第四章 Gibbs 抽样和马尔可夫链

Gibbs 抽样

- Gibbs 抽样是一种迭代抽样算法，随着样本数增加，样本的分布会收敛到目标分布，首先以一个 Bayesian 模型的估计为例介绍如何使用 Gibbs 抽样

- 使用 Bayesian 方法分析数据三要素:
  1. 模型设定：为数据的抽样分布设定具体形式 $p(\boldsymbol{y} \mid \boldsymbol{\theta})$，通常需要引入一些参数 $\boldsymbol{\theta}$
  2. 设定参数的先验分布 $p(\boldsymbol{\theta})$：一般是主观设定，可以加入参数的先验信息，其样本空间应覆盖参数所有可能的取值
  3. 计算参数的后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 并做统计推断：估计参数的后验期望 $E(\boldsymbol{\theta} \mid \boldsymbol{y})$、后验方差 $Var(\boldsymbol{\theta} \mid \boldsymbol{y})$、置信区间等. 参数的后验分布可如下计算:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta}) \tag{1}$$

但 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 对应的分布一般很难识别或很难直接对其抽样

A Bayesian Normal Model

假设数据独立地服从正态分布:

$$Y_i \overset{iid}{\sim} N(\mu, \phi^{-1}), \quad i = 1, \ldots, n \tag{2}$$

参数的 (共轭) 先验分布:

$$\phi \sim \text{Gam}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$$

$$\mu \mid \phi \sim N(\mu_0, \frac{1}{\kappa_0 \phi})$$

其中 ν_0, σ_0^2, μ_0, κ_0^2 都是确定的常数

A Bayesian Normal Model

参数 (μ, ϕ) 的联合后验分布仍是一个 normal-gamma 分布：

$$\phi \mid y_1, \ldots, y_n \sim \text{Gam}(\frac{\nu_n}{2}, \frac{S_n}{2})$$

$$\mu \mid \phi, y_1, \ldots, y_n \sim N(\mu_n, \frac{1}{\kappa_n \phi})$$

其中

$$\nu_n = \nu_0 + n$$

$$S_n = \nu_0 \sigma_0^2 + \frac{n\kappa_0}{\kappa_0 + n}(\mu_0 - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_0 + n}$$

$$\kappa_n = \kappa_0 + n$$

- 计算 $E(\mu \mid y_1, \ldots, y_n)$，可以采用如下的 Monte Carlo 方法：
  (1) 独立抽样 $\phi^{(s)} \sim \text{Gam}(\nu_n/2, S_n/2)$, $s = 1, \ldots, T$
  (2) 对每个 $\phi^{(s)}$, 抽 $\mu^{(s)} \mid \phi^{(s)} \sim N(\mu_n, (\kappa_n \phi^{(s)})^{-1})$, $s = 1, \ldots, T$

  则 $E(\mu \mid y_1, \ldots, y_n) \approx \sum_{s=1}^{T} \mu^{(s)}/T$

A Bayesian Normal Model

如果为模型(2)选取如下的先验分布:

$$\begin{aligned} \mu &\sim N(\mu_0, \tau_0^2) \\ \phi &\sim \text{Gam}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned} \tag{3}$$

- 此时 ϕ 的边际后验分布既不是 Gamma 分布，也不是任何常见的分布
- 方法一：使用数值方法得到 (μ, ϕ) 的近似联合后验分布
  - 首先为各参数选取足够大的取值范围 $\mu \in [\mu_L, \mu_H]$, $\phi \in [\phi_L, \phi_H] \subseteq (0, \infty)$
  - 然后对区域 $[\mu_L, \mu_H] \times [\phi_L, \phi_H]$ 做网格离散
  - 根据(1)，点 (μ_i, ϕ_j) 处的后验概率密度为

$$p(\mu_i, \phi_j \mid y_1, \ldots, y_n) \propto p(\mu_i, \phi_j) p(y_1, \ldots, y_n \mid \mu_i, \phi_j)$$

  - 网格中大部分点的后验概率都很接近 0，造成计算的浪费，且该方法只适用于参数较少的情况

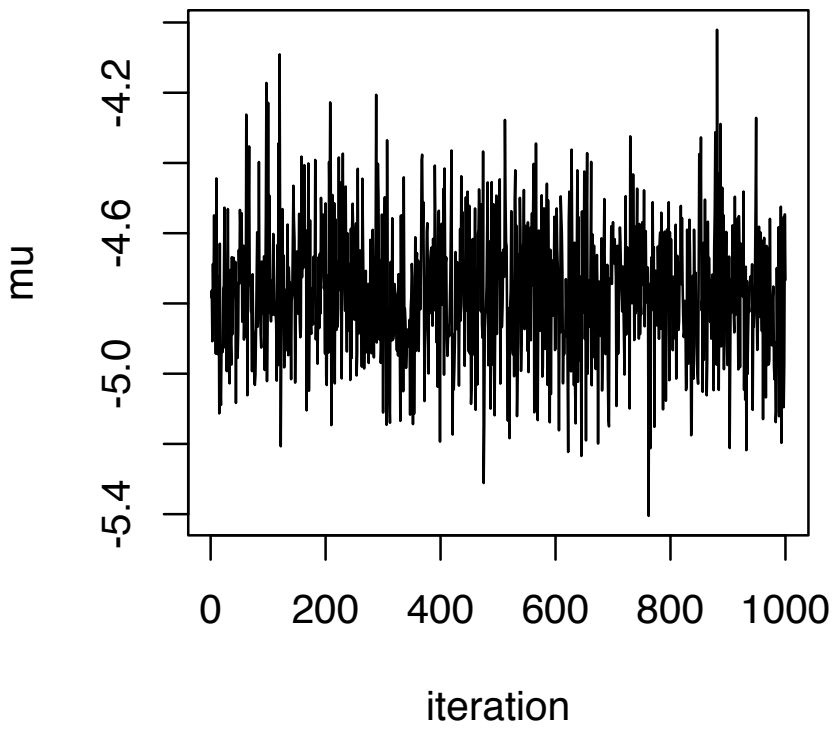
A Bayesian Normal Model

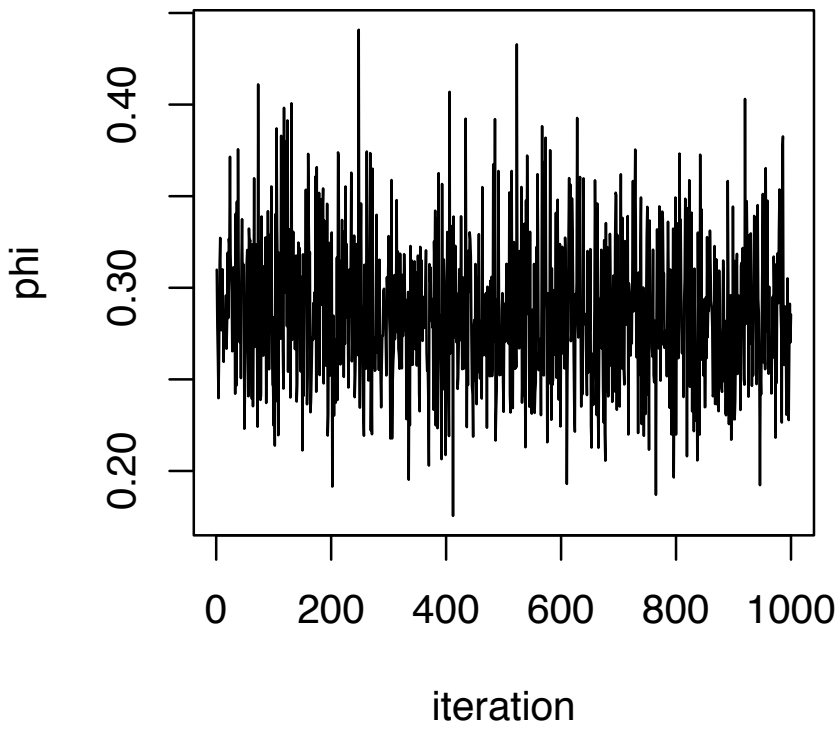
- 方法二：Gibbs 抽样
  - 模型(2) - (3) 中 μ 的完全条件分布为：$\mu \mid \phi, y_1, \ldots, y_n \sim N(\mu_n, \tau_n^2)$，其中 $\mu_n = (\tau_0^{-2}\mu_0 + \phi\sum_{i=1}^n y_i)/(\tau_0^{-2} + n\phi)$, $\tau_n^2 = (\tau_0^{-2} + n\phi)^{-1}$
  - 模型(2) - (3) 中 ϕ 的完全条件分布为：$\phi \mid \mu, y_1, \ldots, y_n \sim \text{Gam}(\nu_n/2, S_n/2)$, 其中 $\nu_n = \nu_0 + n$, $S_n = \nu_0\sigma_0^2 + \sum_{i=1}^n (y_i - \mu)^2$
  - 如何利用 μ 和 ϕ 的完全条件分布得到 (μ, ϕ) 的联合后验分布的样本？
    - ★ 假设 $\phi^{(1)}$ 是边际后验分布 $p(\phi \mid y_1, \ldots, y_n)$ 的一个样本，给定 $\phi^{(1)}$，从 μ 的完全条件分布抽样：

      $$\mu^{(1)} \sim p(\mu \mid \phi^{(1)}, y_1, \ldots, y_n)$$

      则 $(\mu^{(1)}, \phi^{(1)})$ 可以看作联合后验分布 $p(\mu, \phi \mid y_1, \ldots, y_n)$ 的一个样本
  - 只要给定 μ 或 ϕ 的一个初始值，然后轮流从 μ 和 ϕ 的完全条件分布抽样，就可以得到一列来自联合后验分布 $p(\mu, \phi \mid y_1, \ldots, y_n)$ 的样本 $\{(\mu^{(s)}, \phi^{(s)}) : s = 1, \ldots, T\}$

A Bayesian Normal Model





样本在参数空间中“移动”的很快，表明样本之间的相关性较小，此时样本均值可以很好地近似后验分布的期望

A Bayesian Normal Model

- 样本之间的相关性如何影响样本均值对目标期望的近似？
  - 假设 $\{\theta^{(s)} : s = 1, \ldots, T\}$ 是由 Gibbs 抽样得到的一列服从目标分布 $p(\theta)$ 的样本，样本均值 $\bar{\theta}$ 的方差为

$$Var_G(\bar{\theta}) = E\left[\left(\bar{\theta} - E(\theta)\right)^2\right]$$
$$= Var_{MC}(\bar{\theta}) + \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t\neq s} E\left[\left(\theta^{(s)} - E(\theta)\right)\left(\theta^{(t)} - E(\theta)\right)\right]$$

  - $Var_G(\bar{\theta}) > Var_{MC}(\bar{\theta})$，且 Gibbs 样本之间的相关性越高，$Var_G(\bar{\theta})$ 就越大，均值的近似效果越差

- 一个衡量 Gibbs 样本相关性的指标是有效样本数 (ESS)：

$$ESS = \frac{Var(\theta)}{Var_G(\bar{\theta})}$$

为达到与 Gibbs 样本估计量相同精度所需的独立样本个数

Gibbs 抽样

- 假设模型的参数向量为 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, 抽样的目标分布为 $p(\boldsymbol{\theta})$, 如果知道每个分量 θ_j 的完全条件分布 $p(\theta_j \mid \{\theta_k\}_{k \neq j})$, $j = 1, \ldots, p$, 给定初始值 $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$, Gibbs 抽样可以如下从当前样本 $\boldsymbol{\theta}^{(s-1)}$ 产生新的样本 $\boldsymbol{\theta}^{(s)}$:

1. 抽样 $\theta_1^{(s)} \sim p(\theta_1 \mid \theta_2^{(s-1)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)})$
2. 抽样 $\theta_2^{(s)} \sim p(\theta_2 \mid \theta_1^{(s)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)})$
3. 抽样 $\theta_3^{(s)} \sim p(\theta_3 \mid \theta_1^{(s)}, \theta_2^{(s)}, \theta_4^{(s-1)}, \ldots, \theta_p^{(s-1)})$

   $\vdots$

p. 抽样 $\theta_p^{(s)} \sim p(\theta_p \mid \theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_{p-1}^{(s)})$

Gibbs 抽样

- 不断重复上述过程，Gibbs 抽样可以产生一列不独立的样本 $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)}$，且每个样本 $\boldsymbol{\theta}^{(s)}$ 与之前的样本 $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(s-1)}$ 的相关性只取决于 $\boldsymbol{\theta}^{(s-1)}$，因此这一列样本是一条马尔可夫链
- 在满足一些条件后，从任何初始值 $\boldsymbol{\theta}^{(0)}$ 出发，Gibbs 抽样产生的样本 $\boldsymbol{\theta}^{(s)}$ 的分布在 $s \to \infty$ 时收敛到目标分布 $p(\boldsymbol{\theta})$，即

$$P(\boldsymbol{\theta}^{(s)} \in A) \to \int_A p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

- 此时对任意可积函数 g 有

$$\frac{1}{T} \sum_{s=1}^{T} g(\boldsymbol{\theta}^{(s)}) \to E[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad T \to \infty$$

上述过程被称为Markov chain Monte Carlo (MCMC) 方法

马尔可夫链

- 给定初始值 $\boldsymbol{x}^{(0)} \in \mathcal{X}$，按照某个条件分布 $p(\boldsymbol{x} \mid \boldsymbol{x}')$ 依次抽样

$$\boldsymbol{x}^{(t)} \sim p(\boldsymbol{x} \mid \boldsymbol{x}^{(t-1)}) \quad \text{且} \quad \boldsymbol{x}^{(t)} \perp\!\!\!\perp \boldsymbol{x}^{(t-k)} \mid \boldsymbol{x}^{(t-1)},\ k > 1$$

  称这样得到的序列 $\{\boldsymbol{x}^{(t)}\}$ 为状态空间 $\mathcal{X}$ 上的一条马尔可夫链

- 在马尔可夫链中，对任意正整数 k，定义如下的 k 步转移密度函数：
  - $p^1(\boldsymbol{x}^{(t+1)} \mid \boldsymbol{x}^{(t)}) = p(\boldsymbol{x}^{(t+1)} \mid \boldsymbol{x}^{(t)})$
  - $p^2(\boldsymbol{x}^{(t+2)} \mid \boldsymbol{x}^{(t)}) = \int_{\mathcal{X}} p(\boldsymbol{x}^{(t+2)} \mid \boldsymbol{x}^{(t+1)}) p(\boldsymbol{x}^{(t+1)} \mid \boldsymbol{x}^{(t)}) d\boldsymbol{x}^{(t+1)}$
  - $p^3(\boldsymbol{x}^{(t+3)} \mid \boldsymbol{x}^{(t)}) = \int_{\mathcal{X}} p(\boldsymbol{x}^{(t+3)} \mid \boldsymbol{x}^{(t+2)}) p^2(\boldsymbol{x}^{(t+2)} \mid \boldsymbol{x}^{(t)}) d\boldsymbol{x}^{(t+2)}$
  - $\vdots$
  - $p^k(\boldsymbol{x}^{(t+k)} \mid \boldsymbol{x}^{(t)}) = \int_{\mathcal{X}} p(\boldsymbol{x}^{(t+k)} \mid \boldsymbol{x}^{(t+k-1)}) p^{k-1}(\boldsymbol{x}^{(t+k-1)} \mid \boldsymbol{x}^{(t)}) d\boldsymbol{x}^{(t+k-1)}$

  如果 $p(\boldsymbol{x} \mid \boldsymbol{x}') > 0,\ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$，那么对任意正整数 k，k 步转移密度函数 $p^k(\boldsymbol{x} \mid \boldsymbol{x}') > 0,\ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$

马尔可夫链

- 如果存在一个分布 $\pi(\boldsymbol{x})$ 使马尔可夫链满足

$$\pi(\boldsymbol{x}) = \int_{\mathcal{X}} p(\boldsymbol{x} \mid \boldsymbol{x}')\pi(\boldsymbol{x}')d\boldsymbol{x}'$$

  称分布 $\pi(\boldsymbol{x})$ 为该马尔可夫链的一个平稳分布

  - 如果马尔可夫链的初始值服从平稳分布, 整个过程的边际分布会永远“保持” 平稳分布

$$\pi(\boldsymbol{x}) = \int_{\mathcal{X}} p^k(\boldsymbol{x} \mid \boldsymbol{x}')\pi(\boldsymbol{x}')d\boldsymbol{x}', \quad \forall k \in \mathbb{N}^+$$

  - Gibbs 抽样产生的马尔可夫链的转移密度函数是

$$p(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{\theta}^{(s-1)}) = p(\theta_1^{(s)} \mid \theta_2^{(s-1)}, \ldots, \theta_p^{(s-1)}) \cdots p(\theta_p^{(s)} \mid \theta_1^{(s)}, \ldots, \theta_{p-1}^{(s)})$$

    如果 $\forall \boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^{(s)}$, $p(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{\theta}^{(s-1)}) > 0$, 该过程会收敛到唯一的平稳分布 (参数的后验分布)

马尔可夫链

- 如果从任意初始状态 $\boldsymbol{x}' \in \mathcal{X}$ 出发，马尔可夫链可以在有限步到达任意其它状态 $\boldsymbol{x} \in \mathcal{X}$，称马尔可夫链是不可约的 (irreducible)

定义 (不可约性)

如果 $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $\exists k < \infty$ 使得 k 步转移密度 $p^k(\boldsymbol{x} \mid \boldsymbol{x}') > 0$, 称该马尔可夫链是不可约的.

- 如果 $p(\boldsymbol{x} \mid \boldsymbol{x}') > 0$, $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, 那么该马尔可夫链是不可约的
- **一个可约的马尔可夫过程**. 令状态空间 $\mathcal{X} = (-1, 1)$ 上的一个马尔可夫链有如下的转移分布:

$$p(x \mid x') = \begin{cases} x \sim U(0,1), & \text{if } x' \geq 0 \\ x \sim U(-1,0), & \text{if } x' < 0 \end{cases}$$

马尔可夫链

- 如果 $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $\gcd\{k : p^k(\boldsymbol{x} \mid \boldsymbol{x}') > 0\} = 1$, 称该过程具有非周期性
  - 当 $p(\boldsymbol{x} \mid \boldsymbol{x}') > 0, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$，马尔可夫链是非周期的
  - **一个周期性的马尔可夫链**. 令状态空间 $\mathcal{X} = (-1, 1)$ 上的一个马尔可夫链有如下的转移分布：

$$p(x \mid x') = \begin{cases} x \sim U(0,1), & \text{if } x' < 0 \\ x \sim U(-1,0), & \text{if } x' \geq 0 \end{cases}$$

- 称一个不可约且非周期的马尔可夫链具有遍历性(Ergodicity), 具有遍历性的马尔可夫链有以下重要性质：
  1. 存在唯一的平稳分布 $\pi(\boldsymbol{x})$
  2. 从任意初始值出发，该过程都会收敛到平稳分布 $\pi(\boldsymbol{x})$, 即对 $\forall \boldsymbol{x}' \in \mathcal{X}$, 当 $k \to \infty$,

$$p^k(\boldsymbol{x} \mid \boldsymbol{x}') \to \pi(\boldsymbol{x})$$

马尔可夫链

- 对遍历的马尔可夫链，当 $k \to \infty$，该过程产生的 $\boldsymbol{x}^{(k)} \sim \pi(\boldsymbol{x})$
- 此时样本均值会 (almost surely) 收敛到平稳分布的期望，且对任何可积函数 $g(\cdot)$，当 $T \to \infty$，

$$\frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{x}^{(t)}) \to \int g(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x}$$

- 在 Bayesian 分析中通过 Gibbs 抽样产生的马尔可夫链一般是遍历的，通过产生一条很长的马尔可夫链，该过程的样本可以近似描述后验分布 (平稳分布)
- 如果初始值选取得不好，马尔可夫链可能会移动很长时间才收敛到平稳分布概率密度较高的区域，称这段时间为burn-in，用样本均值估计目标期望时通常舍弃 burn-in 阶段的样本