THE ISBA BULLETIN

Vol. 14 No. 4 December 2007

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE PRESIDENT

by Peter Green
ISBA President
P.J.Green@bristol.ac.uk

Greetings of the Season to all ISBA members! Wherever you are in the world, I hope this is a time for celebration and recreation for you, and that you have a great New Year ahead.

I want to begin by congratulating the winners in the recent elections: Mike West, who will be President in 2009, Gabriel Huerta, who takes over as Treasurer next month, and the new Board members: Sylvia Frühwirth-Schnatter, Lurdes Inoue, Hedibert Lopes and Sonia Petrone. You supported them by voting, I hope you will support them in office!

After a fairly quiet year for ISBA conference activity, 2008 brings two major meetings - January in Bormio and July in Queensland. I hope to see many of you at one or both of these events - details can be found later in this Bulletin. The organisers have worked hard to devise exciting programmes, and I hope you'll try to get there to join in the scientific and social interaction. There are also important regional and specialist meetings coming up.

An important 'behind-the-scenes' activity of the Executive recently has been negotiating a joint membership deal with IMS, the Institute of Mathematical Statistics. This involves no loss of autonomy, for either party, but should help us increase our membership, as well as offering a discounted IMS subscription deal to existing ISBA members. This option will be offered to you when you next renew your ISBA membership. *Continue in page 2.*

A MESSAGE FROM THE EDITOR

by Raphael Gottardo
raph@stat.ubc.ca

It has been over four months since I have become the new editor of the ISBA bulletin, and I am pleased to be able to send you a new issue. Of course, all of the credit should go to the people who actually do the work, that is the associate editors. In this December issue, you will find a report on the status of the recently created journal, Bayesian Analysis (BA), written by Brad Carlin, editor-in-chief of BA. To complement this report, Beatrix Jones compiled a bibliography of all applied papers that have been published in BA up to today. As you will see, there are many interesting papers that have been published in BA covering a wide range of applications. After reading this bulletin, I hope you will consider submitting your next article to BA. *Continue on page 2.* ▲

In this issue

WORDS FROM THE PRESIDENT, *Continued from page 1.*

Later in this Bulletin, Brad Carlin reports on some important developments in our journal, *Bayesian Analysis*. I am very grateful to Brad for leading on this; it's a mjor contribution to promoting our discipline - let's all reciprocate by making *BA* a destination of choice for our own work!

This is my last message to you, as I am stepping down now, and passing over the presidential inbox, with all good wishes, to Christian Robert. It has been a rewarding time, and my only regret is that I didn't succeed in getting any specialist Sections started. Perhaps this will happen in 2008. I want to thank all those who have helped me over the course of the year, too numerous to list in full, but let me highlight Executive Secretary Robert Wolpert and Treasurer Bruno Sanso. Both of these are showing real long-term commitment to the success of ISBA, and of course do far more than the president to keep the show on the road; Bruno is stepping down now, so especial thanks to him. All the very best to all for 2008.▲

WORDS FROM THE EDITOR, *Continued from page 1.* In addition to this information about BA, this issue of the bulletin also contains a nice article written by Jeff Morris on "Statistical Issues in Proteomic Research", as well as a great review of software for graphical models written by Kevin Murphy. As usual, we (the editorial board) rely on ISBA members for great contributions, so if you have something interesting you would like to publish in the bulletin please contact us (contact information is available on the last page). So enjoy the bulletin, and of course, greetings to you all and best wishes for the new year! ▲

BAYESIAN ANALYSIS - A MESSAGE FROM THE EDITOR

Bayesian Analysis 2007 SUMMARY REPORT

by Brad Carlin
***Bayesian Analysis* Editor-in-Chief**
brad@biostat.umn.edu

First of all, thanks to editor Raphael Gottardo for allowing me to publish this report in the *ISBA Bulletin*. Normally the report of a journal editor would be published in that journal, and we indeed have placed a copy on our own homepage, `ba.stat.cmu.edu`. But since improving our communication with ISBA membership and the Bayesian community at large remains a key concern for us, I am grateful for this opportunity.

While this article will be largely a report on activities in 2007 and plans for 2008, I will also use it as an opportunity to respond to a couple of skeptical remarks about *Bayesian Analysis* and its future made by two prominent Bayesians and published in this *Bulletin*. I do this partly since I feel it is my responsibility, but also because, as a former skeptic myself, I feel qualified.

Introduction

Bayesian Analysis (BA) is the official journal of ISBA. New issues are published online quarterly, with each issue consisting of roughly 10 papers, one of which is usually a discussion paper with rejoinder from the author(s). Of course, as a free online journal, the "quarterly publication" aspect is somewhat artificial, since articles are posted continuously at `ba.stat.cmu.edu/forthcoming.php` as they are accepted. *BA* also has an RSS feed (joining instructions at `ba.stat.cmu.edu/rssfeed.php`) so ISBA members may be notified the minute a newly accepted paper is posted.

The journal's basic review procedure is this: papers are submitted to the website, and come to the editor-in-chief (EiC) for initial review. Virtually all of these go to one of the journal's six ed-

itors, who in turn either reject the paper or send it on to an associate editor (AE). All papers assigned to an AE receive two full reviews, typically one by the AE and another by a referee of the AE's choosing. Our turnaround time (submission to decision) has been quite good, generally between 8 and 12 weeks. Certainly the online system has been a big help in this regard, coordinating reviewers from around the globe, but the biggest reason for our success has been the diligence of the editorial and administrative staff.

I was named EiC by the ISBA Board of Directors during the summer of 2006, replacing founding EiC Rob Kass. While Rob actually got to preside over just one publication year, he actually invested much closer to three years of effort in getting the journal up and running. Also instrumental in the start-up process were chief programmer Adrian Rollett (who designed and built the online review system), Pantelis Vlachos (who manages the system and oversees the production side of the journal), and Herbie Lee (the managing editor who gently keeps us on schedule). All three of these gentlemen continue to invest enormous hours in the journal, and I am extremely grateful to all three for their ongoing efforts.

Rob also assembled an editorial staff of internationally known scholars. At the time I joined the team, the board included six editors: Alicia Carriquiry, Phil Dawid, David Heckerman, Michael Jordan, Fabrizio Ruggeri, and Dalene Stangl. Alicia stepped down when Rob did in January of this year, and I was fortunate enough to attract David Dunson as a new editor. Similarly, Dalene has decided to step down this year (in order to become Reviews Editor for *JASA* and *TAS*), and as a result Antonietta Mira and Kerrie Mengersen will begin terms as editors in January and August of 2008, respectively. In addition, Marina Vannucci is our deputy editor, giving good advice in many "emergency" situations, and Angelika van der Linde is production editor, making last-minute repairs to accepted papers. Last but certainly not least the entire staff of AEs and referees have been crucial to our success to date; they are too numerous to list here, but I hope you'll go to `ba.stat.cmu.edu/associates.php` and thank the persons on this list the next time you see them for their great work on behalf of our profession.

Review of 2007

The past year has seen the publication of Volume 2 of *BA*. While the December issue is not yet finalized, we expect to publish just under 1000 pages, roughly the same as in Volume 1 (2006). In addition, we continued our mission of publishing outstanding work presented at ISBA-sponsored regional and international conferences. In particular, *BA* handled *all* of the contributed paper submissions to the ISBA 2006 (Valencia 8) world meeting, freeing conference president José Bernardo from having to form his own editorial board to handle these submissions. The result was 17 papers, many by outstanding junior researchers, that appeared in *BA* during 2007, with shorter summaries also appearing in the traditional Valencia volume. Other meetings from which papers have or will be obtained include the Case Studies in Bayesian Statistics (Carnegie Mellon) meetings in 2005 and 2007, the joint IMS-ISBA meetings in Bormio, Italy in 2005 and 2008 (MCMSki and MCMSki II), the Bayesian Inference in Stochastic Processes (BISP-5) meeting in Valencia, Spain last summer, and the upcoming ISBA 2008 meeting in Hamilton Island, Australia.

My own quick look back at the articles *BA* has published to date reveals a wide mix of theory, methods, and applications, and this brings me to the first "skeptical remark" alluded to above. At the very end of his otherwise excellent and entertaining interview in the September 2007 issue of the *Bulletin*, Prof. Don Berry expresses his opposition to ISBA in general, and *BA* in particular. Specifically, he worries that ISBA-sponsored publications like *BA* "would be regarded as second-rate statistics journals by our peers." Perhaps it is unfair to single out Prof. Berry for having these thoughts since they have been expressed ever since the debate over ISBA's formation began by many Bayesians – including myself! I too at one point feared that having our own Bayesian journal would push us into the wacky fringe, away from real science, and that the journal might attract only second-rate papers, since Bayesian authors would continue to send their best work to *JASA, Biometrics*, and other well-established statistical journals.

But now in my (admittedly somewhat ironic) new role as *BA* EiC, I can tell you that I have found most of those fears to be unfounded. Good Bayesians understand that they should not all go to a corner (or some beach in Spain) and congratulate each other; they should do good sci-

ence and publish in the top journals of both our profession and those of our substantive area colleagues. And this is reflected in the quality and timeliness of the papers *BA* is attracting. I also think *BA* fills a niche in publishing longer papers (e.g., case studies), and papers that grow out of exchanges within the Bayesian community itself (the "objective versus subjective" debate by Berger and Goldstein and associated discussion that we published in Volume 1 Number 3 comes immediately to mind). Another pro-BA argument is that, if we (ISBA) don't publish a journal like *BA*, some private outfit will. To me, it seems far preferable to have *BA* run by a professional society, like the well-established journals I mentioned above. Yet another advantage of *BA* is its firm commitment to free open access to its content online; in this regard see also Jim Pitman's fine IMS presidential address on the subject from the 2007 JSM.

Finally at the risk of "piling on" Prof. Berry, I must also take issue with his statement that, "ISBA has ignored us [Bayesian biostatistics] and we have ignored it." I am not sure if by "us" he meant *all* of Bayesian biostatistics or just his group at M.D. Anderson Cancer Center, but I would like to assure ISBA members that neither interpretation of this statement is true. A quick check through the 7 issues published online to date reveals a wealth of papers on topics well within the purview of biostatistics, including gene expression, species distribution patterns, censored lifetime data, microarrays, spatiotemporal neuronal networks, neonatal survival, multiple disease boundary analysis, DNA mixture analyses, and early stage breast cancer modeling. These papers were associated with such prominent Bayesian biostatistical authors as Clyde, Gelfand, Sebastiani, Dominici, Zeger, Parmigiani, Spiegelhalter, and Müller, to name only a few. (This last author is of course a member of Prof. Berry's own group!) Two of the journal's seven discussion papers to date have been explicitly biostatistical, with others illustrating their main points with biostatistical datasets. And of course, the journal has published papers in many other applied areas not directly connected to biology or medicine, ranging from astrophysics to economics. I therefore conclude that there is overwhelming evidence that the journal is not simply publishing articles only of interest to internally-focused Bayesians. Rather, we are succeeding in bringing our expertise to a wealth of applied fields, resulting in both better statistics and, more importantly, better science.

Plans for 2008

And so, editorially, *BA* is in excellent shape: the editorial board is strong, the online review system is functioning well, and our stream of regular and discussion papers is solid and growing. However, on the production side, several bothersome issues remain. As a freely-available online journal, *BA* lacks a revenue stream from which to pay for system improvements and other expenses. Also, while *BA* has recently been added to the Current Index to Statistics (CIS), the journal does not yet have an "impact factor," something that is increasingly important to junior researchers in Europe seeking to advance professionally. Addressing these problems has been and will continue to be high on my to-do list as EiC in 2008.

Regarding the revenue issue, ISBA Past President Alan Gelfand and I thought we had a preliminary solution to this problem when we verbally agreed to grant a private publisher exclusive rights to sell a hard copy version of *BA* to libraries and other interested parties. We had hoped that the publisher would also help us get set up with indexing services, which would ultimately lead to impact factor scores. But after six months of delays, the deal finally fell through when the publisher could not deliver on his promises.

As a result, we are now looking into a much more promising partnership with IMS, a sister professional organization with whom we have an existing relationship (through our joint sponsorship of the MCMSki meetings) and which has an established track record in publishing and indexing. To begin with, the ISBA and IMS executive committees are currently considering a proposal to make *BA* a "supported journal", as described at `www.imstat.org/publications/supported.html`. At the outset, this amounts to listing each other's journals on our webpages, resulting in broader visibility and better submissions. But after some modifications to *BA*'s online publication system, it will also grant us much easier access to the major statistics and other science indices. IMS is a member of CrossRef, maintainer of the DOI (Digital Object Identifier) system. We are working on developing machine readable copies of *BA* data

and metadata. This should make it easy for other publishers to link to *BA* articles in their reference lists. Down the road, we hope this will lead to machine harvesting of *BA* articles and their easy piping to CIS and other prominent math-related indices (MR, ZMATH, etc.).

Of course, getting *BA* into the major indices is only the first step toward acquiring an impact factor; this further requires the journal to acquire high quality articles that then get cited in other indexed mainstream journals. This is a long process; many scientific indices won't even consider a journal until it has published for 3 years (one year longer than *BA* has to date). But free online journals can ultimately obtain good impact factors (e.g., the *Electronic Journal of Probability*), so I am convinced this is the right strategy.

Regarding the revenue stream issue, here again a partnership with IMS would be beneficial. IMS has experience with a contractor, VTEX, that can provide single "print on demand" copies of *BA* given the complete set of PDFs that we already serve online. IMS would manage the whole process for some minimal fee that would be recovered from sales of the volumes. IMS already does this for the free electronic journal *Probability Surveys* for individuals at $40 per year; see `imstat.org/secure/orders/2007.html`.

Another revenue stream we are investigating is joint memberships. The ISBA and IMS executive committees are currently considering listing each other on our membership pages and approving a small discount for members of one society who want to join the other. Again, there is precedent for this: IMS has a very similar relationship with the Applied Probability Society of Informs, which is comparable in size to ISBA (though does not yet have an electronic journal). Finally, institutional memberships and perhaps even advertising could offer further revenue streams, though these ideas have yet to be explored.

Conclusion

I would like to close with a response to a comment made by the second of the aforementioned "skeptical prominent Bayesians," 2008 ISBA President-Elect candidate Prof. Tony O'Hagan. In his candidate statement, Prof. O'Hagan states, "I want to see *Bayesian Analysis* become one of the top journals in Statistics; it currently doesn't have that feel for me." I certainly agree with the first part of this statement! Regarding the second part, I hope that this report has shown that *BA* has a more robust present than it perhaps has gotten credit for, and a bright future based on continued excellence on the editorial side and exciting initiatives on the production side. I look forward to your submissions at `ba.stat.cmu.edu`, and to your more personal thoughts and reactions via email to `brad@biostat.umn.edu`.

ANNOTATED BIBLIOGRAPHY

APPLICATIONS PAPERS IN *Bayesian Analysis*

Edited by: Beatrix Jones
m.b.jones@massey.ac.nz

To complement this issue's feature about *Bayesian Analysis*, the annotated bibliography focuses on applications papers in that journal. The papers were selected from the first seven issues of Bayesian Analysis: Volume 1 issue 1-Volume 2 issue 3. Of course which papers are 'applications papers' is subjective. Many methodological papers include non-trivial examples, and many applications papers include larger lessons about the techniques they employ. The thirteen papers listed here were selected using an easily applied criterion: they feature the application in the title.

In most cases, the annotation was kindly contributed by the lead author, as indicated by the initials at the end of the entry. Where the authors were unable to provide an annotation, I've included a shortened version of the paper's abstract. Rather than strictly summarizing their papers, I asked the authors to include thoughts on why Bayesian methods were particularly suitable for the application discussed in the paper. Those of you wanting more details (and reading this issue on your screen) will notice that the title of each paper is 'clickable,' taking you to the full text of the paper on the Bayesian Analysis web site.

- Gelfand AE, Silander Jr. JA, Wu S, La-

timer A, Lewis PO, Rebelo AG, Holder M (2006) Explaining Species Distribution Patterns through Hierarchical Modeling. *Bayesian Analysis* **1**: 41-92. *This article is followed by comments and a rejoinder.* Arguably, understanding spatial patterns of species diversity and the distribution of individual species is the most consuming problem in ecology. A Bayesian approach seems most natural, almost critical in studying this issue. In order to address the multiple facets of this process, e.g., suitability, availablity, presence/absence, abundance associated with locations, and the way that information comes in at different levels, e.g., environmental factors and species attributes, implies that the synthesis can only be adequately achieved through hierarchical modeling. Additionally, the Bayesian framework enables scaling up to study aspects such as species richnesss, species turnover, and competition between species. All of these issues are studied in the context of a very large dataset (more than 60,000 sampling sites) in the Cape Floristic Region in South Africa. *AEG*

- House LL, Clyde MA, Huang YT (2006) Bayesian Identification of Differential Gene Expression Induced by Metals in Human Bronchial Epithelial Cells. *Bayesian Analysis* **1**:105-120. Given only one observation per treatment and gene (five treatments, 1185 genes), we identified differentially expressed genes by using a latent variable, multi-response Bayesian mixture model. The model enabled us to both pool information across treatments and assess the posterior probability of differential expression per gene by treatment. To account for model uncertainty and hyperparameter specification, we implemented model averaging, MCMC, and Rao-Blackwell parameter estimation. *LLH*

- van Dyk DA, Connors A, Esch DN, Freeman P, Kang H, Karovska M, Vinay Kashyap, Aneta Siemiginowska, Andreas Zezas (2006) Deconvolution in High-Energy Astrophysics: Science, Instrumentation, and Methods. *Bayesian Analysis* **1**:139-236. *This article is followed by a comment and rejoinder.* In recent years, technological advances have dramatically increased the quality and quantity of data available to astronomers. This paper explores a number of data-analytic challenges arrising in high-energy, high-resolution astronomy, where new instrumentation designed to detect and map ultra-violet, X-ray, and γ-ray electromagnetic emission are opening a whole new window to study the cosmos. Because the production of high-energy electromagnetic emission requires temperatures of millions of degrees and is an indication of the release of vast quantities of stored energy, these instruments give a completely new perspective on the hot and turbulent regions of the universe. We describe the new instrumentation, data collection techniques, and scientific goals, along with statistical solutions that use model-based Bayesain deconvoluiton and sophisticated statistical computation to aid in very high resolution imaging, spectral analysis, and time series analysis. *DAvD*

- Lee HKH, Sanso B, Zhou W, Higdon D (2006) Inferring Particle Distribution in a Proton Accelerator Experiment. *Bayesian Analysis* **1**:249-264. This article deals with an inverse problem of attempting to learn about the inputs of a particle accelerator using a computer simulator and data which are complicated nonlinear functions of the inputs. The Bayesian approach is useful because the problem is underspecified, so we can impose some structure while still exploring a range of possible parameter values that are compatible with the observed data. *HKHL*

- Buck CE, Gomez D, Aguilar P, Litton CD, O'Hagan A (2006) Bayesian nonparametric estimation of the radiocarbon calibration curve *Bayesian Analysis* **1**:265-288. The radiocarbon calibration curve is used by archaeologists and other researchers to map radiocarbon determinations onto the calendar scale. The relationship cannot be modelled mechanistically and so it is estimated using high-precision calibration data. This paper takes a nonparametric Bayesian approach to radiocarbon calibration curve estimation, adopting a Gaussian process prior structure. This structure captures the basic a priori knowledge about the curve, but allows the detail to come from the data themselves. A variant on this approach was adopted for constructing

the 2004 internationally-agreed estimates of the calibration curve (Buck and Blackwell, 2004, *Radiocarbon* **46**:1093-1102, and Blackwell and Buck, 2007, forthcoming in the special issue of *BA* from Case Studies in Bayesian Statistics Nine). *CEB*

- Airoldi EM, Anderson AG, Fienberg SE, Skinner KK (2006) Who Wrote Ronald Reagan's Radio Addresses? *Bayesian Analysis* **1**:289-320. This study suggests that President Reagan may not have had a scriptwriter at all; he was much more the author of his words than many have believed. The centerpiece of the analysis is a hierarchical model of words and semantic features that extends and complements the classical work of Mosteller & Wallace. A novel (parametric) statistic to identify relevant words and feature is introduced, and its power and asymptotics explored in the Poisson and Negative-Binomial cases. *EMA*

- Sebastiani P, Xie H, Ramoni MF (2006) Bayesian Analysis of Comparative Microarray Experiments by Model Averaging *Bayesian Analysis* **1**:707–732. A major challenge to the analysis of microarray data is the small number of samples–limited by both cost and sample availability–compared to the large number of genes, now soaring into the tens of thousands per experiment. This situation is made even more difficult by the complex distribution of gene expression measurements and the necessity to limit the number of false detections due to multiple comparisons. This paper introduced a Bayesian method for the discovery of genes with differential expression that uses model averaging to gain robustness over misspecification of the distribution of gene expression data. Another advantage of the method is the use of objective prior distributions that can incorporate information about the background noise of gene expression data to reduce the false positive rate. *PS*

- Rigat F, de Gunst M, van Pelt J (2006) Bayesian Modelling and Analysis of Spatio-Temporal Neuronal Networks. *Bayesian Analysis* **1**:733-764. This paper illustrates a novel hierarchical Bayesian framework modelling functional networks of spiking neurons over time. The hierarchical Bayesian approach is key to derive simulation-based inferences for the parameters characterizing the discrete-time multivariate spiking process, the unknown structure of the functional connections among the analysed neurons and its dependence on fixed-time covariates. The adequacy of the model is investigated using raw residuals and the time-rescaling theorem. When applied to the analysis of experimental multiple spike trains obtained from a culture of neurons in vitro, the model emphasizes the pivotal role or one particular neuron for the initiation of each cycle of network activity and the significant dependence of the estimated network structure on the spatial arrangement of the neurons. *FR*

- Williams B, Higdon D, Gatticker J, Moore L, McKay M, Keller-McNulty S (2006) Combining Experimental Data and Computer Simulations, With an Application to Flyer Plate Experiments. *Bayesian Analysis* **1**:765–792. A flyer plate experiment involves forcing a plane shock wave through stationary test samples of material and measuring the free surface velocity of the target as a function of time. These experiments are conducted to learn about the behavior of materials subjected to high strain rate environments. Computer simulations of flyer plate experiments are conducted with a two-dimensional hydro- dynamic code developed under the Advanced Strategic Computing (ASC) program at Los Alamos National Laboratory. This code incorporates physical models that contain parameters having uncertain values. The objectives of the analyses presented in this paper are to assess the sensitivity of free surface velocity to variations in the uncertain inputs, to constrain the values of these inputs to be consistent with experiment, and to predict free surface velocity based on the constrained inputs. We implement a Bayesian approach that combines detailed physics simulations with experimental data for the desired statistical inference.

- Dominici F, Zeger SL, Parmigiani G, Katz J, Christian P (2007) Does the effect of micronutrient supplementation on neonatal survival vary with respect to the percentiles

of the birth weight distribution? *Bayesian Analysis* **2**:1-30. *This article is followed by comments and a rejoinder.* In developing countries, higher infant mortality is caused in part by poor maternal and fetal nutrition. Infant mortality is greatest among low birth weight infants (2500 g or under). Although it has been demonstrated that nutritional supplementation increases birth weight, there is inconclusive evidence that supplementation improves one year survival rates. It has been hypothesized that a potential benefit to survival among the low birth weight infants is partly compensated by a null or even harmful effect among the largest infants, where in the absence of obstetrical care, increased size would increase the risk of mortatility for both infants and mothers. In this paper we have developed a Bayesian approach to causal inference where we estimated "causal" effects of nutritional supplementation on infant mortality which are allowed to vary with the percentiles of the birth weight. We implement data augmentation methods for estimating the marginal posterior distributions of all parameters of interest accounting for the uncertainty about the missing counterfactuals. A Bayesian approach is particularly suitable here for exploring the sensitivity of the results to unverifiable modeling assumptions and other prior beliefs. We found that the treatment (folic acid, iron, and vitamin A) increased the birth weight of smaller babies and had no apparent effect on larger babies. *FD*

- Cowell RG, Lauritzen SL, Mortera J (2007) A gamma model for DNA mixture analyses. *Bayesian Analysis* **2**:333–348. This paper presents a simple probability model for the peak area values associated with alleles that arise from the amplification of DNA mixtures. It is shown how the model may be represented using a Bayesian network, and thus allow inference to be performed using standard exact propagation algorithms. Applications are (i) separating the DNA mixtures into the genetic profiles of the individual contributors, and (ii) calculating the weight of evidence for a given suspect to have contributed to a mixture under different circumstances. *RGC*

- Xing EP, Sohn K (2007) Hidden Markov Dirichlet Process: Modeling Genetic Inference in Open Ancestral Space *Bayesian Analysis* **2**:491–528. This paper develops a dynamic extension of the Dirichlet process that models the stochastic state-transitions of a hidden Markov model in countably infinite dimensional state space. This new approach represents an alternative to the popular PAC approximation to the coalescent process, and offers an exchangeable and parsimonious generative model for population genotypes originated from recombination and mutation from an unknown number of founders; and it supports joint inference of a recombination hotspot and population structure. On both simulated and real data, this method is competitive in estimating the recombination rates and hotspot, and it generates an ancestral spectrum for representing population structures that reveals details of the genetic diversity of each individual. *EPX*

- Dukic V, Dignam J (2007) Bayesian Hierarchical Multiresolution Hazard Model for the Study of Time-Dependent Failure Patterns in Early Stage Breast Cancer. *Bayesian Analysis* **2**:591–610. The Bayesian multiresolution hazard estimator (MRH) has been recently adapted for estimation of hazard functions (Bouman, Dukic and Meng, 2005, *Statistica Sinica* **15**:325-357; Bouman et al. 2007, forthcoming in *JASA*). In this paper, we have extended the previously proposed MRH methods into the hierarchical multiresolution hazard setting (HMRH). The benefit of HMRH is its ability to model flexibly hazard functions within each of several patient strata while borrowing strength and allowing for common covariate effects (such as of age etc.) across all strata, and accounting for within-stratum correlation. The stratum-specific hazard rates also readily facilitate comparison and testing across strata. The HMRH method was used to examine patterns of recurrence after treatment for early stage breast cancer within four treatment strata, using data from two large-scale randomized clinical trials. *VD*

APPLICATIONS

STATISTICAL ISSUES IN PROTEOMIC RESEARCH

Jeffrey S. Morris
jefmorris@mdanderson.org

Over the past decade, advances in the genomic revolution have also fueled increasing interest in the field of proteomics. Proteomics differs from genomics in that proteins are directly measured instead of their precursors, genes and messenger RNA. In this article, I will overview some work I have done in various statistical aspects of proteomics data, including experimental design, preprocessing, and analysis of the data.

Like many other high throughput technologies, proteomic methods can be very sensitive to varying experimental conditions and sample preprocessing, frequently leading to systematic differences in data obtained at different times or from samples with different handling conditions. As a result, randomized block designs should be used in running the samples, and care must be taken to avoid confounding factors of interest with experimental factors like run time or sample processing protocol (Baggerly, Morris, Coombes 2004; Baggerly, et al. 2004; Baggerly, Coombes, Morris, 2005; Baggerly, et al. 2005; Cooombes, et al. 2005b; Hu, et al. 2005; Morris, et al. 2007a; Gutstein and Morris, 2007; Gutstein, et al. 2007). Otherwise, processing artifacts can show up as strong apparent treatment effects that turn out to be embarrassing false positives.

Most proteomic technologies yield complex functional data with local features corresponding to proteins present in the sample. For example, matrix assisted desorption and ionization time of flight mass spectrometry (MALDI-TOF) yields spectra, spiky functions with x-axis molecular mass (per unit charge) and y-axis intensity (see Figure 1a). 2D gel electrophoresis (2DE) yields image data, with x-axis isolectric point (pH), y-axis molecular mass, and z-axis intensity (see Figure 1b). The spikes in MALDI-TOF and the spots in 2DE data roughly correspond to proteins and peptides present in the biological sample, and their intensities are rough measures of the corresponding abundance, at least in a relative sense. Overviews of MALDI-TOF can be found in Coombes, et al. (2005a), Baggerly, Coombes, and Morris (2006), and Morris, et al. (2007b). An overview of 2DE can be found in Morris and Gutstein (2007).

There are important preprocessing steps that should be performed on the data before subsequent analysis, including denoising, background correction, and normalization (Baggerly, Morris, and Coombes 2003; Coombes, et al. 2003; Coombes, et al. 2005c; Coombes, Baggerly, and Morris 2007; Morris and Gutstein 2007). After preprocessing, there are two major approaches to analyzing these data: feature extraction and functional modeling.

In feature extraction, discrete relevant features of the data (e.g. spikes and spots) are detected and quantified for each sample, resulting in an $N \times p$ matrix containing protein intensities for p features from each of N spectra/gels. This matrix is subsequently analyzed using any of a number of Bayesian or frequentist methods to figure out which features are related to the biological factors being studied. Various frequentist and Bayesian methods are available to estimate and/or control the false discovery rate (FDR) when flagging a feature as significant. We have found that performing feature detection on the average spectrum/gel tends to result in greater sensitivity and specificity, since in the average noise washes out while real features tend to be reinforced across spectra/gels (Morris, et al. 2005; Morris and Gutstein 2007). Also, we have found that using peak or pinnacle intensities for feature quantification has advantages over computing peak or spot volumes (Morris and Gutstein 2007). Algorithms for feature extraction for MALDI-TOF can be found in Coombes, et al. (2005c), Morris, et al. (2005), and Coombes, et al. (2007). A stand alone executable *PrepMS* implementing these methods is described in Karpievitch, et al. (2007). *Pinnacle*, a spot detection and quantification algorithm for 2DE, is described in Morris and Gutstein (2007), and stand alone code is currently being developed. However, no feature detection method is perfect, and since group comparisons are only performed on the detected features, important discoveries can be missed for features falling below the detection threshold. The functional modeling approach is free of this concern.

The functional modeling approach involves

modeling the entire spectrum or image as a function using functional data analysis techniques. Since the functions underlying proteomics data tend to be very irregular with many local features, they require flexible modeling approaches and spatially adaptive regularization for accurate represention. The wavelet-based functional mixed model (WFMM, Morris and Carroll 2006, Morris, et al. 2006) possesses these properties, and seems to be a suitable functional modeling technique for these data. Morris, et al. (2007b) applied this methodology to MALDI-TOF data.

In the functional mixed model, a functional response is related to a set of scalar predictors through *fixed effect functions* $B_i(t)$ of unspecified form, which each models the effect of its corresponding factor across the domain of the function. The model also contains *random effect functions* of unspecified form that provide a convenient mechanism for modeling any correlation between functions imposed by the experimental design, e.g. replicate functions from the same patient. Morris and Carroll (2006) introduce a Bayesian wavelet-based approach for fitting this model. The use of wavelet shrinkage for modeling and function regularization allows the accommodation of very spiky fixed and random effect functions, and also enables parsimonious yet flexible representations of the various between-curve covariance structures. A Metropolis-Gibbs scheme is used to obtain posterior samples for all model parameters, including the fixed effect functions, which can be used to perform Bayesian inference. The only informative priors involve smoothing parameters, which if desired can be estimated from the data using an empirical Bayes approach. Herrick and Morris (2006) discuss stand alone executable code for running this method that is freely available for download (http://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=70).

Given the posterior samples, for each fixed effect function one can compute the posterior probability of a certain effect size of interest for each location within the function. For example, if the $\log_2$ intensities are modeled and we are interested in at least 2-fold expression differences, we could compute $p_i(t) = Pr(|B_i(t)| > 1|data)$ for each t. The quantity $1 - p_i(t)$ is a local false discovery rate (FDR) estimate for calling location t a discovery, defined as a 2-fold expression difference in the factor of interest. Morris, et al. (2007b) describe how to identify a cutpoint on these estimates that maintains a specified average FDR rate across the entire function. This Bayesian approach to FDR is straightforward, and can be applied to any setting where effect sizes are modeled in a Bayesian model.

Figure 2 contains some results from applying the WFMM to two studies described in Morris, et al. (2007b). The first is a pancreatic cancer study in which MALDI-MS was run on serum samples from pancreatic cancer patients and healthy controls. The results in the figure are for the cancer vs. normal fixed effect function, whose interpretation is the difference between cancer and normal mean spectra, after adjusting for block effects. The second study is a mouse study, in which the animals had one of two cancer cell lines injected into one of two organs, their lungs or brain. After the tumors were allowed to grow, MALDI-MS was run on the animals' serum samples. The results in the figure present the organ main effect function, which describes systematic differences between the lung- and brain-injected animals after adjusting for other factors in the model. Looking for regions at least 1.5 fold different with average FDR 0.10 and 2.0 fold different with average FDR 0.05, respectively, the green regions of the curves in the right column of the figure indicate the regions of the spectra we would flag as significant.

There are a number of advantages to using the WFMM to model MALDI-TOF data. It can identify differentially expressed regions of the spectra, which map to differentially expressed proteins, without reliance on peak detection methods. Also, the effects of various factors on the spectra can be modeled simultaneously, enabling one to answer a number of different research questions with a single model fit and also to nonparametrically model out systematic effects of nuisance factors such as run block that can affect both the location and intensities of the peaks (Morris and Carroll, 2007b). Work is currently underway to generalize the method so it can handle higher dimensional functions (e.g. images), which will allow it to be used with 2DE and other multidimensional functional data. Studies need to be conducted to compare the functional modeling and feature extraction approaches to better understand the benefits and drawbacks of each.

Proteomics is an exciting, growing field with many challenges. Careful consideration of the statistical considerations in experimental design, preprocessing, and analysis is essential for the success of the field.

References

- Baggerly, K. A., Coombes, K. R., and Morris, J. S. (2005). Are the NCI/FDA ovarian proteomic data biased? A reply to "Producers and Consumers". *Cancer Informatics*, **1(1)**, April 14, 2005.

- Baggerly, K. A., Coombes, K. R., and Morris, J. S. (2006). An introduction to high-throughput bioinformatics data. In *Bayesian Inference for Gene Expression and Proteomics*, K. A. Do, P. Mueller, and M. Vannucci, editors, Cambridge University Press, pages 1-39.

- Baggerly, K. A., Edmondson, S., Morris, J. S., and Coombes, K. R. (2004). High-resolution serum proteomic patterns for ovarian cancer detection. *Endocrine-Related Cancers*, **11(4)**, 583-584.

- Baggerly K., Morris J. S., Wang J., Gold D., Xiao L. C. and Coombes K. (2003). A comprehensive approach to the analysis of MALDI-TOF proteomics spectra from serum samples. *Proteomics*, **3**, 1667-1672.

- Baggerly, K. A., Morris, J. S., and Coombes, K. R. (2004). Reproducibility of SELDI mass spectrometry patterns in serum: comparing proteomic data sets from different experiments. *Bioinformatics*, **20(5)**, 777-785.

- Baggerly, K. A., Morris, J. S., Edmonson, S., and Coombes, K. R. (2005). Evaluating reported reproducibility of serum proteomic tests for ovarian cancer (with commentary). *Journal of the National Cancer Institute*, **97**, 307-309.

- Coombes, K. R., Baggerly, K. A., and Morris, J. S. (2007). Preprocessing mass spectrometry data. In *Fundamentals of Data Mining in Genomics and Proteomics*, M. Granzow and D. Berrar, editors, Kluwer, *to appear*.

- Coombes K. R., Fritsche H. A. Jr., Clarke C., Cheng J. N., Baggerly K. A., Morris J. S., Xiao L. C., Hung M. C., and Kuerer H. M. (2003). Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid Using Surface Enhanced Laser Desorption and Ionization. *Clinical Chemistry*, **49(10)**, 1615-1623.

- Coombes, K. R., Koomen, J. M., Baggerly, K. A., Morris, J. S., and Kobayashi, R. (2005a). Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics*, **1(1)**, April 14, 2005.

- Coombes, K. R., Morris, J. S., Hu, J., Edmonson, S. R., and Baggerly, K. A. (2005b). Serum proteomics profiling: A young technology begins to mature. *Nature Biotechnology*, **23(3)**, 291-292.

- Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., and Kuerer, H. M. (2005c). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectrua with the undecimated discrete wavelet transform. *Proteomics*, **5**, 4107-4117.

- Gutstein H. B. and Morris J. S. (2007). Laser capture sampling and analytical issues in proteomics. *Expert Reviews in Proteomics*, **4(5)**, to appear.

- Gutstein H. B., Morris J. S., Palani S. B. A., and Sweedler J. V. (2007) Microproteomics: Analysis of protein diversity in small samples. *Mass Spectrometry Reviews*, to appear.

- Herrick R. C. and Morris J. S. (2006). Wavelet-based functional mixed model analysis: Computational Considerations. *Proceedings, Joint Statistical Meetings, ASA Section on Statistical Computing*, 2051-2053.

- Hu, J., Coombes, K. R., Morris, J. S., and Baggerly, K. A. (2005). The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales. *Briefings in Genomics and Proteomics*, **3(4)**, 322-331.

- Karpievitch, Y. V., Hill, E. G., Morris, J. S., Coombes, K. R., Baggerly, K. A., and Almeida, J. S. (2006). PrepMS. *Bioinformatics*, **23(2)**, 264-265.

- Morris, J. S., Arroyo, C., Coull, B. A., Ryan, L. M., Herrick, R. C., and Gortmaker, S. L. (2006a). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *Journal of the American Statistical Association*, **101(476)**, 1352-1364.

- Morris J. S., Baggerly K. A., Gutstein H. B., and Coombes K. R. (2007a). Statistical contributions to proteomic research. *The Urine Proteome*, Humana, to be published.
- Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., and Coombes, K. R. (2007b). Bayesian analysis of mass spectrometry data using wavelet-based functional mixed models. *Biometrics*, doi: 10.1111/j.1541-0420.2007.00895.x.
- Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, **68(2)**, 179-199
- Morris, J. S., Coombes, K. R., Kooman, J., Baggerly, K. A., and Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum. *Bioinformatics*, **21(9)**, 1764-1775.
- Morris, J. S. and Gutstein, H. B. (2007). A fast, automatic method for detecting and quantifying protein spots in 2-d gel electrophoresis data. *Bioinformatics*, under revision.

Figure 1: Sample MALDI-MS spectrum (a) and 2DE image (b). Note how for both of these applications, data can be viewed as functions, with many local features (spikes/spots) corresponding to proteins present in the biological sample.

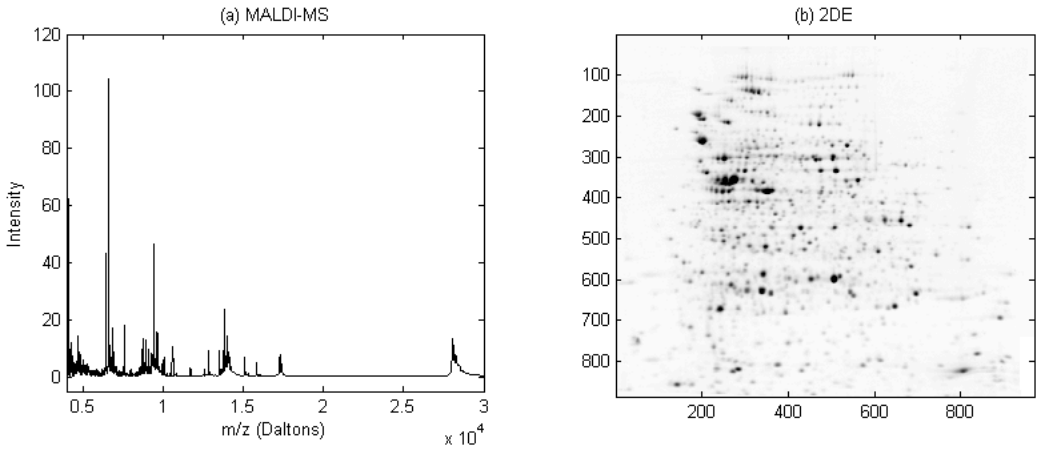
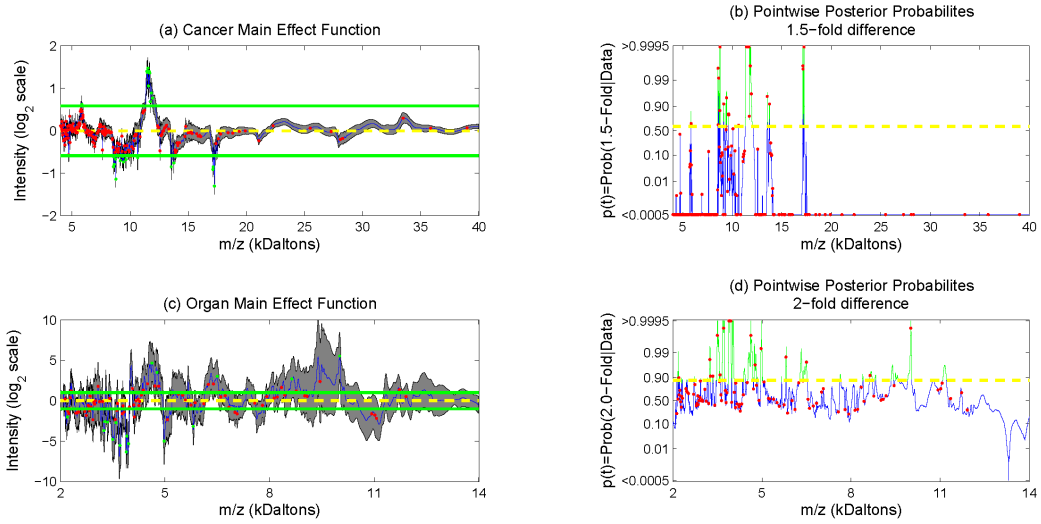


Figure 2: Results from WFMM applied to two MALDI-MS experiments. (a) and (c): Posterior mean and 95% pointwise credible intervals for cancer vs. normal main effect function in pancreatic cancer study, and lung vs. brain organ main effect function in mouse study. The green lines indicate 1.5-fold and 2.0-fold differences in the two examples, respectively, and the dots indicate peaks detected using the average spectrum. (b) and (d): Pointwise posterior probabilities (1-local FDR) of (b) 1.5-fold difference in cancer/normal in pancreatic cancer study, and (d) 2.-fold difference in brain/lung in mouse study. The red dots indicate detected peaks, and the yellow dotted lines indicate the threshold for flagging a location as significant, controlling the expected Bayesian FDR to be less than 0.10 and 0.05 in the two examples, respectively. The green lines mark the regions of the spectra flagged as significant using these criteria.[Figure from Morris, et al. 2007b]



SOFTWARE HIGHLIGHT

SOFTWARE FOR GRAPHICAL MODELS: A REVIEW

by Kevin Murphy
murphyk@cs.ubc.ca

Graphical models (GMs) are a way to represent conditional independence assumptions by using graphs. Specifically, nodes represent random variables and lack of edges represent conditional independencies. The graph is a useful visual representation of complex stochastic systems. The graphical structure is also the basis of efficient inference algorithms.

There are many different kinds of graphical models, but the two most popular ones are based on directed acylic graphs (also called "Bayesian networks") and on undirected graphs (also called "Markov random fields"). In this article, we review some of the more popular and/or recent software packages for dealing with graphical models. A more extensive comparison can be found at http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html.

BUGS

www.mrc-bsu.cam.ac.uk/bugs

BUGS (Bayesian inference using Gibbs Sampling) assumes the model is specified in the form of a DAG (directed acyclic graph), and uses Gibbs sampling for inference. A large number of different conditional distributions (node types) are supported. Internally, various algorithms (such as adaptive rejection sampling and slice sampling) are used to sample from the full conditionals. The software is easy to use, especially since it has recently become possible to call

it directly from R (using R2WinBUGS) and Matlab (using MatBUGS MatBUGS), thus bypassing the rather cumbersome GUI.

Unfortunately, single site Gibbs sampling can be very slow to "mix", resulting in unreliable posterior inferences. In addition, Gibbs sampling cannot be used to find posterior modes, and cannot easily be used to compute the marginal likelihood, which is useful for model selection (although BUGS does return the DIC score of a model).

BUGS is freely available as an executable file. The most recent version, called WinBUGS, only runs on Windows (although one can run it on Linux systems using Wine). Recently, an open-source alternative called OpenBUGS has been created, but it is not nearly as mature as WinBUGS. OpenBUGS is written in a language called "Component Pascal".

JAGS
www-ice.iarc.fr/ martyn/software/jags/

JAGS (Just Another Gibbs Sampler) is very similar in functionality to BUGS. The main difference is that it is fully open source, and works easily on multiple platforms (Windows, unix, etc). The principle advantage over OpenBUGS is that it is written in Java, which is a more widely known language than Component Pascal. In addition, it seems to have a simpler design than OpenBUGS.

VIBES
vibes.sourceforge.net/

VIBES (Variational Inference for Bayesian Networks) is open-source Java, and is designed to be similar to BUGS in functionality, but it uses the variational mean field algorithm for inference. This is potentially much faster, but less accurate than Gibbs sampling. In addition, it is limited to the conjugate exponential family. Note that VIBES is no longer being supported; its author is developing a replacement called Infer.NET (see below).

Infer.NET
research.microsoft.com/mlp/ml/Infer/Infer.htm

Infer.NET is a software package developed at Microsoft Research in Cambridge. They anticipate an initial public release in Spring 2008. The code will not be open source but will be freely available for academic use. The model is specified using a new programming language called Csoft, which allows one to combine stochastic code with standard C# code. Thus one can easily specify graphical models of various kinds. Various Bayesian inference algorithms are supported, including Gibbs sampling, variational mean field, and expectation propagation. The package is designed to generate model-specific code, and to run very fast, even on large models.

BNT
www.cs.ubc.ca/ murphyk/Software/BNT/bnt.html

BNT (Bayes Net Toolbox) is open-source Matlab, and supports many different models and inference algorithms. In particular, it supports DAG models, "dynamic Bayesian networks" (which are DAG models unrolled in time) and influence/ decision diagrams. It also has undocumented and partial support for undirected models.

In terms of inference, BNT, like many other GM packages, can only perform Bayesian inference on discrete or Gaussian random variables. Hence parameter inference is performed using point estimation techniques such as EM or gradient descent. However, conditional on the parameters, inference of the remaining variables can often be performed exactly, using the junction tree algorithm, which includes well-known algorithms, such as the forwards-backwards algorithm, as special cases. If exact methods are too slow, a variety of different approximate inference algorithms are supported, such as "loopy belief propagation".

Hugin
www.hugin.com

Hugin www.hugin.com is a commercial package with functionality similar to BNT. It was one of the first packages for DAG models (including influence diagrams), and it is arguably the most mature. However, there are now a large number of other packages, such as Genie, MSBNx, Netica, PNL, etc. with very similar functionality (see `http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html` for a comparison). These packages focus on exact inference in discrete-state (or conditionally Gaussian) models, using the junction tree or variable elimination algorithm. Some of them also support parameter estimation using EM. These packages are aimed at the business/ data-mining market, and hence they often put more emphasis on user interface and I/O issues than on core functionality.

gR
www.r-project.org/gR

gR (graphical models in R) is a collection of packages rather than a single package. The main package is `gRbase`, which is a way of defining data and models. There is also the `dynamicGraph` package, for visualizing and editing graphs. There are no Bayesian inference alorithms implemented in R. However, R interfaces to several existing model-fitting packages are provided, including `CoCo`, and `mimR`, both of which are designed for fitting contingency tables, which can be represented as undirected GMs.

Blaise
publications.csail.mit.edu/abstracts/abstracts07/bonawitz/bonawitz.html

Blaise is a Java software package that supports efficient Monte Carlo inference (including MCMC and sequential Monte Carlo samplers) in a large class of probabilistic models, including directed graphical models and non-parametric Bayesian models. The plan is to release a first version to the public, under a restricted open source license, in Spring 2008.

Gaussian graphical models

Gaussian graphical models are an important special case of graphical models that support efficient Bayesian inference using techniques from sparse linear algebra. GMRFsim supports inference in undirected GGMs, and GDAGSim supports inference in directed GGMs. Both of these can be used to perform block sampling inside an MCMC sampler. The ggm R package can be used to fit undirected GGMs parameters using point estimation techniques.

Model selection

In addition to inference about states and parameters, there is much interest (especially in the systems biology community) in inference about the graph structure itself. The model selection problem is very difficult, because the space of all graphs on n nodes has size $O(2^{n^2})$. There are basically three main approaches to this: greedy search (and variants), MCMC model averaging, and constraint-based methods. Most of the work has focused on learning DAG models, although the WinMine WinMine package learns dependency networks.

There are many packages that perform greedy search in DAG space: BNT, DAGlearn, Banjo, DealDeal, etc. BNT supports simple hill-climbing. DAGlearn uses L1-penalized logistic regression to reduce the search space. Banjo uses simulated annealing. Deal uses hill-climbing, but can handle conditionally Gaussian models (the other packages assume discrete data).

There are very few publically available packages that perform Bayesian model averaging in the space of DAGs. BNT implements a Metropolis Hastings method with a simple local proposal. BDAGL uses a more sophisticated proposal based on dynamic programming. The GGM package does model averaging in the space of undirected Gaussian GMs, using MCMC and stochastic search techniques. The `http://www.stat.duke.edu/~adobra/hdbcs.html`HdBCS (high dimensional Bayesian covariance selection) package is similar to the GGM package, but searches in the space of DAGs and then converts the result to an undirected GGM.

The constraint-based approach to structure learning, in which one eliminates edges if certain conditional independencies are detected in the data (using some hypothesis testing procedure), is generally faster but more error-prone than the above Bayesian techniques. BNT implements some of the simpler algorithms. Tetrad is a more elaborate package.

Traditionally, the constraint-based approach has been the method of choice for people interested in learning "causal" models from observational data. However, this approach can of course be used to fit "acausal" models, too. For example, the SIN R package uses conditional independency tests to learn the structure of undirected Gaussian GMs. The GeneNet R package uses an FDR approach to threshold the partial correlation coefficients to induce a sparse GGM. The glasso R package uses L1 regularization to estimate a sparse precision matrix.

STUDENTS' CORNER

ENDLESS CONFERENCES

by Luke Bornn
l.bornn@stat.ubc.ca

In this issue we feature an article by Carlos M. Carvalho discussing the benefits of diversifying your thesis work and doing a post-doc. Following this article we present two dissertation abstracts from recent graduates of Duke University. If you would like to see your abstract published or have ideas for future articles in the Students' Corner, I'd be happy to hear from you.

Before the articles and abstracts, however, I want to highlight some references which list upcoming conferences in the statistical sciences. With hundreds of relevant conferences each year, it is easy to overlook many conferences of potential interest. The following links will help guide you in selecting the conferences that best fit your interests.

www.amstat.org/dateline
www.imstat.org/meetings/2008.htm
www.bayesian.org/business/meetings.html
www.conferencealerts.com/statistics.htm

Now the tough part - asking your supervisor to pay for your trip. Maybe we'll cover that topic in a future issue...

SMOOTHING THE TRANSITION

by Carlos M. Carvalho
Carlos.Carvalho@chicagogsb.edu
http://faculty.chicagogsb.edu/carlos.carvalho
Graduate School of Business, The University of Chicago

"What should the next step of my career be?" That's a very common question faced by graduate students completing their doctoral education. Fortunately, unlike many disciplines, we statisticians have an abundance of options both in academia and industry – which in some ways only adds to the complexity of this question. Though my experience was somewhat unique, I have learned a few things that could be helpful to advanced graduate students as they look forward in their careers. In this article I will briefly describe my academic trajectory and focus on a couple of steps that I found to be very important in the transition from graduate school to an assistant professorship.

First, I want to describe my academic trajectory. I completed my Ph.D. in Statistics under the supervision of Mike West at Duke University in February 2006. My graduate work focused on the development of methodological aspects of sparse models for large-scale multivariate problems and associated computational tools for model selection and inference. From an applied perspective, two areas played an important motivating role in my thesis: the study of biological pathways in genome wide expression data and large scale dynamic portfolio problems. After finishing my Ph.D., I started a post-doctoral position at Duke where I was involved with the Duke Integrated Cancer Biology Program. Most of my efforts in that position focused on extending and applying the genomic related aspects of my thesis. Finally, in July 2007, I started as an Assistant Professor of Econometrics and Statistics at The University of Chicago Graduate School of Business.

As you can see, this was somewhat of a unique path (and so is the path of most statisticians) but I will highlight two points in my experience that I believe to be relevant to any job market candidate or assistant professor to be. The first one is the ***exploration of the interdisciplinary opportunities*** of my work. From the very beginning it was clear to me that my work with Mike would focus on the development of statistical methods to deal with large-scale genomic problems. Having an economic undergraduate training meant that I had to invest a lot of time in learning about biology, genetics, etc. I was fortunate to be exposed to a very fruitful cross-disciplinary work environment that allowed me to acquire the necessary knowledge not only through classes but also through constant interaction with biologists, doctors, and other scientists involved in the research efforts. This was not easy nor was it costless; however, it provided me with the necessary "language" to move forward with our projects in a much more efficient pace. In general, having a true understanding of the subject area opens many career options, funding opportunities and enhances the statistician's ability to innovate by approaching problems through new directions.

Given my background and previous interests, I was always trying to build bridges between the methodologies I was developing and possible application in finance and economics. Eventually I found a connection that allowed me to spend a significant proportion of my thesis discussing applications in large-scale dynamic portfolio problems. By the time I was about to defend, it was very clear that my efforts in genetics and finance gave me the ability to apply to a variety of different jobs in many different departments.

The completion of my thesis brings me to the second point in my experience that I want to emphasize: ***the benefits of a post-doctoral position***. Doing a post-doc is not a necessary step in a statistician's career (although it's getting more and more popular!) but it was a fundamental one for me. The first reason is very simple: timing. Graduate students trying to finish their thesis while applying for tenure-track jobs have a lot on their plate. Writing all the applications takes a lot of time and flying around interviewing is pretty much all you have time to do for one or two months. Now, add to that the stress related to the final strides of a thesis and you have yourself a really difficult last year in graduate school. On top of that, in most cases, students are about to submit their thesis work for publication and due to the long turnaround times in our field it means that they will be going into the market with lots of working papers but few publications. Whether we like it or not, the reality is that most successful applicants for assistant professor positions have a few publications by the time they graduate. My decision to take a post-doc position allowed me to finish my thesis more carefully and provided me with time to submit and get most of my papers published before I applied to any tenure-track job. I truly believe that had I gone into the market right after graduate school I would not have had the same opportunities that I had a year or so later.

The second, and perhaps the most important advantage of a post-doc is to give a young researcher time to mature and organize their research agenda for the first few years as an assistant professor. A year as a post-doc gave me plenty of time to think about where I wanted to take my career and my research. It provided me with the opportunity to start a number of new projects that now serve as a base for my research efforts in my first year as a professor. To be perfectly honest, it's not an easy transition: suddenly, as a junior professor, you'll be teaching new classes, participating in committees, getting used to a new social structure and you are also supposed to carry on with your research. My post-doc made this transition much smoother and I significantly improved my chances to succeed in this new stage of my career. An old Brazilian proverb says that "if advice was worth anything, it wouldn't be free"– so take from this what you will... finally, feel free to contact me if you have any questions. Cheers.

Dissertation Abstracts

CONDITIONS FOR RAPID AND TORPID MIXING OF PARALLEL AND SIMULATED TEMPERING ON MULTIMODAL DISTRIBUTIONS

by Dawn Woodard
dawn@stat.duke.edu
http://www.stat.duke.edu/~dawn
Department of Statistical Science, Duke University
PhD Supervisor: Scott Schmidler (Duke)

Stochastic sampling methods are ubiquitous in Bayesian statistics, statistical mechanics, and theoretical computer science. However, when the distribution that is being sampled is multimodal, many of these techniques converge slowly, so that a great deal of computing time is necessary to obtain reliable answers. Parallel and simulated tempering are sampling methods that are designed to converge quickly even for multimodal distributions. This defense addresses the extent to which this goal is achieved.

We give conditions under which a Markov chain constructed via parallel or simulated tempering is guaranteed to be rapidly mixing, meaning that it converges quickly. We provide lower bounds on the convergence rates of parallel and simulated tempering, which imply a single set of sufficient conditions for rapid mixing of both techniques. Similarly, we obtain upper bounds on the convergence rates of parallel and simulated tempering and conditions for torpid mixing. We then give a number of normal mixture and Potts models and show rapid or torpid mixing for each. These examples suggest that similar results might be obtained for broad classes of distributions.

On Evolutionary Theory, Inference, and Simulation: A Geneological Perspective

by Scotland Leman
leman@vt.edu
http://www.stat.vt.edu/~leman
Department of Statistics, Virginia Tech
PhD Supervisor: Michael Lavine (Duke)

This thesis discusses evolutionary inference from both a modeling perspective and the algorithms associated with performing statistical inference.

Genetic data (DNA) takes on a nontraditional form in that a single observation encompass at least hundreds of base pairs and is nonnumeric in nature. Beyond this fact, DNA from individuals that share a common ancestry have similarities in their genetic makeup, so the notion of independent and identically distributed samples does not hold. In turn, a complex network of associations must be employed when modeling the data.

The complexities involved in the modeling procedure directly relate to the complexities involved when reconstructing likelihood functions, or posterior distribution. Many computational methods used during statistical inference involve the idea of drawing samples from proposal distributions. However, such proposal distributions are difficult to construct so that their probability distribution match that of the true target distribution, in turn hampering the efficiency of the overall sampling scheme.

We will describe a general approach to modeling the evolutionary past. Within this framework, we will discuss specific models which address particular phenomena (speciation, introgression and paracentric inversions) which relate to genomic data. The latter part of this thesis will address two simulation methods used for statistical inference. The first will pertain to direct likelihood construction under an Importance Sampling (IS) framework and the second will address a Markov Chain Monte Carlo (MCMC) procedure for posterior sampling.

NEWS FROM THE WORLD

Announcements

ISBA Election results

The 2007 ISBA elections ran 15 Oct - 15 Nov. Election results are available on-line at http://www.bayesian.org/election/election2007.html. Congratulations to the winners and thanks to all participants! Thanks too to the 251 ISBA members who voted this year (a 53% turn-out).

Call for 2007 Mitchell Prize

Nominations are still being accepted for the 2007 Mitchell Prize. The 2007 Mitchell Prize is awarded in recognition of an outstanding paper where a Bayesian analysis has been used to solve an important applied problem. The Prize includes a commemorative plaque an award of $1,000. Eligible papers for the 2007 Mitchell Prize must be published or accepted in a refereed journal or conference proceedings during 2005 or 2006. Deadline for submissions is 31 December 2007. For details on nomination for the 2007 Mitchell Prize please visit http://www.stat.duke.edu/apps/MitchellPrize.

New Membership Partnership with IMS

The Institute of Mathematical Statistics (IMS) and ISBA are pleased to announce that ISBA members can join (or renew) with IMS at 25% off the regular IMS dues rate, and that IMS members can join (or renew) with ISBA at 25% off the regular ISBA dues rate. For example, ISBA members pay US$71 for a year's IMS membership. For all the IMS dues and subscription prices for individual members, see the IMS or (soon) ISBA join/renewal pages.

Degree programs at the University of California at Irvine

The Department of Statistics at the University of California at Irvine would like to announce its new M.S. and Ph.D. programs in Statistics, which were approved as of Fall 2006. Our department was formed in 2002 with the selection of Hal

Stern as founding chair, and since then, seven additional faculty have been recruited. We are now accepting applications (until the January 15, 2007 deadline) for admission for Fall 2008. For information about our department, please go to http://www.ics.uci.edu/statistics/. Once there, click on "Graduate degrees approved" for details about our degrees.

Jobs

Faculty position at Pontificia Universidad Catolica de Chile

The Department of Statistics, Pontificia Universidad Catolica de Chile, invites applications for a tenure-track position at the Assistant Professor level, beginning August 1, 2008. Candidates should have a doctoral degree in Statistics or a related field and exhibit experience in both research and teaching. Appointees will be expected to pursue a vigorous research program and to teach three undergraduate/graduate courses per year.

The Department of Statistics is the leading research group in Chile and offers an undergraduate career, a Master's, and a Ph.D. programs in Statistics. The Pontificia Universidad Catolica de Chile is a highly selective institution and its students are among the top 5% in the country.

Email letter of application, including a statement of research interests, and curriculum vitae with publication list to stat@mat.puc.cl. Send at least three letters of reference, relevant reprints/preprints, and transcripts to Prof. Wilfredo Palma, Director, Departamento de Estadistica, Pontificia Universidad Catolica de Chile, Casilla 306, Santiago 22, Chile.

For full consideration, complete application materials must arrive by March 21, 2008.

Faculty positions at Duke University

Statistics position: The Department of Statistical Science (http://www.stat.duke.edu) at Duke University invites applications for faculty appointment at the level of Associate or Assistant Professor to begin in Fall 2008. Preference will be given to candidates whose core statistical science research interests are complemented with collaborative research interest in systems biology, neurosciences, social sciences, or environmental science.

The Department of Statistical Science is an internationally recognized center of excellence for research and education in the development and application of contemporary statistical methodology. Particular emphasis is directed toward Bayesian modeling in many scientific fields as well as emerging computationally intensive methods. The Department offers outstanding computational facilities and opportunities for interdisciplinary research. It currently has 12 regular rank faculty along with 13 visiting, adjunct, and post doctoral faculty and 33 Ph.D. students.

The Ph.D. program as well as the Department's research agenda benefit from strong connections with the Statistics and Applied Mathematical Sciences Institute (SAMSI) and the National Institute of Statistical Science (NISS), both located nearby in the Research Triangle. A Statistical Science major, started last Fall, provides the primary focus of our undergraduate program. More information about the Department is available at the web site http://www.stat.duke.edu.

All applicants should send a letter, curriculum vitae, and the names of three references. Mail applications to Faculty Search Committee, DSS, Duke University, Box 90251, Durham, NC 27708-0251. For inquiries and e-mail correspondence please write to search@stat.duke.edu. The application pool will remain open until the position is filled but screening will begin on 1 December, 2007.

Duke University is an Equal Opportunity/Affirmative Action Employer. Applications from women and minorities are strongly encouraged.

*System biology positions:*The Department of Statistical Science (http://www.stat.duke.edu) at Duke University is a participating department in the campus-wide systems biology program. New faculty positions in the broad field of cellular systems biology are now being advertised - see the announcement and call for applications below. Appointments can be made in any participating department. Individuals whose interests and expertise are consistent with appointment in the Department of Statistical Science can email stats-sysbiology@stat.duke.edu to indicate their interest and also any enquiries. Formal application is as described in the ad below. We would also appreciate your help in bringing these opportunities to the attention of potentially interested colleagues and past students and postdocs in statistics and allied fields, as well as to the broader scientific community at your institution.

Duke University seeks applications for open rank, tenure track positions in the broad field of cellular systems biology. We seek applicants from both experimental and quantitative/computational disciplines with research interests in the molecular bases of cellular function, development, and evolution. These new appointments will substantially enhance existing Duke strengths in experimental and modeling approaches to understanding the complexity of genetic, metabolic, and signaling networks. Successful applicants will have appointments in one or more Duke departments based on mutual interests. All appointees will be affiliated with the Duke Center for Systems Biology, a cross-school, campus-wide academic center that is also one of the NIH-supported National Systems Biology Centers. Applicants should submit a curriculum vitae, a brief summary of current and proposed research, reprints of 2 or 3 key publications and a statement of teaching interests via the web at www.academicjobsonline.org. Junior candidates should arrange for three letters of recommendation to be uploaded to this website or sent directly to:

Systems Biology Search, Duke University, Box 90338, Durham, NC 27708-0338.

Senior candidates should give the names of three potential referees. Application review will begin on December 1 2007, and continue until the positions are filled.

Duke University is an Equal Opportunity/Affirmative Action Employer; women and members of minority groups are strongly encouraged to apply.

Events

MCMCSki II: Markov Chain Monte Carlo in Theory and Practice, Bormio, Italy, 9-11 Jan. 2008.

The unifying theme of the third joint international meeting of the IMS and ISBA is MCMC and its impact on the theory and practice of statistics, but invited sessions and poster presentations will cover a broad range of statistical topics. Plenary speakers are Peter Green of the University of Bristol, Kerrie Mengersen of the Queensland University of Technology and Xiao-Li Meng of Harvard University.

There will also be a pre-conference "satellite" meeting, from 7-8 January, intended to provide a snapshot of the methodological, practical and theoretical aspects of an emerging group of methods (adaptive MCMC, adaptive population Monte Carlo, and various breeds of adaptive importance sampling amongst others) that attempt to automatically optimize their performance for a given task.

Abstract submission for poster presentations is now open. Limited financial support for the travel of junior (< 5 years since PhD) is anticipated for those presenting in a poster session. For more information visit the website, http://musing.unipv.it/IMS-ISBA-08/, or contact Brad Carlin brad@biostat.umn.edu.

University of Florida Tenth Annual Winter Workshop: Bayesian Model Selection and Objective Methods, Gainsville, Florida, 11-12 Jan. 2008.

The workshop will focus on recent developments in the area of Bayesian model selection and those aspects of objective methods that have to do with model selection. A major purpose of the workshop is to discuss the many recent significant developments in Bayesian model selection; to discuss practical implementation; and to identify important problems and new research directions. All sessions are plenary, and the invited speakers include Merlise Clyde, Ming-Hui Chen, David Draper, Dean Foster, Ed George, Michael Jordan, David Madigan, Glen Meeden, Adrian Raftery, Marina Vannucci, Yuhong Yang. The workshop will also include a contributed poster session. For more information, visit the website http://www.stat.ufl.edu/symposium/2008/index.html.

Bayesian Biostatistics, Houston, Texas, 30 Jan - 1 Feb. 2008.

Current and prospective users of Bayesian biostatistics are invited to join experts in the field for this three-day conference. Attendees will have the opportunity to attend two courses on the first day: Introduction to Bayesian Statistics and Application of Bayesian Methods to Drug and Medical Device Development. On the following two days, invited and contributed talks will cover a range of topics including comprehensive decision modeling (incorporating utilities), prior distributions in clinical studies and drug development, what Bayesian methods can provide traditional methods cannot, Bayesian methods in medical journals, Bayesian

methods in epidemiology, Bayesian methods and medical ethics, how to build a cadre of Bayesian experts, why Bayesian methods are not more widely used, how to ensure quality in Bayesian applications, and guidelines for publishing Bayesian analyses. For more information, visit http://www.mdanderson.org/departments/biostats/ or contact Lydia David lbdavis@mdanderson.org

9th Brazilian Meeting on Bayesian Statistics (EBEB), Maresias Beach Hotel, Maresias, Sao Paulo, Brazil, 24-27 Feb. 2008.

The 9th EBEB will have two special sessions, one dedicated to Professor Carlos A. de Braganca Pereira, to thank him for his many contributions in the development of the Bayesian Statistics in Brazil and, more broadly, in Latin America. Another will pay a tribute to Professor Pilar Iglesias, who passed away in March 2007 and had a close relationship with the Bayesian Brazilian community. You are also invited to visit the website to get more information about 9th EBEB and to appreciate the Atlantic Forest's beauties. We are looking forward to seeing you in Brazil. For more information, visit the website http://www.ime.usp.br/~isbra/ebeb/9ebeb/.

World Meeting of the International Society for Bayesian Analysis, Hamilton Island, Australia, 21-25 Jul. 2008.

ISBA 2008 will combine an excellent scientific program - including five keynote speakers, 90 oral presentations, three parallel sessions and two poster evenings - with an active social schedule. For additional details go to the website, http://www.isba2008.sci.qut.edu.au/, or e-mail isba08@qut.edu.au.