MCMC Tutorial

A short introduction to Bayesian Analysis and Metropolis-Hastings MCMC

Ralph Schlosser

https://github.com/bwv988

February 2017

Overview

- Bayesian Analysis
- Monte Carlo Integration
- Sampling Methods
- Markov Chain Monte Carlo
- Metropolis-Hastings Algorithm
- Example: Linear Regression and M-H MCMC
- Outlook

Bayesian Analysis: Introduction

- Foundation for MCMC: Bayesian Analysis.
- Frequentist – Likelihood model: $p(\underline{x}|\underline{\theta})$
- How likely are the data $\underline{x}$, given the fixed parameters $\underline{\theta}$.
- In general we want to estimate $\underline{\theta}$, e.g. through MLE.
- Bayesian – Bayesian model: $p(\underline{\theta}|\underline{x})$.
- Fundamental difference: In Bayesian analysis, both parameter model and data are treated as random variables.

Thomas Bayes (1707-1761)

Bayesian Analysis: Terminology

- From joint probability distribution to posterior distribution via data likelihood and prior beliefs:

$$
\begin{aligned}
p(x,\theta) &= p(x|\theta)p(\theta) \\
p(\theta|x) &= \frac{p(x,\theta)}{p(x)} \\
&= \frac{p(x|\theta)p(\theta)}{p(x)}
\end{aligned}
$$

- Normalizing term $p(x)$ difficult to get, but often not needed:

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

- Posterior $\propto$ likelihood $\times$ prior

Bayesian Analysis: Pros and Cons

- Pro: Common-sense interpretability of results, e.g. *Credible Intervals* vs. classical Confidence Intervals.
- Pro: Update model parameters as new data becomes available.
- Pro: Create *hierarchical models* through chaining:
  - $p(\phi, \theta|x) = p(x|\phi, \theta)p(\theta|\phi)p(\phi)$
  - Hyperprior: $p(\theta|\phi)p(\phi)$
  - *Yesterdays posterior is tomorrow's prior*
- Con: **Must** have a joint model for parameters, data, and prior.
  - What if we have absolutely no prior information?
- Con: Choice of prior considered to be subjective.
- Con: Subjectiveness makes comparison difficult.

Bayesian Analysis: Applications

- Inferences and predictions in a Bayesian setting:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\theta')p(\theta')d\theta'} \qquad \text{Normalization}$$

$$p(\tilde{y}|y) = \int_\Theta p(\tilde{y}|\theta')p(\theta'|y)d\theta' \qquad \text{Predict new data}$$

- Posterior summary statistics, e.g. expectations:

$$\mathbb{E}_p(g(\theta)|x) = \int_\Theta g(\theta')p(\theta'|x)d\theta'$$

mean: $g(\theta) = \theta$

- Many classical models can be expressed in a Bayesian context, like e.g. linear regression, ARMA, GLMs, etc.
- Missing data: Natural extension.

Monte Carlo Integration: Introduction

- Applied Bayesian analysis asks to integrate over (often analytically intractable) posterior densities.
- Solution: **Monte Carlo Integration**
- Suppose we wish to evaluate $\mathbb{E}_p(g(\theta)|x) = \int_\Theta g(\theta')p(\theta'|x)d\theta'$
- Given a set of N i.i.d. samples $\theta_1, \theta_2, ..., \theta_N$ from the density p:

$$\mathbb{E}_p(g(\theta|x)) \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta_i)$$

- But: Need to be able to draw random samples from p!

Monte Carlo Integration: Example

Simulate $N = 10000$ draws from a univariate standard normal, i.e. $X \sim N(0, 1)$. Let $p(x)$ be the normal density. Then:

$$P(X \leq 0.5) = \int_{-\infty}^{0.5} p(x)dx$$

```r
set.seed(123)
data <- rnorm(n = 10000)
prob.in <- data <= 0.5
sum(prob.in) / 10000
```

```
## [1] 0.694
```

```r
pnorm(0.5)
```

```
## [1] 0.6914625
```

Sampling Methods

- Sampling from the posterior distribution is really important.
- Classical sampling methods:
  - Inversion sampling
  - Importance sampling
  - Rejection sampling
- Drawbacks:
  - Closed-form expression rarely accessible (Method of Inversion).
  - Doesn't generalize well for highly-dimensional problems.
- Metropolis-Hastings MCMC has largely superseded the above.

Markov Chain Monte Carlo (MCMC)

- Unlike pure Monte Carlo, in MCMC we create dependent samples.
- Consider the **target distribution** $p(\theta|x)$ which is only known up to proportionality.
- Construct a Markov Chain in the state space of $\theta \in \Theta$ with stationary distribution $p(\theta|x)$.
- Markov property – New state of chain depends only on previous state (K: transitional kernel d.f.).

$$\theta_{t+1} = K(\theta|\theta_t)$$

- With realizations $\{\theta_t : t = 0, 1, ...\}$ from the chain:

$$\theta_t \rightarrow p(\theta|x)$$

$$\frac{1}{N}\sum_{t=1}^{N} g(\theta_t) \rightarrow \mathbb{E}_p(g(\theta|x)) \text{ a.s.}$$

Metropolis-Hastings MCMC: Intro & some history

- An implementation of MCMC.
- Originally developed by researchers Nicholas Metropolis, Stanislaw Ulam, and co. at Los Alamos National Laboratories in the 1950's.
- Generalized through work done by Hastings in the 1970's.
- Popularized by a 1990 research paper from Gelfand & Smith: `http://wwwf.imperial.ac.uk/~das01/MyWeb/SCBI/Papers/GelfandSmith.pdf`
- M-H MCMC really helped turning Bayesian analysis into practically useful tool.

Metropolis-Hastings MCMC: Terminology

- M-H has two main ingredients.
- A proposal distribution.
  - Dependent on the current chain state θ_t, generate a candidate for the new state ϕ.
  - Written as $q(\theta_t, \phi)$.
  - Can be chosen arbitrarily, but there are caveats (efficiency).
- An acceptance probability.
  - Accept with probability α the move from the current state θ_t to state ϕ.
  - Written as $\alpha(\theta_t, \phi)$.
- Main idea behind M-H: With every step, we want to get closer to the target density (e.g. posterior density).

Metropolis-Hastings MCMC: Intuition

- Let's call our target distribution (from which we want to sample) π.
- At the core of the M-H algorithm we have the calculation of $\alpha(\theta_t, \phi)$:

$$\alpha(\theta_t, \phi) = min\Big(1, \frac{\pi(\phi)q(\phi, \theta_t)}{\pi(\theta_t)q(\theta_t, \phi)}\Big)$$

- Often q is symmetric, in which case it cancels out.
- If $\frac{\pi(\phi)}{\pi(\theta_t)} > 1 \rightarrow$ target density at the proposed **new** value is higher than at current value.
- In this case, we will accept the move from θ_t to ϕ with probability 1.
    - *M-H really loves upward moves* :)
- **Main point**: Working with ratios of π, so only need π up to proportionality!

Metropolis-Hastings MCMC: Algorithm

1. Initialize θ_0, number of iterations.
2. Given the current state θ_t, generate new state ϕ from the proposal distribution $q(\theta_t, \phi)$.
3. Calculate acceptance probability $\alpha(\theta_t, \phi)$.
4. With probability $\alpha(\theta_t, \phi)$, set $\theta_{t+1} = \phi$, else set $\theta_{t+1} = \theta_t$.
5. Iterate
6. Result: Realizations of dependent samples $\{\theta_1, \theta_2, ...\}$ from the target distribution $\pi(\theta)$.

Using these dependent realizations & due to the Monte Carlo approach, we can now look at making inferences and predictions.

Example: Linear Regression and M-H MCMC

- Consider a simple linear model: $y = \beta_1 x + \epsilon$.
- As usual $\epsilon \sim N(0, \sigma^2)$ with σ^2 known.
- We wish to make inferences on, e.g. β_1.
- Bayesian approach:

$$p(\beta_1 | y, x, \sigma^2) = p(y | \beta_1, x, \sigma^2) p(\beta_1)$$

- Let's choose a uniform prior for β_1. We can now create samples using M-H MCMC.
- See R code!

Outlook

- Many more interesting things could be mentioned, e.g. burn-in, choice of q, Gibbs-sampling etc.
- M-H and Monte Carlo in deep learning: http://www.deeplearningbook.org/contents/monte_carlo.html
- Bayesian Deep Learning is a thing (apparently, don't know anything about it!)
- Went way over my head, but looks cool – Finding the Higgs boson, featuring Monte Carlo & Bayes: http://hea-www.harvard.edu/AstroStat/Stat310_1314/dvd_20140121.pdf
- Along the same lines, the amazing NIPS 2016 keynote: https://nips.cc/Conferences/2016/Schedule?showEvent=6195
- M-H in Latent Dirichlet Allocation: http://mlg.eng.cam.ac.uk/teaching/4f13/1112/lect10.pdf