Chapter 1
Introduction

Kerrie L. Mengersen, Pierre Pudlo, and Christian P. Robert

Abstract This chapter is an introduction to this Lecture Note. We briefly describe the contents of this book. Both parts are introduced, namely part A which deals with Bayesian modeling and part B which presents real-world case studies. The last part of the chapter details the organization of the various events related to the Jean-Morlet Chair. It ends with the issues and research directions identified by the participants of the Conference on Bayesian Statistics in the Big Data Era.

Keywords Bayesian inference · Big data · Computational statistics · Conferences · Workshop · Statistical models · Bayesian modeling · Bayesian computation · Case studies

1.1 Overview

The field of Bayesian statistics has exploded over the past 30 years and is now an established field of research in mathematical statistics and computer science, a key component of data science, and an underpinning methodology in many domains of science, business and social science. Moreover, while remaining naturally entwined, the three arms of Bayesian statistics, namely modelling, computation and inference, have grown into independent research fields. Examples of Bayesian models that have matured during this timeframe include hierarchical models, latent variable models, spatial and temporal models, network and systems models, and models

K. L. Mengersen (✉)
Queensland University of Technology, Brisbane, QLD, Australia
e-mail: k.mengersen@qut.edu.au

P. Pudlo
I2M, CNRS, Centrale Marseille, Aix-Marseille University, Marseille, France

C. P. Robert
Université Paris-Dauphine, Paris, France

K. L. Mengersen et al. (eds.), *Case Studies in Applied Bayesian Data Science*, Lecture Notes in Mathematics 2259, https://doi.org/10.1007/978-3-030-42553-1_1

for dimension reduction. Bayesian computational statistics is now an established discipline in its own right, with a wealth of extensions to the original Markov chain Monte Carlo algorithms, likelihood-free approaches such as Approximate Bayesian computation, and optimization methods such as Variational Bayes and Hamiltonian Monte Carlo. In the domain of Bayesian inference, progress continues to be made on many fronts, including the role and influence of priors, model choice and model robustness, hypothesis testing and so on.

While the research arms of Bayesian statistics continue to grow in many directions, they are harnessed when attention turns to solving substantive applied problems. Each such problem set has its own challenges and hence draws from the suite of research a bespoke solution. It is often useful for both theoretical and applied statisticians, as well as practitioners, to inspect these solutions in the context of the problems, in order to draw further understanding, awareness and inspiration.

The aim of this book is to contribute to the field by presenting a range of such problems and their Bayesian solutions. The book arises from a research program at CIRM in France in the second semester of 2018, which supported Kerrie L. Mengersen (Queensland University of Technology, Australia) as a visiting Jean-Morlet Chair and Pierre Pudlo (Aux-Marseille University) as the local Research Professor. Mengersen was also supported by the Australian Research Council (ARC) through a Laureate Fellowship. Various events were held during the course of this semester, including a Masterclass on Bayesian Statistics, a conference on Bayesian Methods in the Big Data Era, a workshop on Bayes and Big Data for Social Good, and a number of Research in Pairs activities. Summaries of the masterclass, conference and workshop are presented later in this chapter.

1.2 Outline of Book

This book comprises two main parts. In Part A, the state of the art of modern Bayesian statistics is reflected through a set of surveys on topics of current interest in the field.

The first chapters of Part A focus on Bayesian modelling, including a general literature survey and evaluation of Bayesian statistical models in the context of big data by Jahan *et al.* and a more in-depth discussion by Goan of a popular modelling approach, namely Bayesian neural networks. The survey by Jahan *et al.* explores the various approaches to Bayesian modelling, in particular those that are motivated by the advent of so-called ‘big data’. The authors conclude their survey by considering the question of whether focusing only on improving computational algorithms and infrastructure will be sufficient to face the challenges of this ‘big data era’. The chapter of Goan complements this general overview. Unlike their frequentist counterparts, Bayesian neural networks can naturally and formally allow for uncertainty in their predictions, which can lead to richer inferences for detection, classification and regression. Goan introduces these models, discusses

the common algorithms used to implement them, compares various approximate inference schemes and highlights opportunities for future research.

The third chapter of Part A focuses on Bayesian computation, with a survey of Markov chain Monte Carlo (MCMC) algorithms for Bayesian computation by Wu and Robert. The authors provide a brief, general overview of Monte Carlo computational methods, followed by a more detailed description of common and leading edge MCMC approaches. These include Metropolis-Hastings and Hamilton Monte Carlo algorithms, as well as scalable versions of these, and continuous time MCMC samplers based on piecewise deterministic Markov processes (PDMP), the Zig-Zag process Sampler and the Bouncy Particle Sampler. Wu and Robert then introduce a generalization of the latter algorithm in terms of its transition dynamics. Their new Generalised Bouncy Particle Sampler is perceived as a bridge between bouncy particle and zig-zag processes that avoids some of the tuning requirements.

The final survey chapters of Part A focus on two illustrative challenge in Bayesian data science that merge the fields of modelling and computation, namely variable selection and model choice in high dimensional regression, and posterior inference for intractable likelihoods. Sutton addresses the first challenge by describing three common priors that are used for sparse variable selection, namely model space priors, spike and slab priors and shrinkage priors, with corresponding computational approaches and software solutions. The second challenge is surveyed by Moores *et al.,* who defines an intractable likelihood as one for which the likelihood function is unavailable in closed form or which is infeasible to evaluate. The approaches covered by Moores *et al.* include pseudo-marginal methods, approximate Bayesian computation (ABC), the exchange algorithm, thermodynamic integration and composite likelihood, with particular attention paid to advancements in scalability for large datasets.

Part B of the book consists of a set of real-world case studies that aim to illustrate the wide variety of ways in which Bayesian modelling and analysis can enhance understanding and inference in practice. Three fields have been chosen for exposition, namely health, environmental health and ecology. The value of Bayesian data science in modelling the brain is described in the first pair of chapters by Cespedes *et al.,* who focus on a joint model of cortical thickness and network connections, and White *et al.*, who aim to cluster action potential spikes. The second pair of chapters address public health issues of vector-borne diseases (Aswi *et al.*), cancer (Cramb *et al.*). The next chapter explores the link between environmental exposures and the neurodegenerative Parkinson's disease (Thomas *et al.*), followed by two chapters that address challenges in workplace health (Harden *et al.,* Tierney *et al.*).

In the last four chapters on ecological applications, the authors showcase the use of diverse data sources to address challenges in conservation and biosecurity. Davis *et al.* use data elicited from experts and citizens to explore factors involved in conservation of cheetahs in Southern Africa and jaguars in South America, while Sequeira *et al.* and Vercelloni *et al.* employ observational data to gain insights into marine conservation. In contrast, Ullah *et al.* employ satellite imagery to model the risk of fire-ant incursion.

Each of these case studies has a complication that motivates a rich range of Bayesian solutions. The models considered by the authors include hierarchical models (Sequeira *et al.*, Tierney *et al.*, Vercelloni *et al.*), parametric and nonparametric mixture models (White *et al.*), spatial models (Aswi *et al.,* Cespedes *et al.,* Cramb *et al.*, Ullah *et al.*) and Bayesian network approaches (Davis *et al.,* Harden *et al.,* Thomas *et al.*).

Cespedes *et al.* propose a new Bayesian generative model for analysis of MRI data that allows for more complete insight into the morphological processes and organization of the human brain. Unlike current models that typically perform independent analyses, their proposed model uses a form of wombling to perform joint statistical inference on biomarkers and connectivity covariance networks. The new model provides posterior probabilities for the connectivity matrix, accounting for the uncertainty of each connection, and enables estimation of the spatial covariance among regions as well as global cortical thickness. These features are critical in the assessment of the pathology of neuro-degenerative diseases such as Alzheimers. White *et al.* also consider a case study in neuroscience research, this time focusing on the analysis of action potentials or 'spike sorting', which aims to characterize neural activities in subjects exposed to different stimuli or other experimental conditions. This problem is cast as an unsupervised clustering problem to which two types of mixture models are applied. The complications in these models include the choice of the number of identified clusters and classification uncertainty.

In a quite different spatial setup, Aswi *et al.* compare the performance of six Bayesian spatio-temporal models in their investigation of dengue incidence in Makassar, Indonesia, taking into account the challenges that are typically faced in practice but not typically catered for by the models, namely a small number of areas and limited number of time periods. These types of geographic spatial models for small area estimation of disease are also considered by Cramb *et al.*, with a focus on the choice of model under different scenarios of rare and common cancers over different types of spatial surfaces. The authors reveal the dramatic impact of model choice on posterior estimates and recommend comparing several different models, especially when analyzing very sparse data.

Returning to neurodegenerative diseases, but from a different perspective, Thomas *et al.* focus on the challenge of understanding the association between environmental exposure to organochloride pesticide (OCP) and age at onset of Parkinson's disease. The authors explore this complicated association via an ensemble model comprised of a meta-analysis and a Bayesian network, whereby odds ratios and other information extracted from the literature are merged with clinical data to probabilistically quantify the network model. The authors acknowledge the limitations of this approach but suggest its merit for future investigation as a mechanism for integrating disparate, sparse data sources to address environmental health questions.

The utility of Bayesian approaches in modelling different aspects of a problem is highlighted in the two chapters by Harden *et al.* and Tierney *et al.,* who both focus on workplace health. Harden *et al.* use an approach similar to that of Thomas *et al.*,

in that they employ published information to characterize, quantify and compare features of workplace health and workplace wellness programs. In contrast, Tierney *et al.* utilise records of routine medical examinations, which are characterized by substantive missing data, to facilitate early detection of disease amongst workplace employees. Whereas Harden *et al.* adopt a Bayesian network approach to combine and quantify their workplace health and wellness systems, Tierney *et al.* use a Bayesian hierarchical regression model to create a workplace health surveillance program.

Turning to the case studies in ecology, one of the major constraints in conservation research and practice is the sparsity of data and the complex interaction of contributing factors. An example is given by Davis *et al.*, who tackle the sensitive issue of human-wildlife conflict and illustrate the utility of a Bayesian network for these types of problems. Two iconic wildlife species are considered and the implications for different conservation management strategies are discussed.

In a quite different ecological setup, Sequeira *et al.* employ a series of uninformed and informed priors to investigate the issue of space-time misalignment of responses and predictors in hierarchical Bayesian regression models. The particular models of interest are predictive biodiversity distribution models, which are used to understand the structure and functioning of ecological communities and to facilitate their management in the face of anthropogenic disturbances. The focal challenge is to predict fish species richness on Australia's Great Barrier Reef. Vercelloni *et al.* also take the Great Barrier Reef as their study area, with the aim of estimating long-term trajectories of habitat forming coral cover as a function of three different spatial scales and environmental disturbances. A hierarchical Bayesian model was also adopted by these authors, but in a semi-parametric framework.

The final chapter in this part, by Ullah *et al.,* focuses on the problem of classification of features of interest in large images. The approach proposed by these authors is to fit a Bayesian non-parametric mixture model to multiple stratified random samples of the image data, followed by the formation of consensus posterior distribution which is used for inference. The method is applied to the challenge of employing remote sensing for plant and animal biosecurity surveillance, with a particular focus on using satellite data to identify high risk areas for fire ants in the Brisbane region of Australia.

Together, these chapters provide a rich tapestry of activity in applied Bayesian data science, motivated by a wide range of real-world problems. It is hoped that these case studies will inspire and expand both research and practice in Bayesian data science.

1.3 Jean-Morlet Research Semester Activities

As described above, the Jean-Morlet semester at CIRM in the second half of 2018 included the organization of a masterclass, conference and workshop at the CIRM Research Centre in Marseille, France. A brief summary of each of these three

activities is given in this section, with a focus on highlighting research directions and illustrating applications of Bayesian statistical modelling and analysis.

1.3.1 Masterclass on Modern Bayesian Statistics

A clear indicator of the establishment of Bayesian statistics is the increased number of graduate courses on the topic. Such courses are more common in statistical science, computer science and data science, but they are also now appearing in a wide range of other fields of science, social science and business. Intensive short courses on Bayesian statistics are also popular mechanisms for training for graduates as well as academics, other researchers and practitioners. These courses are presented either in-person or online.

One such course was the Masterclass in Bayesian Statistics, presented at CIRM on 22–26 October, 2018. Videos and slides of the presentations given in this Masterclass are publicly available on the CIRM website: https://www.chairejeanmorlet.com/2018-2-mengersen-pudlo-1854.html.

The topics presented in this Masterclass can be broadly categorised into Bayesian modelling and Bayesian computation. As an example of the former, Chris Holmes addressed the problem of Bayesian learning at scale. His argument was that Bayesian learning from data is predicated on the likelihood being true, whereas in reality all models are false. If the data are simple and small, and the models are sufficiently rich, then the consequences of model misspecification may not be severe. However, since data are increasingly being captured at scale, Bayesian theory as well as computational methods are required that accommodate and respect the approximate nature of scalable models. A proposed approach is to include the uncertainty of the model in the analysis, via a principled nonparametric representation. Other approaches to model assessment were discussed by Aki Vehtari, who covered cross-validation and projection predictive approaches for model assessment, inference after model selection, and Pseudo-BMA and Bayesian stacking for model averaging. This discussion was complemented by R notebooks using rstanarm, bayesplot, loo, and projpred packages.

Problem-specific Bayesian models were also presented. For example, the presentation by Adeline Samson focused on various types of stochastic models in biology, including point processes, discrete time processes, continuous time processes and models with latent variables, and elaborated on some of the statistical challenges associated with their application.

The many directions of current research in Bayesian computational statistics were highlighted in the presentation by Christian P. Robert, who discussed more efficient simulation via accelerating MCMC algorithms, to approximation of the posterior or prior distributions via partly deterministic Markov processes (PDMP) like the bouncy particle and zigzag samplers. The focus of this presentation was on the evaluation of the normalising constants and ratios of normalising constants in such methods.

Two algorithms of strong current interest are Sequential Monte Carlo (SMC) and Variational Inference (VI) or Variational Bayes (VB). VI algorithms were addressed by Simon Barthelmé, who suggested methods for correcting variational approximations to improve accuracy, including importance sampling and perturbation series. SMC was introduced by Nicolas Chopin, who motivated the approach by state-space (hidden Markov) models and their sequential analysis, and touched on the analysis of non-sequential problems. The presentation also included a description of the formal underpinnings of SMC, building on concepts of Markov kernels and Feynman-Kac distributions, and a discussion of Monte Carlo ingredients including importance sampling and resampling. Standard bootstrap, guided and auxiliary particle filters were then described, followed by estimation methods via PMCMC and SMC^2. SMC was also discussed by Marie-Pierre Etienne in the context of partially observed stochastic differential equations applied to ecology, and by Adam Johansen in the context of defining a genealogy of SMC algorithms. Similarly, Sebastian Reich proposed a unifying mathematical framework and algorithmic approaches for state and parameter estimation of particular types of partially observed diffusion processes.

Scalable algorithms for Bayesian inference are also of great interest. This was reflected in the presentation by Giacomo Zanella, who focused on scalable importance tempering and Bayesian variable selection.

Another indication of the mainstream status of Bayesian statistics is the proliferation of dedicated R packages and analogies in Python and other software, as well as an increase in the number of stand-alone statistical software packages. While many of the Masterclass presentations referred to specific packages, some presentations focused on the stand-alone software. For example, Harvard Rue presented a tutorial on Bayesian computing with INLA, with a focus on estimation of the distribution of unobserved nodes in large random graphs from the observation of very few edges and a derivation of the first non-asymptotic risk bounds for maximum likelihood estimators of the unknown distribution of the nodes for this sparse graphical model. This tutorial was complemented by a presentation on the same topic by Sylvain le Corff. A tutorial on JASP was presented by Eric-Jan Wagenmakers and the software package STAN was used by Bruno Nicenboim to implement a cognitive model of memory processes in sentence comprehension.

Finally, as in all areas of computational statistics, good practice in dealing with data and coding Bayesian algorithms is essential. Julien Stoehr and Guillaume Kon Kam King presented a tutorial on this topic, with reference to writing R code, R packages and R Markdown and knitr documents.

1.3.2 Conference on Bayesian Statistics in the Big Data Era

This conference aimed to bring together an international and interdisciplinary group of researchers and practitioners to share insights, research, challenges and opportunities in developing and using Bayesian statistics in the Big Data era.

As expected, a major focus of the conference was on scalable methods, i.e. models and algorithms that cope with or adapt to increasing large datasets. As illustration, scalable nonparametric clustering and classification were proposed by Peter Muller. Two strategies were discussed: one based on a consensus Monte Carlo approach that splits the data into shards and then combines subset posteriors to recover joint inference, and another that exploits predictive recursion to build up posterior inference for the complete data. Ming-Ngoc Tran canvassed a range of topics such as intractable likelihood and its connection with Big Data problems, subsampling-based MCMC, HMC and SMC for models with tall data, and Variational Bayes estimation methods for extremely high-dimensional models. A quite different compositional approach to scalable Bayesian computation and probabilistic programming was described by Darren Wilkinson.

Sub-sampling, approximations and related methods for dealing with large datasets was discussed by a range of authors. Pierre Alquier proposed techniques for sub-sampling MCMC and associated approximate Bayesian inference for large datasets, while Tamara Broderick proposed a different approach to automated scalable Bayesian inference via data summarisation. David Dunson also contributed to this discussion, describing new classes of scalable MCMC algorithms based on biased subsampling and multiscale representations that, instead of converging to an exact posterior distribution, employ approximations to speed up computation and achieve more robust inference in big data settings. Stéphane Robin also used deterministic approximations to accelerate Sequential Monte Carlo (SMC) for posterior sampling via a so-called shortened bridge sampler. Approximate Bayesian Computation (ABC) was discussed by Pierre Pudlo in the context of model choice, and Jean-Michel Marin described a method of improving ABC through the use of random forests.

With respect to modelling, nonparametric approaches were a popular topic of discussion. In addition to Muller's presentation described above, Amy Herring described centred partition processes for sparse data. Alternative approaches to defining nonparametric priors were also proposed, for example by Antonio Lijoi in the context of covariate-dependent data, and Igor Prünster through the use of hierarchies of discrete random probabilities.

Other models of great international interest included high-dimensional spatial and spatio-temporal models, discussed by Sudipto Banerjee and Noel Cressie, optimal transport described by Marco Cuturi, and high dimensional inference for graphical models presented by Reza Mohammadi. High dimensional regression was addressed by Akihiko Nishimura, who described computational approaches for "large n and large p" sparse Bayesian regression in the context of binary and survival outcomes, and Benoit Liquet, who focused on Bayesian variable selection and regression of multivariate responses for group data. Related design questions were also a priority issue, since efficient sampling, survey and experimental designs can dramatically reduce the number of observations and variables required for inference and the associated computational cost of analysis. To this end, Jia Liu proposed a Bayesian model-based spatiotemporal survey design for log-Guassian Cox processes.

Approaches for high dimensional time series data, motivated by applications in economics, marketing and finance, were promoted by Sylvia Frühwirth-Schnatter, Gregor Kastner and Gary Koop. Fruwirth-Schnatter focused on Markov chain mixture models to describe time series with discrete states, and showed that these models are able to capture both persistence in the individual time series as well as cross-sectional unobserved heterogeneity. Koop described a different approach, focusing on composite likelihood methods for Bayesian vector autoregressive (VAR) models with stochastic volatility, presented by Gary Koop. Kastner also considered VAR models with time-varying contemporaneous correlations that are reportedly capable of handling vast dimensional information sets.

A wide range of applied problems were tackled in the conference, with attendant novel methodology. For example, challenges in public health ranged from nonparametric approaches to modelling sparse health data, by Amy Herring, to methods for including residential history in mapping long-latency diseases such as mesothelioma, by Christel Faes. Graphical models for brain connectivity were also discussed by Reza Mohammadi, as mentioned above. In the genetics field, the problem of high-throughput sequencing data in genomics was addressed using Bayesian multiscale Poisson models by Heejung Shim, while Zitong Li proposed non-parametric regression using Gaussian Processes for analysing time course quantitative genetic data, in particular quantitative trait loci (QTL) mapping of longitudinal traits. Time-course data prediction for repeatedly measured gene expression was also discussed by Atanu Bhattacharjee. Business-related applications included Bayesian preference learning, described by Marta Crispino, Bayesian generalised games in choice form as a new definition of a stochastic game in the spirit of the competitive economy, by Monica Patriche, and econometric models by Fruwirth-Schnatter. As mentioned above, an environmental problem, namely estimating the extent of arctic sea-ice, was addressed by Noel Cressie using a hierarchical spatiotemporal generalised linear model, where data dependencies are introduced through a latent, dynamic spatiotemporal mixed-effects model using a fixed number of spatial basis functions. The model was implemented via a combination of EM and MCMC.

Other issues that were addressed at the conference included data privacy and security (presented by Louis Aslett) and causality in modern machine learning (Logan Graham). In the latter presentation, Graham argued that while much current attention has focused on using machine learning to improve causal inference, there is opportunity for the inverse, namely to use tools from causal inference to improve the learning, efficiency, and generalisation of machine learning approaches to machine learning problems.

1.3.3 Workshop on Bayes, Big Data and Social Good

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. This interest is generating a growing number of workshops on the topic.

One such workshop on "Young Bayesians and Big Data for Social Good" was held at CIRM on 23–26 November 2018. The workshop showcased some of the organisations that are dedicated to social good and are employing data science in general, and Bayesian statistics for this purpose. It also provided opportunity for Bayesian statisticians to discuss methods and applications that are aligned to social good. As indicated by the title of the workshop, the participants were primarily, but not exclusively, early career researchers.

Dedicated social good organisations that were represented at the workshop included Peace at Work (peace-work.org, represented by David Corliss) and Element AI (elementai.com, represented by Julien Cornebise). For example, David Corliss, a spokesperson for Peace at Work, provided 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.

The workshop exposed a wide range of social good problems and associated Bayesian statistical solutions. For example, Jacinta Holloway focused on the utility of satellite imagery to inform the United Nations and World Bank Sustainable Development Goals related to quality of human life and environment by 2030. In a similar vein, Matthew Rushworth described the use of underwater imagery to inform statistical models of the health of the Great Barrier Reef, a UNESCO World Heritage site under threat in Australia. From a computational perspective, Tamara Broderick related her research into the development of simple, general and fast local robustness measures for variational Bayes in order to measure the sensitivity of posterior estimates to variation in choices of choices of priors and likelihoods, to the issue of analysing microcredit data which impacts on small business success in developing countries.

An important issue of trust in data was raised by Ethan Goan in the context of deep learning. Although these models are able to learn combinations of abstract and low level patterns from increasingly larger datasets, the inherent nature of these models remain unknown. Goan proposed that a Bayesian framework can be employed to gain insight into deep learning systems, in particular their attendant uncertainty, and how this information can be used to deliver systems that society can trust.

A number of presentations focused on entity resolution (record linkage and de-duplication of records in one or more datasets) in order in order to accurately estimate population size, with application to estimating the number of victims killed in recent conflicts. Different statistical approaches to address this problem were presented by Andrea Tencredi and Brunero Liseo, Rebecca Steorts, Bihan Zhuang and David Corliss. For example, Bayesian capture-recapture methods were proposed by David Corliss to estimate numbers of human trafficking victims and estimate the size of hate groups in the analysis of hate speech in social media. Tencredi and Liseo took another approach, by framing the linkage problem as a

clustering task, where similar records are clustered to true latent individuals. The statistical model incorporated both the linking process and the inferential process, including the features of the record as well as the variables needed for inference. Paramount to their 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 was used as the prior distribution regarding the cluster assignment of records. The method is able to account for the matching uncertainty in the inferential procedures based on linked data, and can also generate a feedback mechanism of the information provided by the working statistical model on the record linkage process, thereby eliminating potential biases that can jeopardize the resulting post-linkage inference.

The use of Bayesian statistics and big data for health was also a common theme in the workshop. For example, Akihiko Nishimura proposed new sparse regression methods for analyzing binary and survival data; Gajendra Vishwakarma described the use of Bayesian state-space models for gene expression data analysis with application to biomarker prediction; and Antonietta Mira detailed a Bayesian spatio-temporal model to predict cardiac risk, creating a corresponding risk map for a city, and using this to optimize the position of defibrillators.

A different problem tackled by Cody Ross was the resolution of apparent paradoxes in analyses of racial disparities in police use-of-force against unarmed individuals. For example, 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 shoot white relative to black individuals, and use non-lethal as opposed to lethal force on black relative to white individuals. Ross used 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.

Finally, as noted above, good practice in statistical computation can provide substantial benefits for both researchers and practitioners. To this end, Charles Gray described the use of github and the R package ‘tidyverse’ for improved collaborative workflow, with reference to an application in maternal child health research.

1.4 The Future of Bayesian Statistics

Given that Bayesian statistics is now an established field of research, computation and application, it is of interest to consider the future of the profession, particularly in the era of ‘big data’. This was the question posed to the participants of the Conference on Bayesian Statistics in the Big Data Era held at CIRM, Marseille, France on 26–30th November 2018.

The participants collectively identified major issues and directions for Bayesian statistics. These were collated into four key themes: data, computation, modelling; and training. An overall statement on each theme is presented below.

1.4.1 Data

1. Policies like GDPR will need mathematical and statistical formalizations and implementation. This will become an increasingly important issue also beyond the regions where GDPR formally applies.
2. Addressing grand challenges will increasingly require the use of multiple data sources from diverse locations, and need approaches to deal with the resulting heterogeneous.
3. Issues of quality assurance and persistence will increase, as official statistics tend to be replaced by commercial services.
4. While traditional questions of experimental design are becoming less relevant, other experimental design questions will arise, related to subsampling big data.
5. Recognition of the provenance of the data is becoming important, including in particular social media data and derived data from climate models etc.

1.4.2 Computation

1. Despite the exponential growth in computational Bayesian statistics, new algorithms are still required that are targeted to big data.
2. There will be continuing interest in approximations and subsampling strategies, as well as methods for taking advantage of sparse data.
3. Bayesian software will become faster and more intuitive for users to use. This will benefit from active online communities.
4. Current software, such as Tensorflow, R and C++, differ with respect to ease and computational speed, and need to be able to talk to each other better.
5. However, software alone cannot help. Bayesian statisticians will also need to understand more about hardware and decentralised data in order to fine tune algorithms for specific problems.

1.4.3 Models

1. Bayesian models will continue to evolve in the 'big data era'. Three major directions of evolution are in priors, model setup and model choice.
2. With respect to priors, on the one hand, informative priors such as those that induce shrinkage will play an increasingly important role, but on the other hand,

priors on high dimensional data tend to become very influential so development of objective priors for high dimensional data will continue to be of great interest. Overall, we need better ways of choosing priors.
3. With respect to model setup, overall the future will see the development of better Bayesian frameworks, which ignore unnecessary information from data before modelling, determine relevant information for modelling, automatically determine the required complexity of the model, and include generalized methods for Bayesian model selection and diagnostics. There is little doubt that models need to evolve to cope with new kinds and quantities of data. On the other hand, perhaps progress could come from being able to ignore certain aspects of the data. After all, having to fully specify every aspect of a data generating process for a complex dataset can be tedious at best, impossible or harmful at worst.
4. With respect to model choice, we need to learn to handle model misspecification in better ways and develop robust Bayesian modelling approaches. There will be co-existence of parametric and nonparametric models in the future, where the application and utility will depend on specific domains of application. Model-free methodologies will also become more important. On one hand, Bayesian models will become more sophisticated, more flexible (taking advantage of Bayesian nonparametrics), bigger and better, as enabled by data and computational advances.

1.4.4 Training

1. The future will see the development of Bayesian tools for non-expert modellers, with plug and play type models for easy application. When compared to 10 years ago, a huge amount of students now study statistics and machine learning. A few of them are indeed really interested in mathematics, modelling and computer science, but others are more in quest of user-friendly software to use easily in their jobs. If we want to attract these non-specialist students, we need to provide more user-friendly tools for Bayesian learning: Bayesian equivalents of TensorFlow for neural networks. On the other hand, we should not sacrifice the statistical part of the training: modelling, theory, understanding the methods, interpretation of the results. Indeed, there is and should continue to be a role for statisticians and data scientists.
2. We should also talk about artificial intelligence (AI), but also about the world in which AI resides and alternatives to AI. AI will lead to personalized medicine, but this cannot be done without a sound knowledge of biostatistics. Similarly, environmental and economic problems cannot be solved without statistics. We live in a complex world. We should warn people that it will become impossible to understand these topics without a strong statistical background and show them how the Bayesian approach is flexible enough to tackle these problems.