Bayesian Data Analysis Discussion Group

University of Cincinnati

September 1, 2023

Why?

It's been 3 years, but I still cannot do bayesian data analysis well...

Process of Bayesian Data Analysis

1. Set up a full probability model. (Not much experience)

$$P(A, B) = \underbrace{P(B \mid A)}_{\textit{likelihood}} \underbrace{P(A)}_{\textit{prior}} \tag{1}$$

2. Condition on observed data. (ok)

$$\underbrace{P(A \mid B)}_{\textit{posterior}} = \frac{P(B \mid A)P(A)}{P(B)} \tag{2}$$

3. Evaluate the fit of the model (rarely do it) and implications of the resulting posterior distribution.

Search for MH370

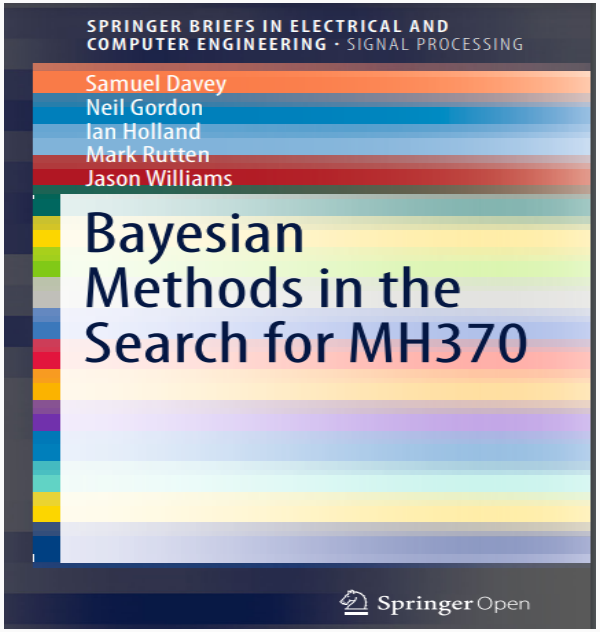


Figure 1: Book Cover

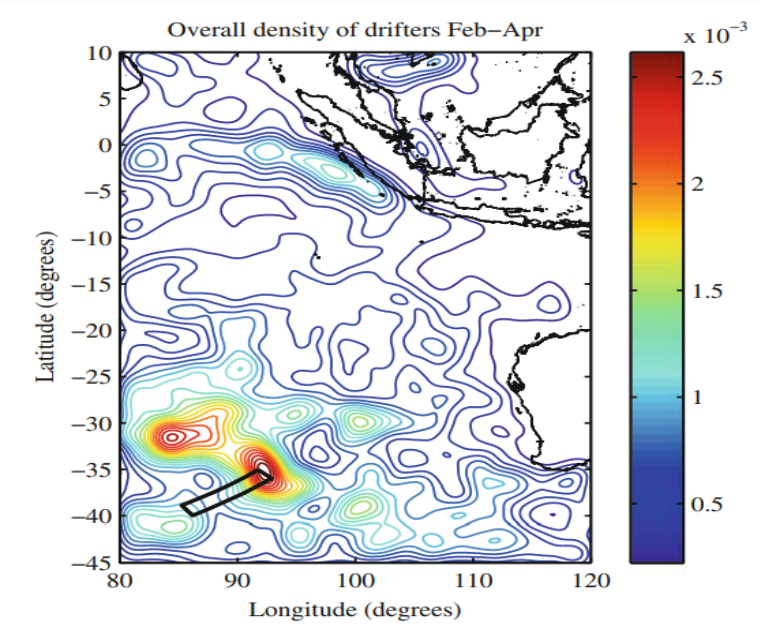


Figure 2: Posterior Density of MH370

- a prior of aircraft state (defined by the Malaysian military radar)
- a likelihood function describing the relationship between available measurements and the aircraft state vector

BDA Book Example and Problems Will Help

1. How to set up priors?

P48: estimate the cancer death rate in each county

$$\begin{aligned} y_j &\sim \text{Poisson}\left(n_j \theta_j\right) \\ \theta_j &\sim \mathit{Gamma}(20, 430000) \end{aligned} \tag{3}$$

And when should we use informative prior?

2. How to check and expand models? (Chapter 6 and 7)

3. How to take data collection process into the analysis? (Chapter 8)

Data Collection Process

3. Capture-recapture problem (P233)

A statistician/fisherman is interested in N, the number of fish in a certain pond. He catches 100 fish, tags them, and throws them back. A few days later, he returns and catches fish until he has caught 20 tagged fish, at which point he has also caught 70 untagged fish, of the 20 'tagged' fish, 15 are definitely tagged, but the other 5 may be tagged—he is not sure.

Go beyond the Book

BDA is good, but we are not gonna be stuck with it forever

Interesting Case Study

1. World Cup - Estimate the team abilities.

worldcup2012.txt

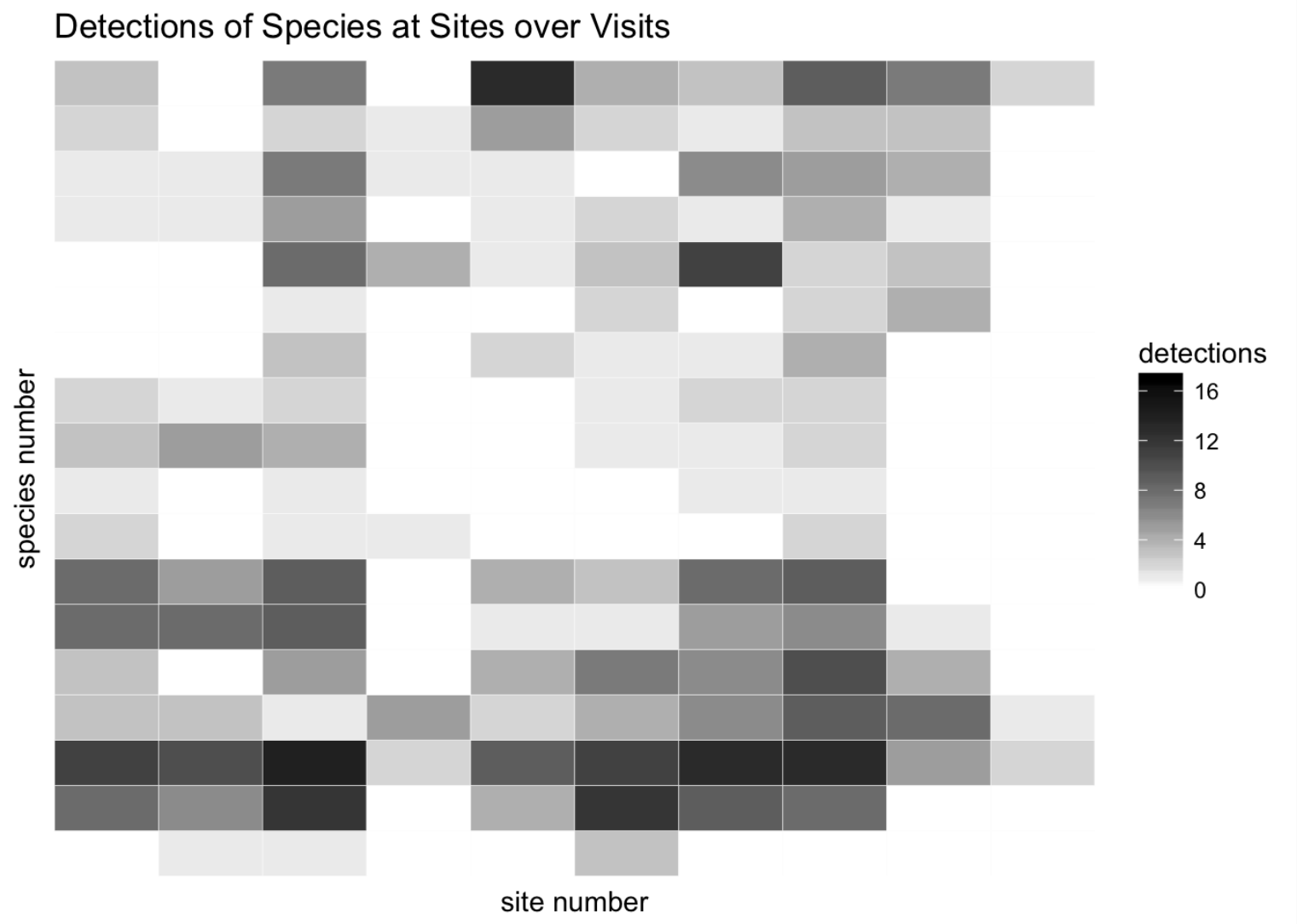
```
Bresil 3 Croatie 1
Mexique 1 Cameroun 0
Bresil 0 Mexique 0
Cameroun 0 Croatie 4
Cameroun 1 Bresil 4
Croatie 1 Mexique 3
Espagne 1 Pays-Bas 5
Chili 3 Australie 1
Espagne 0 Chili 2
Australie 2 Pays-Bas 3
Australie 0 Espagne 3
Pays-Bas 2 Chili 0
Colombie 3 Grece 0
Coted'Ivoire 2 Japon 1
Colombie 2 Coted'Ivoire 1
Japon 0 Grece 0
Japon 1 Colombie 4
Grece 2 Coted'Ivoire 1
```

soccerpowerindex.txt

```
Bresil
Argentine
Allemagne
Espagne
Chili
France
Colombie
Uruguay
Angleterre
Belgique
Pays-Bas
Bosnie
Equateur
Portugal
Coted'Ivoire
Russie
Italie
```

Interesting Case Study

2. Butterfly - Estiamte the species richness (absence or nondetection?)



Interesting Case Study

3. Global temperature - Can you estimate the trend reliably?

Terms of the Contest

The file Series1000.txt contains 1000 simulated time series. Each series has length 135: the same length as that of the most commonly studied series of global temperatures (which span 1880-2014). The 1000 series were generated as follows. First, 1000 random series were obtained (for more details, see below). Then, some of those series were randomly selected and had a trend added to them. Each added trend was either 1°C/century or −1°C/century. For comparison, a trend of 1°C/century is greater than the trend that is claimed for global temperatures.

A prize of $100 000 (one hundred thousand U.S. dollars) will be awarded to the first person who submits an entry that correctly identifies at least 900 series: which series were generated without a trend and which were generated with a trend.

Should you invest 10 dollars in this contest?

Our Next Steps

1. **Get familiar with software and basic workflow**
    - BDA book examples and problems
        - single, multiple parameter, hierarchical models
        - model checking and expansion
        - data collection Process
2. **Move on to more realistic and complicated case studies**
    - learn from strong guys like Andrew Gelman
    - complete code, thoughts, explanation, even discussion

Format of Our Discussion

1. **Meet every Friday at 4pm**
2. **1 or 2 examples will be released in advance**
    - mostly applied
3. **Discussion leader**
    - voluntary (first come, first served)
    - provide Rmarkdown documents can directly run (read data from shared folders)
    - throw questions when necessary
    - take questions not answered to the professors
4. **Software**
    - rstan. (Lots of case studies and active community)