

Computing at Denver: An overview of Stat Comp at JSM 2008

by
Wolfgang Jank
Program Chair

The section on statistical computing has lined up an exciting program for JSM 2008. Our topics fill the entire spectrum between curriculum design, methodological development, and analysis of real-world problems. The individual sessions vary in format and include invited speakers and discussants from around the world, from academia, from research labs and from industry. The common theme among all of the sessions is the computational aspect. This could be in the form of teaching statistical computing to our students, developing new computational algorithms, or using computational methods to analyze challenging and large-scale data. Particular thanks go out to Deepak Agarwal (Yahoo!), David Banks (Duke), Giles Hooker (Cornell), Ravi Vardhan (Johns Hopkins) and Chris Volinsky (AT&T) for putting together such an exciting program. In the following, I will provide a brief overview of the invited sessions at JSM 2008.

Our first invited session is on *Designing Courses on Statistical Computing*. By now, most statistics departments have recognized the need to make computation a central focus of their graduate and undergraduate education. This has been done in different ways, and the most common is a specifically designed course that is taught in the first or second semester of a student's career. The contents of such a course are not well-established, in part because there is no clearly defined common subject matter, and in part because there is no widely-used text that prescribes the content. This session presents several approaches to the problem of designing a broad and sufficient course on modern statistical computing to Ph.D. students in statistics.

Two of our invited sessions deal with the development of new statistical methodology. The first session is on *Global Maximization in EM-type Algorithms*. In likelihood-based modeling, finding the parameter estimates that correspond to the global maximum of the likelihood function is a challenging problem. This problem is especially serious in statistical models with missing data, where popular EM-type algorithms are used to obtain parameter estimates. Despite their hallmark stability and monotone convergence, EM-type algorithms (i.e. the EM algorithm and its variants, such as MM algorithms) frequently converge to local, sub-optimal solutions, especially when the dimension of the parameters space and/or the size of the data are large. This is a particular concern in e.g. finite mixtures and latent variable models. The problem has largely been ignored in statistics and only unprincipled approaches have been used to date (e.g. random multi-starts). Part of the problem is that no practically useful characterization of the global maximum is available. This session presents different solutions to global optimization problems in the context of the EM algorithm.

The second methodological session considers *Advances in Functional Data Analysis (FDA)*. The technological advancements in measurement, collection, and storage of data have led to more and more complex data-structures. Examples include measurements of individuals' behavior over time, digitized 2- or 3-dimensional images of the brain, and

recordings of 3- or even 4-dimensional movements of objects traveling through space and time. Such data, although recorded in a discrete fashion, are usually thought of as continuous objects represented by functional relationships. This gives rise to functional data analysis (FDA), where the center of interest is a set of curves, shapes, objects, or, more generally, a set of *functional observations*. This is in contrast to classical statistics where the interest centers around a set of data vectors. FDA has experienced a rapid growth over the past few years, both in the range of applications for its techniques and in the development of theory for statistical inference. This session will bring together leading researchers in the field to discuss recent advances and applications of FDA. Topics addressed will range from theoretic properties and inference in functional linear regression to new regularization techniques and the application of functional models to the analysis of complex data structures.

Three of our invited sessions address computational challenges driven by large-scale and complex real-world applications. These applications include online advertising, social networks, and Wikipedia data. The Wikipedia is an important unintended consequence of the Internet. It demonstrates how valuable content can be created by distributed volunteer effort. And it also provides statisticians with a rich data source on the connectivity between knowledge domains, the investment/growth trade-offs needed to bootstrap a distributed business, and the difficult problem of quality assurance in a climate of complex error sources. Also, from the standpoint of our professional society, Wikipedia offers a new model for service the ASA can provide to its members, by integrating a range of user-created code, data, discussion, and articles. The session *Analysis of Wikipedia Data* includes analyses of key Wikipedia data sets and a discussion of a possible Wikipedian future for ASA publications.

Another internet phenomenon that poses new and exciting challenges for the statistician are online social networks. Online social networks have become ubiquitous. Sites such as *mySpace* and *Facebook* create online communities where users define their own social networks; other sites such as blogs and online forums create de-facto networks as users link to each other's opinions. There has been a flurry of research on these networks on topics such as dynamic network evolution, community detection, discovering influential actors, etc. However, most of this work has been done *outside* of statistics, in the fields of machine learning, AI, and CS, and published at data mining conferences like KDD. In the session *Analysis of Massive Online Social Networks*, we will discuss a variety of topics about online social networks, blogs, and product review networks. We will also discuss why statisticians have been absent from this field to date, and what contributions our community can make.

Our last invited session deals with the ever-growing phenomenon of online advertising and its implications for the statistician. Online advertising is a multi-billion dollar industry as evident from the phenomenal success of companies like Google, Yahoo, or Microsoft, and it continues to grow at a rapid rate. With broadband access becoming ubiquitous, internet traffic continues to grow both in terms of volume and diversity, providing a rich supply of inventory to be monetized. Fortunately, the surge in supply has also been accompanied by increase in demand with more dollars being diverted to

internet advertising relative to traditional advertising media like television, radio, or newspaper. Marketplace designs that maximize revenue by exploiting billions of advertising opportunities through efficient allocation of available inventory are the key to success. Due to the massive scale of the problem, the only feasible way to accomplish this is by learning the statistical behavior of the environment through massive amounts of data constantly flowing through the system. This gives rise to a series of challenging statistical problems which include prediction of rare events from extremely massive and high dimensional data, experimental designs to learn emerging trends, protecting advertisers by constantly monitoring traffic quality, ranking search engine results, modeling extremely large and sparse social network behavior. The session *Statistical Challenges in Online Advertising and Search* will discuss many of these challenges.

I hope that, by now, you are as excited about next year's program as I am, and I hope that I will see many of you at the sessions sponsored by *Stat Comp*. I would also like to remind you that it is not too late for you to contribute. There are still open slots for topic contributed sessions. If you have any questions, please don't hesitate to contact me at wjank@rhsmith.umd.edu.

See you in Denver,
Wolfgang