THE ISBA BULLETIN

Vol. 23 No. 3 September 2016

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE PRESIDENT

- Steve MacEachern -
ISBA President, 2016
snm@stat.osu.edu

There are many stories from the old days when there was often bitter controversy between various groups of statisticians. In the generations before me, the clashes between frequentists and Bayesians were especially rancorous, and scientific disagreement at times spilled over into personal interactions.

One of my favorite stories is the tale of an editor who took classical statistics, and in particular, the Fisherian view of hypothesis testing, too seriously. Upon reviewing an experiment, the only decisions that could be made were to "reject" or to "fail to reject". In keeping with the editor's views on statistics, he rejected many a Bayesian paper as subjective, non-scientific nonsense. A few years later, with editorial shifts, the former editor found himself submitting to his former journal, now with a Bayesian editor. The Bayesian, finding much good in the submission, but having a long memory and feeling the weight of past injustices, struggled with the decision. Should he take revenge and reject the paper as objective, non-scientific rubbish? Should he merely reject the paper without spiteful commentary? Should he accept the paper?

In this column, I write about refereeing, a topic on which many good columns have been written in recent years. Read those columns, read this one, and develop your own style of refereeing. Having the luxury of working in a large and active department, I spent some time asking colleagues, both Bayesian and not, for their views on refereeing. What appears below is a mix of their views, views I have heard in the past, and my own views.

The refereeing process is similar across most of our professional journals. Authors submit a paper, and the paper is reviewed by an editor. The editor may reject the paper at that stage or may send it to an associate editor for review. The associate editor may return the paper to the editor with a recommendation of rejection or may send it out to referees (usually two) for review. If the paper makes it to referees, the referees provide reports on the paper and a recommendation to the associate editor. The associate editor assembles the referee reports and recommendations, adds a recommendation and perhaps a report of his/her own, and passes all to the editor. The editor makes the final decision on the paper. (Continued p. 2)

In this issue

The advice given here is aimed at the referee, and in particular at those who are early in their career.

Referee for the journal. Some journals focus on a portion of statistics–perhaps computational methods, perhaps applications or applications in a particular area, perhaps probability, perhaps a topic area such as time-series or multivariate analysis–while other journals cover the breadth of the discipline. Some journals cater to a particular technical level or style of presentation. Occasionally, submissions contain fine work but are not a match for the journal. Most of these will be filtered out by the editor and associate editor, but some make it to the referees. Part of your task is to judge the appropriateness of the work for the journal. In a similar vein, various journals occupy different positions on the quality and impact scale. Higher quality journals naturally have an expectation of higher quality, and this should be part of your refereeing. A report that is on target for one journal may miss the mark for another.

Make a value judgement about the paper. Your main task as a referee is to review the submission and to then make a recommendation on the paper. For me, the primary question is whether the paper has high intrinsic value. Value comes in many forms, be it strong mathematics and a collection of relevant and interesting theorems; less relevant theorems, but the development and use of techniques which have broad applicability; creation and development of an interesting class of models; exploration of the properties of existing classes of models; development and implementation of methods for fitting models; high quality modelling on important applications; et cetera; and of course perspective that deepens our understanding of any of the above. In short, papers bring value in many different forms, and your task is to assess this value.

Conversely, I too often see papers refereed with the check-box mentality. Is the topic of the paper a well-established problem? Is the work novel? Have the authors proven a theorem? Does the method improve upon existing methods? Is the technique illustrated with a simulation? Do the authors include a real-data example? Is the paper passably written? Have the authors cited the literature appropriately? While these are good questions to ask and are appropriate for comment in a report, they are secondary questions.

Referee with an open mind. A common complaint from authors is that the referees have not given their paper fair consideration. Often, the perception is that referees have not looked for value in the paper, but have instead looked for reasons to reject the paper. This is sometimes tied to the check-box mentality of refereeing. At other times, it may be due to competitiveness. As a referee, you will most commonly be asked to referee a paper in an area you have worked in. If the authors present a technique that rivals your own, it is essential that you give the submission a fair shake. There is a good chance that the authors are addressing a different aspect of the problem than your work does. Rather than evaluating their technique through your own personal perspective on the problem, your task is to understand the authors' perspective and to make a recommendation on that basis. Your judgements should ideally be based on whether the work brings value to the statistics community, not whether you personally like or dislike the method.

Decision theory has useful lessons for assessing the value of a method. Many papers develop a method and compare it to others, perhaps through simulation. For a pair of fairly good methods, decision theory tells us that their risk functions will cross, with neither method dominating the other. The question is not whether one method is better than another, but under which circumstances one method is better and how large the difference in performance is. It is ridiculous to reject a paper because the method developed in the paper is not uniformly best. More subtly, when a top quality method is compared to several methods, it may not show itself to be the top performer in any of the simulation settings. A method with excellent, but not best, performance across a wide range of settings may be the ultimate winner. This perspective is broadened when one considers questions of implementation: ease of using/modifying the method, computational speed and accuracy, and diagnostics to assess the appropriateness of the method. In short, methods bring value far beyond "best in some simulation." Refereeing with this view encourages better research, as authors have less incentive to limit their simulations to conditions under which their method wins the comparison.

Let the authors write the paper. The authors are, well, the authors. It is their paper, not yours. You may view the material in a different fashion than the authors do. While it is fine to convey your view in your report (the authors may well appreciate this), avoid the dogmatic view that, for the paper to be publishable, the authors must replace their view with yours. One contribution of the paper may be the expression of a novel

view on the material. This view also cascades into choice of content. The authors may develop the ideas that drive the paper in a different fashion than you would. A different development is merely a different paper and is not by itself a reason for rejection.

There are many different styles of writing, and these lead to papers with a different look and feel. The authors' style may differ from your preferred style–perhaps with more text and richer language, perhaps a more technical presentation. Similarly, the organization of the material may differ from your preferred organization. Your task is, once again, to set aside your own preferences in favor of a broad-minded community judgement. Commentary on the differences is fine. Heavy-handed "you didn't write the paper I would have, so reject" is not.

Check the details. Like it or not, part of your job as a referee is to check the details of the work. This is easier for theoretical work, where the mathematics needed to establish results appears in the paper. For me, I first assess the plausibility of the results and then look at the details. I also take this approach when refereeing computational work–for MCMC algorithms, does the algorithm seem like it should mix well? Does it conform to the folklore on how what produces effective algorithms? If not, does the paper explain why? Do simulations and other illustrations of the method seem to be well done? Are the details of conditional distributions (if given) accurate?

Applications papers are arguably the most difficult to referee because so little of the work appears in the paper. For quality applied work, the authors have made many, many decisions that will be almost invisible in the paper: exploratory data analysis to pick up the large-scale patterns, data cleaning and repair, investigation of numerous alternative models, sensitivity of the results to choice of prior distribution, and so on. The work often has the feel of the classic paper in the nonstatistics literature. The bulk of the paper focuses on the scientific implications of the study while there may be only a page or two on methods and results. Your task as a referee is difficult, as the paper won't contain all of the information you need to assess the quality of the analysis. However, there will often be residual signs of high or low quality work. Sloppiness of presentation and overly aggressive claims of the value of the work often go hand-in-hand with lesser quality.

Although this jumps ahead, the authors also have a duty when it comes to accuracy of their work. Authors should submit only after believing all to be correct. Submitting a wild conjecture as a theorem is poor form. Many years ago, I tracked the error rate in papers I saw in the refereeing process. Roughly 2/3 of submissions that I saw had substantial errors, and about half of those were unfixable. I believe the error rate that I saw was somewhat higher than the norm, as many of the papers were in the area of nonparametric Bayes. Troubles included unwitting use of improper posterior distributions, misleading statements about limits, MCMC algorithms with the wrong limiting distribution, incorrect handling of mixed continuous and discrete calculations, and so on. So why the big error rate? The models were at the time difficult to understand, the intuition and mathematics different than for finite-dimensional problems, and MCMC new and poorly understood. Much of research is conducted in this environment of true novelty, as this environment is where many of the big breakthroughs occur. This increases pressure for authors to submit quickly or even prematurely. While submitting early and often has evident benefits for the vita, the hope is that the community tracks when work was done, who did the work, and understands that it is common for different groups to do similar work contemporaneously. The hope is also that the community understands, through time, which researchers are more committed to getting things right.

Reflect on your report. For Bayesians, this is natural: upon reflection, all of those who work with probability models become Bayesian, as $X|\theta$ becomes $\theta|X$. It takes a considerable amount of time to referee a paper well. If you look at the paper reasonably soon after agreeing to referee it, you will have the time to write a draft report and return to it before completing your final report. Do so. For me, assessing the value of a paper often takes a few days of living with its ideas.

This period of reflection will also give you a chance to rethink what you are asking of the authors. When recommending that the authors do extra work for a revision, be judicious in your requests. It is difficult to compare a method to all existing methods, and implementation of some methods would take an enormous investment of time. Ask yourself whether a particular comparison would add value to the paper, and if so, whether the value would justify the time investment on the part of the authors. The same comment holds for requests for additional simulations.

Make a clear recommendation to the Associate Editor. Your final task as a referee is to make a clear recommendation to the associate editor. In addition to the report for the authors, there should be a clear statement directed to the associate editor. The statement should be more than a recommendation of accept or reject. More valuable is a brief summary of the reasoning behind your recommendation. Along with this, you should let the associate editor know which of your comments to the authors must be addressed in a revision and which are optional for revision. Finally, let the associate editor know if you have read quickly over some parts of the paper. In all, you are providing both feedback to the authors and a recommendation on how to proceed to the associate editor.

In a similar spirit, associate editors should be more than a rubber stamp, passing referee reports along to the editor. A good associate editor's letter will synthesize the referee reports and accompanying letters and will contain the associate editor's own commentary on the paper. This is then passed to the editor who makes the final call.

Speed. The typical refereeing advice column emphasizes speed, speed, and more speed. In my mini-survey, few commented on the need for quicker reviews, and then it was relatively low on the list of comments. Timeliness of refereeing is unquestionably important (with these words coming from me, eye rolls of a few editors and associate editors are noted). But, in my opinion, it is not worth sacrificing the quality of the report to produce it a little more quickly. As an associate editor, I would much rather see a quality report after four months than a cursory report in one month. As an author, I am comfortable with longer delays in refereeing, as long as referees and associate editor have read and understood the submission and have rendered their judgement on it.

Authors. For the system to work, authors must also do their part. The authors' responsibilities parallel those of the referee. Authors should make a value judgement on their own work, submitting to a journal which is a match for the work's value and style. They should write with an open mind, describing both strengths and weaknesses of their work. They should put in the time to polish their writing, ensuring decent grammar, spelling, and notation. And the work should be accurate. Scientific honesty is essential. On this last point, if you have a theoretical result with proof, present it as a proposition or theorem; if without proof, present it as a conjecture. Do not try to sneak it through the refereeing process as a theorem. For computational algorithms, know whether you are fitting the model you're attempting to fit. For models, respect the important features of the data and the context from which the data arise. As always, expect to do far more work than will actually appear in the paper.

Returning to the no-doubt apocryphal story with which this column began, the Bayesian editor made a decision with which the frequentist was delighted. He accompanied the positive reviews of the paper with his editorial comment–that, while the paper was not worthy of rejection, in line with the previous editor's views on statistics, he could not, in good conscience, accept the paper. Rather, in spite of his best efforts, he had failed to reject it.

I can only hope that this bulletin's Editor, Beatrix Jones, being a good Bayesian, will not only fail to reject my column, but will in fact accept it.

–Steve MacEachern

A MESSAGE FROM THE EDITOR

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

Thanks to Steve for the endorsement of my Bayesian credentials! One of the more memorable commentaries on the review process that I have heard was a talk by George Casella. In particular I recall the slide heading "Sure, the referees are brain dead monkeys, but..." which was followed by some advice on how to use referee's comments–including misunderstandings–to improve one's argument and the clarity of the paper. Ironically, I now think of this turn of phrase not (usually) when responding to referees, but when acting as a referee myself, trying to produce a report the day before (or the day after) the deadline. But let me assure you it never comes to mind when dealing with my dear *Bulletin* contributors.

Speaking of this, my thanks go to Isadora Antoniano who has been associate editor for the interview section, following on from a long stint with Students' Corner. However, she is taking on

some new challenges in her "real" job, and we are seeking a new AE for the interview section. This role consists of arranging for one Bayesian to interview another—the results are invariably fascinating, as you will have seen in the last issue when Manuel Mendoza was interviewed by Eduardo Gutiérrez-Peña–but the initial matchmaking can take some persistence. Anyone who is interested, or who would like to recommend a colleague for the role, can contact me at the email address above.

As well as our usual features, this issue includes some important information on our 2016 ISBA elections, and a Software Highlight featuring the BFA (Bayesian Factor Analysis) package by Jared Murray, which I can attest is a pleasure to use. Enjoy!

2016 ISBA ELECTION

This year's ISBA nominating committee consisted of Lurdes Inoue, Jaeyong Lee, Peter Mueller, Raquel Prado, Judith Rousseau, Fabrizio Ruggeri, and chair Alexandra M. Schmidt. The committee is pleased to announce the candidates in the 2016 election. Thanks to all these individuals for agreeing to stand. Candidate statements will appear on the website in due course.

President Elect:

Marina Vannucci marina@rice.edu
http://www.stat.rice.edu/~marina

Igor Pruenster igor@unibocconi.it
http://mypage.unibocconi.eu/igorpruenster/

Treasurer:

Robert Gramacy rbg@vt.edu
http://bobby.gramacy.com

Fan Li fli@stat.duke.edu
http://stat.duke.edu/people/fan-li

Board: (4 positions)

Angela Bitto angela.bitto@wu.ac.at
https://www.wu.ac.at/en/statmath/faculty-staff/faculty/abitto/

Natalia Bochkina N.Bochkina@ed.ac.uk
http://www.maths.ed.ac.uk/~nbochkin/

Thais Fonseca thaisf@gmail.com
https://sites.google.com/site/thaisf/

Catherine Forbes Catherine.Forbes@monash.edu
http://monash.edu/research/explore/en/persons/catherine-forbes (2f41dc8e-880f-412d-9867-f8c9bdfe5092).html

Feng Liang liangf@illinois.edu
https://publish.illinois.edu/liangf/

Manuel Mendoza mendoza@itam.mx
http://allman.rhon.itam.mx/~mendoza/mendozaeng.html

Mike So immkpso@ust.hk
http://www.bm.ust.hk/ismt/staff/immkpso.html

Luca Tardella luca.tardella@uniroma1.it
http://www.dss.uniroma1.it/en/node/5700

CONFERENCE REPORT: TIES

EnviBayes and The International Environmetrics Society (TIES) have a long history of collaboration. This year TIES conference held in Edinburgh July 18th-22nd, saw a large and prestigious participation of EnviBayes member both as presenters and organisers (see details at http://www.ed.ac.uk/maths/international-environmetrics-society/about-the-ties-conference/). The conference was a real success in terms of scientific exchange, high level talks and fun. Following this tradition next year TIES conference will see again a strong collaboration with EnviBayes. The event will be held in Bergamo (Italy) 24-26 July 2017 (www.graspa.org/tiesgraspa2017), and it will be a joint meeting of TIES and the Italian environmental statistics research group GRASPA, a section of the Italian Statistical Society (SIS) holding its biennial meeting. The conference will also be a satellite event of the ISI 2017 world conference (http://www.isi2017.org/). The main title of the Bergamo meeting is "Climate and Environment'," scientific and local committees are in progress including members of TIES, GRASPA and EnviBayes.

For future updates on the conference please check the GRASPA web site www.graspa.org/tiesgraspa2017

ENVI-BAYES AWARDS

The EnviBayes section of ISBA this year has granted two best posters awards at the ISBA World Conference in Forte Village (June 13th - 17th, Cagliari, Italy). Awards were granted considering the following set of criteria

1. Relevance to Environmental Science
2. Use of Bayesian Methods
3. Novelty of the proposals.

The winners, out of 29 posters on environment related subjects, were the posters:

A Causal Inference Approach for Estimating an Exposure Response Curve: Estimating Health Effects at Low Pollution Levels by Georgia Papadogeorgou, Phd student at the department of Biostatistics Harvard University (United States). The poster is joint work with Francesca Dominici from the same department. The poster was on display on June 15th.

ABSTRACT: Many methods have been developed to estimate a potentially non-linear exposure response (ER) curve, while accounting for known observed confounders. However, none of these approaches account for the possibility that estimation of the causal effects at low exposure levels might be affected by a different set of confounding variables than estimation of the causal effects at higher exposure levels. Also, none of the existing approaches account for the fact that there is uncertainty regarding which confounders should be included into the model, especially when the number of confounders is large compared to the sample size. Furthermore, it is often the case that the sample size at extreme exposure levels is significantly smaller than at average exposure. Extrapolation and estimation of the ER curve at extreme exposure levels using information from normal levels can lead to significant bias in the estimation of causal effects. Such a situation is met in the study of the health effects of low ambient air pollution. While a lot of information exists for areas of average air pollution, we would like to estimate the causal effect of ambient air pollution at low levels, while using the information of all exposure levels to gain power. Our approach borrows information across exposure levels to identify the important confounding variables at each level separately. Using this information, we estimate the whole ER curve, which will have a causal interpretation, while accounting for the uncertainty in confounder selection at each level of exposure.

Georgia Papadogeorgou, is a 4th year PhD student in Biostatistics at the Harvard T.H. Chan School of Public Health working under the supervision of Dr. Francesca Dominici and Dr. Corwin Zigler. She graduated from the University of Athens, where she studied theoretical and applied mathematics. Her research focuses on the de-

velopment of statistical methods for causal inference, with applications on environmental health science, health policy impact evaluations, and comparative effectiveness research.

Joint Species distribution modeling: dimension reduction using Dirichlet processes by Daniel Taylor-Rodriguez, Post-doc at the Department of Statistical Science Duke University (United States); joint work with Kimberly Kaufield from North Carolina State University, Erin Schliep of University of Missouri, James Clark and Alan Gelfand of Duke University. The poster was on display on June 16th.

ABSTRACT: The primary tool in ecology to learn about where species are and why, is a species distribution model. Historically, such models have been specified individually across species. While marginal models can provide useful information regarding distribution and abundance, they ignore the fact that the distribution and abundance of species is a joint process which involves modeling species simultaneously (e.g., through competition, mutualism, etc.) rather than an independent one for each. As a result, collectively, misleading behaviors, may arise. In particular, individual models often imply too many species per location. Recently, there has been activity in building joint species distribution models. Such models have application to presence-absence, continuous or discrete abundance, abundance with large numbers of zeros, and discrete, ordinal, and compositional data. Here, we deal with the challenge of joint modeling for a large number of species. To appreciate the challenge in the simplest way, with just presence/absence (binary) response and say S species, we have an S-way contingency table with 2^S cell probabilities. Even if S is as small as 100, this is an enormous table, infeasible to work with without some structure to reduce dimension. We develop a computationally feasible approach to accommodate a large number of species (say order 10^3) that allows us to: 1) assess the dependence structure across species; 2) identify clusters of species that have similar dependence patterns; and 3) jointly predict species distributions. To do so, we build hierarchical models capturing dependence between species at the first or "data" stage rather than at a second or "mean" stage. We employ the Dirichlet process for clustering in a novel way to reduce dimension in the joint covariance structure. This last step makes computation tractable. We use Forest Inventory Analysis (FIA) data in the eastern region of the United States to demonstrate our method. It consists of presence-absence measurements for 112 tree species, observed on hectare size plots east of the Mississippi. As a proof of concept for our dimension reduction approach, we also include simulations using continuous and binary data.

Daniel Taylor-Rodriguez Daniel is a Postdoctoral Associate at the Statistical and Applied Mathematical Sciences Institute (SAMSI) and at Duke University. His appointment in SAMSI is in connection with the year-long research program on Mathematical and Statistical Ecology, and at Duke University he has been working with Alan Gelfand and Jim Clark's lab. He obtained

his PhD at the University of Florida in Interdisciplinary Ecology with a concentration in Statistics under the guidance of George Casella, Linda Young and Nikolay Bliznyuk. He is interested in Bayesian selection, estimation and prediction strategies, spurred by methodological challenges found in ecological applications.

The EnviBayes community congratulates the two winners and look forward to see the final publications following on from these early results. In addition, the EnviBayes section would like to congratulate its former chair, Bruno Sanso, from the Department of Applied Mathematics and Statistics of the University of California Santa Cruz who was elected an ISBA fellow during the ISBA World Conference.

FROM THE PROGRAM COUNCIL

- CHRIS HANS -
CHAIR OF THE PROGRAM COUNCIL
program-council@bayesian.org

ISBA at NIPS 2016: ISBA supports initiatives that highlight the importance and impact of Bayesian methods related to current challenges in machine learning and data science. Following the success of the ISBA at NIPS initiatives in 2014 and 2015, this year the Program Council endorsed four proposals for post-conference workshops related to Bayesian methods at NIPS (Neural Information Processing Systems) 2016 https://nips.cc/Conferences/2016. We are pleased to report that three of those four proposals were successful, which means that Bayesian methods will once again be featured prominently at NIPS. The workshops, to be held December 9-10, 2016 in Barcelona, Spain immediately following the NIPS main conference, are:

- *Advances in Approximate Bayesian Inference* Organizers: Tamara Broderick (MIT), Stephan Mandt (Disney Research), James McInerney (Columbia) and Dustin Tran (Columbia) http://approximateinference.org/.
- *Bayesian Deep Learning* Organizers: Yarin Gal (U. of Cambridge), Christos Louizos (U. of Amsterdam), Zoubin Ghahramani (U. of Cambridge), Kevin Murphy (Google) and Max Welling (UC Irvine) http://bayesiandeeplearning.org/.
- *Practical Bayesian Nonparametrics* Organizers: Tamara Broderick (MIT), Trevor Campbell (MIT), Nicholas Foti (U. of Washington), Michael Hughes (Harvard), Jeffrey Miller (Harvard), Aaron Schein (UMass Amherst), Sinead Williamson (UT Austin) and Yanxun Xu (Johns Hopkins) https://sites.google.com/site/nipsbnp2016/.

We encourage individuals attending NIPS to considering participating in the workshops! See the links above for further information.

ISBA at NIPS Travel Awards: As part of the 2016 ISBANIPS initiative, ISBA will provide two ISBA@NIPS Travel Awards to early-career researchers presenting research in the ISBA-endorsed workshops. Qualifying individuals will be nominated by the workshop organizers to the ISBA Program Council, who will select the award recipients. We look forward to reporting

the names of the award winners to the ISBA membership in the December issue of the Bulletin!

Upcoming ISBA events in 2016: In addition to the ISBA-endorsed workshops at NIPS 2016, we would like to highlight the following upcoming meeting that is being co-sponsored by ISBA:

The 10th ICSA International Conference on Global Growth of Modern Statistics in the 21st Century http://www.math.sjtu.edu.cn/conference/2016icsa/Default.aspx, December 19-22, 2016, Shanghai, China.

UPDATE FROM BA

From the BA Editor

- Bruno Sansó -
bruno@soe.ucsc.edu

The Joint Statistical Meetings took place this summer in Chicago from July 30 to August 4. Bayesian Analysis was present with an invited session that highlighted some of the publications in the journal during the previous year. We had a lively session with four invited speakers: Jim Berger, presenting his paper, coauthored with Dongchu Sun and Jose Miguel Bernardo on "Overall Objective Priors"; Phil Dawid presenting his paper, coauthored with Monica Musio on "Bayesian Model Selection Based on Proper Scoring Rules"; Gustavo da Silva Ferreira, who wrote a discussion paper with Dani Gamerman on "Optimal Design in Geostatistics Under Preferential Sampling", and Brian Phillip Weaver, who presented his paper on "Computational Enhancements to Bayesian Design of Experiments Using Gaussian Processes". We are working to make the BA invited session happen at the JSM in Baltimore next year.

I have previously commented on the September issue of the journal, that is now available in full at https://projecteuclid.org/euclid.ba, including the discussion and the rejoinder of the invited paper by Pratola. The paper attracted quite a lot of attention, demonstrated by the fact that we received five contributed discussions. The December issue is already online in a preliminary form. It will be completed in the next months with the discussion paper "Bayesian Solution Uncertainty Quantification for Differential Equations" by Oksana A. Chkrebtii, David A. Campbell, Ben Calderhead, and Mark A. Girolami. It will feature discussions by Sarat Dass, Martin Lysy and Bani Mallick. This paper is already available as advanced publication. You are all welcome to contribute a discussion.

Next year BA will publish a series of review papers, one for each of the thematic sections of ISBA. The idea is to showcase the state of the art for some of the most important topics that are of interest for our sections. The authors that have committed, for now, to write a review are Sudipto Banerjee, for the environmetrics section; Herman van Dijk for the econometrics section and Nicolas Chopin for the section on Bayesian computations.

NEWS FROM THE WORLD

ISBA co-sponsored meetings and conferences

The 10th ICSA International Conference, Shanghai, China. December 19-22, 2016.

The 10th ICSA International Conference will be held at Xuhui campus of Shanghai Jiao Tong University (SJTU), Shanghai, China, during December 19-22, 2016. The theme of this conference is to promote global growth of modern statistics in the 21st century. The purpose of this conference is to bring statisticians from all over the world to Shanghai, China, which is the financial, trade, information and shipping center of China, to share cutting-edge research, discuss emerging issues in the field of modern probability and statistics with novel applications, and network with colleagues from all parts of the world.

The scientific program committee of the 2016 ICSA International Conference is co-chaired by Ming-Hui Chen of University of Connecticut, Zhi Geng of Peking University, and Gang Li of Uni-

versity of California at Los Angeles. James O. Berger of Duke University, Tony Cai of University of Pennsylvania, Kai-Tai Fang of Beijing Normal University - Hong Kong Baptist University United International College (UIC), Zhi-Ming Ma of the Academy of Math and Systems Science, CAS, Marc A. Suchard of the UCLA Fielding School of Public Health and David Geffen School of Medicine at UCLA, Lee-Jen Wei of Harvard University, and C. F. Jeff Wu of Georgia Institute of Technology will deliver keynote presentations. There will be a special session in honor of the receipt(s) of the second Pao-Lu Hsu award. In addition, there will be ample of invited and contributed sessions. All participants including invited speakers are responsible for paying registration fees and booking hotel rooms directly from the hotels listed on the conference website.

For conference logistics, please directly contact Dong Han and Weidong Liu, the co-chairs of the local organizing committee. All inquiries should be sent to Ms. Limin Qin at qinlimin@sjtu.edu.cn. Please visit the conference website http://www.math.sjtu.edu.cn/conference/2016icsa/ for more detailed information. All of you are welcome to participant in this important ICSA conference and to visit Shanghai, one of the most beautiful and historic cities in the world, during December 19-22, 2016.

Bayesian Nonparametrics Conference 26th - 30th June, 2017, Ecole Normale Supérieure, Paris. Please visit the workshop site https://www.ceremade.dauphine.fr/~salomond/BNP11/index.html for further information.

Abstract submission will be open in October.

The Bayesian nonparametrics (BNP) conference is a bi-annual international meeting bringing together leading experts and talented young researchers working on applications and theory of nonparametric Bayesian statistics. It is an official section meeting of the Bayesian nonparametrics section of the International Society for Bayesian Analysis. Past conferences can be found at : https://bayesian.org/sections/BNP/bnp-past-isba-workshops

STUDENTS' CORNER

Shinichiro Shirota
ss571@stat.duke.edu

In this third issue, I introduce a PhD candidate in Statistical Science, Duke University, Ksenia Kyzyurova. She is collaborating with Prof. Jim Berger and Prof. Robert Wolpert. In addition to introducing new researchers, this Students' Corner also features the dissertation abstracts. Issuing abstracts would provide a good opportunity to find collaborators. If you are willing to, don't hesitate to send your dissertation abstract to my email address.

Student Voices

Ksenia Kyzyurova
ksenia@stat.duke.edu

I thank Shinichiro Shirota for giving me the opportunity to present my research in the ISBA Bulletin. I am a fifth year Ph.D. candidate in the Department of Statistical Science at Duke University. Before coming to Duke I received BSc and MSc in Applied mathematics and computer science from ITMO University in St. Petersburg, Russia.

I am lucky to work on my dissertation under the supervision of my adviser Jim Berger. The main focus of my research is statistical approximation and modeling of complex processes. Modern science often requires combining several pieces of information to understand the behaviour in complex composite systems. Such systems are described by mathematical models, usually a system of differential equations, and are solved numerically, resulting in so-called computer models. Emulators are approximations of computer models. Gaussian processes, together with an objective Bayesian implementation of the processes, have become a common tool for emulating complex computer models. Sometimes more than one computer model needs to be utilized for the predictive goal. For instance, to model the true danger of a volcano pyroclastic flow, one might need to combine the flow model (which can produce the flow size and force at a location) with a computer model that provides an assessment of structural damage, for a given flow size and force.

Together with Jim Berger and Robert Wolpert

we have developed the methodology on how to circumvent deterministic coupling of computer models by instead linking their emulators. Direct coupling of computer models is often difficult for computational and logistical reasons (which mainly comes from the fact that they take a lot of time to run: hours, days, or even weeks). We have proposed coupling two computer models by linking independently developed Gaussian process emulators (GaSPs) of these models. We call the resulting emulator of the composite system the linked emulator. We found that in practice the linked emulator results in a smaller epistemic uncertainty than a direct GaSP of the directly coupled computer model would have (if such a model were available). To demonstrate the performance of the linked emulator we have applied the methodology to scientific computer models of volcano pyroclastic flows, volcano eruption columns and volcano ash transport and dispersal model within the long-term collaboration with Elaine Spiller from Marquette University and Bruce Pitman, Abani Patra and Marcus Bursik from the State University of New York at Buffalo.

This research opened up the opportunity for me to collaborate with the Los Alamos National Laboratory (LANL) Statistical Sciences group. This summer during my internship at LANL, together with my mentor, scientist Jim Gattiker, and scientist Sham Bhat, we have extended the methodology of the linked emulator to calibration in systems of computer models; for example, engineering systems for carbon sequestration (part of the Department of Energy carbon capture simulation initiative). The linked emulator allows for development of independent emulators of submodels on their own separately constructed design spaces. This property leads to the advancement of the method for high-dimensional Bayesian inverse problems in the framework of a system of computer models, which may be too computationally expensive otherwise. We demonstrate how calibration works in such a framework for vapour-liquid equilibrium model for the amine-based CO2 capture solvent. I'm grateful to Jim and Sham, and also to Peter Marcy and Troy Holland who helped me understand the problem from statistical, physical, engineering and bureaucratic aspects along the way.

The flexibility and speed of linking emulators allows for development of emulators of heterogeneous and hybrid models. For example, two computer models may have different physical description, or one model may represent a natural (physical) process, and another a social sciences model of a species behaviour. Linking emulators of computer models is an example how remarkably statistics may contribute not only to scientific fields, such as geophysics, engineering systems, astronomy, climate modeling, but also to social sciences, for example, urban and rural development. In my future endeavors I would like to develop and maintain long-lasting great collaborative relationships interdisciplinary and within Bayesian community.

SOFTWARE HIGHLIGHT

BAYESIAN FACTOR ANALYSIS IN R: GAUSSIAN, PROBIT, AND GAUSSIAN COPULA FACTOR MODELING WITH BFA

Jared S. Murray
Carnegie Mellon University
jsmurray@stat.cmu.edu

Latent factor modeling is a useful tool for capturing and understanding dependence in multivariate data. Factor models originated in the social sciences, but have since seen application in areas ranging from genomics to finance. The August 2003 edition of the ISBA Bulletin contains a wonderful annotated bibliography of factor models (compiled by Hedibert Lopes). This note highlights the variety of factor models and associated priors implemented in the `R` package `bfa`, as of version 0.4.

Models implemented in `bfa`

Three classes of factor models are implemented in `bfa`: Gaussian factor models (via the function `bfa_gauss`), mixed Gaussian and probit factor models (`bfa_mixed`), and Gaussian copula factor models (`bfa_copula`).

Gaussian factor models

The most basic model implemented in `bfa` is the Gaussian factor model, given by

$$y_i = \mu + \Lambda\eta_i + \epsilon_i \tag{1}$$

where y_i is a $p \times 1$ vector of observed variables, Λ is a $p \times k$ matrix of factor loadings ($k < p$), $\eta_i \sim N(0, I)$ is a $k \times 1$ vector of latent variables or factor scores, and $\epsilon_i \sim N(0, \Sigma)$ are idiosyncratic disturbances with $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. This is an "exploratory" factor model, to be distinguished from "confirmatory" factor models that posit additional structure in Λ and allow correlation between the latent factors η_i. Confirmatory factor models are not implemented in `bfa`, although it is possible to include some structure in Λ (in the form of zeros and sign constraints).

Marginalizing over η_i yields

$$y_i \sim N(\mu, \Lambda\Lambda' + \Sigma). \tag{2}$$

The model in (1) may reflect a true belief that unobservable factors drive all the observed covariance in y_i, or it may arise simply as a convenient form of data augmentation for (2). The model in (2) is well-motivated even in the absence of "true" latent factors, as it provides a regularized estimate of $\text{Cov}(y_i)$ through the low-rank-plus-diagonal form imposed upon the covariance matrix (see e.g. Bhattacharya and Dunson (2011); Puelz et al. (2016) for applications where regularized estimates of the covariance matrix are of primary interest).

Gaussian-probit factor models for normal and ordinal variables

For data that include binary or ordered categorical variables in addition to Gaussian variables a simple extension is to use a mixed Gaussian-probit model. That is, for a $p \times 1$ dimensional vector z_i we assume that

$$z_i = \mu + \Lambda\eta_i + \epsilon_i \tag{3}$$

with $y_{ij} = z_{ij}$ if the j^{th} variable is continuous, or

$$y_{ij} = \sum_{c=1}^{C_j} c1(\gamma_{jc-1} < z_{ij} \leq \gamma_{jc}) \tag{4}$$

if the j^{th} variable takes ordered values in $\{1, 2, \ldots C_j\}$. The marginal distribution of variable j is parameterized by the ordered collection of "cutpoints" $\gamma_{j0}, \ldots \gamma_{jC_j}$ (with $\gamma_{j0} = -\infty$ and $\gamma_{jC_j} = \infty$. For continuous variables we impose $\sigma_j^2 = 1$ and $\mu_j = 0$ for identifiability. Quinn (2004) provided a Bayesian treatment of this model, and Hahn et al. (2012) introduced a sparse variant for binary observations.

In Murray et al. (2013) we argue that when the idiosyncratic variances σ_j^2 are artificially constrained for identifiability, inference should be based on the scaled loadings

$$\tilde{\lambda}_{jh} = \frac{\lambda_{jh}}{\sqrt{1 + \sum_{h=1}^{k} \lambda_{jh}^2}} \tag{5}$$

so that the correlation between variables j and j' is $c_{jj'} = \sum_{h=1}^{k} \tilde{\lambda}_{jh}\tilde{\lambda}_{j'h}$. Without scaling the factor loadings are not otherwise comparable across the different variables – marginalizing over η_i, the latent variables in z_i are on implicitly different scales since $\text{Cov}(z_{ij}) = 1 + \sum_{h=1}^{k} \lambda_h^2$. Unlike λ_{jh}, $\tilde{\lambda}_{jh}$ is a scale-free measure of variable j's contribution to factor h. This is important for probit models and also for the copula models introduced in the next subsection.

Gaussian copula factor models for general continuous and ordinal variables

In Murray et al. (2013) we extended Gaussian and mixed Gaussian-probit factor models models to collections of arbitrary continuous and ordinal manifest variables using the following copula model:

$$\eta_i \sim N(0, I), \quad z_i | \eta_i \sim N(\Lambda\eta_i,\ I)$$
$$y_{ij} = F_j^{-1}\left(\Phi\left(\frac{z_{ij}}{\sqrt{1 + \sum_{j=1}^{K} \lambda_{jh}^2}}\right)\right) \tag{6}$$

where F_j^{-1} is the pseudo-inverse of the cdf for the j^{th} variable. The Gaussian copula factor model subsumes the Gaussian and mixed Gaussian-probit factor models: when the j^{th} variable is discrete, F_j^{-1} can be described using a collection of cutpoints as in the previous section, and choosing F_j^{-1} to be the quantile function of a normal distribution it is possible to recover Gaussian margins with the appropriate mean and variance. In general, a Gaussian copula factor model is implied by the belief that the data arise as transformations of latent Gaussian random variables that follow a Gaussian factor model. This belief is implicit in the common practice of attempting to monotonically transform individual continuous variables to normality before fitting a factor model.

Often one is primarily interested in characterizing the dependence structure among the observed variables. In this case the collection of marginal distributions (or equivalently, a collection of transformations to joint normality) is a complex, infinite-dimensional nuisance parameter. Hoff (2007) provided an approach for approximate Bayesian inference in this setting using the "extended rank likelihood". Murray et al. (2013) employed the extended rank likelihood for fitting the model in (6) without attempting to infer the marginal distributions F_j, and this is the approach implemented in `bfa`.

Prior distributions

Several prior distributions are available for Λ. These are listed below, along with arguments used to set their parameters in the `bfa_*` functions. Note that all available priors implicitly assume that any manifest Gaussian variables are on the same scale.

- $\lambda_{jh} \sim N(0, b)$ for a fixed b: `prior="normal", loadings.var=b`
- $\lambda_{jh} \sim GDP(a, b)$, where GDP is the generalized double Pareto distribution (Armagan et al., 2013): `prior="gdp", gdp.alpha=a, gdp.beta=b`
- $\lambda_{jh} \sim (1 - \pi_h)\delta_0 + \pi_h N(0, b)$ where δ_0 is a point-mass at zero, and $\pi_h \sim Beta(c, d)$: `prior="pointmass", loadings.var=b, rho.a=c, rho.b=d`

GDP(3,1) priors were recommended by Murray et al. (2013) when scaled loadings are of interest (probit/copula factor models). Normal priors can induce priors on the scaled loadings that have undesirable behavior, especially when the prior variance is large, and should be avoided in that setting.

It is also possible to introduce factor-specific loading scale parameters τ_h with $Gamma(a_\tau, b_\tau)$ priors on $1/\tau_h$ (so that $E(1/\tau_h) = a_\tau/b_\tau$). With these scale parameters the new loadings are given by $\tilde{\lambda}_{jh} = \tau_h^{1/2}\lambda_{jh}$, where the prior on λ_{jh} can be chosen from the list above. These scale parameters can be included by specifying `factor.scales = TRUE, tau.a = a_tau, tau.b=b_tau` in any of the `bfa_*` functions. (Note that `loadings.var` and `tau.b` are redundant, and only one should be specified.) The scale parameters are implemented using the redundant parameterization and transformation approach of Ghosh and Dunson (2009). (The exact model proposed by Ghosh and Dunson (2009) is recovered by using `factor.scales=TRUE` and `prior="normal"` with $a_\tau = b_\tau = 0.5$.)

The loadings matrix can also include positivity restrictions and elements fixed at zero. These are necessary if interest is in Λ or the factor scores (as opposed to the covariance $\Lambda\Lambda' + \Sigma$) as the likelihood in all three models is invariant to rotations of the latent factors. These restrictions can be supplied in two ways: The first is a matrix R of the same size of Λ where $r_{jh} = 0$ if $\lambda_{jh} \equiv 0$, $r_{jh} = 1$ if λ_{jh} is unrestricted, or $r_{jh} = 2$ if $\lambda_{jh} > 0$. The helper function `utri_zero` generates a matrix in this format that encodes the zero upper triangle and positive diagonal identification restriction given in Geweke and Zhou (1996). The second is as a list of triples, where each list element has the format `c("variablename", h, ''0'')` where `"variablename"` is the name of the variable to be constrained in factor (column) h, and valid restrictions are `"0"` or `">0"`.

The idiosyncratic variances in Gaussian factor models (or for the Gaussian variables in a mixed factor model) have independent inverse Gamma priors: $1/\sigma_j^2 \sim Gamma(a_\sigma, b_\sigma)$, where the prior parameters can be set using `sigma2.a=a_sigma, sigma2.b=b_sigma`. Again, since all the idiosyncratic variances have the same prior it is generally advisable to have all the Gaussian variables on the same scale. The means μ_j for the Gaussian variables have independent normal priors with mean `mu.mean` and variance `mu.var`. Mixed factor models use the same prior specification as Gaussian factor models for the means and idiosyncratic variances of continuous variables, with uniform priors on the cutpoints for discrete variables.

Posterior samples

The `bfa_*` functions will always collect posterior samples of the factor loadings. Posterior samples of the factor scores η_i are only collected if the `keep.scores` argument is `TRUE`. When the sample size is relatively large setting `keep.scores` to `FALSE` will track only the posterior mean and variance of the scores, which can save significantly on storage space.

Posterior samples of the loadings and scores can be retrieved using `get_posterior_loadings` and `get_posterior_scores`. The `get_coda` function will return posterior samples of the loadings or scores as `mcmc` objects ready for use with standard

MCMC diagnostics from the `coda` package. Calling `mean` and `variance` on the output of a `bfa_*` function will return just the posterior means and variances of the loadings/scores. All functions include a `scale` option to return the scaled loadings $\tilde{\Lambda}$, which defaults to `FALSE` for Gaussian factor models and `TRUE` for mixed/copula factor models. The `bfa_*` functions return lists that include posterior samples and summaries for other model parameters.

Several other convenience functions are also available: `cov_samp` and `cor_samp` return arrays with posterior samples of the covariance and correlation matrices. The `coef` function returns samples of the implied coefficients from regressing one subvector of y_i onto the remainder of y_i under the Gaussian factor model (see e.g. West (2003); Carvalho et al. (2008) for discussion of latent factor regression models).

References

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(484):1438–1456.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, 9(2):557.

Ghosh, J. and Dunson, D. B. (2009). Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Hahn, P. R., Carvalho, C. M., and Scott, J. G. (2012). A sparse factor analytic probit model for congressional voting patterns. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):619–635.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1(1):265–283.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). Bayesian gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665.

Puelz, D., Hahn, P. R., and Carvalho, C. (2016). Variable selection in seemingly unrelated regressions with random predictors. *Available at SSRN*.

Quinn, K. M. (2004). Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses. *Political Analysis*, 12(4):338–353.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 7*, pages 733–742.