Data Mining and Machine Learning

Guanxu Gleason WANG

University of Glasgow

版本：Past Paper of STAT5099 Data Mining and Machine Learning, UofG

更新：February 2, 2025

目录

0 常见分布及关系

0.1 离散型随机变量

二项分布: $X \sim \text{Binomial}(n, p)$

描述重复独立伯努利试验 n 次后成功 k 次数的分布, 参数为试验次数 n 和成功概率 p.

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ k \in \mathbb{N}. \quad \Longrightarrow \quad \begin{cases} \mathbb{E}[X] = np, \\ \text{Var}[X] = np(1-p). \end{cases}$$

其中,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}.$$

负二项分布: $X \sim \text{NegBin}(r, p)$

描述重复独立伯努利试验中, 经历了 k 次失败才能获得第 r 次成功的分布, 成功概率为 p.

$$\Pr(X = k) = \binom{k+r-1}{k} (1-p)^k p^r,\ k \in \mathbb{N}. \quad \Longrightarrow \quad \begin{cases} \mathbb{E}[X] = \dfrac{r(1-p)}{p}, \\ \text{Var}[X] = \dfrac{r(1-p)}{p^2}. \end{cases}$$

若更关心的是总共进行了多少次试验, 即 $n = k + r$, 最后一次成功 (确定的).

$$\Pr(N = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r},\ n \geqslant r, n \in \mathbb{N}. \quad \Longrightarrow \quad \begin{cases} \mathbb{E}[N] = \dfrac{r}{p}, \\ \operatorname{Var}[N] = \dfrac{r(1-p)}{p^2}. \end{cases}$$

泊松分布: $X \sim \operatorname{Poisson}(\lambda)$

表示单位时间或空间内某事件发生次数的分布, 参数为事件的平均发生率 λ.

$$\Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!},\ k \in \mathbb{N}. \quad \Longrightarrow \quad \begin{cases} \mathbb{E}[X] = \lambda, \\ \operatorname{Var}[X] = \lambda. \end{cases}$$

0.2 连续型随机变量

正态分布: $X \sim \mathcal{N}(\mu, \sigma^2)$

中心在 μ, 标准差为 σ 的钟形分布.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad \Longrightarrow \quad \begin{cases} \mathbb{E}[X] = \mu, \\ \operatorname{Var}[X] = \sigma^2. \end{cases}$$

指数分布: $X \sim \mathrm{Exp}(\lambda)$

中心在 μ, 标准差为 σ 的钟形分布.

$$f(x) = \lambda e^{-\lambda x},\ x > 0. \quad \Longrightarrow \quad \begin{cases} \mathbb{E}[X] = \dfrac{1}{\lambda}, \\ \mathrm{Var}[X] = \dfrac{1}{\lambda^2}. \end{cases}$$

伽马分布: $X \sim \mathrm{Gamma}(\alpha, \lambda)$

形状参数 $\alpha > 0$, 尺度参数 $\lambda > 0$.

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)},\ x > 0. \quad \Longrightarrow \quad \begin{cases} \mathbb{E}[X] = \dfrac{\alpha}{\lambda}, \\ \mathrm{Var}[X] = \dfrac{\alpha}{\lambda^2}. \end{cases}$$

显然, $X \sim \mathrm{Gamma}(1, \lambda)$ 时, $f(x) = \lambda e^{-\lambda x}$, 等价于 $X \sim \mathrm{Exp}(\lambda)$.

贝塔分布: $X \sim \mathrm{Beta}(\alpha, \beta)$

两个形状参数 $\alpha > 0$ 和 $\beta > 0$ 分别控制左侧和右侧区间.

$$f(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)},\ x\in[0,1]. \quad \Longrightarrow \quad \begin{cases}\mathbb{E}\left[X\right]=\dfrac{\alpha}{\alpha+\beta},\\ \mathrm{Var}\left[X\right]=\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},\\ \mathrm{Mode}[X]=\dfrac{\alpha-1}{\alpha+\beta-2},\ \alpha,\beta>1.\end{cases}$$

显然, $X\sim\mathrm{Beta}(1,1)$ 时, $f(x)=1$, 等价于 $X\sim\mathrm{Uniform}(0,1)$.

0.3 伽马函数与贝塔函数

0.3.1 伽马函数

Gamma 函数定义为:

$$\Gamma(n)=\int_0^\infty t^{n-1}e^{-t}\,\mathrm{d}t \quad \Longrightarrow \quad \int_0^\infty x^n e^{-ax}\,\mathrm{d}x=\frac{\Gamma(n+1)}{a^{n+1}}. \tag{1}$$

并且对于正整数 ($n\in\mathbb{N}^*$), 有,

$$\Gamma(n)=(n-1)!$$

对于任意正数 $x>0$, 有,

$$\Gamma(x+1)=x\Gamma(x),\quad \Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}. \quad \Longrightarrow \quad \Gamma\left(x+\frac{1}{2}\right)=(x-1)!\sqrt{\pi}.$$

0.3.2 贝塔函数

Beta 函数定义为:

$$\mathrm{B}(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x \tag{2}$$

用伽马函数表示

$$\mathrm{B}(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

1 贝叶斯模型的推导与 Gibbs 采样

1.1 联合概率密度函数和全条件分布

1.1.1 联合概率密度函数 (Joint Probability Distribution Function)

通过条件概率公式和先验概率递归表示. 公式如下:

$$\begin{aligned} p(\mathbf{y}, \mu, \tau) &= p(\mathbf{y}|\mu, \tau)p(\mu)p(\tau) \\ &= p(\mu)p(\tau)\prod_{i=1}^{n} p(y_i|\mu, \tau) \end{aligned} \tag{3}$$

假设

$$\begin{aligned} y_i|\mu, \tau &\sim \mathcal{N}(\mu, \tau^{-1}), \\ \mu &\sim \mathcal{N}(1, 1), \\ \tau &\sim Ga(2, 1). \end{aligned}$$

因此

$$
\begin{aligned}
p(\mathbf{y},\mu,\tau) &= p(\mathbf{y}|\mu,\tau)p(\mu)p(\tau) \\
&= p(\mu)p(\tau)\prod_{i=1}^{n} p(y_i|\mu,\tau) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\mu-1)^2}{2}\right)\cdot\frac{1^2\tau^{2-1}e^{-\tau}}{\Gamma(2)}\cdot\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\tau^{-1}}}\exp\left(-\frac{(y_i-\mu)^2}{2\tau^{-1}}\right) \\
&= (2\pi)^{-\frac{1}{2}}\exp\left(-\frac{(\mu-1)^2}{2}\right)\cdot\tau\exp(-\tau)\cdot\prod_{i=1}^{n}(2\pi)^{-\frac{1}{2}}\tau^{\frac{1}{2}}\exp\left(-\frac{(y_i-\mu)^2}{2\tau^{-1}}\right) \\
&= (2\pi)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mu-1)^2\right)\cdot\tau\exp(-\tau)\cdot(2\pi)^{-\frac{n}{2}}\tau^{\frac{n}{2}}\exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \\
&= (2\pi)^{-\frac{n+1}{2}}\tau^{\frac{n}{2}+1}\exp\left(-\frac{1}{2}\left[(\mu-1)^2+\tau\sum_{i=1}^{n}(y_i-\mu)^2\right]\right)\exp(-\tau)
\end{aligned}
$$

1.1.2 全条件分布 (Full Conditional Distribution)

根据贝叶斯公式

$$
p(x|y) = \frac{p(y|x)p(x)}{p(y)} \propto p(y|x)p(x) = p(x,y). \tag{4}
$$

接上题

$$
\begin{aligned}
p(\mu|\mathbf{y},\tau) &\propto p(\mathbf{y},\mu,\tau) \\
&= (2\pi)^{-\frac{n+1}{2}}\tau^{\frac{n}{2}+1}\exp\left(-\frac{1}{2}\left[(\mu-1)^2+\tau\sum_{i=1}^{n}(y_i-\mu)^2\right]\right)\exp(-\tau) \\
&\propto \exp\left(-\frac{1}{2}\left[(\mu-1)^2+\tau\sum_{i=1}^{n}(y_i-\mu)^2\right]\right) \\
&= \exp\left(-\frac{(1+\tau n)}{2}\left[\left(\mu-\frac{1+\tau n\overline{y}}{1+\tau n}\right)^2+\text{const.}\right]\right) \\
&\propto \exp\left(-\frac{(1+\tau n)}{2}\left(\mu-\frac{1+\tau n\overline{y}}{1+\tau n}\right)^2\right) \quad \Longrightarrow \quad \mu|\mathbf{y},\tau\sim\mathcal{N}\left(\frac{1+\tau n\overline{y}}{1+\tau n},(1+\tau n)^{-1}\right).
\end{aligned}
$$

同理

$$
\begin{aligned}
p(\tau|\mathbf{y},\mu) &\propto p(\mathbf{y},\mu,\tau) \\
&= (2\pi)^{-\frac{n+1}{2}}\tau^{\frac{n}{2}+1}\exp\left(-\frac{1}{2}\left[(\mu-1)^2+\tau\sum_{i=1}^{n}(y_i-\mu)^2\right]\right)\exp(-\tau) \\
&\propto \tau^{\frac{n}{2}+1}\exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)\exp(-\tau) \\
&= \tau^{\frac{n}{2}+1}\exp\left(-\left[1+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right]\tau\right) \quad \Longrightarrow \quad \tau|\mathbf{y},\mu\sim Ga\left(\frac{n}{2}+21+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right).
\end{aligned}
$$

1.2 Gibbs 采样

Gibbs 采样提供了一种间接的方法, 通过依次从全条件分布中采样来逼近后验分布.

假设我们有一个高维参数向量 $\theta = (\theta_1, \theta_2, \cdots, \theta_n)$，但很难直接从其后验分布 $p(\theta|y)$ 采样. Gibbs 采样的思路是:

1. **分解后验分布:**
   计算每个参数的全条件分布 $p(\theta_i|\theta_{-i}, y)$, 其中 θ_{-i} 表示除 θ_i 以外的所有参数.
2. **循环采样:**
   - 先固定所有其他参数, 从 $p(\theta_1|\theta_2, \theta_3, \cdots, \theta_n, y)$ 中采样新的 θ_1;
   - 再固定新的 θ_2 和其他参数, 从 $p(\theta_2|\theta_1, \theta_3, \cdots, \theta_n, y)$ 中采样新的 θ_2;
   - $\cdots$
   - 最后固定新的 θ_n 和其他参数, 从 $p(\theta_n|\theta_1, \theta_2, \cdots, \theta_{n-1}, y)$ 中采样新的 θ_n.

这形成了一个马尔可夫链, 经过足够长的采样后, 样本会收敛到后验分布 $p(\theta|y)$.

例 1.1 (Gibbs 采样两参数举例) 假设有两个参数 θ_1, θ_2, 目标是从后验分布 $p(\theta_1, \theta_2|y)$ 采样:

1. 选择一个初始值 $\theta_2^{(0)}$.
2. 迭代以下步骤 T 次:
   - **Step 1.** 从条件分布 $p(\theta_1|\theta_2^{(t-1)}, y)$ 采样新的 $\theta_1^{(t)}$.
   - **Step 2.** 从条件分布 $p(\theta_2|\theta_1^{(t)}, y)$ 采样新的 $\theta_2^{(t)}$.
3. 经过一定的"burn-in" (前几个样本丢弃), 最终的样本可以用于逼近后验分布.

接上题:

(1) Initialise $\tau^{(0)}$ in the support of this parameter. E.g. sample them from the prior.
(2) for $k = 1, \cdots, K$
Draw $\mu^{(k)}$ from $p(\mu|\tau^{(k-1)}, \mathbf{y})$
Draw $\tau^{(k)}$ from $p(\tau|\mu^{(k)}, \mathbf{y})$
(3) Discard burn-in and optionally thin-out.

1.2.1 使用 Gibbs 采样样本生成后验预测分布样本

Explain how you could use samples $(\mu^{(t)}, \tau^{(t)}), t = 1, \ldots, T$ from the Gibbs sampler to generate samples from the posterior predictive distribution for a new (independent) observation $\widetilde{y}$.

解释如何利用从 Gibbs 采样器中得到的样本 $(\mu^{(t)}, \tau^{(t)}), t = 1, \ldots, T$, 生成新的（独立的）观测 y 的后验预测分布的样本?

例 1.2 Gibbs 采样已经生成了一组从后验分布 $p(\mu, \tau|y)$ 中抽取的样本:

$$\{(\mu^{(t)}, \tau^{(t)}) : t = 1, \ldots, T\}$$

根据贝叶斯理论, 新的独立观测 $\widetilde{y}$ 的后验预测分布为:

$$p(\widetilde{y}|y) = \int p(\widetilde{y}|\mu, \tau)p(\mu, \tau|y)d\mu d\tau$$

在 Gibbs 采样的情况下, 我们可以通过以下步骤近似生成 y 的样本:

1. 丢弃 **burn-in** 样本 (初始未收敛样本).
2. 使用剩下的样本 $(\mu^{(t)}, \tau^{(t)})$:
   - 对于每一个样本 $(\mu^{(k)}, \tau^{(k)})$, 生成 $\widetilde{y}^{(k)} \sim N(\mu^{(k)}, (\tau^{(k)})^{-1})$.
   - 这表示 $\widetilde{y}$ 的预测分布依赖于当前的 μ 和 τ 值.
3. 最终, 所有生成的 $\widetilde{y}^{(k)}$ 样本 $(k = 1, \ldots, T')$ 将近似于后验预测分布 $p(\widetilde{y}|y)$。

Solution:

Having collected draws from our Gibbs sampler, discard the burn-in, and possibly thin-out. For every remaining draw, take current $(\mu^{(k)}, \tau^{(k)})$, plug them into the likelihood $\widetilde{y}|\mu, \tau \sim \mathcal{N}(\mu, \tau^{-1})$ and generate a sample from this distribution. The resulting set of samples are coming from the posterior predictive distribution.

收集了我们的 Gibbs 抽样结果后, 丢弃预热期数据, 可能还需要进行数据稀疏化. 对于每个剩余的抽样, 取当前的 $(\mu^{(k)}, \tau^{(k)})$, 将它们代入似然函数 $\widetilde{y}|\mu, \tau \sim \mathcal{N}(\mu, \tau^{-1})$, 并从这个分布中生成一个样本. 得到的样本集来自后验预测分布.

1.2.2 为什么丢弃初始样本? 多条链如何帮助决定丢弃数量?

1. 为什么在 Gibbs 采样中通常丢弃初始样本?
2. 运行多条链如何帮助决定丢弃的数量?

例 1.3 1. 为什么丢弃初始样本?

Gibbs 采样是一种马尔可夫链方法, 初始状态通常与目标分布 (后验分布) 不一致. 采样过程需要一些时间达到平稳分布, 这段时间被称为 **burn-in** 阶段.

- 初始样本的问题**:** 在 burn-in 阶段, 样本未收敛到目标分布, 可能会过多依赖初始值, 导致偏差.
- 丢弃的目的**:** 通过丢弃初始样本, 确保最终用于推断的样本接近后验分布.

2. 多条链如何帮助判断丢弃数量?

   运行多条链 (从不同的初始值出发), 可以帮助检测马尔可夫链是否已经收敛到目标分布.

   - 方法**:**
     - 从不同的初始点运行多条链.
     - 对每条链绘制参数值的时间序列图, 观察是否在一定时间后趋于一致.
   - 判断 burn-in 阶段:

     当多条链的采样值开始收敛并在相同区域波动时, 可以确定 burn-in 阶段结束. 在此之前的样本应被丢弃.

Solution:

The Gibbs sampler is a Markov chain. Markov chains take time to reach equilibrium, i.e. settle into their limiting distribution. So one needs to eliminate initial draws which will not yet be sampled from the correct distribution. These initial draws are called the burn-in period.

吉布斯抽样是一种马尔可夫链。马尔可夫链需要时间才能达到平衡，即进入其极限分布。因此，需要消除那些尚未从正确分布中抽取的初始抽取。这些初始抽取被称为预热期。

Multiple chains started from different initial points will all converge to the same limiting distribution along

their own trajectories. A time plot of all the chains will indicate when all the chains come together, and samples prior to that point should be discarded as burn-in.

从不同初始点开始的多个链都将沿着自己的轨迹收敛到相同的极限分布。所有链的时间图将指示所有链何时聚集在一起，并且在那之前的样本应被视为预热期而被丢弃。

2 Poisson 模型与共轭分布及 Jeffreys 先验

给定一个时间段内的公交车到达次数 y, 建模为一个**泊松分布**:

$$p(y|\lambda) = \frac{\lambda^y}{y!}e^{-\lambda}, \quad \lambda > 0,\ y = 0, 1, 2, \ldots$$

2.1 似然函数的推导

似然函数描述的是在参数已知的情况下, 观测数据的联合概率. 泊松分布的概率密度函数为:

$$p(y_i|\lambda) = \frac{\lambda^{y_i}}{y_i!}e^{-\lambda}, \quad i = 1, \ldots, n$$

由于观测值是独立的, 联合似然函数为:

$$\begin{aligned} L(\lambda) = p(y_1, \ldots, y_n|\lambda) &= \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!}e^{-\lambda} \\ &= \frac{\lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} \exp(-n\lambda) \end{aligned}$$

2.2 共轭先验的定义及 Gamma 分布共轭性质的证明

2.2.1 共轭先验

在贝叶斯统计中, 如果后验分布 $p(\lambda|\mathbf{y})$ 和先验分布 $p(\lambda)$ 属于同一分布族, 则称该先验分布为**共轭先验**. 共轭先验的优点是计算简单, 后验分布的参数可以直接通过数据更新.

Suppose the prior distribution comes from a particular family of distributions, e.g. the gamma distributions. If the posterior distribution (obtained via Bayes' theorem) is from the same family of distributions (with parameters modified by the data), then the prior is conjugate to the likelihood.

假设先验分布来自特定的分布族, 例如 Gamma 分布. 如果后验分布 (通过贝叶斯定理获得) 来自同一分布族 (参数由数据修改), 则先验分布与似然函数是共轭的.

2.2.2 先验分布: Gamma 分布

假设先验分布为 Gamma 分布 $\mathrm{Gamma}(\alpha, \beta)$, 先验为:

$$p(\lambda) = \frac{\beta^{\alpha}\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)} \quad \propto \lambda^{\alpha-1}e^{-\beta\lambda}$$

泊松分布的似然函数为:

$$L(\lambda|\mathbf{y}) = \frac{\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}\exp(-n\lambda) \quad \propto \lambda^{\sum_{i=1}^{n} y_i}\exp(-n\lambda)$$

后验分布按贝叶斯公式计算:

$$
\begin{aligned}
p(\lambda|\mathbf{y}) &= \frac{p(y|\lambda)p(\lambda)}{p(y)} \\
&\propto L(\lambda|y)p(\lambda) \\
&\propto \lambda^{\sum_{i=1}^{n} y_i} \exp(-n\lambda) \cdot \lambda^{\alpha-1} e^{-\beta\lambda} \\
&= \lambda^{\alpha-1+\sum_{i=1}^{n} y_i} \exp(-(\beta+n)\lambda)
\end{aligned}
$$

$$
\Rightarrow \lambda|\mathbf{y} \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{n} y_i, \beta + n\right) \tag{5}
$$

代入似然函数和先验分布:

$$
p(\lambda|\mathbf{y}) \propto \left[\lambda^{\sum_{i=1}^{n} y_i} e^{-n\lambda}\right] \cdot \left[\lambda^{\alpha-1} e^{-\beta\lambda}\right]
$$

i.e. the posterior is also a Gamma distribution, therefore the prior is conjugate. (即, 后验分布也是一个 Gamma 分布, 因此先验分布是共轭的.)

2.3 求参数 λ 的 Jeffreys 先验 $p(\lambda)$

Jeffreys 先验是一种非信息先验（Non-informative Prior），定义为:

$$
p(\lambda) \propto \sqrt{J(\lambda)} \tag{6}
$$

其中 $J(\lambda)$ 是参数 λ 的 Fisher 信息量，定义为：

$$J(\lambda) = \mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\log p(\mathbf{y}|\lambda)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\log L(\lambda)\middle|\lambda\right] \tag{7}$$

因此，

$$\begin{aligned}
L(\lambda) &= \frac{\lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}\exp(-n\lambda). \\
\log L(\lambda) &= \log\left(\lambda^{\sum_{i=1}^n y_i}\right) - \log\left(\prod_{i=1}^n y_i!\right) + \log\left(\exp(-n\lambda)\right) \\
&= \left(\sum_{i=1}^n y_i\right)\log\lambda + \text{const.} - n\lambda. \\
\frac{\partial}{\partial\lambda}\log L(\lambda) &= \frac{1}{\lambda}\sum_{i=1}^n y_i - n. \\
\frac{\partial^2}{\partial\lambda^2}\log L(\lambda) &= -\frac{1}{\lambda^2}\sum_{i=1}^n y_i. \\
\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\log L(\lambda)\middle|\lambda\right] &= -\frac{1}{\lambda^2}\sum_{i=1}^n \mathbb{E}\left[y_i\right] = -\frac{n\lambda}{\lambda^2} = -\frac{n}{\lambda}. \\
J(\lambda) &= \frac{n}{\lambda}. \\
p(\lambda) &= \sqrt{\frac{n}{\lambda}} \propto \frac{1}{\sqrt{\lambda}}.
\end{aligned}$$

2.4 Jeffreys 先验是否是一个合法的先验分布?

一个先验分布 $p(\lambda)$ 合法的条件是其积分必须有限:

$$\int_0^\infty p(\lambda)d\lambda < \infty$$

因此,

$$\int_0^\infty \frac{1}{\sqrt{\lambda}}d\lambda = \left[2\lambda^{\frac{1}{2}}\right]_0^\infty = \infty,$$

i.e. the normalisation constant does not exist. 因此 Jeffreys 先验是**不合法的** (improper).

2.5 使用 Jeffreys 先验的后验分布是否合法?

后验分布是否合法取决于后验分布 $p(\lambda|\mathbf{y})$ 是否为正则分布 (proper distribution), 即

$$\int_0^\infty p(\lambda|\mathbf{y})\ \mathrm{d}\lambda = 1.$$

使用 Jeffreys 先验 $p(\lambda) \propto \frac{1}{\sqrt{\lambda}}$ 等价于 $\mathrm{Gamma}\left(\frac{1}{2}, 0\right)$. 由2.2.2可知, 其后验分布为:

$$\lambda|\mathbf{y} \sim \mathrm{Gamma}\left(\frac{1}{2} + \sum_{i=1}^n y_i, n\right)$$

and as long as $n \geqslant 1$, it is a proper (normalised) probability distribution. So, the posterior is proper. 这是因为

- 参数 $\alpha = \frac{1}{2} + \sum_{i=1}^{n} y_i = \frac{1}{2} + n\bar{y}$ 保证了伽玛分布的形状参数始终为正;
- 参数 $\beta = n \geqslant 1$ 保证了规模参数始终有效.

2.6 使用 Jeffreys 先验能否用于推断?

正如2.4 和 2.5所说, 即使 Jeffreys 先验不合法, 也能通过得到合法的后验分布用于推断, 即:

$$
\begin{aligned}
p(\lambda|\mathbf{y}) \propto p(\mathbf{y}|\lambda)p(\lambda) \ &= L(\lambda)\underbrace{p(\lambda)}_{\propto \frac{1}{\sqrt{\lambda}}} \\
&\propto \frac{\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \exp(-n\lambda) \cdot \frac{1}{\sqrt{\lambda}} \\
&\propto \lambda^{-\frac{1}{2}+\sum_{i=1}^{n} y_i} \exp(-n\lambda) \\
&\propto \text{Gamma}\left(\frac{1}{2} + \sum_{i=1}^{n} y_i, n\right)
\end{aligned}
$$

从而再次验证2.5所得到的结论.

2.7 应用

2.7.1 预测概率

假设观测数据的模型 $p(\widetilde{y}|\theta)$, 后验分布为 $p(\theta|\mathbf{y})$, 那么后验预测分布为

$$p(\widetilde{y}|y) = \int_0^\infty p(\widetilde{y}|\theta)p(\theta|\mathbf{y})\,\mathrm{d}\theta.$$

例 2.1 Assume an improper prior $p(\lambda) = \lambda^{-\frac{1}{2}}$. You observe that three buses arrive at this bus stop within the first and the second hours $y_1 = y_2 = 3$. With what probability (down to three places after the decimal point) would you expect four buses to arrive at this bus stop within the third hour $p(\widetilde{y}_3 = 4|y_1, y_2) =$?

Hint: Assume $\Gamma(6.5) \approx 287.8853$ and $\Gamma(10.5) \approx 1133278$.

假设一个不恰当的先验 $p(\lambda) = \lambda^{-\frac{1}{2}}$. 你观察到在前两小时内有三辆公交车到达这个公交车站, $y_1 = y_2 = 3$. 你期望在第三小时内有多少概率 (精确到小数点后三位) 会有四辆公交车到达这个公交车站 $p(\widetilde{y}_3 = 4|y_1, y_2) =$?

使用 Jeffreys 先验 $p(\lambda) \propto \dfrac{1}{\sqrt{\lambda}}$ 等价于 $\mathrm{Gamma}\left(\dfrac{1}{2}, 0\right)$. 由2.2.2中 Eq.5可知, 其后验分布为:

$$\lambda|y_1, y_2 \sim \mathrm{Gamma}\left(\frac{1}{2} + \sum_{i=1}^{2} y_i, 2\right) = \mathrm{Gamma}\left(6\frac{1}{2}, 2\right)$$

Next, we need to derive the posterior predictive distribution $p(\widetilde{y}_3|y_1, y_2)$:

$$\begin{aligned} p(\widetilde{y}_3|y_1,y_2) &= \int_0^\infty p(\widetilde{y}_3|\lambda,y_1,y_2)p(\lambda|y_1,y_2)\,\mathrm{d}\lambda \;=\; \int_0^\infty p(\widetilde{y}_3|\lambda)p(\lambda|y_1,y_2)\,\mathrm{d}\lambda \\ &= \int_0^\infty \frac{\lambda^{\widetilde{y}_3}}{\widetilde{y}_3!}e^{-\lambda}\cdot\frac{2^{6.5}}{\Gamma(6.5)}\lambda^{6.5-1}e^{-2\lambda}\,\mathrm{d}\lambda \\ &= \frac{2^{6.5}}{\widetilde{y}_3!\cdot\Gamma(6.5)}\int_0^\infty \lambda^{\widetilde{y}_3+5.5}e^{-3\lambda}\,\mathrm{d}\lambda \qquad (*) \\ &= \frac{2^{6.5}}{\widetilde{y}_3!\cdot\Gamma(6.5)}\cdot\frac{\Gamma(\widetilde{y}_3+6.5)}{3^{\widetilde{y}_3+6.5}}. \end{aligned}$$

Due to the Eq.1, so that

$$\int_0^\infty \lambda^{\widetilde{y}_3+5.5}e^{-3\lambda}\,\mathrm{d}\lambda \;\overset{n=\widetilde{y}_3+5.5}{=\!=\!=}\; \int_0^\infty \lambda^{n}e^{-3\lambda}\,\mathrm{d}\lambda \;=\frac{\Gamma(n+1)}{3^{n+1}} \;=\frac{\Gamma(\widetilde{y}_3+6.5)}{3^{\widetilde{y}_3+6.5}}.$$

Therefore,

$$\begin{aligned} p(\widetilde{y}_3=4|y_1,y_2) &= \frac{2^{6.5}}{4!\cdot\Gamma(6.5)}\cdot\frac{\Gamma(4+6.5)}{3^{4+6.5}} \\ &= \frac{2^{6.5}}{4!\cdot\Gamma(6.5)}\cdot\frac{\Gamma(10.5)}{3^{10.5}} \;\approx\; \frac{90.5097}{24\cdot 287.8853}\cdot\frac{1133278}{102275.8681} \;\approx 0.145. \end{aligned}$$

2.7.2 后验更新

例 2.2 You observe that 2, 1 and 3 buses arrive at this bus stop within the first three hours $\mathbf{y} = (2, 1, 3)$. You subsequently discover that in the fourth hour of the bus stop there were fewer than 3 buses arrived. Derive the

updated posterior for λ. Make sure that this posterior is properly normalised.

你观察到分别有 2、1 和 3 辆公交车在前三个小时内到达这个公交车站，即 $\mathbf{y} = (2, 1, 3)$。随后你发现，在公交车站的第四个小时内，到达的公交车少于 3 辆。推导出 λ 的更新后验。确保这个后验是正确归一化的。

由2.2.2中 Eq.5可知, 其后验分布为:

$$\lambda|\mathbf{y} \sim \text{Gamma}\left(\frac{1}{2} + \sum_{i=1}^{3} y_i, 2\right) = \text{Gamma}\left(6.5, 3\right).$$

于是有,

$$\begin{aligned} p(\lambda|\mathbf{y} = (2,1,3), z < 3) &\propto p(z < 3|\lambda)p(\lambda|\mathbf{y}) \\ &= \left(\frac{e^{-\lambda}\lambda^0}{0!} + \frac{e^{-\lambda}\lambda^1}{1!} + \frac{e^{-\lambda}\lambda^2}{2!}\right)\frac{3^{6.5}\lambda^{5.5}e^{-3\lambda}}{\Gamma(6.5)} \\ &\propto \left(1 + \lambda + \frac{1}{2}\lambda\right)e^{-\lambda} \cdot \lambda^{5.5}e^{-3\lambda} \\ &= \lambda^{5.5}e^{-4\lambda} + \lambda^{6.5}e^{-4\lambda} + \frac{1}{2}\lambda^{7.5}e^{-4\lambda} \end{aligned}$$

确保后验分布是一个合法的概率分布, 因此要进行归一化 (normalise), 即引入一个归一化常数 C,

使得

$$
\begin{aligned}
C\int_0^\infty \left(\lambda^{5.5}e^{-4\lambda}+\lambda^{6.5}e^{-4\lambda}+\frac{1}{2}\lambda^{7.5}e^{-4\lambda}\right)\,\mathrm{d}\lambda &= 1\\
C\left(\int_0^\infty \lambda^{5.5}e^{-4\lambda}\,\mathrm{d}\lambda+\int_0^\infty \lambda^{6.5}e^{-4\lambda}\,\mathrm{d}\lambda+\frac{1}{2}\int_0^\infty \lambda^{7.5}e^{-4\lambda}\,\mathrm{d}\lambda\right) &= 1\\
C\left(\frac{\Gamma(6.5)}{4^{6.5}}+\frac{\Gamma(7.5)}{4^{7.5}}+\frac{1}{2}\frac{\Gamma(8.5)}{4^{8.5}}\right) &= 1\\
C &= \frac{1}{\frac{\Gamma(6.5)}{4^{6.5}}+\frac{\Gamma(7.5)}{4^{7.5}}+\frac{1}{2}\frac{\Gamma(8.5)}{4^{8.5}}}
\end{aligned}
$$

因此得到每一个部分的权重分别为

$$
C\frac{\Gamma(6.5)}{4^{6.5}},\quad C\frac{\Gamma(7.5)}{4^{7.5}},\quad C\frac{1}{2}\frac{\Gamma(8.5)}{4^{8.5}}
$$

于是得到最终的结果

$$
p(\lambda|\mathbf{y}=(2,1,3),z<3)=C\left(\frac{\Gamma(6.5)}{4^{6.5}}\mathrm{Gamma}(6.5,4)+\frac{\Gamma(7.5)}{4^{7.5}}\mathrm{Gamma}(7.5,4)+\frac{1}{2}\frac{\Gamma(8.5)}{4^{8.5}}\mathrm{Gamma}(8.5,4)\right)
$$

3 Beta 分布与二项分布模型及最大后验估计 (MAP)

1. 某批产品中单个产品不合格的概率为 $\theta, \theta \in (0,1)$.
2. 先验分布: $\theta \sim \mathrm{Beta}(\alpha, \beta)$.
   - 均值为: $\mathbb{E}[\theta] = \dfrac{\alpha}{\alpha+\beta}$.
   - 方差为: $\mathrm{Var}[\theta] = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
   - 众数为: $\dfrac{\alpha-1}{\alpha+\beta-2}$.
3. 采样观测: 从某批次中抽取 $n = 20$ 个产品, 其中 $k = 1$ 个不合格.

3.1 Beta 分布的均值和方差公式

例 3.1 Select a Beta prior for θ such that its mean is 0.01 and its standard deviation is 0.01.

$$\begin{cases} \mu = 0.01 = \dfrac{\alpha}{\alpha+\beta}, \\ \sigma = 0.01 = \sqrt{\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}. \end{cases} \Rightarrow \begin{cases} \beta = 99\alpha, \\ 100\alpha = 98. \end{cases} \Rightarrow \begin{cases} \alpha = 0.98, \\ \beta = 97.02. \end{cases}$$

So, the prior satisfying our requirements is $\mathrm{Beta}(0.98, 97.02)$.

3.2 使用 Beta 先验计算后验分布

We are obviously working with a binomial model for component failure:

$$y|\theta \sim \text{Binomial}(\theta, n).$$

As the prior is conjugate, the posterior is derived as

$$\theta|y \sim \text{Beta}(\alpha + k, \beta + n - k) \ = \text{Beta}(0.98 + 1, 97.02 + 19) \ = \text{Beta}(1.98, 116.02).$$

3.3 计算后验分布的最大后验估计 (maximum a posteriori, MAP)

Key point: MAP 估计是后验分布的众数.

As the posterior is

$$\theta|y \sim \text{Beta}(1.98, 116.02),$$

the MAP estimate of θ is

$$\frac{\alpha - 1}{\alpha + \beta - 2} = \frac{1.98 - 1}{1.98 + 116.02 - 2} = \frac{0.98}{116} \approx 0.0084.$$

3.4 使用 R 模拟估计 95% 的后验区间

First we take a large sample of θ from the posterior:

```
theta <- rbeta(1000000, 1.5, 68.5)
```

The quantiles $(0.025$ and $0.975)$ of this sample then estimate the bounds of the 95% central posterior interval for θ:

```
quantile(theta, probs=c(0.025,0.975))
```

4 分层模型（Hierarchical Model）和 Nimble 编程

Consider a Nimble model defined using the following code:

```
1  exam_code <- nimbleCode({
2    for (j in 1:N) {
3      y[j] ~ dnorm(mu[j], sd = sigma)
4      mu[j] ~ dnorm(theta[j], tau)
5      theta[j] <- beta_0 + beta_1 * log(x[j]) + beta_2 * x[j]
6    }
7    beta_0 ~ dnorm(0, sd = 10)
8    beta_1 ~ dnorm(0, sd = 10)
9    beta_2 ~ dnorm(0, sd = 10)
10   sigma ~ dexp(1)
11   tau ~ dexp(10)
12 })
```

where x and y are vectors from our data set.

4.1 将 Nimble 模型规范转化为标准统计表示

合并 4, 5 行可以写作

$$\mu_j|\beta_0,\beta_1,\beta_2,\tau \sim \mathcal{N}(\beta_0+\beta_1\log(x_j)+\beta_2 x_j,\tau_2),\ j=1,\cdots,N, \text{independently}$$

Row	Standard statistical notation
3	$y_j \mid \mu_j, \sigma \sim \mathcal{N}(\mu_j, \sigma^2),\ j = 1, \cdots, N,$ independently
4	$\mu_j \mid \theta, \tau \sim \mathcal{N}(\theta, \tau_2),\ j = 1, \cdots, N,$ independently
5	$\theta = \beta_0 + \beta_1 \log(x_j) + \beta_2 x_j$
7	$\beta_0 \sim \mathcal{N}(0, 100)$
8	$\beta_1 \sim \mathcal{N}(0, 100)$
9	$\beta_2 \sim \mathcal{N}(0, 100)$
10	$\sigma \sim \text{Exp}(1)$
11	$\tau \sim \text{Exp}(10)$

4.2 分层模型

We consider N independent observations of y coming from a normal distribution with the same variance σ^2, but the mean μ_j is different for every observation. The mean μ_j for observation j has a normal prior with mean $\beta_0 + \beta_1 \log(x_j) + \beta_2 x_j$, where $\beta_0, \beta_1, \beta_2$ are hyperparameters, and the variance for the prior on μ_j is τ^2. Hyperpriors on β are normal with zero mean and variance 100 independently for every β. Hyperpriors for σ and τ are exponential with rates 1 and 10 correspondingly.

我们考虑 N 个独立观测值 y，这些观测值来自一个具有相同方差 σ^2 的正态分布，但每个观测值的均值 μ_j 是不同的。观测值 j 的均值 μ_j 具有一个正态先验分布，其均值为 $\beta_0 + \beta_1 \log(x_j) + \beta_2 x_j$，其中 $\beta_0, \beta_1, \beta_2$ 是超参数，μ_j 的先验分布的方差为 τ^2。超参数 β 的先验分布是均值为零、方差为 100 的独立正态分布。参数 σ 和 τ 的超先验分布是指数分布，其率参数分别为 1 和 10。

One advantage of the hierarchical structure is 'borrowing strengths' (or 'sharing information'): we allow information from other replicates to influence estimation of μ_j for any particular replicate (via their effects on β and τ).

分层结构的一个优点是“信息借用”（或“信息共享”）：我们允许其他重复实验的信息通过其对 β 和 τ 的影响来影响某一特定重复实验的 μ_j 的估计。

In detail,

1. 第一层: 观测模型

$y_j \sim N(\mu_j, \sigma^2)$: 每个观测值 y_j 服从均值为 μ_j、方差为 σ^2 的正态分布.

2. 第二层: 均值的先验分布

   每个均值 μ_j 服从一个正态分布

$$\mu_j \sim \mathcal{N}(\theta_j, \tau^2)$$

   其中 $\theta_j = \beta_0 + \beta_1 \log(x_j) + \beta_2 x_j$.

3. 第三层: 回归系数和参数的超先验分布

   模型中涉及一些超参数, 这些参数本身有先验分布:

   - 回归系数 $\beta_0, \beta_1, \beta_2$: 服从正态分布

$$\beta_0, \beta_1, \beta_2 \sim \mathcal{N}(0, 100).$$

   - 观测误差方差 σ^2: 其平方根 σ 服从指数分布

$$\sigma \sim \text{Exp}(1).$$

   - μ_j 方差的参数 τ^2: 其精度（τ^{-1}）服从指数分布:

$$\tau \sim \text{Exp}(10).$$

答案中提到分层模型的一个核心优点是**信息借用 (borrowing strengths)**, 这也是分层模型的关键优势之一。

什么是信息借用? 在分层模型中, 所有观测值 $y_1, y_2, \ldots, y_N$ 并不是独立的, 而是通过模型的参数（如 $\beta_0, \beta_1, \beta_2, \tau$）建立了联系.

信息共享的作用

- 如果某些观测点的数据较少或不稳定，模型可以通过从其他观测点中“借用信息”来改善这些点的参数估计.
- 例如，观测点 j 的均值 μ_j 的估计会受到其他观测点的影响，因为它们共享相同的全局参数 β 和 τ.

具体体现在本题中的信息借用

- 回归系数 $\beta_0, \beta_1, \beta_2$ 和方差参数 τ 是全局共享的.
- 每个 μ_j 的估计可以通过共享的 β 和 τ 从其他观测点的信息中受益, 特别是在某些 x_j 数据稀缺的情况下.

优点总结

1. **提高稳定性:** 对于某些稀疏数据点，信息借用可以改善参数估计的稳定性。
2. **捕捉层次结构:** 分层模型可以更好地反映数据中的多层次依赖关系。
3. **减少过拟合:** 因为分层模型对全局参数施加了先验约束，从而避免了仅基于个体点的过拟合。

直观总结

1. 模型结构:
   - 第一层: 观测值 y_j 的分布.
   - 第二层: 均值 μ_j 的分布.
   - 第三层: 全局参数（β, σ, τ）的先验分布.
2. 分层模型的核心:

- 通过全局参数（如 β, τ）的共享, 建立了个体观测之间的联系.
- 数据稀疏时, 模型可以利用全局信息来改善个体估计.

3. 分层模型的优点:
   - 信息共享与借用.
   - 捕捉数据中的多层次关系.
   - 提高稀疏数据下的推断效果.

例 4.1 (WinBUGS) Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1/\tau) \quad i = 1, \ldots, n, \quad \text{independently,}$$

where β_0 is the intercept parameter, β_1 is the slope parameter, and τ is the precision of the normally distributed observation noise. We can rewrite the above model as

$$y_i | \beta_0, \beta_1, \tau, x_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \frac{1}{\tau}\right), \quad i = 1, \ldots, n, \quad \text{independently.}$$

Having observed some n data (x_i, y_i), we want to infer the posterior of model parameters β_0, β_1, and τ in a fully Bayesian way, using the following priors:

$$\left.\begin{aligned} \beta_0 &\sim \mathcal{N}(\mu_0, 1/\tau_0) \\ \beta_1 &\sim \mathcal{N}(\mu_1, 1/\tau_1) \\ \tau &\sim \text{Gamma}(\alpha, \beta) \end{aligned}\right\} \text{independently}$$

with fixed real values for hyperparameters $\mu_0, \mu_1, \tau_0, \tau_1, \alpha, \beta$.

Define this model for performing sampling from the parameter posterior using WinBUGS, i.e., provide the model description that goes within model$\{\cdots\}$ in WinBUGS.

```
model {
  for (i in 1:n) {
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1 * x[i] # 显式定义均值 mu[i]
  }
  beta0 ~ dnorm(mu0, tau0)
  beta1 ~ dnorm(mu1, tau1)
  tau ~ dgamma(alpha, beta)
}
```

or

```
model {
  for (i in 1:n) {
    y[i] ~ dnorm(beta0 + beta1 * x[i], tau) # 直接写模型
  }
  beta0 ~ dnorm(mu0, tau0)
  beta1 ~ dnorm(mu1, tau1)
  tau ~ dgamma(alpha, beta)
}
```

5 贝叶斯决策理论与损失函数

Consider a problem of estimating the rate parameter λ of the Poisson model for data $y_1, \cdots, y_n$ with the value ('action') a, using a Gamma prior on the rate parameter λ:

$$y_i|\lambda \sim \text{Poisson}(\lambda), i = 1, \cdots, n, \text{independently},$$
$$\lambda|\gamma, \beta \sim \text{Gamma}(\gamma, \beta).$$

5.1 定义贝叶斯期望损失和贝叶斯动作

贝叶斯决策理论

1. 贝叶斯决策理论通过**损失函数**来度量决策与真实参数值之间的偏差.
2. **贝叶斯期望损失 (Bayesian expected loss)** 是损失函数在后验分布下的期望值:

$$\rho(\pi, a) = \mathbb{E}\left[L(\lambda, a)|\mathbf{y}\right] = \int_\Theta L(\lambda, a)p(\lambda|\mathbf{y})\,\mathrm{d}\lambda \tag{8}$$

   - $L(\lambda, a)$ 是损失函数, 表示在真实参数值为 λ 时选择 a 的代价.
   - π 表示后验分布 $p(\lambda|\mathbf{y})$.
3. **贝叶斯动作 (Bayes action)** 是使期望损失最小化的决策 a_π

$$a^\pi = \arg\min_a \rho(\pi, a) = \arg\min_a \int_\Theta L(\lambda, a)p(\lambda|\mathbf{y})\,\mathrm{d}\lambda \tag{9}$$

例 5.1 Define the Bayesian expected loss $\rho(\pi, a)$ and the Bayes action a^π in the context of making decision in Bayesian framework and using some loss function $L(\lambda, a)$.

解. If $\pi(\lambda) = p(\lambda|\mathbf{y})$ is the believed probability distribution of λ at the time of decision making, the Bayesian expected loss of an action a is Equation (8).

A Bayes action a π is then an action $a \in \mathcal{A}$ that minimises $\rho(\pi, a)$. Or in layman's terms (以通俗的话来说), the best bet for a future action a.

5.2 特定损失函数下的贝叶斯动作

例 5.2 给定损失函数:

$$L(\lambda, a) = \frac{(\lambda - a)^2}{\lambda}$$

证明在观察了 n 个数据点 $y_1, \cdots, y_n$ 后, 贝叶斯行动 a^π 采用 λ 的最大后验估计形式. 假设 $n > 0$ 且 $\sum_{i=1}^{n} y_i > 1$.

解. 步骤如下:

Step 1. 推导贝叶斯期望损失 (根据 Eq.8):

$$\begin{aligned}\rho(\pi, a) &= \mathbb{E}\left[L(\lambda, a)|\mathbf{y}\right] = \mathbb{E}\left[\frac{(\lambda - a)^2}{\lambda}\middle|\mathbf{y}\right] \\ &= \mathbb{E}\left[\frac{\lambda^2 - 2a\lambda + a^2}{\lambda}\middle|\mathbf{y}\right] \\ &= \mathbb{E}\left[\frac{\lambda^2}{\lambda}\middle|\mathbf{y}\right] - 2a\mathbb{E}\left[\frac{\lambda}{\lambda}\middle|\mathbf{y}\right] + a^2\mathbb{E}\left[\frac{1}{\lambda}\middle|\mathbf{y}\right] \\ &= \mathbb{E}\left[\lambda|\mathbf{y}\right] - 2a + a^2\mathbb{E}\left[\frac{1}{\lambda}\middle|\mathbf{y}\right]\end{aligned}$$

Step 2. 求偏导 (根据 Eq.9), 即

$$\frac{\partial \rho(\pi, a)}{\partial a} = 0,$$

由于当前先验分布 (prior distribution) 得到后验分布 (posterior distribution, Eq.5) 为

$$\lambda|\mathbf{y} \sim Ga\left(\gamma + \sum_{i=1}^{n} y_i, \beta + n\right) \quad \Longrightarrow \quad p(\lambda|\mathbf{y}) = \frac{(\beta + n)^{\gamma + \sum_{i=1}^{n} y_i}}{\Gamma\left(\gamma + \sum_{i=1}^{n} y_i\right)} \lambda^{\gamma + \sum_{i=1}^{n} y_i - 1} \exp(-(\beta + n)\lambda).$$

从而得到贝叶斯期望损失最小时的贝叶斯动作:

$$
\begin{aligned}
\frac{\partial \rho(\pi, a)}{\partial a} &= \frac{\partial}{\partial a}\left(\mathbb{E}\left[\lambda|\mathbf{y}\right] - 2a + a^2\mathbb{E}\left[\frac{1}{\lambda}\middle|\mathbf{y}\right]\right) \\
&= -2 + 2a\mathbb{E}\left[\frac{1}{\lambda}\middle|\mathbf{y}\right] \\
&= -2 + 2a\int_0^\infty \frac{1}{\lambda} p(\lambda|\mathbf{y})\,\mathrm{d}\lambda \\
&= -2 + 2a\int_0^\infty \frac{1}{\lambda}\cdot\frac{(\beta+n)^{\gamma+\sum y_i}}{\Gamma\left(\gamma+\sum y_i\right)}\lambda^{\gamma+\sum y_i-1}\exp(-(\beta+n)\lambda)\,\mathrm{d}\lambda \\
&= -2 + 2a\frac{(\beta+n)^{\gamma+\sum y_i}}{\Gamma\left(\gamma+\sum y_i\right)}\int_0^\infty \lambda^{\gamma+\sum y_i-2}\exp(-(\beta+n)\lambda)\,\mathrm{d}\lambda \\
&\overset{Eq.1}{=\!=} -2 + 2a\frac{(\beta+n)^{\gamma+\sum y_i}}{\Gamma\left(\gamma+\sum y_i\right)}\cdot\frac{\Gamma(\gamma+\sum y_i-1)}{(\beta+n)^{\gamma+\sum y_i-1}} \\
&\overset{\Gamma(a+1)=a\Gamma(a)}{=\!=\!=} -2 + 2a\frac{\beta+n}{\gamma+\sum y_i-1} \;= 0
\end{aligned}
$$

解得:

$$
a = \frac{\gamma+\sum_{i=1}^n y_i - 1}{\beta+n}.
$$

Step 3. 求二阶偏导, 以证实上值为最小值.

$$
\frac{\partial^2 \rho(\pi, a)}{\partial a^2} = 2\frac{\beta+n}{\gamma+\sum_{i=1}^n y_i - 1} > 0
$$

as long as $n > 0$, and $\sum_{i=1}^{n} y_i > 1$.

5.3 平方损失函数下的贝叶斯动作

例 5.3 给定损失函数替换为:

$$L(\lambda, a) = (\lambda - a)^2.$$

解. 步骤同例5.2完全一致:

Step 1. 推导贝叶斯期望损失 (根据 Eq.8):

$$\begin{aligned}\rho(\pi, a) &= \mathbb{E}\left[L(\lambda, a)|\mathbf{y}\right] = \mathbb{E}\left[(\lambda - a)^2|\mathbf{y}\right] \\ &= \mathbb{E}\left[\lambda^2 - 2a\lambda + a^2|\mathbf{y}\right] \\ &= \mathbb{E}\left[\lambda^2|\mathbf{y}\right] - 2a\mathbb{E}\left[\lambda|\mathbf{y}\right] + a^2\end{aligned}$$

Step 2. 求偏导 (根据 Eq.9), 即

$$\frac{\partial \rho(\pi, a)}{\partial a} = 0,$$

由于当前先验分布 (prior distribution) 得到后验分布 (posterior distribution, Eq.5) 为

$$\lambda|\mathbf{y} \sim Ga\left(\gamma + \sum_{i=1}^{n} y_i, \beta + n\right)$$

从而得到贝叶斯期望损失最小时的贝叶斯动作:

$$\frac{\partial \rho(\pi, a)}{\partial a} = \frac{\partial}{\partial a}\left(\mathbb{E}\left[\lambda^2|\mathbf{y}\right] - 2a\mathbb{E}\left[\lambda|\mathbf{y}\right] + a^2\right) = -2a\mathbb{E}\left[\lambda|\mathbf{y}\right] + 2a = 0$$

解得:

$$a = \mathbb{E}\left[\lambda|\mathbf{y}\right].$$

Step 3. 求二阶偏导, 以证实上值为最小值.

$$\frac{\partial^2 \rho(\pi, a)}{\partial a^2} = 2 > 0$$

which is strictly positive, it is a minimum, so the Bayes action:

$$a^{\pi} = \mathbb{E}\left[\lambda|\mathbf{y}\right] = \frac{\gamma + \sum_{i=1}^{n} y_i}{\beta + n}.$$