Bayesian Statistics I

Lecture 8 - Markov Chain Monte Carlo. Metropolis-Hastings.

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Lecture overview

- **Markov Chain Monte Carlo**

- **Metropolis-Hastings**

- **MCMC - efficiency, burn-in and convergence**

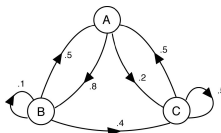# Markov chains

- Let $\mathcal{S} = \{s_1, s_2, ..., s_k\}$ be a finite set of states.
  - ▶ Weather: $\mathcal{S} = \{\text{sunny, rain}\}$.
  - ▶ School grades: $\mathcal{S} = \{A, B, C, D, E, F\}$

- **Markov chain** is a stochastic process $\{X_t\}_{t=1}^{T}$ with **state transitions**
$$p_{ij} = \Pr(X_{t+1} = s_j | X_t = s_i)$$

- School grades: $X_1 = C, X_2 = C, X_3 = B, X_4 = A, X_5 = B$.

- **Transition matrix** for weather example
$$P = \left(\begin{array}{cc} p_{11} & p_{12} \\ p_{21} & p_{22} \end{array} \right) = \left(\begin{array}{cc} 0.9 & 0.1 \\ 0.7 & 0.3 \end{array} \right)$$

Stationary distribution

■ h-**step transition probabilities**

$$P_{ij}^{(h)} = \Pr(X_{t+h} = s_j | X_t = s_i)$$

■ h-**step transition matrix** by **matrix power**

$$P^{(h)} = P^h$$

■ **Unique** equilbrium distribution $\pi = (\pi_1, ..., \pi_k)$ if chain is
  ▶ **irreducible** (possible to get to any state from any state)
  ▶ **aperiodic** (does not get stuck in predictable cycles)
  ▶ **positive recurrent** (expected time of returning is finite)

■ Limiting long-run distribution

$$P^t \rightarrow \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \text{ as } t \rightarrow \infty$$

Stationary distribution, cont.

■ Limiting long-run distribution

$$P^t \rightarrow \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \text{ as } t \rightarrow \infty$$

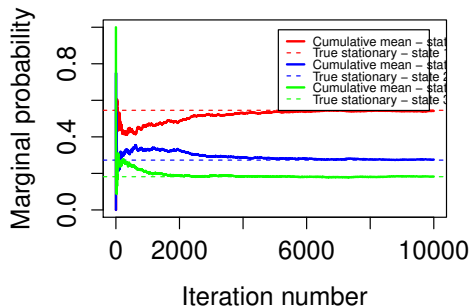■ **Stationary distribution**

$$\pi = \pi P$$

■ Example:

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$
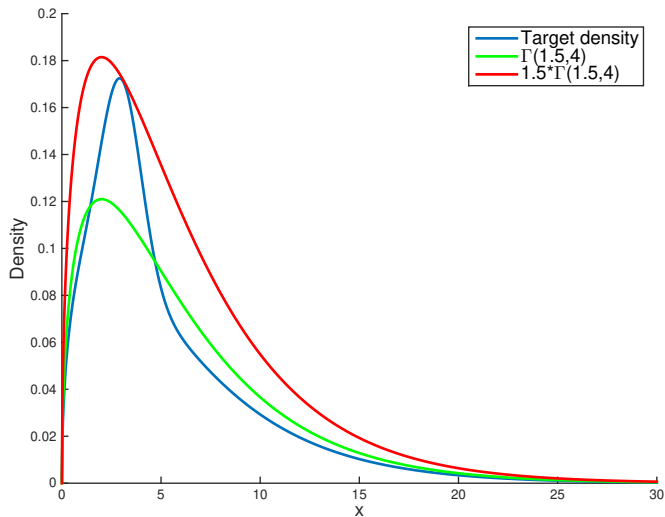
$$\pi = (0.545, 0.272, 0.181)$$

The basic MCMC idea

- Simulate from discrete distribution $p(x)$ when $x \in \{s_1, .., s_k\}$.

- **MCMC**: **simulate a Markov Chain** with a **stationary distribution** that is exactly $p(x)$.

- How to set up the transition matrix P? **Metropolis-Hastings**!

Rejection sampling

Random walk Metropolis algorithm

- **Initialize** $\theta^{(0)}$ and iterate for $i = 1, 2, \ldots$

  1. **Sample proposal**: $\theta_p | \theta^{(i-1)} \sim N\left(\theta^{(i-1)}, c \cdot \Sigma\right)$

  2. Compute the **acceptance probability**

  $$\alpha = \min\left(1, \frac{p(\theta_p|\mathbf{y})}{p(\theta^{(i-1)}|\mathbf{y})}\right)$$

  3. With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

Random walk Metropolis, cont.

- Assumption: we can compute $p(\theta_p|\mathbf{y})$ for any θ.

- Proportionality constants in posterior cancel out in

$$\alpha = \min\left(1, \frac{p(\theta_p|\mathbf{y})}{p(\theta^{(i-1)}|\mathbf{y})}\right).$$

- In particular:

$$\frac{p(\theta_p|\mathbf{y})}{p(\theta^{(i-1)}|\mathbf{y})} = \frac{p(\mathbf{y}|\theta_p)p(\theta_p)/p(y)}{p(\mathbf{y}|\theta^{(i-1)})p(\theta^{(i-1)})/p(y)} = \frac{p(\mathbf{y}|\theta_p)p(\theta_p)}{p(\mathbf{y}|\theta^{(i-1)})p(\theta^{(i-1)})}$$

- **Proportional form of posterior is enough!**

$$\alpha = \min\left(1, \frac{p\left(\mathbf{y}|\theta_p\right)p\left(\theta_p\right)}{p\left(\mathbf{y}|\theta^{(i-1)}\right)p\left(\theta^{(i-1)}\right)}\right)$$

Random walk Metropolis, cont.

■ Common choices of Σ in proposal $N\left(\theta^{(i-1)}, c \cdot \Sigma\right)$:

▶ $\Sigma = I$ (proposes 'off the cigar')
▶ $\Sigma = J_{\hat{\theta}, \mathbf{y}}^{-1}$ (propose 'along the cigar')
▶ **Adaptive**. Start with $\Sigma = I$. Update Σ from initial run.

■ Set c so average acceptance probability is 25-30%.

■ **Good proposal**:

▶ **Easy to sample**
▶ **Easy to compute** α
▶ Proposals should take reasonably **large steps** in θ-space
▶ Proposals should **not be reject too often**.

The Metropolis-Hastings algorithm

■ Generalization when the proposal density is not symmetric.

■ Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, \dots$

1 **Sample proposal**: $\theta_p \sim q\left(\cdot | \theta^{(i-1)}\right)$

2 Compute the **acceptance probability**

$$\alpha = \min\left(1, \frac{p(\mathbf{y}|\theta_p)p(\theta_p)}{p(\mathbf{y}|\theta^{(i-1)})p(\theta^{(i-1)})} \frac{q\left(\theta^{(i-1)}|\theta_p\right)}{q\left(\theta_p|\theta^{(i-1)}\right)}\right)$$

3 With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

The independence sampler

- **Independence sampler**: $q\left(\theta_p|\theta^{(i-1)}\right) = q\left(\theta_p\right)$.

- **Proposal** is **independent of previous draw**.

- Example:

$$\theta_p \sim t_v\left(\hat{\theta}, J_{\hat{\theta},\mathbf{y}}^{-1}\right),$$

  where $\hat{\theta}$ and , $J_{\hat{\theta},\mathbf{y}}$ are computed by numerical optimization.

- Can be very **efficient**, but has a tendency to **get stuck**.

- Make sure that $q\left(\theta_p\right)$ has **heavier tails** than $p(\theta|\mathbf{y})$.

Metropolis-Hastings within Gibbs

- **Gibbs sampling** from $p(\theta_1, \theta_2, \theta_3 | \mathbf{y})$
  - ▶ Sample $p(\theta_1 | \theta_2, \theta_3, \mathbf{y})$
  - ▶ Sample $p(\theta_2 | \theta_1, \theta_3, \mathbf{y})$
  - ▶ Sample $p(\theta_3 | \theta_1, \theta_2, \mathbf{y})$

- When a **full conditional is not easily sampled** we can simulate from it using **MH**.

- Example: at ith iteration, propose θ_2 from $q(\theta_2 | \theta_1, \theta_3, \theta_2^{(i-1)}, \mathbf{y})$. Accept/reject.

- **Gibbs sampling is a special case of MH** when $q(\theta_2 | \theta_1, \theta_3, \theta_2^{(i-1)}, \mathbf{y}) = p(\theta_2 | \theta_1, \theta_3, \mathbf{y})$, which gives $\alpha = 1$. Always accept.

The efficiency of MCMC

- **How efficient** is MCMC compared to iid sampling?
- If $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are **iid** with variance σ^2, then

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{N}.$$

- Autocorrelated $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ generated by **MCMC**

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{N} \left(1 + 2 \sum_{k=1}^{\infty} \rho_k \right)$$

where $\rho_k = Corr(\theta^{(i)}, \theta^{(i+k)})$ is the autocorrelation at lag k.
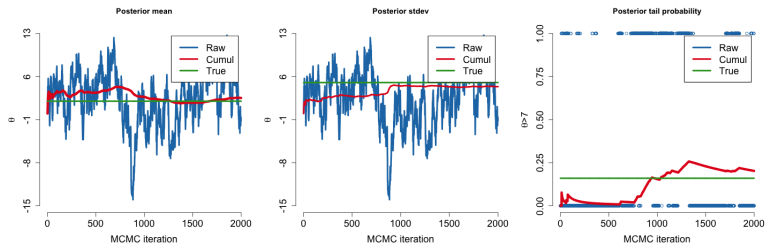
- **Inefficiency factor**

$$\text{IF} = 1 + 2 \sum_{k=1}^{\infty} \rho_k$$

- **Effective sample size** from MCMC

$$\text{ESS} = N/\text{IF}$$

Burn-in and convergence

- How long **burn-in**?

- **How long to sample** after burn-in?

- **Thinning**? Keeping every h draw reduces autocorrelation.

- **Convergence diagnostics**
  - ▶ Raw plots of simulated sequences (trajectories)
  - ▶ CUSUM plots
  - ▶ Potential scale reduction factor, R.

Burn-in and convergence