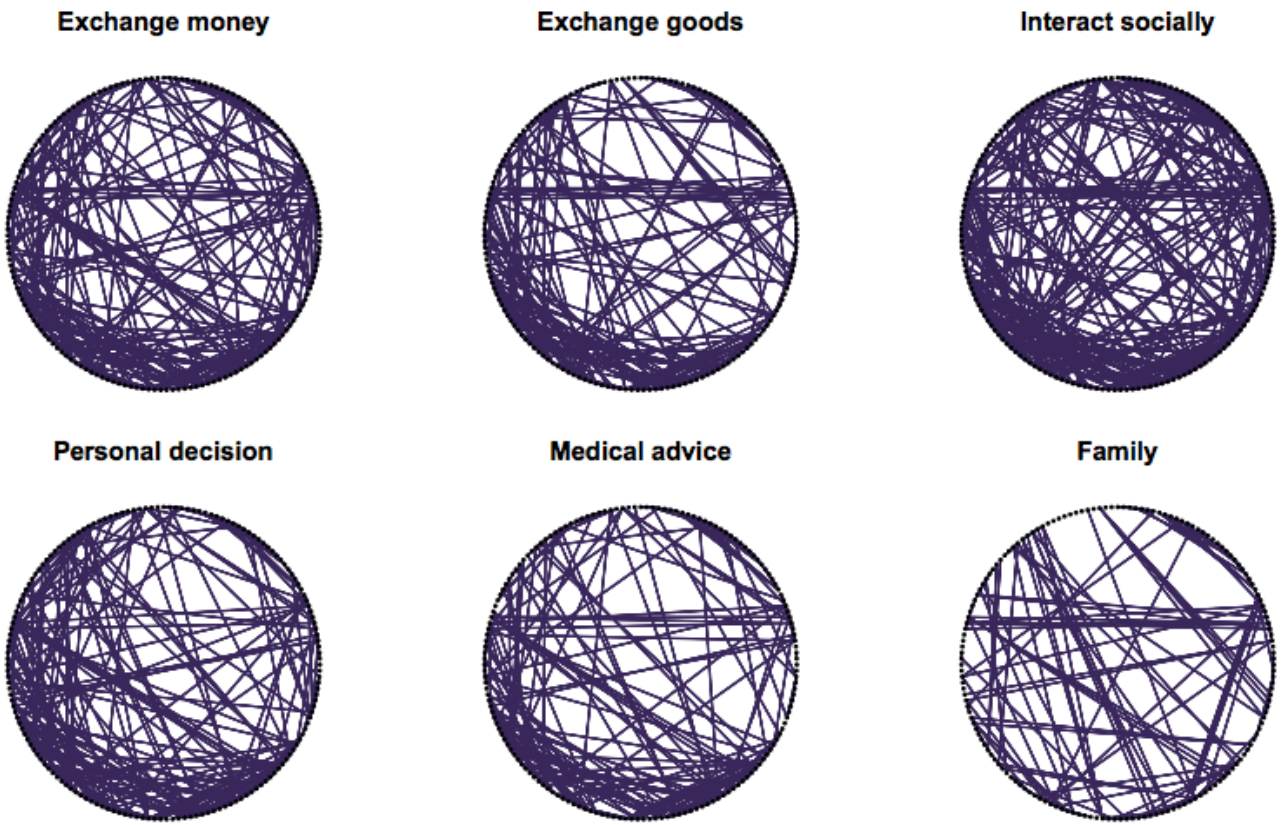
Module 0: Introduction to Bayesian Statistics and Course Expectations

Rebecca C. Steorts

Agenda

- Motivations
- Traditional inference
- Bayesian inference
- Course Expectations
- Homeworks
- Exams
- Questions?

Social networks



Precision Medicine

AI AND HEALTH

Editor: **Daniel B. Neill, H.J. Heinz III College,** Carnegie Mellon University, neill@cs.cmu.edu

A $3 Trillion Challenge to Computational Scientists: Transforming Healthcare Delivery

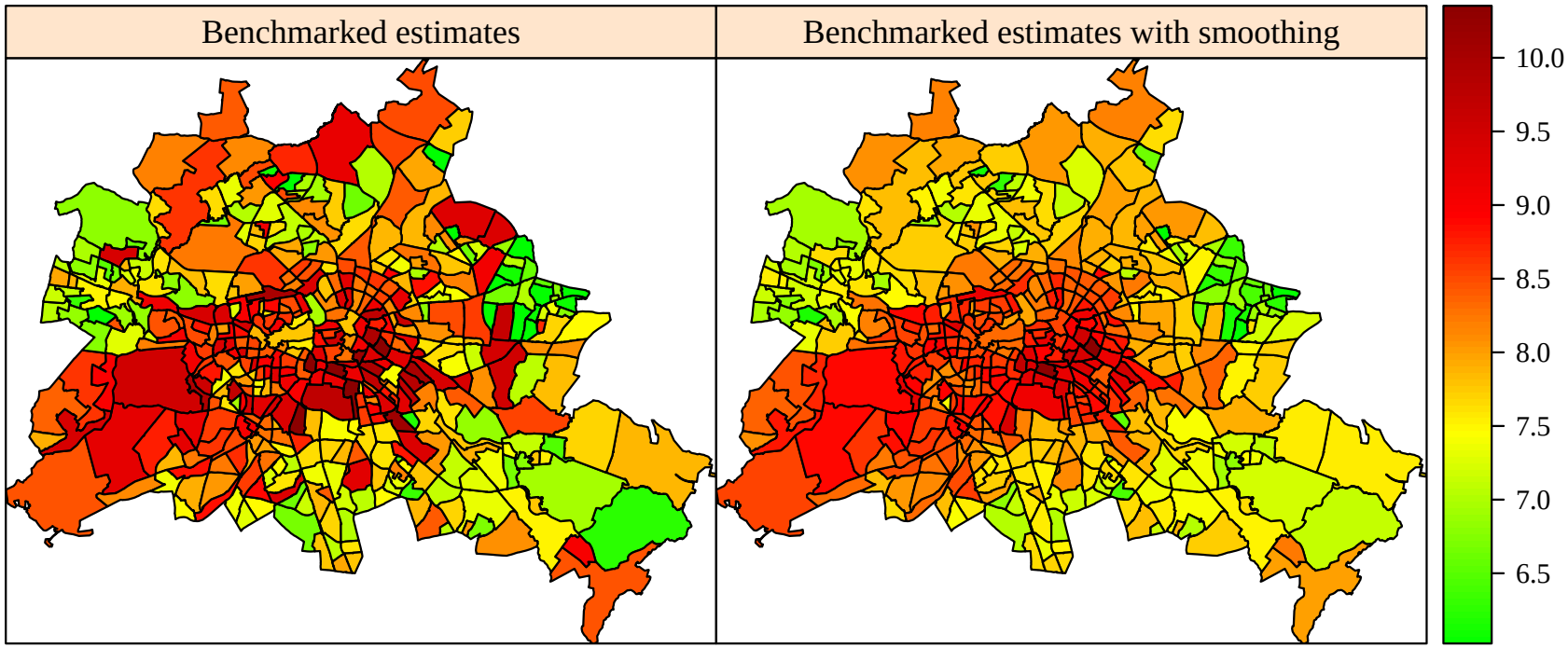
Suchi Saria, *Johns Hopkins University*

Healthcare spending in the US is nearing $3 trillion per year, but in spite of this expenditure, the US is outpaced by most developed countries in terms of health and quality of life outcomes—for example, it ranks 36th internationally in life expectancy.[1] The share of health spending in its gross domestic product has increased sharply, from 5 percent of GDP in 1960 to more than 17 percent today,[2] a rate of increase that's widely believed to be unsustainable.[3]

Policy and regulatory reform have important roles to play in addressing these challenges. Yet one of the largest underexplored avenues is the better use of information derived from the vast amount of health data now being collected in digital format.[4] I believe that one of the most significant open frontier records that weren't amenable to retrospective, automated analyses. The Health Information Technology for Economic and Clinical Health (HITECH) Act, a program that was part of the American Recovery and Reinvestment Act of 2009, incentivized the adoption of Electronic Health Records (EHRs) to encourage the shift from paper to digital records. That program has made more than $15.5 billion available to hospitals and healthcare professionals conditioned on their meeting certain EHR benchmarks for so-called "meaningful use." It's one of the largest investments in healthcare infrastructure ever made by the federal government.

A survey by the American Hospital Association showed that adoption of EHRs has doubled from 2009 to 2011. Today, much of an individual's health

Estimating Rent Prices in Small Domains



Traditional inference

You are given **data** X and there is an **unknown parameter** you wish to estimate θ

How would you estimate θ?

Traditional inference

You are given **data** X and there is an **unknown parameter** you wish to estimate θ

How would you estimate θ?

- In a very simple way, you could assume the data is normally distributed, and estimate the parameter by calculating the maximum likelihood estimator.
- You could alternatively calculate an unbiased estimator.
- In short, using methods from previously classes, there are many probabilistic methods that you could use.

Bayesian methods

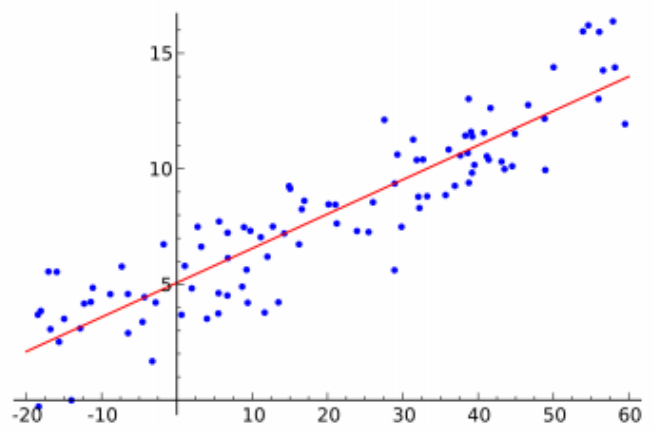
When using a Bayesian method, we assume that we have access to information about the parameter of interest θ.

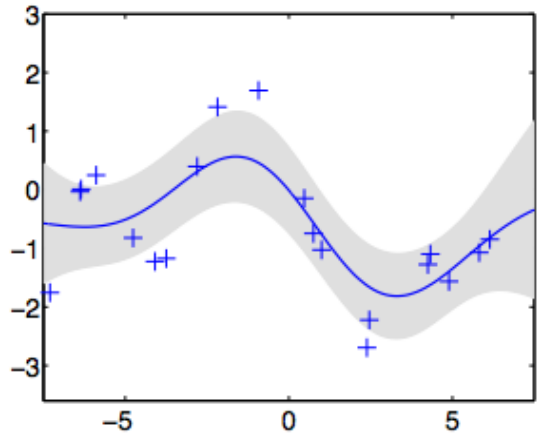
In this sense, the parameter θ is still unknown however we will treat it as a random variable and put a distribution on θ.

Bayesian Motivation

Parameters

$$P(X|\theta) \quad = \quad \text{Probability}[\text{data}|\text{pattern}]$$





Inference idea

$$\text{data} \quad = \quad \text{underlying pattern} \quad + \quad \text{independent noise}$$

[credit: Peter Orbanz, Columbia University]

Bayesian inference

Bayesian methods trace its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call **Bayes' Theorem**

- $p(x \mid \theta)$ likelihood
- $p(\theta)$ prior
- $p(\theta \mid x)$ posterior
- $p(x)$ marginal distribution

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

This course

This course starts with Bayes' Theorem and the basics of Bayesian statistics.

We will then move to many common topics that are used in Bayesian statistics, surveying most of the methods covered in the text by Peter Hoff.

Topics

- Introduction to Bayesian Statistics
- Decision Theory
- Hierarchical Models
- Monte Carlo
- Markov Chain Monte Carlo (MCMC)
- Gibbs Sampling
- Multivariate Bayesian Models
- Linear Regression
- Special Topics

The Lectures

- Lectures will be via slides, code, examples.
- They will be based on the book.
- Expect about 8–10 homeworks throughout the semester.
- All information will be posted on Sakai regarding deadlines.

The Homeworks

1. All code must be written to be reproducible in Markdown.
2. All derivations can be done in any format of your choosing (word, latex, markdown, written by hand) but must be converted to a pdf document. It must be legible.
3. All files must be zipped together and submitted to Sakai as one file. Otherwise, you cannot submit. (Try this early to avoid not submitting anything).
4. Ask questions early if you have a problem to a TA regarding submission issues.
5. Your lowest homework will be dropped.

The Exams

1. All exams will be in class, closed book.
2. They will be cumulative as they go, but they will focus on the most recent material we have covered.
3. Make up exams will not be given.
4. You must take the final exam to pass the class!

Typos

- If you find a typo on a slide, please write it down neatly with the slide number and give it to me after class.
- I do my best to fix typos within 48 hours after lecture and re-post them.
- Thanks for helping me spot typos in advance!

Coding using R, RStudio, and Markdown

- In this class, I will assume that you are very familiar with R.
- If you need to refresh certain R skills, please do this on your own.

Reproducible Code

In this course, all code you turn in will be expected to be reproducible.

What does this mean?

Reproducible Code

- Suppose I write a lecture on linear regression with many plots explaining my analysis.
- You might wonder how exactly I created my analysis.
- If I simply write my code in a way such that you can reproduce my entire lecture, then you can verify all the results that I show you.
- Similarly, if you write your code in this way for your homeworks, then I and the TAs can also verify the results very quickly.

Office Hours and Questions

- Office hours can be found on the course syllabus.

Syllabus

- I encourage you to sign up and use the Google group platform for questions as well. If you don't feel comfortable posting, email myself or a TA and we will post the question for you.

Google group

- When emailing, please CC all TA's and myself.

R versus RStudio

- RStudio mediates your interaction with R.
- RStudio is a driver of the emergergence of R Markdown, knitr, R + Github.

What is Markdown?

- Markdown is a lightweight markup language for creating PDF (or other documents).
- Markup languages produce documents from human readable text (and annotations).
- Some of you might be familiar with LaTeX (less friendly). This is another mark up language for creating PDF documents.

Why I like Markdown

1. It's very easy to learn.
2. The focus is on content rather than coding and debugging of errors.
3. One you have the basics, you can get fancy!

(In fact, my slides right now are in markdown, so you can even make slides)!

R Markdown

This just means that you include R code in your **markdown** document.

```
x<- 2 + 2
x
```

```
## [1] 4
```

Installing RStudio and Markdown

- How do I install Rstudio?
- How do I include markdown?

Use the command

```
install.packages("rmarkdown")
```

More behind R Markdown

R Markdown files are the source code for rich, reproducible documents. You can transform an R Markdown file in two ways.

1. knit: You can knit the file.

- The rmarkdown package will call the knitr package.
- knitr will run each chunk of R code in the document and append the results of the code to the document next to the code chunk.
- This workflow saves time and facilitates reproducible reports.

In the R Markdown paradigm, each report contains the code it needs to make its own graphs, tables, numbers, etc. The author can automatically update the report by re-knitting.

More behind R Markdown

2. convert: You can convert the file.

- The rmarkdown package will use the pandoc program to transform the file into a new format. - For example, you can convert your .Rmd file into an HTML, PDF, or Microsoft Word file.
- You can even turn the file into an HTML5 or PDF slideshow.
- rmarkdown will preserve the text, code results, and formatting contained in your original .Rmd file.

Conversion lets you do your original work in markdown, which is very easy to use. You can include R code to knit, and you can share your document in a variety of formats.

More behind R Markdown

In practice, authors almost always knit and convert their documents at the same time. In this article, I will use the term render to refer to the two step process of knitting and converting an R Markdown file.

You can manually render an R Markdown file with

`rmarkdown::render()`.

Takeaways

1. Make sure that you understand the class policies.
2. Make sure you can access the Google group.
3. Make sure that you can access Sakai, the first homework assignment, and start your assigned readings for the course.
4. If you feel unprepared for the class, please speak with me now, rather than later.
5. Install RStudio and markdown. Use the lab with this lecture to make sure that you can do basic exercises in R and that they are reproducible. (If you have trouble with the lab or first few homework assignments and exam, then please see me quickly to resolve issues.)