Frontiers of Bayesian Inference and Data Science

Maria Fernanda Gil Leyva (IIMAS-UNAM, Mexico),
Alan Riva Palacio Cohen (IIMAS-UNAM, Mexico),
Alejandra Avalos-Pacheco (Technische Universität Wien, Austria),
Fan Bu (University of Michigan, USA),
Peter Müller (University of Texas at Austin, USA)

September 1-6, 2024

1 Overview of the Field

The aim of the proposed workshop was to bring together promising junior researchers in Bayesian statistics to expose current research frontiers. The meeting was organized to dedicate each session to one currently active and promising research area, have an overview talk by a senior research leader, followed by presentations by selected junior researchers, including methodological developments and applications. The schedule was designed to encourage ample interaction and feedback. This was also achieved by a strong and deliberately inclusive social program, including a Slack work space to keep everyone engaged and updated, frequent group outings to the town center in the evenings, and two organzied excursions on Wednesday afternoon and on Friday.

Bayesian statistics has undergone a rapid expansion over the past 30 years, starting with a computational revolution sparked by Markov chain Monte Carlo posterior simulation methods in the 1990s, continuing with the rise of non-parametric Bayesian models and methods, allowing inference for complex random structures, extensive cross-fertilization from rapid related developments in machine learning, leading to variational Bayes methods, approximation methods, and innovative novel inference paradigms including approximate Bayesian inference, generalized Bayes and other methods. Several recent research challenges arise from an explosion of research and successful applications of methods in data science, including in particular deep learning and reinforcement learning.

We organized sessions and invited talks in these areas, together with promising young researchers with related work. The program was substantially enriched by cooperation with ISBA/j-ISBA, the section for junior researchers of the International Society for Bayesian Analysis. Fan BU (U. Michigan) was included as an additional organizer, in her role as Program Chair of j-ISBA.

2 Recent Developments and Open Problems

The workshop program touched on recent developments and open problems across the mentioned research areas. These included in particular sessions on nonparametric Bayesian inference (3 sessions), computational Bayes (2 sessions), applications and challenges in biostatistics (2 sessions), robust and scalable inference, causal inference, networks and graphical models (2 sessions), machine learning and reinforcement learning (1 session).

The sessions were the following: (1) random probability measures, and suitable methods of combining evidence across related samples (related talks mostly on Monday morning); (2) ML and other computational methods to implement actual inference (related talks mostly on Monday afternoon); (3) objective and generally robust Bayesian prior models for mixtures (Tuesday morning); (4) computational challenges for inference in such models (Tu afternoon); (5) alternatives to model-based inference (We morning); (6) graphical models (Thursday morning); (7) applications of machine learning (Thursday afternoon). Due to scheduling constraints some talks were scheduled outside this general schedule of focus themes.

3 Presentation Highlights

The workshop program included 39 talks, with 17 talks delivered on-line. Talks by senior lead presenters were scheduled for 45 minutes. All other talks were 30 minutes. Plus usually 5-10 minutes discussion. Friday morning was reserved for co-working and discussion. Following the intention of the program most speakers were junior researchers. But due to limited funding we could only include 3 Ph.D. students, shifting the focus a bit to junior faculty. The organizers paid attention to including a good number of female speakers and maintaining geographic diversity.

Day 1. The workshop started with an opening talk by *Jeff Miller* (Harvard) who dicussed a problem of matrix factorization, motivated by inference for counts of single base substitutions (SBS) mutations across patients. An important part of the inference approach is a construction that allows for a natural inclusion of informative priors on (mutational) signatures. The talk offered a great balance of principled Bayesian modeling, substantial application, foundational concerns and computing issues.

The program continued with a first session on nonparametric Bayesian inference, starting with a broad expository talk by *Tommaso Rigon* (U. Milano Bicocca), introducing a Gibbs-type family for feature allocation models, paralleling the generalization of DP priors to the much larger class of Gibbs-type priors for species sampling models. The presentation included a good mix of review of important themes in nonparametric Bayesian inference, and extension to novel structures. The session continued with *Isabella Deutsch* (U. Edinburgh) who introduced a generalization of the classic Hawkes process (HP) to a new model called the ancestral HP The model is built for an application to data on group chats (few-to-few conversations), which the speaker pointed out are underexposed in the literature. *Vianey Palacios* (Newcastle U.) talked about heavy tailed NGG-mixture (normalized generalized gamma) models applied to EEG data. The model allows to condition on heavy tails of the random probability measure, as needed to model the known heavy tails of brain wave data. Finally, Ricardo Corradin (U Milano-Bicocca), introduced compound completely random measures, starting with a brief review of CRM's and compound random measures, and characterizing the CMR in terms of a POisson process. The presentation ended with an application to PM10 data from Lombardy, including several downstream analyses.

After lunch break the program continued with a first session on Bayesian computation. The session was lead by *Tamara Broderick* (MIT) talking on evaluating evaluation methods, specifically focusing on spatial data. The discussion was inspiring, offering several insights into limitations of commonly used evaluation methods and related challenges. The session continued with *Dootika Vats* (IIT Kanpur) talking about state of the art MCMC methods, including a novel approach for gradient-based methods for non-differentiable target distributions. *Florian Maire* (U. de Montréal) continued talking about variance reduction for MCMC methods by stratification.

The last session on Monday was on biostatistics and neuroscience application. The lead talk was delivered by *Yanxun Xu* (JHU), taking about dynamic treatment regimens for optiomal treatment combinations for HIV patients. Dr. Xu described an ambitious large project involving many challenges for statistical inference. *Farabi Raihan Shuvo* and *Noirrit Chandra* (U. Texas at Dallas) concluded the day with a shared presentation on drift diffusion models for stop signal reaction times in so called "stop signal task" psychological lab experiments.

Day 2. The program on Tuesday started with a session on computational Bayes, with a lead talk by *Trevor Campbell* (UBC) on a locally adaptive variation of popular Metropolis adjusted Langevin MCMC (MALA) methods. The talk included an excellent introduction to MALA and its limitations. *Panayiota Touloupou*

(U. Birmingham) continued the session with a talk on epidemiologic models that include submodels for patient-level data, thus allowing for the imputation of subject-level status. *Bernardo Flores* (U. Texas at Austin) concluded the computational Bayes session with a talk on core-set methods using prediction-based model-free inference. The talk was well received and prompted many comments.

The second Tuesday morning session was on nonparametric Bayesian inference, starting with *Marta Catalano* (Luiss U., Roma), speaking on optimal transport methods. *Sameer Deshpande* (U. WI) continued with a talk on a variation of the popular Bayesian forests (BART) to allow for smooth response surface estimation. Finally, *Giovanni Rebaudo* (U. Torino) introduced the notion of partially exchangeable sequences of partitions (characterized by a pEPPF), and offered several representation theorems allowing definition of a pEPPF by way of a multivariate variation of a species sampling model, by a predictive probability function and by multivariate species sampling process.

Tuesday afternoon started with a session on robust inference. *Francois-Xavier Briol* (UCL) introduced the notion of conjugate Gaussian process regression. *Eli Weinstein* (Columbia U.) discussed "nonparametrically-perturbed" parametric Bayesian models to address ubiquitous model mis-specfication. *Lorenzo Capello* (U. Pompeu Fabra) concluded the session by introducing scalable implementations of Bayesian inference for coalescence models.

The last Tuesday afternoon session was on causal inference, including *Georgia Papadogeorgou* (U. Florida) talking on causal inference in spatial settings in the presence of spatial interference, *Falco Bargagli Stoffi* (UCLA) discussing models for heterogeneous causal effects, and *Dafne Zorzetto* (Brown U.) on nonparametric methods for principal stratification, that is, when experimental units are stratified by a post-treatment variable.

Day 3. Wednesday morning *Yang Ni* (Texas A & M) started with an overview talk on multi-scale nonparametric models for pan-cancer studies. The multi-scale nature allows for clustering by features that are shared across cancer types while allowing additional refinement of clusters by features that are cancer-specific. *Mario Beraha* (Politenico di Milano) introduced nonparametric Bayesian inference to address the frequency recovery problem for life streaming data (that is, counting the occurrence of tokens). The approach adds a model-based justification and refinement to currently used algorithmic methods.

Bernardo Nipoti (U. Milano Bicocca) delivered the lead talk of a session on random networks, proposing methods for clustering networks, introducing a mixture of CER models. *Francesco Gaffi* (U. Maryland) introduced a very general approach to node clustering across multiple or multi-layer networks. Popular special cases of this general approach reduce to conditional partial exchangeabiltiy, to clustering of networks and more. *Beatrice Fanzolini* (U. Bocconi) introduced the notion of conditional partial exchangeability. A special class of models to implement this notion is the telescoping HDP (hierarchical Dirichlet process).

Day 4. Thursday morning started with another session on random networks and graphical models. *Francsca Panero* (La Sapienza U., Roma) introduced methods for modeling sparse networks. The talk started with a critical review of models based on adjacency matrices, which are restricted to allowing only dense or empty networks. *Deborah Sulem* (U. della Svizzera Italiana) discussed scaleable Bayesian inference for Gaussian graphical models (GGM). The approach exploits a connection between GGM and normal linear regression, using a spike and slab prior implements the desired sparsity. *Felipe Medina* (ITAM, Mexico City) introduced methods exploiting recent advances in reversible jump methods to speed up inference on genetic trees.

A session on Bayesian biostatistics started with a talk by *Tianjian Zhou* (Colorado State U.) on sequential clinical trial design, offering a critical discussion of different views on whether designs with interim analyses required adjustment for error rates. *Roberta de Vito* (Brown U.) discussed Bayesian multi-study techniques to learn reproducible signals across studies, based on clever multi-study versions of matrix factorizations. *Yunshan Duan* (U. of Texas at Austin) concluded the session by discussing self-supervised learning with Gaussian process priors to formalize similarity of latent lower-dimensional representations. An application to the analysis of spatial transcriptomics data provided a good case study.

After a lunch break the program continued with a session on applications. A lead talk was delivered by *Veronica Berrocal* (UC Irvine), talking on Bayesian spatial models. *Francesco Denti* (U. Padua) showed several interesting applications of clustering large-scale grouped data, including clustering of music, grouped by artists, and protein activation across different modalities. The session was concluded by *Arman Oganisian*

(Brown U.) who carefully introduced causal inference concerns for sequential clinical trial designs. The talk included a very helpful review of relevant causal inference notions that arise in sequential design.

The final session of the conference was on machine learning and reinforcement learning (RL). The session started with a talk by *Mauricio Garcia Tec* (Harvard) on an application of RL to decisions about issuing heat alerts. The talk introduced the context for planning heat alerts, discussed the general setup of RL and proposed using Large Language Models (LLMs) as priors for RL. *Gemma Moran* (Rutgers U.) introduced a variational auto encoder (VAE) with an additional feature of introducing sparsity. The latter is achieved by introducing an intermediate step with a mask that restricts network layers to using selected subsets of parent nodes only. A final talk was delivered on-line by *Jack Jewson* (Monash U., Melbourne) who reviewed methods for differentially private statistical inference.

4 Scientific Progress Made and Outcome of the Meeting

There was a general consensus that many challenges and opportunities for Bayesian statistical inference are related to computational bottlenecks, to novel and generalized ways of re-stating inference, to causal inference concerns, and to dynamic, sequential decision problems.

Specific opportunities and links were highlighted in the talks and discussed. After a rapid expansion of research in Bayesian inference and computation over the past 35 years, Bayesian statistics & data analysis in general, and nonparametric Bayesian inference in particular is now a mature research area in statistics and machine learning, with methods and applications way beyond the stylized and simplified research questions in earlier work. The current main challenges are related to computational hurdles and bottlenecks as well as to the need to tackle more complex and highly structured problems. These two sets of challenges are deeply intertwined: any improvement on the computational side leads to developments of richer model structures, which then call for further computational advances. Several talks exposed related recent progress and research opportunities, including talks by *Jeff Miller*, *Tamara Broderick*, *Dootika Vats*, *Florian Maire*, *Trevor Campbell*, and *Sameer Deshpande*.

Another main area of current research activity are non-parametric Bayesian methods. Recent research has started to develop methods for ever more complicated data structures, reflecting experimental conditions in many application areas. However, there are still several gaps in current methods and interesting research opportunities. Related talks included *Tommaso Rigon, Isabella Deutsch, Vianey Palacios, Ricardo Corradin, Giovanni Rebaudo, Yang Ni* and *Mario Beraha*.

Recent years have seen a rapid expansion of causal inference methods, to the point that now most biomedical data analysis short of randomized trials involves some causal aspects. Some talks exposed related research methods, including *Arman Oganisan, Georgia Papadogeorgou, Falco Bargagli Stoffi* and *Dafne Zorzetto*.

Interesting research opportunities are always found in substantial applications. Several talks highlighted the scope of principled Bayesian inference to address important scientific research problems. This includes in particular the presentations by *Yanxun Xu, Farabi Raihan Shuvo, Noirrit Chandra, Panayiota Touloupou, Lorenzo Capello, Tianjian Zhou, Roberta de Vito, Mauricio Garcia Tec, Veronica Berrocal, Jack Jewson* and others.

Sequential decision problems are a traditional topic of Bayesian decision science, and continue to provide outstanding research opportunities. Related talks included *Mauricio Garcia Tec, Yanxun Xu and Tianjian Zhou*.

An interesting very recent direction of research in Bayesian methods is the notion of generalized Bayes, proceeding with stylized versions of Bayesian inference. The talks by *Bernardo Flores* and *Yunshan Duan* explored this opportunity, which was otherwise possibly under-exposed.

Graphical models and inference with network data continues to provide great research problems, as was seen in talks by *Bernardo Nipoti, Francesco Gaffi, Francesca Panero, Deborah Sulem, Felipe Medina* and others.

In summary, a conclusion of this workshop is that there is an abundance of research opportunities for junior investigators. The program highlighted some of the more prominent gaps in existing methods and related current research directions. An important outcome of the workshop was the initiation of new professional networks and links, with junior investigators in related research areas learning about each others' work. And most importantly, the opportunity to meet in person and exchange experiences and ideas. Achieving aims

of this workshop related to networking of junior investigators was critically helped by the collaboration with j-ISBA. The j-ISBA section organized two of the sessions, adding a good number of junior participants whom the organizers might have otherwise overlooked.

Comments and recommendations about programming under restrictive budget constraints Overall the hybrid format with only 15 funded in-person participants can work if needed. Key was to maintain across all sessions a good mix of in-person and on-line talks. However, most on-line participants will only connect for their session, and are unlikely to participate very actively beyond their own talk, simply due to technical constraints and the unavoidable distractions by other committments. Actual workshop-level engagement can only be achieved with the in-person participants on location. The restriced local funding allowed us to offer local funding only to selected junior participants. Some senior participants kindly offered to use their own funds to join, but many could not and had to join on-line only.