

THE ISBA BULLETIN



Vol. 24 No. 4

December 2017

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE ISBA EXECUTIVE

Establishment of a Task Team for a Safe ISBA

As many of you will be aware, over the past few days reports have come out of serious incidents of sexual harassment at ISBA events. The Executive Committee members of ISBA wish to communicate that ISBA takes reports of inappropriate behavior very seriously. We have a great Society and a proud record of including women as members and as leaders. We are in the process of establishing protocols for appropriate behaviour and support mechanisms for our members.

A Task Team for a Safe ISBA has been put in place, with the task of establishing protocols for a safe and inclusive Society.

The team currently comprises the following people:

Kerrie Mengersen (lead)
Marina Vannucci
Clara Grazian
Amy Herring
Daniele Durante
Christian Robert
David Rossell
Dan Simpson

Immediate duties of the Task Team are as follows:

- Establish a code of behaviour for ISBA members;
- Establish a website for safe posting of comments relating to harassment among members;
- Promote the message of safe behaviour and inclusivity among members.

A more complete list of duties will be developed by the Task Team and will be presented to the Board for ratification early in 2018. We are asking our members to work with us to ensure that we have an inclusive society and that our events are safe for all to enjoy. We have established an account at safeisba@bayesian.org and encourage anyone to register concerns or comments about this and related issues. Your input is valued. Regular updates on the activities of the Task Team will be on the ISBA website <https://www.bayesian.org>.

Signatures:

Kerrie Mengersen – ISBA President
Marina Vannucci – President-elect
Steve MacEachern – Past President
Amy Herring – Executive Secretary
Robert Gramacy – Treasurer

In this issue

- FROM THE PROGRAM COUNCIL
 ☛ Page 3
- BA UPDATE
 ☛ Page 4
- THE WITCHES OF BAYES
 ☛ Page 5
- CONFERENCE REPORTS
 ☛ Page 7
- NEWS FROM THE WORLD
 ☛ Page 9
- SOFTWARE HIGHLIGHT
 ☛ Page 10

FROM THE EDITOR

- Beatrix Jones -

m.b.jones@massey.ac.nz

The news addressed in the statement from the executive committee broke just as the final touches were being put on the Bulletin. Kerrie had previously prepared a president's column, mentioning lots of great things ISBA had done in 2017. The *Bulletin* is often like that, a bit of a cheerleader for what's going on in the Bayesian community. However, we are about communicating matters of interest and importance to the Society and its members, not just cheerleading. When there are problems, we want them discussed in the *Bulletin*, and in particular we want to talk about solutions. The *Bulletin* will make it a priority to relay the decisions and recommendations of the task team, and as always we are open to suggestions from readers.

We have, in fact, had correspondence about a professional problem—one of a very different nature. Giulio D'Agostini has written to us about the new Italian fairy tale *The Witches of Bayes*, designed to introduce probabilistic reasoning to children. The problem arises because the lesson of the tale is not entirely accurate. Read his synopsis of the story and let us know what you think—either in article or fairy tale form. We will publish responses in the next issue.

We have some fascinating features in this issue of the bulletin. The Software Highlight features the package *BDgraph*, for fitting graphical models in R; it has been recently updated and we are fortunate to have its authors, Abdolreza Mohammadi and Adrian Dobra, describe its new features for us. As well as updated information about the 2018 World Meeting in Edinburgh, we have conference reports from Bayes on the Beach and BAYES2017, the Bayes Pharma meeting. Your editor must apologise for the late appearance of the Bayes Pharma report, kindly contributed by Virgilio Gómez-Rubio (much) earlier in the year; but perhaps it is well timed for people considering attending BAYES2018, to be held in Cambridge immediately before the World Meeting.

Best wishes for the new year,
Beatrix

FROM THE PROGRAM COUNCIL

program-council@bayesian.org

ISBA 2018 World Meeting

Preparations for the 2018 World Meeting to be held 24th-29th June, 2018, University of Edinburgh are now quite advanced and we expect to be sharing these updates regularly on the website <https://bayesian.org/2018-world-meeting/> over the coming months.

In the latest news, Emeritus Professor Philip Dawid has been selected to present The de Finetti Lecture at the meeting.



Philip Dawid is Emeritus Professor of Statistics, University of Cambridge. For many years he was Professor of Probability and Statistics at University College London. He has been Editor of *Biometrika* and of the *Journal of the Royal Statistical Society (Series B)*. He is an elected Fellow of IMS, ISI and ISBA, and has served as Vice-President of RSS and as President of ISBA, of which he is an honorary lifetime member. Other honours include the Snedecor and DeGroot Prizes, and RSS Guy Medals in Bronze and Silver. His research interests include logical foundations of Bayesian and other schools of probability and statistics, Bayes nets, statistical causality, probability forecasting, and forensic inference.

Confirmed Foundational Lectures at the event will be given by:

- Professor Alan Gelfand, Duke University, USA.
- Professor Edward George, University of Pennsylvania, USA.
- Emeritus Professor Anthony O'Hagan, The University of Sheffield, UK.
- Professor Judith Rousseau, Université Paris Dauphine, France.

The Keynote Lectures series will be presented by:

- Professor Nicolas Chopin, ENSAE, France.
- ProfessorMontserrat Fuentes, Virginia Commonwealth University, USA.
- Professor Michael Jordan, University of California, Berkeley, USA,
- Professor Steve MacEachern, The Ohio State University, USA.

It has been an extended process, but the proposals presented by the ISBA Sections have now been assessed. The organisers of the successful proposals have been notified by email this week. Unsuccessful proposals presented by the Sections were re-assessed with the Member proposed sessions, so if you have not heard the outcome of your proposal, you can expect to hear in the next wave of correspondence.

Member Contributed sessions have also now been assessed by the Scientific Committee. Notifications of these outcomes are currently being prepared and emails should reach proposed session organisers during the week beginning December 11th. All the proposals were of excellent quality, so the Scientific Committee are very excited about the program that is being built for Edinburgh.

Workshop proposals have also been assessed with the outcomes being announced soon. Abstract submission will be opening via the website in the near future.

We will also be offering travel support to students and early career researchers. This support is largely thanks to generous donations from various organisations. At the present, we have confirmed contributions from Google, Stata and RStudio. We are currently seeking support from several other funding bodies which should be finalised in early 2018. We would like to encourage young researchers and students who intend to apply for travel support to submit abstracts at their earliest convenience.

Upcoming ISBA endorsed events in 2017-2018

We would like to highlight the following meetings that are being co-sponsored by ISBA:

O-Bayes17: International Workshop on Objective Bayes Methodology, December 10-13th, 2017.
<https://sites.google.com/site/obayes2017/>

BayesComp 2018, March 26-28th, 2018.
<https://www.maths.nottingham.ac.uk/personal/tk/bayescomp/>

2017 ISBA@NIPS Travel Award

Francesco Locatello is the 2017 recipient of the ISBA@NIPS Travel Award for early career researchers. This award consists of \$US 700 to cover reimbursable expenses associated with attending the meeting.

Francesco is a Ph.D. student at the Max Planck-ETH Center for Learning Systems, supervised by Gunnar Rätsch and Bernhard Schölkopf. He is passionate about optimization for machine learning, where his research focuses on constrained greedy optimization and structured learning applied to biomedical informatics. At NIPS he will be presenting the paper “Boosting Variational Inference: an Optimization Perspective” which outlines an theoretical analysis combining boosting, an old idea from statistical learning theory, with a variational Bayes approach.

UPDATE FROM BA

From the BA Editor

- Bruno Sansó -

bruno@soe.ucsc.edu

The December issue of the journal is available online at <https://projecteuclid.org/current/euclid.ba>. This issue includes the discussion paper: “Uncertainty Quantification for the Horseshoe” by Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. The invited discussants are Ismael Castillo, Ryan Martin and Nic Polson. The paper has been received with great interest by the community. A testament to this interest is the fact that three contributed discussions have been submitted to BA. Up to now the journal had published the discussion paper, the discussions, and the rejoinder as standalone files, but, starting with this issue, we have changed the format. From now on, discussion papers will be published as a bundle that includes the discussions and the rejoinder. The title of the

paper will include the words “with discussion”. In addition to the discussion paper, the December volume will feature the invited review paper on “Deep Learning: A Bayesian Perspective”, by Polson and Sokolov. The paper offers a critical review of the state of the art on the topic of deep learning, highlighting the challenges faced by a Bayesian approach. We believe that this is a timely article that will raise the interest in our community to work in what is one of the hottest topics in the research and application of learning methods.

As has become a tradition, Bayesian Analysis is publishing in its December issue the papers presented at the ISBA 2016, that were submitted to the journal and have successfully completed the editorial process. There are seven papers that went through the usual rigorous vetting standards of the journal. They are grouped together in the volume under the title “Articles from the 2016 ISBA World Meeting”. They will be competing for the Lindley Prize, whose winner will be announced at the closing banquet of the 2018 ISBA World Meeting.

The discussion paper for the March issue will be the paper “Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data”, by Jonathan R. Bradley, Scott H. Holan, and Christopher K. Wikle. We invited Kater Calder, Stefano Castruccon and Dani Gamerman to discuss the paper. The paper is already available in advanced publication at <https://projecteuclid.org/euclid.ba/1507687687>. We welcome contributed discussions. These should be no more than two pages in length, using the BA latex style. They should be submitted to the journal using the EJMS system <https://www.e-publications.org/ims/submission/>, before December 15, 2017.

A word about publication charges. Bayesian Analysis is an open access journal and does not have a mandatory publication charge. We do suggest a donation to offset the cost of running BA. The suggested amount is \$550 and it is collected via the IMS website. This is a small amount when compared to the cost of publishing in many scientific journals. In 2016 the cost of running BA was about \$7,300. We received about \$4,000 from publication charges, which means that ISBA made up for the remaining \$3,300. In view of these numbers, we strongly encourage the authors that publish in Bayesian Analysis to pay the suggested publication charges in order for the journal to stop being a financial burden to ISBA.

THE WITCHES OF BAYES

Giulio D’Agostini has written to the Bulletin about a recent initiative of the Italian National Institute of Statistics (ISTAT), Italy’s body for official statistics. ISTAT staff members have written a book of fairy tales that illustrate probabilistic reasoning, for use with primary school aged children. While in many ways this has been a successful endeavor with a good response from teachers, there have also been issues raised around the correctness of the lessons imparted, particularly by the story “The Witches of Bayes.” Giulio has provided a synopsis of the story, which follows below, followed by the part of the appendix concerning it. Readers of Italian can find the original version here: <https://www.istat.it/it/archivio/204260>. “The Witches of Bayes” starts on page 42, with an appendix on page 62.

We invite responses from Bulletin readers. Topics could include, but are not limited to

- The strengths and shortcomings of the story as a teaching tool.
- When we discover popular materials with shortcomings, how do we best engage to remedy the situation? Do you have a story of successful interactions of this nature?
- An alternate version, sequel or prequel to the story. Illustrations are welcome but not required!

Send your response to the bulletin editor at m.b.jones@massey.ac.nz.

Synopsis A party of 21 witches inhabited a cave just outside the beautiful mountain village of Bayes. A particular attraction of this spot was the delicious smell of sweet and savory goods from the oven of the village bakery.

Every morning the group of witches chose a witch at random, and she placed her hut outside the entrance of the cave. The hut could be black or pink. The villagers from Bayes had to fill the hut with food, but only of one type, sweet or savory. If the food matched the tastes of the witch, the rest of the day would be quiet, otherwise the poor Bayesians suffered from practical jokes of all kinds from the

angry witches. Thus the divination of the tastes of the witch became extremely important. The head of the village was confident of the help of the god Bias, whose response was indicated by the side of a stone coin tossed in the air. One side of the stone was bright, in which case savory food was given; the other side was dark, in which case sweet food was given. But this strategy did not give satisfactory results.

Then Nora, the young daughter of the baker and our heroine, enters the scene. She thought that the preference of each witch had to do with the color of the hut, and remembered an old nursery rhyme—the problem of the witches went back to ancient times—that could be interpreted as follows: the witches with black huts like only savory food; $6/7$ of the witches with pink huts like sweet food, the remaining $1/6$ savory.

Relying on her interpretation of the rhyme, Nora devised the following strategy in order to minimize the nuisance of the angry witches. If the hut was black, then they had to provide savory food. If the hut was pink, they had to draw from a bag containing seven lockets. Six of the seven lockets were marked with D for Dolce, ‘sweet’, and one was marked S for Salato, ‘savory’.

Nora proposed the strategy to the head of the village, but faithful in the god Bias, he was unmoving. Nevertheless, she persisted. Nora asked for permission to bring the food to the witches herself. With the complicity of her family, she used her locket bag strategy instead of the directions of the Bias oracle. The happy end of the story was that days where the witches were angry became much less frequent.

Note: In the story there are a couple of hints which could have an interesting inferential-predictive flavor, but which are not followed up. At a certain point it is written that Nora became curious thinking about which hut will be put outside the cave; just after that, she had the impression that there more pink than black huts and that sweet was the taste preferred by the witches, but she was not sure.

Appendix A smart little girl and a cave which hides greedy and spiteful witches. How will the inhabitants of the village manage to control them? What will be the taste of the witch who has left her hut outside the cave? The old method entrusted the decision to the roll of a coin. The children, instead, will be guided to the discovery of the Bayesian reasoning, which helps in change our own decisions when new pieces of knowledge are acquired. In the story, in fact, the link between the food tastes of the witches and the color of their hut is enclosed in a numeric nursery rhyme, thanks to which Nora will be able to make more reliable forecasts. The theory of probability is the mathematics of uncertainty, it helps us to understand how chance works and to foresee how phenomena we do not understand with certainty will evolve. Thomas Bayes was a British mathematician, famous for the formulation of the theorem that carries his name.

CONFERENCE REPORTS

BAYES ON THE BEACH

Farzana Jahan

Queensland University of Technology
f.jahan@hdr.qut.edu.au

Bayes on the Beach 2017 was a 3-day international conference held during November 13-15, 2017 in the Mantra Legends Hotel, Surfers Paradise, Gold Coast. This conference was supported by the Bayesian Statistics Section of the Statistical Society of Australia (SSA), Australian Chapter of the International Society for Bayesian Analysis (ISBA), ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) and Queensland University of Technology (QUT). This was the 12th International Workshop for the Australasian chapter of the International Society for Bayesian Analysis (ISBA) and the annual meeting of the Bayesian Statistics section of the Statistical Society of Australia (SSA). This is an intentionally small but international forum: there were about 82 participants this year.

The conference featured four distinguished speakers, with a keynote presentation and more general tutorial from each of them. Professor Susan Holmes (Stanford) talked about statistical challenges for microbial ecology in her keynote presentation followed by a tutorial focusing on the proper use of available data for next generation sequencing. Professor Mark Girolami (Imperial College) presented "A Bayesian Perspective for Stochastic Modelling of Urban Structure" in his keynote session, and discussed the role for Bayesian statisticians in probabilistic numerical analysis in the tutorial. Professor Gael Mar-

tin (Monash), in her keynote speech, narrated the history of Bayesian computation beginning in 1763, and followed up with a tutorial on recent advances in approximate Bayesian computation. On the final day, Professor Scott Sisson (University of New South Wales) presented "Symbolic Data Analysis for Modelling Big Data Fast Using a Laptop" and provided a short introduction to modelling extremes in the tutorial. There was also an invited speech from Professor Peter Donnelly (University of Oxford and Director of the Wellcome Centre for Human Genetics) who gave an interesting talk about genomics and big data analytics, highlighting routes to improved drug development.

Four invited presentation sessions broadly focused on the themes of: Modelling, High Dimensional Inference, Priors, and Designs. The Bayesian tradition of lively poster sessions was adhered to during two evenings of the conference, with a total of 39 posters displayed. A unique element of this conference was the workshop at the end of each day. All the participants of the conference had to split into four groups where there were four different problems to consider: modelling of restaurant rating data, Bayesian variable selection, spatio-temporal models, and monitoring the progress of elite swimmers through benchmark data collected during their training. The final day each group had to present a summary of the work they had been doing throughout the workshop. The workshops were a great opportunity of learning and networking.

Detailed information regarding the conference, presentations etc. can be found at: <https://botb2017.wordpress.com/>.



BAYES2017 (BAYES PHARMA MEETING)

Virgilio Gómez-Rubio
University of Castilla-La Mancha
Virgilio.Gomez@uclm.es

This year, the Bayes Pharma conference took place at the Faculty of Pharmacy, University of Castilla-La Mancha in Albacete (Spain) from the 22nd to the 25th of May. As in previous years, the conference focused on different aspects regarding drug development and the pharmaceutical industry. Altogether, the conference gathered 50 people from the academy, industry and regulatory agencies.

Prof. Carmen Armero Cervera (U. of Valencia, Spain) gave a pre-conference course on the 22nd of May about Bayesian survival analysis.



The next day, the welcome speech was given by Prof. Mar Arroyo (dean of the Faculty of Pharmacy), Prof. Emmanuel Lesaffre (U. Luvein, Belgium), on behalf of the scientific committee and Dr. Virgilio Gómez-Rubio, as chair of the local organising committee.



Invited talks comprised different topics on Bayesian analysis. The first one was given by Prof. Maria de Iorio (UCL, UK) on Bayesian variable selection. The next invited talk was about Bayesian models for Big Data by Dr. Steve L. Scott (Google, USA). On the next day, Dr. Harry Yang (Med-

Immune, USA) presented an invited talk on how Bayesian statistics can be used to influence decisions in the regulatory agencies. In the afternoon, Dr. Roseann White (Duke Clinical Research Institute, USA) gave another invited talk about how information from previous studies for the design of medical devices. The last day of the conference, Prof. Jesús López-Fidalgo (U. of Navarre, Spain) presented an introduction to Bayesian optimal experimental design.

In addition to the invited talks, the conference had a number of contributed talks on different topics, regarding the development of new drugs and the pharmaceutical industry. These talks were presented by some of the attendees, including local Ph.D. students and academic staff, attendees from other European universities and a statistician from several pharmaceutical companies.

The conference could have not happened without the support of the Quetelet Society, the Royal Statistical Society and the International Society for Bayesian Analysis (ISBA). Locally, the conference has been supported by the University of Castilla-La Mancha, through its Faculty of Pharmacy and the Department of Mathematics.

The conference slides and abstract booklet can still be found at <https://www.bayes-pharma.org>.



NEWS FROM THE WORLD

Announcements

The AbbVie Statistics Visiting Scholar Program in Honor of Dr. David C. Jordan

AN INSPIRING SABBATICAL OPPORTUNITY FOR WORLD-CLASS STATISTICIANS

This program offers a rare opportunity to collaborate with AbbVie statisticians on research topics that are of interest to both you and our business. The final determination of specific projects does not need to be made prior to arrival, and areas of application may include clinical trials or a number of non-clinical applications.

AbbVie is a global, research based biopharmaceutical company formed in 2013 following separation from Abbott Laboratories. The company's mission is to use its expertise, dedicated people and unique approach to innovation to develop and market advanced therapies that address some of the world's most complex and serious diseases. For further information on the company, visit www.abbvie.com.

We're looking for academic statisticians and biostatisticians to join us in a consulting and research position during their sabbatical (6 to 12 months) in the 2018-2019 academic year. To be considered, you must have an outstanding research record in statistics, biostatistics, bioinformatics or a related field. It's also essential that you have produced relevant publications in methodology and applied research and have excellent communication skills.

AbbVie is located 40 miles north of downtown Chicago. Successful scholars will join us here as part of our Data and Statistical Sciences (DSS) organization. The DSS organization currently employs approximately 150 statisticians and 80 statistical analysts/programmers worldwide, who provide scientific support for all phases of drug development and safety surveillance. Our Data Sciences group has a wide range of accountabilities, which include development and use of technology for data collection and database creation.

Interested applicants should submit a resume and cover letter for consideration to Alan Hartford at Alan.Hartford@abbvie.com. The closing date for applications is January 31, 2018.

Call for ISBA Fellow Nominations

Nominations are now being accepted for new ISBA Fellows. Nominees must be current ISBA members and should have been members for the last three consecutive years, at least. The individual's contributions should have had a significant impact in promoting Bayesian ideas and methods in society, through scientific works and other activities, such as teaching, consulting or service. Submissions and supporting letters should not be made by current members of the Fellows Selection Committee to avoid conflicts of interest. Nominations may be made by any current ISBA member.

Please prepare nominations as a single pdf file following the convention LAST-NAME.FIRSTNAME.pdf so the file name corresponds to the nominee and then email them to fellows@bayesian.org. The nomination should include:

1. a cover letter containing the both the nominator's and nominee's names and email addresses and a citation to be used if the nomination is successful
2. the nominee's current CV
3. three letters of support from ISBA members (other than the nominee) that address the criteria for ISBA Fellow

Nominations must be received on or before February 1, 2018. The new fellows will be announced at the 2018 ISBA World Meeting in Edinburgh, UK (June 24-29, 2018). Please contact Judith Rousseau (rousseau@ceremade.dauphine.fr), Chair of the ISBA Fellows Selection Committee, with any questions.

Call for Zellner Medal Nominations

Nominations are now being accepted for the ISBA Zellner Medal, which recognizes ISBA members who have rendered exceptional and distinguished service to ISBA over an extended period of time, and whose contributions have had an impact on the society beyond the time of his or her incumbency. Nominees must be current ISBA members and should have been members for the last three consecutive years, at least. They should have

served ISBA in a range of leadership roles over an extended period of time.

Please prepare nominations as a single pdf file following the convention LAST-NAME-FIRSTNAME.pdf so the file name corresponds to the nominee and then email them to zellner-medal@bayesian.org. The nomination should include:

1. a cover letter containing the both the nominator's and nominee's names and email addresses and a citation to be used if the nomination is successful;
2. five letters of support from ISBA members (other than the nominee) that address the criteria for the Zellner Medal.

Nominations must be received on or before March 15, 2018. The Zellner Medal recipient(s) will be announced at the 2018 ISBA World Meeting in Edinburgh, UK (June 24-29, 2018). Please contact Fabrizio Ruggeri (fabrizio@mi.imati.cnr.it), Chair of the Zellner Medal Selection Committee, with any questions.

Meetings and conferences

XIV Brazilian Meeting on Bayesian Statistics (EBEB), de Janeiro, Brazil. March 05-09, 2018.

The meeting is organized by ISBrA – the Brazilian Chapter of ISBA and will honor Professor Dani Gamerman for his outstanding contributions to the development and dissemination of Statistics in Brazil and worldwide.

For detailed program and schedule visit the XIV EBEB website: www.redeabe.org.br/eb2018.

Conference Bayes Comp 2018, Barcelona, Spain. March 26-28, 2018

Bayes Comp is a biennial conference sponsored by the ISBA section of the same name. The conference and the section both aim to promote original research into computational methods for inference and decision making and to encourage the use of frontier computational tools among practitioners, the development of adapted software,

languages, platforms, and dedicated machines, and to translate and disseminate methods developed in other disciplines among statisticians.

Details about the conference, the schedule, deadline as well as the list of plenary speakers and details about invited sessions can be found in the conference website <https://www.maths.nottingham.ac.uk/personal/tk/bayescomp/>.

ABS18 Applied Bayesian Statistics Summer School: Bayesian Statistical Modeling and Analysis in Sport, Villa del Grumello, Como, Italy. June 4-8, 2018.

The school will be the 15th in the sequence. It will replicate the successful format of the previous years, and will feature lectures and practical sessions (run by a junior researcher), as well as participants' talks.

The aim of this course is to increase students' ability to develop Bayesian models and computational solutions for real problems in the world of sport. A case study based teaching approach will be taken for the course. Each day, students will be presented with one or two problems posed by Sports Institutes regarding aspects of athlete training for world games. Through participatory problem solving, the students will be challenged to learn about theory, methods and applications of a range of Bayesian models including mixtures, spatio-temporal models, hidden Markov models and experimental design, and computational approaches including Markov chain Monte Carlo and Approximate Bayesian Computation. This hands-on course pays equivalent attention to theory and application, foundation and frontiers in Bayesian modelling and analysis. While the focus of the case studies is on sport, both sporting novices and lovers of sports are welcome, noting that the learning obtained in the course will be widely applicable to many other areas.

The lecturer will be Kerrie Mengersen, Distinguished Professor of Statistics, Queensland University of Technology, Brisbane, Australia. The conference webpage is web.mi.imati.cnr.it/conferences/abs18.html. Registration is now open. Please note that the conference room allows only for a limited number of participants. The ABS18 Secretariat can be contacted at fabrizio@mi.imati.cnr.it.

SOFTWARE HIGHLIGHT

THE R PACKAGE `BDGRAPH` FOR BAYESIAN STRUCTURE LEARNING IN GRAPHICAL MODELS

Abdolreza Mohammadi, Adrian Dobra
 University of Amsterdam, University of Washington
 a.mohammadi@uva.nl, adobra@uw.edu

Graphical models [12] provide a probabilistic framework to characterize the multivariate dependency structure among random variables. These models have received considerable attention in the literature, and have a vast domain of applicability that encompasses all the scientific fields in which the analysis of multivariate datasets is key, e.g. biology, neuroscience, social sciences, and economics. Graphical models make use of graphs to represent relations (e.g., independence or conditional independence) among random variables. A crucial step in data analysis with graphical models is estimating the underlying graph. This is a very difficult computational problem when many random variables are involved. Bayesian methods provide a flexible framework for incorporating uncertainty of the graph structure: inference and estimation are based on averages of the posterior distributions of quantities of interest, weighted by the corresponding posterior probabilities of graphs [9].

The package `BDgraph` [17, 18] for R [22] provides easy-to-use functions for Bayesian structure learning in undirected graphical models for continuous, discrete, and mixed variables. The package implements recent results in the Bayesian literature, including [3, 4, 5, 14, 15, 16]. The package provides several distinctive features: (1) the computationally intensive tasks have efficient parallel implementations in `OpenMP` [19]; (2) all the code is written in C++ and interfaced with R; (3) in addition to functions for Gaussian graphical models (GGMs), `BDgraph` provides functions for fitting Gaussian copula graphical models (GCGMs); (4) `BDgraph` has functions for graph determination for graphical models continuous, discrete and mixed variables based on the marginal pseudo-likelihood (MPL) approach [5, 20]. The MPL gives a practical way to balance computational complexity and accuracy to scale up to large-scale problems.

Over the last couple of years, the `BDgraph` pack-

age has been constantly improved in terms of the speed of computations, and its functionality has been extended. Now at version 2.43, the package offers the end-user a practical option for carrying out Bayesian structure learning in undirected graphical models with several hundreds of variables.

1 Software design and implementation

The `BDgraph` package provides a complete pipeline for multivariate data analysis with graphical models. Its functions are organized in three modules — see Figure 1.

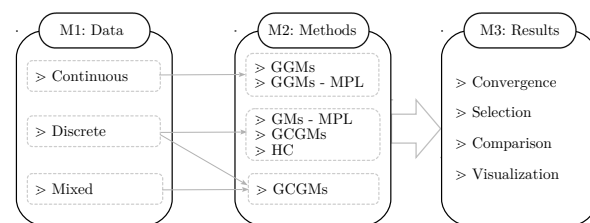


Figure 1: Functional modules of the `BDgraph` package: (M1) data simulation; (M2) methods and search algorithms; and (M3) various functions for convergence check of the search algorithms, graph selection, comparison and goodness-of-fit, and graph visualization.

Module 1 (M1) – Data simulation. The function `bdgraph.sim` generates multivariate Gaussian, discrete and mixed data given different types of undirected graphs, including “random”, “cluster”, “scale-free”, “hub”, “circle”, “AR(1)”, “AR(2)” and “fixed”. Users can employ this function to simulate multivariate data based on these types graphs, or just to simulate graphs of those types.

Module 2 (M2) – Methods and algorithms. The main functions of this module are called `bdgraph` and `bdgraph.mpl`. They implement several estimation methods with different sampling algorithms, as follows.

- Graph estimation in GGMs using the birth-death MCMC (BDMCMC) sampling algorithm described in [15, 16]. This approach is implemented in the function `bdgraph`.
- Graph estimation in GCGMs for non-Gaussian, discrete, and mixed data, using

the BDMCMC sampling algorithm described in [14]. This approach is implemented in the function `bdgraph`.

- Graph estimation based on the reversible jump MCMC (RJMCMC) structural learning algorithms [3, 4]. This approach is implemented in the function `bdgraph`.
- Graph estimation based on MPL for discrete, Gaussian, and non-Gaussian data, using the BDMCMC sampling algorithm described in [5]. This approach is implemented in the function `bdgraph.mpl`.
- Graph estimation based on the hill-climbing (HC) algorithm for discrete data described in [20]. This approach is implemented in the function `bdgraph.mpl`.

Module 3 (M3) – Results. This module includes four types of functions:

- *Graph selection:* The functions `select`, `plinks`, and `pgraph` provide the selected graph, the posterior edge inclusion probabilities, and the posterior probability of each graph, respectively.
- *Convergence check:* The functions `plotcoda` and `traceplot` provide several visualization plots to monitor the convergence of the sampling algorithms.
- *Comparison and goodness-of-fit:* The functions `compare` and `plotroc` provide several comparison measures and a receiver operating characteristic (ROC) plot for model performance comparison.
- *Visualization:* The functions `plot.bdgraph` and `plot.sim` produce graph visualizations of the simulated data and estimated graphs. They are implemented based on `igraph` package [2] for R [22].

While the `BDgraph` package has functions for graph estimation using several different methods including RJMCMC and HC, those methods based on the BDMCMC algorithm have key advantages in terms of their computational performance. A brief introduction to BDMCMC for graphical models is presented in the next section.

2 Bayesian structure learning in graphical models

A graphical model for a random vector $X = (X_1, X_2, \dots, X_p)$ is specified by an undirected graph $G = (V, E)$ where $V = \{1, \dots, p\}$ are vertices

or nodes, and $E \subset V \times V$ are edges or links [12]. A vertex $i \in V$ of G corresponds with variable X_i . The absence of an edge between vertices i and j in G means that X_i and X_j are conditional independent given the remaining variables $X_{V \setminus \{i, j\}}$. The graph G also has a predictive interpretation. Denote by $\text{nbr}_G(i) = \{j \in V : (i, j) \in E\}$ the neighbors of vertex i in G . Then X_i is conditionally independent of $X_{V \setminus (\text{nbr}_G(i) \cup \{i\})}$ given $X_{\text{nbr}_G(i)}$ which implies that, given G , a mean squared optimal prediction of X_i can be made from the neighboring variables $X_{\text{nbr}_G(i)}$.

We focus on the structural learning problem [6, 11] which aims to estimate the structure of G (i.e., which edges are present or absent in E) from the available data $x = (x^{(1)}, \dots, x^{(n)})$. In a Bayesian framework, we explore the posterior distribution of G conditional on the data x , i.e.

$$P(G | x) = \frac{P(G)P(x | G)}{\sum_{G \in \mathcal{G}_p} P(G)P(x | G)}, \quad (1)$$

where $P(G)$ is a prior distribution on the graph space \mathcal{G}_p and $P(x | G)$ is the marginal likelihood of the data conditional on G [11]. Determining the graphs with the highest posterior probabilities (1) is a complex problem since the number of possible undirected graphs $2^{\binom{p}{2}}$ becomes large very fast as p increases. For example, for $p = 50$, the number of possible undirected graphs exceeds the largest possible value in R ($1.8e + 308$). This motivated the development of computationally efficient search algorithms for exploring large spaces of graphs that have the ability to move quickly towards high posterior probability regions by taking advantage of local computation.

Among them, the BDMCMC algorithm [5, 14, 15, 16] is a trans-dimensional MCMC algorithm, and represents an alternative to the well known RJMCMC algorithm [8]. The BDMCMC algorithm is based on a continuous time birth-death Markov process [21]. Its underlying sampling scheme traverses \mathcal{G}_p by adding and removing edges corresponding to the birth and death events. Given that the process is at state $G = (V, E)$, we define the birth and death events as independent Poisson processes as follows:

Birth event – each edge $e \in \bar{E}$ where $\bar{E} = \{e \in V \times V : e \notin E\}$, is born independently of other edges as a Poisson process with rate $B_e(G)$. If the birth of edge e occurs, the process jumps to $G^{+e} = (V, E \cup \{e\})$ which is a graph with one edge more than G .

Death event – each edge $e \in E$ dies independently of other edges as a Poisson process with

rate $D_e(G)$. If the death of edge e occurs, the process jumps to $G^{-e} = (V, E \setminus \{e\})$ which is a graph with one edge less than G .

This birth-death Markov process is a jump process with intensity $\alpha(G) = \sum_{e \in \bar{E}} B_e(G) + \sum_{e \in E} D_e(G)$. Its waiting time to the next jump has an exponential distribution with mean $1/\alpha(G)$. Thus, the birth and death probabilities are proportional to the birth and death rates.

To optimize the convergence speed, following [5, 16], we define the birth and death rates as follows:

$$R_e(G) = \min \left\{ \frac{P(G^* | x)}{P(G | x)}, 1 \right\}, \quad (2)$$

for each $e \in \{E \cup \bar{E}\}$,

where for the birth of edge e we take $G^* = (V, E \cup \{e\})$, and for the death of edge e we take $G^* = (V, E \setminus \{e\})$. The rates are calculated based on the MPL approach as described in [5].

The BDMCMC algorithm is presented in pseudo-code in Algorithm 1. It samples from the target posterior distribution (1) on \mathcal{G}_p based on the above birth-death mechanism.

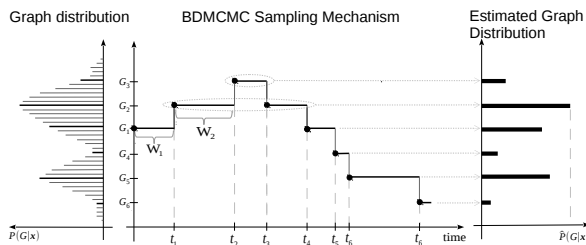


Figure 2: The left and right panels display the true and estimated posterior distribution (1) on the graph space. The middle panel reports sample scheme of Algorithm 1 where $\{W_1, W_2, \dots\}$ denote waiting times, and $\{t_1, t_2, \dots\}$ denote jumping times.

Figure 1 shows how the output from Algorithm 1 can be used to estimate posterior quantities of interest. The output consists of a set of sampled graphs and a set of waiting times $\{W_1, W_2, \dots\}$. Based on the Rao-Blackwellized estimator [1], the estimated posterior probability of each sampled graph is proportional to the expectation of the length of the holding time in that graph which is estimated as the sum of the waiting times in that graph. The posterior inclusion probability of an edge $e \in V \times V$ is estimated by

$$\hat{P}(\text{edge } e | x) = \frac{\sum_{t=1}^N \mathbb{I}(e \in G^{(t)}) W(G^{(t)})}{\sum_{t=1}^N W(G^{(t)})},$$

where N denotes the number of iterations, and $\mathbb{I}(\cdot)$ is an indicator function: $\mathbb{I}(e \in G^{(t)}) = 1$ if $e \in G^{(t)}$, and 0 otherwise. This estimation is implemented in the function `p.links` of `BDgraph`.

3 The user interface of BDgraph

We exemplify the user interface of the `BDgraph` package by analyzing the “reinis” dataset which is available in the package. The dataset consists of 6 binary variables that are potential risk factors of coronary heart disease: smoking (`smoke`), strenuous mental work (`mental`), strenuous physical work (`phys`), systolic blood pressure (`systol`), ratio of beta and alpha lipoproteins (`protein`), and family anamnesis of coronary heart disease (`family`). These data were collected from 1841 men employed in a car factory in Czechoslovakia [7].

```
R> library( BDgraph )
R> data( reinis ) # Load the data
```

Since “reinis” is a binary contingency table, we apply the Bayesian structure learning framework based on the MPL of [5] by calling the function `bdgraph.mpl` with the option `method = "dgm-binary"`. The default prior of `bdgraph.mpl` on the graph space is uniform. This is the prior we use here. We run BDMCMC for 10,000 iterations with 6,000 discarded as burn-in with the call:

```
R> sample <- bdgraph.mpl( data = reinis,
+ method = "dgm-binary",
+ algorithm = "bdmcmc",
+ iter = 10000,
+ burnin = 6000,
+ save.all = TRUE )
```

We specify the option “`save.all = TRUE`” to save all the sampled graphs in order to check the convergence of the algorithm. Running this function takes less than 1 second on a laptop computer, as the computational intensive tasks are performed in C++ in parallel. Users can obtain the adjacency matrix of the selected graph (`selected_g`) and the estimated posterior probabilities of all possible edges (`p.links`) as follows:

```
R> summary( sample )
$selected_g
      smoke mental phys systol protein family
smoke      .      .    1     1      1      .
mental      .      .    1     .      .      .
phys        .      .    .     .      .      .
systol      .      .    .     .      1      .
protein     .      .    .     .      .      .
family      .      .    .     .      .      .
```


Algorithm 1 BDMCMC algorithm

Input: A graph $G = (V, E)$ and data x
for N iterations **do**
 for all the possible edges in parallel **do**
 Calculate the birth and death rates in (2),
 end for
 Calculate the waiting time:

$$W(G) = \frac{1}{\left(\sum_{e \in \bar{E}} B_e(G) + \sum_{e \in E} D_e(G)\right)}.$$

Update G based on birth/death probabilities.

end for

Output: Samples from the posterior distribution (1).

`$p_links`

	smoke	mental	phys	systol	protein	family
[1,]	.	.	1	0.75	1.00	.
[2,]	.	.	1	.	0.11	0.06
[3,]	0.01	.
[4,]	0.99	.
[5,]
[6,]

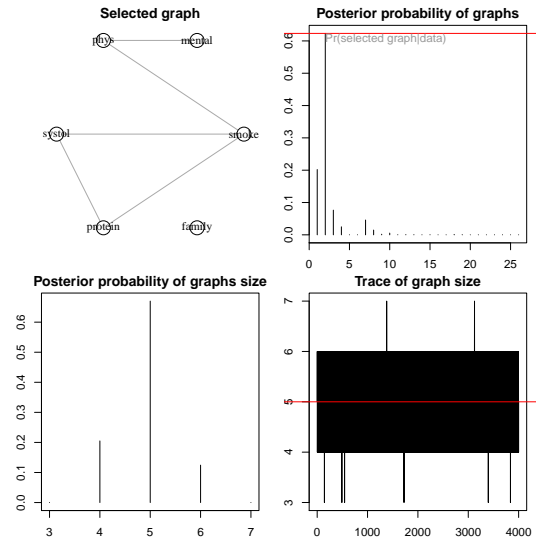


Figure 3: Visualization summary of the BDMCMC algorithm for the “reinis” data.

The function summary also generates a visualization summary – see Figure 3. The top-left panel gives the graph with the estimated highest posterior probability. The top-right panel gives the estimated posterior probabilities of all the graphs which are visited by the BDMCMC algorithm; it indicates that the algorithm visited 27 different graphs, and that the estimated highest posterior probability is around 0.63. The bottom-left panel shows the estimated posterior probabilities of the size of the graphs (number of edges); it indicates that the algorithm visited mainly graphs with sizes 4, 5 and 6. The bottom-right panel shows a trace plot based on the size of the sampled graphs.

We remark that the edges that belong to the highest posterior probability graph from Figure 3 are among the edges of the graphs determined for the “reinis” data using the `gRim` package [10], the MCMC algorithm of [13] or the loglinear model determination method of [7].

The convergence of the sampling algorithm from the posterior distribution of the graphs can be examined with the command:

```
R> plotcoda( sample )
```

This produces Figure 4. This plot shows that the BDMCMC algorithm converges after around 1000 iterations.

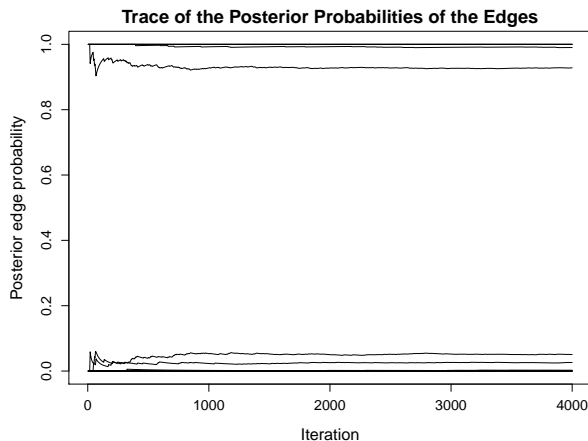


Figure 4: Trace plot showing the convergence of the the BDMCMC algorithm based on the trace of estimated posterior probability of all possible edges for the “reinis” data.

4 Analyzing a hyper-sparse high-dimensional contingency table with BDgraph

We perform structural learning in a $p = 214$ dimensional binary contingency table constructed by mapping geolocated tweets into the district municipalities of South Africa. The complete analysis of this table with BDgraph is presented in [5]. This table is hyper-sparse: only 55015 cells contain positive counts (the logarithm of the percentage of non-zero counts is -132.813). Among the 55015 non-zero counts, there are 46175 (83.93%) counts of 1, 3439 (6.25%) counts of 2, 1411 (2.56%) counts of 3, 747 (1.36%) counts of 4, and 476 (0.87%) counts of 5. The top five largest counts are 58929, 42781, 28731, 28197 and 22313.

The function `bdgraph.mpl` implements the following prior on the space of graphs [11]:

$$P(G) \propto \left(\frac{\beta}{1-\beta} \right)^{|E|}, \quad (3)$$

where $\beta \in (0, 1)$. In this application, we employ the prior (3) with $\beta = 1/\binom{214}{2} = 4.388 \times 10^{-5}$ by setting the `g.prior` argument of `bdgraph.mpl`. With this choice, the expected number of edges is 1, thus sparser graphs receive larger prior probabilities compared to denser graphs.

We run the function `bdgraph.mpl` for 10,000 iterations with 6,000 iterations as burn-in as follows:

```
R> sample <- bdgraph.mpl( data = TwitterData,
+ method = "dgm-binary",
+ multi.update = 1, g.prior= 2/(p*(p-1)),
+ iter = 10000, burnin = 6000 )
```

This function call completed after about 12 hours on a computer with an Intel Xeon 2.6 GHz processor with 48 cores, and a Linux operating system. The resulting median graph that contains the edges with estimated posterior probabilities greater than 0.5 has 1534 edges, and it is determined and visualized with the following command:

```
R> select( sample, cut = 0.5, vis = TRUE )
```

By using option `vis = TRUE`, the function `select` plots the selected graph. The function `plinks` returns the matrix with estimated posterior probabilities of all possible edges in the graph. A visualization of this 214×214 matrix is presented in Figure 5.

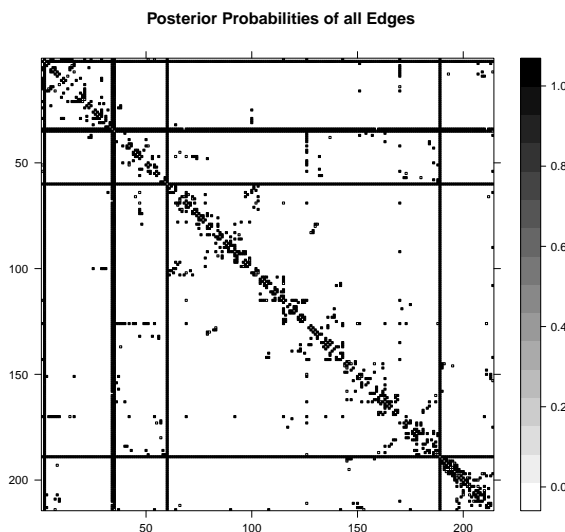


Figure 5: Heatmap of the 214×214 matrix of posterior inclusion probability of edges for the contingency table from Section 4.

5 More information

The BDgraph package is available on the CRAN at:

<https://cran.r-project.org/web/packages/BDgraph>

References

- [1] O. Cappé, C.P. Robert, and T. Rydén. Reversible jump, birth-and-death and more general continuous time markov chain monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):679–700, 2003.
- [2] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*:1695, 2006.
- [3] A. Dobra and A. Lenkoski. Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993, 2011.
- [4] A. Dobra, A. Lenkoski, and A. Rodriguez. Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433, 2011.
- [5] A. Dobra and A. Mohammadi. Loglinear model selection and human mobility. *arXiv preprint arXiv:1711.02623*, 2017.
- [6] M. Drton and M. H. Maathuis. Structure learning in graphical modeling. *The Annual Review of Statistics and Its Application*, 4:365–393, 2017.
- [7] D. Edwards and T. Havranek. A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72(2):339–351, 1985.
- [8] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- [9] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.
- [10] S. Hojsgaard, D. Edwards, and S. Lauritzen. *Graphical Models with R*. Springer-Verlag, New York, 2012.
- [11] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400, 2005.
- [12] S. L. Lauritzen. *Graphical Models*, volume 17. Oxford University Press, USA, 1996.
- [13] H. Massam, J. Liu, and A. Dobra. A conjugate prior for discrete hierarchical log-linear models. *Annals of Statistics*, 37:3431–3467, 2009.
- [14] A. Mohammadi, F. Abegaz Yazew, E. van den Heuvel, and E. C. Wit. Bayesian modelling of dupuytren disease using gaussian copula graphical models. *Journal of Royal Statistical Society-Series C*, 66(3):629–645, 2017.
- [15] A. Mohammadi, H. Massam, and G. Letac. The ratio of normalizing constants for bayesian graphical gaussian model selection. *arXiv preprint arXiv:1706.04416*, 2017.
- [16] A. Mohammadi and E. C. Wit. Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.
- [17] A. Mohammadi and E. C Wit. BDgraph: Bayesian structure learning of graphs in R. *arXiv preprint arXiv:1501.05108v4*, 2017.
- [18] M. Mohammadi and E. C. Wit. *BDgraph: Bayesian Structure Learning in Graphical Models using Birth-Death MCMC*, 2017. R package version 2.43.
- [19] OpenMP Architecture Review Board. OpenMP application program interface version 3.0, 2008.
- [20] J. Pensar, H. Nyman, J. Niiranen, and J. Corander. Marginal pseudo-likelihood learning of discrete markov network structures. *Bayesian Analysis*, 2017.
- [21] C. J. Preston. Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, 46:371–391, 1977.
- [22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

Executive Committee

President: Kerrie Mengersen
Past President: Steve MacEachern
President Elect: Marina Vannucci
Treasurer: Robert Gramacy
Executive Secretary: Amy Herring

Program Council

Chair: Clair Alston-Knox
Vice Chair: David Rossell
Past Chair: Chris Hans

Board Members:

2017–2019:
Natalia Bochkina, Catherine Forbes, Feng Liang, and Luca Tardella

2016–2018:
David Banks, Abel Rodriguez, Marc Suchard, Luke Tierney

2015–2017:
Carlos M. Carvalho, Sudipto Banerjee, Vanja Dukic, Alessandra Guglielmi

EDITORIAL BOARD

Editor

Beatrix Jones
m.b.jones@massey.ac.nz

Associate Editors

News of the World
Xinyi Xu
xinyi@stat.osu.edu

Software Highlight
Anton Westveld
anton.westveld@anu.edu.au

Interviews
Leanna House
lhouse@vt.edu