Markov chain Monte Carlo

Timothy Hanson[1] and Alejandro Jara[2]

[1] Division of Biostatistics, University of Minnesota, USA
[2] Department of Statistics, Universidad de Concepción, Chile

IAP-Workshop 2009 Modeling Association and Dependence in Complex Data

Catholic University of Leuven, Leuven, November, 2009

Outline

1. Simulating posterior distributions
2. Discrete state space Markov chains
3. Continuous state space Markov chains

Obtaining posterior inference

- We start with a full Bayesian probability model. May be hierarchical, involve dependent data, etc.

Obtaining posterior inference

- We start with a full Bayesian probability model. May be hierarchical, involve dependent data, etc.
- Must be possible to evaluate *unnormalized* posterior

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\theta_1, \ldots, \theta_k | y_1, \ldots, y_n).$$

Obtaining posterior inference

- We start with a full Bayesian probability model. May be hierarchical, involve dependent data, etc.
- Must be possible to evaluate *unnormalized* posterior

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\theta_1, \ldots, \theta_k | y_1, \ldots, y_n).$$

- e.g. In simple model $\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ this is usual

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Obtaining posterior inference

- We start with a full Bayesian probability model. May be hierarchical, involve dependent data, etc.
- Must be possible to evaluate *unnormalized* posterior

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\theta_1, \ldots, \theta_k|y_1, \ldots, y_n).$$

- e.g. In simple model $\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ this is usual

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

- e.g. In hierarchical model $\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\tau} \sim p(\mathbf{y}|\boldsymbol{\theta})$, $\boldsymbol{\theta}|\boldsymbol{\tau} \sim p(\boldsymbol{\theta}|\boldsymbol{\tau})$, $\boldsymbol{\tau} \sim p(\boldsymbol{\tau})$ this is

$$p(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{y}) \propto p(\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\tau})p(\boldsymbol{\theta}, \boldsymbol{\tau}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\tau})p(\boldsymbol{\tau}).$$

Monte Carlo inference

- Sometimes it is possible to *sample* directly from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ (or $p(\boldsymbol{\theta},\boldsymbol{\tau}|\mathbf{y})$, etc.): $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^M \stackrel{iid}{\sim} p(\boldsymbol{\theta}|\mathbf{y})$.

Monte Carlo inference

- Sometimes it is possible to *sample* directly from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ (or $p(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{y})$, etc.): $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^M \overset{iid}{\sim} p(\boldsymbol{\theta}|\mathbf{y})$.
- We can use empirical estimates (mean, variance, quantiles, etc.) based on $\{\boldsymbol{\theta}^k\}_{k=1}^M$ to estimate the corresponding population parameters.
  - $M^{-1}\sum_{k=1}^M \boldsymbol{\theta}^k \approx E(\boldsymbol{\theta}|\mathbf{y})$.
  - p^{th} quantile: where $0 < p < 1$, $[\cdot]$ integer function, $\theta_j^{[pM]} \approx q$ such that $\int_{-\infty}^q p(\theta_j|\mathbf{y})d\theta_j = p$.
  - etc.

Markov chain Monte Carlo (MCMC)

- Only very simple models are amenable to Monte Carlo estimation of posterior inference.

Markov chain Monte Carlo (MCMC)

- Only very simple models are amenable to Monte Carlo estimation of posterior inference.
- A generalization of the Monte Carlo approach is Markov chain Monte Carlo.

Markov chain Monte Carlo (MCMC)

- Only very simple models are amenable to Monte Carlo estimation of posterior inference.
- A generalization of the Monte Carlo approach is Markov chain Monte Carlo.
- Instead of independent draws $\{\theta^k\}$ from the posterior, we obtain *dependent* draws.

Markov chain Monte Carlo (MCMC)

- Only very simple models are amenable to Monte Carlo estimation of posterior inference.
- A generalization of the Monte Carlo approach is Markov chain Monte Carlo.
- Instead of independent draws $\{\theta^k\}$ from the posterior, we obtain *dependent* draws.
- Treat them the same as if they were independent though. Ergodic theorems (Tierney, 1994, Section 3.3) provide LLN for MCMC iterates.

Markov chain Monte Carlo (MCMC)

- Only very simple models are amenable to Monte Carlo estimation of posterior inference.
- A generalization of the Monte Carlo approach is Markov chain Monte Carlo.
- Instead of independent draws $\{\theta^k\}$ from the posterior, we obtain *dependent* draws.
- Treat them the same as if they were independent though. Ergodic theorems (Tierney, 1994, Section 3.3) provide LLN for MCMC iterates.
- Let's get a taste of some fundamental ideas behind MCMC.

Discrete state space Markov chain

- Let $S = \{s_1, s_2, \ldots, s_m\}$ be a set of m states. Without loss of generality, we will take $S = \{1, 2, \ldots, m\}$. Note this is a *finite* state space.

Discrete state space Markov chain

- Let $S = \{s_1, s_2, \ldots, s_m\}$ be a set of m states. Without loss of generality, we will take $S = \{1, 2, \ldots, m\}$. Note this is a *finite* state space.
- The sequence of vectors $\{X^k\}_{k=0}^{\infty}$ forms a Markov chain on S if

$$P(X^k = i | X^{k-1}, X^{k-2}, \ldots, X^2, X^1, X^0) = P(X^k = i | X^{k-1}),$$

where $i = 1, \ldots, m$ are the possible states. At time k, the distribution of X^k only cares about the previous X^{k-1} and none of the earlier $X^0, X^1, \ldots, X^{k-2}$.

Discrete state space Markov chain

- Let $S = \{s_1, s_2, \dots, s_m\}$ be a set of m states. Without loss of generality, we will take $S = \{1, 2, \dots, m\}$. Note this is a *finite* state space.
- The sequence of vectors $\{X^k\}_{k=0}^{\infty}$ forms a Markov chain on S if

$$P(X^k = i | X^{k-1}, X^{k-2}, \dots, X^2, X^1, X^0) = P(X^k = i | X^{k-1}),$$

  where $i = 1, \dots, m$ are the possible states. At time k, the distribution of X^k only cares about the previous X^{k-1} and none of the earlier $X^0, X^1, \dots, X^{k-2}$.
- If the probability distribution $P(X^k = i | X^{k-1})$ doesn't change with time k then the chain is said to be *homogeneous* or *stationary*. We will only discuss stationary chains.

Transition matrix

- Let $p_{ij} = P(X^k = j | X^{k-1} = i)$ be the probability of the chain going from state i to state j in one step. These values can be placed into a *transition matrix*:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{m,1} \\ p_{12} & p_{22} & \cdots & p_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1,m} & p_{2,m} & \cdots & p_{m,m} \end{bmatrix}.$$

Transition matrix

- Let $p_{ij} = P(X^k = j | X^{k-1} = i)$ be the probability of the chain going from state i to state j in one step. These values can be placed into a *transition matrix*:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{m,1} \\ p_{12} & p_{22} & \cdots & p_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1,m} & p_{2,m} & \cdots & p_{m,m} \end{bmatrix}.$$

- Each column specifies conditional probability distribution & elements add up to 1.

Transition matrix

- Let $p_{ij} = P(X^k = j | X^{k-1} = i)$ be the probability of the chain going from state i to state j in one step. These values can be placed into a *transition matrix*:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{m,1} \\ p_{12} & p_{22} & \cdots & p_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1,m} & p_{2,m} & \cdots & p_{m,m} \end{bmatrix}.$$

- Each column specifies conditional probability distribution & elements add up to 1.
- **Question**: Describe the chain with each column in the transition matrix is identical.

n-step transition matrix

- You should verify that the transition matrix for $P(X^k = j | X^{k-n} = i) = P(X^n = j | X^0 = i)$ (stationarity) is given by the product $\mathbf{P}^n$.

n-step transition matrix

- You should verify that the transition matrix for $P(X^k = j | X^{k-n} = i) = P(X^n = j | X^0 = i)$ (stationarity) is given by the product $\mathbf{P}^n$.
- This can be derived through iterative use of conditional probability statements, or by using the Chapman-Kolmogorov equations (which follow from iterative use of conditional probability statements).

Initial value X^0

- Say that the chain is started by drawing X^0 from $P(X^0 = j)$. These probabilities specify a the distribution for the *initial value* or *state* of the chain X^0.

Initial value X^0

- Say that the chain is started by drawing X^0 from $P(X^0 = j)$. These probabilities specify a the distribution for the *initial value* or *state* of the chain X^0.
- Silly but important question: What happens when $P(X^0 = j) = 0$ for $j = 1, 2, \ldots, m$? This has implications for choosing a starting value in MCMC.

Example

Let $\mathbf{p}^k$ be vector of probabilities $P(X^k = j) = p_j^k$. Let's look at an example.

- Three states $S = \{1, 2, 3\}$.

Example

Let $\mathbf{p}^k$ be vector of probabilities $P(X^k = j) = p_j^k$. Let's look at an example.

- Three states $S = \{1, 2, 3\}$.
- Initial state X^0 distributed
  $$\mathbf{p}^0 = \begin{bmatrix} P(X^0 = 1) \\ P(X^0 = 2) \\ P(X^0 = 3) \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.4 \\ 0.3 \end{bmatrix}.$$

Example

Let $\mathbf{p}^k$ be vector of probabilities $P(X^k = j) = p_j^k$. Let's look at an example.

- Three states $S = \{1, 2, 3\}$.
- Initial state X^0 distributed $\mathbf{p}^0 = \begin{bmatrix} P(X^0 = 1) \\ P(X^0 = 2) \\ P(X^0 = 3) \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.4 \\ 0.3 \end{bmatrix}.$
- Transition matrix $\mathbf{P} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix}.$

Example chain

$X^0 = 1$, $X^1, X^2, \ldots, X^{10}$ generated according to P.

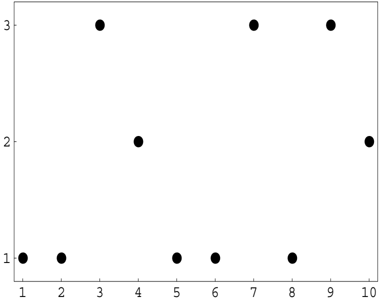

Figure: $X^0, X^1, \ldots, X^{10}$.

Longer chain

Example: Different $X^0 = 1, X^1, X^2, \ldots, X^{100}$ generated according to P.

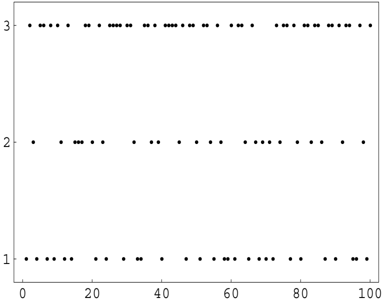

Figure: $X^1, \ldots, X^{100}$.

Limiting distribution

- *Marginal* or *unconditional* distribution of X^1 is given by the law of total probability

$$P(X^1 = j) = \sum_{i=1}^{m} P(X^1 = j | X^0 = i) P(X^0 = i).$$

Here, $m = 3$ states. In general, $\mathbf{p}^k = \mathbf{P}\mathbf{p}^{k-1}$.

Limiting distribution

- *Marginal* or *unconditional* distribution of X^1 is given by the law of total probability

$$P(X^1 = j) = \sum_{i=1}^{m} P(X^1 = j | X^0 = i) P(X^0 = i).$$

Here, $m = 3$ states. In general, $\mathbf{p}^k = \mathbf{P}\mathbf{p}^{k-1}$.

- Simply

$$\mathbf{p}^1 = \mathbf{P}\mathbf{p}^0 = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.4 \\ 0.3 \end{bmatrix} = \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix}.$$

Recursion...

- $$\mathbf{p}^1 = \begin{bmatrix} P(X^1 = 1) \\ P(X^1 = 2) \\ P(X^1 = 3) \end{bmatrix} = \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix}.$$

Recursion...

- $\mathbf{p}^1 = \begin{bmatrix} P(X^1 = 1) \\ P(X^1 = 2) \\ P(X^1 = 3) \end{bmatrix} = \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix}$.
- $\mathbf{p}^2 = \begin{bmatrix} P(X^2 = 1) \\ P(X^2 = 2) \\ P(X^2 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix}$.

Recursion...

- $\mathbf{p}^1 = \begin{bmatrix} P(X^1 = 1) \\ P(X^1 = 2) \\ P(X^1 = 3) \end{bmatrix} = \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix}.$
- $\mathbf{p}^2 = \begin{bmatrix} P(X^2 = 1) \\ P(X^2 = 2) \\ P(X^2 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix}.$
- $\mathbf{p}^3 = \begin{bmatrix} P(X^3 = 1) \\ P(X^3 = 2) \\ P(X^3 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$

Recursion...

- $\mathbf{p}^1 = \begin{bmatrix} P(X^1 = 1) \\ P(X^1 = 2) \\ P(X^1 = 3) \end{bmatrix} = \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix}.$
- $\mathbf{p}^2 = \begin{bmatrix} P(X^2 = 1) \\ P(X^2 = 2) \\ P(X^2 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix}.$
- $\mathbf{p}^3 = \begin{bmatrix} P(X^3 = 1) \\ P(X^3 = 2) \\ P(X^3 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$
- $\mathbf{p}^4 = \begin{bmatrix} P(X^4 = 1) \\ P(X^4 = 2) \\ P(X^4 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$

Recursion...

- $$\mathbf{p}^1 = \begin{bmatrix} P(X^1 = 1) \\ P(X^1 = 2) \\ P(X^1 = 3) \end{bmatrix} = \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix}.$$
- $$\mathbf{p}^2 = \begin{bmatrix} P(X^2 = 1) \\ P(X^2 = 2) \\ P(X^2 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix}.$$
- $$\mathbf{p}^3 = \begin{bmatrix} P(X^3 = 1) \\ P(X^3 = 2) \\ P(X^3 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$$
- $$\mathbf{p}^4 = \begin{bmatrix} P(X^4 = 1) \\ P(X^4 = 2) \\ P(X^4 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$$
- $$\mathbf{p}^5 = \begin{bmatrix} P(X^5 = 1) \\ P(X^5 = 2) \\ P(X^5 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$$

Recursion...

- $$\mathbf{p}^1 = \begin{bmatrix} P(X^1 = 1) \\ P(X^1 = 2) \\ P(X^1 = 3) \end{bmatrix} = \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix}.$$
- $$\mathbf{p}^2 = \begin{bmatrix} P(X^2 = 1) \\ P(X^2 = 2) \\ P(X^2 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix}.$$
- $$\mathbf{p}^3 = \begin{bmatrix} P(X^3 = 1) \\ P(X^3 = 2) \\ P(X^3 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$$
- $$\mathbf{p}^4 = \begin{bmatrix} P(X^4 = 1) \\ P(X^4 = 2) \\ P(X^4 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$$
- $$\mathbf{p}^5 = \begin{bmatrix} P(X^5 = 1) \\ P(X^5 = 2) \\ P(X^5 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$$
- $$\mathbf{p}^{50} = \begin{bmatrix} P(X^{50} = 1) \\ P(X^{50} = 2) \\ P(X^{50} = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}.$$

Recursion...

- $\mathbf{p}^1 = \begin{bmatrix} P(X^1 = 1) \\ P(X^1 = 2) \\ P(X^1 = 3) \end{bmatrix} = \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix}$.
- $\mathbf{p}^2 = \begin{bmatrix} P(X^2 = 1) \\ P(X^2 = 2) \\ P(X^2 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.33 \\ 0.20 \\ 0.47 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix}$.
- $\mathbf{p}^3 = \begin{bmatrix} P(X^3 = 1) \\ P(X^3 = 2) \\ P(X^3 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.21 \\ 0.54 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}$.
- $\mathbf{p}^4 = \begin{bmatrix} P(X^4 = 1) \\ P(X^4 = 2) \\ P(X^4 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}$.
- $\mathbf{p}^5 = \begin{bmatrix} P(X^5 = 1) \\ P(X^5 = 2) \\ P(X^5 = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}$.
- $\mathbf{p}^{50} = \begin{bmatrix} P(X^{50} = 1) \\ P(X^{50} = 2) \\ P(X^{50} = 3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.3 \\ 0.8 & 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}$.
- $\mathbf{p}^\infty = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}$ is *limiting* or *stationary* distribution of Markov chain. Here it's essentially reached within 3 iterations!

Starting value gets lost

Limiting distribution doesn't care about initial value

- When $\mathbf{p}^0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, stationary distribution $p^\infty = \begin{bmatrix} 0.26 \\ 0.23 \\ 0.51 \end{bmatrix}$ essentially reached within 5 iterations (within two significant digits), but is only reached exactly at $k = \infty$.
- Note that stationary distribution satisfies $\mathbf{p}^\infty = \mathbf{P}\mathbf{p}^\infty$. That is, if $X^{k-1} \sim \mathbf{p}^\infty$ then so is $X^k \sim \mathbf{p}^\infty$.

Some important notions

- An *absorbing* state exists if $p_{ii} = 1$ for some i. i.e. once X^k enters state i, it stays there forever.

Some important notions

- An *absorbing* state exists if $p_{ii} = 1$ for some i. i.e. once X^k enters state i, it stays there forever.
- If every state can be reached from every other state in finite time the chain is *irreducible*. This certainly occurs when $p_{ij} > 0$ for all i and j as each state can be reached from any other state in one step!

Some important notions

- An *absorbing* state exists if $p_{ii} = 1$ for some i. i.e. once X^k enters state i, it stays there forever.
- If every state can be reached from every other state in finite time the chain is *irreducible*. This certainly occurs when $p_{ij} > 0$ for all i and j as each state can be reached from any other state in one step!
- **Question:** Can a chain with an absorbing state be irreducible?

Positive recurrence

- Say the chain starts at $X^0 = i$. Consider

$$P(X^k = i, X^{k-1} \neq i, X^{k-2} \neq i, \ldots, X^2 \neq i, X^1 \neq i | X^0 = i).$$

This is the probability that the *first return* to state i occurs at time k. State i is *recurrent* if

$$\sum_{k=1}^{\infty} P(X^k = i, X^{k-1} \neq i, X^{k-2} \neq i, \ldots, X^2 \neq i, X^1 \neq i | X^0 = i) = 1.$$

Positive recurrence

- Say the chain starts at $X^0 = i$. Consider
$$P(X^k = i, X^{k-1} \neq i, X^{k-2} \neq i, \ldots, X^2 \neq i, X^1 \neq i | X^0 = i).$$
This is the probability that the *first return* to state i occurs at time k. State i is *recurrent* if
$$\sum_{k=1}^{\infty} P(X^k = i, X^{k-1} \neq i, X^{k-2} \neq i, \ldots, X^2 \neq i, X^1 \neq i | X^0 = i) = 1.$$
- Let T_i be distributed with the above probability distribution; T_i is the *first return time* to i.

Positive recurrence

- Say the chain starts at $X^0 = i$. Consider

  $$P(X^k = i, X^{k-1} \neq i, X^{k-2} \neq i, \ldots, X^2 \neq i, X^1 \neq i | X^0 = i).$$

  This is the probability that the *first return* to state i occurs at time k. State i is *recurrent* if

  $$\sum_{k=1}^{\infty} P(X^k = i, X^{k-1} \neq i, X^{k-2} \neq i, \ldots, X^2 \neq i, X^1 \neq i | X^0 = i) = 1.$$

- Let T_i be distributed with the above probability distribution; T_i is the *first return time* to i.
- If $E(T_i) < \infty$ the state is *positive recurrent*. The *chain* is positive recurrent if all states are positive recurrent.

Positive recurrence

- Say the chain starts at $X^0 = i$. Consider

  $$P(X^k = i, X^{k-1} \neq i, X^{k-2} \neq i, \ldots, X^2 \neq i, X^1 \neq i | X^0 = i).$$

  This is the probability that the *first return* to state i occurs at time k. State i is *recurrent* if

  $$\sum_{k=1}^{\infty} P(X^k = i, X^{k-1} \neq i, X^{k-2} \neq i, \ldots, X^2 \neq i, X^1 \neq i | X^0 = i) = 1.$$

- Let T_i be distributed with the above probability distribution; T_i is the *first return time* to i.
- If $E(T_i) < \infty$ the state is *positive recurrent*. The *chain* is positive recurrent if all states are positive recurrent.
- **Question:** Is an absorbing state recurrent? Positive recurrent? If so, what is $E(T_i)$?

Periodicity

- A state i is *periodic* if in can be *re*-visited only at regularly spaced times. Formally, define

$$d(i) = \text{g.c.d}\{k : (\mathbf{P}^k)_{ii} > 0\}$$

where g.c.d. stands for greatest common divisor. i is periodic if $d(i) > 1$ with period $d(i)$.

Periodicity

- A state i is *periodic* if in can be *re*-visited only at regularly spaced times. Formally, define

$$d(i) = \text{g.c.d}\{k : (\mathbf{P}^k)_{ii} > 0\}$$

  where g.c.d. stands for greatest common divisor. i is periodic if $d(i) > 1$ with period $d(i)$.
- A state is *aperiodic* if $d(i) = 1$. This of course happens when $p_{ij} > 0$ for all i and j.

Periodicity

- A state i is *periodic* if in can be *re*-visited only at regularly spaced times. Formally, define

$$d(i) = \text{g.c.d}\{k : (\mathbf{P}^k)_{ii} > 0\}$$

where g.c.d. stands for greatest common divisor. i is periodic if $d(i) > 1$ with period $d(i)$.
- A state is *aperiodic* if $d(i) = 1$. This of course happens when $p_{ij} > 0$ for all i and j.
- A chain is *aperiodic* if all states i are aperiodic.

What is the point?

If a Markov chain $\{X^k\}_{k=0}^{\infty}$ is aperiodic, irreducible, and positive recurrent, it is *ergodic*.

Theorem: Let $\{X^k\}_{k=0}^{\infty}$ be an *ergodic* (discrete time) Markov chain. Then there exists a stationary distribution $\mathbf{p}^{\infty}$ such that $\mathbf{p}_i^{\infty} > 0$ for $i = 1, \ldots, m$, that satisfies $\mathbf{P}\mathbf{p}^{\infty} = \mathbf{p}^{\infty}$ and $\mathbf{p}^k \to \mathbf{p}^{\infty}$.

- Can get a draw from $\mathbf{p}^{\infty}$ by running chain out a ways (from *any* starting value in the state space S!). Then X^k, for k "large enough," is approximately distributed $\mathbf{p}^{\infty}$.

Aperiodicity

- Aperiodicity is important to ensure a limiting distribution. e.g. $\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ yields an irreducible, positive recurrent chain, but both states have period 2. There is no limiting distribution! For any initial distribution $\mathbf{p}^0 = \begin{bmatrix} P(X^0 = 1) \\ P(X^0 = 2) \end{bmatrix} = \begin{bmatrix} p_1^0 \\ p_2^0 \end{bmatrix}$, $\mathbf{p}^k$ alternates between $\begin{bmatrix} p_1^0 \\ p_2^0 \end{bmatrix}$ and $\begin{bmatrix} p_2^0 \\ p_1^0 \end{bmatrix}$.

Positive recurrence and irreducibility

- Positive recurrence roughly ensures that $\{X^k\}_{k=0}^{\infty}$ will visit each state i enough times (infinitely often) to reach the stationary distribution. A state is recurrent if it can keep happening.

Positive recurrence and irreducibility

- Positive recurrence roughly ensures that $\{X^k\}_{k=0}^{\infty}$ will visit each state i enough times (infinitely often) to reach the stationary distribution. A state is recurrent if it can keep happening.
- Irreducibility disallows the chain getting “stuck” in certain subsets of S and not being able to get out. Reducibility would imply that the full state space S could not be explored. Actually, if $\mathbf{P}\mathbf{p}^{\infty} = \mathbf{p}^{\infty}$ and the chain is irreducible, then the chain is positive recurrent.

Positive recurrence and irreducibility

- Positive recurrence roughly ensures that $\{X^k\}_{k=0}^{\infty}$ will visit each state *i* enough times (infinitely often) to reach the stationary distribution. A state is recurrent if it can keep happening.
- Irreducibility disallows the chain getting "stuck" in certain subsets of *S* and not being able to get out. Reducibility would imply that the full state space *S* could not be explored. Actually, if $\mathbf{P}\mathbf{p}^{\infty} = \mathbf{p}^{\infty}$ and the chain is irreducible, then the chain is positive recurrent.
- Note that everything is satisfied when $p_{ij} > 0$ for all *i* and *j*!!

Illustration

A reducible chain can still converge to its stationary distribution!

Let $\mathbf{P} = \begin{bmatrix} 1 & 0.5 \\ 0 & 0.5 \end{bmatrix}$. For any starting value, $\mathbf{p}^{\infty} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

Here, $\mathbf{p}^k$ does converge to stationary distribution, we just don't have $p_i^{\infty} > 0$ for $i = 1, 2$.

Markov chain Monte Carlo

MCMC algorithms are cleverly constructed so that the posterior distribution $p(\theta|\mathbf{y})$ is the *stationary distribution* of the Markov chain!

- Since θ typically lives in $\mathbb{R}^d$, there is a continuum of states. So the Markov chain is said to have a continuous state space.

Markov chain Monte Carlo

MCMC algorithms are cleverly constructed so that the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is the *stationary distribution* of the Markov chain!

- Since $\boldsymbol{\theta}$ typically lives in $\mathbb{R}^d$, there is a continuum of states. So the Markov chain is said to have a continuous state space.
- The transition matrix is replaced with a *transition kernel*: $P(\boldsymbol{\theta}^k \in A|\boldsymbol{\theta}^{k-1} = \mathbf{x}) = \int_A k(\mathbf{s}|\mathbf{x})d\mathbf{s}$.

Continuous state spaces...

- Notions such as aperiodicity, positive recurrence, and irreducibility are generalized for continuous state spaces (see Tierney, 1994). Same with ergodicity.

Continuous state spaces...

- Notions such as aperiodicity, positive recurrence, and irreducibility are generalized for continuous state spaces (see Tierney, 1994). Same with ergodicity.
- Stationary distribution now satisfies

$$
\begin{aligned}
P^{\infty}(A) &= \int_A p^{\infty}(\mathbf{s})d\mathbf{s} \\
&= \int_{\mathbf{x}\in\mathbb{R}^d} \left[\int_A k(\mathbf{s}|\mathbf{x})d\mathbf{s}\right] p^{\infty}(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathbf{x}\in\mathbb{R}^d} P(\boldsymbol{\theta}^k \in A|\boldsymbol{\theta}^{k-1} = \mathbf{x})p^{\infty}(\mathbf{x})d\mathbf{x}
\end{aligned}
$$

Continuous state spaces...

- Notions such as aperiodicity, positive recurrence, and irreducibility are generalized for continuous state spaces (see Tierney, 1994). Same with ergodicity.
- Stationary distribution now satisfies

$$\begin{aligned}
P^\infty(A) &= \int_A p^\infty(\mathbf{s})d\mathbf{s} \\
&= \int_{\mathbf{x}\in\mathbb{R}^d} \left[\int_A k(\mathbf{s}|\mathbf{x})d\mathbf{s}\right] p^\infty(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathbf{x}\in\mathbb{R}^d} P(\boldsymbol{\theta}^k \in A|\boldsymbol{\theta}^{k-1} = \mathbf{x})p^\infty(\mathbf{x})d\mathbf{x}
\end{aligned}$$

- Continuous time analogue to $\mathbf{p}^\infty = \mathbf{P}\mathbf{p}^\infty$.

MCMC

Again, MCMC algorithms cleverly construct $k(\theta|\theta^{k-1})$ so that $p^{\infty}(\theta) = p(\theta|\mathbf{y})$!

- Run chain out long enough and θ^k approximately distributed as $p^{\infty}(\theta) = p(\theta|\mathbf{y})$.

MCMC

Again, MCMC algorithms cleverly construct $k(\theta|\theta^{k-1})$ so that $p^\infty(\theta) = p(\theta|\mathbf{y})$!

- Run chain out long enough and θ^k approximately distributed as $p^\infty(\theta) = p(\theta|\mathbf{y})$.
- Whole idea is that kernel $k(\theta|\theta^{k-1})$ is *easy* to sample from but $p(\theta|\mathbf{y})$ is *difficult* to sample from.

MCMC

Again, MCMC algorithms cleverly construct $k(\boldsymbol{\theta}|\boldsymbol{\theta}^{k-1})$ so that $p^{\infty}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$!

- Run chain out long enough and $\boldsymbol{\theta}^k$ approximately distributed as $p^{\infty}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$.
- Whole idea is that kernel $k(\boldsymbol{\theta}|\boldsymbol{\theta}^{k-1})$ is *easy* to sample from but $p(\boldsymbol{\theta}|\mathbf{y})$ is *difficult* to sample from.
- Different kernels: Gibbs, Metropolis-Hastings, Metropolis-within-Gibbs, etc.

MCMC

Again, MCMC algorithms cleverly construct $k(\theta|\theta^{k-1})$ so that $p^\infty(\theta) = p(\theta|\mathbf{y})$!

- Run chain out long enough and θ^k approximately distributed as $p^\infty(\theta) = p(\theta|\mathbf{y})$.
- Whole idea is that kernel $k(\theta|\theta^{k-1})$ is *easy* to sample from but $p(\theta|\mathbf{y})$ is *difficult* to sample from.
- Different kernels: Gibbs, Metropolis-Hastings, Metropolis-within-Gibbs, etc.
- We will mainly consider variants of Gibbs sampling in R and DPpackage. WinBUGS can automate the process for some problems; most of the compiled functions in DPpackage use *Gibbs sampling* with some *Metropolis-Hastings* updates. Will discuss these next...

Simple example, finished...

Example: Back to simple finite, discrete state space example with $m = 3$.

- Run out chain $X^0, X^1, X^3, \ldots, X^{10000}$. Initial value X^0 doesn't matter.

Simple example, finished...

Example: Back to simple finite, discrete state space example with $m = 3$.

- Run out chain $X^0, X^1, X^3, \ldots, X^{10000}$. Initial value X^0 doesn't matter.
- Can estimate $\mathbf{p}^\infty$ by

$$\hat{\mathbf{p}}^\infty = \begin{bmatrix} \frac{1}{10000} \sum_{k=1}^{10000} I_{\{1\}}(X^k) \\ \frac{1}{10000} \sum_{k=1}^{10000} I_{\{2\}}(X^k) \\ \frac{1}{10000} \sum_{k=1}^{10000} I_{\{3\}}(X^k) \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.22 \\ 0.52 \end{bmatrix}.$$

Simple example, finished...

Example: Back to simple finite, discrete state space example with $m = 3$.

- Run out chain $X^0, X^1, X^3, \ldots, X^{10000}$. Initial value X^0 doesn't matter.
- Can estimate $\mathbf{p}^\infty$ by
$$\hat{\mathbf{p}}^\infty = \left[\begin{array}{c} \frac{1}{10000} \sum_{k=1}^{10000} I_{\{1\}}(X^k) \\ \frac{1}{10000} \sum_{k=1}^{10000} I_{\{2\}}(X^k) \\ \frac{1}{10000} \sum_{k=1}^{10000} I_{\{3\}}(X^k) \end{array} \right] = \left[\begin{array}{c} 0.26 \\ 0.22 \\ 0.52 \end{array} \right].$$
- This is *one* approximation from running chain out *once*. Will have (slightly) different answers each time.

Simple example, finished...

Example: Back to simple finite, discrete state space example with $m = 3$.

- Run out chain $X^0, X^1, X^3, \ldots, X^{10000}$. Initial value X^0 doesn't matter.
- Can estimate $\mathbf{p}^\infty$ by

$$\hat{\mathbf{p}}^\infty = \begin{bmatrix} \frac{1}{10000}\sum_{k=1}^{10000} I_{\{1\}}(X^k) \\ \frac{1}{10000}\sum_{k=1}^{10000} I_{\{2\}}(X^k) \\ \frac{1}{10000}\sum_{k=1}^{10000} I_{\{3\}}(X^k) \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.22 \\ 0.52 \end{bmatrix}.$$

- This is *one* approximation from running chain out *once*. Will have (slightly) different answers each time.
- LLN for ergodic chains guarantees this approximation will get better the longer the chain is taken out.