WORKSHOP

Young Bayesians and Big Data for Social Good (event 1911)

Dates: 23-26 November 2018

Place: CIRM (Marseille Luminy, France)

DESCRIPTION

There is increasing international interest and engagement in the concept of 'data and statistics for social good', with volunteers and organisations working on issues such as human rights, migration, social justice and so on. There is commensurate international interest in Bayesian modelling and computation, particularly in the area of Big Data.

The purpose of this workshop is to build on an emerging endeavour to bring these two areas of activity together. Participants in the workshop will share research into Bayesian methods and big data for social good, as well as discuss related problems that might be addressed using Bayesian approaches. In addition to formal presentations, participants will form collaborative groups to work on new problems and present their findings at a final session on the last day of the workshop which overlaps with a more general conference on 'Bayesian Statistics in the Big Data Era'.

The workshop will provide opportunities for early career researchers around the world who are interested in issues of social good to form lifelong international networks for their immediate and future careers. The semi-structured environment will encourage and progress new research opportunities. The anticipated outputs of the workshop include outlines of journal articles and/or chapters of a new book on this topic, as well as a starting place for practical solutions to problems posed.

ORGANIZING COMMITTEE

- Kerrie Mengersen (Queensland University of Technology)
- Pierre Pudlo (Aix-Marseille Université)
- Christian P. Robert (Université Paris-Dauphine)
- Denys Pommeret (Aix-Marseille Université)

SPEAKERS

- Tamara Broderick (MIT)
- David Corliss (Peace Work)
- Julien Cornebise (Element AI)
- Brody Foy (Oxford) Rhodes Artificial Intelligence Lab
- Andrew Gelman (Columbia University)
- Logan Graham (Oxford) Rhodes Artificial Intelligence
- Antonietta Mira (Università della Svizzera italiana and University of Insubria)
- Cody Ross (Max Planck Institute for Evolutionary Anthropology, Leipzig)
- Rebecca Steorts (Duke University)

https://www.chairejeanmorlet.com/2018-2-mengersen-pudlo-1911.html

Bayesian Statistics and Big Data for Social Good: 23-26 November 2018

Bayesian Statistics and Big Data for Social Good - November 2018

Friday	
12:00 PM	Workshop Registration
12:30PM	12:30-2pm Lunch
2:20 PM	Welcome
2:30 PM	Rebecca Steorts (V)
3:20 PM	Alex Volfonsky
3:40 PM	Break
4:10 PM	Kerrie Mengersen
4:30 PM	Workshop Discussion
6:00 PM	Break
7:30 PM	7:30 Dinner

Saturday	
9:30 AM	Meetup
9:40 AM	David Corliss
10:30 AM	Break
11:00 AM	Julien Cornebise (V)
11:30 AM	Andrea Tancredi
11:50 AM	Brunero Liseo
12:10 PM	Matthew Rushworth
12:30PM	12.30-2.30 pm Lunch
2:30 PM	Logan Graham & Brody Foy (V)
3:20 PM	Charles Gray
3:40 PM	Break
4:10 PM	Gajendra Vishwararma
4:30 PM	Andrew Gelman (V)
5:00 PM	Workshop Discussion
6:00 PM	Free time and dinner Dinner outside CIRM (own arrangements)

Sunday	
9:30 AM	Meetup
9:40 AM	Cody Ross
10:30 AM	Break
11:00 AM	Bihan Zhuang
11:20 AM	Jacinta Holloway
11:40 AM	Workshop Discussion
12:30PM	12.30-2pm Lunch
2:00 PM	Free Time
5:30 PM	Tamara Broderick
6:20 PM	Ethan Goan
6:40 PM	Workshop Discussion
7:30 PM	7:30 Dinner

Monday	
9:00 AM	David Corliss
9:50 AM	Farzana Jahan
10:10 AM	Atanu Bhattacharjee
10:30 AM	Break
11:00 AM	Antonietta Mira
11:30 AM	Workshop Presentations and Concluding Discussions
12:30PM	12.30-2pm Lunch

SPONSORS

Workshop Program: Friday 23rd November 2018, 12:00 – 18:00

Presenter	Title
Rebecca Steorts, Duke University	*Video*
Alex Volfovsky, Duke University	*Design of experiments, networks and social good*
Kerrie Mengersen, Queensland University of Technology	*A Review of Bayesian Statistical Models for Big Data*

Presenter: **Rebecca Steorts**
Via Video
E-mail: beka@stat.duke.edu
Authors: Rebecca Steorts
Affiliation: Duke University

Presenter: **Alex Volfovsky**,
Design of experiments, networks and social good
E-mail: alexander.volfovsky@duke.edu
Authors: Alex Volfovsky
Affiliation: Duke University

There is mounting concern that social media sites contribute to political polarization by creating "echo chambers" that insulate people from opposing views about current events. In this talk I will describe an experiment that addressed political polarization and echo chambers on Twitter by exposing individuals to those of opposite political ideologies. I will then address the difficulties of running complex experiments on large social networks and provide a novel randomization scheme for resolving some of these issues. Lastly, we'll see how a Bayesian approach can help address these difficulties in an observational setting.

Presenter: **Kerrie Mengersen**
Bayesian Networks for conservation
E-mail: k.mengersen@qut.edu.au
Authors: Kerrie Mengersen
Affiliation: Queensland University of Technology

Many conservation issues are complex problems involving environmental, social, political, industrial, commercial and biological issues. The data associated with these problems can be big, messy or nonexistent. How can Bayesian statisticians contribute to these world issues? In this presentation I will describe some of our efforts to use Bayesian networks to address questions of cheetah conservation in Southern Africa, dredging on the Great Barrier Reef and creation of a jaguar corridor in Peru. I will also discuss our use of diverse data sources such as virtual reality, acoustics and citizen science.

Workshop Program: Saturday 24th November 2018, 9:30 – 18:00

Presenter	Title
David Corliss, Peace-Work	*Developing the Future of Data For Special Good*
Julien Cornebise, Element AI	*Live call: Bayes/Big Data/Social Good*
Andrea Tancredi, Sapienza Università di Roma	*A unified framework for de-duplication and population size estimation*
Brunero Liseo, Sapienza Università di Roma	*A Bayesian approach for Normal regression with deduplicated data*
Matthew (Em) Rushworth, Queensland University of Technology	*Working With Constraint: Towards a Bayesian Approach for Compositional Data*
Logan Graham & Brody Foy University of Oxford	*TBA*
Charles Gray, La Trobe University	*Open and reproducible data analysis*
Gajendra Vishwakarma, Indian Institute of Technology Dhanbad	*Bayesian State-Space Modeling in Gene Expression Data Analysis: An Application with Biomarker Prediction*
Andrew Gelman Columbia University	*Honesty and transparency are not enough*

Presenter: **David Corliss**
Developing the Future of Data For Special Good
E-mail: davidjcorliss@peace-work.org
Authors: David Corliss
Affiliation: Peace-Work

In Data For Social Good, statisticians, data scientists, and other researchers work together for greater good of society. People early in their career, especially students, are playing important roles in this growing movement. At the same time Data For Social Good has been gaining strength, new technological developments are expanding the frontiers of what science can do to make a better world. Among emerging statistical methods, Bayesian statistics are especially important due to its ability to leverage informed priors arising from case histories so important to advocacy for justice and social good. This presentation gives an overview of the state of Data For Good, Bayesian methodology as an important area of new technological development, and experiences and opportunities for students to get involved in making a difference by applying their developing analytic skills in projects for the greater good.

Presenter: **Julien Cornebise**
Bayes/Big Data/Social Good
E-mail: julien@elementai.com
Authors: Julien Cornebise
Affiliation: Element AI

Presenter: **Andrea Tancredi**
A unified framework for de-duplication and population size estimation
E-mail: andrea.tancredi@uniroma1.it
Authors: Andrea Tancredi, Brunero Liseo
Affiliation: Sapienza Università di Roma

Data de-duplication is the process of finding records in one ore more datasets belonging to the same entity.
In this paper we tackle the de-duplication process via a latent entity model, where the observed data are perturbed versions of a set of key variables drawn from a finite population of N different entities. The main novelty of our approach is to consider the population size N as an unknown model parameter. As a result, one salient feature of the proposed method is the capability of the model to account for the de-duplication uncertainty in the population size estimation.
As by-products of our approach, we obtain a more adequate prior distribution on the linkage structure and a novel simulation algorithm for the posterior distribution based on the marginalization of the key variables at the population level.
We apply our approach to a synthetic data set comprising German names \added{and we show an application matching records from two data sets reporting victims killed in the recent Syrian conflict.

Presenter: **Brunero Liseo**,
A Bayesian approach for Normal regression with deduplicated data
E-mail: brunero.liseo@uniroma1.it
Authors: Brunero Liseo, Andrea Tancredi
Affiliation: Sapienza Università di Roma

We propose a Bayesian approach for performing record linkage and regression across arbitrarily many lists, while simultaneously considering duplicate detection. We frame the linkage problem as a clustering task, where similar records are clustered to true latent individuals. We propose a statistical model to incorporate both the linking process and the inferential process, including the features of the record as well as the variables needed for inference. Paramount to our approach is the key observation that the prior over the space of linkages can be written as a random partition model. In particular, the Pitman-Yor process will be used as the prior distribution regarding the cluster assignment of records. By the joint modeling of the record linkage and the inferential process, one is able to account for the matching uncertainty in the inferential procedures based on linked data. Moreover, one is able to generate a feedback mechanism of the information provided by the working statistical model on the record linkage process. This feedback mechanism is essential to eliminate potential biases that can jeopardize the resulting post-linkage inference. We apply our methodology to the case of multiple regression, and illustrate empirically that the feedback mechanism improves the performance of the record linkage process.

Presenter: **Em Rushworth**
Working With Constraint: Towards a Bayesian Approach for Compositional Data
E-mail: matthew.rushworth@hdr.qut.edu.au
Authors: Matthew (Em) Rushworth
Affiliation: Queensland University of Technology

Coral reefs are under threat worldwide and the XL Catlin Seaview Survey is one effort to create a consistent record of their status for researchers. We can interpret such data as being compositional and this imposes important constraints upon any analysis. I will explore the current state of the compositional data analysis literature, covering recent developments and discuss how these analyses could be implemented using Bayesian approaches.

Presenter: **Logan Graham and Brody Foy**
To be advised
E-mail: logan@robots.ox.ac.uk
Authors: Logan Graham
Affiliation: University of Oxford

Presenter: **Charles Gray**
Open and reproducible data analysis
E-mail: charlestigray@gmail.com
Authors: Charles Gray
Affiliation: La Trobe University

In an era of ever-bigger data and computational power, analyses are increasingly ambitious. And metaresearchers have been successful in convincing scientists that with great power comes great responsibility; scientists can and should reach for something more nuanced than the default macro in Excel. But if we're going to ask applied scientists to use more sophisticated algorithms, we need to lower the programmatic barrier to the implementation of better (i.e., open, reproducible, accessible, interpretable) methods. In this talk I'll reflect on what I've learnt so far in investigating *good enough* (Bryan, 2018) scientific computing practices and *opinionated* (Parker, 2017) data analysis. I'm interested in hearing thoughts from participants on good enough practices in data handling. What are your horror and success stories? What do you think would facilitate open but ethical data handling? If there is interest, we could form a discussion group and produce a radix website in R to collect our comments..

Presenter: **Gajendra Vishwakarma**
Bayesian State-Space Modeling in Gene Expression Data Analysis: An Application with Biomarker Prediction
E-mail: vishwagk@rediffmail.com vishwagk@iitism.ac.in
Authors: Gajendra K. Vishwakarma
Affiliation: Department of Applied Mathematics, Indian Institute of Technology Dhanbad,

The advancement in computational biology and statistical modeling help to identify the genes which cause the disease like cancer by comparing its expression levels in diseased and healthy people. Bayesian state space modeling is a new advancement in statistics which can estimate unobserved values of a process using the information from an observed outcome and its relationship. Using these two ideas together, we are trying to model and estimate the expression values of genes which are significantly different among two groups.
The complicated integration of posterior densities is carried out using the Markov Chain Monte Carlo (MCMC) simulations. The study shed light on the use of Bayesian State Space modeling to elucidate the behavior of Bio-markers.

Keywords: Bayesian state space modeling, Bio-marker prediction, MCMC

Presenter: **Andrew Gelman**
Honesty and transparency are not enough
E-mail: gelman@stat.columbia.edu
Authors: Andrew Gelman
Affiliation: Department of Statistics and Department of Political Science, Columbia University

Consider this paradox: statistics is the science of uncertainty and variation, but data-based claims in the scientific literature tend to be stated deterministically ("We have discovered ... the effect of X on Y is ... hypothesis H is rejected"). Is statistical communication about exploration and discovery of the unexpected, or is it about making a persuasive, data-based case to back up an argument? Traditional advice on statistics and ethics focuses on professional integrity, accountability, and responsibility to collaborators and research subjects. All these are important, but when considering ethics, statisticians must also wrestle with fundamental dilemmas regarding the analysis and communication of uncertainty and variation. We discuss in the context of examples in various fields of research and policy.

Workshop Program: Sunday 25th November 2018, 09:30 – 19:30

Presenter	Title
Cody Ross, Max Planck Institute for Evolutionary Anthropology	*Resolution of apparent paradoxes in the race-specific frequency of use-of-force by police*
Bihan Zhuang, Duke University	*Entity Resolution with an Application the El Salvadorian Conflict Dat*
Jacinta Holloway, Queensland University of Technology	*Statistical analysis of big data to measure sustainable development goals*
Tamara Broderick, University	*Covariances, robustness, and variational Bayes*
Ethan Goan, Queensland University of Technology	*Deep Learning we can Trust*

Presenter: **Cody Ross**
Resolution of apparent paradoxes in the race-specific frequency of use-of-force by police
E-mail: ctross@ucdavis.edu
Authors: Cody Ross
Affiliation: Max Planck Institute for Evolutionary Anthropology

Analyses of racial disparities in police use-of-force against unarmed individuals are central to public policy interventions; however, recent studies have come to apparently paradoxical findings concerning their existence and form. Although anti- black racial disparities in U.S. police shootings have been consistently documented at the population level, new work has suggested that racial disparities in encounter-conditional use of lethal force by police are reversed relative to expectations, with police being more likely to: 1) shoot white relative to black individuals, and 2) use non-lethal as opposed to lethal force on black relative to white individuals. In this talk, I use a generative stochastic model of encounters and use-of-force conditional on encounter to demonstrate that if even a small subset of police more frequently encounter and use non-lethal force against black individuals than white individuals, then analyses of pooled encounter-conditional data can fail to correctly detect racial disparities in the use of lethal force. In more technical terms, statistical assessments of racial disparities conditioned on problematic intermediate variables, such as encounters, which might themselves be a causal outcome of racial bias, can produce misleading inferences. Population-level measures of use-of-force by police are more robust indicators of the overall severity of racial disparities than encounter-conditional measures---since the later neglect the differential morbidity and mortality arising from differential encounter rates---and should be used when evaluating the local-level public health implications of racial disparities in police use-of-force.

Presenter: **Bihan Zhuang**,
Entity Resolution with an Application the El Salvadorian Conflict Dat
E-mail: bz44@duke.edu
Authors: Bihan Zhuang
Affiliation: Duke University

Entity resolution (record linkage or duplicate detection) is the process of merging noises databases together to remove duplicate entities often in the absence of a unique identifier. Our motivation in this talk is a case study of homicide registries in El Salvador, where from 1980 to 1991, the Republic of El Salvador, in Central America, underwent a civil war between the Salvadoran Government and the left-wing guerrilla Farabundo Marti National Liberation Front (FMLN). The parties signed a peace agreement in 1992, which later led to the creation of the Commission on the Truth (UNTC) for El Salvador by the United Nations. Between 1992 and 1993, the UNTC summoned the Salvadoran society to report violations that occurred during the war, mainly focusing on homicides and disappearances of noncombatants. In 1993 the UNTC published a report with the results of their investigations, including a list of homicides directly obtained from testimonials (friends and family). In addition to the names of the victims, this list contains the reported locations and dates of the killings. As a result of the data collection process, there are many challenges with noise, duplication in the data, and also missing values. Motivated as such, we analyze this data set in a completely unsupervised framework, where we providing other unsupervised comparisons when feasible.

Presenter: **Jacinta Holloway**,
Statistical analysis of big data to measure sustainable development goals
E-mail: j1.holloway@qut.edu.au
Authors: Jacinta Holloway
Affiliation: Queensland University of Technology

The United Nations and World Bank have set Sustainable Development Goals (SDGs) related to quality of human life and environment by 2030. Big data sources, including satellite images, are a low cost and global scale data source for measuring these SDGs. I will describe the types of SDGS that can be measured from satellite images, and how Statistical Machine Learning methods, including Bayesian approaches, can be applied to these data to produce statistics and monitor progress towards the SDGs.

Presenter: **Tamara Broderick**
Covariances, robustness, and variational Bayes
E-mail: tbroderick@csail.mit.edu
Authors: Tamara Broderick
Affiliation: CSAIL, Massachusetts Institute of Technology

In Bayesian analysis, the posterior follows from the data and choices of both prior and likelihood. These choices may be somewhat subjective and reasonably vary over some range. We wish to measure the sensitivity of posterior estimates to variation in these choices. Since VB casts posterior inference as an optimization problem, its methodology is built on the ability to calculate derivatives of posterior quantities with respect to model parameters.
We use this insight to develop local prior robustness measures for VB.
We show our method can be extended to correct VB mis-estimation of posterior covariances by use of a classic result relating derivatives of posterior expectations to posterior covariances. In our experiments, we demonstrate that our methods are simple, general, and fast; they provide accurate posterior robustness measures and uncertainty estimates with runtimes that can be an order of magnitude smaller than MCMC. We illustrate with an analysis of microcredit data.

Presenter: **Ethan Goan**
Deep Learning we can Trust
E-mail: ej.goan@qut.edu.au
Authors: Ethan Goan
Affiliation: Queensland University of Technology

Deep learning has provided state-of-the-art performance in many complex machine learning applications. These models are able to learn combinations of abstract and low level patterns from increasingly larger data sets. Despite promising empirical results, the inherent nature of these models remains unknown. This presentation discusses how a Bayesian framework can be employed to gain insight into deep learning systems. Current research into uncertainty estimation in deep learning is reviewed, along with how this information can be used to deliver systems that society can trust.

Workshop Program: Monday 26th November 2018, 09:30 – 12:30

Presenter	Title
David Corliss, Peace-Work	*Bayesian Capture-Recapture in Social Justice Research*
Farzana Jahan, Queensland University of Technology	*Making More of Spatial Maps: a Bayesian meta-analysis approach*
Atanu Bhattacharjee, Tata Memorial Centre	*tba*
Antonietta Mira Università della Svizzera italana and University of Insubria	*Big data for health: a Bayesian spatio-temporal analysis for predicting cardiac risk in Ticino and optimal defibrillators positioning*

Presenter: **David Corliss**
Bayesian Capture-Recapture in Social Justice Research
E-mail: davidjcorliss@peace-work.org
Authors: David Corliss
Affiliation: Peace-Work

Capture-Recapture (RC) methodology provides a way to estimate the size of a population from multiple, independent samples. While the was developed more than a century ago to count animal populations, it has only recently become important in Data For Social Good. The large number of samples with varying amounts of intersection and developed over a period of time, so often found in Data For Social Good projects, can greatly complicate conventional RC methodology. These conditions are ideal, however, for Bayesian Capture Recapture. This presentation describes the use of Bayesian Capture Recapture to estimate populations in Data for Social Good. Examples illustrating this method include include new work by the author in estimating numbers of human trafficking victims and in estimating the size of hate groups from the analysis of hate speech in social media.

Presenter: **Farzana Jahan**
Making More of Spatial Maps: a Bayesian meta-analysis approach
E-mail: .jahan@hdr.qut.edu.au
Authors: Frazana Jahan
Affiliation: Queensland University of Technology

Analysis of spatial patterns of cancer incidence is a significant field of research. Modelling observational cancer incidence data have been done in many instances, but modelling estimated cancer data available online, has not been explored using meta-analysis model. The use of Bayesian meta-analysis model widens the scope of research utilising the large amount of online cancer data sources. The present study proposes a hierarchical Bayesian meta-analysis model to analyse the point and interval estimates cancer incidence available online, for instance, Australian Cancer Atlas (ACA). The proposed model will focus to reveal the patterns of cancer incidence for the cancers included in ACA in major cities, regional and remote areas. It is observed that the proposed models can generate similar patterns of cancer incidences based urban/rural status of small areas with those revealed by the analysis of raw data . Further investigations can be made including covariates to the proposed model to explore the reasons behind the varying patterns of each cancer and adding spatial components to the model considering the presence of the spatial autocorrelation.

Presenter: **Atanu Bhattacharjee**,
tba
E-mail: atanustat@gmail.com
Authors: Atanu Bhattacharjee,
Affiliation: [1]Centre for Cancer Epidemiology,The Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Tata Memorial Centre,Navi Mumbai-410210.Tata Memorial Centre

Presenter: **Antonietta Mira**
Big data for health: a Bayesian spatio-temporal analysis for predicting cardiac risk in Ticino and optimal defibrillators positioning
E-mail: Antonietta.mira@usi.ch
Authors: Antonietta Mira
Affiliation: Università della Svizzera italiana and University of Insubria

The term 'Public Access Defibrillation' (PAD) is referred to programs based on the placement of Automated External Defibrillators (AED) in key locations along cities' territory together with the development of a training plan for users (first responders). PAD programs are considered necessary since time for intervention in cases of sudden cardiac arrest outside of a medical environment (out-of-hospital cardiocirculatory arrest, OHCA) is strongly limited: survival potential decreases from a 67% baseline by 7 to 10% for each minute of delay in first defibrillation. However, it is widely recognized that current PAD performance is largely below its full potential. We provide a Bayesian spatio-temporal statistical model for predidicting OHCAs. Then we construct a risk map for Ticino, adjusted for demographic covariates, that explains and forecasts the spatial distribution of OHCAs, their temporal dynamics, and how the spatial distribution changes over time. The objective is twofold: to efficiently estimate, in each area of interest, the occurrence intensity of the OHCA event and to suggest a new optimized distribution of AEDs that accounts for population exposure to the geographic risk of OHCA occurrence and that includes both displacement of current devices and installation of new ones.