Bayesian Theory and Computation

Lecture 1: Introduction

Cheng Zhang

School of Mathematical Sciences, Peking University

March 09, 2021

General Information

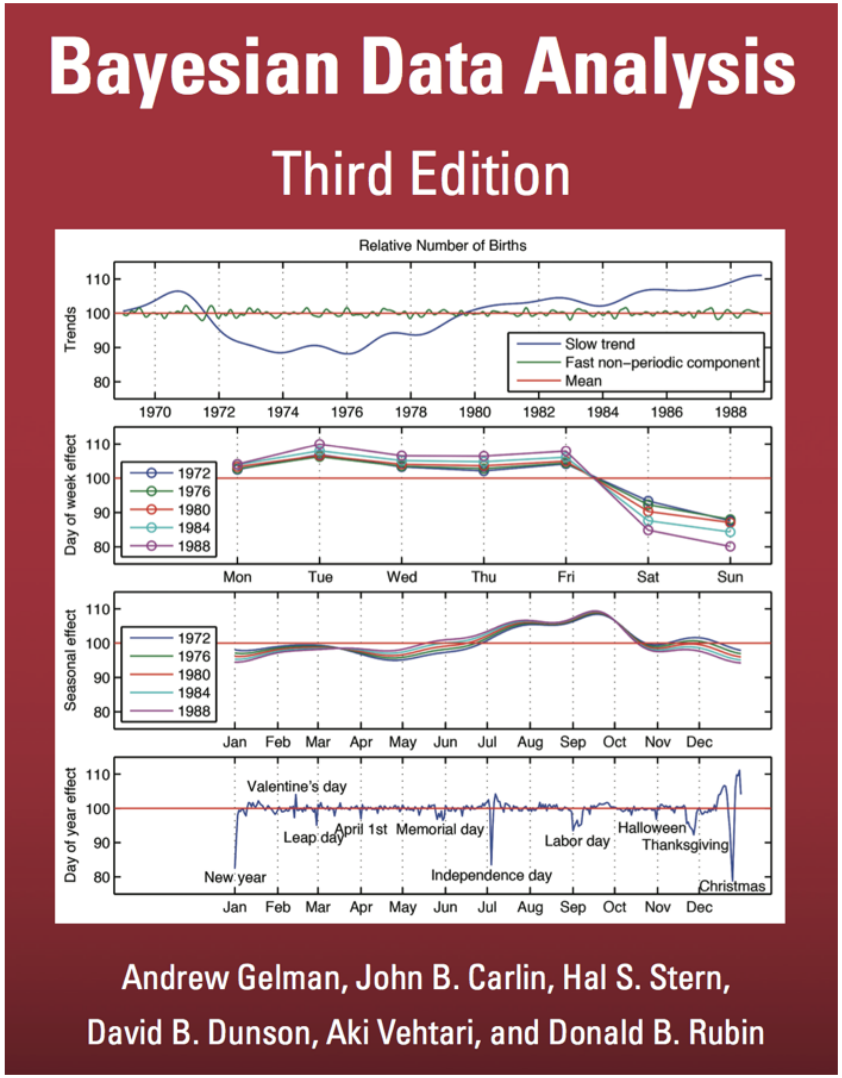
- Class times:
  - Odd Tuesday 1:00-2:50pm, Friday 3:10-5:00pm
  - Science Classroom Building, Room 407
- Instructor:
  - Cheng Zhang: `chengzhang@math.pku.edu.cn`
- Teaching assistants:
  - Weijian Luo: `2001110057@stu.pku.edu.cn`
- Tentative office hours:
  - 1279 Science Building No.1
  - Thursday 3:00-5:00pm or by appointment
- Website:

  `https://zcrabbit.github.io/courses/btc-s21.html`

- A branch of statistical sciences focusing on Bayesian approaches, an alternative to frequentist approaches.
- The focus lies on modern Bayesian statistical methods and theory, and various statistical models with Bayesian formulation.
- With the rise of modern computational power, Bayesian approaches have been developing rapidly due to the ease of handling complicated models and the ability of providing uncertainty quantification.

Goals

- Learn how to formulate a scientific question by constructing a Bayesian model and perform Bayesian statistical inference to answer that question.
- Develop a deeper understanding of the mathematical theory of Bayesian statistical methods and modeling.
- Learn several computational techniques, and use them for Bayesian analysis of real data using a modern programming language (e.g., python).



Download it here:
`http://www.stat.columbia.edu/~gelman/book/`

Other interesting references:

- Liu, J. (2001). Monte Carlo Strategies in Scientific Computing, Springer-Verlag.
- Lange, K. (2002). Numerical Analysis for Statisticians, Springer-Verlag, 2nd Edition.
- Christian, P. R. (2004). The Bayesian Choice, Springer.

Tentative Topics

- Approximate Bayesian Inference Methods
  - Sequential Monte Carlo
  - Markov chain Monte Carlo
  - Variational Inference
  - Scalable Approaches
- Bayesian Theory
  - Decision Theory
  - Convergence Analysis of Bayesian Inference Methods
- Bayesian Models
  - Regression and Classification Models
  - Hierarchical Models
  - Non-parametric Models

Familiar with at least one programming language (with python preferred!).

- All class assignments will be in python (and use numpy).
- You can find a good Python tutorial at

  `http://www.scipy-lectures.org/`

  You may find a shorter python+numpy tutorial useful at

  `http://cs231n.github.io/python-numpy-tutorial/`

Familiar with the following subjects

- Probability and Statistical Inference
- Stochastic Processes

Grading Policy

- 4 Problem Sets: $4 \times 15\% = 60\%$
- Final Course Project: 40%
  - up to 4 people for each team
  - Teams should be formed by the end of week 4
  - Midterm proposal: 5%
  - Oral presentation: 10%
  - Final write-up: 25%
- Late policy
  - 7 free late days, use them in your ways
  - Afterward, 25% off per late day
  - Not accepted after 3 late days per PS
  - Does not apply to Final Course Project
- Collaboration policy
  - Finish your work independently, verbal discussion allowed

Final Project

- Structure your project exploration around a general problem type, theory, algorithm, or data set, but should explore around your problem, testing thoroughly or comparing to alternatives.
- Present a project proposal that briefly describe your teams' project concept and goals in one slide in class at midterm.
- There will be in class project presentation at the end of the term. Not presenting your projects will be taken as voluntarily giving up the opportunity for the final write-ups.
- Turn in a write-up (< 10 pages) describing your project and its outcomes, similar to a research-level publication.

Today's Agenda

- Why Bayesian?
- Basic concepts in Bayesian Statistics

- Statistical methods are mainly inspired by applied scientific problems.
- The overall goal of statistical analysis is to provide a robust framework for designing scientific studies, collecting empirical evidence, and analyzing the data, in order to understand unknown phenomena, answer scientific questions, and make decisions.
- To this end, we rely on the *observed data* as well as our *domain knowledge*.

Domain Knowledge

- Our domain knowledge, which we refer to as prior information, is mainly based on previous empirical evidence.
- For example, if we are interested in the average normal body temperature, we would of course measure body temperature of a sample of subjects from the population, but we also know, based on previous data, that this average is a number close to 37°C.
- In this case, our prior knowledge asserts that values around 37 are more plausible compared to values around 34 or 40.

Objective vs. Subjective

- We could of course attempt to minimize our reliance on prior information (e.g., use weakly informative prior).
- Most frequentist methods follow this principle and use the domain knowledge to decide which characteristics of the population are relevant to our scientific problem (e.g., we might not include height as a risk factor for cancer), but avoid using priors when making inference.
- Note that this should not give us the illusion that the frequentist methods are entirely objective.

- Bayesian methods on the other hand provide a principled framework that to incorporate prior knowledge in the process of making inference.
- **Bayes' Theorem** (Thomas Bayes)

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

  - $p(\mathcal{D}|\theta)$ is the model probability function, also known as the **likelihood** function when viewed as a funciton of θ.
  - $p(\theta)$ is the prior
  - $p(\mathcal{D})$ is the normalizing constant, also known as the model evidence.
- This can be viewed as an inverse probability formula.

- If the prior is in fact informative, this should lead to more accurate inference and better decisions. Also, the way we incorporate our prior knowledge in the analysis is explicit (e.g., $p(\theta)$)
- The counterargument is that this makes our analysis more prone to mistakes.
- While the underlying concept for Bayesian statistics is quite simple, implementing Bayesian methods might be difficult compared to their frequentist counterparts, due to integration of many parameters.

- Recall that we define the underlying mechanism that generates data, $\mathcal{D}$, using a probability model $p(\mathcal{D}|\theta)$, which depends on the unknown parameter of interest, θ.
- Frequentist method typically use this probability for inference
- To estimate the model parameter, we can find θ that maximizes the probability of the observed data.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(\mathcal{D}|\theta) = \arg\max_{\theta} \log p(\mathcal{D}|\theta)$$

Often, the log-likelihood function is denoted as $L(\theta)$, and

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta)$$

This is known as the Maximum likelihood estimate (MLE).

- The gradient of L with respect to θ is called the **score**

$$s(\theta) = \frac{\partial L}{\partial \theta}$$

  The expected value of the score is zero: $\mathbb{E}(s) = 0$.
- The variance of the score is known as **Fisher information**

$$\mathcal{I}(\theta) = \mathbb{E}(ss^T)$$

  Under mild assumptions (e.g., exponential families),

$$\mathcal{I}(\theta) = -\mathbb{E}\left(\frac{\partial^2 L}{\partial \theta \partial \theta^T}\right)$$

  Fisher information is a measure of the expected curvature of the Log-likelihood function.

- **Consistency**. Under weak regularity condition, $\hat{\theta}_{MLE}$ is consistent: $\hat{\theta}_{MLE} \to \theta_0$ in probability as $n \to \infty$, where θ_0 is the "true" parameter
- **Asymptotical Normality**.

$$\hat{\theta}_{MLE} - \theta_0 \to \mathcal{N}(0, \mathcal{I}^{-1}(\theta_0))$$

  See Rao 1973 for more details.
- **Efficiency**. $\mathcal{I}^{-1}(\theta_0)$ is the minimum variance that can be achieved by any unbiased estimator, which is known as **Cramér-Rao Lower Bound**.

- For any unbiased estimator $\hat{\theta}$ of θ_0 based on independent observations following the true distribution, the variance of the estimator is bounded by the reciprocal of the Fisher information

$$\mathbb{V}\mathrm{ar}(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta_0)}$$

- Sketch of proof: Consider a general estimator $T = t(X)$ with $\mathbb{E}(T) = \psi(\theta_0)$. Let s be the score function,

$$\mathbb{C}\mathrm{ov}(T, s) = \mathbb{E}(Ts) = \psi'(\theta_0)$$

Therefore,

$$\mathbb{V}\mathrm{ar}(T) \geq \frac{[\psi'(\theta_0)]^2}{\mathbb{V}\mathrm{ar}(s)} = \frac{[\psi'(\theta_0)]^2}{\mathcal{I}(\theta_0)}$$

Example: Poisson Distribution

$$L(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} y_i \log \theta - n\theta - \sum_{i=1}^{n} \log y_i!$$

$$s(\theta) = \frac{\sum_{i=1}^{n} y_i}{\theta} - n, \quad \mathcal{I}(\theta) = \frac{n}{\theta}$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} y_i \log \theta - n\theta = \frac{\sum_{i=1}^{n} y_i}{n}$$

By the **Law of large numbers**

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$$

By **central limit theorem**

$$\hat{\theta}_{MLE} - \theta_0 \xrightarrow{d} \mathcal{N}\left(0, \frac{\theta_0}{n}\right)$$

- We can also use the likelihood funciton to device standard tests (Wald test, score test, and likelihood ratio test) to perform hypothesis testing.
- **Strong Likelihood Principle**: The relevant information in any inference about θ after $\mathcal{D}$ is observed is contained entirely in the likelihood function.
- In other words, if the corresponding likelihood function for two observed samples x, y are proportional,

$$f_1(\theta, x) \propto f_2(\theta, y)$$

then inference for θ should be the same whether we observe x or y.

- The following example is from David MacKay's book.
- A scientist has just received a grant to examine whether a specific coin is fair (i.e., $p(H) = p(T) = 0.5$) or not.
- He sets up a lab and starts tossing the coin. Of course, because of his limited budget, he can only toss the coin a finite number of times. Suppose he tosses the coin 12 times, of which only 3 are heads.
- He hires a frequentist statistician and ask him to estimate the p-value hoping that the result could be published in one of the journals that only publish if the p-value is less than 0.05.
- The statistician says: "you tossed the coin 12 times and you got 3 heads. The one-sided p-value is 0.07".

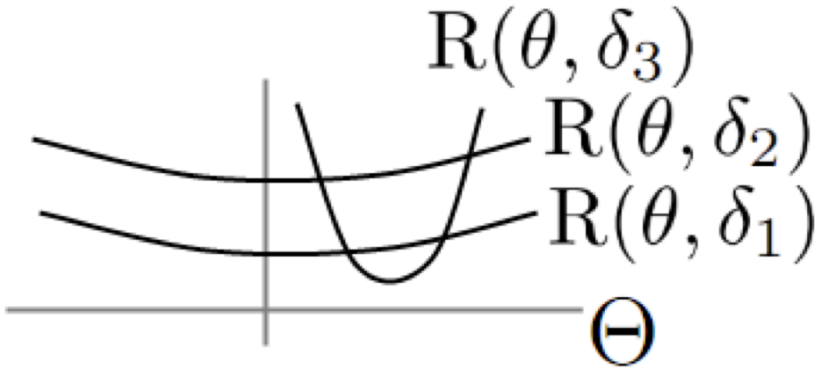
Violation of The Strong Likelihood Principal

- The scientist says: "Well, it wasn't exactly like that... I actually repeat the coin tossing experiment until I got 3 heads and then I stopped".
- The statistician say: "In that case, your p-value is 0.03".
- Note that in the first scenario, we use a binomial model, and in the second scenario, we use a negative-binomial model with the following likelihood functions respectively

$$f_1(\theta, x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad f_2(\theta, x) = \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x}$$

  which are proportional.
- We'll see the answer by a Bayesian statistician later.

- In statistical decision theory, we need more than just probability: we need a measure of loss or gain for each possible outcome, i.e., a *loss function*
- A loss function $\ell(\theta, \delta(X))$ is a function that assigns to each possible outcome of a decision $\delta(X)$ a number that represents the cost and the amount of regret (e.g., loss of profit) we endure when that outcome occurs.
- The loss function determines the penalty for predicting $\delta(X)$ if θ is the true parameter. E.g., 0-1 loss in the discrete case, or the square loss $\ell(\theta, \delta(X)) = \|\theta - \delta(X)\|^2$.
- Note that in general, $\delta(X)$ does not necessarily have to be an estimate of θ.
- To make decisions, we need to calculate which procedure is the best even though we cannot observe the true nature of the parameter space and data.



- The frequentist risk is

$$R(\theta, \delta) = \mathbb{E}_\theta(\ell(\theta, \delta(X)))$$

  where θ is held fixed and the expectation is taken over $\mathcal{X}$.
- Often, one decision does not dominate the other everywhere (e.g., δ_1, δ_2).
- The challenge is how should we decide which one is better when they overlap, e.g., δ_1, δ_3?

- A decision δ is *inadmissible* if it is dominated everywhere, i.e., there exist δ' such that

$$R(\theta, \delta') \leq R(\theta, \delta), \quad \forall \theta$$

  E.g., δ_2 compared to δ_1 in the previous figure.

- Variance-bias Tradeoff

$$\mathbb{E}\|\theta - \delta(X)\|^2 = \|\theta - E\delta(X)\|^2 + \mathbb{E}\|\delta(X) - \mathbb{E}\delta(X)\|^2$$

- Unbiased procedures are not always good. In fact, some unbiased procedures can be inadmissible.

- Consider the following model

$$x_i|\mu_i \sim \mathcal{N}(\mu_i, 1), \quad i = 1, \ldots, N$$

  where $x_i,\ i = 1, \ldots, N$ are independent.
- A reasonable estimate of μ is $\hat{\mu}_{\text{MLE}} = x$, i.e., the maximum likelihood estimator.
- The MLE is unbiased, $\mathbb{E}x = \mu$. Moreover, it also achieves the Cramér-Rao Lower Bound, meaning that it is the best unbiased estimator with the minimum variance.
- The expected squared error (i.e., frequentist risk when ℓ is the squared error) is

$$\mathbb{E}\|\mu - x\|^2 = N$$

- While the above estimator is commonly used (e.g., ANOVA, regression), the statistics community was shocked when Stein and James showed that the following estimator, known as James-Stein estimator, dominates MLE for $N > 2$

$$\hat{\mu}_{\text{JS}} = \left(1 - \frac{N-2}{\|x\|^2}\right) x$$

- **Theorem** (James and Stein, 1961). For $N \geq 3$, the James-Stein estimator everywhere dominates the MLE $\hat{\mu}_{\text{MLE}}$ in terms of the expected total squared error,

$$\mathbb{E}_\mu \|\hat{\mu}_{\text{JS}} - \mu\|^2 < \mathbb{E}_\mu \|\hat{\mu}_{\text{MLE}} - \mu\|^2$$

for every choice of μ.

- Suppose that μ follows some prior distribution

$$\mu_i \sim \mathcal{N}(0, A), \quad i = 1, \ldots, N$$

- The posterior is also Gaussian

$$\mu|x \sim \mathcal{N}(Bx, BI), \quad B = \frac{A}{A+1}$$

- The Bayes estimator is

$$\hat{\mu}_{\text{Bayes}} = \left(1 - \frac{1}{A+1}\right) x$$

- Unfortunately, we do not know A, how can we deal with it?

- Now that we observe x, we can somehow estimate A from x. In fact, the marginal distribution of x now is

$$x \sim \mathcal{N}(0, (A+1)I)$$

- Therefore, $\|x\|^2$ has a scaled chi-square distribution with N degrees of freedom

$$\|x\|^2 \sim (A+1)\chi^2_N$$

- This gives the following unbiased estimate of $1/(A+1)$

$$\mathbb{E}\left(\frac{N-2}{\|x\|^2}\right) = \frac{1}{A+1}$$

- Plug it back into $\hat{\mu}_{\text{Bayes}}$ gives $\hat{\mu}_{\text{JS}}$.

- We will show later the Bayes estimator have even lower "risk".
- Such shrinkage estimators are the main inspiration behind the field of empirical Bayes, which was created to help frequentist methods to achieve full Bayesian efficiency in large scale studies.
- For more details, see Efron's book on Large-Scale Inference.
- We will focus on more formal and principled Bayesian framework, which provides similar benefits through the shrinkage of parameter estimates.

- Bayesian inference starts by defining the joint probability for our prior opinion and the mechanism based on which the data are generated.
- To make inference, we refer to this updated opinion as our posterior opinion, which itself is expressed in terms of probabilities.
- Probability has a central role in Bayesian statistics, and provides a coherent and axiomatic framework for deriving Bayesian methods and making statistical inference.

- In the Bayesian paradigm, probability is a measure of uncertainty.
- A Bayesian statistician would use probability models for random variables that change and those that might not change (e.g., the population mean) but we are uncertain about their value.
- Consider the well-known coin tossing example. What is the probability of head in one toss?
- There are only two possibility for the outcome: head and tail. Assuming symmetry (i.e., a fair coin), head and tail equal probability 1/2.

- In the frequestist view, probability is assigned to an event by regarding it as a class of individual events (i.e., trials) all equally probable and stochastically independent.
- For the coin tossing example, we assume a sequence of iid tosses, and the probability of head is 1/2 since the number of times we observe head divided by the number of trials reaches 1/2 as the number of trials grows.
- Note that while Bayesians and frequentists provide the same answer, there is a fundamental and philosophical difference in how they view probability.
- Bayesian feel comfortable to assign probabilities to events that are not repeatable.
- For example, I can show you a picture of a car and ask "what is the probability that the price of this car is less than $5000?"

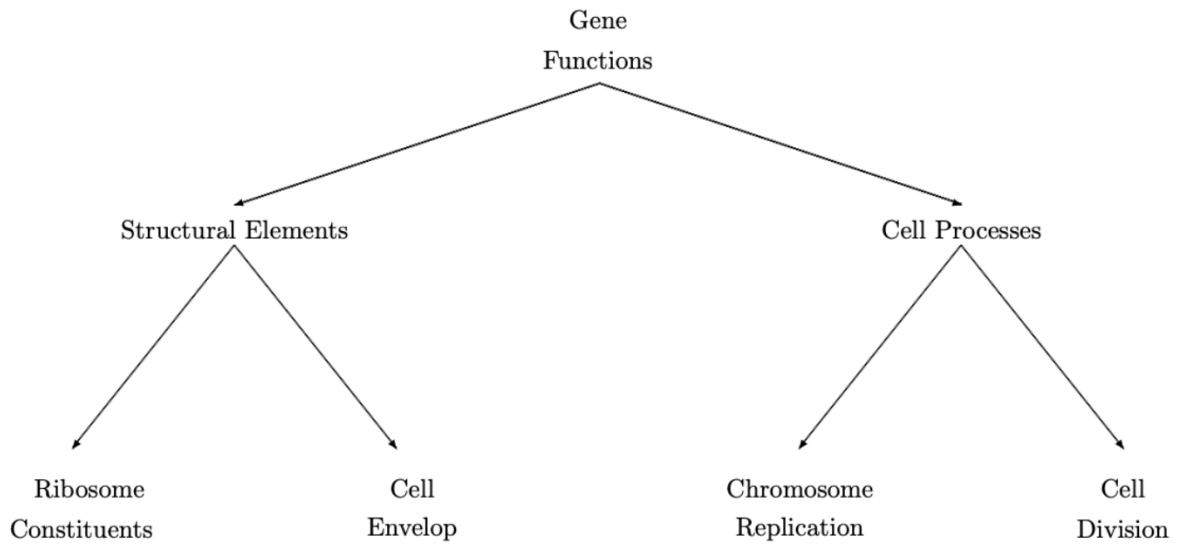
- As mentioned above, within the Bayesian framework, we use probability not only for the data, but also for the model parameters, since it is the population parameter and is almost always unknown.
- We usually use our (or others') domain knowledge, which is accumulated based on previous scientific studies.
- We almost always have such information, although it could be vague.

- For example, consider the study conducted by Mackowiak, et al. to find whether the average normal body temperature is the widely accepted value of 37°C.
- Let's denote the average normal body temperature for the population as θ. We know that θ should be close to 37°C; that is, values close to 37°C are more plausible than values close to 34°C for example.
- We assume that as we move away from 37°C the values become less likely in a symmetric way (i.e., it does not matter if we go higher or lower).

- Base on the above assumptions (and ignoring the fact that body temperature cannot be negative), we can set $\theta \sim \mathcal{N}(37, \tau^2)$.
- In the above prior, τ^2 determines how certain we are about the averge normal body temperature being around 37°C.
- If we believe that it is almost impossible that the average normal body temperature is above 22 and below 52, we can set $\tau = 5$ so the approximate 99.7% interval includes all the plausible values from 22 to 52.
- A general advise is that we should keep an open mind, consider all possibilities, and avoid using very restrictive priors.

- Sometimes, our prior opinion is based on what we know about the underlying structure of the data.
- For example, in many classification problems, we have prior knowledge about how classes can be arranged in a hierarchy.
- Hierarchical classification problems of this sort are abundant in statistics and machine learning.
- On such example is prediction of genes biological functions.

- As shown in the following figure, gene functions usually are presented in a hierarchical form, starting with very general classes (e.g., cell processes) and becoming more specific in lower levels of the hierarchy (e.g., cell division)



Adapted from Riley 1993

- In the Bayesian framework, we can incorporate such information in our model.

Course Outline

- So far, we have tried to establish why we use Bayesian analysis and introduced some basic concept of it.
- Throughout this course, we will discuss different Bayesian methods and theory, modern Bayesian models and their applications for analyzing scientific problems.
- We first start with simple and classic models, then move to more complicated models (e.g., hierarchical models), followed by advanced computational methods.
- Finally, we will discuss Bayesian nonparameteric, with an emphasis on Gaussian process models and Dirichlet process models.

References

- D. MacKay. Information Theory, Inference, and Learning Algorithms, Cambridge University Press. 2003
- Bradley Efron. Large-Scale Inference. Cambriage University Press. 2010.