Chapter 5 Markov Chain Monte Carlo

5.1 Markov Chains

- **Definition:** Consider a sequence of random variables $X^{(0)}, X^{(1)}, \cdots$. The sequence (or stochastic process) is called a *Markov chain* if

$$P\big(X^{(t+1)} = y \,|\, X^{(t)} = x, X^{(t-1)} = x^{(t-1)}, \cdots, X^{(0)} = x^{(0)}\big)$$
$$= P\big(X^{(t+1)} = y \,|\, X^{(t)} = x\big) \qquad \text{for all } t.$$

where $P(\cdot \,|\, \cdot)$ denotes the conditional PDF/PMF.

- **Remarks:**
  - In a Markov chain, the conditional distribution of the *future state* $X^{(t+1)}$ given the *past states* $X^{(0)}, \cdots, X^{(t-1)}$ and the *present state* $X^{(t)}$ **only depends on the present state**.
  - In a Markov chain, given $X^{(t)}$, the future state $X^{(t+1)}$ and the past states $X^{(0:t-1)} := (X^{(0)}, \cdots, X^{(t-1)})$ are independent, that is,

$$P\big(X^{(t+1)} = y, X^{(0:t-1)} = x^{(0:t-1)} \,|\, X^{(t)} = x\big)$$
$$= P\big(X^{(t+1)} = y \,|\, X^{(t)} = x\big) \cdot P\big(X^{(0:t-1)} = x^{(0:t-1)} \,|\, X^{(t)} = x\big).$$

5.1 Markov Chains

- – If $X^{(t)}$, $t = 0, 1, \cdots$, are discrete random variables, $\{X^{(t)}\}$ is called a discrete-state Markov chain; if $X^{(t)}$, $t = 0, 1, \cdots$, are continuous, $\{X^{(t)}\}$ is called a continuous-state Markov chain.

  – For a Markov chain, we have

$$(1)\ P\big(X^{(t+1)} = x^{(t+1)} \,|\, X^{(t)} = x^{(t)}, X^{(t-1)} = x^{(t-1)}\big) = P\big(X^{(t+1)} = x^{(t+1)} \,|\, X^{(t)} = x^{(t)}\big);$$

$$(2)\ P\big(X^{(t+1:t+s)} = x^{(t+1:t+s)} \,|\, X^{(0:t)} = x^{(0:t)}\big) = P\big(X^{(t+1:t+s)} = x^{(t+1:t+s)} \,|\, X^{(t)} = x^{(t)}\big);$$

$$\begin{aligned}(3)\ P\big(X^{(t+1)} = x^{(t+1)} \,|\, X^{(0:t)} = x^{(0:t)}, X^{(t+2:t+s)} = x^{(t+2:t+s)}\big)& \\ = P\big(X^{(t+1)} = x^{(t+1)} \,|\, X^{(t)} = x^{(t)}, X^{(t+2)} = x^{(t+2)}\big)&.\end{aligned}$$

  – For any $k > l$ and $t_k > \cdots > t_{l+1} > t_l > t_{l-1} > \cdots > t_1 \geq 0$, we can show that

$$\begin{aligned}P\big(X^{(t_l)} = x^{(t_l)} \,|\, X^{(t_k)} = x^{(t_k)}, \cdots, X^{(t_{l+1})} = x^{(t_{l+1})}, X^{(t_{l-1})} = x^{(t_{l-1})}, \cdots, X^{(t_1)} = x^{(t_1)}\big)& \\ = P\big(X^{(t_l)} = x^{(t_l)} \,|\, X^{(t_{l+1})} = x^{(t_{l+1})}, X^{(t_{l-1})} = x^{(t_{l-1})}\big)&.\end{aligned}$$

5.1 Markov Chains

- – A Markov chain is called *homogeneous* if

$$P\big(X^{(t+1)} = y \,|\, X^{(t)} = x\big) = P\big(X^{(t)} = y \,|\, X^{(t-1)} = x\big)$$
$$= \cdots = P\big(X^{(1)} = y \,|\, X^{(0)} = x\big).$$

  – **We focus on homogeneous Markov chain in the following.**

  – $T(x,y) := P\big(X^{(t+1)} = y \,|\, X^{(t)} = x\big)$ is called the one-step *transition probability* (or *transition kernel*) of a homogeneous Markov chain.

  – Suppose that $X^{(t)} \in \{0, 1, 2, \cdots\}$. The matrix

$$\mathbb{T} = \begin{pmatrix} T(0,0) & T(0,1) & T(0,2) & \cdots \\ T(1,0) & T(1,1) & T(1,2) & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ T(i,0) & T(i,1) & T(i,2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

  is called the one-step *transition matrix*.

5.1 Markov Chains

- – Obviously, we have $T(x,y) \geq 0$ and $\int T(x,y)\,dy = 1$.

  – For simplicity, we also use notations $p(x^{(0:t)})$ and $p(x^{(t+1)} \mid x^{(0:t)})$ to denote

$$P\big(X^{(0:t)} = x^{(0:t)}\big) \quad \text{and} \quad P\big(X^{(t+1)} = x^{(t+1)} \mid X^{(0:t)} = x^{(0:t)}\big),$$

  respectively.

  – For a homogeneous Markov chain, the joint distribution of $X^{(0:t)}$ is

$$\begin{aligned} p(x^{(0:t)}) &= p(x^{(0)})\, p(x^{(1)}|x^{(0)})\, p(x^{(2)}|x^{(0:1)}) \cdots p(x^{(t)}|x^{(0:t-1)}) \\ &= p(x^{(0)})\, p(x^{(1)}|x^{(0)})\, p(x^{(2)}|x^{(1)}) \cdots p(x^{(t)}|x^{(t-1)}) \\ &= p(x^{(0)})\, T(x^{(0)}, x^{(1)})\, T(x^{(1)}, x^{(2)}) \cdots T(x^{(t-1)}, x^{(t)}). \end{aligned}$$

5.1 Markov Chains

- **Example:** Consider daily precipitation outcomes in San Francisco. The following table gives the rainfall status for 1814 pairs of consecutive days. A day is considered to be wet if more than 0.01 inch of precipitation is recorded and dry otherwise.

	Wet Today	Dry Today	Total
Wet Yesterday	418	256	674
Dry Yesterday	256	884	1140
	674	1140	1814

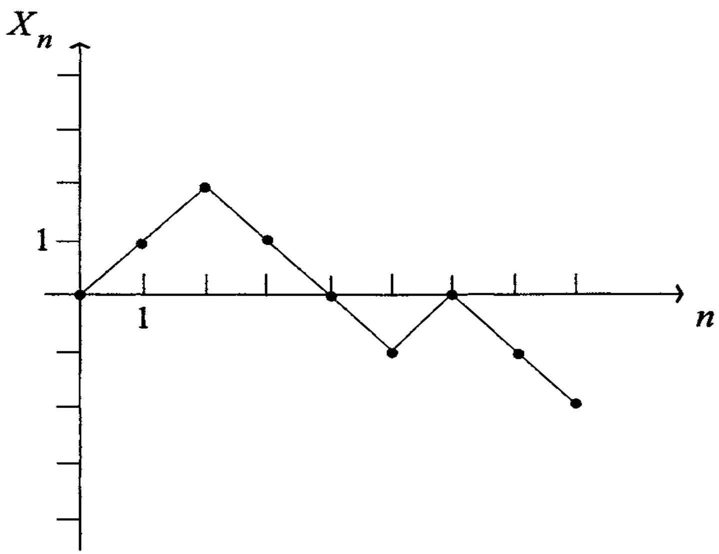
  - We use $X^{(t)}$ to describe the rainfall status for day t, where $X^{(t)} \in \{0, 1\}$, and 0 and 1 denote wet day and dry day, respectively.
  - **We assume** $\{X^{(t)}, t = 0, 1, \cdots\}$ **is a Markov chain**. From the data, we can estimate the one-step transition matrix as $\hat{\mathbb{T}} = \begin{pmatrix} 0.620 & 0.380 \\ 0.224 & 0.775 \end{pmatrix}$.

5.1 Markov Chains

- **Example: Random Walk.** Consider a particle that, at time 0, is at the origin. At each time unit, a coin is tossed. If "tail" (respectively, "head") is obtained, the particle moves one unit to the top (resp., bottom). Let $X^{(t)}$, $t = 0, 1, \cdots$, be the position of the particle after t tosses of the coin.
  - Note that $X^{(t)} \in \{0, \pm 1, \pm 2, \cdots\}$.
  - We have

$$
\begin{aligned}
& P\big(X^{(t+1)} = j \,|\, X^{(t)} = i, X^{(0:t-1)} = x^{(0:t-1)}\big) \\
= \; & P\big(X^{(t+1)} - X^{(t)} = j - i \,|\, X^{(t)} = i, X^{(0:t-1)} = x^{(0:t-1)}\big) \\
= \; & P\big(X^{(t+1)} - X^{(t)} = j - i\big) \\
= \; & \begin{cases} 0.5, & \text{if } j - i = \pm 1; \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

- – Hence, $\{X^{(t)}, t = 0, 1, \cdots\}$ is a Markov chain with one-step transition probabilities (or conditional probabilities) $T(i, j) = 0.5$ if $j = i \pm 1$ and $T(i, j) = 0$ otherwise.



- **Example: Markovian State Space Model:** A state space model is (homogeneous) Markovian if

$$p(x_t \,|\, x_{0:t-1}, y_{1:t-1}) = g(x_t \,|\, x_{t-1}) \quad \text{and} \quad p(y_t \,|\, x_{0:t}, y_{1:t-1}) = \zeta(y_t \,|\, x_t).$$

  – Note that

$$\begin{aligned} p(x_{t+1}, y_{t+1} \,|\, x_{0:t}, y_{0:t}) &= p(x_{t+1} \,|\, x_{0:t}, y_{0:t}) \cdot p(y_{t+1} \,|\, x_{0:t}, x_{t+1}, y_{0:t}) \\ &= g(x_{t+1} \,|\, x_t)\zeta(y_{t+1} \,|\, x_{t+1}) \\ &= p(x_{t+1}, y_{t+1} \,|\, x_t, y_t). \end{aligned}$$

  The model is a two-dimensional Markov chain if we let $X^{(t)} = (x_t, y_t)$.

5.1 Markov Chains

- **n-Step Transition Probabilities:** Consider a homogeneous Markov chain $\{X^{(t)}, t = 0, 1, 2, \cdots\}$, we can show that the n-step transition probabilities $P\big(X^{(t+n)} = y \mid X^{(t)} = x\big)$ **do not** depend on t. We use $T^{(n)}(x, y)$ to denote it, that is

$$T^{(n)}(x, y) = P\big(X^{(t+n)} = y \mid X^{(t)} = x\big).$$

- **Remarks:**
  - By definition, $T^{(1)}(x, y) = T(x, y)$.
  - Usually, $T^{(n)}(x, y) \neq [T(x, y)]^n$, where $T(x, y)$ is the one-step transition probability.
  - We have $T^{(n)}(x, y) \geq 0$ and $\int T^{(n)}(x, y)\, dy = 1$.

- **Theorem: Chapman-Kolmogorov Equations.** The n-step transition probabilities satisfy

$$T^{(n+m)}(x,y) = \int T^{(n)}(x,z)T^{(m)}(z,y)\,dz$$

for all $n, m \geq 0$ and all x, y.

- **Proof.** We have

$$\begin{aligned}
T^{(n+m)}(x,y) &= P\big(X^{(n+m)} = y \,|\, X^{(0)} = x\big) \\
&= \int P\big(X^{(n+m)} = y, X^{(n)} = z \,|\, X^{(0)} = x\big)\,dz \\
&= \int P\big(X^{(n)} = z \,|\, X^{(0)} = x\big) \cdot P\big(X^{(n+m)} = y \,|\, X^{(n)} = z, X^{(0)} = x\big)\,dz \\
&= \int P\big(X^{(n)} = z \,|\, X^{(0)} = x\big) \cdot P\big(X^{(n+m)} = y \,|\, X^{(n)} = z\big)\,dz \\
&= \int T^{(n)}(x,z)T^{(m)}(z,y)\,dz.
\end{aligned}$$

- **Remarks:** Suppose that $X^{(t)} \in \{0, 1, 2, \cdots\}$.
  - We use $\mathbb{T}^{(n)}$ to denote the n-step transition matrix, that is,

$$\mathbb{T}^{(n)} = \begin{pmatrix} T^{(n)}(0,0) & T^{(n)}(0,1) & T^{(n)}(0,2) & \cdots \\ T^{(n)}(1,0) & T^{(n)}(1,1) & T^{(n)}(1,2) & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ T^{(n)}(i,0) & T^{(n)}(i,1) & T^{(n)}(i,2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

  - The Chapman-Kolmogorov equation becomes

$$T^{(n+m)}(i,j) = \sum_k T^{(n)}(i,k)\, T^{(m)}(k,j)$$

    which is equivalent to $\mathbb{T}^{(n+m)} = \mathbb{T}^{(n)} \cdot \mathbb{T}^{(m)}$.
  - From the Chapman-Kolmogorov equations, we also have that

$$\mathbb{T}^{(n)} = \mathbb{T}^{(n-1)} \cdot \mathbb{T} = \left(\mathbb{T}^{(n-2)} \cdot \mathbb{T}\right) \cdot \mathbb{T} = \cdots = (\mathbb{T})^n .$$

- **Marginal Distribution of $X^{(t)}$:** Let $g_0(x) := P(X^{(0)} = x)$ be the distribution of $X^{(0)}$. Then we have

$$\begin{aligned} P\big(X^{(0)} = x, X^{(t)} = y\big) &= P\big(X^{(0)} = x\big)P\big(X^{(t)} = y \,|\, X^{(0)} = x\big) \\ &= g_0(x)T^{(t)}(x, y) \end{aligned}$$

  and

$$P\big(X^{(t)} = y\big) = \int g_0(x)T^{(t)}(x, y)\, dx.$$

- **Invariant Distribution:** A distribution $\kappa(x)$ (PDF/PMF) is called the *invariant distribution* (or *stationary distirbution*) of a homogeneous Markov chain $\{X^{(t)}, t = 0, 1, \cdots\}$ if

$$\int \kappa(x)T(x, y)\, dx = \kappa(y) \quad \text{for all } y.$$

5.1 Markov Chains

- **Remarks:**
  - Let $\kappa(x)$ be an invariant distribution. Then if $X^{(0)} \sim \kappa$, we have $X^{(1)} \sim \kappa$, $X^{(2)} \sim \kappa, \cdots$.
  - Suppose that $X^{(t)} \in \mathcal{S} = \{0, 1, 2, \cdots\}$.
    * Assume that the distribution of $X^{(0)}$ is $\boldsymbol{\nu} = (\nu_0, \nu_1, \cdots)$, that is, $P\big(X^{(0)} = i\big) = \nu_i$, $i = 0, 1, \cdots$, denoted by $X^{(0)} \sim \boldsymbol{\nu}$.
    * Here $\boldsymbol{\nu}$ is called a *distributional vector*, which satisfies $\nu_i \geq 0$ and $\sum_i \nu_i = 1$.
    * If $X^{(0)} \sim \boldsymbol{\nu}$, then $P\big(X^{(t)} = j\big) = \sum_i \nu_i T^{(t)}(i, j)$, which is equivalent to $X^{(t)} \sim \boldsymbol{\nu}\mathbb{T}^{(t)} = \boldsymbol{\nu} \cdot (\mathbb{T})^t$.
    * Distribution $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \cdots)$ is invariant for a discrete-state Markov chain if

$$\boldsymbol{\kappa}\mathbb{T} = \boldsymbol{\kappa}.$$

5.1 Markov Chains

- **Definition:** Let $\{X^{(t)}, t = 0, 1, \cdots\}$ be a discrete-state homogeneous Markov chain. Suppose that $X^{(t)} \in \mathcal{S} = \{0, 1, 2, \cdots\}$.
  - The chain is called *irreducible* if any state j can be reached from any state i in a finite number of steps. In other words, for each $i, j \in \mathcal{S}$ there must exist $t > 0$ such that $P\big(X^{(t)} = j \mid X^{(0)} = i\big) > 0$.
  - A state $i \in \mathcal{S}$ is said to have *period d*, where d is the *greatest common divisor* of the set $\big\{t : T_{i,i}^{(t)} > 0\big\}$. If every state in $\mathcal{S}$ has period 1, then the Markov chain is called *aperiodic*.
  - A state $i \in \mathcal{S}$ is called *positive recurrent* if starting in state i, the expected number of steps that the Markov chain returns to i is finite.
  - The chain is called **ergodic** if it is irreducible, aperiodic, and all states in $\mathcal{S}$ are positive recurrent.
  - The definitions can be extended to continuous-state Markov chains.

5.1 Markov Chains

- **Theorem:** Let $\{X^{(t)}, t = 0, 1, \cdots\}$ be an ergodic (discrete-state or continuous state) Markov chain. Then the chain has a **unique** invaraint distribution $\kappa(x)$. Starting from **any initial distribution**, $X^{(t)}$ converges in distribution to κ as $t \to \infty$, denoted by

$$X^{(t)} \xrightarrow{d} \kappa,$$

and for any function $h(\cdot)$ with finite expectation, we have

$$\frac{1}{m}\sum_{t=1}^{m} h(X^{(t)}) \xrightarrow{a.s.} E_{\kappa}\big[h(X)\big] = \int h(x)\kappa(x)\, dx.$$

- **Remarks:**
  - The theorem is a generalization of the strong law of large numbers (SLL-N) for i.i.d. samples. When t is large, we know that $X^{(t)}$ approximately follows the distribution κ (although $X^{(t)}$, $t = 1, 2, \cdots$, are not independent.)

- – *Example:* Consider a three-state Markov chain $X^{(t)} \in \{0, 1, 2\}$ with the transition matrix
$$\mathbb{T} = \begin{pmatrix} 3/4 & 1/4 & 0 \\ 1/8 & 2/3 & 5/24 \\ 0 & 1/6 & 5/6 \end{pmatrix}.$$
  * The two step transition matrix is
$$\mathbb{T}^{(2)} = \begin{pmatrix} 19/32 & 17/48 & 5/96 \\ 17/96 & 49/96 & 5/16 \\ 1/48 & 1/4 & 35/48 \end{pmatrix}.$$
  * Let $n = 50$, the n-step transition matrix is
$$\mathbb{T}^{(50)} = \begin{pmatrix} 0.182 & 0.364 & 0.454 \\ 0.182 & 0.364 & 0.454 \\ 0.182 & 0.364 & 0.454 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 2/11 & 4/11 & 5/11 \end{pmatrix}.$$

- – Suppose that $X^{(t)} \in \mathcal{S} = \{0, 1, 2, \cdots\}$. If the chain is ergodic, we can prove that

$$\lim_{t\to\infty} \mathbb{T}^{(t)} = \lim_{t\to\infty} \begin{pmatrix} T^{(t)}(0,0) & T^{(t)}(0,1) & T^{(t)}(0,2) & \cdots \\ T^{(t)}(1,0) & T^{(t)}(1,1) & T^{(t)}(1,2) & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ T^{(t)}(i,0) & T^{(t)}(i,1) & T^{(t)}(i,2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} \kappa_0 & \kappa_1 & \kappa_2 & \cdots & \\ \kappa_0 & \kappa_1 & \kappa_2 & \cdots & \\ \vdots & \vdots & \vdots & \cdots & \\ \kappa_0 & \kappa_1 & \kappa_2 & & \cdots \\ \vdots & \vdots & \vdots & \ddots & \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix} \begin{pmatrix} \kappa_0 & \kappa_1 & \kappa_2 & \cdots \end{pmatrix}.$$

  – For any distribution $\boldsymbol{\nu}$, if $X^{(0)} \sim \boldsymbol{\nu}$,

$$X^{(t)} \sim \boldsymbol{\nu}\mathbb{T}^{(t)} \to \boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2, \cdots).$$

5.2 Metropolis-Hastings Algorithm

- **Markov Chain Monte Carlo (MCMC):** We want to generate random samples from a target distribution $f(x)$.
  - MCMC generates an ergodic Markov chain $X^{(0)}, X^{(1)}, X^{(2)}, \cdots$, for which the invariant distribution is $f(x)$.
  - We can estimate approximate $f(x)$ by $\hat{f}(x) = \frac{1}{m}\sum_{t=1}^{m} \delta(x - X^{(t)})$ and estimate $E_f\big[h(X)\big]$ by

$$\frac{1}{m}\sum_{t=1}^{m} h(X^{(t)}).$$

- **Detailed Balance Condition:** For a homogeneous Markov chain $\{X^{(t)}, t = 0, 1, \cdots\}$, we call a distribution $f(x)$ satisfies the *detailed balance condition* if

$$f(x)T(x,y) = f(y)T(y,x) \quad \text{for all } x, y.$$

- **Remarks:**
  - If $f(x)T(x,y) = f(y)T(y,x)$ for all x, y, we have
  $$\int f(x)T(x,y)\,dx = \int f(y)T(y,x)\,dx = f(y),$$
  and $f(x)$ is an invariant distribution of $\{X^{(t)}, t = 0, 1, \cdots\}$.
  - The detailed balance condition is a **sufficient condition** for invariant distribution.
  - The invariant distribution requires that
  $$\int \kappa(x)T(x,y)\,dx = \kappa(y) = \int \kappa(y)T(y,x)\,dx,$$
  which achieves "overall" balance.

5.2 Metropolis-Hastings Algorithm

- **Metropolis-Hastings (MH) Algorithm:** We want to generate random samples from a target distribution $f(x)$.
  - Assign an initial state $X^{(0)}$.
  - For $t = 1, 2, \cdots,$
    * Generate a candidate sample X^* from a *proposal distribution* $q(x \mid X^{(t-1)})$, where $q(\cdot \mid \cdot)$ is a conditional PDF/PMF.
    * Compute the Metropolis-Hastings ratio

$$R(X^{(t-1)}, X^*) := \frac{f(X^*)/q(X^* \mid X^{(t-1)})}{f(X^{(t-1)})/q(X^{(t-1)} \mid X^*)} = \frac{f(X^*) \cdot q(X^{(t-1)} \mid X^*)}{f(X^{(t-1)}) \cdot q(X^* \mid X^{(t-1)})}.$$

    * Generate U from the uniform$(0, 1)$ distribution and let

$$X^{(t)} = \begin{cases} X^*, & \text{if } U \leq R(X^{(t-1)}, X^*); \\ X^{(t-1)}, & \text{otherwise.} \end{cases}$$

5.2 Metropolis-Hastings Algorithm

- **Remarks:**

  – In the MH algorithm, each X^* is accepted as $X^{(t)}$ with probability

$$\min\left\{1, R(X^{(t-1)}, X^*)\right\} = \min\left\{1, \frac{f(X^*)\cdot q(X^{(t-1)}\,|\,X^*)}{f(X^{(t-1)})\cdot q(X^*\,|\,X^{(t-1)})}\right\}.$$

  – Obviously, $X^{(0)}, X^{(1)}, X^{(2)}, \cdots$ generated by the MH algorithm is a homogeneous Markov chain. (**Why?**)

  – The one-step transition probability is

$$\begin{aligned} T(x,y) &= P\big(X^{(t)} = y\,|\,X^{(t-1)} = x\big) \\ &= \begin{cases} q(y\,|\,x)\min\left\{1, \frac{f(y)\cdot q(x\,|\,y)}{f(x)\cdot q(y\,|\,x)}\right\}, & \text{if } y \neq x; \\ q(x\,|\,x) + \int_{z\neq x} q(z\,|\,x)\left[1 - \min\left\{1, \frac{f(z)\cdot q(x\,|\,z)}{f(x)\cdot q(z\,|\,x)}\right\}\right] dz, & \text{if } y = x. \end{cases} \end{aligned}$$

  – Detailed balance: when $y \neq x$,

$$f(x)T(x,y) = \min\left\{f(x)q(y\,|\,x), f(y)q(x\,|\,y)\right\} = f(y)T(y,x).$$

  So $f(x)$ is an **invariant distribution** of the generated Markov chain.

5.2 Metropolis-Hastings Algorithm

- **Example: Bayesian Inference.** Under the Bayesian setting, we assume that the parameter θ has a prior distribution with PDF $\pi(\theta)$. Given θ, the observed data Y follows a distribution with PDF $p(y \mid \theta)$.
  - The posterior distribution of θ is
$$p(\theta \mid Y = y) = \frac{p(\theta, Y = y)}{p(Y = y)} = \frac{\pi(\theta) p(y \mid \theta)}{\int \pi(\theta) p(y \mid \theta)\, d\theta}.$$
  - We can use the MH algorithm to generate a Markov chain $\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(m)}$, for which the invariant distribution is $p(\theta \mid Y = y)$.
  - Usually we **discard the first m_0 samples** that greatly depend on the initial sample, which is called the *burn-in period*. Then we approximate the posterior distribution $p(\theta \mid Y = y)$ by
$$\frac{1}{m - m_0} \sum_{t=m_0+1}^{m} \delta(\theta - \theta^{(t)})$$
and estimate $E\big[\theta \mid Y = y\big]$ by $\frac{1}{m-m_0} \sum_{t=m_0+1}^{m} \theta^{(t)}$.

5.2 Metropolis-Hastings Algorithm

- – **MH Algorithm for Bayesian Inference:**
  - ∗ Assign an initial state $\theta^{(0)}$.
  - ∗ For $t = 1, 2, \cdots,$
    - · Generate a candidate sample θ^* from a proposal distribution $q(\theta \mid \theta^{(t-1)})$.
    - · Accept θ^* as $\theta^{(t)}$ with probability

$$\begin{aligned} &\min\left\{1, \frac{p(\theta^* \mid Y = y) \cdot q(\theta^{(t-1)} \mid \theta^*)}{p(\theta^{(t-1)} \mid Y = y) \cdot q(\theta^* \mid \theta^{(t-1)})}\right\} \\ = \; &\min\left\{1, \frac{\pi(\theta^*) p(y \mid \theta^*) \cdot q(\theta^{(t-1)} \mid \theta^*)}{\pi(\theta^{(t-1)}) p(y \mid \theta^{(t-1)}) \cdot q(\theta^* \mid \theta^{(t-1)})}\right\}. \end{aligned}$$

    - · If θ^* is rejected, let $\theta^{(t)} = \theta^{(t-1)}$.

5.2 Metropolis-Hastings Algorithm

- **Random Walk Proposal:** At time t, we generate ϵ_t from a given distribution with PDF/PMF $g(\epsilon)$, and let

$$X^* = X^{(t-1)} + \epsilon_t.$$

Then the proposal distribution is

$$q(x \mid X^{(t-1)}) = g\big(x - X^{(t-1)}\big)$$

and X^* is accepted as $X^{(t)}$ with probability

$$\min\left\{1, \frac{f(X^*) \cdot g\big(X^{(t-1)} - X^*\big)}{f(X^{(t-1)}) \cdot g\big(X^* - X^{(t-1)}\big)}\right\}.$$

- **Remarks:**

  - When $g(\epsilon)$ is symmetric around 0, X^* is accepted with probability $\min\left\{1, \frac{f(X^*)}{f(X^{(t-1)})}\right\}$.

5.2 Metropolis-Hastings Algorithm

- – When the target distribution $f(x)$ is continuous, we often let $\epsilon_t \sim N(0, \sigma^2)$ with a given σ^2.

  – We need to choose an appropriate σ when using random walk proposal. If σ is too large, the acceptance probability could be very low; if σ is too small, the acceptance probability is high, but the chain moves very slow.

  – Sometimes, we may let

$$X^* = X^{(t-1)} + a \cdot \nabla \log f(X^{(t-1)}) + \epsilon_t,$$

  where $\epsilon_t \sim N(0, \sigma^2)$, and a is a given constant. We will expect $f(X^*)$ to be larger than $f(X^{(t-1)})$. The acceptance probability becomes

$$\min\left\{1, \frac{f(X^*) \cdot g\big(X^{(t-1)} - X^* - a \cdot \nabla \log f(X^*)\big)}{f(X^{(t-1)}) \cdot g\big(X^* - X^{(t-1)} - a \cdot \nabla \log f(X^{(t-1)})\big)}\right\}.$$

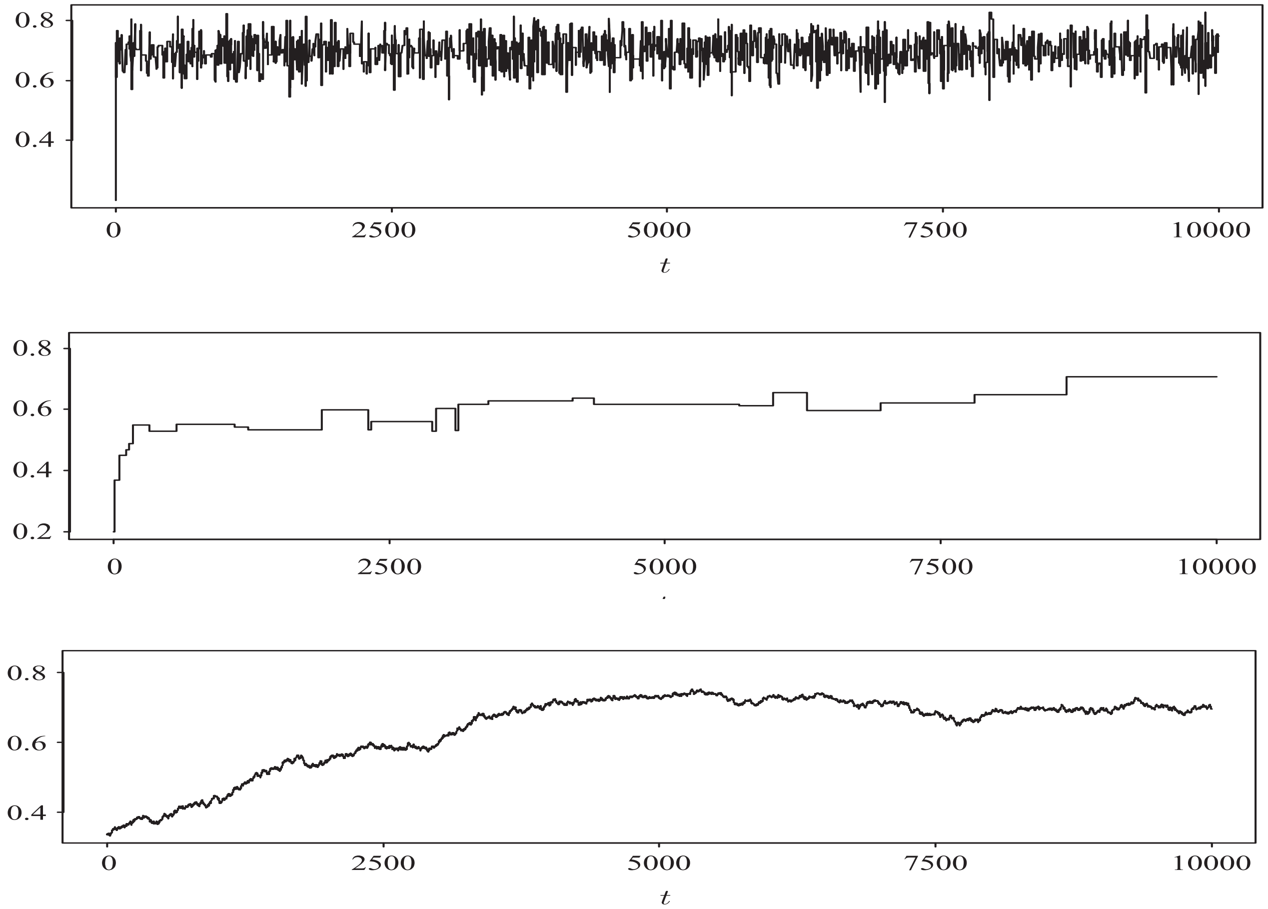


Figure: Traceplots for random walk proposals using appropriate σ(upper panel), large σ(middle panel), and small σ(lower panel).

- **Independent Proposal:** At time t, we generate X^* from a proposal distribution $q(x \mid X^{(t-1)}) = g(x)$, which does not dependent on $X^{(t-1)}$. Then X^* is accepted as $X^{(t)}$ with probability

$$\min\left\{1, \frac{f(X^*) \cdot g\big(X^{(t-1)}\big)}{f(X^{(t-1)}) \cdot g\big(X^*\big)}\right\} = \min\left\{1, \frac{f(X^*)/g\big(X^*\big)}{f(X^{(t-1)})/g\big(X^{(t-1)}\big)}\right\}.$$

- **Remarks:**
  - It requires that the support of the proposal g covers the support of the target distribution f, otherwise the Markov chain may not be irreducible.
  - The proposal g should have a fatter tail than the target distribution f, otherwise $f(X^{(t-1)})/g\big(X^{(t-1)}\big)$ could be very large and the chain will tend to get stuck at $X^{(t-1)}$ for long periods.
  - Usually we only use independent proposal to update a sub-vector of X.

5.3 Gibbs Sampling

- **Gibbs Sampling:** We want to generate random samples from a target distribution $f(x)$, where $x = x_{1:n} = (x_1, \cdots, x_n)$ is a high dimensional vector. We generate Markov chain $X_{1:n}^{(0)}, X_{1:n}^{(1)}, X_{1:n}^{(2)}, \cdots$ as follows.
  - Assign an initial state $X_{1:n}^{(0)}$.
  - For $t = 1, 2, \cdots$, generate $X_{1:n}^{(t)} = (X_1^{(t)}, X_2^{(t)}, \cdots, X_n^{(t)})$ sequentially as follows.
    * Generate $X_1^{(t)}$ from $f\big(x_1 \,|\, X_2^{(t-1)}, \cdots, X_n^{(t-1)}\big)$;
    * Generate $X_2^{(t)}$ from $f\big(x_2 \,|\, X_1^{(t)}, X_3^{(t-1)}, \cdots, X_n^{(t-1)}\big)$;
    * $\cdots\cdots$;
    * Generate $X_n^{(t)}$ from $f\big(x_n \,|\, X_1^{(t)}, X_2^{(t)}, \cdots, X_{n-1}^{(t)}\big)$.

  Then we have, for any function h with finite expectation,

$$E_f\big[h(X_{1:n})\big] = \int h(x_{1:n}) f(x_{1:n})\, dx_{1:n} \approx \frac{1}{m - m_0} \sum_{t=m_0+1}^{m} h\big(X_{1:n}^{(t)}\big).$$

- **Remarks:**
  - Suppose that T_1 and T_2 are transition kernels, define
    $$T_1 \circ T_2(x, y) := \int T_1(x, z) T_2(z, y)\, dz.$$
    Then $T_1 \circ T_2$ is also a transition kernel, that is, $T_1 \circ T_2(x, y) \geq 0$ and $\int T_1 \circ T_2(x, y)\, dy = 1$. Similarly, if $T_1, T_2, \cdots, T_n$ are transition kernels, then $T_1 \circ T_2 \circ \cdots \circ T_n$ is also a transition kernel.
  - If a distribution $f(x)$ is invariant for both T_1 and T_2, then
    $$\begin{aligned}\int f(x) \cdot T_1 \circ T_2(x, y)\, dx &= \int f(x) \left[\int T_1(x, z) T_2(z, y)\, dz \right] dx \\ &= \int \left[\int f(x) T_1(x, z)\, dx \right] T_2(z, y)\, dz \\ &= \int f(z) T_2(z, y)\, dz = f(y),\end{aligned}$$
    so $f(x)$ is invariant for $T_1 \circ T_2$.

5.3 Gibbs Sampling

- – Similarly, if a distribution $f(x)$ is invariant for $T_1, T_2, \cdots, T_n$, then $f(x)$ is also invariant for $T_1 \circ T_2 \circ \cdots \circ T_n$.

  – For the Gibbs sampling algorithm, define transition kernel

  $$T_i(x_{1:n}, y_{1:n}) = f(y_i \,|\, y_{1:i-1}, y_{i+1:n}) \cdot I(y_{1:i-1} = x_{1:i-1}) \cdot I(y_{i+1:n} = x_{i+1:n}),$$

  where $I(\cdot)$ is the indicator function.

  – For each T_i, we have

  $$\begin{aligned} & f(x_{1:n})T_i(x_{1:n}, y_{1:n}) \\ = \; & f(x_{1:n})f(y_i \,|\, y_{1:i-1}, y_{i+1:n}) \cdot I(y_{1:i-1} = x_{1:i-1}) \cdot I(y_{i+1:n} = x_{i+1:n}) \\ = \; & f(x_{1:n}) \cdot \frac{f(y_{1:n})}{f(y_{1:i-1}, y_{i+1:n})} \cdot I(y_{1:i-1} = x_{1:i-1}) \cdot I(y_{i+1:n} = x_{i+1:n}) \\ = \; & f(y_{1:n})f(x_i \,|\, x_{1:i-1}, x_{i+1:n}) \cdot I(y_{1:i-1} = x_{1:i-1}) \cdot I(y_{i+1:n} = x_{i+1:n}) \\ = \; & f(y_{1:n})T_i(y_{1:n}, x_{1:n}), \end{aligned}$$

  the target distribution $f(x_{1:n})$ satisfies the details balance condition for T_i. Hence, $f(x_{1:n})$ is invariant for T_i.

5.3 Gibbs Sampling

- – The one-step transition kernel for the Gibbs sampling algorithm is

$$\begin{aligned} T(x_{1:n}, y_{1:n}) &= P\Big(X_{1:n}^{(t)} = y_{1:n} \,|\, X_{1:n}^{(t-1)} = x_{1:n}\Big) \\ &= T_1 \circ T_2 \circ \cdots \circ T_n(x_{1:n}, y_{1:n}). \end{aligned}$$

  So the target distribution $f(x_{1:n})$ is invariant for $T = T_1 \circ T_2 \circ \cdots \circ T_n$.

  – When $n = 1$, the Gibbs sampling algorithm proposes to draw $x^{(t)}$ from the target distribution $f(x^{(t)})$.

  – In the MH algorithm, if we let the proposal distribution $q(x \,|\, X^{(t-1)}) = f(x)$, the generated X^* will be accepted as $X^{(t)}$ with probability

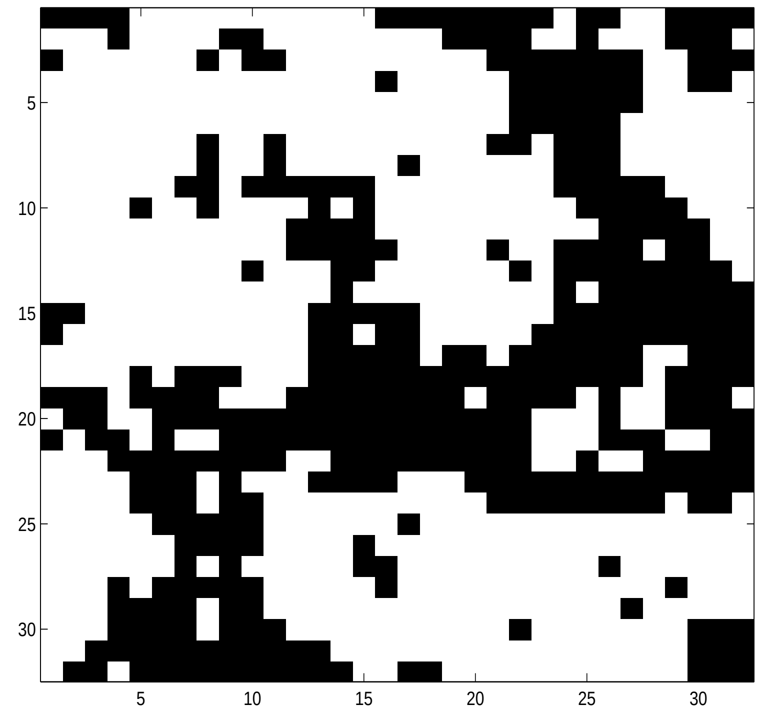
$$\min\left\{1, \frac{f(X^*) \cdot f\big(X^{(t-1)}\big)}{f(X^{(t-1)}) \cdot f\big(X^*\big)}\right\} = 1.$$

  – In the Gibbs sampling algorithm, we can also use the Metropolis-Hastings method to update X_i, it is called the *MH-within-Gibbs* sampler.

5.3 Gibbs Sampling

- **MH-within-Gibbs Sampling:** We want to generate random samples from a target distribution $f(x_{1:n})$.
  - Assign an initial state $X_{1:n}^{(0)}$.
  - For $t = 1, 2, \cdots$, generate $X_{1:n}^{(t)} = (X_1^{(t)}, X_2^{(t)}, \cdots, X_n^{(t)})$ sequentially as follows.
    - Generate X_1^* from a proposal distribution $q\big(x_1 \,|\, X_1^{(t-1)}, X_{2:n}^{(t-1)}\big)$. Accept X_1^* as $X_1^{(t)}$ with probability $\min\left\{1, \frac{f(X_1^* \,|\, X_{2:n}^{(t-1)}) q(X_1^{(t-1)} \,|\, X_1^*, X_{2:n}^{(t-1)})}{f(X_1^{(t-1)} \,|\, X_{2:n}^{(t-1)}) q(X_1^* \,|\, X^{(t-1)}, X_{2:n}^{(t-1)})}\right\}$. If X_1^* is rejected, let $X_1^{(t)} = X_1^{(t-1)}$.
    - $\vdots$
    - Generate X_n^* from a proposal distribution $q\big(x_n \,|\, X_{1:n-1}^{(t)}, X_n^{(t-1)}\big)$. Accept X_n^* as $X_n^{(t)}$ with probability $\min\left\{1, \frac{f(X_n^* \,|\, X_{1:n-1}^{(t)}) q(X_n^{(t-1)} \,|\, X_{1:n-1}^{(t)}, X_n^*)}{f(X_n^{(t-1)} \,|\, X_{1:n-1}^{(t)}) q\big(X_n^* \,|\, X_{1:n-1}^{(t)}, X_n^{(t-1)}\big)}\right\}$. If X_1^* is rejected, let $X_n^{(t)} = X_n^{(t-1)}$.

- **Example: 2D Ising Model.** In a magnet field, the atomic spins on a $N \times N$ lattice space, $\mathcal{L} = \{(i,j) : i,j = 1, \cdots, N\}$, can be represented by a random matrix $\boldsymbol{X} = \{X_{i,j}\}_{N \times N}$. Each $X_{i,j}$ is either 1 or -1.



5.3 Gibbs Sampling

- – The random matrix X follows a distribution with the form

  $$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{S} e^{-U(\boldsymbol{x})/kT},$$

  where $\boldsymbol{x} = \{x_{i,j}\}_{N\times N}$, k is the Boltzmann constant, T is the temperature, $S = \sum_{\boldsymbol{x}} e^{-U(\boldsymbol{x})/kT}$ is the normalizing constant.

  – The potential function is

  $$U(\boldsymbol{x}) = -J \sum_{(i,j)\sim(i',j')} x_{i,j}\, x_{i',j'} + \sum_{i,j} h_{i,j}\, x_{i,j},$$

  where the symbol $(i,j) \sim (i',j')$ means that the two sites are neighbors, J is called the *interaction strength*, $\{h_{i,j}\}_{N\times N}$ is the magnetic field.

  – We want the calculate the *internal energy*, which is defined as

  $$E\big[U(\boldsymbol{X})\big] = \sum_{\boldsymbol{x}} U(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}).$$

5.3 Gibbs Sampling

- – For simplicity, we write $\boldsymbol{X}$ as a vector $Z_{1:n} = (Z_1, \cdots, Z_n)$, where $n = N^2$. We use the Gibbs sampling algorithm to generate a Markov chain $Z_{1:n}^{(0)}, Z_{1:n}^{(1)}, Z_{1:n}^{(2)}, \cdots$ as follows.
  - ∗ Assign an initial state $Z_{1:n}^{(0)}$.
  - ∗ For $t = 1, 2, \cdots$, generate $Z_{1:n}^{(t)} = (Z_1^{(t)}, Z_2^{(t)}, \cdots, Z_n^{(t)})$ as follows.
    - · Generate $Z_1^{(t)} \in \{0, 1\}$ from

$$P\big(Z_1 = z_1 \,|\, Z_{2:n}^{(t-1)}\big) = \frac{P\big(Z_1 = z_1, Z_{2:n}^{(t-1)}\big)}{P\big(Z_1 = 1, Z_{2:n}^{(t-1)}\big) + P\big(Z_1 = -1, Z_{2:n}^{(t-1)}\big)};$$

    - · $\vdots$
    - · Generate $Z_n^{(t)} \in \{0, 1\}$ from

$$P\big(Z_n = z_n \,|\, Z_{1:n-1}^{(t)}\big) = \frac{P\big(Z_{1:n-1}^{(t)}, Z_n = z_n\big)}{P\big(Z_{1:n-1}^{(t)}, Z_n = 1\big) + P\big(Z_{1:n-1}^{(t)}, Z_n = -1\big)}.$$

  Then we can estimate $E\big(U(\boldsymbol{X})\big)$ by $\frac{1}{m-m_0}\sum_{t=m_0+1}^{m} U(Z_{1:n}^{(t)})$.

- **Example: Stochastic Volatility Model.** Let $y_t = \log(P_t/P_{t-1})$ be the observed log-return of a financial asset at time t. Consider the following state space model

$$\begin{aligned} \text{state equation} &: \quad \log \sigma_t^2 = b_0 + b_1 \log \sigma_{t-1}^2 + u_t, \\ \text{observation equation} &: \quad y_t \sim N(0, \sigma_t^2), \end{aligned}$$

  where $u_t \sim N(0, \delta^2)$. We also assume $\log \sigma_0^2 \sim N(\mu_0, \eta_0^2)$ with known μ_0 and η_0^2.

  – For simplicity, we let $z_t = \log \sigma_t^2$ and let $\theta = (b_0, b_1, \delta^2)$ be the model parameters.

  – Given the observations $y_{1:T}$, we want to estimate θ and $z_{0:T}$.

  – We estimate θ and $z_{0:T}$ under the Bayesian setting.

5.3 Gibbs Sampling

- – **Priors:** Assume the prior distribution of the vector $(b_0, b_1)'$ is a bivaraite normal distribution $N(\mu_b, \Sigma_b)$, and the prior distribution of δ^2 is an inverse-Gamma(α, β) distribution with density $p(\delta^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\delta^2\right)^{-\alpha-1} \exp\{-\beta/\delta^2\}$ for $\delta^2 > 0$. Here μ_b, Σ_b, α and β are given *hyperparameters*.

  – The **target distribution** is

$$
\begin{aligned}
p(z_{0:T}, \theta \,|\, y_{1;T}) &\propto p(z_{0:T}, y_{1:T}, b_0, b_1, \delta^2) \\
&= p(b_0, b_1)\, p(\delta^2)\, p(z_0) \prod_{t=1}^{T} p(z_t \,|\, z_{t-1}, b_0, b_1, \delta^2) p(y_t \,|\, z_t).
\end{aligned}
$$

  – Let $X = (z_{1:T}, \theta)'$. We generate a Markov chain $X^{(0)}, X^{(1)}, \cdots$ with invariant distribution $p(z_{0:T}, \theta \,|\, y_{1;T})$. Then we can use the generated samples $\{X^{(s)}\}_{s=m_0+1}^{m}$ to estimate θ and $z_{0:T}$.

5.3 Gibbs Sampling

- **MCMC for Stochastic Volatility Model:**
  - Assign an initial value $X^{(0)} = \left(z_{0:T}^{(0)}, \theta^{(0)}\right)$.
  - For $s = 1, 2, \cdots$:
    * **Gibbs Updating for z_0:** Generate $z_0^{(s)}$ from

$$\begin{aligned} p(z_0 \,|\, z_{1:T}^{(s-1)}, \theta^{(s-1)}, y_{1:T}) &= p(z_0 \,|\, z_1^{(s-1)}, \theta^{(s-1)}) \\ &\propto p(z_0 \,|\, \theta^{(s-1)}) p(z_1^{(s-1)} \,|\, z_0, \theta^{(s-1)}) \sim N\left(\mu_{z_0}, \Sigma_{z_0}\right), \end{aligned}$$

where

$$p(z_0 \,|\, \theta^{(s-1)}) = p(z_0) \sim N(\mu_0, \eta_0^2)$$

and

$$p(z_1^{(s-1)} \,|\, z_0, \theta^{(s-1)}) \sim N(b_0^{(s-1)} + b_1^{(s-1)} z_0, \delta^2).$$

(*Note: We can first calculate* $p(z_0, z_1^{(s-1)} \,|\, \theta^{(s-1)})$, *then find* $p(z_0 \,|\, z_1^{(s-1)}, \theta^{(s-1)})$.)

- – * **MH Updating for z_1:** Generate z_1^* from

$$\begin{aligned} q(z_1 \mid z_0^{(s)}, z_{1:T}^{(s-1)}, \theta^{(s-1)}, y_{1:T}) &= p(z_1 \mid z_0^{(s)}, z_{2:T}^{(s-1)}, \theta^{(s-1)}) \\ &= p(z_1 \mid z_0^{(s)}, z_2^{(s-1)}, \theta^{(s-1)}) \\ &\propto p(z_1 \mid z_0^{(s)}, \theta^{(s-1)})\, p(z_2^{(s-1)} \mid z_1, z_0^{(s)}, \theta^{(s-1)}) \\ &\propto p(z_1 \mid z_0^{(s)}, \theta^{(s-1)})\, p(z_2^{(s-1)} \mid z_1, \theta^{(s-1)}) \\ &\sim N\big(\mu_{z_1}, \Sigma_{z_1}\big). \end{aligned}$$

Accept z_1^* as $z_1^{(s)}$ with probability

$$\begin{aligned} &\min\left\{1, \frac{p(z_1^* \mid z_0^{(s)}, z_{2:T}^{(s-1)}, \theta^{(s-1)}, y_{1:T})/p(z_1^* \mid z_0^{(s)}, z_{2:T}^{(s-1)}, \theta^{(s-1)})}{p(z_1^{(s-1)} \mid z_0^{(s)}, z_{2:T}^{(s-1)}, \theta^{(s-1)}, y_{1:T})/p(z_1^{(s-1)} \mid z_0^{(s)}, z_{2:T}^{(s-1)}, \theta^{(s-1)})}\right\} \\ = \ &\min\left\{1, \frac{p(y_1 \mid z_1^*, \theta^{(s-1)})}{p(y_1 \mid z_1^{(s-1)}, \theta^{(s-1)})}\right\}, \end{aligned}$$

where $p(y_1 \mid z_1, \theta) \sim N(0, e^{z_1})$. If z_1^* is rejected, let $z_1^{(s)} = z_1^{(s-1)}$.

- – * **MH Updating for** $z_2, \cdots, z_n$**:** Generate $z_2^{(s)}, \cdots, z_T^{(s)}$ sequentially using the MH-within-Gibbs algorithm.
    * **Gibbs Updating for** (b_0, b_1)**:** Define $\boldsymbol{b} = (b_0, b_1)'$, generate $\boldsymbol{b}^{(s)} = (b_0^{(s)}, b_1^{(s)})'$ from

$$
\begin{aligned}
& p(\boldsymbol{b} \,|\, z_{0:T}^{(s)}, \theta^{(s-1)}, y_{1:T}) \\
= \;& p\big(\boldsymbol{b} \,|\, z_{0:T}^{(s)}, (\delta^2)^{(s-1)}\big) \propto p\big(z_{0:T}^{(s)}, \boldsymbol{b}, (\delta^2)^{(s-1)}\big) \\
\propto \;& p(\boldsymbol{b}) \prod_{t=1}^{T} p\big(z_t^{(s)} \,|\, z_{t-1}^{(s)}, \boldsymbol{b}, (\delta^2)^{(s-1)}\big) \\
\propto \;& \exp\Big\{ -\frac{1}{2}(\boldsymbol{b} - \mu_b)'\Sigma_b^{-1}(\boldsymbol{b} - \mu_b) - \frac{1}{2(\delta^2)^{(s-1)}} \sum_{t=1}^{T} \big[z_t^{(s)} - (1, z_{t-1}^{(s)})\boldsymbol{b}\big]^2 \Big\} \\
\propto \;& \exp\Big\{ -\frac{1}{2}\boldsymbol{b}'A\boldsymbol{b} + \boldsymbol{b}'d \Big\} \sim \mathbf{N}\big(\mathbf{A}^{-1}\mathbf{d}, \mathbf{A}^{-1}\big).
\end{aligned}
$$

where $A = \Sigma_b^{-1} + \frac{1}{(\delta^2)^{(s-1)}} \sum_{t=1}^{T}(1, z_{t-1}^{(s)})'(1, z_{t-1}^{(s)})$ and $d = \Sigma_b^{-1}\mu_b + \frac{1}{(\delta^2)^{(s-1)}} \sum_{t=1}^{T} z_t^{(s)}(1, z_{t-1}^{(s)})'$.

5.3 Gibbs Sampling

- – * **Gibbs Updating for δ^2:** Generate $(\delta^2)^{(s)}$ from

$$
\begin{aligned}
& p(\delta^2 \mid z_{0:T}^{(s)}, b_0^{(s)}, b_1^{(s)}, y_{1:T}) \\
= \; & p(\delta^2 \mid z_{0:T}^{(s)}, b_0^{(s)}, b_1^{(s)}) \propto p\big(z_{0:T}^{(s)}, b_0^{(s)}, b_1^{(s)}, \delta^2\big) \\
\propto \; & p(\delta^2) \prod_{t=1}^{T} p\big(z_t^{(s)} \mid z_{t-1}^{(s)}, b_0^{(s)}, b_1^{(s)}, \delta^2\big) \\
\propto \; & \big(\delta^2\big)^{-\alpha-1} \exp\{-\beta/\delta^2\} \cdot \prod_{t=1}^{T} \frac{1}{\sqrt{\delta^2}} \exp\{-(z_t^{(s)} - b_0^{(s)} - b_1^{(s)} z_{t-1}^{(s)})^2/2\delta^2\} \\
= \; & \big(\delta^2\big)^{-\alpha-\frac{T}{2}-1} \exp\Big\{ -\frac{1}{\delta^2}\Big[\beta + \frac{1}{2}\sum_{t=1}^{T}(z_t^{(s)} - b_0^{(s)} - b_1^{(s)} z_{t-1}^{(s)})^2\Big]\Big\} \\
\sim \; & \text{inverse-Gamma}(\alpha^*, \beta^*),
\end{aligned}
$$

where $\alpha^* = \alpha + \frac{T}{2}$ and $\beta^* = \beta + \frac{1}{2}\sum_{t=1}^{T}(z_t^{(s)} - b_0^{(s)} - b_1^{(s)} z_{t-1}^{(s)})^2$.

- **Remarks:**

  – We can estimate the parameters by

$$\hat{b}_0 = \frac{1}{m - m_0} \sum_{s=m_0+1}^{m} b_0^{(s)}, \quad \hat{b}_1 = \frac{1}{m - m_0} \sum_{s=m_0+1}^{m} b_1^{(s)}, \quad \hat{\delta}^2 = \frac{1}{m - m_0} \sum_{s=m_0+1}^{m} (\delta^2)^{(s)},$$

  and estimate the latent states by

$$\hat{z}_t = \frac{1}{m - m_0} \sum_{s=m_0+1}^{m} z_t^{(s)}.$$

  – Compared with the particle filter, the MCMC algorithm is more convenient for parameter estimation, but it can not be used for online estimation. When a new observation y_{T+1} is received, we need rerun the MCMC algorithm.

5.4 Implementation

- Consider the Bayesian inference problem. Assume that the parameter $\theta_{1:p} = (\theta_1, \cdots, \theta_p)$ has a prior distribution with the PDF $\pi(\theta_{1;p})$. Given $\theta_{1;p}$, the observed data Y follows a distribution with the PDF $p(y \mid \theta_{1;p})$.
  - Our target distribution is the posterior distribution

$$p(\theta_{1:p} \mid Y = y) \;=\; \frac{p(\theta_{1:p}, Y = y)}{p(Y = y)} = \frac{\pi(\theta_{1:p})p(y \mid \theta_{1:p})}{\int \pi(\theta_{1:p})p(y \mid \theta_{1:p})\, d\theta_{1:p}}.$$

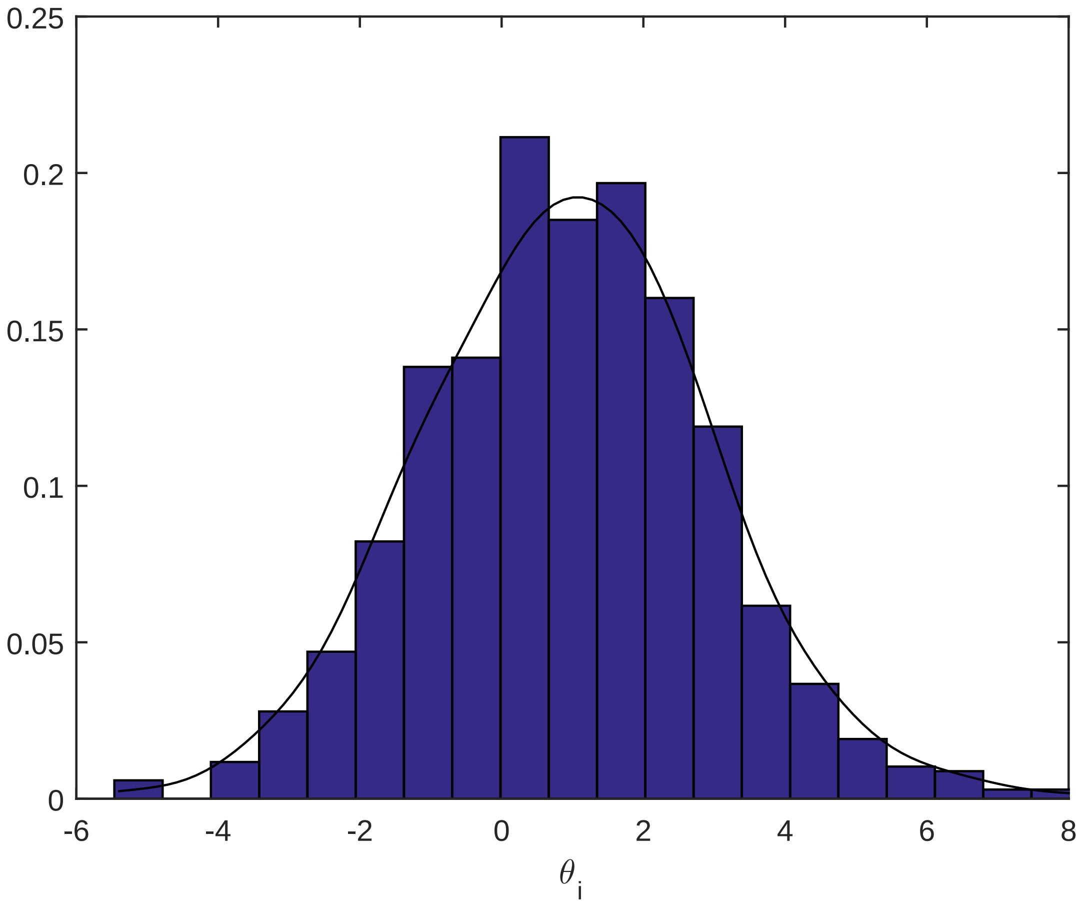
  - We focus on inference of θ_i. The marginal posterior distribution of θ_i is

$$p(\theta_i \mid Y = y) \;=\; \frac{\int \pi(\theta_{1:p})p(y \mid \theta_{1:p})\, d\theta_{1:i-1}\, d\theta_{i+1:p}}{\int \pi(\theta_{1:p})p(y \mid \theta_{1:p})\, d\theta_{1:p}},$$

    which does not have a closed-form expression in most cases.
  - Suppose we generated a Markov chain $\theta_{1:p}^{(0)}, \theta_{1:p}^{(1)}, \cdots, \theta_{1:p}^{(m)}$ whose invariant distribution is $p(\theta_{1:p} \mid Y = y)$. We can use the samples $\theta_i^{(m_0+1)}, \cdots, \theta_i^{(m)}$ to make inference of θ_i.

5.4 Implementation

- **Posterior Density:** We want to estimate $p(\theta_i \,|\, Y = y)$.
  - **Histogram Estimator:**



- – **Kernel Estimator:** Estimate $p(\theta_i \,|\, Y = y)$ by

$$\hat{p}(\theta_i \,|\, Y = y) = \frac{1}{m - m_0} \sum_{t=m_0+1}^{m} \frac{1}{h} K\left(\frac{\theta_i - \theta_i^{(t)}}{h}\right),$$

where $K(u)$ is a *kernel function* satisfying $K(u) \geq 0$ and $\int K(u)\,du = 1$, for example, $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ for $-\infty < u < \infty$.

  * It is easy to verify that

$$\int \frac{1}{h} K\left(\frac{\theta_i - \theta_i^{(t)}}{h}\right) d\theta_i = \int K(u)\,du = 1.$$

  * We can show that as $h \to 0$,

$$\frac{1}{h} K\left(\frac{\theta_i - \theta_i^{(t)}}{h}\right) \to \delta(\theta_i - \theta_i^{(t)}),$$

where $\delta(\cdot)$ is the Dirac delta function. Hence,

$$\hat{p}(\theta_i \,|\, Y = y) = \frac{1}{m - m_0} \sum_{t=m_0+1}^{m} \frac{1}{h} K\left(\frac{\theta_i - \theta_i^{(t)}}{h}\right) \approx \frac{1}{m - m_0} \sum_{t=m_0+1}^{m} \delta(\theta_i - \theta_i^{(t)}).$$

5.4 Implementation

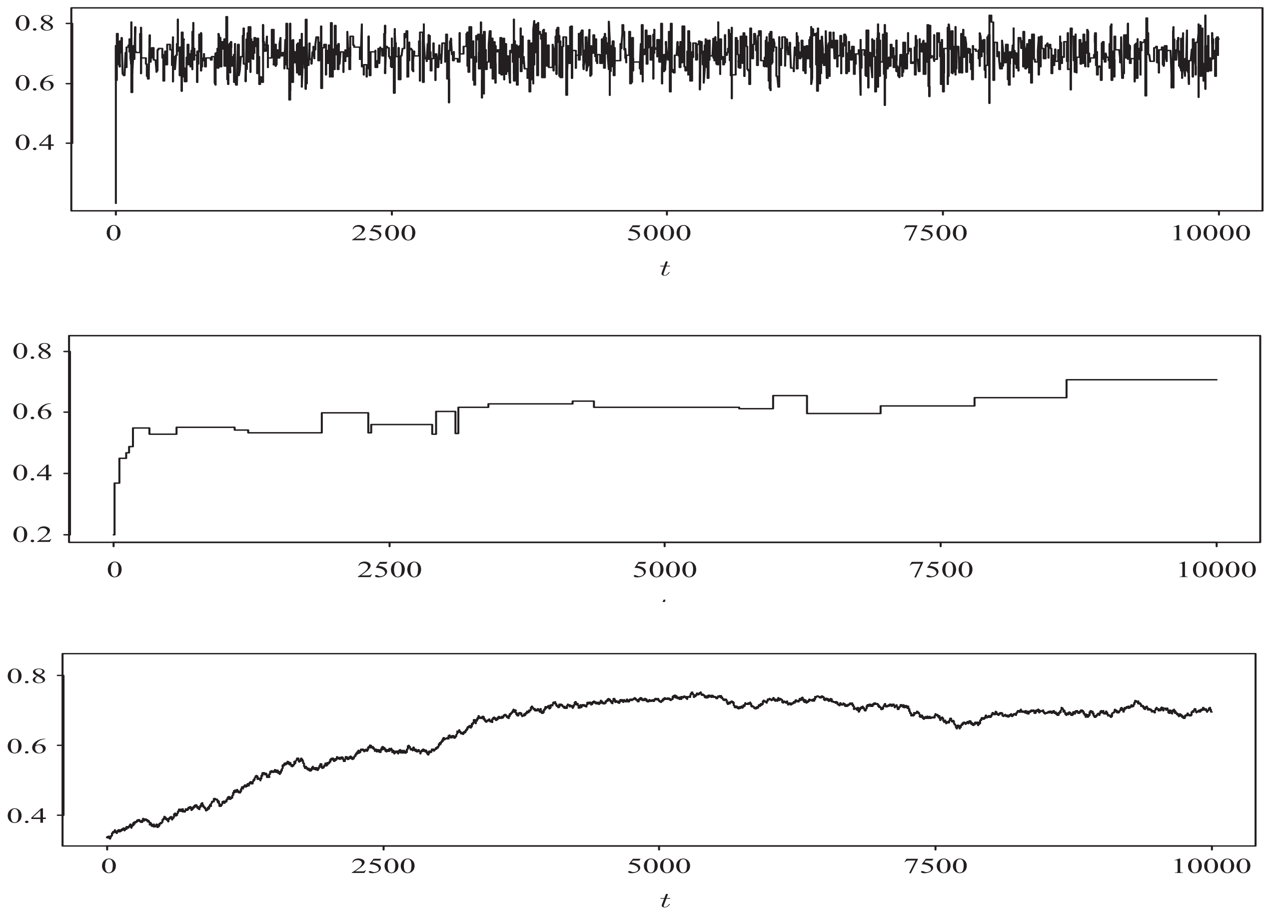
- **Posterior Mean:** We can estimate θ_i by

$$\hat{\theta}_i = \frac{1}{m - m_0} \sum_{t=m_0+1}^{m} \theta_i^{(t)} \approx E\big(\theta_i \,|\, Y\big).$$

- **Highest Posterior Density (HPD) Interval:** The shortest interval contains $100(1-\alpha)\%$ of the posterior probability.
  - For simplicity, we assume $m_0 = 0$ and arrange the samples in the ascending order as $\theta_i^{*(1)} \leq \cdots \leq \theta_i^{*(m)}$.
  - If we believe the posterior distribution is unimode and symmetric, then the $100(1-\alpha)\%$ HPD interval is $\big(\theta_i^{*(m\alpha/2)}, \theta_i^{*(m(1-\alpha/2))}\big)$.
  - If the posterior distribution is unimode but not symmetric, we need to find the interval $\big(\theta_i^{*(k)}, \theta_i^{*(k+m(1-\alpha))}\big)$, $k = 1, \cdots, m\alpha$, with the shortest length.

5.4 Implementation

- **Traceplots:** We often use *traceplot* to investigate mixing rate of the MCMC samples $\theta_i^{(0)}, \theta_i^{(1)}, \cdots, \theta_i^{(m)}$.



5.4 Implementation

- **Effective Sample Size:** For simplicity, assume $m_0 = 0$, θ_i is estimated by $\hat{\theta}_i = \frac{1}{m}\sum_{t=1}^{m} \theta_i^{(t)}$.
  - Assume that $\theta_i^{(1)}$ (approximately) follows the invariant distribution $p(\theta_i \mid Y = y)$, $\theta_i^{(2)}, \theta_i^{(3)}, \cdots$ also follow the invariant distribution. Then $\mathrm{Var}\big(\theta_i^{(t)}\big)$ does not depend on t.
  - We can further show that

$$\rho_i(k) := \frac{\mathrm{Cov}\big(\theta_i^{(t)}, \theta_i^{(t+k)}\big)}{\left[\mathrm{Var}\big(\theta_i^{(t)}\big)\mathrm{Var}\big(\theta_i^{(t+k)}\big)\right]^{1/2}} = \frac{\mathrm{Cov}\big(\theta_i^{(t)}, \theta_i^{(t+k)}\big)}{\mathrm{Var}\big(\theta_i^{(t)}\big)}$$

    does not depend on t.

- – Then we have

$$
\begin{aligned}
\mathrm{Var}\big(\hat{\theta}_i\big) &= \frac{1}{m^2}\sum_{t=1}^{m}\sum_{s=1}^{m}\mathrm{Cov}\big(\theta_i^{(t)},\theta_i^{(s)}\big) \\
&= \frac{1}{m^2}\sum_{t=1}^{m}\mathrm{Var}\big(\theta_i^{(t)}\big) + \frac{2}{m^2}\sum_{t<s}\mathrm{Cov}\big(\theta_i^{(t)},\theta_i^{(s)}\big) \\
&= \frac{1}{m^2}\mathrm{Var}\big(\theta_i^{(t)}\big)\Big[m + 2(m-1)\rho_i(1) + 2(m-2)\rho_i(2) + 2(m-3)\rho_i(3) + \cdots\Big] \\
&\approx \frac{1}{m}\mathrm{Var}\big(\theta_i^{(t)}\big)\Big[1 + 2\rho_i(1) + 2\rho_i(2) + 2\rho_i(3) + \cdots\Big]
\end{aligned}
$$

  – The *effective sample size* of $\theta_i^{(1)},\cdots,\theta_i^{(m)}$ is defined as

$$
\mathrm{ESS}_i := m/\Big[1 + 2\sum_{k=1}^{\infty}\rho_i(k)\Big],
$$

which indicates that $\theta_i^{(1)},\cdots,\theta_i^{(m)}$ perform as ESS_i i.i.d. samples drawn from the target distribution. In practice, we can use $\hat{\mathrm{ESS}}_i := m/\big[1 + 2\sum_{k=1}^{K}\hat{\rho}_i(k)\big]$, where $\hat{\rho}_i(k)$ is the sample correlation coefficient.

- **Estimating Normalizing Constants:** Consider the target distribution $f(x) \propto \bar{f}(x)$, where $\bar{f}(x)$ is known. We want to calculate the normalizing constant $\int \bar{f}(x)\,dx$.
  - **Example:** Consider a state space model with the joint distribution $p(x_{0:T}, y_{1:T}; \theta)$. We want to calculate $p(y_{1:T}; \theta) = \int p(x_{0:T}, y_{1:T}; \theta)\,dx_{0:T}$, which is the likelihood function for the observed data $y_{1:T}$.
  - **Importance Sampling/Particle Filter:** We generate $X^{(1)}, \cdots, X^{(m)}$ from a trial distribution $q(x)$ with $\mathcal{X}_q \supset \mathcal{X}_f$, and calculate $w^{(j)} = \bar{f}(X^{(j)})/q(X^{(j)})$. Then

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} w^{(j)} &\xrightarrow{a.s.} E\big(w^{(j)}\big) \\ &= \int \frac{\bar{f}(x)}{q(x)}\,q(x)\,dx = \int \bar{f}(x)\,dx.\end{aligned}$$

5.4 Implementation

- – **MCMC:** We use the MCMC algorithm to generate a Markov chain $X^{(1)}, \cdots, X^{(m)}$ with invariant distribution $f(x)$. Choose a density function $g(x)$ satisfying $\mathcal{X}_g \subset \mathcal{X}_f$. Then

$$
\begin{aligned}
\frac{1}{m}\sum_{t=1}^{m}\frac{g(X^{(t)})}{\bar{f}(X^{(t)})} &\xrightarrow{a.s.} E_f\Big[\frac{g(X^{(t)})}{\bar{f}(X^{(t)})}\Big] \\
&= \int_{\mathcal{X}_f}\frac{g(x)}{\bar{f}(x)}\,f(x)\,dx \\
&= \int_{\mathcal{X}_f} g(x)\,\frac{1}{\int_{\mathcal{X}_f}\bar{f}(u)\,du}\,dx \\
&= \frac{1}{\int_{\mathcal{X}_f}\bar{f}(u)\,du}\int_{\mathcal{X}_f} g(x)\,dx \\
&= \frac{1}{\int_{\mathcal{X}_f}\bar{f}(u)\,du}.
\end{aligned}
$$

Homework

1. For a Markov chain $X^{(0)}, X^{(1)}, \cdots, X^{(t)}, \cdots$, prove that

$$P\big(X^{(t+1)} = x^{(t+1)} \,|\, X^{(t)} = x^{(t)}, X^{(t-1)} = x^{(t-1)}\big) = P\big(X^{(t+1)} = x^{(t+1)} \,|\, X^{(t)} = x^{(t)}\big)$$

and

$$\begin{aligned} P\big(X^{(t+2)} &= x^{(t+2)}, X^{(t+1)} = x^{(t+1)} \,|\, X^{(0:t)} = x^{(0:t)}\big) \\ &= P\big(X^{(t+2)} = x^{(t+2)}, X^{(t+1)} = x^{(t+1)} \,|\, X^{(t)} = x^{(t)}\big) \end{aligned}$$

for all t.

2. Let $\{X^{(t)}, t = 0, 1, \cdots\}$ be a homogeneous Markov chain, where $X^{(t)} \in \{0, 1, 2\}$ and

$$\mathbb{T} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.6 & 0 & 0.4 \\ 0.5 & 0 & 0.5 \end{pmatrix}.$$

Suppose that $P(X^{(0)} = 0) = P(X^{(0)} = 1) = P(X^{(0)} = 2) = 1/3$. Determine $\mathbb{T}^2$ and $E(X^{(3)})$.

Homework

3. Consider a 2D Ising model with

$$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{-U(\boldsymbol{x})}}{\sum_{\boldsymbol{x}} e^{-U(\boldsymbol{x})}},$$

where

$$U(\boldsymbol{x}) = -0.2 \times \sum_{(i,j)\sim(i',j')} x_{i,j}\, x_{i',j'} + 0.3 \times \sum_{i,j} x_{i,j},$$

(1) Implement the Gibbs sampling algorithm to estimate the internal energy, $E(U(\boldsymbol{X}))$, defined on a 4×5 grid. Compare the estimated value of $E\big(U(\boldsymbol{X})\big)$ with its true value.

(2) Calculate $E(U(\boldsymbol{X}))$ of an Ising model defined on a 20×20 grid.

4. In the MCMC algorithm for stochastic volatility model, we can show that $p(z_t \mid z_{t-1}, z_{t+1}, b_0, b_1, \delta^2)$ is a normal distribution $N(\mu_{z_t}, \sigma^2_{z_t})$. Determine μ_{z_t} and $\sigma^2_{z_t}$ for $t = 0, 1, \cdots, T$.