Introduction to Bayesian Modeling

Contents

1 Overview **3**

1.1 Goal 3
1.2 Date and Time 3
1.3 Target Audience 3
1.4 Prerequisites 3
1.5 Textbook 3
1.6 Structure 3
1.6.1 Philosophy 3
1.6.2 Format 3

2 Topics **4**
2.1 Why Bayes? 4
2.2 Fully conjugate models 4
2.2.1 Methods 4
2.2.2 Models 4
2.3 Semi-conjugate models 5
2.3.1 Methods 5
2.3.2 Models 5
2.4 Non-conjugate models 5
2.4.1 Methods 5
2.4.2 Models 6

A Student mini-presentations **6**
A.1 Intro 6
A.2 Model Checking 6

B Student mini projects **7**
B.1 Missing data 7
B.2 Batting average dataset 7

C Some topics we won’t get to **7**

D Notes to self: some details for when slides are constructed **8**
D.1 Why Bayes 8
D.2 Introduction to inference 8
D.3 Hierarchical models 8

1 Overview

1.1 Goal

The goal of this workshop is to introduce students to the concepts and practice of Bayesian modeling. We will begin by motivating Bayesian approaches. Next, we will introduce and apply models with conjugate priors, such as Bayesian normal, Bayesian binomial, and Bayesian linear regression. We will then introduce the two primary techniques for approximate Bayesian inference, namely Markov Chain Monte Carlo (MCMC) and variational inference. Using these techniques, and in some cases clever trickery, we will then tackle models for which there are not conjugate priors, such as Bayesian logistic regression, Bayesian multiclass regression, Bayesian mixture models, and Bayesian hidden Markov models. Finally, we will very briefly discuss Bayesian deep learning. For applications, we will use Python; namely, a combination of pymc3, scikit-learn, and code we write ourselves.

1.2 Date and Time

The workshop will be held via Zoom, June 7-11, from 2pm-5pm EST.

1.3 Target Audience

We expect that the typical student will be a graduate student, faculty member, staff member, or researcher in a quantitative field (such as computer science, statistics, engineering, or biology), who would like to learn more about Bayesian modeling.

1.4 Prerequisites

Prerequisites include calculus, some familiarity with introductory linear algebra (matrix multiplications, determinants, and traces), and some familiarity with introductory probability (e.g., we will assume prior fluency with concepts such as expectation, conditional probability, and commonly used distributions, such as Gaussian and Poisson). The class will use Python as a common language. The workshop will employ student-centered components; for maximum benefits, we strongly encourage setting aside 1-2 hours per day outside of the workshop to work on material.

1.5 Textbook

A textbook is not needed for this workshop. However, three excellent resources are:

1. Peter Hoff's textbook [1]
2. Christopher Bishop's textbook [2]
3. Andrew Gelman et al.'s textbook [3], available online at `http://www.stat.columbia.edu/~gelman/book/`.

We will draw especially heavily from [1].

1.6 Structure

1.6.1 Philosophy

Learning in order to create is both more fun and more effective than *learning for some extrinsic purpose*. Hence, the workshop is structured so as to (a) be student-centered and (b) allow self-determination and autonomy in how students engage with the material.

1.6.2 Format

A portion of the workshop will be interactive, including:

- Interactive exercises:
  - Google Collab exercises using iPython notebook

  - Constructing and/or working with Python (rather than R) implementations of [1] and [3].
- Student “lightning chat" (5 minute) presentations. To allow for student-centered direction and autonomy, students may choose any of the following:
  - Selecting an idea from the "Lightning Chat ideas" section of the Appendix.
  - Reviewing a demo with us from `https://github.com/avehtari/BDA_py_demos`.
  - Presentation of Python implementations of models from [1] , [3], or the workshop.
  - Presentation of an exercise, e.g. from [3] or [1].
  - Presentation of a reading section, blog, etc. of interest.
  - Presentation of a mathematical derivation of something relevant to the course.
  - Presentation of how a concept relates to something from their research area.
- (Maybe) Mini reading group discussions – We likely won’t have time to read a whole paper, but we could discuss sections of a text at the very beginning, or perhaps sections of a relevant paper.

2 Topics

Below are topics we plan to cover in the workshop.

We organize the workshop around three different modeling situations: (a) fully conjugate models, (b) semi-conjugate models, and (c) non-conjugate models. Note that models are typically named after the likelihood component (e.g. “Bayesian multivariate normal"), and so a named model often fit into multiple categories, depending on the choice of prior.

2.1 Why Bayes?

We motivate the Bayesian approach.

2.2 Fully conjugate models

We present everything in here using conjugate models with closed-form posteriors. The models are useful in and of themselves, as well as to build intuition for more complicated models.

2.2.1 Methods

- **Belief functions, Bayes rule** – Sections 2.1, 2.2 of [1]. [4] briefly overviews the Bayesian framework. For important mistakes in real life in medicine and law, see [7]. *Why most published research findings are false* [8] provides nice additional motivation in science. Could perhaps cover exchangeability here.
- **Exponential family techniques** - If time. See [9].

2.2.2 Models

- **Single-parameter conjugate models** - Chapter 11.2 of [10] has a nice very brief introduction to inference. Sections 3.1 and 3.2 of [1] cover the binomial and Poisson models. Introduce the exponential family formalism (see [9]; see also Section 5.2 of [10]) for much greater breadth.
- **Bayesian multivariate normal** - Here we introduce the Bayesian MVN, a fundamental model. In this section, we present the Bayesian MVN as a conjugate model, but note that it (like many other models) can also be semi-conjugate, depending on the choice of prior.

2.3 Semi-conjugate models

To do inference with semi-conjugate models, we need to introduce MCMC and VI. These are the two primary methods for doing inference on Bayesian models without closed-form posteriors. In Section 2.4, we will push these methods further to the non-conjugate case.

2.3.1 Methods

- **Intro to MCMC sampling** - Basic sampling methods - inverse cdf, rejection sampling. Basic Markov Chain Monte Carlo (MCMC) - Metropolis Hastings, Gibbs sampling.
- **Introduction to variational inference** - Coordinate ascent variational inference (CAVI) [11].

2.3.2 Models

We provide semi-conjugate formulations of the following models. With such formulations, inference prototypically proceeds via Gibbs sampling or CAVI.

- **Bayesian multivariate normal** - The Bayesian MVN can be semi-conjugate with a certain choice of prior. Section 7 of [1] covers this model. The semi-conjugate formulation extends naturally to support imputation for missing data.
- **Mixture models** - Brief presentation of Gaussian mixture models; will occur alongside CAVI.
- **Hidden markov models** - We will probably need to skip this – or make it super fast – in the interest of time. But might give brief presentation of hidden markov models[1]. May give short overview to probabilistic graphical models here.
- **Bayesian linear regression** – Section 9 of [1]. I have notes on this. There are some nice slides here which also illustrate the use of kernels.[2] Introduce model selection here as well (Section 9.3 of [1] or Section 3.4 of [2]). Section 11.2.2 of [10] also has a nice, quick summary of Bayes factors for model comparison.
- **Simple hierarchical models** - Section 11.4 of [10] has a nice brief overview. Hierarchical normal model (e.g. Gelman's 5 schools example). Hierarchical linear regression (Chapter 13 of [3], Secs 11.1-11.3 of [1].). Figure 11.1 and 11.3 (right) of [1] nicely shows the beneficial effect of sharing statistical strength in a hierarchical linear regression, as compared to many separate linear regressions.

2.4 Non-conjugate models

Here, again, we use MCMC and VI, but we need to employ more advanced (or at least more general) techniques.

2.4.1 Methods

- **Gradient-based sampling** - A brief introduction to Hamiltonian Monte Carlo (HMC) and No U-Turn Sampler (NUTS). Introduction to the python package `pymc3`.
- **Model augmentation** - An auxilliary variable may be introduced to transform a non-conjugate model into one that is conjugate or semi-conjugate.
- **Variational inference for non-conjugate models** - We will probably need to skip this in the interest of time. Common strategies are Laplace method VI, Delta method VI, and automatic differentiation VI.

2.4.2 Models

- **Categorical models** - Includes various kinds of regressions (logistic, probit regression, binomial, etc.) Cover the auxilliary variable trick, if there is time. See also pp. 390 of [1] for a useful warm-starting strategy.
- **A racially polarized voting model** - The model loses conjugacy, because you take a linear combination of your beta-distributed parameters to get the parameter you feed into a binomial.

[1] Nice reference for frequentist HMM's: `http://jwmi.github.io/ASM/5-HMMs.pdf`

[2] Nice Bayesian linear regression slides: `https://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/slides/lec19-slides.pdf`

- **Bayesian deep learning models** - Bayes and neural networks. 20-30 min w/ guest presenter, Kyle Heuton, Ph.D. student, computer science.

References

[1] Peter D Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009.

[2] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[3] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[4] Zoubin Ghahramani. Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110553, 2013.

[5] Leonhard Held and Chris C Holmes. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168, 2006.

[6] Andrew Gordon Wilson. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

[7] *The obscure maths theorem that governs the reliability of Covid testing*. Available at `https://www.theguardian.com/world/2021/apr/18/obscure-maths-bayes-theorem-reliability-covid-lateral-flow-tests-probability`.

[8] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[9] Michael Wojnowicz. *The exponential family*. Available (with permission).

[10] Anthony Christopher Davison. *Statistical models*, volume 11. Cambridge university press, 2003.

[11] Michael Wojnowicz. *Foundations of variational inference*. Available (with permission) at `https://github.com/mikewojnowicz/vi_foundations`.

[12] Ray Hill. Multiple sudden infant deaths–coincidence or beyond coincidence? *Paediatric and perinatal epidemiology*, 18(5):320–326, 2004.

[13] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

[14] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019.

A Student lightning chat ideas

A.1 Intro

- Present *Why most published research findings are false.* [8].
- Present Bayesian analysis of multiple sudden infant deaths. [12].
- Present a piece of Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521(7553), 452-459.

A.2 Model Checking

- Textbook sections TBD.
- Present a section of Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ... Modrak, M. (2020). Bayesian workflow. arXiv preprint arXiv:2011.01808.

B Student mini projects

These are small problems to work on. Some are open-ended. These presumably would not be part of the workshop.

A.2 Model Checking

- Textbook sections TBD.
- Present a section of Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ... Modrak, M. (2020). Bayesian workflow. arXiv preprint arXiv:2011.01808.

B Student mini projects

These are small problems to work on. Some are open-ended. These presumably would not be part of the workshop.

B.1 Missing data

This project covers missing data and imputation. Imputation is made part of inference, rather than a separate step.

- Inference with a Bayesian normal model in the presence of missing data is given, in R code, on pp. 119 of [1]. Implement, and test, the code in Python.

B.2 Batting average dataset

The hierarchical normal model for (arcsine-transformed) batting average data on pp. 163 of [3] has some serious deficiencies, as exposed in Table 6.1 in the section on model checking.

Can you construct (and learn) a better model which makes predictions closer to the true final batting average?

Examples:

- Add an extra layer to the hierarchy, so that player p's 1970 batting average inherits from player p's overall batting average which in turn inherits from a population batting average. (Of course, I am speaking of the arcsine-transformed batting averages, so that we can use a hierarchical normal model.)
- Add an autoregressive component, because, as mentioned by Gelman, player batting averages *DO* change over time.

The text also does a poor job of checking the modeling assumption violations that were of concern. Can you do a better job of checking them, and if necessary, address them?

Examples:

- If batting averages are indeed heavy tailed or skewed, move from a normal distribution to something else. For example, could try a t-distribution with Laplace inference to handle the non-conjugacy.
- If the variance is indeed too high for a binomial model, try something that can handle the overdispersion.

C Some topics we won't get to

- **Bayesian workflow** – Lots of nice resources for Bayesian workflow. For example: [13] or [14]. Section 6 of [3] covers model checking, as does Section 11.2.3 of [10]. Some points to make re: model checking
  - *Samples from the posterior predictive should capture important properties of the observed dataset.* For a violation of this, see the normal model for Newcomb's speed of light measurements. (Compare Figures 6.2 and 3.1 of [3].)
- **Bayesian GLM's and GLMM's** - Bayesian GLM's would extend the section on "Regression models for binary and multi-class data" to other distributions, and present models in

- **VI for nonconjugate models** - For instance, we could cover Laplace and Delta Method VI as applied to categorical models. We could also cover black box / automatic differentiation VI.

D Notes to self: some details for when slides are constructed

D.1 Why Bayes

Original sketch See Section 1.2 of [1]. [2] has some nice plots motivating why use Bayesian linear regression over standard linear regression. [4] has some nice plots illustrating the Bayesian approach and how it mitigates overfitting. I can provide a nice example with biometric profiling of human typing dynamics. [5] has a nice simple example of obtaining non-standard functionals from the posterior that can be of interest. [6] presents the case for Bayesian deep learning.

Bayes law in real life The primary motivation here is the importance of not "flipping the conditional" during interpretation. Sally Clark was convicted of murdering her two children, because the chance of two babies dying of SIDS in one family was one in 73m. But the expert witness ignored the prior probability that someone was a double murderer [7]. (See [12] for discussion.) Could also refer to the classic example with positive tests - maybe in reference to COVID [7]. *Why most published research findings are false* [8] provides a third example.

Modeling Application: Estimating the probability of a rare event See Section 1.2.1 of [1]. The problem is to estimate the proportion of people in a small town that have a disease given a small sample of 20 individuals. This is a nice believable example in which prior information is natural: we use prevalence in similar towns. (Make a note that really this is foreshadowing hierarchical models and hierarchical regressions!) This also illustrates another nice property of Bayes – we can get lots of functionals from the posterior. The plot on the right of Figure 1.1 – showing the shift between the prior and posterior distributon – is a classic. The sensitivity analysis is pretty cool – see the right hand plot of Figure 1.2, as well as the last couple sentences on pp.7.

The REAL kicker, I think, is the comparison to non-Bayesian methods. The frequentist confidence interval is complete garbage here, and the "adjusted" Wald interval is clearly just a (very specific) choice of .prior. Nice opportunity for discussion here. Ask: what's the advantage of Bayes over that? Possible answers: That is seemingly ad hoc; it's not flexible to other choice of priors; Bayes makes it clear how it relates to priors; and it doesn't allow for sensitivity analysis (or investigation of various functionals).

D.2 Introduction to inference

Bayes factors Very nice brief discussion in Section 11.2.2 of [10].

Multivariate normal - missing data and imputation See Section 7.5 of [1].

D.3 Hierarchical models

Some motivations:

- A way to use "surrounding data" as a prior in a more formal way. Think back to Hoff's disease prevalence example. In that case, we constructed our beta prior manually, by taking a couple of basic facts about similar towns and then converting that into beta parameters. A hierarchical model could let the prior expectation be tied more exactly to those surrounding towns (including, if a regression is involved, similarity w.r.t. relevant characteristics, such as size, SES, etc.), to automatically set the strength of the prior expectation according to the relative uncertainty within and between towns, and to automatically adapt as data rolls in.
- Can be convenient. Consider Wand's construction of the half-t distribution as a hierarchical model of inverse gammas, which makes for a conditionally conjugate scheme. Basically

any auxilliary variable trick (think Polya Gamma augmentation) can be considered as the construction of a hierarchical model for computational convenience.

- By adding layers to the hierarchy, we can escape bad assumptions – e.g., the beta-binomial model handles over-dispersion. (This might be a better example for "why Bayes".)