Bayesian Statistics and its Applications

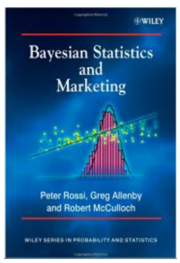
Instructor's Introduction

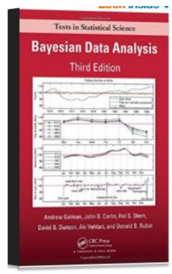
- 董晓静 xdong1@scu.edu

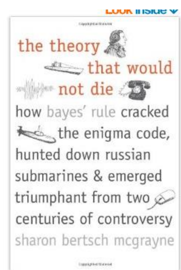
- Santa Clara University
- Associate Professor of Marketing and Business Analytics
- Founding Director of MS in Business Analytics

Course Introduction

- Introduction to Bayesian Modeling
  - Text book: Bayesian Statistics and Marketing
  - Reference book: Bayesian Data Analysis, Gellman et al.
  - Fun book: The Theory that would not Die
  - Programming: R, RStudio
  - Package: bayesm







Topics in this Course

- Topics to be covered
  - Bayesian concepts and Bayesian modeling
    - Experimental design
  - Markov Chain Monte Carlo
  - Gibbs sampling and data augmentation
    - Application in Probit model
  - Metropolis Hastings with random walk
    - Application in MNL model
  - Hierarchical Bayesian modeling
    - Application in individual level MNL model

Basic Concepts

Probability Definition

- Frequentist
  - If you repeat the same events many many times, the ratio between (1) number of times that the event happens and (2) the total number of trials.
- Bayesian
  - Subjective probability: personal belief

The Goal of Statistical Inference

Make inferences about unknown quantities using available information.

- Inference: make probability statements
- Unknowns: parameters, functions of parameters, states or latent variables, "future" outcomes, outcomes conditional on an action
- Information
  - data-based
  - non data-based
    - Theories of behavior; "subjective views" there is an underlying structure
    - Expert knowledges not included in the data
    - Parameters are finite or in some range

7

Bayes Theorem

$$p(\theta|D) = \frac{p(D,\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{p(D)} \propto p(D|\theta)p(\theta)$$

- In the context of model estimation
  - D is data
  - θ is model parameter
  - $p(D)$ is a constant
  - $p(D|\theta)$ is called the *likelihood* function, $l(\theta)$
  - $p(\theta)$ is the *prior* distribution of the model parameters
  - $p(\theta|D)$ is the *posterior* distribution of the model parameters

Posterior $\propto$ Likelihood $\times$ Prior

Bayesian Modeling

$$p(\theta|D) \propto l(\theta)p(\theta)$$

The posterior distribution of the model parameter conditional on the data is proportional to the product of

- the likelihood of the data and
- the prior distribution of the model parameters

- The result
  - Is a distribution rather than a point
  - Is conditional on the data collected, rather than based on sampling theory

Likelihood Function

- Likelihood function is proportional to the joint probability of the data to happen
- Likelihood functions
  - Normal random variables
  - Regression model
  - Binary Logit model
  - Multinomial Logit model
  - Multinomial Probit model

The likelihood Principle

$$p(D\,|\,q)\ \mu\ \ell(q)$$

Note: any function proportional to data density can be called the likelihood.

- LP: the likelihood contains **all** information relevant for inference. That is, as long as I have the same likelihood function, I should make the same inferences about the unknowns.

- In contrast to modern econometric methods (GMM) which does not obey the likelihood principle

Identification

$$R = \left\{ q : p\left(\text{Data}\middle| q\right) = k \right\}$$

If dim(R) >= 1, then we have an "identification" problem. That is, there are a set of observationally equivalent values of the model parameters. The likelihood is "flat" or constant over R.

Practical Implications

likelihood can have flats or ridges.

Issue for both the Bayesian (is it?) and non-Bayesian.

Identification

Is this a problem?

no, I have a proper prior

no, I don't maximize

"Classical" solution:

impose enough constraints so that constrained parameter space is identified.

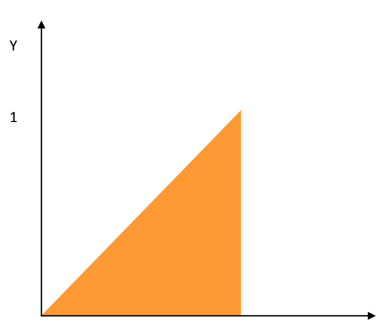
"Bayesian" solution:

use proper prior and recognize that some functions of θ are determined entirely by prior

13

Distribution Theory 101

- Marginal and Conditional Distributions:



$$p_X(x) = \int p_{X,Y}(x,y)\,dy$$

$$= \int_0^x 2\,dy = 2y\Big|_0^x = 2x$$

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{2}{2x}$$

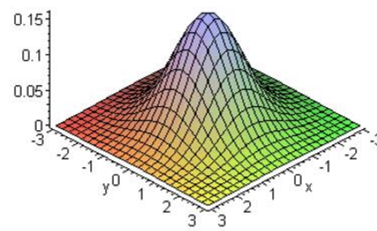
$y \in (0,x)$ uniform

Normal Distribution

- x and y follows a Bivariate Normal distribution

$$f(x,y) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left(-\frac{1}{2}\left(\binom{x}{y}-\mu\right)^T \Sigma^{-1}\left(\binom{x}{y}-\mu\right)\right)$$

- Marginal and conditionals are all Normal distributions



Conjugacy

How to Obtain Posterior

- Posterior distribution is proportional to a product of two distributions (prior and likelihood)
  - Conjugate – we know the class of the posterior distribution
  - Not conjugate – we don't know the distribution, simulation methods have to be adopted, Markov Chain Monte Carlo

Conjugacy

- A prior is conjugate to the likelihood, if the posterior derived from this prior and likelihood is in the same class of distribution as the prior.
- In modeling:
  - Step 1: write out the likelihood function
  - Step 2: see if it has conjugate prior. Use conjugate, if yes.
  - Step 3: estimate the model

Exponential Family

- Bernoulli model

$$\theta = \text{prob}(y_i = 1)$$

Likelihood function of the data is

$$p(y|\theta) = \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}$$

Prior: beta distribution

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \sim Beta(\alpha,\beta)$$

Posterior: beta distribution

$$p(\theta|y) \propto \theta^{\alpha+\sum_i y_i-1}(1-\theta)^{\beta+n-\sum_i y_i-1} \sim Beta(\alpha',\beta')$$

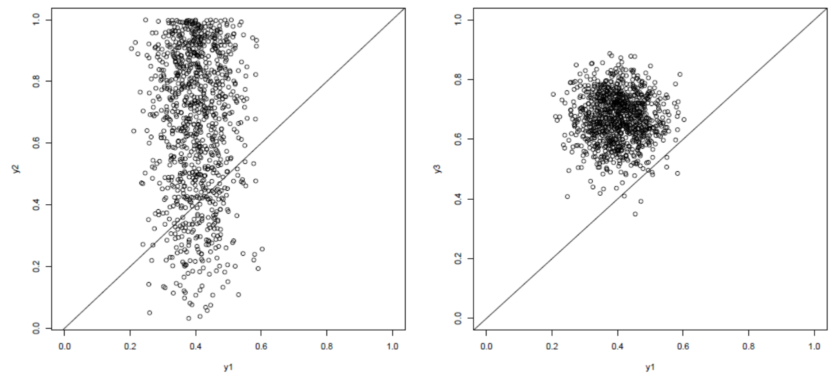
$$\alpha' = \alpha + \sum_i y_i\,, \qquad \beta' = \beta + n - \sum_i y_i$$

- Normal model
  - "completing the square"
  - Inverted Gamma distribution

Application: Thompson Sampling

- A/B/C test – three designs of a webpage to attract users to convert
  - Day 1, randomly assign 10 users to each design
    - Calculate the probability of conversion rate θ_A is the highest among the two, call it p_{1A}
    - Calculate the probability of conversion rate θ_B is the highest among the two, call it p_{1B}
    - Calculate the probability of conversion rate θ_C is the highest among the two, call it p_{1C}
  - Day 2, randomly assign users to each design based on the ratio of p_{1A}, p_{1B}, p_{1C}
    - Collect data, calculate the probability of p_{2A}, p_{2B}, p_{2C}
  - Day 3, randomly assign users to each design based on the ratio of p_{2A}, p_{2B}, p_{2C}

You can compute the probability that one is better than another



$X = P(\theta_1 | 20 \text{ successes}, 50 \text{ trials})$
$Y = P(\theta_2 | 2 \text{ successes}, 3 \text{ trials})$
$P(\theta_1 > \theta_2 | \mathbf{y}) \approx .16$

$X = P(\theta_1 | 20 \text{ successes}, 50 \text{ trials})$
$Y = P(\theta_2 | 20 \text{ successes}, 30 \text{ trials})$
$P(\theta_1 > \theta_2 | \mathbf{y}) \approx .009$

Benefits of Thompson Sampling

- It is dynamic
  - At each stage, collect some information about the performance of the designs, if design B is not attracting lots of conversions, the conversion rate θ_B could be low, therefore, not assign that many future users any more
- Therefore it reduces regrets by dynamically updating the design based on limited information collected, rather than waiting for the whole experiment finishes

Summary

- Basic concepts of Bayes theorem
- Conjugacy in the case of Binomial distribution
- Next class:
  - Markov Chain Monte Carlo method
  - conjugacy in the case of Normal distribution

Normal Model

- A set of random variable from normal distribution, knowing these variables $x_1, x_2, \dots, x_N$, use conjugate prior to calculate its mean μ, and variance σ^2

Likelihood function of the data, from the normal PDF

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(\sum_{i=1}^{N} -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

Prior distribution for $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$

Prior distribution for σ

$$p(\sigma^2) \propto (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\nu_0 s_0^2}{2\sigma^2}\right) \sim \frac{\nu_0 s_0^2}{\chi^2_{\nu_0}}$$

Posterior for $\mu|\sigma^2 \sim N\left(\mu_1, \sigma_\mu^2\right)$

$$\mu_1 = \frac{\frac{\sum_1^N x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}, \qquad \sigma_\mu^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

Posterior for $\sigma^2|\mu \sim \frac{\nu_1 s_1^2}{\chi^2_{\nu_1}}$, also called $scaled - inv - \chi^2(\nu_1, s_1^2)$, or $Inv - Gamma\ \left(\frac{\nu_1}{2}, \frac{\nu_1 s_1^2}{2}\right)$

$$\nu_1 = \nu_0 + N, \qquad s_1^2 = \frac{\nu_0 s_0^2 + N s^2}{\nu_0 + N}$$