THE ISBA BULLETIN

Vol. 11 No. 4 December 2004

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE EDITOR

by J. Andrés Christen
jac@cimat.mx

I hope you will enjoy reading this issue of the ISBA Bulletin. I have included a section on general information of ISBA awards, that might be of interest to everyone, besides the more regular sections of Applications, Software Review and News of the World. Bruno Sansó and Gabriel Huerta are leaving as AE's. We thank them for their enthusiastic participation in creating this Bulletin. Also, I thank Marina Vannucci, Catherine Calder, Alexandra Schmidt and Ramses Mena to have accepted being AE's for the ISBA Bulletin. They are Annotated Bibliography, Applications, News from the World and Software Review AE's, respectively.

Recently I have recieved e-mails suggesting to advertize job positions, of particular interest for Bayesian statisticians, in the ISBA Bulletin. The Bulletin currently does not charge for advertisements. Including paid publicity (academic job positions, books etc. paid to the ISBA treasury) will represent, to say the least, a slight change in style of this Bulletin. It is my feeling that the ISBA community might have comments on this change and I will be very glad to hear your opinions, before a final decision is taken. I have started a discussion group where you may express your opinions, that will be active until 29th Feb. 2005. You may join the dicusion group at `http://groups-beta.google.com/group/ISBA-Bulletin-adds`. I will be very glad to hear your opinions.

Contents

*ISBA Bulletin, **11**(4), December 2004* **ISBA AWARDS**

GENERAL INFORMATION ON SOME OF THE ISBA RELATED AWARDS

The lindley Prize

by Phil Dawid
dawid@stats.ucl.ac.uk

The Lindley Prize is awarded for innovative research in Bayesian Statistics delivered as a contributed oral or poster presentation at either a Valencia International Meeting on Bayesian Statistics or a ISBA World Meeting, and published in the refereed proceedings of that meeting. The Prize is named for Dennis V. Lindley, and recognises the impact and importance of his work in the foundations, theory and application of Bayesian Statistics, and his marked influence on the evolution and spread of the discipline over many decades.

Award winning papers will present research in Bayesian statistics that is judged important, timely and notably original; truly innovative work will be judged more highly than successful development of ideas previously exposed. The Prize may be awarded for work in foundations, theory, methodology or applications of Bayesian statistics.

The Prize was established in 2000 by the Lindley Prize Founders, and is administered on their behalf by ISBA. The Founders provided initial financial contributions to the Prize Foundation and established the Charter governing the administration and award of the Lindley Prize.

The winner is announced and presented with the Prize every two years, at the alternating Valencia International Meetings and ISBA World Meetings. The award is made to the winning paper from the previous meeting. The winner receives a plaque and a cheque for $1,500.

The procedures for administration of the Prize, including details of the biennial process of review and selection of the Prize winner and the identities and roles of the Lindley Prize Founders, are covered by the Lindley Prize Charter. This is available on the ISBA website at http://www.bayesian.org/awards/Lindleycharter.

2002 Winner

The first competition for the Lindley Prize was based on the contributed papers presented at the 7th Valencia International Meeting on Bayesian Statistics held in Tenerife in 2002, and published in its Proceedings volume "Bayesian Statistics 7". The Selection Committee consisted of Philip Dawid (University College London, Chair), Jun Liu (Harvard), Kerrie Mengersen (Queensland), Julia Mortera (Roma Tre), and Mark Schervish (Carnegie-Mellon). The very high quality of all the papers made the Committee's task extremely demanding but also extremely rewarding.

The winner of the 2002 Lindley Prize was RADFORD NEAL (Toronto), in respect of his paper "Density Modeling and Clustering Using Dirichlet Diffusion Trees", in recognition of its originality and elegance, and the importance of its contributions to theoretical and applied Bayesian analysis. A plaque and cheque were presented to Radford at the ISBA 2004 World Meeting in Viña del Mar, Chile in May 2004.

2004 Competition

The competition for the second Lindley Prize will be based on contributed papers presented at ISBA 2004 and selected for publication as such EITHER in the Special Issue on Bayesian methods in Business and Industry of "Applied Stochastic Models in Business and Industry" (ASMBI), being edited by Ed George and Fabrizio Ruggeri, OR in ISBA's new electronic journal "Bayesian Analysis". Details of how to submit to these journals have been sent to all ISBA 2004 speakers.

The Selection Committee for the 2004 competition will comprise Philip Dawid (University College London, Chair), Carmen Fernandez (Lancaster), Val Johnson (Texas), Jun Liu (Harvard) and Mark Schervish (Carnegie-Mellon).

Deadlines

To be considered for the competition, papers must be submitted to one of the above journals before 31 DECEMBER 2004. All ISBA 2004 contributed papers that (i) meet this submission deadline, (ii) are accepted for publication, and (iii) whose FINAL version is submitted before 1 AUGUST 2005, will automatically be considered for the Lindley Prize. The announcement and award of the 2004 Prize will take place during Valencia 8 in June 2006.

The DeGroot Prize 2004

by Stephen E. Fienberg
fienberg@stat.cmu.edu>

The DeGroot Prize is awarded every two years by the International Society for Bayesian Analysis (ISBA) to the author or authors of a published book in Statistical Science. The Prize is named for Morris ("Morrie") H DeGroot, and recognizes the impact and importance of his work in Statistics and Decision Theory, and his marked influence on the evolution of the discipline over several decades through his personal scholarship, educational and professional leadership. The prize particularly recognizes DeGroot's authorship and editorship of major books that had marked impact on the development of Statistics and Decision Theory, and the value he placed on the role of books generally. Morrie's book, Optimal Statistical Decisions, helped educate a generation of statisticians and is one of the great books in the field. Published in 1970 and subsequently translated into both Russian and Polish, it provided an elegant and comprehensive treatment of a subject that has since come to be recognized as an essential part of statistics and of science as a whole. In 1975, his undergraduate text Probability and Statistics was published. A model of what a textbook should be, it played an important role in mathematical statistics curricula throughout the country. These books served as the inspiration for the creation of the DeGroot Prize.

2004 Winner

The 2004 winner of the DeGroot Prize is "The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation" (2nd edition) by Christian P. Robert, and published by Springer-Verlag in 2001. Prize winners receive a plaque and a check for $1,500. The winner was announced on May 27 at the ISBA 2004 banquet in Viña Del Mar, Chile.

Robert's book was selected from 22 stellar submissions; the entries represented the full spectrum of topics in statistical science including introductions to probability and statistics (Bayesian and non-Bayesian), historical books, non-Bayesian methods, through to applications of Bayesian methods in astrophysics, biology, data mining, finance, and risk assessment. The selection committee believes that Robert's book sets a new standard for modern textbooks dealing with Bayesian methods, especially those using MCMC techniques, and that it is a worthy successor to DeGroot's and Berger's earlier texts.

The members of the 2004 selection committee were: Kathryn Chaloner, Stephen Fienberg (chair), Ed George, Steffen Lauritzen, and Sylvia Richardson. The DeGroot Prize will be awarded again in 2006 with nominations solicited in the fall of 2005.

*ISBA Bulletin, **11**(4), December 2004* **APPLICATIONS**

ADAPTIVE EXPLORATION OF COMPUTER EXPERIMENT PARAMETER SPACES

by Robert B. Gramacy, Herbert K. H. Lee and William G. Macready
{rbgramacy,herbie}@ams.ucsc.edu
wgm@email.arc.nasa.gov

Many complex phenomena are difficult to investigate directly through controlled experiments. Instead, computer simulation is becoming a commonplace alternative to providing insight into such phenomena. The drive towards higher fidelity simulation continues to tax the fastest of computers, even in highly distributed computing environments. Computational fluid dynamics simulations in which fluid flow phenomena are modeled are an excellent example—fluid flows over complex surfaces may be modeled accurately but only at the cost of supercomputer resources. In this article, we discuss the problem of fitting a response surface for a computer model when we also have the ability to design the experiment adaptively, updating the experiment as we learn about the model– a task which we feel the Bayesian approach is particularly well-suited. Much of what is presented here follows our work in Gramacy et al. (2004).

Computational expense of the simulation and/or high dimensional inputs often prohibit the naive approach of running the experiment over a dense grid of possible inputs. However, computationally inexpensive surrogate models can often provide accurate approximations to the simulation, especially in regions of the input space where the response is easily predicted.

For example, consider a model for the computational fluid dynamics of flight conditions for a proposed reusable NASA launch vehicle called the Langley Glide-Back Booster. The simulations involve the integration of the inviscid Euler equations over a mesh of 1.4 million cells. Each run of the Euler solver for a given set of parameters takes on the order of 5-20 hours on a high end workstation. There are three input parameters (side slip angle, Mach number, angle of attack). Six outputs are monitored (lift, drag, pitch, side-force, yaw, roll). The Figure 1(a) shows lift as a function of speed and angle of attack. Of note is the large ridge at Mach 1, where the flight abruptly transitions from subsonic to supersonic. While most of the output space is rather smooth, the ridge is clearly not. Thus there is interest in being able to automatically explore this surface, learning about the ridge and spending relatively more effort there than in the smooth regions.

The above experiment is an example of a situation where surrogate models combined with active learning techniques could direct future sampling, dramatically reducing the size of the final experimental design, saving thousands of hours of computing time. Sampling can be focused on input configurations where the surrogate model is least sure of its predicted response, either because the output response is changing significantly or because there are relatively few nearby data points already examined.

The traditional surrogate model used to approximate outputs to computer experiments is the Gaussian process (GP). GPs are conceptually straightforward, easily accommodate prior knowledge in the form of covariance functions, and return a confidence around predictions. In spite of its simplicity, there are three important disadvantages to standard GPs in our setting. Firstly, inference on the GP scales poorly with the number of data points, typically requiring computing time that grows with the cube of the sample size. Secondly, GP models are usually stationary in that the same covariance structure is used throughout the entire input space. In the applications we have in mind, where subsonic flow is quite different than supersonic flow, this limitation is unacceptable. Thirdly, the error (standard deviation) associated with a predicted response under a GP model does not directly depend on any of the previously observed output responses.

All of these shortcomings may be addressed by partitioning the input space into regions, and fitting separate GPs within each region. Partitioning allows for modeling of non-stationary behavior, and can ameliorate some of the computational demands by fitting models to less data. Finally, a fully Bayesian approach yields uncertainty measures for predictive inference which can help direct future sampling.

Bayesian Treed GP Models

A tree model partitions the input space and infers a separate model within each region. Partitioning is accomplished by making (recursive) binary splits on the value of a single variable (e.g., speed > 0.8) so that partition boundaries are parallel to coordinate axes. These sorts of models are often referred to as Classification and Regression Trees (CART). CART has become popular because of its ease of use, clear interpretation, and ability to provide a good fit in many cases. The Bayesian approach is straightforward to apply to tree models, provided that one can specify a meaningful prior for the size of the tree. We follow Chipman et al. (1998) who specify the prior through a tree-generating process. Starting with a null tree (all data in a single partition), the tree, $\mathcal{T}$, is probabilistically split recursively with each partition, η, being split with probability $p_{\mathrm{SPLIT}}(\eta, \mathcal{T}) = a(1+q_\eta)^{-b}$ where q_η is the depth of η in $\mathcal{T}$ and a and b are parameters chosen to give an appropriate size and spread to the distribution of trees. We expect a relatively small number of partitions, and choose a and b accordingly. Extending the work of Chipman et. al (2002), we fit a stationary GP with linear trend independently within each of R regions, $\{r_\nu\}_{\nu=1}^R$, depicted by the tree, $\mathcal{T}$. The GP correlation structure for each partition is chosen either from the isotropic power family, or separable power family of unknown (random) parameterization. In both cases, the correlation function takes the form $K_\nu(\mathbf{x}_j, \mathbf{x}_k) = K_\nu^*(\mathbf{x}_j, \mathbf{x}_k) + g_\nu \delta_{j,k}$ where $\delta_{\cdot,\cdot}$ is the Kronecker delta function, and K_ν^* is a *true* correlation representative from a parametric family. Priors which encode our belief that the global covariance structure is non-stationary are chosen for parameters to K_ν^* and g_ν.

Most literature on the *Design and Analysis of Computer Experiments* [6, 5] deliberately omits the nugget parameter (g) on grounds that computer experiments are deterministic. However, there are many reasons why one may wish to study a computer experiment with a model that includes an explicit noise component. In particular, the experiment may, in fact, be non-deterministic. Our collaborators tell us that their CFD solvers are often started with random initial conditions, involve forced random restarts when diagnostics indicate that convergence is poor, and that input configurations arbitrarily close to one another often fail to achieve the same estimated convergence. Thus a conventional GP model without a small-distance noise process (e.g. a nugget) can be a mismatch to such inherently non-smooth data.

The data $\{\mathbf{X}, \mathbf{t}\}_\nu$ in region r_ν are used to estimate the parameters $\boldsymbol{\theta}_\nu$ of the model active in the region. Parameters to the hierarchical priors depend only on $\{\boldsymbol{\theta}_\nu\}_{\nu=1}^R$. Samples from the posterior distribution are gathered using Markov chain Monte Carlo (MCMC). Integrating out dependence on the tree structure $\mathcal{T}$ is accomplished by reversible-jump MCMC. We implement the tree operations *grow, prune, change*, and *swap* similar to those in Chipman et al. (1998).

Adaptive Sampling

In the world of Machine learning, adaptive sampling would fall under the blanket of a research focus called *active learning*. Active learning techniques are currently being applied successfully in areas such as computational drug design/discovery by aiding in the search for compounds that are active against a biological target. However, we are not aware of any other active learning algorithms that use non-stationary modeling to help select small designs.

In the statistics community, the traditional approach to sequential data solicitation goes under the general heading of *(Sequential) Design of Experiments* [6]. Depending on whether the goal of the experiment is inference or prediction (as described by a choice of utility), different algorithms for obtaining optimal designs can be derived. For example, one can choose the Kullback-Leibler distance between the posterior and prior distributions (with parameters $\boldsymbol{\theta}$) as a utility. For Gaussian process models with correlation function $\mathbf{K}$, this is equivalent to maximizing $\det(\boldsymbol{K})$. Subsequently chosen input configurations are called $D-$optimal designs. Choosing quadratic loss leads to what are called $A-$optimal designs. An excellent review of Bayesian approaches to the design of experiments is contained in Chaloner & Verdinelli (1995) .

A hybrid approach to designing experiments employs active learning techniques. The idea is to consider a set of candidate input configurations and choose a rule for deciding the order in which they should be added to the design. For example, consider an approach which maximizes the information gained about model parameters by selecting the location $\tilde{\mathbf{x}}$ which has the greatest standard deviation in predicted output. This approach has been called ALM for Active Learning–Mackay, and has been shown to approximate maximum expected information designs. Given its simplicity this is the method we explored first. MCMC posterior predictive samples provide a convenient estimate of location-specific variance; namely the width of predictive quantiles.

An alternative algorithm is to select $\tilde{\mathbf{x}}$ minimizing the resulting expected squared error averaged over the input space, called ALC for Active Learning–Cohn. Conditioning on $\mathcal{T}$, the reduction in variance at a point $\mathbf{y}$ given that the location $\mathbf{x}$ is added into the data has a simple closed form. Averaging over $\mathbf{y}$ gives an estimate of the reduction in predictive variance obtained by adding $\mathbf{x}$ into the design—easily computed using MCMC methods. A comparison between ALC and ALM using standard GPs appears in [7].

Given these two hybrid approaches to sequential design, constructing a list of input configurations to send to available computing agents is simply a matter of sorting candidate locations ranked via either ALM or ALC. That way, the most informative locations are first in line for simulation when agents become available. Candidates could come from a pre-defined grid, a random sub-sample, a Latin Hypercube (LH) sample, an optimal design (e.g. a sequentially *D*-optimal design), or some combination (e.g. LH sub-sample of a *D*-optimal design).

Experimental Results

Bayesian adaptive sampling (BAS) proceeds in trials. Suppose N samples and their responses have been gathered in previous trials (or from a small initial grid, before the first trial). In the current trial the model is estimated for data $\{\mathbf{X}_i, t_i\}_{i=1}^N$. In accordance with the ALM algorithm, MCMC predictive quantiles are gathered, and sorted. Since our current experiments are based on pre-calculated pairs of input configurations and responses delivered by NASA, candidates (for now) must be chosen via random-subsample from the available data. We developed an artificial clustered simulation environment with a fixed number of agents in order to simulate the parallel and asynchronous evaluation of input configurations. After refreshing the sorted list of candidates, BAS gathers finished and running input configurations and adds them into the design. Predictive mean estimates are used as surrogate responses for unfinished (running) configurations until the true response is available. New trials start with fresh candidates.

Figure 1(a) shows one of the six outputs (lift) plotted as a function of speed (Mach) and angle of attack (alpha) based on the full design of more than 3000 input configurations. The third input, side slip angle (beta), is fixed at zero. A fitted surfaced based upon 750 total samples is shown in Figure 1(b). Configurations gathered using BAS (for beta=0) are shown in the Figure 1(c). Also shown in Figure 1(c) is a representative sample of the partitions obtained by integrating over the tree ($\mathcal{T}$). BAS has the desired behavior in that it fits different models around and on either side of the Mach 1 regions, and focuses most of the adaptive sampling around Mach 1. Further partitioning and sampling occurs for large angle of attack (alpha) near Mach 1 as indeed the response is changing most rapidly in this region.

Visually, there is little difference between the true surface, Figure 1(a), and the estimated surface, Figure 1(b). However, using a Bayesian treed GP model with adaptive sampling requires fewer than $1/4$ as many samples compared to a simple gridding, saving thousands of hours of computing time. For a more detailed analysis of these results, experiments on other data, and comparisons with other approaches, the interested reader is referred to a paper we presented at ICML 2004 [4]. Our future work includes running a live experiment on the NASA supercomputers.

In conclusion, creating a surrogate model for computer experiments is a problem that will continue to be of interest, as additional computing resources are put toward more accurate simulations rather than faster results. The Bayesian approach allows a natural mechanism for creating a sequential design based on the current estimated uncertainty.

References

[1] Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design, a review. *Statistical Science*, *10 No. 3*, 273–1304.

[2] Chipman, H., George, E., & McCulloch, R. (1998). Bayesian CART model search (with discussion). *Journal of the American Statistical Association*, *93*, 935–960.

[3] Chipman, H. A., George, E. I., & McCulloch, R. E. (2002). Bayesian treed models. *Machine Learning*, *48*, 303–324.

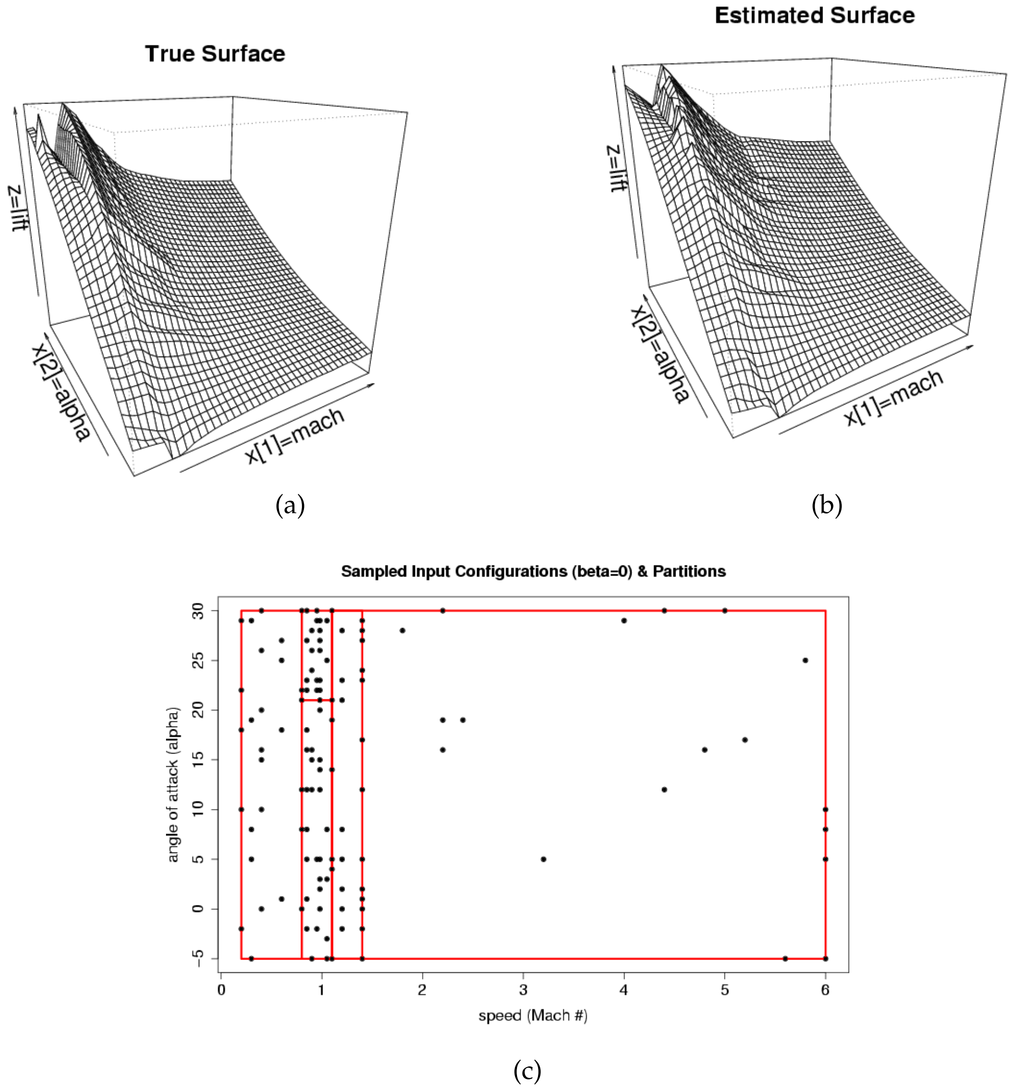
[4] Gramacy, R. B., Lee, H. K. H., & Macready, W. (2004). Parameter space exploration with Gaussian process trees. *Proceedings of the International Conference on Machine L earning* (pp. 353–360). Omnipress & ACM Digital Library.

[5] Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, *4*, 409–435.

[6] Santner, T. J., Williams, B. J., & Notz, W. I. (2003). *The design and analysis of computer experiments*. New York, NY: Springer-Verlag.

[7] Seo, S., Wallat, M., Graepel, T., & Obermayer, K. (2000). Gaussian process regression: Active data selection and test point rejection. *Proceedings of the International Joint Conference on Neural Networks IJCNN 2000* (pp. 241–246). IEEE.

Figure 1: (a) CFD projections, true surface based on ∼3000 data points. (b) CFD projections, fitted surface based on 750 adaptive samples over 100 trials. (c) Adaptively sampled input locations (for a slice of side-slip-angle (beta = 0)).



(a)

(b)

(c)

BAYESIALAB: THE DECISION SUPPORT AND DATA MINING TOOL

by Lionel Jouffe
jouffe@bayesia.com

Introduction

From a statistical viewpoint, a Bayesian network efficiently encodes the joint probability distribution of the variables describing an application domain. It is represented in a graphical annotated form that seems quite natural to human experts for a large variety of applications. The nodes of a Bayesian network correspond to domain variables and the arcs that connect the nodes correspond to direct probabilistic relations between these variables.

BayesiaLab is a complete cross-platform laboratory that allows one to design and to use Bayesian networks. The decision models can be designed through expertise and/or automatically extracted from data; the represented knowledge can be quickly assimilated by using a set of original analytical tools; the models can be used in interactive or in batch mode; dynamic systems can be modeled by using Dynamic Bayesian networks; action policies can be discovered thanks to reinforcement learning algorithms.

Bayesian network editing

The creation of nodes and arcs is realized thanks to mere clicks and drags. The graphical area comes with a magnetic grid, some alignment tools, and a set of powerful layout algorithms are available for a helpful assistance in the layout of the graph. An ergonomic editing tool allows an easy access to all the necessary information related to the nodes: type (Label for symbolic state nodes and Interval for continuous variables), values (a wizard is available for the automatic generation of values) and probability distributions. These distributions can be directly described in the tables (the editor has completion and normalization functions), or can be quickly and concisely described by using the powerful equation editor (discrete and continuous probability distributions, arithmetic and trigonometric functions, all the usual operators). Lastly, the Man-Machine interface has a comprehensive search tool that enables one to search for nodes and arcs (wildcards characters are available for a flexible description of the search), and supports full Cut & Paste functionalities. Networks, nodes, tables, equations, charts and reports can be pasted into BayesiaLab and into all the external tools that implement the clipboard functions.

Communication and Traceability

In order to make it easier to understand the knowledge represented by the networks, hypertext with links to external documents (text, images) can be associated to the network, to nodes and arcs. Color tags can also be associated to group nodes and arcs into semantic sets. Finally, an image can be set to the background of the network to quickly indicate the studied domain.

Inference

Two kinds of inferences are available for static Bayesian networks. For moderately connected networks, it is possible to use an exact inference algorithm based on the construction of a junction tree. When the connectivity of the network prevents the construction of this intermediate structure, approximate inference based on Monte Carlo simulations can be used. The algorithm implemented is the likelihood weighting algorithm.

Two kinds of inference are also available for dynamic Bayesian networks:

- Inference based on a junction tree: whereas this algorithm is exact for static networks, it returns approximate results in the dynamic case since at each time step, only the marginal probabilities of the nodes of the $t+1$ slice are back propagated to the nodes of the t slice.

- Inference based on Monte Carlo simulations (particle filtering): the approximation is of the same order as in the static case, i.e. the approximation is not related to the dependence of the nodes but it is only due to the random character of the simulation

Analysis tool box

Thanks to their graphical structure, Bayesian networks are often considered as a very readable formalism. However, this readability quickly decreases while the size of the network increases. Furthermore, even for reasonably sized networks, many inferences are necessary to really understand the knowledge encoded by the network. In order to improve the understanding of the networks, BayesiaLab offers a complete analysis toolbox.

- The first analysis tool allows to compute the force of the relations. Based on the Kullback-Liebler distance between the joint probability distribution with and without the arc, arcs can be printed with a thickness proportional to their force. The exact values of the KL distance can be obtained by generating a HTML analysis report. By sorting the relations in a decreasing order, this tool makes it possible to focus the analysis only on the important relations (especially useful when the network has been automatically learnt).

- The analysis can also be focused toward a target node. Colored squares are then printed inside the nodes that are correlated to the target. The brightness of the square color is proportional to the information brought by the node to the knowledge of the target (mutual information). This tool is very convenient for the illustration of the D-Separation concept. A HTML report can also be generated. This report returns an ordered list of the correlated nodes and describes the probabilistic profile of each value of the target.

- The target value analysis focuses on a particular value of the target variable. The squares are replaced by some “ smileys ”, whose brightness is proportional to the mutual information. The “ smileys ” characterize the type of the probabilistic relation that links the node and the target value.

- When various hard evidences are set, an analysis of the set of observed values returns a contradiction measure. Three sets of observations are defined: those that confirm a predefined root observation, those that contradict it, and the independent observation set.

- Influence path analysis is a tool that highlights the paths taken by the information from a node to the target node. Once again, this tool is indeed very helpful to understand the D-Separation concept.

- Bayesian networks represent dependence and independence probabilistic relations. Networks made out of the same skeleton but with different arcs orientation can represent the same set of relations. They are said to belong to the same Equivalence Class. This last analysis tool displays this Equivalence Class by removing the orientation of the arcs that can be inverted without modifying the joint probability distribution. If there is no hidden variable, relations that remain oriented can be considered as causal.

Learning

Whereas Bayesian Networks are historically designed by expertise, BayesiaLab offers a broad set of learning algorithms to automatically induce Bayesian networks from data. It is thanks to such algorithms that BayesiaLab stands out as a powerful Data Mining tool. Data can be imported from text files or directly extracted from databases with SQL requests by using a JDBC/ODBC bridge. Missing values can be rigorously processed by using the current Bayesian network to infer their probabilities. The used algorithm is an Expectation/Maximization algorithm. Note that hidden/latent variables are also processed thanks to that algorithm. Three discretization algorithms are available to deal with continuous variables: equal distances to create equal length intervals, equal frequencies for intervals with the same a priori distributions, and Decision Tree, an Entropy-based algorithm that induce the best thresholds with respect to a discrete target variable. Once data is imported and discretized, learning algorithms can be used. All these algorithms are based on the Minimum Description Length score for the qualification of the candidate networks. This score takes into account the structure complexity and the network fitness to data. BayesiaLab offers four kinds of Bayesian network learning algorithms, the first one relative to the parameters only, and the three others being used for the induction of the structure and the parameters. These learning algorithms allow to:

- Estimate the parameters of a given Bayesian network, i.e. the conditional probability tables;
- Induce a Bayesian network representing all the probabilistic relations that hold in data. Three algorithms are proposed: SopLEQ, an algorithm that searches the space defined by the equivalent classes (i.e. the networks representing the same set of dependences and independences), Taboo, that directly searches the space of Bayesian networks, and Taboo Order, an algorithm searching the space defined by the ordered list of nodes (indeed, it is trivial to find the best Bayesian network for a given order). The starting point in the search space corresponds to the fully unconnected network, except for Taboo that starts from the current network and supports expert knowledge expressed with fixed arcs.
- Design a Bayesian network dedicated to the characterization of a target variable. Five algorithms are available: the Naïve Structure is based on two strong hypotheses: (H1) all the variables depend on the target, and, (H2) all the variables are independent of all the others knowing the target value; The Augmented Naïve algorithm relaxes the second hypothesis of the Naïve (H2) by searching the relations between the variables knowing the target; Son & Spouse relaxes the two hypothesis (H1) and (H2) by searching the target dependent variables and the dependence between those variables conditionally to the target; The Markov Blanket algorithm finds the minimal subset of variables that are really important for the target characterization: its parents, its children and its co-parents. Knowing the values of these variables makes the target variable independent of all the others; Finally, the Augmented Markov Blanket algorithm uses the previous one and searches the relations that hold between the variables of the Markov Blanket.
- Cluster the data into groups of points sharing the same characteristics. This segmentation algorithm consists in adding a hidden variable to the network, the Cluster, and finding the number of clusters and their characterization. The used structure is the Naïve one where the target variable is the Cluster, and an Expectation/Maximization algorithm is used to estimate the conditional probabilities.

Note that all these learning algorithms take rigorously into account a priori knowledge expressed by an initial Bayesian network associated to an equivalent number of cases quantifying the expertise.

Decision Support

BayesiaLab also offers a broad set of Decision Support tools

- Adaptive Questionnaire proposes dynamic series of questions. The order of the questions takes into account the relevance of the information given to the target variable and the corresponding cost of those questions. This set of questions is updated after each answer.
- The lift curve is plotted by setting on the X-axis the individuals and on the Y-axis the percentage of individuals having the target value with respect to a given potential number of target values. The individuals are sorted according to their probability, returned by the Bayesian network, to belong to the target value. Since this curve is interactive, it allows to find the optimal threshold, i.e. the one corresponding to the best compromise between the number of considered individuals and the obtained target value rate: treating x% of individuals allows obtaining y% of the potential target values.
- Imputation of missing values is naturally and rigorously dealt by Bayesian networks. By associating a data base to the network, BayesiaLab automatically imputes a value to all the encountered missing values. The imputed value is randomly sampled from the conditional probability distribution returned by the network by taking into account the evidence that is set on the other variables.
- Utility nodes describe cost/gain functions on their associated variables (parents). By using the parents' probability distribution, they allow to compute expected costs/gains associated to the states defined by all the variables, and also to evaluate policies. For example, by setting evidences on a set of variables (e.g. some actions), what are the consequences on the global state value, i.e. the sum over all the defined costs/gains?
- Decision nodes are used to define policies. They do not have an associated probability

table but a quality table that returns a numerical value for each action modality with respect to each combination of its parents' values. The parents' combination represents the perception states of the Decision node. The optimal action policy is then directly available since it is described in the quality table: for each state, it corresponds to the action modality with the best quality. These quality tables can be directly entered by the user or automatically learnt by using reinforcement learning. Based on various trials and errors and on a scheme of rewards and punishments modeled by the Utility nodes, the learner will interact with the system described by the network to find a policy that optimizes the reinforcement received signals. Reinforcement learning is available for Decision nodes belonging to static as well as dynamic Bayesian networks.

Dynamic Bayesian networks

In order to model systems that evolve with time, as for example those studied in reliability analysis, it is necessary to take the temporal dimension into account. Even if it is possible to do it with a static Bayesian network, by unrolling it on the desired number of time steps (i.e. by duplicating the network for each time step), this solution is possible only for a limited and previously known number of time steps. Dynamic Bayesian networks provide a much more compact representation for stochastic dynamic systems. This compactness is based on the following assumptions: (A1) the process is Markovian, i.e. the variables of time step t depend only on the variables of the preceding time step $t-1$; (A2) the system is time invariant, i.e. the probability tables do not evolve with respect to time. This last assumption can be partially relaxed in BayesiaLab by using the time variable to modify the probability distributions according to the value of the current time step by the means of the equations. This representation allows representing stochastic dynamic systems only with two time slices. The first slice describes the initial network at time step t_0 and the second one describes the temporal transitions $t+1$.

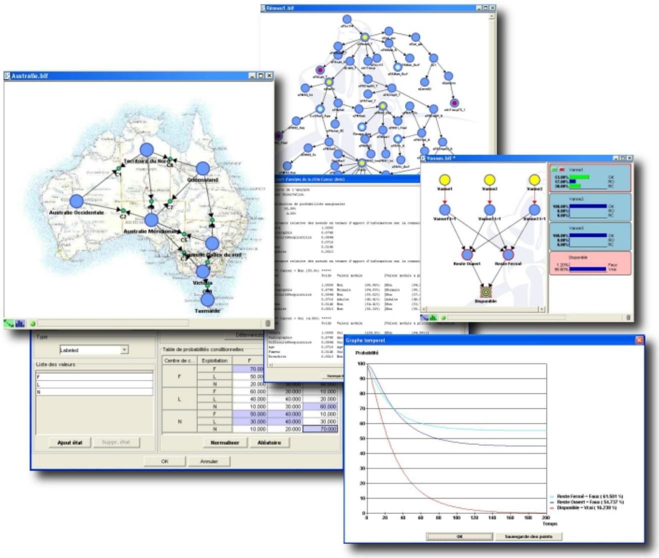
Conclusion

An evaluation version of BayesiaLab and a dynamic presentation are available on

`http://www.bayesia.com`.

This website also contains some application examples that describe the use of BayesiaLab in various domains: Modeling and simulation of complex systems, Risk analysis, Mining customer data bases, Intrusion detection, Text Mining, MicroArrays analysis and Health Trajectory analysis.

Figure 2: Working environment in BayesiaLab



NEWS FROM THE WORLD

by Gabriel Huerta
ghuerta@stat.unm.edu

✻ *denotes an ISBA activity*

➤ Events

Statistics Conference in Honor of Jim Press. May 14, 2005. University of California Riverside, USA.

The Department of Statistics at the University of California, Riverside (UCR) will be hosting a one-day conference in honor of Professor S. James (Jim) Press to commemorate the occasion of his 28 years of distinguished service at UCR, and his 50 years in the mathematics/statistics profession.
The conference will take place at UCR, on Saturday, May 14, on the UCR campus, approximately 8:30-5:30. Then, a sit-down banquet, followed by a poster session. The night before (May 13), there will be a party at his home in Riverside for colleagues, students, visitors, friends, and family.
Topics for the conference reflect the wide-ranging interests of Professor Press, particularly in Bayesian Analysis, Multivariate Analysis, and Cognitive Aspects of Sample Surveys.
The keynote speakers for the conference are:

- Ingram Olkin, Professor of Education and Statistics, Stanford University
- Judith Tanur, Distinguished Teaching Professor of Sociology, State University of New York at Stony Brook
- Arnold Zellner, H.G.B. Alexander Distinguished Service Professor Emeritus of Economics and Statistics, University of Chicago

Participants already include: Barry Arnold, Bob Beaver, Hamparsum Bozdogun, Norman Bradburn, Mark Ghamsary, Wesley Johnson, Jay Kadane, Ruben Klein, Sang Lee, Ingram Olkin, Dale Poirier, John Rolph, Kazuo Shigemasu, Hal Stern, Judy Tanur, Liangwei Wang, and Arnold Zellner.
The Chair of the conference is Professor Subir Ghosh (ghosh@ucrac1.ucr.edu). Information about the program, travel, accommodations, and registration will be available on the conference website: `http://statistics.ucr.edu`, when the website has been completed.

First Call For Submissions

This meeting will consist of invited talks and a poster session. If you would like to present at the conference please submit a title and abstract to cecelias@ucr.edu

Workshop on recent advances in modeling spatio-temporal data. May 25-26, 2005. Southampton Statistical Sciences Research Institute (S3RI) University of Southampton, Southampton, UK.

The aim of the workshop is to discuss recent developments in modeling spatio-temporal data across a wide range of application areas and identify directions for future work.
The workshop will feature oral presentations by several international experts, poster presentations and round table discussion groups. For further information see the workshop website
`http://www.maths.soton.ac.uk/s3riwshop/`
The workshop is to be preceded by a two-day short-course on hierarchical modeling of spatial and temporal data which will be given by Professor Alan Gelfand (Duke University, USA). For further information about the short-course, see the website:
`http://www.maths.soton.ac.uk/s3ricourse/`
The numbers of participants on both the workshop and the short-course are strictly limited, and for this reason we encourage you to register as soon as possible. Information and registration forms are available from the above websites.
There are a small number of Royal Statistical Society Student Bursaries available for full-time registered research students who make a poster presentation at the workshop. Students may apply by completing the appropriate part of the registration form.
Please contact the workshop organizers with any queries:

Sujit Sahu and Sue Lewis
S3RI
School of Mathematics
University of Southampton
Southampton, SO17 1BJ
UK
E-Mail: s3riwshop@maths.soton.ac.uk

✻ **Bayesian Inference in Stochastic Processes** June 2-4, 2005. Varenna, Italy.

The workshop will bring together experts in the field to review, discuss, and explore directions of development of Bayesian inference in stochastic processes for Bayesian Inference. There will be session on Markov processes, state-space models, spatial, empirical, birth-death, and branching processes. Theoretical and applied contributions (for example, queuing, population models, signal processing) are welcome. For details, contact Antonio Pievatolo at antonio.pievatolo@mi.imati.cnr.it, or visit `www.mi.imati.cnr.it/conferences/bisp4.html`

Joint annual meeting of the Western North America Region (WNAR) of the International Biometric Society and the Institute of Mathematical Statistics (IMS) June 21-24, 2005. Fairbanks, Alaska, U.S.A.

The University of Alaska Fairbanks will host the 2005 joint annual meeting WNAR and IMS. Contributed, invited, and student paper sessions are planned as well as one or more continuing education workshops. Participants are encouraged to register and make airline and lodging reservations early, as June is peak tourist season. For those wanting to see more of Alaska before or after the conference, the conference web page provides a link to the Fairbanks Convention and Visitors Bureau where you can plan your Alaskan adventure. Hiking, white water rafting and sight seeing at Denali National Park is about a two-hour drive from Fairbanks. The planned deadline for abstracts is April 15, 2005. Please check the web site `www.uaf.edu/wnar` for updated information as it becomes available.
Contact information:
For WNAR - Gilbert W. Fellingham(gwf@byu.edu)
For IMS - Thomas Lee (tlee@stat.colostate.edu)
submitter: Dana Thomas (ffdlt@uaf.edu)

Recent Advances in Biostatistics, Bioinformatics and Markov Chain Monte Carlo July 7-8, 2005. Sydney, Australia.

This symposium will focus on statistical issues in both Biostatistical and MCMC fields. Cross-disciplinary research will also be presented. Registration deadline: July 1, 2005. Estimated attendance: 100. Contact Yanan Fan, Department of Statistics, School of Mathematics, University of New South Wales, Sydney, 2052, Australia, yanan@maths.unsw.edu.au, `www.maths.unsw.edu.au/~scott/symposium`

✻ Eighth Workshop on Case Studies in Bayesian Statistics. September 16-17, 2005. Carnegie Mellon University, Pittsburgh, PA, USA.

The Eighth Workshop on Case Studies of Bayesian Statistics will take place on September 16th and 17th 2005 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, poster presentations of contributed papers on applied Bayesian work and, new this year, contributed presentations by young researchers. In conjunction with the workshop, the Department of Statistics' Eighth Morris H DeGroot memorial lecture will be delivered. Abstract due January 25. We are calling for proposals for major case studies 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 Tuesday, January 25, 2005. 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 on the basis of all abstracts received. This year we are also soliciting detailed abstracts (roughly 1 page) of proposed 15-minute presentations by young researchers (students or completed PhD within five years). These abstracts will be due July 1, and the organizing committee will select among them in constructing the final program. Abstracts should emphasize the scientific problems, and the way in which the statistical work solves the problems. Abstracts not selected for talks will be considered as posters. Anyone interested in submitting a case study abstract should look at the web page, What makes a good case study? `http://www.stat.cmu.edu/bayesworkshop/2005/goodstudy.html`
Contributed paper abstracts for posters will be due September 1, 2005.
The organizing committee of the Eighth Workshop includes Alicia Carriquiry, Elena Erosheva, Constantine Gatsonis, Rob Kass, Herbie Lee, and Isa Verdinelli.
Please submit abstracts via our webpage `http://www.stat.cmu.edu/bayesworkshop` which contains additional information, including abstracts of previous, successful case studies.
If you have questions, please contact Rob Kass at kass@stat.cmu.edu, or any of the other organizers.

➤ Miscellanea

✻ Bayesian Analysis journal

I am very pleased to announce that the new ISBA electronic journal Bayesian Analysis is taking submissions at

`http://ba.stat.cmu.edu`

For a brief description of the journal and its editorial board, see that web page. We are interested in outstanding research and scholarship. Please submit your work!! Some comments follow. Bayesian Analysis will be published on our web site and will be freely available. It will be dedicated to rapid editorial turnaround of manuscripts, which will be facilitated by (1) a large board of editors and associate editors who will handle most refereeing, and (2) an electronic manuscript-handling system that will greatly reduce the book-keeping overhead for the editorial board. Much effort has gone into creation of the automated manuscript-handling system. Written in PHP/MySQL, it not only enables reviewers to get papers, but also keeps track of all editorial activities and allows instant access to the status and complete history of manuscripts. My hope is that the system will not only reduce the organizational effort required of editors and associate editors but that it will, in addition, relieve the editorial assistant from most of these chores as well (thereby reducing the assistant's job to only a few hours per week, and making the cost of running the journal very small).
The system has the following features:

- Articles are submitted in pdf format and are accessible to relevant referees and editorial board members on the system web site.
- Articles are tracked by a unique article reference number, and authors may use this number to check on the status of a submitted article.
- Editors, AEs and referees may view a list of the articles assigned to them, and may then examine the history and current status of any of these articles.
- Editors have access to an editorial load monitor so that editors and AEs can be picked taking account of load over the past 12 months.
- The system automatically sends email reminders of response due dates to editors, AEs, and referees.
- Letters to authors are composed by Editors, checked by the Editor-in-Chief, and sent to authors using the system. They are then archived by the system and are accessible to relevant past and future reviewers.
- The system allows editors, AEs, and referees to compose messages for other users of the system and archives all such correspondence. This is intended to help with organization of all internal email discussion concerning an article.
- The system maintains logs of all editorial activities related to each article.

The system could not have been constructed without the supervisory advice of our Electronic Production Manager Pantelis Vlachos, the extremely proficient programming of Adrian Rollett, and the miscellaneous help of our very capable editorial assistant Heather Wainer. I am personally grateful to all of them and pleased to acknowledge their work. As you use the system, please don't hesitate to let us know of any comments or suggestions for improvement.
I'm excited to be involved in a much-needed vehicle for communication across the ever-widening network of people interested in Bayesian methods. I expect Bayesian Analysis to begin publishing sometime in 2005 and will send an announcement when this occurs.

Rob Kass
Editor-in-Chief
Bayesian Analysis

ISBA
INTERNATIONAL SOCIETY FOR BAYESIAN ANALYSIS

Executive Committee

Program Council

Board Members

EDITORIAL BOARD

Editor

Associate Editors