Bayesian Computations

PhD course in Statistical Inference

Mattias Villani

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

Overview of the lecture

- **Gibbs sampling**
- **Data augmentation**
  - **Mixture models**
  - **Probit regression**
- **Markov Chain Monte Carlo**
- **Metropolis-Hastings**
- **MCMC - efficiency, burn-in and convergence**

- If $\theta^{(1)}, \ldots, \theta^{(N)}$ is an **iid sequence** from $p(\theta)$, then

$$\frac{1}{N}\sum_{t=1}^{N}\theta^{(t)} \quad \rightarrow \quad E(\theta)$$
$$\frac{1}{N}\sum_{t=1}^{N}g(\theta^{(t)}) \quad \rightarrow \quad E[g(\theta)]$$

  for some function $g(\theta)$ of interest.

- Easy to compute **tail probabilities** $\Pr(\theta \leq c)$ by letting

$$g(\theta) = I\left(\theta \leq c\right)$$

  and

$$\frac{1}{N}\sum_{t=1}^{N}g(\theta^{(t)}) = \frac{\#\ \theta\text{-draws smaller than } c}{N}.$$

- Let $F(x)$ be the CDF of X. **Inverse CDF method**:
  1. Generate u from the uniform distribution on $[0, 1]$.
  2. Compute $x = F^{-1}(u)$.

- **Exponential distribution**:

$$u = F(x) = 1 - \exp(-\lambda x)$$

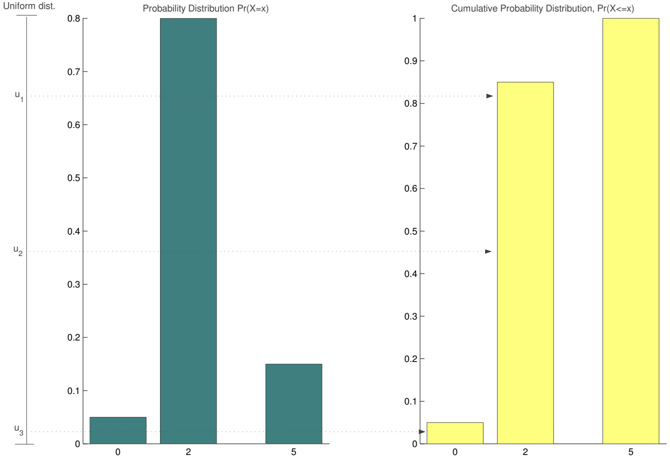
  Inverting gives

$$x = -\ln(1 - u)/\lambda$$

- So:

$$u \sim U(0, 1) \text{ and } x = -\ln(1 - u)/\lambda \Rightarrow x \sim Expon(\lambda)$$

Inverse CDF method, discrete case



- **Cauchy distribution**:

$$
\begin{aligned}
f(x) &= \frac{1}{\pi}\frac{1}{1+x^2} \\
u &= F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)
\end{aligned}
$$

Inverting ...

$$x = \tan[\pi(u - 1/2)].$$

- Use **relations**:

$$y, z \text{ are indep } N(0,1) \Rightarrow \frac{y}{z} \sim \text{Cauchy}(0,1)$$

- **Chi-square**. If $x_1, \ldots, x_\nu \overset{iid}{\sim} N(0,1)$, then $\sum_{i=1}^{\nu} x_i^2 \sim \chi^2_\nu$.

- Easily implemented methods for **sampling from multivariate distributions**, $p(\theta_1, \ldots, \theta_k)$.

- Requirements: Easily sampled **full conditional distributions**:
    - $p(\theta_1|\theta_2, \theta_3 \ldots, \theta_k)$
    - $p(\theta_2|\theta_1, \theta_3, \ldots, \theta_k)$
    - $\vdots$
    - $p(\theta_k|\theta_1, \theta_2, \ldots, \theta_{k-1})$

- Gibbs sampling is a special case of **Metropolis-Hastings** (see Lecture 8).

- Metropolis-Hastings is a **Markov Chain Monte Carlo** (**MCMC**) algorithm.

- Choose initial values $\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}$.
- Repeat for $j = 1, \ldots, N$:
  - Draw $\theta_1^{(j)}$ from $p(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_k^{(j-1)})$
  - Draw $\theta_2^{(j)}$ from $p(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_k^{(j-1)})$

    $\vdots$
  - Draw $\theta_k^{(j)}$ from $p(\theta_k | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{k-1}^{(j)})$
- Return draws: $\theta^{(1)}, \ldots, \theta^{(N)}$, where $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_k^{(j)})$.

- Gibbs draws $\theta^{(1)}, ..., \theta^{(N)}$ are **dependent**, but

$$\bar{\theta} = \frac{1}{N}\sum_{t=1}^{N}\theta_j^{(t)} \quad \rightarrow \quad E(\theta_j)$$

$$\frac{1}{N}\sum_{t=1}^{N} g(\theta^{(t)}) \quad \rightarrow \quad E[g(\theta)]$$

- $\theta^{(1)}, ..., \theta^{(N)}$ **converges in distribution** to the target $p(\theta)$.
- $\theta_j^{(1)}, ..., \theta_j^{(N)}$ converges to the marginal distribution of θ_j, $p(\theta_j)$.
- **Dependent draws** $\rightarrow$ **less efficient** than iid sampling.
- **IID samples**: $\theta^{(1)}, ..., \theta^{(N)}$: $\mathrm{Var}(\bar{\theta}) = \frac{\sigma^2}{N}$.
- **Autocorrelated samples**: $\mathrm{Var}(\bar{\theta}) = \frac{\sigma^2}{N}\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$, where ρ_k is the autocorrelation at lag k.
- **Inefficiency factor**: $1 + 2\sum_{k=1}^{\infty}\rho_k$.

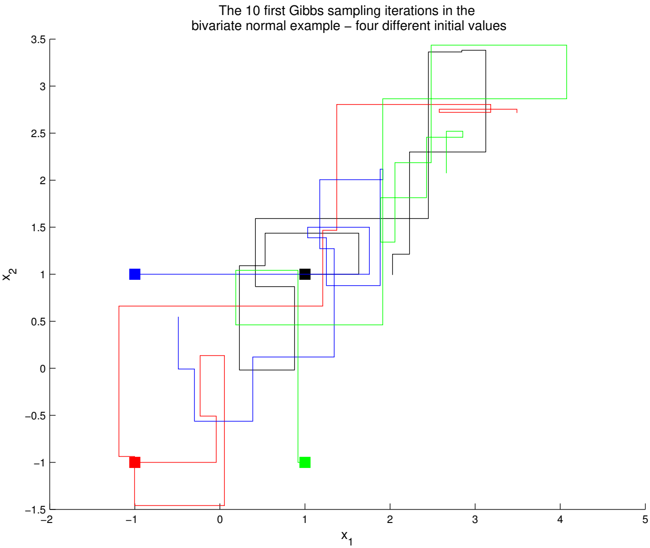
- **Joint distribution**

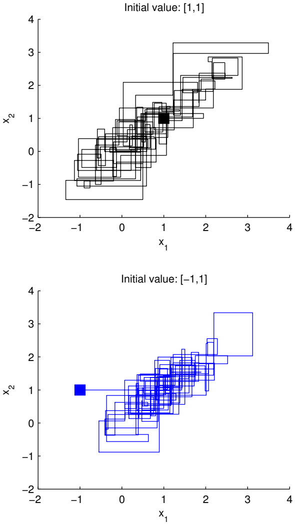
$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

- **Full conditional posteriors**

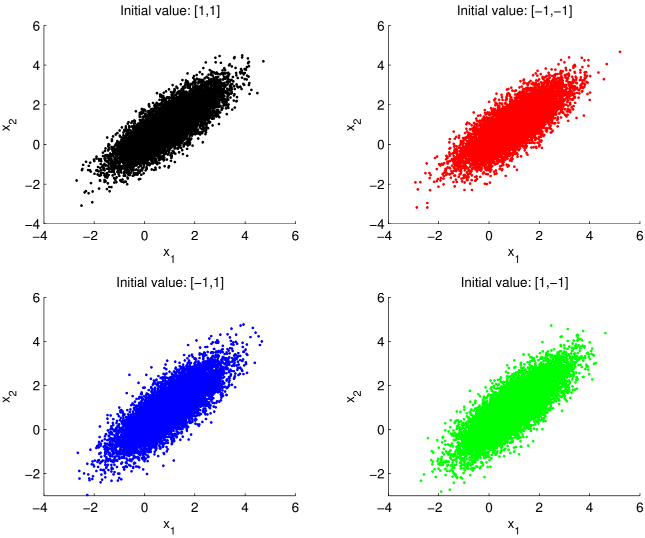
$$\begin{aligned} \theta_1|\theta_2 &\sim N[\mu_1 + \rho(\theta_2 - \mu_2), 1 - \rho^2] \\ \theta_2|\theta_1 &\sim N[\mu_2 + \rho(\theta_1 - \mu_1), 1 - \rho^2] \end{aligned}$$

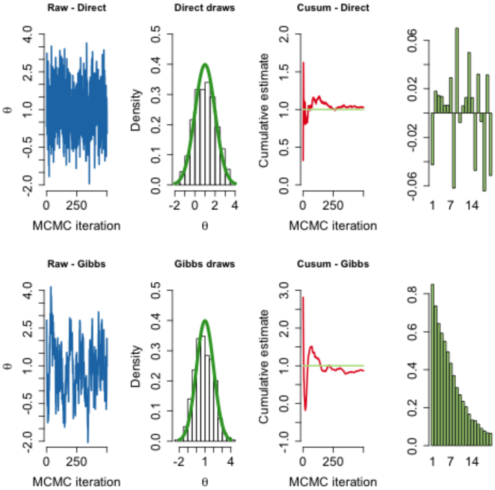
Gibbs sampling - Bivariate normal




Initial value: [1,1]
Initial value: [−1,−1]
Initial value: [−1,1]
Initial value: [1,−1]
x_1
x_2
−2 −1 0 1 2 3 4

Gibbs sampling - Bivariate normal



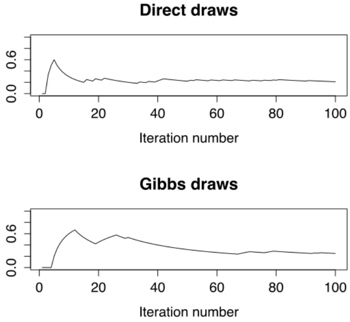

Raw - Direct
Direct draws
Cusum - Direct
θ
MCMC iteration
Density
Cumulative estimate
Raw - Gibbs
Gibbs draws
Cusum - Gibbs

Estimating $Pr(\theta_1 > 0, \theta_2 > 0)$

- Joint probability by counting:

$$Pr(\theta_1 > 0, \theta_2 > 0) \approx N^{-1} \sum_{i=1}^{N} 1(\theta_1^{(i)} > 0, \theta_2^{i)} > 0)$$

.



- **Normal model** with **semi-conjugate prior**

$$\mu \sim N(\mu_0, \tau_0^2)$$
$$\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$$

- **Full conditional posteriors**

$$\mu|\sigma^2, x \sim N\left(\mu_n, \tau_n^2\right)$$
$$\sigma^2|\mu, x \sim Inv - \chi^2\left(\nu_n, \frac{\nu_0\sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2}{n + \nu_0}\right)$$

with μ_n and τ_n^2 defined the same as when σ^2 is known (Lecture 2).

- **AR(p) process**

$$x_t = \mu + \phi_1(x_{t-1} - \mu) + \ldots + \phi_p(x_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2).$$

- Let $\phi = (\phi_1, \ldots, \phi_p)'$.
- **Prior**:
    - $\mu \sim$Normal
    - $\phi \sim$Multivariate Normal
    - $\sigma^2 \sim$Scaled Inverse χ^2.
- The **posterior** can be simulated by **Gibbs sampling**[1]:
    - $\mu | \phi, \sigma^2, x \sim$ Normal
    - $\phi | \mu, \sigma^2, x \sim$ Multivariate Normal
    - $\sigma^2 | \mu, \phi, x \sim$ Scaled Inverse χ^2

[1] Villani (2009). Steady State Priors for Vector Autoregressions. *Journal of Applied Econometrics.*

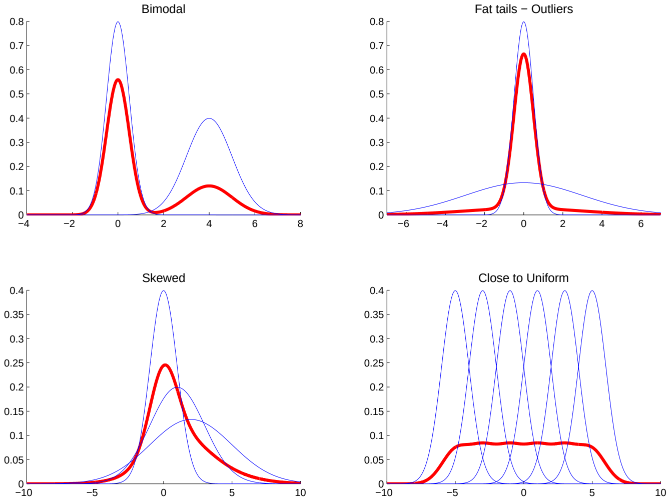
- Let $\phi(x|\mu, \sigma^2)$ denotes the **PDF** of $x \sim N(\mu, \sigma^2)$.

- Two-component **mixture of normals** [MN(2)]

$$p(x) = \pi \cdot \phi(x|\mu_1, \sigma_1^2) + (1 - \pi) \cdot \phi(x|\mu_2, \sigma_2^2)$$

- **Simulate** from a MN(2):
  - Simulate a **membership indicator** $I \in \{1, 2\}$: $I \sim Bern(\pi)$.
  - If $I = 1$, simulate x from $N(\mu_1, \sigma_1^2)$
  - If $I = 2$, simulate x from $N(\mu_2, \sigma_2^2)$.

Illustration of mixture distributions



- The **likelihood** is a product of sums. **Messy** to work with.
- **Assume** that we know where each observation comes from

$$I_i = \begin{cases} 1 \text{ if } x_i \text{ came from Density 1} \\ 2 \text{ if } x_i \text{ came from Density 2} \end{cases}.$$

- Given $I_1, \dots, I_n$ it is easy to estimate $\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ by separating the sample according to the I's.
- But we do **not** know $I_1, \dots, I_n$!
- **Data augmentation**: add $I_1, \dots, I_n$ as unknown parameters.
- **Gibbs sampling**:
  - Sample $\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ **given** $I_1, \dots, I_n$
  - Sample $I_1, \dots, I_n$ **given** $\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$

- Prior: $\pi \sim Beta(\alpha_1, \alpha_2)$. Conjugate prior for (μ_j, σ_j^2), see L5.
- Define: $n_1 = \sum_{i=1}^{n}(I_i = 1)$ and $n_2 = n - n_1$.
- **Gibbs sampling**:
  - $\pi \mid \mathbf{I}, \mathbf{x} \sim Beta(\alpha_1 + n_1, \alpha_2 + n_2)$
  - $\sigma_1^2 \mid \mathbf{I}, \mathbf{x} \sim Inv\text{-}\chi^2(\nu_{n_1}, \sigma_{n_1}^2)$ and $\mu_1 | \mathbf{I}, \sigma_1^2, \mathbf{x} \sim N\left(\mu_{n_1}, \frac{\sigma_1^2}{\kappa_{n_1}}\right)$
  - $\sigma_2^2 \mid \mathbf{I}, \mathbf{x} \sim Inv\text{-}\chi^2(\nu_{n_2}, \sigma_{n_2}^2)$ and $\mu_2 | \mathbf{I}, \sigma_2^2, \mathbf{x} \sim N\left(\mu_{n_2}, \frac{\sigma_2^2}{\kappa_{n_2}}\right)$
  - $I_i \mid \pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \mathbf{x} \sim Bern(\theta_i)$, $i = 1, ..., n$,

$$\theta_i = \frac{(1-\pi)\phi(x_i; \mu_2, \sigma_2^2)}{\pi\phi(x_i; \mu_1, \sigma_1^2) + (1-\pi)\phi(x_i; \mu_2, \sigma_2^2)}.$$

- **K-component mixture of normals**

$$p(x) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \sigma_k^2)$$

- Multi-class indicators: $I_i = k$ if x_i comes from component k.
- **Gibbs sampling**
  - $(\pi_1, ..., \pi_K) \mid \mathbf{I}, \mathbf{x} \sim Dirichlet(\alpha_1 + n_1, \alpha_2 + n_2, ..., \alpha_K + n_K)$
  - $\sigma_k^2 \mid \mathbf{I}, \mathbf{x} \sim Inv\text{-}\chi^2$ and $\mu_k | \mathbf{I}, \sigma_k^2, \mathbf{x} \sim Normal, \quad for\ k = 1, ..., K,$
  - $I_i \mid \pi, \mu, \sigma^2, \mathbf{x} \sim Multinomial(\theta_{i1}, ..., \theta_{iK}), for\ i = 1, ..., n,$

$$\theta_{ij} = \frac{\pi_j \phi(x_i; \mu_j, \sigma_j^2)}{\sum_{r=1}^{k} \pi_r \phi(x_i; \mu_r, \sigma_r^2)}.$$

- Gibbs sampling is very powerful for **missing data** problems.
- **Semi-supervised learning**.

- **Probit regression**:

$$\Pr(y_i = 1 \mid x_i) = \Phi(x_i^T \beta)$$

- **Random utility formulation**:

$$\begin{aligned} u_i &\sim N(x_i^T \beta, 1) \\ y_i &= \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{if } u_i \leq 0 \end{cases}. \end{aligned}$$

- Check: $\Pr(y_i = 1 \mid x_i) = \Pr(u_i > 0) = 1 - \Pr(u_i \leq 0) = 1 - \Pr(u_i - x_i^T \beta < -x_i^T \beta) = 1 - \Phi(-x_i^T \beta) = \Phi(x_i^T \beta)$.
- Given $u = (u_1, \ldots, u_n)$, β can be analyzed by linear regression.
- u is **not observed**. Gibbs sampling to the rescue![2]

[2] Albert and Chib (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *JASA*.

- Simulate from **joint posterior** $p(u, \beta|y)$ by iterating between
  - $p(\beta|u, y)$ is multivariate normal (linear regression)
  - $p(u_i|\beta, y), \ i = 1, ..., n.$
- The full conditional posterior distribution of u_i

$$
\begin{aligned}
p(u_i|\beta, y) &\propto p(y_i|\beta, u_i)p(u_i|\beta) \\
&= \begin{cases} N(u_i|x_i'\beta, 1) & \text{truncated to } u_i \in (-\infty, 0] \text{ if } y_i = 0 \\ N(u_i|x_i'\beta, 1) & \text{truncated to } u_i \in (0, \infty) \text{ if } y_i = 1 \end{cases}
\end{aligned}
$$

- A histogram of β-draws approximates $p(\beta|y) = \int p(u, \beta|y)du$.

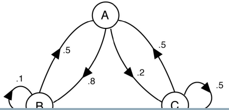
- Let $\mathcal{S} = \{s_1, s_2, ..., s_k)$ be a finite set of **states**.
  - Weather: $\mathcal{S} = \{\text{sunny}, \ \text{rain}\}$.
  - School grades: $\mathcal{S} = \{A, B, C, D, E, F\}$

- **Markov chain** is a stochastic process $\{X_t\}_{t=1}^{T}$ with random **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

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{pmatrix}$$



- ***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**

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

- Limiting long-run distribution

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

- **Stationary distribution**

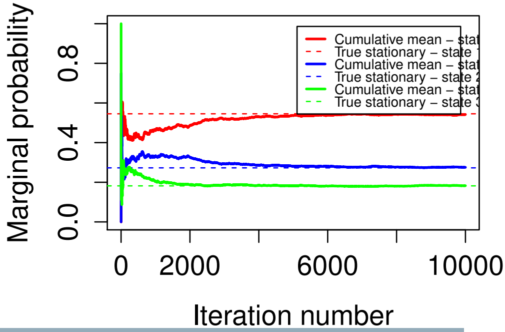
$$\pi = \pi P$$

- Example:

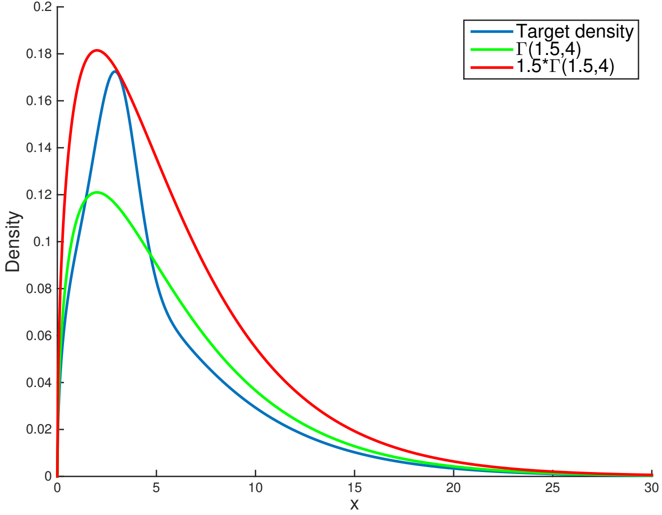
$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

$$\pi = (0.545, 0.272, 0.181)$$

- 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



- **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.

- 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 is enough**

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

- Common choices of Σ in proposal $N\left(\theta^{(i-1)}, c \cdot \Sigma\right)$:
  - $\Sigma = I$ (proposes 'off the cigar')
  - $\Sigma = J^{-1}_{\hat{\theta},\mathbf{y}}$ (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**.

- Generalization when the proposal density is not symmetric.

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

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.

- **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

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

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

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

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

- **Inefficiency factor**

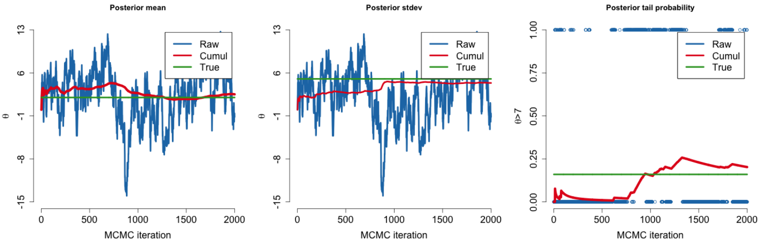
$$\mathrm{IF} = 1 + 2\sum_{k=1}^{\infty}\rho_k$$

- **Effective sample size** from MCMC

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

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