THE ISBA BULLETIN

Vol. 13 No. 2 *June 2006*

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE NEW PRESIDENT

by Alan Gelfand
ISBA President
alan@stat.duke.edu

The Valencia/ISBA 2006 meeting will be remembered as a magical gathering. Despite the size (nearly 550 delegates) and the bottlenecks (at the elevators, posters, dinners), the warmth, collegiality, and feel-good ambience were evident, far exceeding expectations for a meeting of this size. Under most any customary demographic measure, we had a very broad range of attendance. Moreover, there seemed to be a substantial amount of interaction across demographic groups.

Scientifically, while some may question the viability of the single session format with such a large, diverse crowd, we can certainly agree that the invited and contributed papers were of very high quality. And, we would also agree that the posters set an exceptionally high standard. The scope and content was outstanding, providing another demonstration of the consequential presence of Bayesian thinking across a very broad range of scientific inquiry.

It must also be noted that the cabaret was incredible! In terms of talent level, originality, and technological prowess, a new bar was set. Thanks to all who participated truly a spectacular ending to a special week. Finally, thanks again to the Valencia scientific committee, to the ISBA program committee, and especially to Jose B. for an event we will recall fondly and be proud to say we attended.

A MESSAGE FROM THE EDITOR

by J. Andrés Christen
jac@cimat.mx

This June issue of the bulletin is a bit late since I wanted to wait after the Valencia meeting had taken place so as our President (Alan) could send you s short review. I am also very happy to have attended and most also congratulate all the people involved (scientific committees, local committee, office staff and student volunteers) in the complex organization of such a big meeting.

We don't have an interview section in this issue. However, Arnold Zellner pointed out to me an interview made a couple of years ago, talking about Arnold's life and his outstanding career. Instead of trying to include it in the Bulletin I simply prompt you to read this quite comprehensive interview. The link is `http://www.strategy2market.com/downloads/InterviewwithArnoldZellner.pdf`.

Finnally, an important piece of information, that most of you probably know already, is that the FDA (US Food and Drug Administration) has issued a "Draft Guidance for Industry and Food and Drug Administration Staff; Guidance for the Use of Bayesian Statistics in Medical Device". This draft was released on May 23 this year for public comment. There will be a public meeting to address the draft at The Universities of Shady Grove, Rockville, MD, from 8:30 - 5:00 on Thursday, July 27, 2006. Please check `http://www.fda.gov/cdrh/meetings/072706-bayesian.html` for more information.

Contents

➤ APPLICATIONS
☛*Page 1*

➤ NEWS FROM THE WORLD
☛*Page 4*

SUGGESTIONS

PLEASE, FEEL COMPLETELY FREE TO SEND US SUGGESTIONS THAT MIGHT IMPROVE THE QUALITY OF THE BULLETIN

jac@cimat.mx

High-dimensional Regression in Cancer Genomics

by Chris Hans and Mike West
hans@stat.ohio-state.edu
mw@isds.duke.edu

Cancer-genomic data present modeling challenges due to their high-dimensional nature. A typical dataset consists of several (tens or hundreds, typically the former) tumor samples which are then used to generate tens of thousands of estimated gene expression levels. Associated with each patient is a particular outcome of interest, e.g. survival time, presence/absence of lymph node invasion, or response to a particular treatment. Two goals are the identification of the complex multivariate patterns of association between key genes and the outcome, and prediction for future patients.

Shotgun Stochastic Search

Because the sample size is often much smaller than the total number of candidate predictor variables, we focus on generating lists of *sparse* regression models — models comprised of only a few genes. By scoring each model by its unnormalized posterior probability, we can build lists with potentially millions of models that exhibit good fit to the data, and then perform model averaging to identify dominant genes and to provide predictive distributions. Emphasis on sparse models is made possible through the use of prior distributions that highly penalize the addition of predictor variables to a model. Indeed, sparsity is emerging as an important component in modeling such high-dimensional datasets.

In order to perform the search necessary to construct such model lists, we introduce the "shotgun stochastic search" (SSS), a parallel-computing based method for exploring large model spaces (see Hans, Dobra and West (2005) for details). SSS is related to MCMC approaches for traversing regression model spaces, however due to the use of a distributed computing environment, SSS is able to more quickly explore local regions of model space, rapidly identifying promising models and providing new directions in which the search may evolve.

Brain Cancer Survival Study

We analyzed gene expression data from a survival study in brain cancer based at the W.M. Keck Center for Neuro-Oncology at Duke University. A detailed description of the data along with an initial analysis can be found in Rich *et al.* (2005). The study consists of 41 patients diagnosed with glioblastoma, a form of brain cancer associated with relatively short survival times. Although survival times are generally short, significant variation is observed and it is of interest to explore possible biological explanations for this variability by analyzing the gene expression data.

For each patient we have survival time (in days) measured from initial diagnosis along with a tumor specimen. Due to the nature of the disease, all of the patients in the study are deceased and hence there is no censoring information. Gene expression data is available on Affymetrix human U133A microarrays, processed using the current standard RMA method to provide expression estimates for each gene. After an initial screening to remove probes whose expression levels were clearly "in the noise", a total of $8,408$ genes were included in the analysis.

We used SSS to explore the model space of sparse regression models: $\log \text{survival time} \sim \mathrm{N}(X_\gamma \boldsymbol{\beta}_\gamma, \sigma^2)$, where $\boldsymbol{\gamma}$ represents a particular (small) subset of genes. The best one million models found by SSS identified several genes that were associated with variability in survival time (see Rich *et al.* (2005) for details). In particular, one gene known to have increased expression levels in several other types of cancer was found to be influential in discriminating survival time, especially in the context of regression models including two other genes that have specific neural functions. The top panel of Figure 1 displays model averaged fitted values with 95% credible intervals.

Versatility of SSS

While this application concered the normal linear model, SSS can be used to search any discrete space. For example, we performed a separate analysis of the brain cancer data using Weibull regression models with survival time in months as the outcome. A plot of leave-one-out cross-validated predictions of one year survival probability is given in the second panel of Figure 1 — combining information across thousands of Weibull regression models seems to well-discriminate twelve month survival (approximately the general population median). We have also applied SSS to binary regression models (Dressman *et al.*, 2006) and to Gaussian graphical models (Jones *et al.*, 2005), both in contexts of high-dimensional can-

cer genomic datasets. We anticipate applications in other complicated modeling contexts designed to coherently extract information from these inherently high-dimensional problems. ▲

References

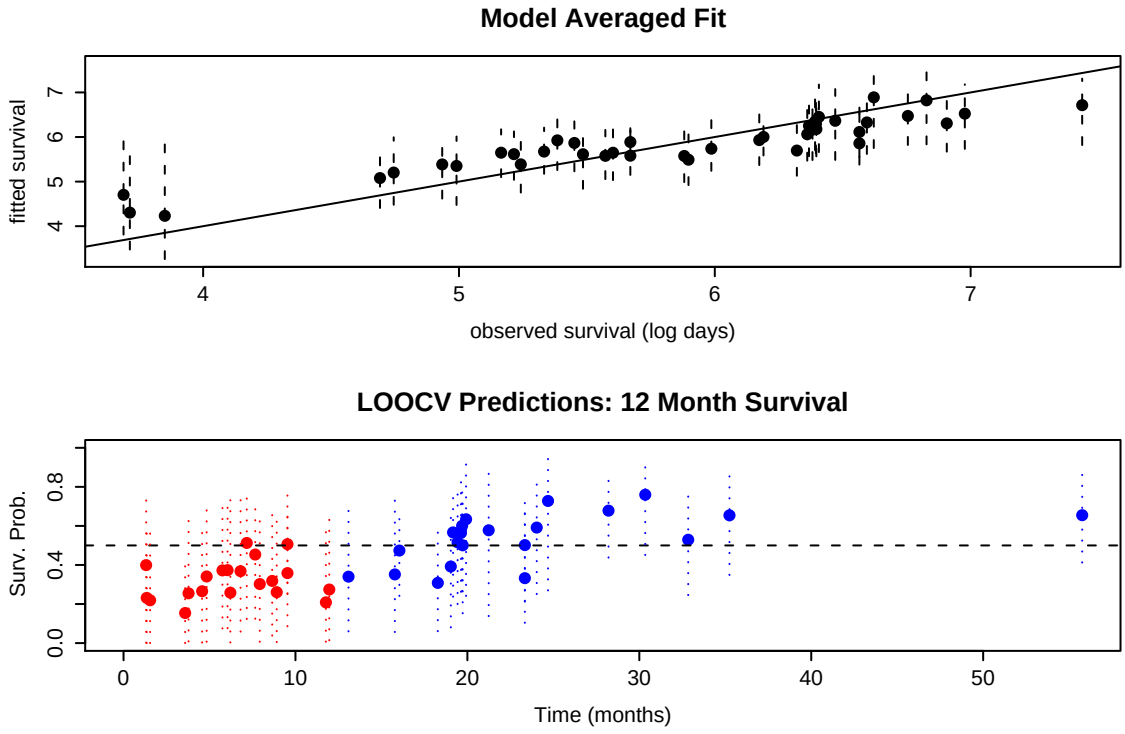
- Dressman, H., Hans, C., Bild, A., Olson, J., Rosen, E., Marcom, P., Liotcheva, V., Jones, E., Vujaskovic, Z., Marks, J., Dewhirst, M., West, M., Nevins, J. and Blackwell, K. (2006). "Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy." *Clinical Cancer Research*, **12**, 819–826.

- Hans, C., Dobra, A. and West, M. (2005). "Shotgun stochastic search for 'large p' regression." ISDS Discussion Paper 2005-10 (Submitted).

- Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. and West, M. (2005). "Experiments in stochastic computation for high-dimensional graphical models." *Statistical Science*, **20**, 388–400.

- Rich, J., Hans, C., Jones, B., Iversen, E., McClendon, R., Rasheed, B., Dobra, A., Dressman, H., Bigner, D., Nevins, J. and West, M. (2005). "Gene expression profiling and genetic markers in glioblastoma survival." *Cancer Research*, **65**, 4051–4058.

Figure 1: The top panel displays model-averaged fitted log survival time vs. observed log survival time (in days) under the normal linear model setup. The bottom panel displays leave-one-out cross-validated model-averaged predicted twelve month survival probabilities for the Weibull regression model; red (blue) points indicate patients with actual survival time less (greater) than one year.



NEWS FROM THE WORLD

by Alexandra M. Schmidt
alex@im.ufrj.br

I would like to encourage those who are organizing any event around the World, to get in touch with me to announce it here.

Events

Conference on High Performance Computing for Statistical Inference, Trinity College, Dublin, August 23rd-25th 2006.

The main topics of the conference are: implementation of statistical analysis with distributed computing; Computation with very large stochastic systems; Very large scale applications; The suitability to parallelization of statistical methods; MCMC; Monte Carlo simulation in a distributed computing environment; Grid technologies for statistical analysis; Quantum computing and statistical inference. Please, go to conference webpage (`http://www.tcd.ie/Statistics/hpcsi/`) or contact organizer Simon Wilson at simon.wilson@tcd.ie for further information.

Nonlinear Statistical Signal Processing Workshop: Classical, Unscented and Particle Filtering Methods, Cambridge, UK, September, 13-15, 2006

A workshop in nonlinear sequential estimation and processing, bringing together classical, modern unscented and particle filtering methods as well as newly emerging techniques; topics include filtering, smoothing and the sequential treatment of batch problems. Theory, computational methods and applications are welcome. The workshop will be held in the beautiful and historic surroundings of Corpus Christi College, Cambridge, UK. For more details visit `http://www-sigproc.eng.cam.ac.uk/NSSPW/`.

International Summer School on Recent Developments in Spatial Statistics, University of Klagenfurt, Austria, September 11 - 16, 2006.

The International Summer School addresses most of all to young researchers (Master students, PhD students, PostDoc researchers) in academia, "extra-university" research, and industry who are interested to learn about recent developments, new methods and applications in spatial statistics and related areas. The Summer School will present the following actual topics:

* Spatial Design by W. Mueller, Vienna
* MCMC Methods by Ch. Robert, Paris
* Bayes Maximum Entropy Methods by G. Christakos, San Diego
* Spartan Random Fields Modelling by D. Hristopulos, Chania
* Markov Random Fields Modelling and Applications by H. Rue, Trondheim
* Recent Developments in spatio-temporal Geostatistics by T. Gneiting, Seattle, and M. Schlather, Hamburg
* Spatial Point Processes and Applications in Ecology and Epidemiology by R. Waagepetersen, Aalborg
* Geostatistical Simulation by Ch. Lantuejoul, Fontainebleau
* Data Assimilation by D. Cornford, Aston

Detailed information can be found at `http://www.uni-klu.ac.at/statgis06`.

Workshop on recent advances in Monte Carlo based inference, Isaac Newton Institute, Cambridge, UK, 30th October to 3rd November 2006.

This workshop is part of a two month visitor programme at the Isaac Newton Institute on "Stochastic Computation in the Biological Sciences". The focus of the workshop is on Monte Carlo methods; primarily on recent advances/ideas in terms of methodology, but also to include theoretical results for these methods (which have practical importance); and work on applying these modern computationally-intensive methods to important scientific applications. There will be a relatively small number of long talks, covering areas from MCMC and sequential Monte Carlo to Variational methods, Indirect Inference and Quasi-Monte Carlo methods amongst others. For further information, and details of how to apply, visit `httpa://www.newton.cam.ac.uk/programmes/SCB/scbw01.html`. ▲

ISBA/SBSS Archive for Abstracts

All authors of statistics papers and speakers giving conference presentations with substantial Bayesian content should consider submitting an abstract of the paper or talk to the ISBA/SBSS Bayesian Abstract Archive. Links to e-prints are encouraged. To submit an abstract, or to search existing abstracts by author, title, or keywords, follow the instructions at the abstract's web site,

http://www.isds.duke.edu/isba-sbss/