new research area are also welcome. The papers, once accepted, will be published together in a future issue of SII. Some of accepted papers may be chosen as invited discussion papers in this special issue. The submission deadline for this SII special issue is November 1, 2014. The Guest Editors for the SII special issue include Ming-Hui Chen (University of Connecticut), Radu V. Craiu (University of Toronto), Faming Liang (Texas A&M University), Chuanhai Liu (Purdue University), and Heping Zhang (Editor-in-Chief) (Yale University).

The potential follow-up workshops were discussed. There were extensive email communications among the workshop participants on developing a textbook for big data after the workshop.

References

[1] A. Kleiner, A. Talwalkar, P. Sarkar and M. I. Jordan, A scalable bootstrap for massive data. Journal of the Royal Statistical Society, Series B (2014). doi:10.1111/rssb.12050.

[2] N., Lin and R. Xi, Aggregated estimating equation estimation. Statistics and Its Interface **4** (2011), 73–83.

[3] R. Xi, N. Lin, and Y. Chen, Compression and aggregation for logistic regression analysis in data cubes. IEEE Transactions on Knowledge and Data Engineering **21**, 479–492.

[4] X. Chen and M. Xie, A split and conquer approach for extraordinarily large data analysis. Statistica Sinica (2014). In press.

[5] F. Liang, Y. Cheng, Q. Song, J. Park, and P. Yang, A resampling-based stochastic approximation method for analysis of large geostatistical data. Journal of the American Statistical Association **108** (2013), 325–339.

[6] M.-H. Chen, Q.-M. Shao, and J.G. Ibrahim, Monte Carlo methods in Bayesian computation, Springer, New York, 2000.

[7] J.S. Liu, Monte Carlo strategies in scientific computing, Springer, New York, 2001.

[8] F. Liang, C. Liu, and R.J. Carroll, Advanced Markov chain Monte Carlo methods: learning from past samples, Wiley, New York, 2010.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39** (1977), 1–38.

[10] P. Li, A. Owen, and C.-H. Zhang, One permutation hashing. Neural Information Processing Systems (NIPS) **25** (2012), 3122–3130.

[11] P. Li, G. Samorodnitsky, and J. Hopcroft, Sign Cauchy projections and Chi-square kernel. Advances in Neural Information Processing Systems **26**, 2013.

[12] P. Li, C.-H. Zhang, and T. Zhang, Compressed counting meets compressed sensing. arXiv:1401.0201v1, 2013.

[13] C. Andrieu, A. Lee, and M. Vihola, Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. arXiv:1312.6432, 2013.

[14] R.N., Henson, G. Flandin, K.J. Friston, and J. Mattout, A parametric empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction. Human Brain Mapping **31** (2010), 1512–1531.

[15] P. Bickel and E. Levina, E., Covariance regularization by thresholding. Annals of Statistics **36** (2008), 2577–2604.

[16] J. Zhang, C. Liu, and G. Green, Source Localization with MEG Data: A beamforming approach based on covariance thresholding. Biometrics (2013), doi: 10.1111/biom.12123.

[17] R. Martin and C. Liu, Inferential models: a framework for prior-free posterior probabilistic inference. Journal of the American Statistical Association **108** (2013), 301–313.

[18] R. Martin and C. Liu, Comment: foundations of statistical inference. Statistical Science (2014a), in press.

[19] R. Martin and C. Liu, 3. Martin, R. and Liu, C. (2014b). Conditional inferential models: combining information for prior-free probabilistic inference. Journal of the Royal Statistical Society, Series B (2014b), in press.

[20] R. Martin and C. Liu, Marginal inferential models: prior-free probabilistic inference on interest parameters. unpublished manuscript, 2014c.

[21] P. Du and X. Wang, Penalized likelihood functional regression. Statistica Sinca (2013), in press.

[22] X. Wang, P. Du, and J. Shen, (2013). Smoothing splines with varying smoothing parameter. Biometrika **100** (2013), 955–970.

[23] L. Dai, M. Tian, J.Y. Wu, J.F. Xiao, X.M. Wang, J.P. Townsend, and Z. Zhang, AuthorReward: increasing community curation in biological knowledge wikis through automated authorship quantification. Bioinformatics **29** (2013), 1837–1839.

[24] Z. Zhang, J. Sang, L.N. Ma, G. Wu, H. Wu, D.W. Huang, D. Zou, S.Q. Liu, A. Li, L.L. Hao, M. Tian, C. Xu, X.M. Wang, J.Y. Wu, J.F. Xiao, L. Dai, L.-L. Chen, S.N. Hu, and J. Yu, RiceWiki: a wiki-based database for community curation of rice genes. Nucleic Acids Research **42** (2014), D1222–D1228.

[25] B. Xie, B. Jankovic, V. Bajic, L. Song, and X. Gao, Poly(A) motif prediction using spectral latent features from human DNA sequences. Bioinformatics **29** (2013), i316–i325.

[26] R. Grosse, C. Maddison, and R. Salakhutdinov, Annealing between distributions by averaging moments. In Neural Information Processing Systems (NIPS 27), 2013, www.cs.toronto.edu/∼rsalakhu/papers/nips2013_moment.pdf

[27] A.Y. Shestopaloff and R.M. Neal, MCMC for non-linear state space models using ensembles of latent sequences. Technical Report, 2013, http://arxiv.org/abs/1305.0320.

[28] Y.S. Niu and H. Zhang, The screening and ranking algorithm to detect DNA copy number variations. The Annals of Applied Statistics **6** (2012), 1306–1326.