Lecture 8: Intro to Markov Chain Simulation

UNIVERSITY OF
SAN FRANCISCO

James D. Wilson

MSAN 628

Computational Statistics

Outline

- Review of Monte Carlo, Rejection, and Importance Sampling
- Intro to Markov Chains
- Lots of Examples

Simulations so far: Ingredients

- Target density: (posterior density) $p(\theta \mid y)$
  - posterior is calculated using Bayes rule from prior $\pi(\theta)$ and density $f(y \mid \theta)$
  - everything generalizes to *any* target density!
- Unnormalized density: $q(\theta \mid y) = \pi(\theta)f(y \mid \theta)$
  - the kernel of Bayes rule
- Proposal distribution: $g(\theta)$ - must be an integrable and non-negative function for all θ.
  - We'd like something "close" to $p(\theta \mid y)$
  - The further away $g(\theta)$ is from $p(\theta \mid y)$, the less efficient sampling will be!

Simulations so far: Monte Carlo Simulations

Goal: Estimate expectations of functions of θ given y: $\mathbb{E}[h(\theta) \mid y]$

- We know $p(\theta \mid y)$ and know how to simulate from it!
- Want to calculate $\mathbb{E}[h(\theta) \mid y = \int h(\theta)p(\theta \mid y)d\theta$

  Strategy:

  - Simulate S values $\theta^1, \ldots, \theta^S$ from $p(\theta \mid y)$
  - Approximate integral with $\frac{1}{S}\sum_{s=1}^{S} h(\theta^s)$

- Estimate gets better with larger S!

Simulations so far: Rejection Sampling

Goal: Simulate from some “funny" (but known) posterior distribution $p(\theta \mid y)$ whose random variable we do not understand.

Examples:

- $p(\theta \mid y) = \frac{1}{\pi y}\left(1 + \left(\frac{\theta - 1}{y}\right)^2\right)$
- $p(\theta \mid y) = \frac{1}{y}\exp\left(-\frac{2|\theta - y|}{y}\right)$
- $p(\theta \mid y) = \frac{1}{4!y^5}\theta^4\exp\left(-\frac{\theta}{y}\right)$

Each of the above are well-defined probability density functions, but you may not recognize which ones..

Simulations so far: Rejection Sampling

Goal: Simulate from some “funny" (but known) posterior distribution $p(\theta \mid y)$ whose random variable we do not understand.

Strategy:

- Choose a “nice” proposal distribution $g(\theta)$ (that you know how to simulate from)
- Choose an $M < \infty$ such that $Mg(\theta) > p(\theta \mid y)$ for all θ

  Two-step algorithm for simulating a single value from $p(\theta \mid y)$

  - Simulate θ from $g(\theta)$
  - Accept θ with probability $\dfrac{p(\theta \mid y)}{Mg(\theta)}$. If rejected, repeat the first step and try again!

Simulations so far: Importance Sampling

Goal: Calculate expectations of some function of θ given y **without** knowing $p(\theta \mid y)$.

- What?!
- OK, let's use a proposal density that I **do** know: $g(\theta)$
- Ah right, I can re-write $\mathbb{E}[h(\theta) \mid y]$ in the following way:

$$\mathbb{E}[h(\theta) \mid y] = \frac{\int [h(\theta) q(\theta \mid y)/g(\theta)] \, g(\theta) d\theta}{\int [q(\theta \mid y)/g(\theta)] \, g(\theta) d\theta}$$

- Which, using Monte Carlo simulations can be estimated with sums... (on next slide)

Simulations so far: Importance Sampling

Goal: Calculate expectations of some function of θ given y **without** knowing $p(\theta \mid y)$.

Strategy:

- Simulate S samples $\theta^1, \ldots, \theta^S$ from the "nice" function $g(\theta)$
- Estimate $\mathbb{E}[h(\theta) \mid y]$ using Monte Carlo:

$$\mathbb{E}[h(\theta) \mid y] \approx \frac{\sum_{s=1}^{S} h(\theta^s) w(\theta^s)}{\sum_{s=1}^{S} w(\theta^s)}$$

Here, $w(\theta^s) = q(\theta^s)/g(\theta^s)$ are importance weights.

Summary

- We can now simulate from any “funny" density!!! (simply use rejection sampling)
- We can readily approximate expectations of functions of θ given y with **known** posterior densities (Monte Carlo Simulations) and **unknown** posterior densities (Importance Sampling)
- What else could we want?!

It's all about Efficiency!

- Well, we have little intuition on how efficient the methods we've looked at so far. They all (at least) rely on S random draws. (how big must S be?)
- Further, they depend on a "good" proposal distribution (how "good" must $g(\theta)$ be?)
- We avoid going deeply into the analysis here (it's not easy and actually still not fully understood). Such an analysis is known as mixing time analysis
- We can improve efficiency using smart sampling based on Markov Chains! So first, we must learn the basics of Markov Chains..

Stochastic Processes

Stochastic Process

A stochastic process with state space S is a dynamic sequence of random variables $X = \{X_t : t \in [0, \infty)$, where at each time t the random variable $X_t \in S$

- Continuous time process: X is a sequence of an uncountably infinite number of random variables
- Discrete time process: X is a sequence of a countably (infinite or finite) number of random variables
- We will only be concerned with discrete time processes, where $X = \{X_t : t \in \mathbb{Z}\}$

Stochastic Processes: Examples

Time Series Models!

- AR(p) models: $X_t = C + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$
- MA(q) models: $X_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$
- ARMA(p,q) models: $X_t = C + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$

Above, ϵ_t is typically assumed to be white noise, which brings about the randomness in X_t

Analyzing Stochastic Processes

Major Aim: Calculate joint probabilities:

$$\mathbb{P}(X_{t_1} = s_1, X_{t_2} = s_2, \ldots, X_{t_k} = s_k)$$

- In general, this is not easy!
- To solve this, we have to either have to assume some sort of “step-wise” dependence (as in the ARMA - based models) or some sort of conditional independence
- Markov Chains have a nice representation that allows easy calculation of joint probabilities

Markov Chains: History

- Founded in 1906 by Andrey Markov, a prominent Russian mathematician
- Discovered to end a dispute about whether or not independence was needed for the weak law of large numbers to hold

Markov Chains: Preliminaries

- **Basic Idea**: "if I knew what happened yesterday, nothing before that matters"
- **Mathematically speaking**: For a stochastic process $X = \{X_t : t \in \mathbb{Z}\}$, we assume conditional independence:

$$\mathbf{P}(X_t = s_t \mid X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = \mathbf{P}(X_t = s_t \mid X_{t-1} = s_{t-1}) \qquad (1)$$

- Equation (1) is called the Markov property

Markov Chains: Preliminaries

Markov Chains

A stochastic process $X = \{X_t : t \in \mathbb{Z}\}$ on state space S that satisfies the Markov property is called a Markov chain.

Notes:

- Technically, X is a Markov chain of order 1. A Markov chain of order q depends on the last q observations in the process.
- Also, technically this is a *Discrete time Markov chain* (DTMC) since time is observed in discrete increments. (There is also a *continuous time Markov chain*, but we won't be talking about those.)

Markov Chains

Aim (revisited): calculate $\mathbb{P}(X_0 = s_0, X_1 = s_1, \ldots, X_k = s_k)$

- For a Markov chain of order 1, we can use conditional independence to calculate joint probabilities:

$$\mathbb{P}(X_0 = s_0, X_1 = s_1, \ldots, X_k = s_k) = \mathbb{P}(X_0 = s_0) \prod_{t=1}^{k} \mathbb{P}(X_t = s_t \mid X_{t-1} = s_{t-1})$$

- The above decomposition suggests that we need the following quantities to fully characterize the stochastic process X:
    - Initial state probabilities: $\{p_j : j \in S\} = \{\mathbb{P}(X_0 = j)\}$
    - Transition probabilities: $\{p_{i,j} : i, j \in S\}$, where $p_{i,j} = \mathbb{P}(X_t = i \mid X_{t-1} = j)$

Transition Probabilities

- The previous slide shows that to know everything about a Markov chain on state space S, we need initial state probabilities and transition probabilities (so we need $|S| + |S^2|$ probabilities)
- For finite S, we call the one-step transition probability matrix $\mathbb{T}$:

$$\mathbb{T} = (p_{i,j})_{i,j \in S}$$

- **Important Feature**: For each $i \in S$, we *must* have $\sum_{j \in S} p_{i,j} = 1$. In other words, the rows of $\mathbb{T}$ sum to 1! (why?)

Example

Example 1

Suppose that the chance of rain tomorrow depends on previous weather only through whether or not it is raining today. Suppose also that if it rains today, it will rain tomorrow with probability α, and if it does not rain today, it will rain tomorrow with probability β.

1. Is the chance of rain on day *t* a Markov chain?
2. If so, what transition probability matrix describes this weather process and its state space?

Example

Example 2

On any given day Eric is either cheerful (C), so-so (S), or glum (G). If he is cheerful today, then he will be C, S, or G tomorrow with repective probabilities 0.5, 0.4, 0.1. If he is feeling so-so today, then he will be C, S, or G tomorrow with probabilities 0.3, 0.4, 0.3. If he is glum today, then he will be C, S, or G tomorrow with probabilities 0.2, 0.3, 0.5. Let X_n denote Gary's mood on day n. Then, is $\{X_n, n \geq 0\}$ a Markov chain? If so, what is its transition probability matrix and state space?

Example

Example 3

Consider a gambler who, at each play of a game, either wins \$1 with probability p or loses \$1 with probability $1 - p$. Suppose the gambler quites playing either when he goes broke or he attains a fortune of $\$N$. What transistion probability matrix and state space characterizes the gambler's winnings?

Example: Google's Page Rank

Page Rank

Aim: given n interlinked webpages, rank them in order of importance. Key desire: assign pages importance scores $\mathbf{x} = (x_1, \ldots, x_n)$, where $x_i \geq 0$. To come up with importance scores, Google founders created a Markov chain which describes the likelihood of linking from website i to website j. In particular, they specified:

$$p_{i,j} = \frac{1}{n_j}\mathbb{I}(\text{page } j \text{ links to page } i),$$

where n_j = number of outgoing links on page j. Subsequently, the founders identified the ranking vector as the vector $\mathbf{x}$ that solves: $\mathbb{T}\mathbf{x} = \mathbf{x}$.

- Is $\mathbb{T} = (p_{i,j})_{i,j\leq n}$ a valid transition probability matrix?

Next Time:

- Markov chain Simulation
  - Gibbs Sampler
  - Metropolis Algorithm
  - Metropolis Hastings Algorithm