THE ISBA BULLETIN

Vol. 10 No. 1 *March 2003*

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE PRESIDENT

by Ed George
ISBA President
edgeorge@wharton.upenn.edu

As we begin our 11th year of existence, it gives me great pleasure to report that ISBA is strong and healthy. Our current membership, now more than 650, is growing rapidly. To all of you who have recently joined ISBA, a special welcome - we are delighted to have you as colleagues. If you haven't already done so, I highly recommend that you visit the ISBA website (http://www.bayesian.org/). There you will find comprehensive information on ISBA and what we do. I would also like to welcome the new ISBA Board members Michael Goldstein, Jun Liu, Christian Robert and Marina Vannucci, the new Vice-Chair of the Program Council Jose Bernardo, and the new President-Elect Jim Berger. And let me gratefully acknowledge former ISBA Board members Deborah Ashby, Dani Gamerman, Dalene Stangl and Mark Steel, Past-Chair of the Program Council Tony O'Hagan and Past-President Alicia Carriquiry, who are all stepping down after three years of devoted service. ISBA is a collaborative effort, and generous service such as theirs has been instrumental to our continuing success. And while I'm on the topic of generous service, I know we are all extremely grateful to Hedibert Freitas Lopes and the entire Editorial Board of the Bulletin for their Herculean efforts. As you can readily see, they have once again produced a superb top quality issue. (NB: It was only my tardiness in putting this message together that delayed the timeliness of this issue - mea culpa).

So what's coming up on the ISBA horizon? To begin with, ISBA is now more involved than ever in Bayesian meetings around the world. I am delighted to announce that planning is now underway for ISBA 2004, our 7th World meeting, which is officially set to be held at Vina Del Mar, Chile, May 23-27, 2004. This summer, our first joint IMS-ISBA Conference will be held in San Juan, Puerto Rico, June 24-26, 2003. Satellite workshops on Model Selection and on Bioinformatics and Biostatistics have been added before and after this exciting meeting. In enthusiastic support of establishing a tradition of joint IMS-ISBA meetings, both the IMS and the ISBA Executive Committees have recently approved a plan for a Second Joint IMS/ISBA International Conference to be held in the Italian Alps in Bormio, Italy, Winter, 2005. Further details will be announced soon. ISBA is also cosponsoring both the Fourth International Workshop on Objective Prior Methodology in Aussois, France, June 15-20, 2003, and the International Workshop on Bayesian Data Analysis, Santa Cruz, CA, August 7-10, 2003. Finally, ISBA has officially endorsed The 23rd annual conference on Bayesian methods and maximum entropy in science and engineering, Jackson Hole, WY, August 3-8, 2003, A Conference in Honor of Arnold Zellner: Recent Developments in the Theory, Method, and Application of Information and Entropy Econometrics, Washington. D.C., September 19-21, 2003, and Current Trends in Survey Sampling and Official Statistics, Calcutta, India, January 2-3, 2004.

Beyond meetings, there is also exciting news on the ISBA publication front. It's looking more and more like the ISBA Bulletin will soon have company, a new ISBA electronic journal. In response to last year's overwhelmingly positive vote, (see David Draper's column in the June 2002 issue of the Bulletin), Rob Kass has been hard at work putting together a final proposal for the organization and production of a truly state-of-the-art electronic journal. If all goes according to plan, this new home for innovative research on Bayesian analysis will serve as the flagship for ISBA. Stay tuned for further details.

And the regular functions of ISBA continue to run smoothly. In addition to production of our mainstay, the ISBA Bulletin, the awarding committees for the DeGroot Prize, the Lindley Prize, the Mitchell Prize and the Savage Prize, which are now all under the aegis of ISBA, have been assembled and are hard at work. The five ISBA local chapters around the world, in Australia-New Zealand, Brazil, Chile, India and South Africa, are thriving.

ISBA continues to be very successful in fostering a sense of community among people who have a fundamental interest in the development and application of Bayesian methods. Of course, this includes full-fledged Bayesians who embrace the comprehensive treatment of all uncertainty using probability But it also includes those who simply have an interest in using aspects of Bayesian methods to a more limited extent. The appreciation of the potential for Bayesian methods is growing both inside and outside the statistics community. Through our tradition of inclusiveness, ISBA is doing a wonderful job of nurturing this appreciation. The first encounter with Bayesian ideas by many people, simply entails the discovery that a particular Bayesian method superior on a particular problem or question. Nothing succeeds like success, and this observed superiority often leads to a further pursuit of Bayesian analysis. For scientists with little or no formal statistical background, Bayesian methods are being discovered as the only viable method for approaching their problems. For many of them, statistics has become synonymous with Bayesian analysis. That ISBA is growing is not surprising given the worldwide explosion of interest in Bayesian methods.

In concluding my first message to you, I hope I have conveyed my enthusiasm for the very encouraging current state of ISBA. But as ISBA as continues to move forward, I am also mindful that the evolution of ISBA must be guided in positive directions. Initiatives to increase membership, visibility and the overall influence of ISBA will be considered by the ISBA Board in the coming months. Needless to say, ISBA is a democratic organization that belongs to all of us, so please don't hesitate to send us your ideas and suggestions. I look forward to reporting to you on our progress in a future issue of the Bulletin.

Contents

➤ Interview with Adrian
☛ *Page 3*

➤ Applications
☛ *Page 5*

➤ Applications
☛ *Page 7*

➤ Annotated bibliography
☛ *Page 12*

➤ Student's corner
☛ *Page 14*

➤ Software review
☛ *Page 17*

➤ News from the world
☛ *Page 18*

ADRIAN F.M. SMITH

by Brunero Liseo
brunero.liseo@uniroma1.it

There is hardly any need to introduce Adrian Smith. He is one of the most important (Bayesian) statisticians in the world, and he has contributed in several ways to the development and to the popularity of our discipline. He is currently the Principal of the Queen Mary's College, University of London. We e-mailed Professor Smith a number of questions about his professional story and his personal view of Statistics. Here are his responses.

1. Adrian, while preparing this interview I have noticed that your name is invariably linked to at least one seminal paper on different topics: from hierarchical Bayes modelling to default Bayes factors, from MCMC to exchangeability. From your personal posterior *viewpoint what has been your most important contribution?*

Modern applied statistical work is now overwhelmingly centred around a really powerful and elegant modelling and computational synthesis that combines graphical models with Markov chain Monte Carlo simulation. In so far as my thesis work on hierarchical models with Dennis Lindley and my work on computation with Alan Gelfand triggered some of this, these may have been the most useful contributions.

2. At the time when you started your Ph.D., how popular was Bayesian statistics in the UK? Apart from Dennis Lindley, who were the main influential scientists for a young student at that time?

At the end of the 1960's it was very lonely being a Bayesian and there was a slight sense of being a persecuted minority. At Royal Statistical Society discussion meetings and international conferences, people would spend at least as much time being rude about Bayesian ideas as they did contributing to the paper under discussion. I didn't really mind this; at least in career terms you got noticed by being one of those guys with weird ideas. And, looking back, there is incredible pleasure in having - as it were - "got it right". Dennis Lindley was admirable in the way he stuck to guns. I also learnt a lot at University College London from Mervyn Stone, a very original thinker and immensely stimulating teacher and colleague. And I read a lot of Jack Good's work.

3. I know you speak Italian quite fluently. How come? Why did you decide to translate the de Finetti's book?

I'm not sure my Italian friends would agree with that! Anyway, around summer 1970 I heard someone at University College, perhaps Dennis, talking about de Finetti and how L. J. Savage thought that this man had really important things to say about Bayesian Statistics. So I subsequently bought de Finetti's two volume work on probability and - not speaking a word of Italian - began the painstaking job of working through the 700 plus pages with the aid of an Italian dictionary. This was a slow business! Then, in 1971, I moved to Oxford to my first academic job as a lecturer in the Mathematics Institute. One day, I got chatting in the tea room to a visiting Italian group theorist and discovered two things. First, he had nowhere to stay; secondly, he was from Rome and knew de Finetti. So he moved in with me and we began the joint project of properly translating de Finetti. By the time we'd finished, I was beginning to feel my way in Italian. And then the Italian mathematics community got to hear about the translation and, wrongly, assumed I must be an Italian speaker. So I began to received invitations to spend the summers in Italy - lecturing in Italian. So I had to do it. No way out.

4. How did you become interested in exchangeability? What is your present day opinion about the foundational debate (among Bayesians) between the predictivist approach and the hypothetical one, which is usually based on parametric models?

I thought the simple 0-1 form of the de Finetti representation theorem was just the most beautiful example of mathematical elegance combined with this fabulous intepretational possibility that swept up in one go the unravelling of the mysteries of how probability relates to frequency and symmetric dependence to independence, and how "parameters" can be given a meaning. I still find it quite thrilling. I seem to remember first seeing it in an experimental "advanced inference" course that Phil Dawid was giving in London. Is there a debate about the predictivist approach? I'm clearly out of touch. Anyone who doesn't get should re-read de Finetti. I still get royalties.

5. Your book with Josè Bernardo on "Bayesian theory" strictly relies on a decisional framework; do you still believe that this should be the natural scheme for the theory of statistical inference?

I'm not sure about "natural scheme", but the decision framework seems a compelling way of motivating the "discipline" that we need to impose on otherwise potentially unruly separate uncertainty judgements. And it is certainly one very satisfying way of trying to get a unified overview - viewing inference simply as a special case of decision making.

6. In the preface of your book "Bayesian theory" you wrote that it was supposed to be the first of a three volume series, the others being on Bayesian Methods and Bayesian Computation: is it still an ongoing project or Bayesian methodology evolves too fast to be able to keep track of it?

The world moved on. MCMC has become the standard way of carrying out Bayesian computation in serious problems and there is a fast moving industry of refining the methods to best deal with specific cases. The distinction between methods and computation no longer seems clear and rather than dwell on these issues abstractly, people are getting on with the business of using the powerful approaches we now have to tackle important and exciting applied problems.

7. Many statisticians consider your seminal papers with Alan Gelfand in JASA (1990) and with Gareth Roberts in JRSS-B (1993), both on MCMC methods, as fundamental steps in the revolution of Bayesian statistics over the last years. Another common feeling among Bayesians is that this revolution has had the great merit of turning many classical statisticians into Bayesians. The only bad news is that nowadays Bayesian ideas too often seem to be used sine grano salis, prior hyperparameters being chosen to tune the computations. Would you like to comment on that?

Am I concerned about people tuning hyperparameters? Two things occur to me: First, Bayes theorem - like any other - merely says that the left-hand side must equal the right-hand side; there is no chronology, so the subjectivist task could really be (re-)defined to be that of finding the prior-posterior pairing with which one feels most comfortable. Secondly, however wonderful and compelling the Bayesian formalism might be, when it comes to handling really complex problems, time, energy and intellectual fire-power constraints necessitate a degree of pragmaticism. When the problem is too big or messy to permit jesuitical adherence to Bayesian perfection, think Bayes - but with a large pinch of salt.

8. In the introduction of your latest book on "Bayesian methods for nonlinear classification and regression" you say that those methods, which were once exclusive domain of statisticians, are nowadays used by many other researchers. What is, in your opinion, the peculiar role of the statistician in this respect?

To worry about ad hoc approaches that don't (at least roughly) correspond to "Bayes with a pinch of salt".

9. Tell us something about your position at Queen Mary's.

I am what most UK universities call the *Vice-Chancellor*, most European universities the *Rector*, which is a combination into one job of what most US universities would call separately the *President* and the *Provost*. I have the ultimate responsibility for the satisfactory functioning of the institution, both academically and operationally, strategically and day-to-day. In short: non-stop Bayesian decision- making!

10. If a grad student in applied statistics asked you what are the must *courses to take in statistical and mathematical theory, what would you suggest?*

I think it is important, both for being able to access the widest possible range of research material and for personal self-confidence, to have the broadest possible range of mathematical knowledge and skills, both analytic and algebraic - and, I suspect, increasingly important to keep up with what is happening in computer science. Statistically, graphical models, stochastic simulation are currently centre stage for applications. But there is a lot of exciting stuff going on under headings like AI and the problem is that you never know what's around the corner. The most dangerous thing would be just to get absorbed in what's going on in a narrowly defined world of "Statistics".

Thanks to Adrian for his stimulating and informative answers.

SUGGESTIONS

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

`hedibert@im.ufrj.br`

Sequential Monte Carlo Methods for Bayesian Estimation

by Arnaud Doucet
ad2@eng.cam.ac.uk

1 Introduction

Numerical integration methods are necessary to perform Bayesian inference in complex models. The introduction at the beginning of the 90's of Markov chain Monte Carlo (MCMC) methods in Bayesian statistics has had a huge impact in the field and MCMC methods have now become standard tools. One thing MCMC are not good at is performing sequential Bayesian inference; i.e. estimating recursively in time a sequence of posterior distributions. Such sequential estimation problems arise in many applications such as speech processing, telecommunications or target tracking.

Consider for example a wireless communications problem. When someone calls you on your mobile phone, a sequence of symbols is sent by the transmitter through the atmosphere and arrives distorted at the receiver because of multiple paths, additive noise, etc. This problem can be set in a Bayesian framework. Given some samples of the received signal, one could estimate the posterior distribution of the transmitted symbols using MCMC. Unfortunately, this would be completely unrealistic in a telecommunications context; you do not want to wait for an MCMC sampler to converge so as to be able to listen to your correspondent; we are interested in estimating the symbols in real-time and to be able to update the posterior distributions of the symbols as data become available.

In the mid 90's a new class of Sequential Monte Carlo (SMC) methods was introduced in statistics and engineering to address these problems. The objective of this brief article is to review this emerging field.

2 Sequential Monte Carlo Methods

2.1 A Bit of Methodology

For any sequence $\{Z_k\}$, we denote $Z_{i:j} \triangleq (Z_i, Z_{i+1}, \ldots, Z_j)$. SMC methods are a set of Monte Carlo methods generating recursively in time a large collection of N ($N >> 1$) weighted random samples (called particles) $\left\{W_n^{(i)}, X_{0:n}^{(i)}; i = 1, \ldots, N\right\}$ where $W_n^{(i)} > 0$, $\sum_{i=1}^N W_n^{(i)} = 1$ such that

$$\sum_{n=1}^{N} W_n^{(i)} f\left(X_{0:n}^{(i)}\right) \underset{N\to\infty}{\to} \int f(x_{0:n})\, \pi_n(x_{0:n})\, dx_{0:n}$$

where $\{\pi_n\}_{n\geq 0}$ is the sequence of probability distributions of interest; each distribution is only known up to a normalizing constant. The basic elements of SMC are sequential importance sampling and resampling.

Assume at time[1] $n-1$ the weighted samples $\left\{W_{n-1}^{(i)}, X_{0:n-1}^{(i)}; i = 1, \ldots, N\right\}$ approximating π_{n-1} are available. At time n, one extends each path $X_{0:n-1}^{(i)}$ by sampling $X_n^{(i)}$ according to an importance distribution $q_n\left(\cdot \middle| X_{0:n-1}^{(i)}\right)$. To correct for the discrepancy between the new target distribution π_n and the importance distribution, a simple importance sampling argument shows that it is necessary to update the weights according to

$$W_n^{(i)} \propto \frac{\pi_n\left(X_{0:n}^{(i)}\right)}{\pi_{n-1}\left(X_{0:n-1}^{(i)}\right) q_n\left(X_n^{(i)} \middle| X_{0:n-1}^{(i)}\right)} W_{n-1}^{(i)}.$$

This method is the *Sequential Importance Sampling* (SIS) method. Essentially, it consists of doing importance sampling using at time n the importance distribution

$$q_0(x_0) \prod_{k=1}^{n} q_k\left(x_k \middle| x_{0:k-1}\right).$$

Typically, in most applications, the computational complexity required to compute $W_n^{(i)}$ given $W_{n-1}^{(i)}$ is independent of n and the method is truly recursive. The main problem with this method though is that it is just a special instance of importance sampling and does not work if n is large!

The key idea of SMC is the *Resampling step*. Assume at time n, a collection of particles $\left\{W_n^{(i)}, X_{0:n}^{(i)}; i = 1, \ldots, N\right\}$ approximating π_{n-1} is available. If the variance of the weights is too high, particles with small weights are killed and particles with high weights are copied multiple times. The underlying idea is to focus the computational efforts on the promising zones of the space. Finally one assigns equal weights to each copy. The resampling step is what makes SMC work.

[1] This variable is simply a counter and need not have any relation with "real time".

Clearly it introduces errors at each time n but it can be shown both practically and theoretically that this ensures that the algorithm does not "degenerate" over time. More formally, it consists of performing the following approximation

$$\sum_{i=1}^{N} W_n^{(i)} \delta_{X_{0:n}^{(i)}} (dx_{0:n}) \approx \sum_{i=1}^{N} \frac{N_n^{(i)}}{N} \delta_{X_{0:n}^{(i)}} (dx_{0:n})$$

where $N_n^{(i)} \in \mathbb{N}$ is the number of times the particles $X_{0:n}^{(i)}$ is copied and $\sum_{i=1}^{N} N_n^{(i)} = N$ so as to keep the size of the population constant. To minimize the error introduced by the resampling scheme, one usually selects a stochastic mechanism to obtain $\left\{ N_n^{(i)}; i = 1, \ldots, N \right\}$ such that $E\left[N_n^{(i)}\right] = NW_n^{(i)}$ (unbiased approximation) and one wants small variances $var\left[N_n^{(i)}\right]$. Several resampling schemes have been proposed in the literature including multinomial, residual and stratified resampling.

Note that one obtains an estimate of the joint distribution $\pi_n (x_{0:n})$ at index n. However, one can only expect to obtain a "good" approximation of the most "recent" marginal distributions $\pi_n (x_{k:n})$ for $n - k$ say below 10. This is because if particles are resampled many times between time k and n there are very few distinct paths $X_{0:k}^{(i)}$ at index n. Fortunately, this is all that is necessary in many applications.

We have presented here a simple generic SMC method. However like MCMC methods, SMC methods are not a black box and it is necessary to design carefully the algorithm so as to obtain good performance for a reasonable number of particles. Recently many papers have proposed improved SMC methods to improve this basic scheme: construction of efficient importance sampling distributions, Rao-Blackwellised estimates, use of MCMC moves etc; see (Doucet, De Freitas and Gordon; 2001).

2.2 Application to Sequential Bayesian Inference

Consider the following problem. One is interested in estimating the state of a Markov process $\{X_k\}_{k\geq 0}$ given some observations $\{Y_k\}_{k\geq 1}$. The unobserved (hidden) Markov process is defined by

$$X_0 \sim \mu, \quad X_k | X_{k-1} \sim f\left(\cdot | X_{k-1}\right)$$

whereas the observations are assumed independent conditional upon $\{X_k\}_{k\geq 0}$ having marginal distribution

$$Y_k | X_k \sim g\left(\cdot | X_k\right).$$

Estimating the posterior distribution of X_k given $Y_{1:k}$ is a very important problem known as optimal filtering. When the model is linear and Gaussian, the posterior distribution is Gaussian and its statistics can be computed using the Kalman filter (West & Harrison, 1997). However in many real-world applications, these linearity and Gaussianity assumptions are not valid and one needs to use numerical methods. SMC methods can be applied directly to this problem by setting π_n as the posterior distribution of the collection of states $X_{0:n}$ given the observations $Y_{1:n}$. Indeed this posterior distribution satisfies

$$\pi\left(x_{0:n}\right) \propto \mu\left(x_0\right) \prod_{k=1}^{n} f\left(x_k | x_{k-1}\right) g\left(y_k | x_k\right)$$

and is typically known up to a normalizing constant.

2.3 A Bit of History and a Few References

As with MCMC, SMC were first developed in a physics context and then rediscovered and improved recently by statisticians and engineers.

The SIS method can be attributed to Hammersley and Morton (1954) who introduced it so as to simulate self-avoiding random walks for modelling polymers. In the context of sequential Bayesian inference, it was rediscovered in engineering by Handschin and Mayne (1969).

The key resampling step has been first introduced in physics by Hetherington (1984) where a multinomial resampling scheme is explicitly used. It was rediscovered in the sequential Bayesian inference context in (Gordon, Salmond & Smith; 1993) and it is now widely acknowledged that this paper is the seminal article on SMC (for non-physicists!). Kitagawa (1993) published essentially the same algorithm at the same time.

Later Pitt & Shephard (1999) proposed a variant of the SMC algorithm which is one of the most popular methods in current use. Doucet, Godsill & Andrieu (2000) and Liu & Chen (1998) are two survey papers; the first one emphasizes the sequential Bayesian inference context whereas the second is much more general. Many theoretical results are available for SMC methods; see (Del Moral & Miclo; 2000). An up-to-date booklength survey of the literature from the statistical/engineering viewpoint is the volume edited by Doucet, De Freitas and Gordon (2001) whereas Iba (2001) is another recent paper describing the connections between the SMC algorithms used in physics and statistics.

Finally, many papers on the subject can be downloaded on the SMC preprint service maintained by Elena Punskaya and Nando De Freitas:
`http://www-sigproc.eng.cam.ac.uk/smc/`

References

[1] Del Moral, P. and Miclo, L. (2000) Branching and Interacting Particle Systems Approximations of Feynman-Kac Formulae with Applications to Non-linear Filtering. In *Séminaire de Probabilités XXXIV*, Ed.Azéma, J., Emery, M., Ledoux, M. and Yor, M., *Lecture Notes in Mathematics*, Berlin: Springer-Verlag, **1729**, 1-145.

[2] Doucet A., Godsill S.J. and Andrieu C. (2000) On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. *Statist. Comput.*, **10**, 197-208, 2000.

[3] Doucet, A., De Freitas, J.F.G. and Gordon, N.J. (editors) (2001) *Sequential Monte Carlo Methods in Practice*. Springer Series in Statistics for Engineering and Information Science. New York: Springer-Verlag.

[4] Hammersley J.M. and Morton K.W. (1954) Poor Man's Monte Carlo. *Journal of the Royal Statistical Society* B, 16, pp. 23-38.

[5] Handschin J.E. and Mayne D.Q. (1969) Monte Carlo Techniques to Estimate the Conditional Expectation in Multi-stage Non-linear Filtering. *International Journal of Control*, 9, 547-559.

[6] Hetherington, J.H. (1984) Observations on the Statistical Iteration of Matrices. *Physical Review A*, 30, pp. 2713-2719.

[7] Kitagawa G. (1993). A Monte Carlo Filtering and Smoothing Method for Non-Gaussian Nonlinear State Space Models. in *Proc. 2^{nd} US-Japan Joint Seminar on Statistical Time Series Analysis*, Honolulu, Hawaii, 110-131.

[8] Iba Y. (2000) Population Monte Carlo Algorithms. *Transactions of the Japanese Society for Artificial Intelligence*, 16, 279-286.

[9] Liu J.S. and Chen R. (1998) Sequential Monte Carlo Methods for Dynamic Systems. *Journal of the American Statistical Association*, 93, 1032-1044.

[10] Pitt M.K. and Shephard N. (1999) Filtering via Simulation: Auxiliary Particle Filters. *Journal of the American Statistical Association*, 94, 590-599.

[11] West M. and Harrison P.J. (1997) *Bayesian Forecasting and Dynamic Models* (2nd Edition). New York: Springer-Verlag.

Bayesian Research at the NASA Ames Research Center, Computational Sciences Division

by Robin D. Morris
`rdm@email.arc.nasa.gov`

[I am writing this in the week following the break-up of the space shuttle Columbia on re-entry. Our thoughts are with the families of Rick Husband, Michael Anderson, Laurel Clark, David Brown, William McCool, Kalpana Chawla and Ilan Ramon.]

NASA Ames Research Center is one of NASA's oldest centers, having started out as part of the National Advisory Committee on Aeronautics, (NACA). The site, about 40 miles south of San Francisco, still houses many wind tunnels and other aviation related departments. In recent years, with the growing realization that space exploration is heavily dependent on computing and data analysis, its focus has turned more towards Information Technology. The Computational Sciences Division has expanded rapidly as a result. In this article, I will give a brief overview of some of the past and present projects with a Bayesian content. Much more than is described here goes on with the Division. The web pages at `http://ic.arc.nasa.gov` give more information on these, and the other Division projects.

AUTOCLASS: Bayesian research at Ames began in 1985. The first major project, lead by Peter Cheeseman, was AUTOCLASS, a system for performing unsupervised classification of data, where the number and description of the natural classes of the data is not known. AUTOCLASS handles missing data, mixed real and discrete attributes, and estimates the posterior probability over a range of model structures. It is one of the earliest examples of a restricted class of Bayes Net system. AUTOCLASS has proved extremely useful in practice, and has found subtly different classes that were unknown to the investigators, as well as many previously known classes (but unknown to AutoClass). AUTOCLASS is publicly available.

IND: Another early project, lead by Wray Buntine, was the IND system, which was concerned with Bayesian software for supervised classification using decision trees. A tree is ”grown” from data using a recursive partitioning algorithm to create a tree which (hopefully) has good prediction of classes on new data. As well as reimplementing parts of some of the standard Decision Tree algorithms (e.g. C4) and offering experimental control suites, IND also introduced Bayesian and MML methods and more sophisticated search in growing trees. These produce more accurate class probability estimates that are important in applications like diagnosis

The approach used in IND has subsequently been adapted to learning Bayesian networks from data, to learning n-grams for language modeling, and to a classification model known as Alternating Decision Trees . The data structures and algorithms have been quiet influential. Moreover, rumor has it that Breiman, an influential Bayesian antagonist, was motived by INDs apparent successes to develop the Bagging approach to classification trees that subsequently became the empirical champion in the field.

IND has seen widespread use in empirical and applied studies, and is publicly available.

AUTOBAYES: An ongoing project of general applicability in Bayesian analysis is the AutoBayes project.

AutoBayes is an automatic program synthesis system for the machine learning domain under development by the Automated Software Engineering group since 1999. From the outside, AutoBayes is essentially a compiler for a modeling language similar to the BUGS language; inside, however, it employs sophisticated code generation methods and is one of the more complex synthesis systems produced by the Automated Software Engineering community. AutoBayes takes as its input a statistical model, extracts a Bayesian network from it, and then generates a program which solves the learning task specified in the model. Unlike BUGS, however, AutoBayes is not restricted to a single generic algorithm (i.e., Gibbs sampling) but can generate different algorithms which are specialized for the model.

AutoBayes contains a comprehensive schema library. A schema contains two parts, a definition of when it is applicable, and a code template. During synthesis, AutoBayes finds the schemas that are applicable, then instantiates a code fragment in a model-specific way (e.g. an EM schema is instantiated if a sub-problem is recognised as a finite mixture model). These code fragments can spawn new, simpler, synthesis tasks, which are solved recursively; the recursion terminates if subproblems can be solved either numerically or symbolically. An important aspect here is the interaction of the schemas with the symbolic subsystem (i.e., a simple Mathematica-like symbolic-algebraic kernel) which allows the identification and efficient solution of tractable subproblems, even if they are embedded in the original model.

This divide-and-conquer approach allows AutoBayes to synthesize larger programs in a bottom-up fashion, using both schemas and symbolic solutions as building blocks. After synthesis, the code is optimized and translated into a C/C++ program which can be run standalone or linked dynamically into the Matlab or Octave environments.

AutoBayes has been used to generate code for a spectrum of models, ranging from textbook examples (e.g., normal models with various priors) to ”almost state-of-the-art” machine learning algorithms; it has also been applied successfully to some data analysis problems within NASA.

Autonomy

A major research area in the Computational Sciences Division is to provide autonomous capabilities for spacecraft and rovers - communication bandwidth to space exploration vehicles and onboard storage are limited; the spacecraft collect vastly more data than can be returned, and cannot be controlled in real time[2]. The need for autonomy, both for spacecraft operations and scientific discovery, is obvious. One great success in this area was the *Remote Agent* system on the Deep Space One spacecraft. This was a “traditional” AI system, based around planning and scheduling, modeling the state of the spacecraft, and a smart executive module. Currently, research is underway to address some of the limitations of that system, and many of the approaches being pursued are Bayesian.

Diagnosis: Diagnosis is the problem of detecting and identifying any faults or unexpected events that occur in a system from observations of that system. Bayesian belief updating methods are being applied to this problem, maintaining a belief distribution over the state of the system, and updating the distribution based on a model of the evolution of the system and on new observations as they arrive.

[2]Communication with spacecraft on Mars typically occurs twice per day.

The models used are probabilistic hybrid automata – they contain a mixture of discrete states and continuous variables. The evolution of the system is governed by a transition function which gives the probability of a transition from one discrete state to another, and a set of differential equations which model the behavior of the continuous variables, and are dependent on the discrete state.

Optimal approaches to this problem are computationally infeasible, particularly on-board a spacecraft or planetary rover[3] . Particle filters can be used to track the state in reasonable computation time. However, diagnosis problems present some interesting challenges for particle filter algorithms, particularly because the fault states have very low probability of occurring. Several variants of particle filters have been developed, tuned to solving diagnosis problems.

Scheduling: New research is applying Bayesian techniques to scheduling problems. The domain here is one in which a number of tasks must be scheduled, given a set of constraints on when the tasks must be performed. Completing certain tasks, or subsets of the tasks results in a numerical reward. There is uncertainty about the duration of the individual tasks so the problem becomes one of building the schedule that maximizing the expected reward obtainable.

Autonomous Exploration: This project investigated the application of Bayesian statistics to the problem of autonomous geological exploration with a robotic vehicle. It concentrated on the subproblem of classifying rock types while addressing the issues associated with operating onboard a mobile robot. The Bayesian paradigm was used in a natural way to solve the more general robotic problems of autonomously profiling an area and allocating scarce sensor resources. Major considerations are the need to use of multiple sensors and the ability of a robotic vehicle to acquire data from different locations. Needless sensor use must be curtailed if possible, such as when an object is sufficiently well identified given sensor data acquired so far. Furthermore, by investigating rocks in many locations, the robot has the opportunity to profile the environment. Different rock samples are statistically dependent on each other. These dependencies can be exploited to substantially improve classification accuracy.

The classification system was been implemented onboard the Nomad robot developed at Carnegie Mellon University, and applied to the task of recognizing meteorites amongst terrestrial rocks in Antarctica. In January 2000 A.D., Nomad was deployed to Antarctica where it made the first autonomous robotic identification of a meteorite.

Data Analysis

NASA has been described as a *data collection agency* – each mission returns huge quantities of data, and Earth observing satellites return data at such a rate that it is difficult to archive, let alone analyze. Naturally, therefore, there are a number of data analysis projects within the Division.

Planetary Nebula Modeling: Stars like our sun end their lives as swollen red giants surrounded by cool extended atmospheres. The nuclear reactions in their cores create carbon, nitrogen and oxygen, which are transported by convection to the outer envelope of the stellar atmosphere. As the star finally collapses to become a white dwarf, this envelope is expelled from the star to form a planetary nebula (PN) rich in organic molecules. The physics, dynamics, and chemistry of these nebulae are poorly understood and have implications not only for our understanding of the stellar life cycle but also for organic astrochemistry and the creation of prebiotic molecules in interstellar space.

This project is working toward generating three-dimensional models of planetary nebulae, which include the size, orientation, shape, expansion rate and mass distribution of the nebula, as well as the distance from earth. Such a reconstruction of a PN is a challenging problem for several reasons. First, the data consist of images obtained over time from the Hubble Space Telescope and long-slit spectra obtained from Kitt Peak National Observatory and Cerro Tololo Inter-American Observatory. These images are of course taken from a single viewpoint in space, which amounts to a very challenging tomographic reconstruction. Second, that there are two disparate data types requires that we utilize a method that allows these data to be used together to obtain a solution. Bayesian model estimation is applied using a parameterized physical model that incorporates much prior information about the known physics of the PN. By modeling the nebula in three-dimensions it is possible reconcile the observed tangential expansion observed as an angular size change of the object with the radial expansion velocity determined from the Doppler shift in the spectral lines thus providing accurate estimates of the objects expansion velocity, dynamical age, and distance from earth.

[3] The computational capacity of the Mars rovers scheduled for launch later this year is equivalent to a 25MHz PowerPC

Event analysis for GLAST: The Gamma Ray Large Area Space Telescope is a project to map the incidence of gamma rays from the entire sky. It is an orbiting telescope, scheduled for launch in 2006. It works by converting an incident gamma ray into an electron-positron pair in one of a stack of tungsten layers, and then detecting the positions where these charged particles cross layers of silicon microstrip detectors. However, the analysis is complicated by numerous secondary processes – the electron and positron are scattered each time they traverse the layers, and can also knock out further electrons, which cause the microstrips to fire as they cross them. We are studying the feasibility of using a detailed model of the physics of the detector to define importance sampling distributions to enable a particle filter type approach to be used to estimate, for each event, the direction from which the gamma ray came, and its energy.

Analysis of hyper-spectral solar flux data: This effort aims at developing a Bayesian framework for analyzing hyper-spectral data on solar radiation in the atmosphere, collected with a custom-built NASA radiometer in various field campaigns around the world. This framework is expected to allow efficient and accurate determination, from heterogeneous data, of the chemical composition and the physical state of the atmosphere, thus significantly enhancing our understanding of, as well as our capability to model and predict, the Earth system. Specific goals include: retrieval of cloud physical parameters for understanding their evolution and for assessing their impact on weather and global climate; identification of composition, size, shape, and distribution of aerosols for evaluating their effects on solar radiation budget; quantification of the influence of tropospheric ozone and carbon-based trace gases on radiative forcing.

The main thrust of present research is toward developing forward physical models - one for the atmosphere and one for the instrument - suitable for use as likelihood functions within a Bayesian parameter estimation scheme.

Computer Vision: The low-level vision problem is conceived as the construction of a 3-D surface model of the local world, where the model is represented as a triangulated mesh with reflectance parameters associated with each triangle. In addition to inferring the 3-D mesh, the lighting and camera parameters must also be inferred. This is an extremely hard inference problem, because an observed image depends on the 3-D model geometry and reflectance as well as the camera and lighting parameters. The likelihood function is essentially the computer graphics problem: given all the model information, what would the image look like. Bayes theorem inverts this function and allows the 3-D model to be inferred given the images. The forward model is well understood; the physics of light scattering and camera optics is well known.

The simplified 2-D problem was first investigated. If the camera and lighting are essentially constant, there is no significant parallax, and it is impossible to separate out the effects of surface geometry from surface reflectance. Instead, model "images" at super-resolution are reconstructed from multiple images of the same area. Super-resolution is possible because each image is an independent sample of the unknown surface. This program yielded spectacular improvements in resolution, enabling features to be seen that were completely invisible in the individual images. Currently, this research is being extended to full 3-D surface reconstruction from images with different camera views and lighting. Although straight forward in principle, it is extremely difficult in practice because there are typically millions of model parameters to be inferred, and because simultaneous estimation of camera, lighting and 3-D surface model from images is greatly complicated by their mutual dependence. For this simultaneous inference to work, it is necessary to use techniques such as feature matching between images to "bootstrap" the inference procedure. Once such heuristic information initializes the joint model search sufficiently close to the global maximum, standard gradient methods to find the MAP estimate (and its associated covariance matrix) seem to work well in practice. This is ongoing research that should give a full Bayesian foundation to the problem of low-level computer vision.

Separation of Neural Signals: The electric potentials and magnetic fields generated by ensembles of synchronously active neurons in response to external stimuli provide information essential to understanding the processes underlying cognitive and sensorimotor activity. Interpreting the recordings of these potentials and fields can be problematic as each detector records signals that have been simultaneously generated by various regions throughout the brain. Separating these signals into a set of components each originating from a synchronous ensemble has proven to be a very difficult problem.

The differential variability component analysis (dVCA) algorithm relies on a more physiologically realistic source model that accounts for variability of response amplitude and latency across multiple experimental trials.

Rather than making any unrealistic assumptions of independence of components, this algorithm utilizes the differential variability of the evoked waveforms to aid their characterization. By applying the Bayesian methodology to this new source model, we derive an algorithm that uses EEG data simultaneously recorded from multiple electrodes to identify multiple components each representing synchronous neuronal activity from an ensemble of neurons displaying a distinct trial-to-trial variability pattern. In addition, this algorithm estimates the single-trial amplitude and latency of each component active during any particular evoked response.

Analysis of Earth Observing Data: A couple of projects involved in the analysis of Earth Observing Data are the following.

One project is looking at using naive Bayes classifiers applied to MODIS (Moderate Resolution Imaging Spectroradiometer) data for generating a cloud mask product. The current methods of generating the cloud mask products from MODIS data at the DAACs (Distributed Active Archive Centers) are too slow to allow for the product to be included in the broadcast stream, and so are not used in other data products, limiting their accuracy. The goal is to use naive Bayes to produce a quick product which could be sent out along with the data.

A second project is looking at the uncertainty present in the data products themselves, many of which are derived from the raw satellite observations. The derivation of these data products from the observations and other data is often via some empirically determined relationships (e.g. the production of Leaf Area Index maps from Normalised Difference Vegetation Index maps). The Earth Science community then uses these derived quantities with little appreciation of the range of uncertainty present, and the effect of that uncertainty on predictions made using these derived data products. In this project we are analyzing the relationships used to generate certain data products, with a view to quantifying the uncertainty, and making it available together with the data product.

Novel Interfaces: This builds on work done on Monte-Carlo methods for mixture modeling. In particular a Bayesian approach to the parameterization of Gaussian mixture models, looking at the case where the distributions change over time. This work will be applied to "virtual keyboards", where electrical signals from the muscles in the users forearm are captured by dry electrodes on the skin, and decoded to recognize the gestures associated with pressing a particular key. It will also used to enable a "virtual joystick", used to fly a high-fidelity aircraft simulator. These models are being developed to augment HMM-based models to improve performance.

Contacts

- AUTOCLASS - Peter Cheeseman, cheesem@email.arc.nasa.gov
- IND - Will Taylor, taylor@email.arc.nasa.gov
- AutoBayes - Bernd Fischer, fisch@email.arc.nasa.gov
- Diagnosis - Richard Dearden, dearden@email.arc.nasa.gov
- Scheduling - Jeremy Frank, frank@email.arc.nasa.gov
- Autonomous Exploration - Liam Pedersen, pedersen@email.arc.nasa.gov
- Planetary Nebula - Kevin Knuth, kevin.h.knuth@nasa.gov
- GLAST - Robin Morris, rdm@email.arc.nasa.gov
- Solar Flux - Dogan Timucin, dogan.a.timucin@nasa.gov
- Vision - Peter Cheeseman
- Neural Signals - Kevin Knuth
- EOS Data - Kevin Wheeler, kevin.r.wheeler@nasa.gov and Robin Morris
- Novel Interfaces - Kevin Wheeler

MODEL AVERAGING

by Keying Ye
keying@vt.edu

Model averaging methods, especially Bayesian Model Averaging (BMA) have been used in many applications. The purpose of model averaging in data analysis is to incorporating model uncertainty. Some of the methods and applications are considered in this note.

Bayesian model avering in general statistics

1. Brown, P.J., Vannucci, M., and Fearn, T. (2002) Bayes model averaging with selection of regressors, JRSS B 64, 519-536. In the context of the multivariate general linear model, decision theory is applied in incorporating variable selection for prediction using Bayesian model averaging.

2. Clyde, M. (1999) Bayesian Model Averaging and Model Search Strategies (with discussion), in Bayesian Statistics 6. J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith eds., 157-185. Oxford University Press. Approximations to the posterior model probabilities are introduced to develop efficient methods for deterministic or stochastic sampling from high dimensional model spaces in the use of Bayesian model averaging method.

3. Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion), Journal of the Royal Statistical Society Series B, 57, 45-97. Describing the danger of not incorporating model uncertainty in regression analysis, the author used examples to discuss the Bayesian model averaging in both discrete and continuous situations.

4. Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999) Bayesian model averaging: A tutorial, Statistical Science 14, 382-401. This is an important review paper of BMA in statistics. It covers most of the recent statistical methods and applications using BMA up to date.

5. Leamer, E. E. (1978). Specification Searches, Wiley, New York. In this book, the formula of BMA for the posterior mixture distribution was stated.

6. Liang, F., Truong, Y.K., Wong, W.H. (2001). Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting, Statistica Sinica, 11, 1005-1030. The numerical results show that the Bayesian model averaging procedure resulted from the automatic prior setting, in which there is no parameter to be specified by users, provides a significant improvement in predictive performance over other two procedures proposed in the literature. The procedure is extended to the problem of Bayesian curve fitting with regression splines.

7. Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression, JASA, 83, 1023-1036. Variable activation probability used in Bayesian variable assessment is mentioned in the paper, although primarily the authors dealt with variable selection.

8. Raftery, A.E.; Madigan, D. and Hoeting, J.A. (1997). Bayesian Model Averaging for Linear Regression Models, JASA 92, 179-191. In this article, the authors use Occam's window and MCMC approach that directly approximates the exact solution to deal with computational problems in BMA applications.

9. Wasserman, L (2000). Bayesian model selection and model averaging, J Math Psychology, 44, 92-107. In this paper the author reviewed the Bayesian model averaging with emphasis on objective Bayesian methods based on noninformative priors.

BMA Applications in Economics and Business

10. Bunnin, F.O., Guo, Y. and Ren, Y. (2000). Option pricing under Model and Parameter Uncertainty using Predictive Densities, Statistics and Computing, 12, 37-44. A European Call option pricing on a share index is used to apply Bayesian model averaging by constructing a model's predictive density by integrating its transition density function using posterior distributions.

11. Chua, C. L., Griffiths, W.E. and O Donnell, C.J. (2001). Bayesian model averaging in consumer demand systems with inequality constraints, Canadian Journal of Agricultural Economics, 49, 269-292. Bayesian model averaging is used to choose between the results from the two alternative functional forms in an application to USDA data for beef, pork and poultry.

12. Fernandez, C., Ley, E. and Steel, M.F. (2001). Benchmark priors for Bayesian model averaging, Journal of Econometrics, 100, 381-427. Partially noninformative prior structure related to a natural conjugate g-prior specification is proposed to economics models using Bayesian model averaging.

13. Fernandez, C., Ley, E. and Steel, M.F. (2001). Model Uncertainty in Cross-Country Growth Regressions, Journal of Applied Econometrics, 16, 563-576. Bayesian model averaging is used to investigate the issue of model uncertainty in cross-country growth regressions.

14. Murphy, M.; Wang, D. (2001). Do previous birth interval and maternal education influence infant survival? A Bayesian model averaging analysis of Chinese data, Population Studies, 55, 37-48. A Bayesian model averaging strategy that takes account of model uncertainty as well as parameter uncertainty is used to study the effect of socio-economic covariates on infant mortality in China in the 1980s.

BMA Applications in bioinformatics, biostatistics and engineering

15. Chickering, D.M., Heckerman, D. (2000). A comparison of scientific and engineering criteria for Bayesian model selection, Stat Comput 10, 55-62. For Bayesian-network models containing discrete variables only, the predictive performance of the model average can be significantly better than those of single models selected by either posterior-probability and engineering criteria. Furthermore, differences between models selected by the two criteria can be substantial.

16. Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles, Bioinformatics, 18, 1194-1206. A clustering procedure based on the Bayesian infinite mixture model is developed using Bayesian model averaging and it is applied to clustering gene expression profiles.

17. Meyer, R.D. and Box, G. (1992). Finding the active factors in fractionated screening experiments, Technical Report 80, Center for Quality and Productivity Improvement, University of Wisconsin. Although a formal Bayesian model averaging is not developed, the idea of model averaging is actually used in computing the variable activation probability.

18. Millis, S.R. and Volinsky, C.T. (2001). Assessment of Response Bias in Mild Head Injury: Beyond Malingering Tests, Journal of Clinical and Experimental Neuropsychology, 23, 809-828. Bayesian model averaging as a statistical technique to derive optimal prediction models is performed in assessment of response in mild head injury with a clinical data set.

19. Seaman, S.R., Richardson, S., Stucker, I., and Benhamou, S. (2002). A Bayesian partition model for case-control studies on highly polymorphic candidate genes, Genetic Epidemiology 22, 356-368. BMA is used to Bayesian partition model clustering genotypes according to risk.

20. Viallefont, V.; Raftery, A. E.; Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies, Statistics in Medicine, 20, 3215-3230. Bayesian model averaging method is shown to better account for model uncertainty than classic variable selection method in case-control studies.

21. Volinsky, C. T.; Madigan, D.; Kronmal, R. A. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke, Applied statistics, 46, 433-448. Bayesian model averaging method is used to the selection of variables in Cox proportional hazard models in the context of a cardiovascular health study.

22. Wan, Y, Nowak, RD (2000). A new Bayesian model averaging framework for wavelet-based signal processing, IEEE Inter. Conf. on Acoustics, Speech, and Signal processing, Proc. I-VI : 476-479. A new signal modeling framework using Bayesian model averaging and the redundant or translation-invariant wavelet transform is developed.

BMA Applications in Environmental and Ecological Sciences

23. Morales, K.H., Ibrahim, J.G., Ryan, L.M., Chen, C.J. (2001). Bayesian model averaging with applications to the risk assessment for arsenic in drinking water, Arsenic Exposure and Health Effects IV, 145-151. Importance of accounting for model uncertainty in the risk assessment process for study arsenic in drinking water is studied using BMA. Data fitting is shown equally well using both BMA and classic approach. Yet, risk estimates are shown different.

24. Lipkovich, I.A. (2002). Bayesian Model Averaging and Variable Selection in Multivariate Ecological Models, Ph.D. dissertation, Virginia Tech. Correspondence Analysis, Canonical Correspondence Analysis and Redundancy Analysis for Multivariate Ecological Models are discussed using Bayesian model averaging.

25. Noble, R. (2000). Multivariate Applications of Bayesian Model Averaging, Ph.D. dissertation, Virginia Tech. In this dissertation, the author studied multivariate analyses of Principal Components, Correlation Analysis and Canonical Correlation Analysis in environmental and ecological sciences using Bayesian model averaging.

Frequentist views in model averaging

Although the model averaging formula is clearly a Bayes' formula, the way to calculate model probabilities can be done using frequentist methods. The followings are several of those papers.

26. Breiman, L. (1996). Stacked regression, Machine Learning, 24, 49-64. Using stepwise regression and cross-validation method to find weights of the models which minimizing the validation sum of squares.

27. Breiman, L. (1996). Bagging predictors, Machine Learning, 24, 123-140. The author implements computing intensive re-sampling algorithm to find the weights of models using bootstrapping and forward selection.

28. Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997). Model selection: an integral part of inference, Biometrics, 53, 275-290. Bootstrapping the data set and determining the best model in each bootstrap sample can be used to evaluate model weights and can be used to apply model averaging idea.

29. Freedman, D.A., Navidi, W. amd Peters, S.C. (1986). On the impact of variable selection in fitting regression equations. In On model uncertainty and its statistical implications: proceedings of a workshop held in Groningen, Dijkstra, T.K. ed.. Due to the uncertainty on how to bootstrap data set to evaluate model weights, the authors proposed to use the full model as a basis for bootstrapping residuals.

Academy and Industry

by Lilla Di Scala and Luca La Rocca
lilla@dimat.unipv.it luca@dimat.unipv.it

The highlight of this STUDENT'S CORNER is again an extensive interview, which we believe will be of great interest to all students, featuring a person who has had the opportunity of seeing the best (and worst) of both worlds: Academia and Industry. We are talking about Mauro Gasparini, Full Professor at the Department of Mathematics of the Turin Polytechnic. The talk centers on his experience within Novartis, one of the biggest pharmaceutical companies in the world. Professor Gasparini obtained his degree in Statistics at the University of Bologna, Italy, and then went on, as he himself explains, to obtain a Ph.D at the University of Michigan. Later, after a post as Assistant Professor at Purdue University, he returned to Europe to work at Novartis Pharma in Basel on Clinical Trials' issues, within their Pharmacokinetics and Pharmacodynamics group. Now he is in Turin, where he does applied research on biostatistics and lots more. Before giving him the floor, we would like to emphasize a remark which came out during our talk: that is, never to look down on applied work. In fact, although it may sometimes become routine-work, it often provides challenging mathematical problems: one should not forget that the real world is the main driving force of Statistics.

1. Dear Professor Scalia Tomba, first of all thank you for being here with us. As an appetizer, can you tell us about your experience in Sweden before and after having obtained your Ph.D.? Sweden was, and still is, a good place to be a student. While I was studying for my Ph.D., I was able to do a lot of teaching assistance and statistical consultancy work; after I got my Ph.D., I was able to go on doing more or less the same things, but with a better salary.

2. Dear Professor Gasparini, thank you for having accepted to talk to the Bulletin. First of all, can you tell us about your experience in the States before and after having obtained your Ph.D.? I landed at the University of Michigan in 1987 after having obtained a Fulbright scholarship. A few months earlier I did not even know where Michigan was, I had to check on the Atlas. The Fulbright organization provided me with the right links to Michigan and other schools, but Michigan was the only one offering me a teaching assistantship. So I went to Michigan, I did not have any surplus money to spend in higher education.

The Ph.D experience turned out to be very enjoyable and enriching. I met important teachers, like Michael Woodroofe, who left a mark in all my scientific career. After my Ph.D., I went to Purdue as an Assistant Professor. There again I learned a lot from my colleagues, notably James Berger, and started my research in nonparametric Bayes as well as some applied projects.

3. After how many years in academia did you decide to move into the private sector and why? Once again, the decision was more of a necessity than a choice. The Fulbright scholarship visa came with a "two-year residency requirement", that is, I had to leave the US, sooner or later, for two years before applying for a green card (permanent residency in the US). So, I left the States after three years as an Assistant Professor at Purdue. A good friend of mine, Jeffrey Eisele, an American himself, was working in Novartis Switzerland. He told me about an opening position, I applied and got the job. Not that I particularly wanted it, but things turned out in the best way for me and resulted in a very refreshing change in my career.

4. What was your first impression of the private sector? Was anything particularly surprising to you? The first impression was definitely better than anything I expected. There are a lot of unjustified prejudices about the private sector and in particular about pharmaceutical companies. I was positively surprised to see that Amy Racine, the person who made the final decision about hiring me and became my boss, was doing very sophisticated Statistics within a private company. Another positive surprise was that the job of (SAS) programmer was quite separate from the statistician's. The latter solely analysed data already preprocessed by the programmer into the required database. A common joke in the pharma industry is that a SAS programmer does PROC MEANS, a statistician does PROC GLM and a senior statistician does PROC MIXED.

5. With regards to your experience in the private sector, where have you worked precisely, for how long and with which assignments? I have worked in the Pharmacokinetics and Pharmacodynamics (PK/PD) group of the company Ciba-Geigy, which then became Novartis after a merger with Sandoz. My assignments were to participate in writing clinical development plans, write the statistical part of clinical trial protocols and analyse data from clinical trials. The data I worked with were mainly PK/PD, for example data from phase III supporting and expanding knowledge about how a drug is absorbed, distributed and excreted from the human body. But also response variables from Phase I, II and IV clinical trials, for example dose finding trials, equivalence trials, postmarketing surveillance trials.

6. What kind of background/interests/personal qualities should someone wanting to work for a pharmaceutical company have? Somebody said that (applied) Statistics is the best way of sticking your nose in someone else's business. So, you would have to have all the characteristics that allow you to do that properly, for example curiosity, a multidisciplinary view, ability to relate to co-workers very different from yourself, patience, openness. Patience is especially required to put up with writing reports, often of a repetitive nature, and to conform to SOPs (Standard Operating Procedures) for doing almost everything, from writing protocols to sending email. Luckily, nowadays you can also cut and paste almost everything.

7. Is it possible to conduct methodological research within the pharmaceutical industry, or standard techniques are preferred? Yes, it is definitely possible to do high quality methodological work in the pharma industry, and Amy Racine is a good example of that. She participated to the Bayes MCMC explosion of the early nineties providing important applied motivations, like patient to patient variability in the analysis of PK/PD data and in general clinical data. If you want other examples, just open specialized journals like Statistics in Medicine or Controlled Clinical Trials. Of course, before they let you do methodological work, you have to prove that you can use standard techniques in the best possible way and, especially, faster than other colleagues who do not want to engage in methodological work. It is not that impossible for a Ph.D. in Statistics, since often your colleagues, many of them MAs or graduates from other disciplines, will have to struggle more than you.

8. When and why did you consider coming back to an academic job? There are some drawbacks to working within an industry: a good amount of routine work, time constraints and confidentiality. You really have to respect deadlines to deliver a protocol or a statistical analysis, otherwise the budget will run in the red. Then, even if you're doing some decent research, often you can not publish it because of confidentiality.

In a University, you also have time constraints and routine work due to teaching, but the game is much more under your control. And there is no confidentiality in research, so that Ambition and Vanity, the main driving forces of an academic, make you publish maybe more than you should.

9. How do academic research and industrial reasearch differ? What are the pros and cons of both? I've told you the cons of industrial research: the pros, apart from more money, are that you really feel you are doing something useful and interesting for somebody. Some years ago, I would have said "relevant". It is true that in academia you have more freedom, but you also have a weaker stimulus to renovate your research interests and produce something that more than a few people are willing to read.

10. Is there anything you would like to say to young PhD students? Well, first that I am not that old. Second, don't look down on applications and industry. Statistics is in danger of being bypassed as a discipline, as it happened to Operation Research or other branches of Mathematics. Take for example the current genomic rave. Ask a biologist what interdisciplinary help he or she can get from outside biology to analyse data, and the answer will be "Bioinformatics", not Statistics. Then you look well, and half of contemporary Bioinformatics is really Statistics in disguise. The advice is then: be nice to biologists, be nice to computer people, be nice to engineers. You really need to come closer to their language and understand what they need in order for your work to be recognized as an important contribution to modern science and technology. Never go to a scientist and ask "Give me some data, I want to try my new wonderful method". It never works. There are many more things in the real world than your little statistical philosophy can imagine. Besides, it is usually more fun to try to serve science rather than your private publication scheme.

11. What has been, in your opinion, the main contribution of Bayesians to clinical biostatistics (or to biostatistics in general)? Reporting the results of a statistical test or of an interval estimation using the Bayesian language is much more easy and straightforward. Classical statistics, whether Fisher or Neyman-Pearson like, involves conditional statements about the truth of a hypothesis, for example, that are seldom reported correctly. For this reason, Bayes Biostatistics is now much more accepted than a few years ago, see for example recent editorials in the British Medical Journal or even in the popular press. And there are official moves directed to eliminate the simple attachment of p-values to results in scientific journals.

Also, Bayesian hierarchical modeling and DAGs are very flexible and simple tools to model complexity. Scientists like them. David Spiegelhalter and his group, writing, documenting and mantaining BUGS, did more for Bayesian Biostatistics, in my opinion, than any other research enterprise.

Finally, the dynamical character of the Bayes reasoning, which is continually updated in the light of new evidence, fits well the actual flow of research and the accrual of knowledge. Scientists understand that, and they like it. Contributions of Bayesians to meta-nalysis and random effect modelling shed the correct light on the issue.

On the other hand, Clinical Biostatistics has given back to Bayesians lots of motivations and stamina. Take for example the complex relationship between a clinical statistician and a Health Authority, like the FDA. Nobody (well, almost nobody) has ever said you can not do a Bayesian analysis in a clinical trial. But you have to weigh and discuss your priors with the FDA ahead of time, before the trial begins, to gain credibility and make the results of some use for submission of a New Drug Application. This is one of the few examples I know of an external, objective, not (not so much) antagonistic and intelligent referee for your priors.

➤ Phd Thesis

Let us now introduce a doctorate thesis completed at the Department of Statistical Science of the University of Padua, Italy. The Ph.D program consists of three years, the first of which is devoted to attending institutional courses as well as seminars aimed at providing an outlook of potential research themes, while giving students the opportunity of getting in contact with experts in the field. More detailed information is available on-line at www.stat.unipd.it/dott/infodott.php.

Catia Scricciolo Dept. of Statistical Science, University of Padua catia@stat.unipd.it Consistency and rate of convergence of a sequence of posterior distributions in non-parametric problems. Advisors: Adriana Brogini and Larry Wasserman

In this thesis we investigate sufficient conditions for consistency and assess the rate of convergence of a sequence of pseudo-posterior distributions arising from an independent and identically distributed sample. The distribution can be viewed either as the pseudo-posterior corresponding to a pseudo-likelihood or as the posterior of a data-modified prior on a space of probability measures having density with respect to a common dominating measure, the space being equipped with the Hellinger metric. A posterior is strongly consistent if it asymptotically accumulates in Hellinger neighborhoods of the true distribution along almost all sample paths when sampling from the true density. Sufficient conditions for posterior consistency in this set-up have recently appeared in the literature, the underlying idea being to approximate the parameter space with sample-size-dependent sequences of finite-dimensional subsets satisfying both an entropy and a tail condition. On other hand, the sequence of distributions at study has been lately devised to allow investigation into posterior consistency of common non-parametric priors, i.e. Polya trees, infinite-dimensional exponential families and location mixtures of normal densities, avoiding recourse to sieves. We consider a generalized pseudo-posterior and show it is Hellinger consistent if only the Schwartz's support condition is fulfilled. This entails consistency also for the derived pseudo-Bayes density estimator. The pseudo-posterior concentrates on shrinking Hellinger neighborhoods of the true distribution of size a large multiple of the prior concentration rate, thus converging at least as fast as the true posterior. Leaving out the trivial case when the parameter space is totally bounded, hence the pseudo and true posteriors converge at the same rate, whether optimal or sub-optimal, in other cases results on large sample properties of the pseudo-posterior can be conveniently exploited to study the asymptotic behaviour of the corresponding true posterior. Thus, for location mixtures of normals, when the scale parameter is distributed independently of the mixing measure and is given an ad hoc sample-size-dependent prior, sufficient conditions for consistency that are present in the recent literature can be weakened. Furthermore, when the mixing distribution is the trajectory of a Dirichlet process and the true density a mixture of normals with mixing measure having either compact support or sub-Gaussian tails, without employing conditions on the tail behaviour of the base measure, the rate of convergence can be improved to the best rate currently known, the optimal rate for the problem being unknown.

GaussianWaveDen Toolbox for MATLAB 5.2

by Anestis Antoniadis, Jeremie Bigot
Laboratoire IMAG-LMC,
University Joseph Fourier,
BP 53,
38041 Grenoble Cedex 9,
France.
and
Theofanis Sapatinas,
Department of Mathematics and Statistics,
University of Cyprus,
P.O. Box 20537,
CY 1678 Nicosia,
Cyprus

Wavelet analysis has been found to be a powerful tool for the nonparametric estimation of spatially-variable objects. Gaussian Wavelet Denoising (`GaussianWaveDen`) is a *free of charge* software which implements various wavelet shrinkage estimators appearing in the literature for denoising Gaussian measurements. These estimators arise from a wide range of classical and empirical Bayes methods treating individual or blocks of empirical wavelet coe.cients either globally or in a level-dependent fashion.

`GaussianWaveDen` is a toolbox for `Matlab 5.2` and its subroutines should be installed by specifying its path in `Matlab 5.2`. Each function in `GaussianWaveDen` has an accompanying `html` help documentation in the directory `html-help`. `GaussianWaveDen` and the accompanying paper of Antoniadis, Bigot & Sapatinas (2001) which describes in detail the various wavelet shrinkage estimators that have been implemented in `GaussianWaveDen` can be downloaded from `http://www.jstatsoft.org/v06/i06`.

`GaussianWaveDen` makes an extensive use of the `MatLab` routines available in the *free of charge* `WaveLab 8.2` software. `WaveLab 8.2` is a toolbox for `Matlab 5.2` developed by Buckheit, Chen, Donoho, Johnstone & Scargle (1995). `WaveLab 8.2` has over 800 subroutines which are well documented, indexed and cross-referenced, and it is available as a compressed archive, in a format suitable for the machine in question: `.zip` (for MSWindows), `.tar.Z` (for Unix) and `.sea.hqx` (for Macintosh). The archives may be accessed from `http://www-stat.stanford.edu/~wavelab`.

References

[1] Antoniadis, A., Bigot, J., Sapatinas, T. (2001). Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study. *Journal of Statistical Software*, Vol. 6, Issue 6, 1-83.

[2] Buckheit, J.B., Chen, S., Donoho, D.L., Johnstone, I.M., Scargle, J. (1995). About WaveLab. Technical Report, Department of Statistics, Stanford University, USA.

NEWS FROM THE WORLD

by Gabriel Huerta
ghuerta@stat.unm.edu

✻ *denotes an ISBA activity*

➤ Events

International Workshop on Statistical Modelling
July 7-11, 2003,Leuven

Dear all,

It is my great pleasure to announce the '18th International Workshop on Statistical Modelling (IWSM)' to be held in Leuven (Belgium), from Monday July 7, to Friday July 11, 2003, with a short course on Sunday, July 6.

I also would like to refer you to the website http://www.luc.ac.be/censtat/IWSM2003/ which is continuously updated with the latest information concerning scientific and social programmes.

Confirmed invited speakers are Prof. Ron Brookmeyer (The Johns Hopkins University, U.S.A.), Prof. Marie Davidian (North Carolina State University, U.S.A.), Prof. Anastasios Tsiatis (North Carolina State University, U.S.A.), Prof. Henry Wynn (The University of Warwick, U.K.), and Prof. Chris Chatfield (The University of Bath, U.K.).

On Sunday July 6, prior to the start of the conference, a short course on Smoothing Methods will be given by Prof. Brian Marx (Louisiana State University, U.S.A.) and Prof. Paul Eilers (University of Leiden, The Netherlands).

Furthermore, the Workshop is organized back to back with the 'Theme Conference of the Royal Statistical Society on Statistical Genetics and Bioinformatics,' to be held in Diepenbeek (Belgium), July 14-17, 2003.

The IWSM has always been a rich environment for applied statisticians with different backgrounds, working in different environments, and with interests in statistical modelling in its most general sense. We hope to continue this tradition next year in Leuven, and we are very much looking forward to meeting many of you on this event.

Best regards,

On behalf of the local organizing committee, Geert Verbeke Biostatistical Centre K.U.Leuven Kapucijnenvoer 35 B-3000 Leuven Belgium Tel: +32 16 33 68 91 Tel (sec): +32 16 33 68 92 Fax: +32 16 33 69 00
Website: http://www.med.kuleuven.ac.be/biostat/

Workshop on Statistical Analysis of Gene Expression Data*July 11-14,2003, Wye College Conference Center, Kent, UK*

Organised by Sylvia Richardson (Imperial College) and Phil Brown (University of Kent)

This workshop is sponsored by EPSRC and the Royal Statistical Society. Its aim is to foster statistical research at the interface between new methodological developments and the biological and experimental context. Some of the topics to be covered are: – Data transformation and normalisation, – Differential gene expression, – Profile clustering and pattern recognition, – Discrimination and clinical profiling, – Experimental design, Implementation and algorithmic issues.

The workshop will be held at the conference centre of Imperial College situated in Wye (between Ashford and Canterbury, Kent), with easy access from London and on the Eurostar high speed rail route to Belgium from nearby Ashford. The workshop will stop at lunchtime on the 14th July so that participants of the RSS conference on ;Statistical genetics and Bioinformatics in Limburgs can easily travel there.
Website: http://www.med.ic.ac.uk/divisions/60/BGX/bgx/july2003workshop.html

International Conference of the Royal Statistical Society *July 15-17,2003, Limburgs Universitair Centrum, Diepenbeek, Belguim*

The International Conference of the Royal Statistical Society in 2003 will be hosted by the Center for Statistics of the Limburgs Universitair Centrum, Diepenbeek, Belguim; co-organized by the Biostatistical Centre of the Katholieke Universiteit Leuven. The main conference will be held from 15 July to 17 July, 2003. There will be a short course on 14 July 2003. The Chairman of the Local Organizing Committee for RSS 2003 is Geert Molenberghs.

Website: http://www.luc.ac.be/censtat/rss2003

The 23rd Annual Conference on Bayesian Methods and Maximum Entropy *August 3-8, 2003, Wyoming, USA*

Dear Colleagues,

The twenty-third workshop on "Bayesian and Maximum-Entropy Methods", for brevity "MaxEnt 23", will be held in Jackson Hole, Wyoming., August 3 - 8, 2003. The meeting is supported, in part, by the Edwin T. Jaynes International Center for Bayesian methods and Maximum Entropy, located at Boise State University and endowed by Dr. John Parker Burg (who, in the 1960's, developed the well-known Burg Algorithm). The meeting is being organized by professor Gary J. Erickson (of Boise State University) and Dr. C. Ray Smith, with assistance by the session chairman and others.

MaxEnt 23 will honor Professor Myron Tribus, whose many research papers on Bayesian and maximum Entropy Methods in science and engineering were of fundamental and technical importance but who single-handedly saw to a wide dissemination to the scientific and engineering communities of the new developments (of Jaynes and Tribus), via summer courses (starting in 1960), two textbooks (1961, 1969) and a major workshop (1978) on the Maximum Entropy journalism, all of which were instrumental in bringing the whole field to its current status. His continued presence at and support of the MaxEnt Workshops are also gratefully acknowledged.

For more on the list of speakers, registration and reservations please visit our web site. Anyone interested in organizing a session should contact Gary Erickson, gerickson@boisestate.edu, phone: 208 426 4401

Website: http://www.maxent23.org

✻ **First IMS-ISBA Joint Meeting** *July 24-26, 2003, Isla Verde, San Juan, Puerto Rico*

The 1st joint statistical meeting of IMS (Institute of Mathematical Statistics) and ISBA(International Society for Bayesian Analysis) will be held in Isla Verde, San Juan, PR, USA.

This joint meeting focuses on topics that are undergoing rapid development, and are of interest of members of IMS and ISBA. The three selected topics are: 1) Causal-Graphical Modeling; 2) Spatial Statistics and 3) Analysis of Extremes. Starting from an overview of the topic given by leaders in the field, presentations highlighting recent advances in each topic have been organized. There will also be poster presentations.

The registration fee is $160 (U.S currency) to be paid before April 23, 2003, through a check (USA bank) or money order payable to IMS, or by credit card (Master or Visa card only). Address: IMS/ISBA meeting registration, P.O. Box 22718, Beachwood OH 44122 FAX: 1-(216) 921-6703 (Credit card registration) - Tel: 1-(216) 295-2340 - E-mail: erg@imstat.org. For Puerto Rican participants registration is $80.

Pre-registration is available in two format: Microsoft word or PDF version. After you fill in the form, please send it back to us before April 23, 2003 by E-mail or by mail or by fax to the address below:

Dr. Luis R Pericchi
Director and Professor
Dept. of Mathematics and Computer Science
University of Puerto Rico, Rio Piedras Campus
P.O. Box: 23355
San Juan, PR00931-3355
Tel: +1(787)764-0000 EXT 1238
FAX: +1(787)281-0651 or +1(787)772-1437
E-mail: imsisba@orion.cnnet.clu.edu

Website: http://www.cnnet.clu.edu/math/ims-isba-pr2003

✻ **International Workshop on Bayesian Data Analysis** *7-10 August 2003, University of California, Santa Cruz, CA*

The focus of the workshop will be Bayesian data analysis: starting with a real problem in science or decision-making, formulating the problem in statistical terms, using Bayesian methods to solve the original problem, and discussing the strengths and weaknesses of the solution both statistically and substantively, with plenty of attention to the interplay between the real-world context and the Bayesian model-building, checking, and reformulating. The workshop is intended for statisticians, scientists, and engineers involved in applications requiring statistical inference, prediction, and decision-making and using Bayesian methods.

The meeting will be held on the campus of the University of California, Santa Cruz (UCSC). The goal is to bring together about 100 people interested in Bayesian applications in a variety of disciplines, including (but not limited to) bioinformatics, biostatistics, econometrics, engineering, epidemiology, computer science, machine learning, and statistics. Limited financial support is available and should be requested at time of registration. The registration deadline is 9 June 2003. As long as places are still available, registration will continue after this date up to and including the first day of the meeting, but after 9 June it may be more difficult for you to (a) have your contributed paper listed in the program and (b) receive full consideration for funding support. Participation in the workshop will be limited, and consideration will be given to program balance. Special consideration will be given to young investigators and Ph.D. students, and students and members of under-represented groups are especially encouraged to apply.
Website: link: http://www.ams.ucsc.edu/bayes03/

We expect that the Workshop will offer invited sessions in at least the following areas: bioinformatics, biostatistics/epidemiology, engineering applications, machine learning/computer science, nonparametric and semiparametric methods, simulation-based computation, and spatiotemporal modeling.

We hope to achieve significant participation by both young and more established researchers, and to bring people together from internationally leading research centers in as many continents as possible (at least Australia, Europe, North America, and South America).

Subject to the availability of sufficient travel funds, partial travel funding to the meeting will be available, in part on the basis of need, and with emphasis on funding to permit young researchers and members of groups underrepresented in science and engineering (e.g., underrepresented minorities, women, and persons with disabilities) to participate (we are hoping that the proximity of the Workshop in space and time to the Joint Statistical Meetings will permit people to attend both conferences with a fairly modest amount of travel support from the Workshop itself).
Website: http://www.ams.ucsc.edu

✻ **Seventh Workshop on Case Studies in Bayesian Statistics** *September, 12-13 2003, Carnegie Mellon University, Pittsburgh, PA*

The Seventh Workshop on Case Studies of Bayesian Statistics will take place on September 12th and 13th 2003 at Carnegie Mellon University, Pittsburgh, PA. The Workshop will feature in-depth presentations and discussions of substantial applications of Bayesian statistics to problems in science and technology, and poster presentations of contributed papers on applied Bayesian work. In conjunction with the workshop, the Department of Statistics' Seventh Morris H DeGroot memorial lecture will be delivered by Stephen Stigler.

We are calling for proposals in the form of detailed abstracts (about 2 pages) from those interested in presenting one of the main invited papers for discussion. To be considered for a presentation, abstracts are due by January 20, 2003. Abstracts should emphasize scientific and technological background, and should clarify the extent to which the statistical work will address key components of the problems articulated. They should also include statements that make clear the amount of work that will be accomplished by the time the manuscripts are due, which is July 1, and clearly identify the collaborators and particularly the non-statisticians who will be involved in the presentation. Case studies to be presented at the meeting will be selected by the organizing committee.

The organizing committee of the Seventh Workshop includes Alicia Carriquiry, Constantine Gatsonis, David Higdon, Rob Kass, Donna Pauler, and Isa Verdinelli.

Contributed paper abstracts will be due August 15, 2003. Please submit abstracts via our webpage which contains additional information, including abstracts of previous, successful case studies. If you have questions, please contact Rob Kass at the kass@stat.cmu.edu, or any of the other organizers.
Website: http://www.stat.cmu.edu/bayesworkshop