universität wien

Statistics for Data Science, Winter 2023

Chapter 2:

Montecarlo and Bayesian Methods

Goals

Analyze statistical properties of inference methods (e.g., precision of estimates, coverage of CIs, power of tests) ...

- mathematically
- numerically

Use numerical/simulation methods for data analysis, e.g, ...

- compute quantiles of analytically intractable distributions
- MCMC for Bayesian data analysis
- Bootstrap and resampling

Overview

- Montecarlo methods
  - Numerical integration
  - Random number generation
- Introduction to Bayesian analysis
- Refined MC methods

Montecarlo Methods

Montecarlo Integration

Suppose we want to (approximately) compute the mean of a real random variable X with pdf or pmf f.

If we can generate a sample $X_1, \ldots, X_N \overset{iid}{\sim} f$, then by the LLN,

$$\frac{1}{N}\sum_{j=1}^{N} X_j \quad \xrightarrow[N\to\infty]{i.p.} \quad \mathbb{E}[X].$$

If X takes values in $\mathcal{X}$ and $S : \mathcal{X} \to \mathbb{R}$, then $S(X_1), \ldots, S(X_N)$ are also iid and

$$\frac{1}{N}\sum_{j=1}^{N} S(X_j) \quad \xrightarrow[B\to\infty]{i.p.} \quad \mathbb{E}[S(X)] = \mathbb{E}_f[S],$$

provided that the expectation $\mathbb{E}[S(X)]$ is finite.

Idea: For large $\cancel{B}$ N, generate (pseudo) random numbers $X_1, \ldots, X_N \overset{iid}{\sim} f$ and compute $\frac{1}{N}\sum_{j=1}^{N} S(X_j)$ to approximate $\mathbb{E}[S(X)]$.

To compute the probability $P(X \in A)$ we can use

$$\frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_A(X_j) \xrightarrow[N\to\infty]{i.p.} \mathbb{E}[\underbrace{\mathbb{1}_A(X)}_{= \begin{cases} 1, & X \in A \\ 0, & \text{else} \end{cases}}] = P(X \in A).$$

How to approximately evaluate the variance of a real $X \sim f$?

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

$$\uparrow \qquad\qquad \uparrow$$

$$\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)^2$$

The trick also works for quantiles of f, provided that they are unique.

- Suppose the corresponding cdf $F(x) := \int_{-\infty}^{x} f(y)dy$ is invertible.
- $X_1, \dots, X_N \overset{iid}{\sim} f$
- $\hat{F}_N^{\dagger}(\alpha) \xrightarrow[N\to\infty]{i.p.} F^{-1}(\alpha). = q_\alpha$ HW

R Example:
Simulate the quantile function of the χ^2-distribution.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 NOVEMBER 5, 2020 VOL. 383 NO. 19

Remdesivir for the Treatment of Covid-19 — Final Report

J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R.W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T.F. Patterson, R. Paredes, D.A. Sweeney, W.R. Short, G. Touloumi, D.C. Lye, N. Ohmagari, M. Oh, G.M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M.G. Kortepeter, R.L. Atmar, C.B. Creech, J. Lundgren, A.G. Babiker, S. Pett, J.D. Neaton, T.H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H.C. Lane, for the ACTT-1 Study Group Members*

ABSTRACT

BACKGROUND

Although several therapeutic agents have been evaluated for the treatment of coronavirus disease 2019 (Covid-19), no antiviral agents have yet been shown to be efficacious.

METHODS

We conducted a double-blind, randomized, placebo-controlled trial of intravenous remdesivir in adults who were hospitalized with Covid-19 and had evidence of lower respiratory tract infection. Patients were randomly assigned to receive either remdesivir (200 mg loading dose on day 1, followed by 100 mg daily for up to 9 additional days) or placebo for up to 10 days. The primary outcome was the time to recovery, defined by either discharge from the hospital or hospitalization for infection-control purposes only.

RESULTS

A total of 1062 patients underwent randomization (with 541 assigned to remdesivir and 521 to placebo). Those who received remdesivir had a median recovery time of 10 days (95% confidence interval [CI], 9 to 11), as compared with 15 days (95% CI, 13 to 18) among those who received placebo (rate ratio for recovery, 1.29; 95% CI, 1.12 to 1.49; P<0.001, by a log-rank test). In an analysis that used a proportional-odds model with an eight-category ordinal scale, the patients who received remdesivir were found to be more likely than those who received placebo to have clinical improvement at day 15 (odds ratio, 1.5; 95% CI, 1.2 to 1.9, after adjustment for actual disease severity). The Kaplan–Meier estimates of mortality were 6.7% with remdesivir and 11.9% with placebo by day 15 and 11.4% with remdesivir and 15.2% with placebo by day 29 (hazard ratio, 0.73; 95% CI, 0.52 to 1.03). Serious

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Beigel at the National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Ln., Rm. 7E60, MSC 9826, Rockville, MD 20892-9826, or at jbeigel@niaid.nih.gov.

*A complete list of members of the ACTT-1 Study Group is provided in the Supplementary Appendix, available at NEJM.org.

A preliminary version of this article was published on May 22, 2020, at NEJM.org. This article was published on October 8, 2020, and updated on October 9, 2020, at NEJM.org.

N Engl J Med 2020;383:1813-26.
DOI: 10.1056/NEJMoa2007764

Consider the model:

- Treatment group: $X_1, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_t, \sigma_t^2)$, $\mu_t \geq 0$
- Control group: $Y_1, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_c, \sigma_c^2)$, $\mu_c \geq 0$
- Treatment and control groups are independent

X_i, Y_j are observed times to recovery.

$H_0 : 0 \leq \mu_c \leq \mu_t, \sigma_t^2 > 0, \sigma_c^2 > 0$ vs.
$H_1 : \mu_c > \mu_t \geq 0, \sigma_t^2 > 0, \sigma_c^2 > 0$

Example: Behrens-Fisher Problem

- Treatment group: $X_1, \dots, X_{n_1} \overset{iid}{\sim} N(\mu_t, \sigma_t^2)$, $\mu_t \geq 0$
- Control group: $Y_1, \dots, Y_{n_2} \overset{iid}{\sim} N(\mu_c, \sigma_c^2)$, $\mu_c \geq 0$

$H_0 : 0 \leq \mu_c \leq \mu_t, \sigma_t^2 > 0, \sigma_c^2 > 0$ vs.
$H_1 : \mu_c > \mu_t \geq 0, \sigma_t^2 > 0, \sigma_c^2 > 0$

$$\bar{Y}_{n_2} - \bar{X}_{n_1} \sim N\left(\mu_c - \mu_t, \frac{\sigma_t^2}{n_1} + \frac{\sigma_c^2}{n_2}\right)$$

Test statistic:

$$\cancel{Z = \frac{\bar{Y}_{n_2} - \bar{X}_{n_1}}{\sqrt{\frac{\sigma_t^2}{n_1} + \frac{\sigma_c^2}{n_2}}}} \quad ?! \qquad S = \frac{\bar{Y}_{n_2} - \bar{X}_{n_1}}{\sqrt{\frac{\hat{\sigma}_t^2}{n_1} + \frac{\hat{\sigma}_c^2}{n_2}}}.$$

$$\hat{\sigma}_t^2 = \frac{1}{n_1 - 1} \underbrace{\sum_{i=1}^{n_1} \left(X_i - \bar{X}_{n_1}\right)^2}_{\sim \sigma_t^2 \chi^2_{n_1-1}}, \quad \hat{\sigma}_c^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left(Y_j - \bar{Y}_{n_2}\right)^2$$

universität wien

- Treatment group: $X_1, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_t, \sigma_t^2)$, $\mu_t \geq 0$
- Control group: $Y_1, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_c, \sigma_c^2)$, $\mu_c \geq 0$

$H_0 : 0 \leq \mu_c \leq \mu_t, \sigma_t^2 > 0, \sigma_c^2 > 0$ vs.
$H_1 : \mu_c > \mu_t \geq 0, \sigma_t^2 > 0, \sigma_c^2 > 0$

$\sigma_t = \sigma_c = 1$

$S \sim N(\Delta_\mu, 1)$

$\hat{\sigma}_t = \hat{\sigma}_c = 1$

Test statistic:

$$S = \frac{\bar{Y}_{n_2} - \bar{X}_{n_1}}{\sqrt{\frac{\hat{\sigma}_t^2}{n_1} + \frac{\hat{\sigma}_c^2}{n_2}}}.$$

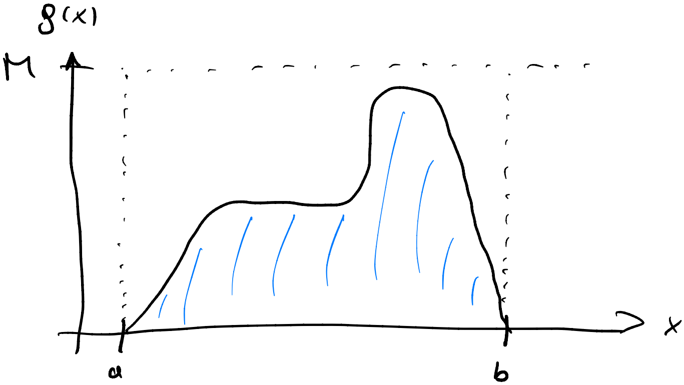
No closed form pdf of S exists! Depends on $\mu_t, \mu_c, \sigma_t^2, \sigma_c^2$!

Reject H_0 if $S \geq q_{1-\alpha}$

Simulate $1 - \alpha$ quantile! But for which $\mu_t, \mu_c, \sigma_t^2, \sigma_c^2$?

Note: $\mathbb{E}[S(X)] = \int S(x)f(x)dx$.
How about general (definite) integrals?

$$\int_a^b g(x)dx, \qquad 0 \leq g(x) \leq M$$



How about general (definite) integrals?

$$I := \int_a^b g(x)dx, \qquad 0 \leq g(x) \leq M$$

- Generate $(X_i, Y_i) \overset{iid}{\sim} Unif([a,b] \times [0,M]), i = 1, \ldots N.$
- Compute $\hat{I}_N := \frac{M(b-a)}{N} \sum_{i=1}^N \mathbb{1}_{\{Y_i \leq g(X_i)\}} \xrightarrow[N\to\infty]{i.p.} I$

$$\mathbb{E}\left[\mathbb{1}\{Y \leq g(X)\}\right] = P(Y \leq g(X))$$

$$= \int_0^b \int_0^{g(x)\wedge M} \frac{1}{M(b-a)} \, dy \, dx$$

$$= \frac{1}{M(b-a)} \int_0^b \int_0^{g(x)} 1 \, dy \, dx = \frac{1}{M(b-a)} \int_0^b g(x) \, dx$$

We want to approximately evaluate $\mathbb{E}_f[S]$. $= \int S(x) f(x)\, dx$

Let $g : \mathcal{X} \to [0, \infty)$ be a pdf or pmf such that

$$g(x) = 0 \quad \Rightarrow \quad S(x)f(x) = 0.$$

Define

$$T(x) := \begin{cases} S(x)\frac{f(x)}{g(x)}, & \text{if } g(x) > 0, \\ 0, & \text{if } g(x) = 0. \end{cases}$$

- Generate $Y_1, \ldots, Y_N \overset{iid}{\sim} g$
- $\tilde{\mu}_N := \frac{1}{N}\sum_{i=1}^N T(Y_i) \xrightarrow[N\to\infty]{i.p.} \mathbb{E}_g[T]$. $= \mathbb{E}_f[S]$

Claim: $\mathbb{E}_g[T] = \mathbb{E}_f[S]$.

IMPORTANCE SAMPLING

We want to approximately evaluate $\mathbb{E}_f[S]$.

Let $g : \mathcal{X} \to [0, \infty)$ be a pdf or pmf such that

$$g(x) = 0 \quad \Rightarrow \quad S(x)f(x) = 0.$$

Define

$$T(x) := \begin{cases} S(x)\frac{f(x)}{g(x)}, & \text{if } g(x) > 0, \\ 0, & \text{if } g(x) = 0. \end{cases}$$

- Generate $Y_1, \ldots, Y_N \stackrel{iid}{\sim} g$
- $\tilde{\mu}_N := \frac{1}{N}\sum_{i=1}^N T(Y_i) \xrightarrow[N\to\infty]{i.p.} \mathbb{E}_g[T]$.

- Useful when it is easier to sample from g than from f.
- Can lead to improved efficiency.

EXAMPLE: EVALUATE $P(N(0,1) \geq 4)$

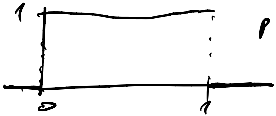
By classical MC integration:

- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $S(x) = \mathbb{1}_{[4,\infty)}(x)$
- $X_1, \dots, X_N \overset{iid}{\sim} f$
- $\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^N S(X_i)$.

By importance sampling:

- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $S(x) = \mathbb{1}_{[4,\infty)}(x)$
- $g(x) = e^{-(x-4)}$, if $x \geq 4$, and $g(x) = 0$ else.
- $Y_1, \dots, Y_N \overset{iid}{\sim} g$
- $\tilde{\mu}_N = \frac{1}{N} \sum_{i=1}^N T(Y_i)$.

$$P(U \le t) = \begin{cases} 0 & t < 0 \\ t & 0 \le t \le 1 \\ 1 & t > 1 \end{cases}$$



pdf

1. The inversion method

We want to sample from a distribution with cdf $F : \mathbb{R} \to [0, 1]$.

Claim: If $U \sim Unif(0, 1)$ then $F^\dagger(U)$ has cdf F.

$t = F(x)$
$\in [0,1]$

Claim: $F^\dagger(\alpha) \le x \iff \alpha \le F(x)$

$$P(F^\dagger(U) \le x) = P(U \le F(x)) = F(x)$$

i.e. cdf of $F^\dagger(U)$ is F.

recall: if F is invertible, then $F^\dagger = F^{-1}$

universität wien

2. Rejection sampling

We want to sample from a distribution with pdf $f : \mathcal{X} \to [0, \infty)$, where possibly $f(x) = c \cdot h(x)$ for unknown norming constant $c > 0$.

$c = \left(\int_{\mathcal{X}} h(x)\, dx \right)^{-1}$

Pick a pdf $g : \mathcal{X} \to [0, \infty)$ and $M > 0$ such that $h(x) \leq Mg(x)$ for all $x \in \mathcal{X}$. Define

$$T(x) := \left. \begin{cases} \frac{h(x)}{Mg(x)}, & \text{if } g(x) > 0, \\ 0, & \text{if } g(x) = 0. \end{cases} \right\} \leq 1$$

Algorithm:

1. Generate $U \sim Unif(0, 1)$ and $Y \sim g$ indep.
2. If $U \leq T(Y)$, return $X = Y$, otherwise discard Y.
3. Repeat sufficiently many times to produce $X_1, \ldots, X_N$.

Claim: $X_1 \sim f$

Introduction to Bayesian Data Analysis

statistical model: $(\mathcal{X}, \Theta, (p_\theta)_{\theta\in\Theta})$, $\Theta \subseteq \mathbb{R}^p$

prior distribution: $\pi : \Theta \to [0, \infty)$ a pdf or pmf

"prior knowledge about the parameter θ"

The 'likelihood'

$$p(x|\theta) = p_\theta(x)$$

is understood as the conditional distribution of the data $X_1, \ldots, X_n$ given θ, and $\theta \sim \pi$.

Goal: update our belief about θ using the data, i.e., compute (features of) posterior distribution

$$p(\theta|x) := \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|\theta)\pi(\theta)d\theta}$$

Bayes formula

Notation 'proportional to' omitting norming constants:

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|\theta)\pi(\theta)d\theta} = c(x)p(x|\theta)\pi(\theta)$$

$$\boxed{p(\theta|x) \propto p(x|\theta)\pi(\theta)}$$

$$c(x) = \left(\int_\Theta p(x|\theta)\pi(\theta)\,d\theta\right)^{-1}$$

- Very popular in applications because of conceptual simplicity and powerful simulation techniques (see later).
- What about the true parameter?
- How to choose the prior?
  - cheating?
  - non-informative prior
  - conjugate prior
  - Jeffrey's prior
- Conclusions of frequentist and Bayesian analysis often almost identical.

Likelihood:

$$p(x|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right), \qquad \theta \in \Theta = \mathbb{R}$$

prior:

$$\pi_\mu(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - \mu)^2\right)$$

$\mu \in \mathbb{R} \ldots$ prior mean

Find posterior distribution!

Likelihood:

$$p(x|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right), \qquad \theta \in \Theta = \mathbb{R}$$

Gaussian prior:

$$\pi_\mu(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - \mu)^2\right)$$

$\mathcal{N}(\hat\theta_n, \hat\sigma_n^2)$

$\mu \in \mathbb{R} \ldots$ prior mean

Gaussian posterior distribution: ($\pi_\mu \ldots$'conjugate prior')

$$p(\theta|x,\mu) \propto p(x|\theta)\pi(\theta) \propto \frac{1}{\sqrt{2\pi\hat\sigma_n^2}} \exp\left(-\frac{1}{2}\frac{(\theta - \hat\theta_n)^2}{\hat\sigma_n^2}\right)$$

$\hat\theta_n = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}\mu \quad \ldots$ posterior mean $\hat\theta_n(x) = \int_{\mathbb{R}} \theta \cdot p(\theta|x,\mu) d\theta$

$\hat\sigma_n^2 = \frac{1}{n+1}$

Likelihood:

$$p(x|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right), \qquad \theta \in \Theta = \mathbb{R}$$

non-informative prior / improper prior ($N(0, \infty)$):

$$\boxed{\pi(\theta) = 1, \quad \theta \in \Theta = \mathbb{R} \quad \text{not a pdf !!!}}$$

Can still find Gaussian posterior density

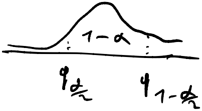
$$p(\theta|x) \propto p(x|\theta) \cdot 1 \propto \frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}} \exp\left(-\frac{1}{2}\frac{(\theta - \bar{x}_n)^2}{\hat{\sigma}_n^2}\right), \quad \hat{\sigma}_n^2 = \frac{1}{n}$$

Here: posterior mean $= \bar{x}_n =$ frequentist MLE

Bayesian Credible Intervals

For $\alpha \in (0,1)$ and a sample $x \in \mathcal{X}$, a $1-\alpha$ Bayesian credible interval is an interval

$$\boxed{\begin{gathered} BI_\alpha(x) = [L_\alpha(x), U_\alpha(x)] \\ \text{such that} \\ P(\theta \in BI_\alpha(x) | X = x) \geq 1 - \alpha. \end{gathered}}$$



Note: Here, the probability is over the randomness in θ given the data, i.e., w.r.t. the posterior distribution,

$$P(L_\alpha(x) \leq \theta \leq U_\alpha(x) | X = x) = \int_{L_\alpha(x)}^{U_\alpha(x)} p(\theta|x) d\theta.$$

posterior distrib.

Thus, $L_\alpha(x)$ and $U_\alpha(x)$ are quantiles of the posterior distribution.

universität wien

Gaussian Likelihood and Gaussian prior

$$p(x|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right),$$

$$\pi_\mu(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - \mu)^2\right)$$

Gaussian posterior distribution: ($\hat{\theta}_n = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}\mu$, $\hat{\sigma}_n^2 = \frac{1}{n+1}$)

$$p(\theta|x,\mu) = \frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}} \exp\left(-\frac{1}{2}\frac{(\theta - \hat{\theta}_n)^2}{\hat{\sigma}_n^2}\right)$$

$$BI_\alpha(x) = ?$$

<u>Point estimation:</u> compute posterior mean (or median or mode), e.g., $\hat{\theta}_n = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}\mu$.

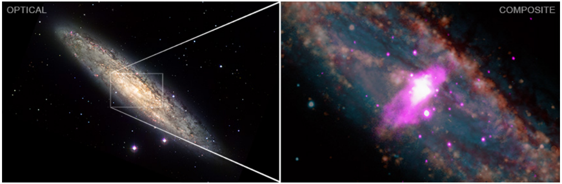
<u>Credible intervals:</u> compute quantiles of the posterior distribution.

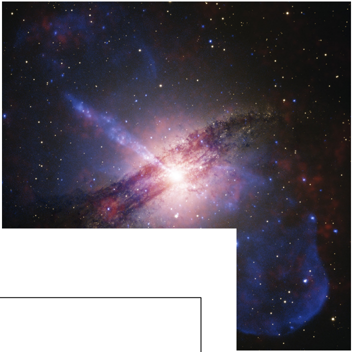
All of Bayesian statistics is concerned with computing features of the posterior distribution!

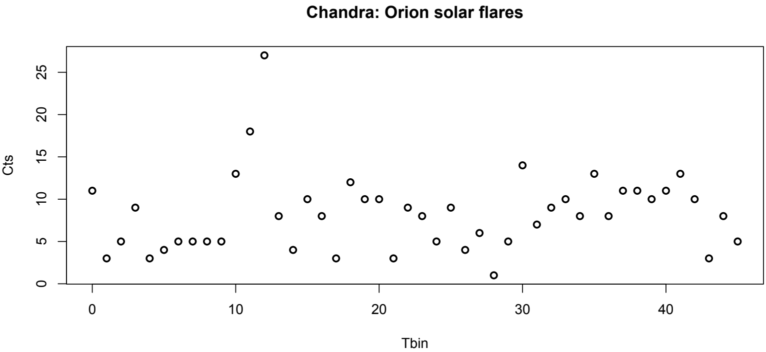
$$\boxed{p(\theta|x) := \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)d\theta}}$$

Can be hard in complex models! But all we need is to be able to sample from the posterior distribution!

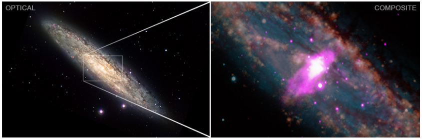
Example: Bayesian change point detection

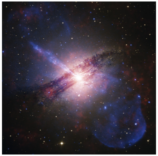






EXAMPLE: BAYESIAN CHANGE POINT DETECTION





The raw data gives the **individual photon arrival times (in seconds)** and their energies (in keV). The processed data we consider here is obtained by **grouping the events into evenly-spaced time bins (10,000 seconds width)**.

Our goal for this data analysis is to identify the change point and estimate the intensities of the process before and after the change point.

Source:
http://www.iiap.res.in/astrostat/School07/R/MCMC.html

Example: Bayesian change point detection

Data:

$Y_i \in \mathbb{N}$ ... counts of events in time interval $i = 1, \dots n$

" $k \in \{1, \dots, n-1\}$ ... change point " not observed!

How to model this?

$$Y_i = \sum_{t=1}^{10000} X_t$$

$$X_t = \begin{cases} 1, & \text{photon arrival at time } t, \\ 0, & \text{no arrival at time } t. \end{cases}$$

$p = P(X_t = 1)$ is small! Counting rare events.

$$Y_i \sim \text{Binomial}(10000, p)$$

EXAMPLE: BAYESIAN CHANGE POINT DETECTION

Data:

" $Y_i \in \mathbb{N} \ldots$ counts of events in time interval $i = 1, \ldots n$

$k \in \{1, \ldots, n-1\} \ldots$ change point " not observed

Model:

$$\left.\begin{array}{rcl} Y_1, \ldots, Y_k | k, \theta, \lambda & \overset{iid}{\sim} & \text{Poisson}(\theta) \\ Y_{k+1}, \ldots, Y_n | k, \theta, \lambda & \overset{iid}{\sim} & \text{Poisson}(\lambda) \end{array}\right\} \text{indep.}$$

prior:

$$\left.\begin{array}{lcl} \theta | b_1, b_2 & \overset{iid}{\sim} & \text{Exp}(b_1) = \text{Gamma}(1, b_1) \\ \lambda | b_1, b_2 & \overset{iid}{\sim} & \text{Exp}(b_2) \\ k | b_1, b_2 & \overset{iid}{\sim} & \text{Unif}(\{1, \ldots, n-1\}) \end{array}\right\} \text{indep.}$$

hyper-prior : $b_1, b_2 \overset{iid}{\sim} \text{Exp}(1)$

Advanced MC methods

Gibbs sampler

Goal: sample from a multivariate pdf $\boxed{f(x), x \in \mathbb{R}^p.}$

Algorithm:

$x = (x_1 \dots x_p)$

Choose a starting value $X^{(0)} \in \mathbb{R}^p$ with $f(X^{(0)}) > 0$.

For $t = 1, 2, \dots$

1.) generate $X_1^{(t)}$ from density $x_1 \mapsto f_1(x_1 | X_2^{(t-1)}, \dots, X_p^{(t-1)})$

k.) generate $X_k^{(t)}$ from density

$$x_k \mapsto f_k(x_k | X_1^{(t)}, \dots, X_{k-1}^{(t)}, X_{k+1}^{(t-1)}, \dots, X_p^{(t-1)})$$

p.) generate $X_p^{(t)}$ from density $x_p \mapsto f_p(x_p | X_1^{(t)}, \dots, X_{p-1}^{(t)})$

… to get $X^{(t)} = (X_1^{(t)}, \dots, X_p^{(t)})^T \in \mathbb{R}^p$.

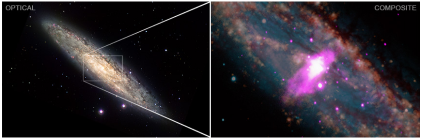
Theoretical properties:

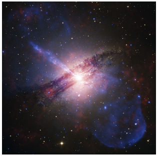
- $(X^{(t)})_{t\in\mathbb{N}}$ is a Markov chain in $\mathbb{R}^p$, i.e.,

  $$P(X^{(t)} \in A | X^{(0)}, \dots, X^{(t-1)}) = P(X^{(t)} \in A | X^{(t-1)}), \quad \forall A \subseteq \mathbb{R}^p.$$

- $X^{(t)} \xrightarrow[t\to\infty]{d.} X \sim f$
-
  $$\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) \xrightarrow[T\to\infty]{i.p.} \mathbb{E}_f[h]$$

EXAMPLE: BAYESIAN CHANGE POINT DETECTION





posterior distribution:

$$p(\theta, \lambda, k, b_1, b_2 | Y) \propto$$
$$\propto \theta^{\sum_{i=1}^{k} Y_i} e^{-k\theta} \lambda^{\sum_{i=k+1}^{n} Y_i} e^{-(n-k)\lambda} b_1 e^{-n_1 \theta} b_2 e^{-b_2 \lambda} e^{-b_1} e^{-b_2}.$$

For Gibbs we need all five univariate conditional distributions:

$p(\theta|\lambda, k, b_1, b_2, Y)$, $p(\lambda|\theta, k, b_1, b_2, Y)$, $p(k|\theta, \lambda, b_1, b_2, Y)$,
$p(b_1|\theta, \lambda, k, b_2, Y)$, $p(b_2|\theta, \lambda, k, b_1, Y)$.

But that is easy!

Metropolis-Hastings

Goal: sample from a multivariate pdf $f(x)$, $x \in \mathbb{R}^p$.

Algorithm:

Choose a starting value $X^{(0)} \in \mathbb{R}^p$ with $f(X^{(0)}) > 0$.

For $t = 1, 2, \dots$

A.) generate Z from a proposal density $y \mapsto q(z|X^{(t-1)})$.

B.) compute

$$\alpha := \alpha(Z, X^{(t-1)}) := \min\left(1, \frac{f(Z)}{f(X^{(t-1)})} \frac{q(X^{(t-1)}|Z)}{q(Z|X^{(t-1)})}\right)$$

C.) set

$$X^{(t)} = \begin{cases} Z, & \text{with probability } \alpha, \\ X^{(t-1)}, & \text{with probability } 1 - \alpha. \end{cases}$$