统计学笔记

Ruijun Shi[1]

2021 年 11 月 21 日

[1] GitHub：https://github.com/RuijunShi

摘要

简单的统计学笔记，主要是在天文学，特别是引力波和pulsar timing中遇到的统计学，现在没写多少内容。缓慢更新中[1]。同时这也是我第一次用latex编写书籍，学习过程艰难啊。有错误请大家指出！
latex资料参考https://github.com/Ali-loner

Contents

Chapter 1

数学和物理基础

未完待续。。。图片测试 SVD分解 1.1

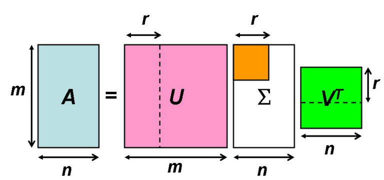

图 1.1: tSVD分解

1.1 线性变换

1.1.1 线性相关

1.1.2 特征值与特征向量

1.1.3 线性代数几何意义

正交矩阵

1.2 误差分析

1.3 特殊函数

1.4 偏微分方程

1.5 玻尔兹曼分布

Chapter 2

数理统计基础

2.1 概率

1. 概率的定义（略）概率满足：非负性，规范性，可列可加性

2. 概率的性质：

   重点：逆事件概率；加法公式；有限可加性

3. 条件概率：

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{2.1}$$

4. 乘法定理：

$$P(AB) = P(A|B)P(A) \tag{2.2}$$

5. 全概率公式：

$$P(A) = P(A \mid B_1) P(B_1) + P(A \mid B_2) P(B_2) + \ldots + P(A \mid B_n) P(B_n) \tag{2.3}$$

6. 独立性：满足

$$P(AB) = P(A)P(B) \tag{2.4a}$$

$$P(B|A) = P(B) \tag{2.4b}$$

$$\tag{2.4c}$$

2.2 单变量分布和多变量分布

1. 随机变量的概念（略）

2. 分布函数的概念（略）和性质：不减函数；$0 \le F(x) \le 1$; $F(x+0) = F(x)$

3. 概率密度函数

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{dt} \tag{2.5}$$

性质 2.1 概率分布的性质

$$f(x) \le 0 \tag{2.6a}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{2.6b}$$

$$P\{x_1 < X < x_2\} = \int_{x_1}^{x_2} f(x)dx \tag{2.6c}$$

$$F'(x) = f(x) \tag{2.6d}$$

2.3 随机变量的数学特征

2.3.1 数学期望与方差

定义 2.1 数学期望
积分*:*

$$E(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x \tag{2.7}$$

为连续性随机变量的数学期望，离散状态下为*:*

$$E(X) = \sum_{k=1}^{\infty} x_k p_k \tag{2.8}$$

性质 2.2 方差的性质

- 设C为常数，则有$E(C) = C$
- 设C为常数，X为随机变量，则有
$$E(CX) = CE(X)$$
- 设X, Y两个随机变量，则有：
$$E(X+Y) = E(X) + E(Y)$$
- 设X, Y是两个相互独立的随机变量，则有：
$$E(XY) = E(X) + E(Y)$$

定义 2.2 方差
设X是一个随机变量，若$E\{[X-E(X)]^2\}$存在，则称为$E\{[X-E(X)]^2\}$为随机变量X的方差，记为$D(X)$或者$\mathrm{Var}(X)$

根据定义，我们把方差写为：

$$D(X) = \int_{-\infty}^{\infty} [x - E(x)]^2 f(x) \mathrm{d}x \tag{2.9}$$

随机变量的方差可以写为：

$$D(X) = E(X^2) - [E(X)]^2 \tag{2.10}$$

性质 2.3 方差的性质

- 设C为常数：*D(C)=0*
- 设C为常数，X为随机变量，有：

$$D(CX) = C^2 D(X), \qquad D(X+C) = D(X)$$

- 设X, Y为两个随机变量，有：

$$D(X+Y) = D(X) + D(Y) + 2E\{(X - E(X))(Y - E(Y))\}$$

  若X, Y相互独立，则有：

$$D(X+Y) = D(X) + D(Y)$$

- $D(X) = 0$的充要条件是X以概率为*1*取常数$E(X)$，即：

$$P\{X = E(X)\} = 1$$

2.3.2 矩和协方差矩阵

定义 2.3 协方差 随机变量$E\{(X - E(X))(Y - E(Y))\}$称为变量$X, Y$的协方差，记为$\mathrm{Cov}(X, Y)$：

$$\mathrm{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\} \tag{2.11}$$

协方差是变量误差的一种描述。若随机变量X,Y完全独立则有$\mathrm{Cov}(X, Y) = 0$

定义 2.4 相关系数

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{2.12}$$

当二维随机变量的二阶中心矩存在：

$$\begin{aligned}
c_{11} &= E\{[X_1 - E(X_1)]^2\} \\
c_{12} &= E\{[X_1 - E(X_1)][X_2 - E(X_2)]\} \\
c_{21} &= E\{[X_2 - E(X_2)][X_1 - E(X_1)]\} \\
c_{22} &= E\{[X_2 - E(X_2)]^2\}
\end{aligned} \tag{2.13}$$

则矩阵

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

称为协方差矩阵。若有n维随机变量，则矩阵：

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \tag{2.14}$$

该矩阵是一个对称矩阵。

2.3.3 多元正态分布及协方差矩阵的直观理解

协方差矩阵描述随机变量的总体误差，而方差是协方差的一种特殊形式。协方差可以用多元正态分布直观理解其意义。

2.4 常见分布及其期望方差

2.4.1 离散型分布

1. (0-1)分布:

$$P(X = k) = p^k(1-p)^{1-k}, 0 < p < 1, k = 0, 1 \tag{2.15}$$

2. 二项分布:

$$P(X = k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k} \tag{2.16}$$

3. 泊松分布:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, \ldots \tag{2.17}$$

2.4.2 连续型分布

4. Beta分布:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{2.18}$$

其中：$0 \le x \le 1,\ \alpha > 0,\ \beta > 0,\ \Gamma(z) = \int_0^{+\infty} t^{z-1}e^{-t}dt$

5. 均匀分布:

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

6. 指数分布:

$$f(x) = \begin{cases} \frac{1}{\theta}e^{-x/\theta} & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.20}$$

7. 正态分布:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.21}$$

$f(x)$关于μ对称；$f(\mu) = max(f(x)) = \frac{1}{\sqrt{2\pi}\sigma}$。

8. *Gamma*分布:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\beta x} \tag{2.22}$$

其中：$x > 0,\ \alpha > 0,\ \beta > 0$

9. Inv-*Gamma*分布:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)}e^{-\frac{\beta}{x}} \tag{2.23}$$

其中：$x > 0,\ \alpha > 0,\ \beta > 0$

10. χ^2分布:

$$f_k(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1}e^{-\frac{x}{2}} \tag{2.24}$$

等价$\alpha = k/2, \beta = 1/2$的Gamma分布

11. Inv-χ^2分布:

$$f(x) = \frac{2^{-\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{-\left(\frac{k}{2}+1\right)}e^{-\frac{1}{2x}} \tag{2.25}$$

等价$\alpha = k/2, \beta = 1/2$的Inv-Gamma分布

12. Scaled Inv-χ^2分布:

$$f(x) = \frac{\frac{k}{2}^{-\frac{k}{2}}s^k}{\Gamma\left(\frac{k}{2}\right)} x^{-\left(\frac{k}{2}+1\right)}e^{-\frac{ks^2}{2x}} \tag{2.26}$$

等价$\alpha = k/2, \beta = ks^2/2$的Inv-Gamma分布。

2.5 大数定律与数理统计

2.6 误差传递

2.7 参数估计

2.8 假设检验

Chapter 3

贝叶斯统计

3.1 贝叶斯定理

1. 贝叶斯公式

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_i)P(B_j)} = \frac{P(A|B_i)P(B_i)}{P(A)} \tag{3.1}$$

先验：$P(B)$

似然: $P(A|B)$

后验: $P(B|A)$

证据（归一化）: $P(A)$

2. 贝叶斯公式含义：通过数据推算模型参数的概率。即：

$$P(\text{Model}(\theta)|\text{Data}) = P(\text{Data}|\text{Model}(\theta))P(\theta) \tag{3.2}$$

3. 贝叶斯统计的优势：将这个某种程度上是主观性的信息明确表达在先验概率中，而不是隐藏在没有明确指出的假设中; 让数据说话，减少主观性的先验概率。

3.2 贝叶斯单参数估计

3.2.1 二项分布估计

1. **无信息先验**

- 似然：

$$p(y|\theta) = \begin{pmatrix} n \\ y \end{pmatrix} \theta^y (1-\theta)^{n-y} \tag{3.3}$$

- 先验：均匀分布
- 后验：

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y} \sim Beta(y+1, n-y+1) \tag{3.4}$$

- 预测：

$$Pr(\widetilde{y}=1|y) = \int_0^1 Pr(\widetilde{y}=1|\theta, y) = \int_0^1 \theta p(\theta|y) d\theta = E(\theta|y) \tag{3.5}$$

1. **有信息先验**

- 先验：$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- 似然：

$$p(y|\theta) = \begin{pmatrix} n \\ y \end{pmatrix} \theta^y (1-\theta)^{n-y} \tag{3.6}$$

- 后验：

$$p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \sim Beta(\alpha+y, \beta+n-y)$$

- 后验期望：$E(\theta|y) = \frac{\alpha+y}{\alpha+\beta+n}$
- 先验期望：$E(y) = \frac{\alpha}{\alpha+\beta}$

  当$n \to \infty$: $E(\theta|y) \to y/n$

  数据很大的时候可以用正态分布近似后验分布

3.2.2 正态分布参数估计

1. **已知方差求均值**

- 似然：

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y-\theta)^2\right] \sim N(\theta, \sigma^2) \tag{3.7}$$

- 先验：

$$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right) \sim N(\mu_0, \tau_0^2) \tag{3.8}$$

- 后验：

$$p(\theta|y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right) \sim N(\mu_1, \tau_1) \tag{3.9}$$

  其中：

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}$$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

  精度：方差的倒数

- 预测：

$$\begin{aligned} p(\widetilde{y}|y) &= \int p(\widetilde{y}|\theta)p(\theta|y)d\theta \\ &\propto \int \exp\left(-\frac{(\widetilde{y}-\theta)^2}{2\sigma^2}\right)\exp\left(-\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right)d\theta \\ E(\widetilde{y}|\theta) &= \theta \\ D(\widetilde{y}|\theta) &= \sigma^2 + \tau_1^2 \end{aligned} \tag{3.10}$$

- 多个相互独立数据：

$$\begin{aligned} p(\theta \mid y) &\propto p(\theta)p(y \mid \theta) \\ &= p(\theta)\prod_{i=1}^{n} p\left(y_i \mid \theta\right) \\ &\propto \exp\left(-\frac{1}{2\tau_0^2}\left(\theta-\mu_0\right)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}\left(y_i-\theta\right)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2}\left(\theta-\mu_0\right)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\theta\right)^2\right]\right) \end{aligned} \tag{3.11}$$

可得：

$$p(\theta|y_1,\cdots,y_n) = p(\theta|\bar{y}) = N(\theta|\mu_n, \tau_n^2)$$

其中：

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

若$n \to +\infty$，τ_0不变，则：$\theta|y \sim N(\bar{y}, \sigma^2/n)$

若$\tau \to +\infty$，n 不变，则：$\theta|y \sim N(\bar{y}, \sigma^2/n)$

1. **已知均值求方差**

由于有n个服从$\sim N(\theta, \sigma^2)$的分布，因此：

- 似然：

$$p(y|\sigma^2) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right) = (\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2}v} \tag{3.12}$$

其中：

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$$

- 共轭先验：$p(\sigma^2) \propto (\sigma^2)^{-\alpha+1} e^{\beta/\sigma^2} \sim \mathrm{Inv} - \chi^2(v_0, \sigma_0^2)$
- 后验：

$$\begin{aligned} p(\sigma^2|y) &\propto p(\sigma^2)p(y|\sigma^2) \\ &\propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{v_0/2+1} \exp\left(-\frac{v_0\sigma_0^2}{2\sigma^2}\right) \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2}\frac{v}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-((n+v_0)/2+1)} \exp\left(-\frac{1}{2\sigma^2}(v_0\sigma_0^2 + nv)\right) \\ &\sim Inv - \chi^2(v_0 + n, \frac{v_0\sigma_0^2 + nv}{v_0 + n}) \end{aligned} \tag{3.13}$$

3.3 贝叶斯多参数模型

3.3.1 多参数模型处理

1. **置之不理**

2. **边缘化**

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2 \tag{3.14}$$

用贝叶斯公式展开：

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$$

将θ_2边缘化积分，得到θ_1的后验分布。

3. **平均化**

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2$$

3.3.2 无信息先验的正态分布

- 先验：

$$p(\mu, \ln\sigma^2) \sim U(\mu, \ln\sigma^2) \tag{3.15}$$

或者先验写为：

$$p(\mu, \sigma^2) \sim \frac{1}{\sigma^2}$$

- 似然（**有n次观测**）：

$$p(y|\mu, \sigma^2) = \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

- 联合后验：

$$p(\mu, \sigma^2|y) \propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]\right)$$

其中：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- 边缘后验$p(\sigma^2|y)$：

$$
\begin{aligned}
p(\sigma^2|y) &\propto \int p(\mu,\sigma^2|y)d\mu \\
&\propto (\sigma^2)^{-\frac{n+1}{2}} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \\
&\sim Inv-\chi^2(n-1,s^2)
\end{aligned}
\tag{3.16}
$$

- 边缘后验$p(\mu|y)$：

$$
\begin{aligned}
p(\mu|y) &= \int_0^{\infty} p(\mu,\sigma^2|y)d\sigma^2 \\
&\propto \left[1+\frac{n(\mu-\bar{y})^2}{(n-1)s^2}^{-n/2}\right] \\
&\sim t_{n-1}(\bar{y},s^2/n)
\end{aligned}
$$

- 预测后验分布

$$
p(\widetilde{y}|y) = \iint p(\widetilde{y}|\mu,\sigma^2,y)p(\mu,\sigma^2|y)d\mu d\sigma^2
$$

3.3.3 共轭先验分布

- 先验分布：

  $\mu|\sigma^2 \sim N(\mu_0,\sigma^2/\kappa_0)$

  $\sigma^2 \sim Inv-\chi^2(\nu_0,\sigma_0^2)$

- 联合先验分布

$$
\begin{aligned}
p(\mu,\sigma^2) &= p(\mu|\sigma^2)p(\sigma^2) \\
&\propto N-Inv-\chi^2(\mu_0,\sigma_0^2/\kappa_0\ ;\ \nu_0,\sigma_0^2)
\end{aligned}
$$

- 似然分布

$$
p(y|\mu,\sigma^2) = \sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)
$$

- 联合后验分布

$$\begin{aligned} p\left(\mu, \sigma^2 \mid y\right) & \propto \sigma^{-1}\left(\sigma^2\right)^{\left(\nu_0 / 2+1\right)} e^{-\frac{1}{2 \sigma^2}\left[\nu_0 \sigma_0^2+\kappa_0\left(\mu-\mu_0\right)^2\right]}\left(\sigma^2\right)^{-n / 2} e^{-\frac{1}{2 \sigma^2}\left[(n-1) s^2+n(\bar{y}-\mu)^2\right]} \\ & \propto \mathrm{N}-\operatorname{Inv}-\chi^2\left(\mu_n, \sigma_n^2 / \kappa_n ; \nu_n, \sigma_n^2\right) \end{aligned} \tag{3.17}$$

其中参数为：

$$\begin{gathered} \mu_n=\frac{\kappa_0}{\kappa_0+n} \mu_0+\frac{n}{\kappa_0+n} \bar{y} \\ \kappa_n=\kappa_0+n \\ \nu_n=\nu_0+n \\ \nu_n \sigma_n^2=\nu_0 \sigma_0^2+(n-1) s^2+\frac{\kappa_0 n}{\kappa_0+n}\left(\bar{y}-\mu_0\right)^2 \end{gathered} \tag{3.18}$$

- 条件后验分布$p(\mu|\sigma^2, y)$

$$\left(\mu \mid \sigma^2, y\right) \sim \mathrm{N}\left(\mu_n \frac{\sigma^2}{\kappa_n}\right)$$

- 方差边缘后验分布$p(\sigma^2|y)$

$$\left(\sigma^2 \mid y\right) \sim \operatorname{Inv}-\chi^2\left(\nu_n, \sigma_n^2\right)$$

- 均值边缘后验分布$p(\mu|y)$

$$\begin{aligned} p(\mu \mid y) & \propto\left[1+\frac{\kappa_n\left(\mu-\mu_n\right)^2}{\nu_n \sigma_n^2}\right]^{-\left(\nu_n+1\right) / 2} \\ & =t_{\nu_n}\left(\mu \mid m u_n, \sigma_n^2 / \kappa_n\right) \end{aligned} \tag{3.19}$$

3.4 层次化贝叶斯模型

3.4.1 参数化先验分布

1. 先验分布[2]

   先验分布是由某个未知参数的分布ϕ给出：

$$p(\theta|\phi)=\prod_{j=1}^{J} p(\theta_j|\phi) \qquad \theta=(\theta_1, \theta_2, \cdots)$$

   边缘化：

$$p(\theta) = \int \left(\prod_{j=1}^{J} p(\theta_j|\phi) \right) p(\phi) d\phi$$

2. 联合先验分布 $p(\phi, \theta) = p(\theta|\phi)p(\phi)$

3. 超先验分布：$p(\phi)$

4. 后验分布

   - 联合后验：$p(\phi, \theta|y) = p(y|\theta)p(\theta|\phi)p(\phi)$
   - 条件后验：$p(\theta|\phi, y)$
   - 边缘后验：$p(\phi|y)$

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}$$

5. 层次化贝叶斯完整表述

$$\begin{aligned} p(\phi, \theta|y) &\propto p(y|\phi, \theta)p(\phi, \theta) \\ &= p(y|\theta)p(\phi, \theta) \\ &= p(y|\theta)p(\theta|\phi)p(\phi) \end{aligned} \tag{3.20}$$

1. 层次化贝叶斯计算步骤

- 写出联合后验分布$p(\theta, \phi|y)$：即超先验分布，总体分布和似然分布的乘积
- 确定条件后验分布$p(\theta|\phi, y)$：

$$p(\theta|\phi, y) = \prod_{j=1}^{J} p(\theta_j|\phi, y)$$

- 边缘化给出ϕ的贝叶斯估计

3.4.2 二项分布的分层贝叶斯模型

1. y_i先验（组内模型） $y_j \sim Bin(n_j, \theta_j)$

2. θ_j先验（组间模型）：$\theta_j \sim Beta(\alpha, \beta)$

3. 联合先验：$p(\alpha, \beta, \theta) = p(\alpha, \beta)p(\theta|\alpha, \beta)$

4. 似然：$p(y|\theta, \alpha, \beta)$

5. 联合后验：$p(\theta,\alpha,\beta|y)$

$$
\begin{aligned}
p(\theta,\alpha,\beta|y) &\propto p(\alpha,\beta)p(\theta|\alpha,\beta)p(y|\theta,\alpha,\beta)\\
&= p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\prod_{j=1}^{J}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j}
\end{aligned}
\tag{3.21}
$$

6. 条件后验： $p(\theta|\alpha,\beta,y)$ ：单参数模型给定的后验

$$
\begin{aligned}
p(\theta\mid\alpha,\beta,y) &= \prod_{j=1}^{J}\frac{\Gamma\left(\alpha+\beta+n_j\right)}{\Gamma\left(\alpha+y_j\right)\Gamma\left(\beta+n_j-y_j\right)}\theta_j^{\alpha+y_j-1}\left(1-\theta_j\right)^{\beta+n_j-y_j-1}\\
&\sim\prod_{j=1}^{J}Beta(\alpha+y_j,\beta+n_j-y_j)
\end{aligned}
\tag{3.22}
$$

7. 边缘后验：$p(\alpha,\beta|y)$

$$
\begin{aligned}
p(\alpha,\beta\mid y) =&\frac{p(\theta,\alpha,\beta\mid y)}{p(\theta\mid\alpha,\beta,y)}\\
&\propto\frac{p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}\left(1-\theta_j\right)^{\beta-1}\prod_{j=1}^{J}\theta_j^{y_j}\left(1-\theta_j\right)^{n_j-y_j}}{\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}\left(1-\theta_j\right)^{\beta+n_j-y_j-1}}\\
=&p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)\Gamma\left(\alpha+y_j\right)\Gamma\left(\beta+n_j-y_j\right)}{\Gamma(\alpha)\Gamma(\beta)\Gamma\left(\alpha+\beta+n_j\right)}
\end{aligned}
\tag{3.23}
$$

3.4.3 正态分布的分层贝叶斯模型

1. **数据结构**

假设J个独立试验，每个实验都由θ_j给出其参数估计，估计n_j个$i.i.d$正态分布的数据点y_{ij}，每个点方差为σ^2：

$$
y_{ij}|\theta_j\sim N\left(\theta_j,\sigma^2\right)\quad i=1,\ldots,n_j\quad j=1,\ldots,J
$$

样本均值（充分统计量）：

$$
\bar{y}_{\cdot j}=\frac{1}{n_j}\sum_{i=1}^{n_j}y_{ij}
$$

样本均值的分布

$$
\bar{y}_{\cdot j}\sim N(\theta_j,\sigma_j^2)
$$

样本方差：

$$\sigma_j^2 = \frac{\sigma^2}{n_j}$$

样本的均值是从θ中估计，样本方差是从σ中估计。

相当于 θ 的似然分布：

$$\bar{y}_{\cdot j}|\theta \sim N(\theta_j, \sigma_j^2)$$

由于 σ 是已知的，下面所有的分布都是在 σ 已知情况下成立。

1. **层次化模型：无信息先验**

θ是从参数(μ, τ)中抽取：

$$p\left(\theta_1, \ldots, \theta_J \mid \mu, \tau^2\right) = \prod_{j=1}^{J} p\left(\theta_j \mid \mu, \tau^2\right)$$

边缘化：

$$p\left(\theta_1, \ldots, \theta_J\right) = \iint \prod_{j=1}^{J}\left[p\left(\theta_j \mid \mu, \tau^2\right)\right] p\left(\mu, \tau^2\right) d\mu d\tau$$

- **先验和似然**

组内抽样：$\bar{y}_{\cdot j} \sim N(\theta_j, \sigma_j^2)$

θ的先验（组内模型）：$\theta|\mu, \tau \sim N(\mu, \tau^2)$

μ的先验： $p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$

θ_j似然分布：$p(y|\theta) \sim (\bar{y}_{\cdot j}|\theta) \sim N(\theta_j, \sigma_j^2)$

- **联合后验：**$p(\theta, \phi|y)$

$$\begin{aligned} p(\theta, \mu, \tau|y) &\propto p(\mu, \tau)p(\theta|\mu, \tau)p(y|\theta) \\ &\propto p(\mu, \tau)\prod_{j=1}^{J} p(\theta_j|\mu, \tau^2)\prod_{j=1}^{J} p(\bar{y}_{\cdot j}|\theta_j, \sigma_j^2) \end{aligned} \tag{3.24}$$

其中：$\bar{y}_{\cdot j} \sim N(\theta_j, \sigma_j^2)$

可以忽略只依赖 y 和 σ_j 的参数，因为其已知。

- **θ条件后验**：$p(\theta_j \mid \mu, \tau, y_{\cdot,j})$

$$\theta_j \mid \mu, \tau, y_{\cdot,j} \sim N\left(\hat{\theta}_j, V_j\right)$$

其中：

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$$V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

- **超参数边缘后验**：$p(\mu, \tau \mid y)$

$$p(\mu, \tau \mid y) \propto p(\mu, \tau) p(y \mid \mu, \tau)$$

对于正态分布：

$$\bar{y}_{\cdot j} \sim N(\mu, \tau^2 + \sigma^2)$$

因此：

$$p(\mu, \tau \mid y) \propto p(\mu, \tau) \prod_{j=1}^{J} p\left(\bar{y}_{\cdot j} \mid \mu, \tau^2 + \sigma_j^2\right) \tag{3.25}$$

$$\bar{y}_{\cdot j} \mid \mu, (\tau^2 + \sigma_j^2) \sim N\left(\mu, \tau^2 + \sigma_j^2\right) \tag{3.26}$$

- **给定τ下μ的边缘后验分布**： $p(\mu \mid \tau, y)$从单参数模型中得出的结论

$$\mu \mid \tau, y \sim N\left(\hat{\mu}, V_\mu\right) \tag{3.27}$$

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}\bar{y}_{\cdot j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}}$$

$$V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}$$

- τ的后验：$p(\tau \mid y)$

$$
\begin{aligned}
p(\tau \mid y) &= \frac{p(\mu, \tau \mid y)}{p(\mu \mid \tau, y)} \\
&\propto \frac{p(\tau) \prod_{j=1}^{J} N\left(\bar{y}_{\cdot j} \mid \mu, \sigma_j^2 + \tau^2\right)}{N\left(\mu \mid \hat{\mu}, V_\mu\right)} \\
&\propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} \left(\sigma_j^2 + \tau^2\right)^{-1/2} \exp\left(-\frac{\left(\bar{y}_{\cdot j} - \hat{\mu}\right)^2}{2\left(\sigma_j^2 + \tau^2\right)}\right)
\end{aligned} \tag{3.28}
$$

- 从后往前采样，即先算出τ的后验，然后依次采样算出μ的后验，超参数联合后验，θ的条件后验，联合后验等。

3.5 贝叶斯回归

3.6 贝叶斯模型选择

3.7 费舍尔信息矩阵Fisher Information

Chapter 4

随机过程

4.1 随机过程及其统计描述

4.2 平稳随机过程

4.3 马尔科夫链

Chapter 5

MCMC

5.1 蒙特卡罗法 Monte Carlo Method

5.1.1 随机采样和接受-拒绝采样

蒙特卡罗法是通过概率模型的随机抽样进行随机抽样的方法。假设概率分布已知，通过概率分布得到随机样本，并通过得到的随机样本得到概率分布的随机性质，因此蒙特卡洛方法的核心是随机抽样。接下来我们介绍**接受-拒绝采样**。

已知概率密度分布为$f(x)$，但是这个概率密度分布复杂，各个变量并不独立，无法直接采样或者积分，因此可以通过蒙特卡罗方法进行抽样，得到样本X，得到其随机分布。我们在这介绍**接受-拒绝采样**。我们需要一个辅助的**建议分布**，记为$q(x)$。这个建议分布可以产生我们的候选样本，但建议分布要满足：

$$c * q(x) \geq f(x)$$

之后我们对样本按照建议分布$q(x)$进行抽样，得到样本x^*，同时对均匀分布$U(0,1)$进行抽样，得到u。之后计算$\frac{f(x)}{c*q(x)}$（这个值一定在0~1之间，对应图1的绿色部分比例），若

$$u \leq \frac{f(x^*)}{c * q(x^*)} \tag{1.1}$$

则x^*接受作为样本，否则拒绝。

怎么理解这个过程呢？简单来说就是我们先对建议分布$q(x)$的概率密度进行采样，因为这个比我们的$f(x)$更容易采样。假如这个分布很复杂，维度很高，直接算的话浪费计算资源，因此要先用一个简单的建议分布$q(x)$进行采样，得到建议的采样，但是这建议分布的采样终究不是我们需要的采样，所以我们需要在利用均匀分布$U(0,1)$，由我们算出来的$f(x^*)/cq(x^*)$进行接受或者拒绝。在图1中，按照

图 5.1:

绿色比例进行接受。如果在第x_i^*个抽样刚好落在中间红色区域比较大的点，那么拒绝的概率就高，反之绿色部分的比例更大，则我们接受的概率就越高。

听到这是不是有点迷糊了？别着急！我们看看图1，在是不是x^*处红色部分占比越大，与目标分布$f(x)$相差就越远了？所以我们在这里就必须剔除一些点了，不然远离我们的真实分布了！这时候可能会有其他疑问了，那在图1两端概率很小时候岂不是都接受率很高？是的，但是那两端概率很低呀！因为我们在使用建议分布抽样的时候概率那里的点已经是很少了，所以我们不用拒绝很多样本点也就和目标分布类似了。所以这时候就要用到一个均匀分布$U(0,1)$，在该点上随机生成一个u，然后按照(1.1)则接受，否则拒绝。

所以假设我们抽了n个样本，对样本进行拒绝，就是要生成和判断n次u的取值，也就是对每个样本点进行计算和判断是否拒绝，完成一次拒绝-接受采样。

接受-拒绝采样的缺点也是有的，主要是接受率比较低，抽样效率低。

5.1.2 数学期望和蒙特卡罗积分

如果我们要算目标函数为$f(x)$，其概率密度为$p(x)$，我们记函数$f(x)$关于密度函数$p(x)$的数学期望为$E_{p(x)}[f(x)]$。我们按照密度函数$f(x)$独立抽取n个样本$x_1, x_2 \cdots, x_n$，之后计算样本的均值：

$$\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

作为$f(x)$的近似值。根据大数定理，当样本容量增大，样本均值以概率1收敛于数学期望。因此我们可以用上述方法得到我们的数学期望。

$$E_{p(x)}[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

而在$\mathcal{X}$上数学期望的积分形式为：

$$E(x) = \int_{\mathcal{X}} f(x)p(x)dx$$

如果我们的目标积分为：

$$h(x) = f(x)q(x)$$

我们可以写为：

$$\int_{\mathcal{X}} h(x)dx = \int_{\mathcal{X}} f(x)q(x)dx = E_{p(x)}[f(x)]$$

对于复杂的函数，可以给定一个概率密度函数$p(x)$，只要取

$$f(x) = \frac{h(x)}{p(x)}$$

就可以用$p(x)$进行抽样算出积分：

$$\int_{\mathcal{X}} h(x)dx = E_{p(x)}[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

因此我们可以先用$p(x)$抽样。然后再进行积分。

5.2 蒙特卡罗法 Monte Carlo Method

5.2.1 随机采样和接受-拒绝采样

蒙特卡罗法是通过概率模型的随机抽样进行随机抽样的方法。假设概率分布已知，通过概率分布得到随机样本，并通过得到的随机样本得到概率分布的随机性质，因此蒙特卡洛方法的核心是随机抽样。接下来我们介绍**接受-拒绝采样**。

已知概率密度分布为$f(x)$，但是这个概率密度分布复杂，各个变量并不独立，无法直接采样或者积分，因此可以通过蒙特卡罗方法进行抽样，得到样本X，得到其随机分布。我们在这介绍**接受-拒绝采样**。我们需要一个辅助的**建议分布**，记为$q(x)$。这个建议分布可以产生我们的候选样本，但建议分布要满足：

$$c * q(x) \geq f(x)$$

之后我们对样本按照建议分布$q(x)$进行抽样，得到样本x^*，同时对均匀分布$U(0,1)$进行抽样，得到u。之后计算$\frac{f(x)}{c*q(x)}$（这个值一定在0~1之间，对应图1的绿色部分比例），若

$$u \leq \frac{f(x^*)}{c * q(x^*)} \tag{1.1}$$

则x^*接受作为样本，否则拒绝。

怎么理解这个过程呢？简单来说就是我们先对建议分布$q(x)$的概率密度进行采样，因为这个比我们的$f(x)$更容易采样。假如这个分布很复杂，维度很高，直接算的话浪费计算资源，因此要先用一个简单的建议分布$q(x)$进行采样，得到建议的采样，但是这建议分布的采样终究不是我们需要的采样，所

图 5.2:

以我们需要在利用均匀分布$U(0,1)$，由我们算出来的$f(x^*)/cq(x^*)$进行接受或者拒绝。在图1中，按照绿色比例进行接受。如果在第x_i^*个抽样刚好落在中间红色区域比较大的点，那么拒绝的概率就高，反之绿色部分的比例更大，则我们接受的概率就越高。

听到这是不是有点迷糊了？别着急！我们看看图1，在是不是x^*处红色部分占比越大，与目标分布$f(x)$相差就越远了？所以我们在这里就必须剔除一些点了，不然远离我们的真实分布了！这时候可能会有其他疑问了，那在图1两端概率很小时候岂不是都接受率很高？是的，但是那两端概率很低呀！因为我们在使用建议分布抽样的时候概率那里的点已经是很少了，所以我们不用拒绝很多样本点也就和目标分布类似了。所以这时候就要用到一个均匀分布$U(0,1)$，在该点上随机生成一个u，然后按照(1.1)则接受，否则拒绝。

所以假设我们抽了n个样本，对样本进行拒绝，就是要生成和判断n次u的取值，也就是对每个样本点进行计算和判断是否拒绝，完成一次拒绝-接受采样。

接受-拒绝采样的缺点也是有的，主要是接受率比较低，抽样效率低。

5.2.2 数学期望和蒙特卡罗积分

如果我们要算目标函数为$f(x)$，其概率密度为$p(x)$，我们记函数$f(x)$关于密度函数$p(x)$的数学期望为$E_{p(x)}[f(x)]$。我们按照密度函数$f(x)$独立抽取n个样本$x_1, x_2 \cdots, x_n$，之后计算样本的均值：

$$\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

作为$f(x)$的近似值。根据大数定理，当样本容量增大，样本均值以概率1收敛于数学期望。因此我们可以用上述方法得到我们的数学期望。

$$E_{p(x)}[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

而在$\mathcal{X}$上数学期望的积分形式为：

$$E(x) = \int_{\mathcal{X}} f(x)p(x)dx$$

如果我们的目标积分为：

$$h(x) = f(x)q(x)$$

我们可以写为：

$$\int_{\mathcal{X}} h(x)dx = \int_{\mathcal{X}} f(x)q(x)dx = E_{p(x)}[f(x)]$$

对于复杂的函数，可以给定一个概率密度函数$p(x)$，只要取

$$f(x) = \frac{h(x)}{p(x)}$$

就可以用$p(x)$进行抽样算出积分：

$$\int_{\mathcal{X}} h(x)dx = E_{p(x)}[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

因此我们可以先用$p(x)$抽样。然后再进行积分。

5.3 MCMC原理

5.3.1 MCMC原理

我们简单介绍了蒙特卡罗方法和马尔可夫链，接下来我们介绍马尔可夫链蒙特卡罗方法（简称MCMC方法）。MCMC方法适用于随机变量多元的、密度函数是非标准形式的、随机变量不相互独立的情况。

假设多元随机变量$x = [x_1, x_2, x_3, \cdots]$，满足$x \in \mathcal{X}$且其概率密度为$p(x)$，$f(x)$是定义在$x \in \mathcal{X}$上的函数，我们的目标是获得概率分布$p(x)$的抽样以及$f(x)$的数学期望$E_{p(x)}[f(x)]$。在随机变量x的状态空间$\mathcal{S}$上满意遍历定理（上一章2.2马尔科夫链性质）的马尔可夫链$X = \{X_0, X_1, \cdots, X_t, \cdots\}$，当这个马尔可夫链平稳时的分布就是其抽样的目标分布$p(x)$。

怎么通俗地解释这个原理呢？首先构造一个马尔可夫链，在状态空间$\mathcal{S}$上进行随机游走。根据遍历原理，总有一个时刻m之后，这个马尔可夫链近于平稳分布，也就是在期望附近游走。假设我们随机游走了n步，取我们平稳分布之后的样本集合$\{x_{m+1}, x_{m+2}, x_{m+3}, \cdots, x_n\}$，就是我们目标抽样分布的结果。

这里又要唠叨一句，平稳分布只是各个状态的期望。我们用天气由于预报的例子来说，假设当天气平稳分布的时候，其平稳分布为$[0.7\ 0.3]^T$，我们之后观测天气的结果符合我们这个平稳分布，也就是说我们接下来100天中有70天是晴天，30天是雨天。当然这只是一个简单的假设模型罢了！

所以当到达平稳之后样本集合为$\{x_{m+1}, x_{m+2}, x_{m+3}, \cdots, x_n\}$就是我们目标分布的结果。在时刻m之前的时期我们称为**燃烧期（burn-in）**。

更晕的还在后边，我们怎么构造这样一个马尔可夫链？这里我们需要一个转移核（连续）或者转移矩阵（离散）。如何构造这转移核/矩阵，构成一个可逆的马尔可夫链，使得遍历定理成立是很关键

的。如果该马尔可夫链成立，由于遍历定理成立，因此初始值的选取最终会收敛到同一平稳分布；燃烧期之前的样本都要丢弃，因为燃烧期之前的样本都不是服从样本的分布。当然目前MCMC收敛的判断是经验性的。

MCMC方法比拒绝-接受采样更容易实现，虽然丢弃了燃烧器之前的样本，但其效率仍然比拒绝-接受采样的效率高。目前常用的MCMC 方法主要是Metropolis-Hasting算法（M-H算法）和吉布斯抽样。

5.3.2 MCMC算法

根据上面的介绍，MCMC方法可以是以下的步骤：

1. 在随机变量x的状态空间$\mathcal{S}$上构造一个满足遍历定理的马尔可夫链，使得其平稳分布为$p(x)$；
2. 在状态空间某一点x_0出发，构造随机游走，产生样本$x_0, x_1, x_2, \cdots, x_t, \cdots$；
3. 应用遍历定理，确定燃烧期m，求得函数$f(x)$的均值

$$\hat{E}f = \frac{1}{n-m}\sum_{i=m+1}^{n} f(x_i)$$

在这其中有几个问题：

1. 如何定义马尔可夫链
2. 如何确定收敛步骤
3. 如何确定迭代步数确保精度

5.4 Metropolis-Hastings采样

5.4.1 M-H采样原理

上一章我们讲了MCMC抽样的一些问题，这一章我们介绍MCMC的一种代表算法。这一小节我们介绍M-H采样的原理。

我们需要构造一条马尔可夫链；要构造一个转移核，使得平稳分布就是我们要的抽样分布。参考第二章2.2节的细致平衡：

$$p_{ji}\pi_j = p_{ij}\pi_i$$

当然要构造这个细致平衡条件很难。假设**目标分布**为$\pi(x)$，转移核为$q(i,j)$，通常我们只能得到这样的结果：

$$\pi(i)q(i,j) \neq \pi(j)q(j,i)$$

因此我们要构造一个分布能使得其分布是细致平衡。

假设我们通过建议分布$q(i,j)$中随机抽取一个**候选状态**x_j**(后面简写为**j**)**，我们可以在两端乘以一个$\alpha(i,j)$使得其变为平稳分布，即：

$$\pi(i)q(i,j)\alpha(i,j) = \pi(j)q(j,i)\alpha(j,i)$$

在这里$\alpha(i,j)$为**接受分布**。建议分布$q(i,j)$是马尔可夫链转移核，且该马尔可夫链是不可约的，同时这个分布是容易采样的。那这个接受分布怎么构建呢？

我们对$\pi(i)q(i,j)\alpha(i,j) = \pi(j)q(j,i)\alpha(j,i)$ 移项：

$$\alpha(i,j) = \frac{\pi(j)(j,i)}{\pi(i)q(i,j)}\alpha(j,i)$$

实际上我们需要把两边的接受分布扩大到1，这样接受率才会达到最大。如果$\alpha(j,i) = 1$，有：

$$\alpha(i,j) = \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}$$

相反的，如果上式$\alpha(i,j) > 1$，我们令$\alpha(i,j) = 1$：

$$1 = \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}\alpha(j,i)$$

即我们取$\alpha(i,j) = 1$。这时候接受分布α扩大到最大。因此得到接受分布：

$$\alpha(i,j) = \min\left\{1, \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}\right\}$$

这时候的接受率最高，且容易证明这个转移核$q(i,j)\alpha(i,j)$是平稳分布的。之后我们从区间(0fi1)中均匀采样，得到一个随机数u，按照以下判断：

$$x_t = \begin{cases} x_j, & u \leq \alpha(i,j) \\ x_i, & u > \alpha(i,j) \end{cases}$$

决定其是否接受下一步。

图 5.3:

怎么理解这个接受分布呢？前面提到，当我们从状态i以概率$p(i,j)$转移到状态j的时候，不一定是平稳的，而加入α之后就可以得到一个新的平稳的马尔可夫链。而我们是否要转移到下一步就是要考虑这个接受分布了。我们定性的解释这个问题。假设我们的建议分布$q(i,j) = q(j,i)$，我们的接受分布$\alpha = \min\{1, \frac{\pi_j}{\pi_i}\}$，如果$\alpha$太小，说明我们下一步抽取的$j$所对应的概率是远小于我们上一步抽取的概率$i$，所以要舍弃；而如果是1的话说明抽取的$j$很符合我们的目标分布$\pi$，因此可以更好地接近我们要抽样的分布。这个判断步骤类似于本文开头的拒绝-接受采样。α类似于接受-拒绝采样的$\frac{f(x)}{c*q(x)}$。综上，其实这个转移核是要根据我们的抽样i,j共同确定的。

5.4.2 M-H采样算法

常见的采样有ptmcmc[3]

5.5 吉布斯采样

5.5.1 满条件分布

MCMC要抽样的函数一般都是多变量的联合概率分布$p(x) = p(x_1, x_2, \cdots, x_k)$，其中$x = (x_1, x_2, \cdots, x_k)^T$是$k$维变量。若条件概率分布：$p(x_I|x_{-I})$中出现了所有的变量$k$，其中:

$$x_I = \{x_i, i \in I\},\ x_{-I} = \{x_i, i \notin I\} \qquad I \subset K = \{1, 2, \cdots, k\}$$

那么称这个分布为**满条件分布**。

5.5.2 Gibbs采样原理

当然M-H采样有接受分布的存在，因此效率还是不够高。Gibbs采样可以避免这个问题。吉布斯采样的基本原理是从满条件概率分布出发，从满条件概率分布中抽样，得到一个样本序列。基本原理是：吉布斯抽样过程是在一个马尔可夫链上随机游走，平稳分布就是目标联合分布。接下来我们介绍Gibbs采样的细节。

我们考虑二维情况：假设有一个二维分布$p(x,y)$，我们发现：

$$p(x_1, y_1)p(y_2|x_1) = p(x_1)p(y_1|x_1)p(y_2|x_1)p(x_1, y_2)p(y_1|x_1) = p(x_1)p(y_2|x_1)p(y_1|x_1)$$

整理得：

$$p(x_1, y_1)p(y_2|x_1) = p(x_1, y_2)p(y_1|x_1)$$

图 5.4:

假设点A为(x_1,y_1)，点B为(x_1,y_2)，我们可以改写为：

$$p(A)p(y_2|x_1) = p(B)p(y_1|x_1)$$

也就是说当点A转移到点B的时候服从上述的马尔可夫链。而上式的条件概率就是我们的转移矩阵或者转移核。也就是说对维度y的满条件分布即为上述马尔可夫链的转移核或者转移矩阵。如果把二维扩展到n维，即可以得到n维的吉布斯抽样。假设建议分布是当前变量x_j，$j=1,2,\cdots,k$（也就是抽取第j维变量）的满条件分布：$q(x',x)=p(x'_j|x_{-j})$，这里的x指的是当前的抽样，x'指的是下一步的抽样。扩展到维的情多维：

$$p(x'_j,x_{-j})p(x'_j|x_{-j}) = p(x_j,x_{-j})p(x_j|x_{-j})$$

这时候的接受率$\alpha=1$。因此吉布斯采样可以认识是M-H采样的一种特殊情况。这个建议分布就是我们的目标分布的满条件分布。（这里和上一章的符号有所变化，不过内容是一样的，只是为了方便表达。）下面证明如何得到接受率$\alpha=1$：

根据M-H采样的接受分布公式，有：

$$q(x,x') = p(x'_j|x_{-j})$$

代入接受分布：

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

因为我们是抽取当前变量j，因此有：

$$\begin{aligned}\alpha(x,x') &= min\left\{1, \frac{p(x')q(x',x)}{p(x)q(x,x')}\right\} \\ &= min\left\{1, \frac{p(x'_j,x_{-j})p(x'_j|x_{-j})}{p(x_j,x_{-j})p(x_j|x_{-j})}\right\} = 1\end{aligned} \tag{5.1}$$

之后抽取k维，循环n次，抛去燃烧期m，得到我们的抽样，计算均值。

5.5.3 Gibbs采样算法

5.6 Nested采样

5.7 数值贝叶斯方法

Chapter 6

高斯随机过程

6.1 高斯随机过程及其统计描述

6.2 核密度估计

6.3 高斯混合模型

6.4 高斯学习

Chapter 7

统计算法

7.1 奇异值分解及主成分分析

7.2 退火算法

7.3 遗传算法

7.4 支持向量机

7.5 聚类算法

7.6 简单的神经网络

参考文献

[1] Piotr Jaranowski and Andrzej Królak. *Analysis of gravitational-wave data*. Number 29 in Cambridge monographs on particle physics, nuclear physics and cosmology. Cambridge University Press, Cambridge ; New York, 2009.

[2] Andrew Gelman, John B Carlin, Hal Steven Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. 2014. OCLC: 1063654237.

[3] Justin Ellis and Rutger van Haasteren. jellis18/ptmcmcsampler: Official release, October 2017.