Statistical Computing
Chap. 4: Markov Chain Monte Carlo

LIU, Ran

March 29, 2025

Summary

介绍

- MCMC 方法的抽样策略就是要构造一个非周期不可约的马氏链使得其平稳分布等于我们的目标分布 f.
- 对于足够大的 t, 由这样的马氏链得到 X_t 的边际分布近似 f.
- MCMC 方法的一个非常流行的应用是帮助简便 Bayes 推断，这时 f 就是参数 X 的 Bayes 后验分布.
- MCMC 方法的精髓在于构造一适当的链 (转移矩阵).

Monte Carlo 方法的发展

- Monte Carlo 方法的起源可以追溯到 Metropolis 和 Ulam (1949)，他们首次提出了使用随机采样进行数值计算的思想。
- Metropolis et al. (1953) 进一步发展了 Monte Carlo 方法，提出了 Metropolis 算法，用于模拟物理系统的统计行为，例如 Ising 模型。
- Hastings (1970) 推广了 Metropolis 算法，提出了 Metropolis-Hastings (MH) 算法，使其适用于更一般的目标分布。
- Geman and Geman (1984) 在图像处理领域引入了 Gibbs 采样，专门用于从高维联合分布中采样。
- Gelfand et al. (1990) 进一步推广了 Gibbs 采样，使其成为贝叶斯统计计算的重要工具，在 MCMC（Markov Chain Monte Carlo）方法中得到了广泛应用。

马尔科夫链

考虑在每个时间段有一个值的随机过程. 令 X_n 表示它在时间段 n 的值, 如果

$$\begin{aligned}&\mathrm{P}\{X_{n+1}=j|X_n=i,X_{n-1}=i_{n-1},\cdots,X_1=i_1\}\\=&\mathrm{P}\{X_{n+1}=j|X_n=i\}\\=&P_{ij}\end{aligned}$$

这样的随机过程称为马尔科夫链. P 为一步转移概率 P_{ij} 的矩阵. n 次转移后的矩阵为 P^n.

Markov 链

对随机变量序列 $\{X_0, X_1, X_2, \cdots\}$，在任一时刻 $n \geq 0$, 序列中下一时刻 $n+1$ 的 X_{n+1} 由条件分布 $P(x|X_n)$ 产生，它只依赖于时刻 n 的当前状态，而与时刻 n 以前的历史状态 $\{X_0, \cdots, X_{n-1}\}$ 无关。满足这样条件的随机变量序列称为 Markov 链。

若转移概率不随 n 改变，则称此链为时间齐性的，否则为时间非齐性。

一些性质

如果从任一状态 i 经有限步后都可到达任一状态 j. 也就是说，对于任两个状态 i, j, 都存在 $m > 0$ 使得 $p(X_{m+n} = j | X_n = i) > 0$（连通的），则称这一条马氏链是不可约的（irreducible）.

简单来说就是，任意两个状态都是连通的。

我们称能以概率 1 回来的状态为常返的 (recurrent state),

$$f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)} = 1$$

其中

$$f_{ij}^{(n)} = p(X_n = j, X_k \neq j, k = 1, 2, ..., n-1 | X_0 = i)$$

为马氏链在 0 时从状态 i 出发, 经 n 步转移后, 首次到达状态 j 的概率 (首达概率). 若返回概率小于 1, 即 $f_{ii} < 1$, 则为非常返态。

简单来说就是，有限步能回到原地。

常返态又可以分为正常返和零常返。一个状态的平均返回时间 μ_i 为

$$\mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$$

若 $\mu_i < \infty$，则为正常返；反之 $\mu_i = \infty$ 为零常返。如果状态空间有限的话，其常返状态都是非零常返的 (正常返状态 positive recurrent state).

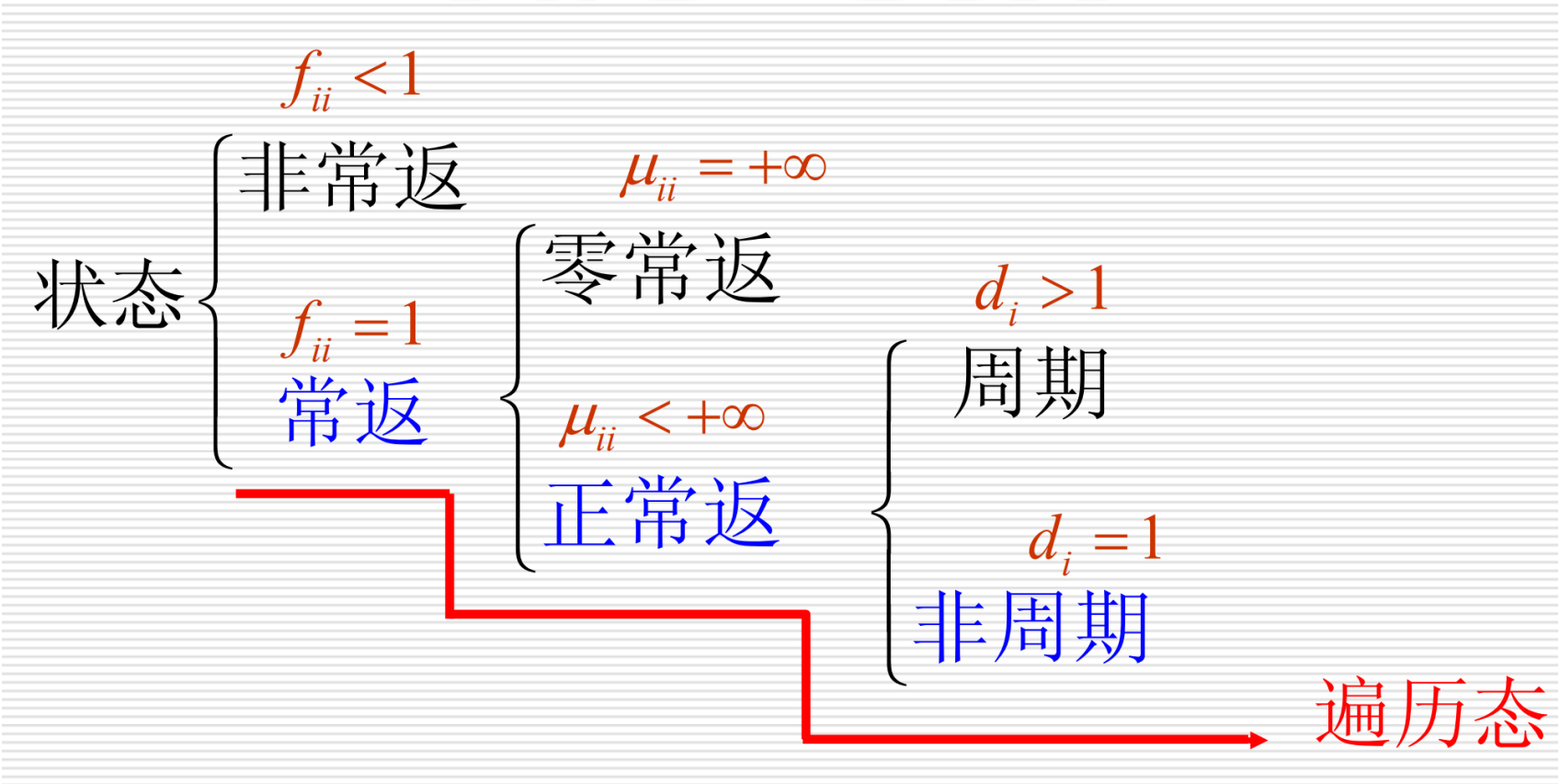
简单来说就是，有限步能回到原地，然后判断回到原地的步数的期望是否有限。

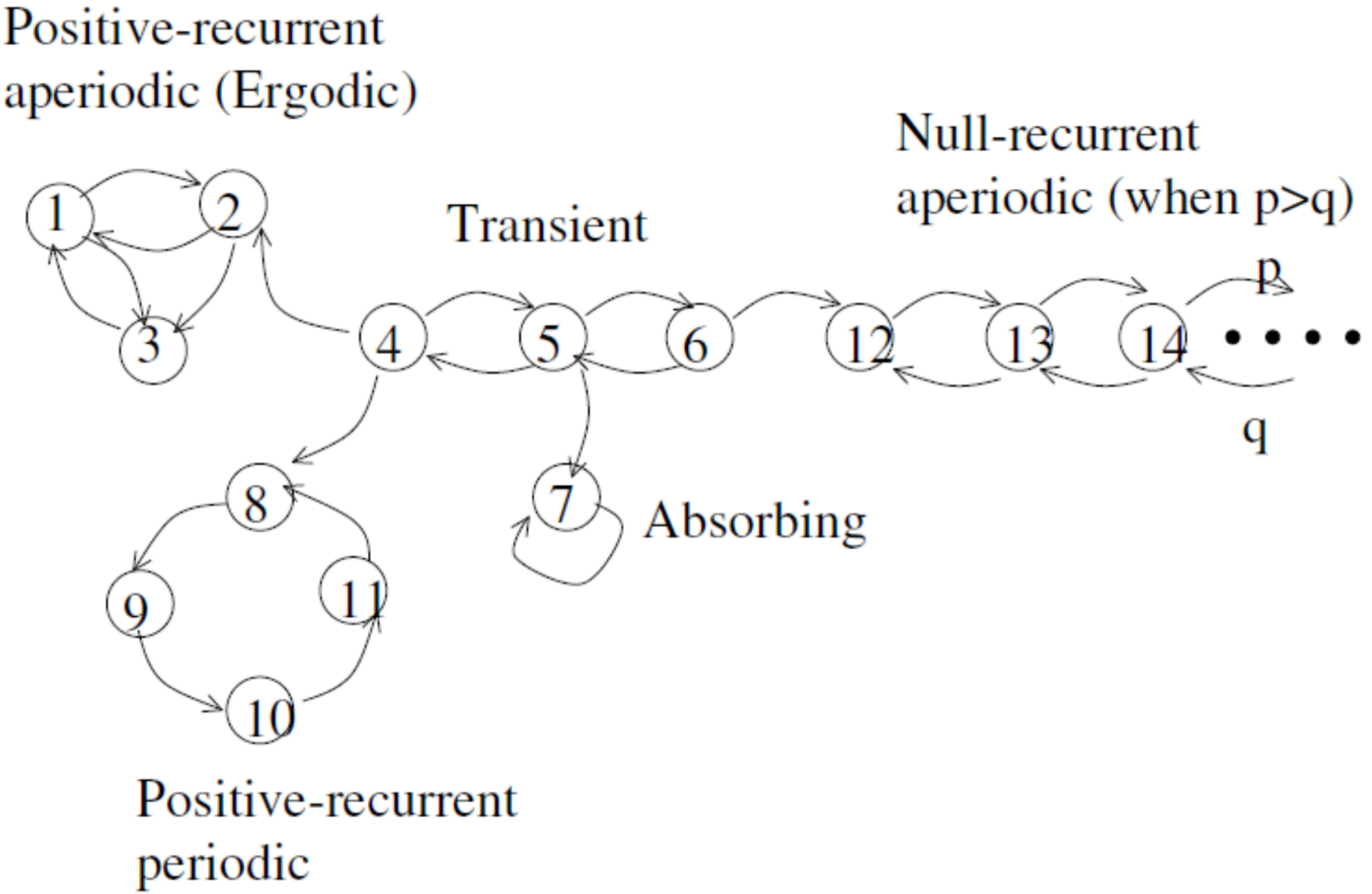
如果马尔可夫链只能以一定的规则间隔访问状态空间的某些部分，则它是周期性的. 称状态 j 具有周期 d, 如果由状态 j 经非 d 整数倍步到达 j 的概率为 0. 周期可定义为集合 $\{n : n \geq 0, f_{ii}^{(n)} > 0\}$ 的最大公约数，即所有返回可能的次数的最大公约数。

如果一条马氏链的每一个状态的周期都为 1，则称此链为非周期的。

简单来说就是看，状态达到存不存在某种规则间隔。

如果一条马氏链是不可约、非周期，且其所有状态都是非零常返的，则称之为遍历的 (ergodic).




Positive-recurrent
aperiodic (Ergodic)
Null-recurrent
aperiodic (when p>q)
Transient
p
q
Absorbing
Positive-recurrent
periodic
1
2
3
4
5
6
12
13
14
7
8
9
11
10

平稳分布

平稳分布定义：若一离散分布 π 满足 $\pi^T P = \pi^T$，则称之为转移概率矩阵为 P 的马氏链的平稳分布.

(充分条件) 如果一条时间齐性的马氏链满足

$$\pi_i p_{ij} = \pi_j p_{ji}, \forall i, j \in \mathcal{S},$$

则 π 是此链的平稳分布，且称此链为可逆的. 上述方程也称为细致平衡 (detailed balance).

如果一个转移概率阵为 P, 平稳分布为 π 的马氏链是不可约的，正常返的，且非周期的 (遍历), 则 π 唯一，且满足

$$\lim_{n\to\infty} P\left[X_n = j\right] = \pi_j,$$

也就是极限分布即为平稳分布。其中 π_j 是 π 的第 j 个元素，且满足如下方程组:

$$\pi_j \geqslant 0, \sum_{i\in\mathcal{S}} \pi_i = 1, 且 \pi_j = \sum_{i\in\mathcal{S}} \pi_i p_{ij}, \forall j \in \mathcal{S}.$$

推论：如果 $X_1, X_2, \ldots$ 是一不可约，正常返的，非周期的平稳分布为 π 的马氏链值，则 $X^{(n)}$ 依分布收敛到分布为 π 的随机变量.

MCMC 原理

- MCMC 的核心是构建转移矩阵，使得我们的目标分布 (π) 满足细致平衡，也即目标分布是构建出的马尔科夫链的平稳分布。
- 又因为遍历性，这条链的极限分布就是唯一的平稳分布。
- 所以我们只要迭代次数足够大，就能假设达到了平稳分布，X_n 即为目标分布的样本.

连续状态

若状态是连续的，我们有相似的定义：

Markov 链转移核

在连续分布情况下，对任一可测集 $\mathcal{B}$，一步转移概率定义为

$$P(x \to \mathcal{B}) = \int_{\mathcal{B}} p\left(x, x'\right) dx'.$$

转移概率 $p(\cdot, \cdot)$ 称为 Markov 链转移核。通常假定 $p(\cdot, \cdot)$ 与 t 无关，即基于该转移核的 Markov 链是时间齐次的。

例

根据转移核 $P(X_{t+1} \mid X_t) \sim N(0.5X_t, 1)$, 产生平稳分布是 $N(0, 4/3)$ 的 Markov 链。

只要验证细致平稳条件即可。注意到此时平稳分布为

$$\pi(x) = \frac{1}{\sqrt{4/3}\sqrt{2\pi}} \exp(-3x^2/8),$$

转移核为

$$p(x, x') = \frac{1}{\sqrt{2\pi}} \exp\{-(x' - x/2)^2/2\}.$$

从而，

$$\begin{aligned}\pi(x)p(x,x') &= \frac{1}{\sqrt{4/3}\cdot 2\pi}\exp(-3x^2/8)\exp\{-(x'-x/2)^2/2\}\\ &= \frac{1}{\sqrt{4/3}\cdot 2\pi}\exp\{-x^2/2-(x')^2/2+xx'/2\}\\ &= \pi(x')p(x',x),\end{aligned}$$

其中最后一步注意到 x 和 x' 在式子中的对称性就可以得到。

MCMC 方法估计 $E_\pi f(X) = \int f(x)\pi(x)dx$ 步骤概括如下:

基本流程

1. 选择转移核 $p(\cdot,\cdot)$（参数更新公式），使得其平稳分布是 $\pi(x)$;
2. 从某一点 X_0 出发，用上述转移核 $p(\cdot,\cdot)$ 产生 Markov 链序列 $X_0, X_1, \cdots, X_n$。
3. 对较大的 n，选择合适的 m，$E_\pi f(X) = \int f(x)\pi(x)dx$ 的估计为

$$\hat{E}_\pi f = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t).$$

- 根据上述步骤，构造的转移核 $p(\cdot,\cdot)$ 使得概率分布 $\pi(x)$ 为其平稳分布最为重要。
- 如何构造合适的转移核是 MCMC 方法主要研究的问题，不同的 MCMC 方法主要区别就是转移核的构造方法不同。

Summary

满条件分布

MCMC 方法中转移核 $p(x, x')$ 的构造大多建立在形如 $\pi(x_T \mid x_{-T})$ 的条件分布上，其中 $x_T = \{x_i, i \in T\}$, $x_{-T} = \{x_i, i \notin T\}$. 对于集合 T，我们有 $T \subset I = \{1, \cdots, p\}$，$p$ 是变量 x 的维度。

在条件分布 $\pi(x_T \mid x_{-T})$ 中，p 个分量或者出现在条件中，或者出现在变元中，这种条件分布称为满条件分布。

对任意 $T \subset I$，记 $x = (x_T, x_{-T}) \in X$. 满条件分布 $\pi(x_T \mid x_{-T})$ 具有如下性质

$$\pi\left(x_T \mid x_{-T}\right) = \frac{\pi(x)}{\int \pi(x) dx_T} \propto \pi(x).$$

一般情况下 $y = (x, z, \theta)$，这里 x 表示观测数据，z 表示缺失参数，θ 表示参数.

令 $p(x, z|\theta)$ 表示完全数据的密度函数，$\pi(\theta)$ 表示 θ 的先验分布，注意到 $f(y) = p(x, z|\theta)\pi(\theta)$, 则 y 的满条件分布如下

$$f(z_i|z_{-i}, x, \theta) \propto p(x, z|\theta)$$
$$\pi(\theta_i|\theta_{-i}, x, z) \propto f(y) \propto p(x, z|\theta)\pi(\theta_i|\theta_{-i})$$
$$f(x_i|x_{-i}, z, \theta) \propto p(x, z|\theta),$$

其中 $\theta_{-i} = \{\theta_j : j \neq i\}$, $z_{-i} = \{z_j : j \neq i\}$, $x_{-i} = \{x_j : j \neq i\}$.

例

设 $x = (x_1, x_2)$ 的密度函数为(黑板推导)

$$\pi(x_1, x_2) \propto \exp\{-\frac{1}{2}(x_1 - 1)^2 (x_2 - 1)^2\},$$

则满条件分布为

$$\begin{aligned}\pi(x_1 \mid x_2) &\propto \pi(x_1, x_2) \\ &\propto \exp\{-\frac{1}{2}(x_1 - 1)^2 (x_2 - 1)^2\} = N(1, (x_2 - 1)^{-2}),\end{aligned}$$

以及

$$\begin{aligned}\pi(x_2 \mid x_1) &\propto \pi(x_1, x_2) \\ &\propto \exp\{-\frac{1}{2}(x_2 - 1)^2 (x_1 - 1)^2\} = N(1, (x_1 - 1)^{-2}).\end{aligned}$$

伽马分布

伽马分布 $X \sim \Gamma(\alpha, \lambda)$ 的密度函数为

$$f(x) = \frac{\lambda^{\alpha} x^{(\alpha-1)} e^{(-\lambda x)}}{\Gamma(\alpha)}, \mathrm{x} > 0$$

其中 Gamma 函数之特征为:

$$\begin{cases} \Gamma(\alpha) = (\alpha - 1)! & \text{if } \alpha \text{ is } \mathbb{Z}^{+} \\ \Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) & \text{if } \alpha \text{ is } \mathbb{R}^{+} \\ \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \end{cases}$$

指数分布为 $\alpha = 1$ 的伽玛分布。

例子

设 $y_1, \cdots, y_n$ 独立同分布，来自正态分布 $N(\mu, \tau^{-1})$. 参数 μ, τ^{-1} 的先验分布分别为正态分布 $\mu \sim N(0,1)$, 伽马分布 $\tau \sim \Gamma(2,1)$, 且 μ 与 τ 独立. 计算满条件分布.

记 $y = (y_1, \cdots, y_n)$, 则 (y, μ, τ) 的联合密度函数为(黑板推导)

$$\begin{aligned} p(y, \mu, \tau) &= p(y \mid \mu, \tau) p(\mu) p(\tau) \\ &= (2\pi)^{-\frac{n+1}{2}} \tau^{\frac{n}{2}+1} \exp\{-\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \mu)^2 - \frac{\mu^2}{2} - \tau\}. \end{aligned}$$

参数的后验分布为

$$\pi(\mu, \tau \mid y) = \frac{p(y, \mu, \tau)}{\int p(y, \mu, \tau) d\mu d\tau} \propto p(y, \mu, \tau).$$

满条件分布 $\pi(\mu \mid \tau, y)$, $\pi(\tau \mid \mu, y)$ 分布为

$$\pi(\mu \mid \tau, y) \propto \exp\{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{\mu^2}{2}\}$$

$$\propto \exp\{-\frac{\tau}{2}n\mu^2 - \frac{\mu^2}{2} + \tau\mu\sum_{i=1}^{n}y_i\}$$

$$\propto \exp\{-\frac{1}{2}(1+n\tau)(\mu - \frac{\tau\sum_{i=1}^{n}y_i}{1+n\tau})^2\},$$

由上可得 $\pi(\mu \mid \tau, y)$ 的分布为

$$N(\tau\sum y_i/(1+n\tau), (1+n\tau)^{-1})$$

我们在看 τ 的满条件分布

$$\pi(\tau \mid \mu, y) \propto \tau^{\frac{n}{2}+1} \exp\{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2 - \tau\}$$

$$= \tau^{\frac{n}{2}+1} \exp\{-\tau(1+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2)\}.$$

则 $\tau \mid \mu, y$ 服从伽马分布

$$\Gamma(2+n/2,\ \ 1+1/2\sum_{i=1}^{n}(y_i-\mu)^2)$$

满条件分布并不都能表示为显示形式。

例子

样本 $y_i, i = 1, \cdots, n$ 独立且 $y_i \sim B(1, p_i)$ (伯努利分布)，其中 $p_i = (1+\exp(-(\alpha+\beta x_i)))^{-1}$. x_i 假设是固定的，已知的。参数 α, β 的先验分布分别为 $\alpha \sim N(0,1)$，$\beta \sim N(0,1)$，且 α, β 独立。

记 $y = \{y_1, y_2, \cdots, y_n\}$, 则 (y, α, β) 的联合分布为(黑板推导)

$$\pi(y, \alpha, \beta) = p(y \mid \alpha, \beta) p(\alpha) p(\beta)$$

$$\propto \prod_{i=1}^{n} \{(1+\exp\{\alpha+\beta x_i\})^{-(1-y_i)} (1+\exp\{-(\alpha+\beta x_i)\})^{-y_i}\}$$

$$\times \exp(-\frac{1}{2}\beta^2 - \frac{1}{2}\alpha^2).$$

其中上式因为

$$1 - p_i = 1 - \frac{1}{1+\exp(-(\alpha+\beta x_i))} = \left(\frac{1+\exp(-(\alpha+\beta x_i))}{\exp(-(\alpha+\beta x_i))}\right)^{-1}$$

则满条件分布 $\pi(\beta \mid \alpha, y)$ 和 $\pi(\alpha \mid \beta, y)$ 分别为:

$$
\begin{aligned}
& \pi(\beta \mid \alpha, Y) \\
\propto \quad & \exp\{-\frac{1}{2}\beta^2\} \prod_{i=1}^{n} \{(1+\exp\{\alpha+\beta x_i\})^{y_i-1}(1+\exp\{-(\alpha+\beta x_i)\})^{-y_i}\}, \\
& \pi(\alpha \mid \beta, y) \\
\propto \quad & \exp\{-\frac{1}{2}\alpha^2\} \prod_{i=1}^{n} \{(1+\exp\{\alpha+\beta x_i\})^{y_i-1}(1+\exp\{-(\alpha+\beta x_i)\})^{-y_i}\}.
\end{aligned}
$$

Summary

Metropolis–Hastings 算法

Metropolis-Hastings 算法是马尔可夫链蒙特卡洛 (MCMC) 方法的一种，用于从难直接抽样的概率分布中获取一系列随机样本。它通过构建一个 Markov 链来实现，使其平衡分布等于所需分布。

Metropolis-Hastings 方法转移核的构造如下

$$p\left(x, x'\right) = q\left(x' \mid x\right) \alpha\left(x \to x'\right), \tag{1}$$

潜在的转移核 $q(x' \mid x)$ 作为 x' 的函数是一个概率密度或概率分布，被称为提案分布。提案分布可以取各种形式，常把它取为易于产生随机数的分布。

Metropolis-Hastings 方法

Metropolis-Hastings 方法的具体实施方法为，如果链在时刻 t 处于状态 x，即 $X_t = x$. 首先由 $q(\cdot \mid x)$ 产生一个潜在的转移 $x \to x'$，然后根据概率 $\alpha(x \to x')$ 决定是否转移。

也就是说，在潜在转移点 x' 找到后，以概率 $\alpha(x \to x')$ 接受 x' 作为链在下一时刻 $t+1$ 的状态值，而以概率 $1 - \alpha(x \to x')$ 拒绝转移到 x'，从而链在下一时刻 $t+1$ 仍处于状态 x。

其中 $\alpha(x \to x')$ 称为接受概率，满足 $0 < \alpha(x \to x') \leq 1$。

实际计算中，产生区间 $[0,1]$ 上均匀分布的随机数 u，令

$$X_{t+1} = \begin{cases} x' & u \leq \alpha(x \to x') \\ x & u > \alpha(x \to x') \end{cases}$$

假设 $\pi(x)$ 为目标概率分布。MH 算法的过程为：

1. 初始化：选定初始状态 x_0，令 $t = 0$；
2. 迭代过程：
   1. 生成：从某一容易抽样的分布 $q(x'|x_t)$ 中随机生成候选状态 x'；
   2. 计算：计算是否采纳候选状态的概率
      $\alpha(x_t \to x') = \min\left(1, \frac{\pi(x')}{\pi(x_t)} \frac{q(x_t|x')}{q(x'|x_t)}\right)$
   3. 接受或拒绝
      1. 从 [0,1] 的均匀分布中生成随机数 u；
      2. 如 $u \leq \alpha(x_t \to x')$，则接受该状态，并令 $x_{t+1} = x'$；
      3. 如 $u > \alpha(x_t \to x')$，则拒绝该状态，并令 $x_{t+1} = x_t$ （复制原状态）；
   4. 增量：令 $t = t + 1$.

推导

MH 算法的推导开始于细致平衡，我们想在满足细致平衡的条件下构造转移概率：

$$p(x' \mid x)\pi(x) = p(x \mid x')\pi(x'),$$

这里的 $\pi(x)$ 是我们想要采样的目标分布。MH 算法是将这个转移概率分成了两个子步骤；提案和接受拒绝. 提案分布 $q(x'|x)$ 是在给定 x 后提出 x' 的条件概率，而接受分布 $\alpha(x \to x')$ 是要不要接受新提出的这一状态 x'. 转移概率被拆分成如下形式:

$$p(x'|x) = q(x'|x)\alpha(x \to x'),$$

* 其中 $q(x'|x)$ 提案产生候选值这一步，可包含决定性步骤 (概率为 1)，但一定要保证 $q(x|x')$ 也能由同样的决定性步骤到达。

我们再来看看，MH 的算法构造的转移概率满不满足细致平衡：

$$\pi(x)q(x' \mid x)\alpha(x \to x') = \min\left(\pi(x)q(x'|x), \pi(x')q(x|x')\right)$$

同样的我们有

$$\pi(x')q(x \mid x')\alpha(x' \to x) = \min\left(\pi(x')q(x|x'), \pi(x)q(x'|x)\right).$$

代入细致平衡的公式，即证：

$$p(x' \mid x)\pi(x) = p(x \mid x')\pi(x').$$

Metropolis 选择

提案分布 $q(x' \mid x)$ 可以取各种形式，接下来介绍几种常用的建议分布。

Metropolis **建议 $q(x' \mid x)$ 为对称分布，即**

$$q(x' \mid x) = q(x \mid x'), \quad \forall x, x'$$

此时，$\alpha(x \to x')$ 简化为

$$\alpha\left(x \to x'\right) = \min\{1, \frac{\pi\left(x'\right) q\left(x \mid x'\right)}{\pi(x) q\left(x' \mid x\right)}\} = \min\{1, \frac{\pi\left(x'\right)}{\pi(x)}\}.$$

常用的对称分布包括形式为 $q(x, x') = f(|x - x'|)$ 的分布，比如 random-walk Metropolis algorithm(RWM)：

$$q\left(x' \mid x\right) \propto \exp\{-\left(x' - x\right)^2 / (2\sigma_0^2)\}.$$

σ_0^2 可以自行选择，若接受率过大，可能会导致更新移动的太慢，可调大 σ_0；若接受率过小，采样效率就低，可调小 σ_0，一般接受率在 $20 - 40\%$ 左右比较好。

* 有理论证明 RWM 高维极限下的最佳接受率为 0.234。A. Gelman, W. R. Gilks, G. O. Roberts "Weak convergence and optimal scaling of random walk Metropolis algorithms," The Annals of Applied Probability, 7(1), 110-120, (February 1997)

例子

生成一个 Markov 链，使得其平稳分布为柯西分布，

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

选定的建议分布为 $q(x' \mid x)$ 是 $N(x, b^2)$，其中 b 为任意常数，如 $0.1, 1, 10$. **则此时**

$$\alpha\left(x, x'\right) = \min\{1, \frac{\pi\left(x'\right)}{\pi(x)}\} = \min\{1, \frac{1+x^2}{1+\left(x'\right)^2}\}$$

独立抽样

如果 $q\left(x' \mid x\right)$ 与当前状态 x 无关，即 $q\left(x' \mid x\right) = q\left(x'\right)$，则由此建议分布导出的 Metropolis-Hastings 算法称为独立抽样。此处，$\alpha\left(x, x'\right)$ 为

$$\alpha\left(x, x'\right) = \min\{1, \frac{\pi(x')/q(x')}{\pi(x)/q(x)}\}.$$

如果 $q(x)$ 接近 $\pi(x)$，基于独立抽样获取的 Markov 链的收敛效果更好。

* 不太建议，因为接受率可能很低，不如小量更新的 random walk MH。

例子

给定数据 $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\theta, 1)$, 先验分布 $\pi(\theta) = 1/\{\pi(1+\theta^2)\}$. 此时我们的后验分布为

$$\begin{aligned}
\pi(\theta \mid Y_1, \ldots, Y_n) &\propto p(Y_1, \ldots, Y_n \mid \theta)\pi(\theta) \\
&\propto \exp\{-(\sum_{i=1}^{n} \frac{(y_i-\theta)^2}{2})\} \times \frac{1}{1+\theta^2} \\
&\propto \exp\{-n(\theta-\overline{y})^2/2\} \times \frac{1}{1+\theta^2}.
\end{aligned}$$

假设已有数据给定 $n = 40$, $\overline{y} = 0.14$, 此时, 使用 x 和 x' 的记号, θ 的后验分布为

$$\pi(x) \propto \exp\{-40(x-0.14)^2/2\} \times \frac{1}{1+x^2}.$$

求后验期望 $E(x)$.

选定的建议分布为 $q(x' \mid x) = q(x') = 1/(\pi\{1 + (x')^2\})$. 此时,

$$\begin{aligned} \alpha\left(x, x'\right) &= \min\{1, \frac{\pi\left(x'\right) q(x)}{\pi(x) q(x')}\} \\ &= \min\left\{1, \frac{\exp\{-40(x' - 0.14)^2/2\}}{\exp\{-40(x - 0.14)^2/2\}}\right\}. \end{aligned}$$

采样之后求平均值, 即得后验期望.

单元素 Metropolis-Hastings 算法

在 X 是 p 维的情况，同时产生整个 X 有时是困难的 (接受率特别低)，而将 X 根据其分量逐个进行抽样则简单得多，这就要用到条件分布，特别是满条件分布性质。

单元素 Metropolis-Hastings 算法的想法是，对于 p 维变量 X，基于 $p-1$ 维变量 X_{-i} 的条件分布 $X_i \mid X_{-i}, i = 1, \cdots, p$，选择转移核 $q_i\left(x_i' \mid X_{-i} = x_{-i}\right)$。由转移核 $q_i\left(x_i' \mid X_{-i} = x_{-i}\right)$ 产生可能的 x_i'，以概率

$$\alpha_i\left(x_i' \mid x_{-i}\right) = \min(1, \frac{\pi\left(x'\right) q_i\left(x_i \mid x_{-i}\right)}{\pi(x) q_i\left(x_i' \mid x_{-i}\right)})$$

决定是否接受 x' 作为链的下一状态。

即每次只更新一个元素，其他保持不变：

$$\left\{x_1^{(t+1)}, ..., x_{i-1}^{(t+1)}, x_i^{(t)}, x_{i+1}^{(t)}, ..., x_n^{(t)}\right\}$$
$$\downarrow\downarrow$$
$$\left\{x_1^{(t+1)}, ..., x_{i-1}^{(t+1)}, x_i^{*}, x_{i+1}^{(t)}, ..., x_n^{(t)}\right\}$$

类似 coordinate asent.

Summary

Gibbs 算法

Gibbs 抽样是一种单元素 Metropolis-Hastings 算法的特殊情况，单元素 Metropolis-Hastings 算法中取 $q_i(x_i' \mid x_{-i})$ 为 $\pi\left(x_i|x_{-i}\right)$。此时，不难验证

$$\begin{aligned}\alpha_i(x_i' \mid x_{-i}) &= \min(1, \frac{\pi\left(x'\right)\pi\left(x_i \mid x_{-i}\right)}{\pi(x)\pi\left(x_i' \mid x_{-i}\right)}) \\ = \quad &\min(1, \frac{\pi(x_{-i})\pi\left(x_i' \mid x_{-i}\right)\pi\left(x_i \mid x_{-i}\right)}{\pi\left(x_{-i}\right)\pi(x_i \mid x_{-i})\pi\left(x_i' \mid x_{-i}\right)}) = 1,\end{aligned}$$

这里 $\pi(x_{-i})$ 是 x_{-i} 的密度函数。接受率等于 1 意味着，我们不需要舍弃样本，每个更新后的值都为样本。

Gibbs 步骤

吉布斯采样的过程则为:

1. 确定初始值 $\mathbf{X}^{(1)}$.
2. 假设已得到样本 $\mathbf{X}^{(i)}$, 记下一个样本为 $\mathbf{X}^{(i+1)} = \left(x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_n^{(i+1)}\right)$.

   对其中某一分量 $x_j^{(i+1)}$ 可通过在其他分量已知的条件下该分量的概率分布来抽取该分量。

   对于此条件概率，我们使用样本 $\mathbf{X}^{(i+1)}$ 中已得到的分量 $x_1^{(i+1)}$ 到 $x_{j-1}^{(i+1)}$ 以及上一样本 $\mathbf{X}^{(i)}$ 中的分量 $x_{j+1}^{(i)}$ 到 $x_n^{(i)}$，即

   $$f\left(x_j^{(i+1)} \middle| x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)}\right).$$
3. 重复上述过程.

$$X_1^{(t+1)}|\cdots f\left(x_1|x_2^{(t)},\cdots,x_p^{(t)}\right),$$

$$X_2^{(t+1)}|\cdots f\left(x_2|x_1^{(t+1)},x_3^{(t)},\cdots,x_p^{(t)}\right),$$

$$\cdots$$

$$X_{p-1}^{(t+1)}|\cdots f\left(x_{p-1}|x_1^{(t+1)},x_2^{(t+1)},\cdots,x_p^{(t)}\right),$$

$$X_p^{(t+1)}|\cdots f\left(x_p|x_1^{(t+1)},x_2^{(t+1)},\cdots,x_{p-1}^{(t+1)}\right),$$

更新排序

- X 元素的更新顺序对于不同的循环是可以变化的.
- 有时候对每个循环而言, 使用随机顺序是比较合理的. 这被称作为随机扫描 Gibbs 抽样.
- 事实上, 甚至没有必要在每个循环中对每个元素都进行更新, 而只要每个元素的更新足够地频繁就可以了.

区组化

当 X 的元素相关时，区组化特别有用，用其构造的算法能够使更相关的元素在同一个区组中被一起抽样出来.

$$X_1^{(t+1)} \mid \cdot \sim f\left(x_1 | x_2^{(t)}, x_3^{(t)}, x_4^{(t)}\right),$$

$$X_2^{(t+1)}, X_3^{(t+1)} \mid \cdot \sim f\left(x_2, x_3 | x_1^{(t+1)}, x_4^{(t)}\right),$$

$$X_4^{(t+1)} \mid \cdot \sim f(x_4 | x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}).$$

混合 Gibbs

我们可以在适当的时候使用不同的 MH 采样，比如

1. 用某 Gibbs 迭代更新 $X_1^{(t+1)} \mid \left(x_2^{(t)}, x_3^{(t)}, x_4^{(t)}, x_5^{(t)}, x_6^{(t)}\right)$;
2. 用某 MH 迭代更新 $\left(X_2^{(t+1)}, X_3^{(t+1)}\right) \mid \left(x_1^{(t+1)}, x_4^{(t)}, x_5^{(t)}, x_6^{(t)}\right)$;
3. 用某 MH 迭代更新 $X_4^{(t+1)} \mid \left(x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}, x_5^{(t)}, x_6^{(t)}\right)$;
4. 用某 Gibbs 迭代更新 $\left(X_5^{(t+1)}, X_6^{(t+1)}\right) \mid \left(x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}, x_4^{(t+1)}\right)$.

当 X 的一个或者多个元素的一元边际密度没有显示表达的时候，Gibbs 算法中的 Metropolis-Hastings 迭代可以使用. 有时也是 Gibbs 跳出局部最优的好方法.

值得注意的是，MH 在 Gibbs 中使用来更新某一些分量，专有名词叫 MH-within-Gibbs，它在实际使用时，MH 步骤不需要接受了样本再向前。但也要注意是否有变量因接受率极低，导致整个迭代过程都没有更新。

例子：已知参数，但假设我们只能单独产生单变量正态分布的随机数，如何从二维正态采样？给定一个目标分布：

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix})$$

我们有他们的联合概率分布

$$f(x \mid \mu, \Sigma) \propto \det(\Sigma)^{-1/2} \exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)).$$

如果我们用 Gibbs 算法的话：

$$f(x_1 \mid x_2) \sim N(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}),$$

$$f(x_2 \mid x_1) \sim N(\mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(x_1 - \mu_1), \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}).$$

如果用 Metropolis 算法选择 $q(X'|X^{(t)})$ 的话

$$X \mid X^{(t)} \sim N(\begin{bmatrix} X_1^{(t)} \\ X_2^{(t)} \end{bmatrix}, \begin{bmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{bmatrix}).$$

注意到我们是用的 Metropolis 算法，$q(X'|X) = q(X'|X)$，则接受率为

$$\alpha(X, X') = \min\{\frac{f(X')}{f(X)}, 1\} = \min\{\frac{\exp(-(X'-\mu)^\top \Sigma^{-1}(X'-\mu)/2)}{\exp(-(X-\mu)^\top \Sigma^{-1}(X-\mu)/2)}, 1\}.$$

多项分布例子

例

对多项分布

$$P\{X_1 = m_1, X_2 = m_2, \cdots, X_n = m_n\} = \frac{N!}{m_1! m_2! \cdots m_n!} p_1^{m_1} p_2^{m_2} \cdots p_n^{m_n},$$

$N = \sum_{i=1}^{n} m_i$. **某实验服从上述多项分布，** $N = 22$, $n = 7$, **7 个结果出现的概率分别为**

$$p := (p_1, p_2, \cdots, p_7) = (\frac{\theta}{4}, \frac{1}{8}, \frac{\theta}{4}, \frac{\eta}{4}, \frac{\eta}{4}, \frac{3}{8}, \frac{1}{2}(1 - \theta - \eta)).$$

现有观测数据为 $y = (y_1, y_2, y_3, y_4, y_5) = (14, 1, 1, 1, 5)$, 缺失数据为 $z = (z_1, z_2)$, 且

$$(z_1, y_1 - z_1, y_2, y_3, z_2, y_4 - z_2, y_5) \sim M(22; p),$$

其中 M 表示多项分布.

取平坦分布作为 (θ, η) 的先验分布，即 $\pi(\theta, \eta) \propto 1$. (y, z, θ, η) 的联合分布为(黑板推导)

$$\pi(y, z, \theta, \eta) \propto \left(\frac{\theta}{4}\right)^{z_1+y_2} \left(\frac{1}{8}\right)^{y_1-z_1} \left(\frac{\eta}{4}\right)^{z_2+y_3} \left(\frac{3}{8}\right)^{y_4-z_2} \left(\frac{1-\theta-\eta}{2}\right)^{y_5}. \quad (2)$$

参数 θ, η 的后验分布为

$$\pi(\theta, \eta \mid y, z) \propto \theta^{z_1+y_2} \eta^{y_3+z_2} (1-\theta-\eta)^{y_5}.$$

根据 (2), 可得如下满条件分布

$$\begin{aligned}
& \pi(\theta \mid y, z, \eta) \\
\propto \quad & \theta^{z_1+y_2}((1-\eta)-\theta)^{y_5} = (1-\eta)^{z_1+y_2+y_5}(\frac{\theta}{1-\eta})^{z_1+y_2}(1-\frac{\theta}{1-\eta})^{y_5} \\
\propto \quad & (\frac{\theta}{1-\eta})^{z_1+y_2}(1-\frac{\theta}{1-\eta})^{y_5} \sim (1-\eta)\mathrm{Beta}\left(z_1+y_2+1, y_5+1\right),
\end{aligned}$$

其中 $\mathrm{Beta}(\alpha, \beta)$ 的密度函数为:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in (0,1).$$

类似的，可得

$$
\begin{aligned}
\pi(\eta \mid y, z, \theta) &\propto \eta^{y_3+z_2}(1-\theta-\eta)^{y_5} \sim (1-\theta)\mathrm{Beta}\left(y_3+z_2+1, y_5+1\right) \\
\pi\left(z_1 \mid y, \theta, \eta\right) &\propto \left(\frac{\theta}{4}\right)^{z_1+y_2}\left(\frac{1}{8}\right)^{y_1-z_1} \propto \left(\frac{2\theta}{8}\right)^{z_1}\left(\frac{1}{8}\right)^{y_1-z_1} \\
&\propto \left(\frac{2\theta}{2\theta+1}\right)^{z_1}\left(\frac{1}{2\theta+1}\right)^{y_1-z_1} \sim \mathrm{Bino}(y_1, \frac{2\theta}{2\theta+1}) \\
\pi\left(z_2 \mid y, \theta, \eta\right) &\propto \left(\frac{\eta}{4}\right)^{z_2+y_3}\left(\frac{3}{8}\right)^{y_4-z_2} \propto \left(\frac{2\eta}{8}\right)^{z_2}\left(\frac{3}{8}\right)^{y_4-z_2} \\
&\propto \left(\frac{2\eta}{2\eta+3}\right)^{z_2}\left(\frac{3}{2\eta+3}\right)^{y_4-z_2} \sim \mathrm{Bino}(y_4, \frac{2\eta}{2\eta+3}).
\end{aligned}
$$

上式给出了参数和隐变量的更新公式，由此可进行 Gibbs 抽样.

上述证明有什么问题？(Hint: 归一化常数，先验)

分类消费者例子

数据：X_{gj} 代表第 j 个消费者在上个月，一共从第 g 个类别中买了 X_{gj} 个物品。

问题：怎么通过消费偏好，将消费者归为 K 类？

	Customer 1	c_2	c_3		c_j		c_n
I_1 food							
I_2 cloth							
…					X_{gj}		
I_G movie							

令 c_j 代表第 j 个消费者属于的类别，取值范围为 $1, 2, ..., K$.

简单来看，其实就是有 n 个样本，G 个特征，目标是将这些样本分成 K 组。

先设置缺失数据 c_j 的边际分布：

$$p(c_j = k) = \pi_k, \quad \sum_{k=1}^{K} \pi_k = 1,$$

这里 π_k 表示第 k 个类别的比例。再令

$$X_{gj} \mid c_j = k \sim Pois(\lambda_{gk}),$$

其中 λ_{gk} 代表第 k 类的消费者买第 g 类物品的均值 (期望)。

模型：

$$c_j \sim Categorical(\pi_1, ..., \pi_K), \quad j = 1, ..., n,$$
$$X_{gj} \mid c_j = k \sim Pois(\lambda_{gk}), \quad g = 1, ..., G, \quad j = 1, ..., n.$$

Q1: 属于 k 组的消费者比例是多少？[parameter]

$$\pi = (\pi_1, ..., \pi_K).$$

Q2: k 组的消费者上个月平均购买了多少件属于 g 类别的商品？[parameter]

$$\Lambda = \{\lambda_{gk}\}, \quad g = 1, ..., G, \quad k = 1, ..., K.$$

Q3: 第 j 个消费者属于哪组？[missing data]

$$c_j, \quad j = 1, ..., n.$$

观测数据：

$$\{X_{gj}\}, \quad j = 1, ..., n, \quad g = 1, ..., G.$$

完整数据的似然函数为：(一般需要先设置隐变量的边际分布和观测数据基于隐变量的条件分布才能得到)

$$f(x, c|\pi, \Lambda) = \prod_{j=1}^{n} \prod_{k=1}^{K} \left[\pi_k \prod_{g=1}^{G} \frac{\lambda_{gk}^{x_{gj}}}{x_{gj}!} e^{-\lambda_{gk}} \right]^{I(c_j=k)}$$

在此章的推断 (inference) 中，我们使用 Bayesian 的框架，

$$\text{Prior} + \text{Likelihood} \longrightarrow \text{Posterior}.$$

我们设定参数的共轭先验分布为 Dirichlet 分布 (多元 Beta 分布):

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

$$p(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_K)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

这里的 $\alpha_1, ..., \alpha_K$，我们称他为超参数 (hyperparameters).

Dirichlet 分布的期望是

$$E(\pi_k) = \frac{\alpha_k}{\alpha_1 \cdots + \alpha_K} = \frac{\alpha_k}{\sum_l \alpha_l}.$$

它的最大概率值点 (mode) 是

$$\tilde{\pi_k} = \frac{\alpha_k - 1}{\sum_l \alpha_l - K}, \quad \alpha_k > 1$$

当我们有先验信息的时候，可以根据信息设置 α_k 的值，否则一般都设置为 1. 当 α_k 都为 1 时，dirichlet 分布退化成均匀分布.

若数据服从 Poisson 分布，则参数的共轭先验是 gamma 分布，即

$$\lambda_{gk} \sim Gamma(\alpha, \beta), \quad p(\lambda_{gk}) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} (\lambda_{gk})^{\alpha-1} e^{-\beta\lambda_{gk}}, \quad E(\lambda_{gk}) = \frac{\alpha}{\beta}$$

我们推出差个比例常数的联合后验分布是

$$p(\pi, \Lambda | c, x) \propto f(x, c | \pi, \Lambda) p(\pi) p(\Lambda) \propto f(x, c | \pi, \Lambda) p(\pi) \prod_{g=1}^{G} \prod_{k=1}^{K} p(\lambda_{gk})$$

$$\propto \prod_{j=1}^{n} \prod_{k=1}^{K} \left[\pi_k \prod_{g=1}^{G} \frac{(\lambda_{gk})^{x_{gj}}}{x_{gj}!} e^{-\lambda_{gk}} \right]^{\mathbb{I}\{c_j = k\}} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \prod_{g=1}^{G} \prod_{k=1}^{K} (\lambda_{gk})^{\alpha-1} e^{-\beta\lambda_{gk}}$$

全条件后验

我们首先推导 π_k 的全条件后验分布:

$$p(\pi|-) \propto \prod_{j=1}^{n}\prod_{k=1}^{K}(\pi_k)^{\mathbb{I}\{c_j=k\}}\prod_{k=1}^{K}(\pi_k)^{\alpha_k-1}$$

$$\sim Dirichlet\left(\sum_{j=1}^{n}\mathbb{I}\{c_j=1\}+\alpha_1,\cdots,\sum_{j=1}^{n}\mathbb{I}\{c_j=K\}+\alpha_K\right)$$

只要看联合分布里面跟 π 相关的部分就行.

再推导 λ_{gk} 的全条件后验分布:

$$
\begin{aligned}
p\left(\lambda_{gk} \mid -\right) &\propto \prod_{j=1}^{n}\left[\left(\lambda_{gk}\right)^{x_{gj}} e^{-\lambda_{gk}}\right]^{\mathbb{I}\{c_j=k\}}\left(\lambda_{gk}\right)^{\alpha-1} e^{-\beta\lambda_{gk}} \\
&\propto \left(\lambda_{gk}\right)^{\sum_{j=1}^{n} x_{gj}\mathbb{I}\{c_j=k\}} e^{-\sum_{j=1}^{n}\lambda_{gk}\mathbb{I}\{c_j=k\}}\left(\lambda_{gk}\right)^{\alpha-1} e^{-\beta\lambda_{gk}} \\
&\propto \left(\lambda_{gk}\right)^{\sum_{j=1}^{n}\mathbb{I}\{c_j=k\}x_{gj}+\alpha-1} e^{-\left(\sum_{j=1}^{n}\mathbb{I}\{c_j=k\}+\beta\right)\lambda_{gk}} \\
&\sim Gamma\left(\sum_{j=1}^{n}\mathbb{I}\left\{c_j=k\right\}x_{gj}+\alpha, \sum_{j=1}^{n}\mathbb{I}\left\{c_j=k\right\}+\beta\right)
\end{aligned}
$$

然后是 c_j:

$$
p\left(c_j=k \mid -\right)=\frac{\pi_k \prod_{g=1}^{G}\frac{\left(\lambda_{gk}\right)^{x_{gj}}}{x_{gj}!}e^{-\lambda_{gk}}}{\sum_{\ell=1}^{K}\pi_\ell \prod_{g=1}^{G}\frac{\left(\lambda_{g\ell}\right)^{x_{gj}}}{x_{gj}!}e^{-\lambda_{g\ell}}}.
$$

实际迭代

给定 $\pi^{(t)}, \Lambda^{(t)}, c^{(t)}$,

$$\pi_1^{(t+1)}, .., \pi_k^{(t+1)}|- \sim \text{Dirichlet}\left(\alpha_1 + \sum_{j=1}^{n} \mathbb{I}(c_j^{(t)} = 1), \ldots, \alpha_k + \sum_{j=1}^{n} \mathbb{I}(c_j^{(t)} = 1)\right),$$

$$\lambda_{gk}^{(t+1)}\Big| \sim \text{Gamma}\left(c + \sum_{j=1}^{n} \mathbb{I}\left(c_{j=1}^{(t)}\right) x_{gj}, d + \sum_{j=1}^{n} \mathbb{I}\left(c_j^{(t)}\right)\right),$$

$$p(c_j^{(t+1)} = k|-) = \frac{\pi_k^{(t+1)} \prod_{g=1}^{G} (\lambda_{gk}^{(t+1)})^{x_{gj}} \exp\{-\lambda_{gk}^{(t+1)}\}}{\sum_l \pi_l^{(t+1)} \prod_{g=1}^{G} (\lambda_{gl}^{(t+1)})^{x_{gj}} \exp\{-\lambda_{gl}^{(t+1)}\}}$$

Summary

实施

本节将研究链的长期运行的表现问题. 例如,

- 链是否已经运行地足够长了;
- 链的前面部分是否受初始值的强烈影响;
- 是否该使用多个不同的初始值来运行;
- 如何用链的输出得到估计并衡量其近似精度, 等等.

混合和收敛

在马尔可夫链蒙特卡罗（Markov Chain Monte Carlo，MCMC）算法中，混合（mixing）和收敛（convergence）是相关但不同的概念。

混合指的是马尔可夫链在状态空间中有效地探索和转移的能力。

混合良好的链可以自由、快速地在状态空间中移动，访问不同的区域并从目标分布的不同部分进行采样。这表明马尔可夫链能够高效地探索分布并生成代表性的样本。

相反，混合不佳意味着链在某些区域陷入困境，无法充分探索分布的整个范围，并可能产生偏误或不准确的样本。(样本自相关性太强)

收敛是指马尔可夫链随着迭代次数增加，逐渐接近并稳定在目标分布周围的特性。收敛意味着马尔可夫链生成的样本随着算法的进行越来越能够代表目标分布。它表明算法已经达到一种状态，在进一步迭代中估计到的分布不会显著改变。

需要注意的是，混合是收敛的前提条件。如果马尔可夫链混合不好，它将无法收敛到目标分布。然而，即使链混合良好，也不能保证收敛。收敛需要良好的混合以及足够的迭代次数，以确保链充分探索状态空间并稳定在目标分布周围。

相关术语

1. 预烧 (burn-in): 我们通常假定 MCMC 要经过一段时间的迭代才能收敛到平稳分布，这段过程我们称为 burn-in.
2. 轨迹图 (trace plot): 画出每次迭代时参数的值.
3. 对数似然函数或后验分布函数图：随着迭代，对数似然函数的变化.
4. 多链: 使用不同初值的多条短链画出变量的轨迹图，观测 f 的主要特征 (比如多峰，高度集中的支撑域). 之后选取一个好的初始值，运作一个相当长的单链计算并公布结果. (一般由最终 likelihood 大小筛选)
5. 自相关性图: 描述样本序列在不同迭代延迟下的相关性.

$$\rho_k = \frac{\sum_{i=k+1}^{n}(x_i - \bar{X})(x_{i-k} - \bar{X})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

Gelman-Rubin 统计量

为确定预烧期和运行长度，Gelman 和 Rubin 提出一种统计量判断 MCMC 是否已经收敛到平稳分布.

假设感兴趣的变量是 X. 其中 $x_1^{(j)}, x_2^{(j)}, \ldots$ 是第 j 个马尔可夫链的样本，并假设有 J 个链并行运行.

- 对于每个链，首先丢弃 D 值作为“预烧”并保留剩余的 L 值，$x_D^{(j)}, x_{D+1}^{(j)}, \ldots x_{D+L-1}^{(j)}$.

- 计算:

$$\bar{x}_j = \frac{1}{L}\sum_{t=1}^{L} x_t^{(j)} \quad \text{(chain mean)}, \quad \bar{\boldsymbol{x}}_\cdot = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j \quad \text{(grand mean)}$$

$$\mathrm{B} = \frac{L}{J-1}\sum_{j=1}^{J}(\bar{x}_j - \bar{x}_\cdot)^2 \quad \text{(between chain variance)}$$

$$s_j^2 = \frac{1}{L-1}\sum_{t=1}^{L}(x_t^{(j)} - \bar{x}_j)^2 \quad \text{(within chain variance)}$$

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

- Gelman-Rubin 统计量是

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$$

.

我们能看到当 $L \to \infty$ 并且 $B \to 0$ 时，R 是趋近于 1 的. 实际应用中，某些学者建议可以接受 $\sqrt{R} < 1.2$.

但使用这种方法有一些潜在的困难. 当 f 是多峰分布的情况下, 如何选择合适的初始值也许较为困难, 如果选择不恰当, 则会导致大部分的链都长期停留在同样的子域或者峰的附近.

稳妥方法：结合轨迹图和对数似然函数图，多条链进行肉眼观测分析.

Geweke 检验

Geweke（1992）提出一种基于样本序列前后段统计量比较的 MCMC 收敛诊断方法，该方法适用于单条链。

- 基本思想：如果马尔可夫链已收敛到平稳分布，则其前段与后段的样本应来自同一分布。
- 做法是将一条长链 (burn-in 后的) 分成前段（如前 10%）和后段（如后 50%），然后对感兴趣的统计量（如均值）计算两个子样本的均值差，并标准化为 Z 分数。

Geweke 检验的步骤

1. 将马尔可夫链分为两个不重叠的部分：
   - 前段：通常取前 10
   - 后段：通常取后 50
2. 分别计算两个子样本的均值与方差。
3. 计算标准化差异（Geweke Z-score）：

$$Z = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\hat{S}_a^2/n_a + \hat{S}_b^2/n_b}}$$

其中：

- $\bar{X}_a$, $\bar{X}_b$ 分别为前段与后段的样本均值；
- $\hat{S}_a^2$, $\hat{S}_b^2$ 为对应的谱方差估计；
- n_a, n_b 为对应的样本大小。

谱方差估计（Spectral Variance Estimation）

在 MCMC 中，样本之间通常存在自相关，不能简单使用样本方差。

- 设序列为平稳过程，其样本均值的方差为：

$$\mathrm{Var}(\bar{x}) = \frac{1}{n^2}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{Cov}(x_t, x_s) = \frac{1}{n}\left[\gamma_0 + 2\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\gamma_k\right]$$

  其中 γ_k 是延迟 k 的自协方差：$\gamma_k := \mathrm{Cov}(x_t, x_{t+k})$

- 当样本足够多（$n \to \infty$）时，上式趋近于：

$$\mathrm{Var}(\bar{x}) := n\,\mathrm{Var}(\bar{x}) = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k$$

谱方差是样本均值方差乘以 n（跟原来的样本方差的计算过相似）。

我们乘以 n 是因为我们要从" 样本均值的方差" 推回到" 原序列的有效方差"，也就是谱方差。它等价于假设每个样本是独立的情况下，你需要多少个样本才能达到当前样本均值的方差。

谱方差还可以用于估计有效样本数（Effective Sample Size, ESS），它衡量相关样本等价于多少个独立样本：

$$\text{ESS} = \frac{n \cdot \gamma_0}{\sigma^2_{\text{spectral}}}$$

说明：自相关越强，谱方差越大，ESS 越小，信息越冗余。

为了在有限样本下估计谱方差，常采用以下方法：

1. **窗口函数法（Windowed Estimator）**：
   - 常用 Bartlett 窗口（线性衰减）：

$$\hat{\sigma}^2 = \hat{\gamma}_0 + 2\sum_{k=1}^{M}\left(1 - \frac{k}{M+1}\right)\hat{\gamma}_k$$

   - M 是截断点，控制考虑的最大延迟阶数。
2. **批量均值法（Batch Means）**：
   - 将样本划分为 b 个 batch，每个 batch 大小为 m。
   - 计算每批均值 $\bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_b^*$。
   - 用这些均值计算方差，再乘以 m 得到估计：

$$\hat{\sigma}^2 = m \cdot \frac{1}{b-1}\sum_{i=1}^{b}(\bar{x}_i^* - \bar{x})^2$$

Geweke Z 分数的解释

- 如果链已收敛，则 Z 应服从标准正态分布。
- 可以画出多个参数的 Z 分数，观察其是否大致落在 $[-1.96, 1.96]$ 区间内。
- 若大部分 Z 值超出此区间，说明链的前后段来自不同分布，可能尚未收敛。
- 注意：谱方差估计需要考虑样本之间的自相关性，不能简单使用样本方差。

Geweke 检验的优缺点

- 优点：
  - 实现简单，适用于单链；
  - 可用于诊断特定参数的收敛性。
- 缺点：
  - 结果依赖于子样本的选取比例；
  - 不能检测多峰分布下的全局混合性；
- 建议与轨迹图、对数似然曲线图、Gelman-Rubin 诊断等方法联合使用。