BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Bayesian Statistical Methods (BAY)

Semester 2, 2022

Prepared by:

Dr Koen Simons
Centre for Epidemiology and Biostatistics,
Melbourne School of Population and Global Health
The University of Melbourne

THE UNIVERSITY OF MELBOURNE

Contents

Bayesian Statistical Methods (BAY)

Semester 2, 2022

Instructor contact details

Dr Koen Simons

School/Dept,
University

(03) 9035 7927

Koen.simons@unimelb.edu.au

If you have any general BCA queries, please contact: Karolina Kulczynska-Le Breton or Emily Higginson at the BCA Coordinating Office on 02 9562 5076/54 or email bca@sydney.edu.au

Background

Bayesian inference is concerned with fitting full probability models to data and summarizing the results as a probability distribution for the parameters that define the model and any other unobserved quantities that can be predicted (prospectively or retrospectively) as a consequence of the model. The Bayesian approach allows the data analyst to make probabilistic statements about both quantities they observe and quantities about which they wish to learn, including model parameters, imputations for missing data and predictions for future observations. In contrast, the standard frequentist approach makes modelling assumptions only about observable quantities and cannot benefit from the explicit use of subjective probability to quantify uncertainty about inferences. Bayesian inference is important in modern statistical practice, including biostatistics. This is in part due to a belief by some statisticians that it has philosophical and logical advantages. It is, however due much more to the ability of the Bayesian modelling process to tackle complex hierarchical problems and, most importantly, the availability of new computational techniques and user-friendly software tied to the ever-increasing processing power of desktop computers.

Prerequisites

Epidemiology (EPI), Mathematical Foundations for Biostatistics (MFB), Principles of Statistical Inference (PSI), Regression Modeling for Biostatistics 1 (RM1)

Unit summary

This unit introduces Bayesian statistical concepts and methods, with emphasis on practical applications in biostatistics. We begin with a discussion of subjective probability in quantifying uncertainty in the scientific process. Subsequently, the concept of full probability modelling is introduced and developed through single- and multi-parameter models with conjugate prior distributions.

The connection with frequentist approaches is examined in light of the relationship between non-informative and informative prior distributions and their effect on posterior estimates. We discuss the specification of appropriate prior distributions, including the concepts of non-informative and weakly informative priors. We consider the frequentist properties of Bayesian procedures.

This unit explains why Bayesian statistics requires advanced computational techniques, and introduces the class of Markov Chain Monte Carlo (MCMC) algorithms to explore the full posterior distribution. The challenges of and intuition behind the algorithms are shown. Subsequently, we focus on assessing whether the algorithm converged and how to remediate failures. We discuss posterior predictive checks and information criteria as means to assess model fit and compare alternative models.

We illustrate the richness of Bayesian modelling software as a means to estimate hierarchical models and discuss the role of shrinkage. Finally, we revisit the full Bayesian workflow.

This unit is taught exclusively using the R software, using its RStan library and related packages. All examples and tutorials are provided in R only.

Workload requirements

The expected workload for this unit is 10-12 hours per week on average, consisting of discussion posts, independent study and completion of assessment tasks. For the first five modules, one Zoom session per module is planned. There are no required readings other than the subject notes.

Learning Outcomes

At the completion of this unit students should be able to:

1. Explain the difference between Bayesian and frequentist concepts of statistical inference.
2. Demonstrate how to specify and fit simple Bayesian models with appropriate attention to the role of the prior distribution and the data model.
3. Explain how these generative models can be used for inference, prediction and model criticism.
4. Demonstrate proficiency in using statistical software packages (R) to specify and fit models, assess model fit, detect and remediate non-convergence, and compare models.
5. Engage in specifying, checking and interpreting Bayesian statistical analyses in practical problems using effective communication with health and medical investigators.

Unit content

The unit is divided into 6 modules, summarised in more detail below. Each module will involve approximately 2 weeks of study and generally includes the following material:

1. Module notes describing concepts and methods, and including some exercises of a more "theoretical" nature.
2. One or more extended examples illustrating the concepts/methods introduced in the notes and including more practically oriented exercises.
3. Classroom discussion of the module guided by the coordinator.

Study materials for all Modules are downloadable from the eLearning unit site. Assignments and supplementary material, such as datasets will be posted to the unit site. Please note that we are not able to post copies of copyright material (journal articles and book extracts)—for these you will have to rely on the hard copy mail-out or resources from your home university's library.

Recommended approaches to study

Students should work through each module systematically, following the module notes and any readings referred to, and working through the accompanying exercises. *You will learn a lot more efficiently if you tackle the exercises systematically as you work through the notes.* You are encouraged to post any content-related questions to eLearning, whether they relate directly to a given exercise, or are a request for clarification or further explanation of an area in the notes. You should also work through all of the computational examples in the notes for yourself on your own computer.

Outline solutions to the exercises in each module (except those to be submitted for assessment, as described below) will be posted online at the midway point of the allocated time period for the module. This is intended to encourage you to attack the exercises independently (or via the eLearning site), and yet not make you wait too long to see the sketch solutions.

Make the most of this unit by engaging with coordinators and fellow students on the Discussion Board and in Tutorials. These are safe spaces to discuss the course material and related ideas and students are encouraged to make the most of them by engaging in respectful discussion.

Method of communication with coordinator

Questions about administrative aspects or course content can be emailed to the coordinator, and when doing so please use "[BAY]:" in the Subject line of your email to assist in keeping track of our email messages. Coordinator/s will be available to answer questions related to the module notes and practical exercises, and to address any other issues that require clarification. However, please note that instructors are not necessarily available every day of the week and you should expect that it may take a day or so to respond to questions (possibly longer over weekends and during breaks!).

We strongly recommend that you post content-related questions to the Discussions tool in the BAY area of BCA's eLearning site. In 2022 we are using the Learning Management system hosted by the University of Sydney. You may be familiar with the system from previous BCA units, and will receive any specific instructions on using the eLearning site this semester from the BCA Coordinating Office. There is also a "Getting Started" document available on the Student Resources page of the BCA website.

Relying on Canvas for content-related communication and problem-solving will enable other students to benefit from responses and indeed to respond themselves, and we try to encourage as much interaction as possible within the class through this medium. We will also use Canvas for posting all course materials.

Module Descriptions

Below is an outline of the study modules, followed by a timetable and assessment description table

Each module is scheduled to begin on a Monday and conclude on the Sunday of the following week, except Module 6. **The due date for submission of the required exercises from each module is 11:59pm on the Monday immediately following the completion of the module, as indicated below. There are no short exercise sets for Modules 3 and 6. The Major assessments cover multiple Modules, respectively Modules 1-3 and Modules 1-5.**

Module 1: **Bayesian Concepts**

- Schools of thought
- Probability, information, uncertainty and beliefs
- Bayes' rule
- Bayesian statistical inference

This module discusses probability and philosophical views on probability. Bayes' rule is introduced and demonstrated with examples using discrete parameters. Bayesian Workflow is introduced. Common myths about confidence intervals are discussed. What is prior information? Examples of prior elicitation are provided.

Module 2: **Continuous priors – conjugate distributions**

- Are integrals just sums?
- The normal model with known variance
- Conjugate priors for single parameter families
- The linear model; or the beginning of the end
- Limitations of conjugate priors

This module extends from discrete parameters to continuous parameters. The problem of integration is discussed, with a brief reminder what integrals are, and it is shown that analytical solutions exist for a select subset of problems. Limitations of conjugate priors are discussed.

Module 3: **Model fitting with Markov Chain Monte Carlo**

- The curse of dimensionality
- Souvenirs from Monte Carlo
- Markov Chain Monte Carlo
- Assessing convergence
- How good is the approximation?
- Using R for MCMC

This module discusses the impossibility of using simple methods to determine the entire posterior distribution. A brief refresher on Monte Carlo methods is provided, before introducing Markov Chain Monte Carlo. Both the Gibbs sampler and Metropolis-Hastings are introduced, culminating in the No-U-Turn Sampler. (Note that these algorithms have all been programmed by experts.)

Pragmatically, the problem of assessing convergence in distribution is explained, as well as methods to determine the accuracy of the numerical approximation. Worked examples using R are provided.

Module 4: **Model checks, comparisons and stacking**

- Posterior predictive checks
- Out-of-sample predictive performance
- Stacking

This module deals with pragmatic questions. Supposing the algorithm has converged and we have a representative sample from the posterior, how do we address the questions: ”Is this model any good? Where does it fall short? What alternative models could be better? Which of these models fit the data best?”

Module 5: **Hierarchical models and shrinkage**

This module introduces hierarchical models and partial pooling. We contrast partial pooling with complete pooling and no pooling, and the bias-variance trade-off.

Module 6: Bayesian Workflow and other Open Problems (Not for assessment)

This module serves as a book-end to Module 1: why Bayes, why not Bayes? Is Bayes consistent? Objectivity versus subjectivity? As the title implies, these are open questions, that is questions with no definitive answer (and therefore not for assessment).

Unit schedule

Semester 2, 2022 starts on Monday 25th of July

Week	Week commencing	Module	Topic	Assessment
1	25 July	Intro	Introduction Students have one week to refresh R independently	Students must install Rstan
2	01 August	Module 1	Bayesian concepts	
3	08 August	Module 1	Bayesian concepts	
4	15 August	Module 2	Continuous priors – conjugate distributions	M1 exercises due 15/08
5	22 August	Module 2	Continuous priors – conjugate distributions	
6	29 August	Module 3	MCMC fitting and assessing convergence	M2 exercises due 29/08
7	05 September	Module 3	MCMC fitting and assessing convergence	**No** M3 exercises
8	12 September	Module 4	Model evaluation and comparison	Major assignment 1 due 12/09
9	19 September	Module 4	Model evaluation and comparsion	
	26 September	-	**mid semester break**	M4 exercises due 26/09
10	03 October	Module 5	Hierarchical models and shrinkage	
11	10 October	Module 5	Hierarchical models and shrinkage	
12	17 October	Module 6	Bayesian workflow and open problems	M5 exercises due 17/10
13	24 October	-	-	**NO** M6 exercises
				Major assignment 2 due 11/11

Assessment

Assessment will include 2 written assignments worth 30% each, to be made available in the middle and at the end of the semester, and to be completed within approximately 3 weeks. These assignments will be posted on the eLearning site together with an online Announcement broadcasting their availability. In addition, students will be required to submit solutions to selected practical exercises (one assignment for each for modules 1, 2, 4 and 5), worth a total of 40%, by deadlines specified throughout the semester (see table below).

Assessments are due by 11:59pm on the stated day.

Assessment name	Assessment type	Coverage	Learning objectives	Weight
Module 1 exercises	Assignment	Module 1	1,2,3	10%
Module 2 exercises	Assignment	Module 2	1,2,3	10%
Major Assignment 1	Assignment	Modules 1-3	1,2,3,5	30%
Module 4 exercises	Assignment	Module 4	3,4	10%
Module 5 exercises	Assignment	Module 5	3,4	10%
Major Assignment 2	Assignment	Modules 1-5	3,4,5	30%
Online quizzes	Non-assessed	Various	various	-

In general you are required to submit your work typed in Word or similar (e.g. using Microsoft's Equation Editor for algebraic work) and we strongly recommend that you become familiar with equation typesetting software such as this. If extensive algebraic work is involved you may submit neatly handwritten work, however please note that marks will potentially be lost if the solution cannot be understood by the markers due to unclear or illegible writing. This handwritten work should be scanned and collated into a single pdf file and submitted via the eLearning site. See the BCA Assessment Guide document for specific guidelines on acceptable standards for assessable work.

The instructors will generally avoid answering questions relating directly to the assessable material until after it has been submitted, but we encourage students to discuss the relevant parts of the notes among themselves, via eLearning. However **explicit solutions to assessable exercises should not be posted for others to use**, and each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source.

We acknowledge the boundary between seeking clarification to help understand the question and seeking help to answer the question is not perfectly clear. Questions such as 'is … a typo and should it be … instead' can posted on the eLearning and/or emailed to the coordinator. Question such as 'is … the correct value' are clearly off-limits, however, issues with software are more ambiguous 'code … fails with error …' may reveal your attempted solution even you didn't intend for that. If you are unsure which category your question belongs to, email is recommended. In addition, we encourage students to discuss the relevant parts of the notes among themselves, via eLearning, as long as explicit solutions to assessable exercises are not posted for others to use, and each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source.

Submission of assessments and academic honesty policy

You should submit all your assessment material via eLearning unless otherwise advised. The use of Turnitin for submitting assessment items has been instigated within unit sites. For more detail please see pages 3-5 the BCA Student Assessment Guide.

The BCA pays great attention to academic honesty procedures. Please familiarise yourself with the procedures and policies at your university of enrolment. Links to these are available in the BCA Student Assessment Guide. When submitting assessments using Turnitin you will need to indicate your compliance with the plagiarism guidelines and policy at your university of enrolment before making the submission.

Late submission of assessments and extension procedure

The standard BCA policy for late penalties for submitted work is a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 10 days (including weekends and public holidays). Extensions are possible, but these need to be applied for (by email) as early as possible. The Unit Coordinator is not able to approve extensions beyond three days; for extensions beyond three days you need to apply to your home university, using their standard procedures.

Learning resources

There is **no** required textbook. The subject notes are self-contained.

Content-wise the closest, **optional**, textbook is Lambert B. A Student's Guide to Bayesian Statistics. SAGE Publications Ltd, 2018. Where possible, it is indicated in the subject notes which chapter/section/subject of Lambert's textbook corresponds to the subject notes.

The subject notes were also inspired by McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and Stan, CRC Press / Taylor and Francis / Chapman and Hall, 2016.

Software

For this subject you will need to have access to:

R, including the packages rstanarm and brms. These will in turn pull in RStan. ggplot2 and other libraries. All this software is free and gratis. RStan requires R version 3.4 or later. The subject notes were prepared using R 4.2.1.

The RStan team "strongly" recommends that those who use RStudio use version 1.4 or later. The coordinator does not use RStudio but assumes that students who use RStudio are better off upgrading to 1.4.

Linux users should be aware that R's install.package() occasionally creates conflicts when mixed with the distribution's package manager (apt, rpm, ...). The unit coordinator prefers to have all R packages installed via the linux distribution. Some distributions have limited packages available from the package manager by default. Please see https://cran.r-project.org/bin/linux/ for instructions on adding software repositories. It is plausible that installing everything using install.package() also avoids conflicts.

Note that installing brms **may** require installing the developer tools (Rtools) for R if these are not yet installed on your system. It is more likely that they are missing on MS Windows but even then, often the install.packages() command will pull them in as a dependency. If you don't use linux, please try installing by using via the RStudio interface or by install.packages(c('rstanarm','brms')) **before** attempting instructions from other tutorials found online.

Please install RStan and verify installation in the first week of the unit, so there is sufficient time to troubleshoot issues. To verify, load the rstan library and then enter the command below in an R session:

Example (stan_model, package = "rstan", run.dontrun = TRUE)

This will spit out strings of text, will take a few minutes to start sampling, and may provide a warning. It should NOT result in an error.

Feedback

Our feedback to you:
The types of feedback you can expect to receive in this unit are

- Formal individual feedback on submitted exercises/assignments
- Feedback from non-assessed online quizzes
- Responses to questions posted on the Discussion Board and in Tutorials

Your feedback to us:
One of the formal ways students provide feedback on teaching and their learning experience is through the BCA student evaluations at the end of each unit. The feedback is anonymous and provides the BCA with evidence of aspects that students are satisfied with and areas for improvement.

Required mathematical background

Students should be familiar with integration and approximation. Fortunately, for almost all practical applications, the analytical problems have already been translated into the R software by the developers of Stan, Rstan, rstanarm, loo and brms.

Unit Changes, including response to recent student evaluation

BAY was last delivered in Semester 2 2020.

For the 2020 delivery we changed from pure Stan (C++ like syntax) to rstanarm as this reduces the complexity of using the software. At the time, the subject notes were **adapted** to facilitate this, however, the emphasis and order of topics remained unchanged.

In 2020, it was allowed to use Stata for some of the early assignments. Prior to 2020 this made sense as both R and Stata users had to switch to Stan for the second half of the subject. Using rstanarm and brms, both in R syntax, reduces the workload for students. Alas, switching from Stata to R mid-way does little to reduce the double burden of learning new theory and learning new syntax.

BAY now uses the R software from the first module until the last module, from easier tasks to posterior prediction from hierarchical linear models. To guide students in using R and refresh their memory, more examples of “easy” tasks are included in the subject notes. Students are encouraged to spend the first week of the semester refreshing their knowledge of R independently. Finally, all scripts included in the subject notes are available on canvas, to facilitate experimenting with different settings.

Since 2020, the subject notes have been completely revised. There are no changes in learning outcomes and generally the same topics are covered. There are changes in emphasis, corresponding to changes in available software, and some reordering of content. The following were reduced:
1. There is now only one module on conjugate priors and analytical derivations of the formulae for conjugate priors, down from two modules. More complex conjugate families are mentioned but the derivations are omitted as they are rarely used in practice.
2. Generalised Least Squares is no longer mentioned. This option is rarely used and is not an intermediate step between techniques mentioned earlier or later; it is not useful as a stepping stone towards MCMC.

Instead, more attention is given to:
3. The explanation on different views on probability is extended in module 1.
4. An early example of Bayesian workflow, including prior elicitation is included in module 1
5. Examples of model comparison are extended, that is more examples of using models fitted with MCMC
6. Model stacking is added to Module 4
7. Bayesian workflow and contemporary issues are discussed more extensively in Module 6 (not for assessment)

The new subject notes follow roughly the same progression as the previous versions, however, hierarchical models are deferred to Module 5. They will be covered to the same extent in said module. By postponing hierarchical models to module 5, students have had more time to become familiar with MCMC software. This separates learning about R usage and MCMC from learning about hierarchical models and shrinkage; it aims to smooth the learning curve.

Acknowledgments

New subject notes are based on the old subject notes written by the previous coordinator of the unit, by Professor Lyle Gurrin. Additional inspiration was drawn from Ben Lambert's textbook and Richard McElreath's textbook, and almost certainly influenced by mc-stan.org tutorials and Andrew Gelman's blog. The new subject notes were edited and proofread by Professor Lyle Gurrin.