VE414 Lecture 1

Jing Liu

UM-SJTU Joint Institute

September 9, 2019

Course Information

- Course Description:

  The aim of this course is to introduce students to the Bayesian statistical modelling and inference, and to the related computational strategies and algorithms. Topics covered include: Principles of Bayesian statistics, Bayesian linear, hierarchical models, and generalised linear models, Bayesian Networks. Bayesian computational methods, including Gibbs sampler and Metropolis-Hastings algorithms, are presented with an emphasis to the issues related to their implementation and monitoring of convergence. Programming languages Julia, R and Stan are introduced.

- Who should take this class?

  The prerequisite for this class is computer/programming knowledge at the level of VE101 (or above), and statistics knowledge at the level of VE401 (or above). Both undergraduates and graduate ECE students are welcome to take the course.

Contact Information

- Instructor:

  Jing Liu

- Lectures:

Monday	(12:10pm – 1.50pm)	in F-404 (Odd weeks only)
Tuesday	(02:00pm – 3.40pm)	in F-404
Thursday	(02:00pm – 3.40pm)	in F-404

- Office Hours:

Tuesay	(09:00am – 11:00am)	in JI-Building 441A
Thursday	(09:00am – 11:00am)	in JI-Building 441A

- Email:

  stephen.liu@sjtu.edu.cn

- Teaching Assistant/s:

  See Canvas for his/her contact information

- To improve communication between the students and the teaching team please observe the following guidelines:
  - Any student facing a special situation likely to impact his studies, such as serious illness or official duty, is expected to contact the instructor as early as possible in order to discuss it and see if any solution can be found.
  - When sending an email related to this course please include the tag [VE414] in the subject e.g. Subject: [VE414] Question Gibbs.
  - When contacting your TA for a grade issue or any other major problem send a carbon copy (cc) to the instructor as well. Not doing it might result in omissions, not up-to-date grades etc. If such problem occurs and there is no record of the issue the request will be **automatically rejected**.
  - Never attach a large file ($>$ 2 MB) to an email, use a Dropbox type of service instead and only include a link in the email.
  - Keep in touch with the teaching team, feedbacks and suggestions will be much appreciated.

Grading Policy

- Assignment:

  25% There will be eight assignments.

- Project:

  25% There will be a project in the form of a challenge.

- Exam:

  50% There will be two exams:

Midterm	Final
25%	25%

- Quiz (Optional):

  15% Quizzes will be given frequently in class.

- For those who attempt all quizzes, their grade is whichever is the higher of:

  0. 25% Assignment + 25% Project + 0% Quiz + 50% Exam
  1. 25% Assignment + 10% Project + 15% Quiz + 50% Exam
  2. 25% Assignment + 40% Project + 15% Quiz + 20% Exam

- For this course, the grade will be curved to achieve a median grade of "B".

Project

- Each of you need to be in one and only one 3-member team for the project.
- The project will be graded according to the following three aspects:
    1. Oral Presentation of your model
    2. Poster Presentation of your model
    3. Prediction Accuracy of your model

  each of those three aspects has an equal weight.
- Each member of the same team will receive the same project mark.
- You will be working on a simulated data, part 1 of which will be given to you,

part 2 will not be given to you, instead it will be used to assess your model.

Honour Code

- Honesty and trust are important. Students are responsible for familiarising themselves with what is considered as a violation of honour code.
- Assignments/projects are to be solved by each student individually. You are encouraged to discuss problems with other students, but you are advised not to show your written work to others. Copying someone else's work is a very serious violation of the honour code.
- Students may read resources on the Internet, such as articles on Wikipedia, Wolfram MathWorld or any other forums, but you are not allowed to post the original assignment question online and ask for answers. It is regarded as a violation of the honour code.
- Since it is impossible to list all conceivable instance of honour code violations, the students has the responsibility to always act in a professional manner and to seek clarification from appropriate sources if their or another student's conduct is suspected to be in conflict with the intended spirit of the honour code.

Programming Language

From the creators of Julia

We are power Matlab users. Some of us are Lisp hackers. Some are Pythonistas, others Rubyists, still other Perl hackers. We generate more R plots than any sane person should. C is our desert island programming language.

We love all of these language; they are wonderful and powerful. For the work we do: scientific computing, machine learning, data mining, large-scale linear algebra, distributed and parallel computing, each one is perfect for some aspects of the work and terrible for others. Each one is a trade-off. We are greedy, we want more.

We want a language that is open source, with a liberal license. We want the speed of C with the dynamism of Ruby. We want a language with true macros like Lisp. But with obvious, familiar mathematical notation like Matlab. We want something as usable for general programming as Python, as easy for statistics as R, as natural for string processing as Perl, as powerful for linear algebra as Matlab, as good at gluing programs together as the shell. Something that is dirt simple to learn, yet keeps the most serious hackers happy. We want it interactive and we want it compiled. Did we mention it should be as fast as C? In 2012, we set out to create the language of our greed—the language we have created is called Julia.

- Julia is a general-purpose programming language.

  `https://julialang.org/`

- R is a programming language for statistical computing and graphics.

  `https://mirrors.shu.edu.cn/CRAN/`

- Stan is a programming language for Bayesian inference and diagnostics.

  `https://mc-stan.org/`

- Julia

  We will use it to implement new, or intrinsically slow algorithms.

- R

  We will use it to do small tasks, generate plots and as an interface.

- Stan

  We will use it to run existing algorithms and diagnostics in ShinyStan.

Textbook

- Textbook:

  Gelman et al. (2014), Bayesian Data Analysis

- Some Additional Material:

  Liu (2001), Monte Carlo Strategies in Scientific Computing

  Cox (2006), Principles of Statistical Inference

  Scutari and Denis (2015), Bayesian Networks

  Press et al. (2002), Numerical Recipes

  Grolemund and Wickham (2016), R for Data Science

- Other course related materials will be available on Canvas.

Teaching Schedule

Week	Topics	Others
1	Introduction Probability and Inference Conjugate Prior	
2	Noninformative and Weakly informative Prior Hierarchical Models	
3	Decision Analysis Quadrature and Laplace Rejection Sampling	A1 due
4	**National Holiday**	
5	Importance Sampling Case Study (Optional) Stochastic process	A2 due
6	Discrete Markov Chain Continuous Markov Chain	A3 due

7	Gibbs Sampling Metropolis-Hastings Algorithm Expectation-Maximisation Algorithm	A4 due
8	Case Study (Optional) **Midterm**	A5 due
9	Hierarchical Model revisited Bayesian Linear Models Generalised Linear Models	
10	Nonparametric Models Splines	A6 due
11	Finite Mixture Models Hidden Markov Models (Optional) Graphical Models	A7 due
12	Bayesian Networks Inference in Bayesian Networks	A8 due Poster
13	Learning in Bayesian Networks Algorithms for Bayesian Network (Optional) Project Presentation	
14	**Final Exam**	

- Bayesian analysis is actually not based on any typical introductory topics:
  - Modern probability theory (Kolmogorov 1931)
  - t-test (Gosset 1908)
  - Maximum likelihood estimation (Fisher 1912)
  - Confidence interval (Neyman 1937)
  - Regression analysis (Legendre 1805 and Gauss 1809)

  which are typical frequentist approach/data analysis.
- Not only the two approaches do not build on each other, there is actually a religious war between the two communities for more than a century!
- The Reverend Thomas Bayes (1702-1761) is believed to be the first person to come up with the essential idea of Bayesian analysis.
- The idea was also developed independently by someone far more prominent,

  Pierre Simon Laplace (1749-1827),

  who gave its modern mathematical form and scientific application.

- However, Laplace mistakingly believed Bayesian is unnecessary since it tends to give the same results as frequentist in the presence of a large dataset.
- Laplace at age 62, likely to be the world's first true Bayesian, converted to be a frequentist, and he remained so for the remaining 16 years of his life.
- After Laplace's death, many mathematicians/statisticians, most noticeably

  Sir Ronald Fisher (1890-1962),

  succeeded in marginalising the approach over several controversial debates.
- In addition to one-sided philosophical debates, Bayesian approach can be notoriously expensive to compute, thus difficult to illustrate in early days.
- Bayesian data analysis became increasingly accepted during the second half of the 20th century, with modern computing power and new computational approaches, Bayesian analysis became mainstream in the past 20 years.
- However, misunderstandings towards Bayesianism still remain, despite being popular in decision theory, machine learning, and artificial intelligence.

Theorem (Bayes' Theorem)

Let $\mathcal{A}$ and $\mathcal{B}$ be two events, and $\Pr(\mathcal{B}) \neq 0$, then

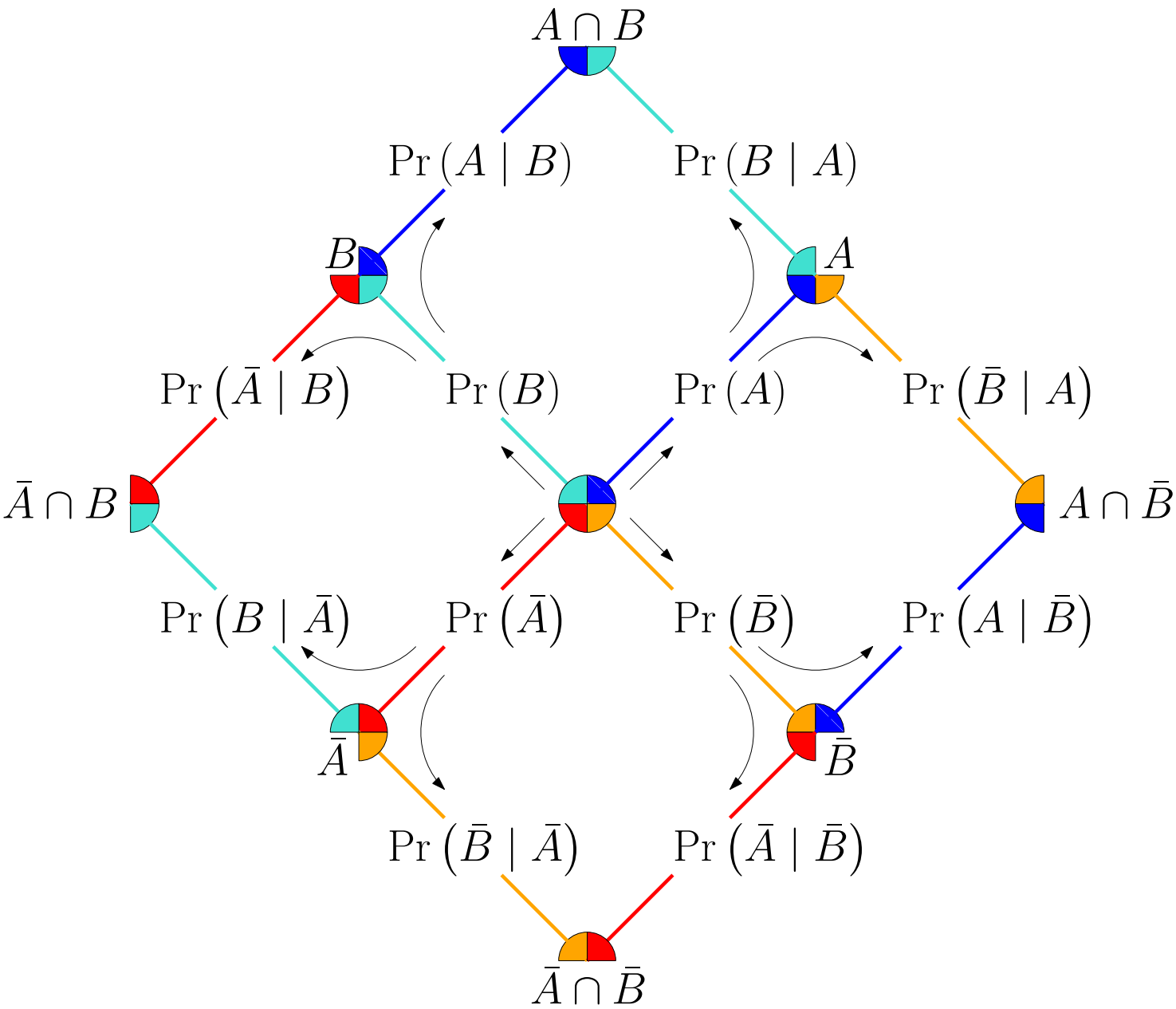
$$\Pr(\mathcal{A} \mid \mathcal{B}) = \frac{\Pr(\mathcal{B} \mid \mathcal{A}) \Pr(\mathcal{A})}{\Pr(\mathcal{B})}$$

In general, for some partition $\{\mathcal{A}_j\}$ of the sample sample, then

$$\Pr(\mathcal{A}_i \mid \mathcal{B}) = \frac{\Pr(\mathcal{B} \mid \mathcal{A}_i) \Pr(\mathcal{A}_i)}{\sum_j \Pr(\mathcal{B} \mid \mathcal{A}_j) \Pr(\mathcal{A}_j)}$$

- It is named after Thomas Bayes, however, he actually has very little to do with this formula, it was Laplace that worked out the mathematical form.
- It is not difficult to understand using the concept of conditional probability.

Q: Can you give a tree diagram of the probability space for the simple form?


$A \cap B$
$\Pr(A \mid B)$
$\Pr(B \mid A)$
B
A
$\Pr(\bar{A} \mid B)$
$\Pr(B)$
$\Pr(A)$
$\Pr(\bar{B} \mid A)$
$\bar{A} \cap B$
$A \cap \bar{B}$
$\Pr(B \mid \bar{A})$
$\Pr(\bar{A})$
$\Pr(\bar{B})$
$\Pr(A \mid \bar{B})$
$\bar{A}$
$\bar{B}$
$\Pr(\bar{B} \mid \bar{A})$
$\Pr(\bar{A} \mid \bar{B})$
$\bar{A} \cap \bar{B}$

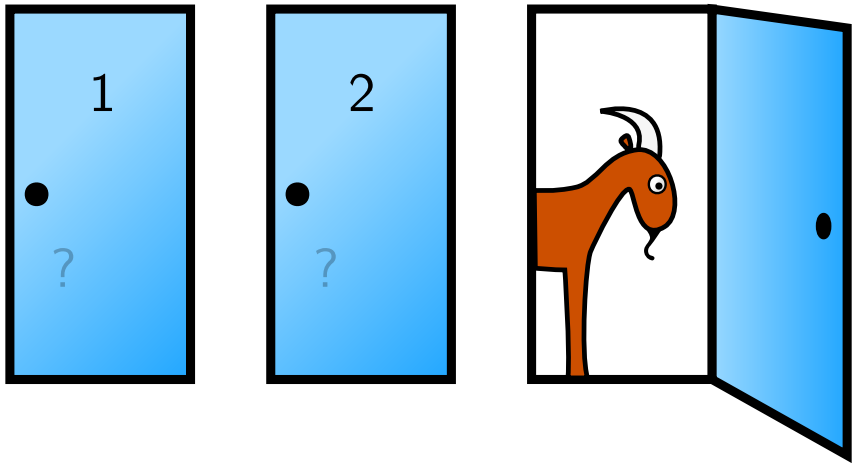
- The famous Monty Hall problem arises from a popular television game show

Let's Make a Deal

is often used to illustrate the Bayes' theorem and Bayesian inference.

Monty Hall Problem

Suppose you are on a game show, and you are given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 2, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to switch to door No. 1?"



Is it to your advantage to switch your choice?

- A typical solution:

  Let $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ denote the events in which the car is placed behind door No.1, No.2 and No.3, respectively. Without loss of generality, suppose we select door No.2 and Monty opens No.3.

$$\Pr(\mathcal{C}_2 \mid \mathcal{O}_3) = \frac{\Pr(\mathcal{O}_3 \mid \mathcal{C}_2)\Pr(\mathcal{C}_2)}{\sum_{j=1}^{3}\Pr(\mathcal{O}_3 \mid \mathcal{C}_j)\Pr(\mathcal{C}_j)} = \frac{1/2 \cdot 1/3}{1 \cdot 1/3 + 1/6 + 0 \cdot 1/3} = \frac{1}{3}$$

$$\Pr(\mathcal{C}_1 \mid \mathcal{O}_3) = \frac{\Pr(\mathcal{O}_3 \mid \mathcal{C}_1)\Pr(\mathcal{C}_1)}{\sum_{j=1}^{3}\Pr(\mathcal{O}_3 \mid \mathcal{C}_j)\Pr(\mathcal{C}_j)} = \frac{1 \cdot 1/3}{1 \cdot 1/3 + 1/6 + 0 \cdot 1/3} = \frac{2}{3}$$

  where $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$ denote the events in which Monty opens door No.1, No.2 and No.3, respectively.

uses Bayes' theorem to see switching is the right choice. But it often fails to attract one's attention to the reasoning behind this solution thus this choice.

- Before Monty opens door No.3 and reveals the goat behind it,

$$\Pr(\mathcal{C}_2) = \frac{1}{3} = \Pr(\mathcal{C}_1)$$

we have no reason to switch, however, the information provided by Monty,

$$\Pr(\mathcal{C}_2 \mid \mathcal{O}_3) = \frac{1}{3} \qquad \text{and} \qquad \Pr(\mathcal{C}_1 \mid \mathcal{O}_3) = \frac{2}{3}$$

allows us to improve our chance of winning the car.

- Using Bayes' theorem, we essentially updated our understanding regarding the likelihood of where the car is from the data (not behind door No. 3).
- We could introduce some random variables to make this more concise:

$$Y = \begin{cases} -1, & \text{if } \mathcal{C}_1 \text{ happens,} \\ 0, & \text{if } \mathcal{C}_2 \text{ happens,} \\ 1, & \text{if } \mathcal{C}_3 \text{ happens,} \end{cases} \qquad \text{and} \qquad X = \begin{cases} -1, & \text{if } \mathcal{O}_1 \text{ happens,} \\ 0, & \text{if } \mathcal{O}_2 \text{ happens,} \\ 1, & \text{if } \mathcal{O}_3 \text{ happens.} \end{cases}$$

- Notice how the information provided by Monty updated our understanding

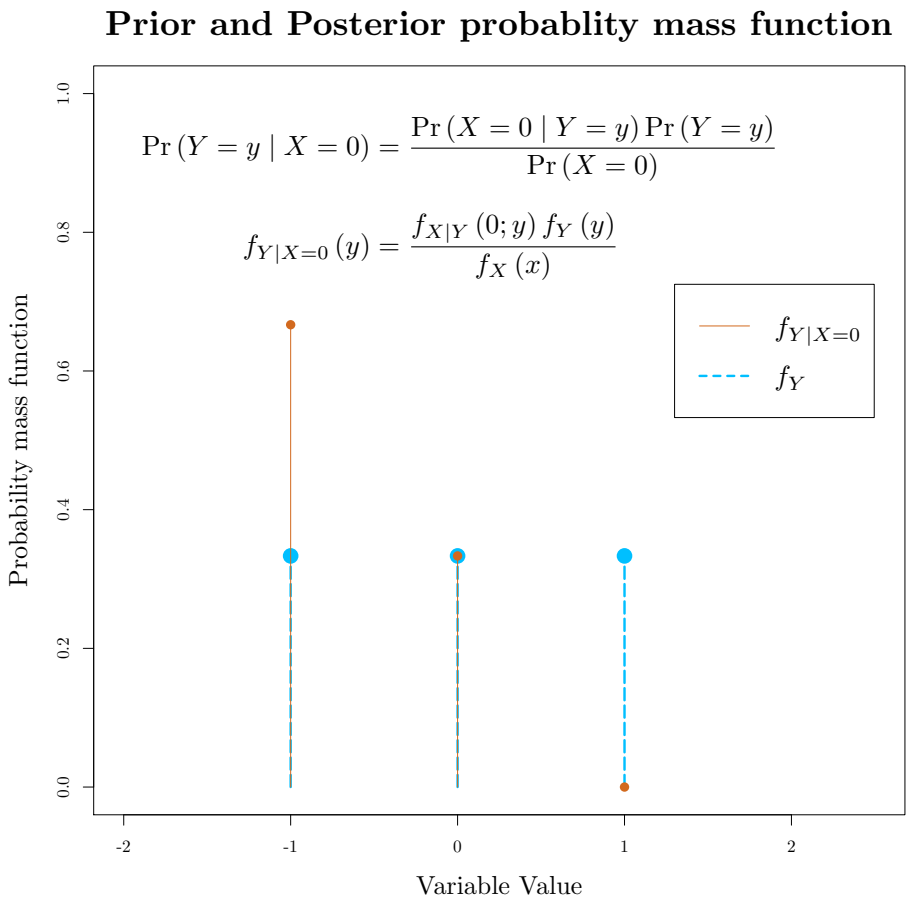


Figure: R Code: `monty_hall_prior_posterior_414.R`

Definition

In Bayesian analysis, without any data, the unobserved parameter or variable is modelled as a random variable Y having a marginal distribution

$$f_Y$$

This marginal distribution is known as the prior. The data is modelled as a random variable X given Y having a conditional distribution

$$f_{X|Y}$$

Given a particular realisation of the data $X = x$ is observed, the function

$$\mathcal{L}(y; x) = f_{X|Y}(x \mid y)$$

as a function of y only is called the likelihood. The conditional distribution of Y,

$$f_{Y|X=x}$$

given the observed $X = x$, is known as the posterior.

- This tiny problem gives a taste of what Bayesian inference is about, namely, updating one's understanding when data become available, i.e.

$$f_Y \longrightarrow f_{Y|X=x}$$

using Bayes' theorem

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x \mid Y = y)\Pr(Y = y)}{\Pr(X = x)}$$

$$\Longrightarrow \quad f_{Y|X=x}(y) = \frac{f_{X|Y}(x \mid y)\, f_Y(y)}{f_X(x)}$$

- However, using Bayes' theorem does not make someone a frequentist, using it for everything does! That is, Bayesianism is about creating models that are very much like the mechanics behind learning from experience, being a Bayesian means the learning rule is always some form of Bayes' theorem.