第四章 先验分布的确定

Wang Shujia

Contents

Motivation

本章目标：如何选择先验分布？或如何把**先验信息**转换为**先验分布**？

1. 先验分布是贝叶斯分析的最大争议。因此如何选择令人信服的先验分布是贝叶斯分析的重要工作。
    - “事实上，多数读者都不愿意接受那些在分析数据之前一无所知的作者所指定的模型”
2. 如何权衡影响后验分布的两个因素？
    - 先验分布的强度和数据的多寡
3. 选择先验分布有哪些常用的方法？
    (a) 参考前人的研究成果
    (b) 根据观察和预测性的思考，导出先验分布
    (c) 用部分数据确定先验分布的超参数
    (d) 用无信息先验和共轭先验

先验分布的一般类型

一般不能严格区分，常见类型有：

1. 无信息先验 (uninformative)
2. 共轭先验 (conjugate)
3. 有信息先验 (informative)

Andrew Gelman 推荐

Prior Choice Recommendations:https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

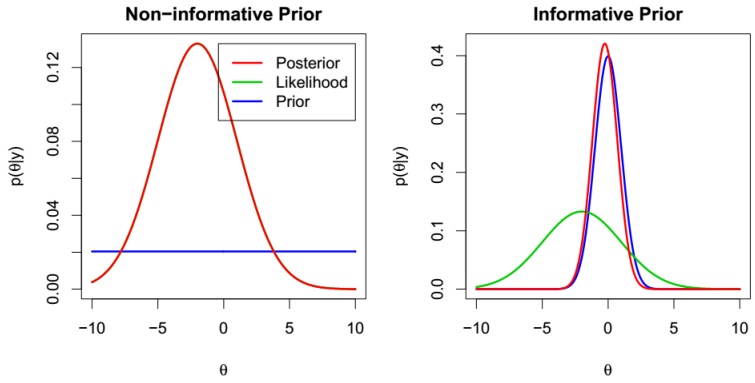
- Flat prior;
- Super-vague but proper prior: normal(0, 1e6);
- Weakly informative prior, very weak: normal(0, 10);
- Generic weakly informative prior: normal(0, 1);
- Specific informative prior: normal(0.4, 0.2) or whatever. Sometimes this can be expressed as a scaling followed by a generic prior: theta = 0.4 + 0.2*z; z ~ normal(0, 1);

1 无信息先验 (Uninformative prior)

无信息先验 (Uninformative prior)

- 没有先验信息或不清楚，如何确定先验分布？
- **无信息先验** (Uninformative,Noninformative): 先验分布 $\pi(\theta)$ 为常数或均匀分布，先验信息对结果无影响

- Also known as: *vague, flat, ignorance*, or *diffuse* priors



均匀先验 (Uniform Priors)

- **正常先验** (Proper): 先验分布 $\pi(\theta)$ 是一个密度函数 (即非负，积分等于 1)，如 $\pi(\theta) \propto 1,\ 0 < \theta < 1$
- **非正常先验** (Improper prior): 先验分布 $\pi(\theta)$ 在定义域内积分不存在，即

$$\pi(\theta) \geq 0 \text{ and } \int_{\Theta} \pi(\theta) \mathrm{d}\theta = +\infty$$

  如 $\mathrm{N}(\theta, 1):\ \pi(\theta) \propto 1, -\infty < \theta < \infty$
- **均匀先验**：先验分布 $\pi(\theta)$ 是一个均匀分布
- Comments
  - 无信息先验（无限支撑）常常是非正常先验
  - 即使是非正常先验，后验分布一般也是正常分布
  - 经常作为正常先验的极限分布 (先验方差 $\sigma_0^2 \to \infty$)
  - 在 WinBUGS，非正常无信息先验常用 $N(0, 1000^2)$ 近似
  - 用非正常先验可使误用先验信息更稳健 (Robust)

Jeffreys 先验

无信息先验的参数变换：不满足不变性

$$x_1, x_2, \ldots, x_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2),\ \mu \text{ known}$$

先验

$$\pi(\sigma) \propto 1$$

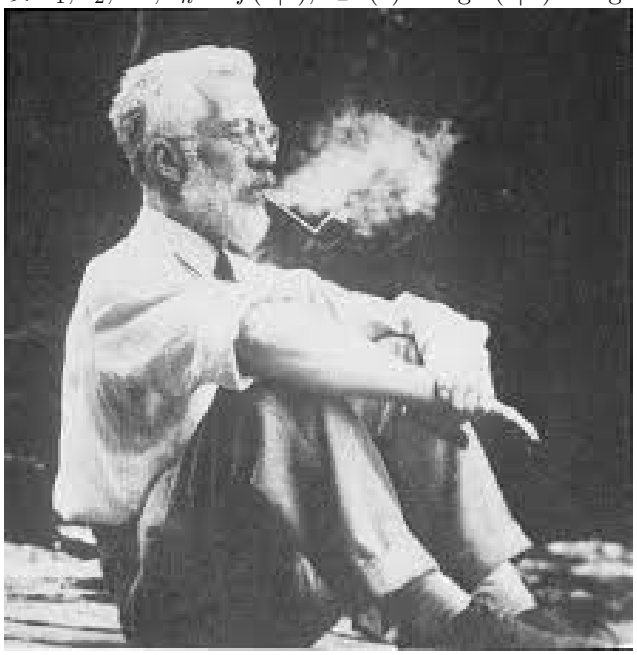
则方差 $\phi = \sigma^2$ 的先验分布: $\pi(\sigma^2) \propto 1/\sigma \neq 1$

定理 1 (随机变量的变换). *设 r.v.θ 的 pdf 为 $f(\theta)$, $\phi = g(\theta)$ 是 θ 的严格单调函数，则 ϕ 的分布密度为*

$$p(\phi) = f(\theta)|\frac{d\theta}{d\phi}| = f(\theta)|g'(\theta)|^{-1},\ (\theta = g^{-1}(\phi))$$

Fisher 信息

设 $x_1, x_2, \ldots, x_n \overset{iid}{\sim} f(x|\theta)$, 记 $l(\theta) = \log L(\theta|\boldsymbol{x}) = \log \Pi f(x_i|\theta)$ 为似然函数的对数。



则关于 θ 的 **Fisher 信息**定义为：

$$J(\theta) = \mathrm{E}\left[\left(\frac{\partial l(\theta)}{\partial \theta}\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2 l(\theta)}{\partial \theta^2}\right]$$

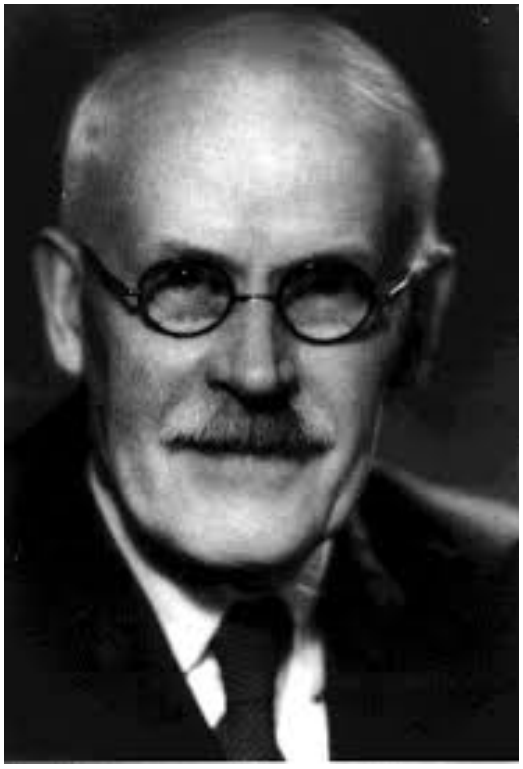
如果 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, 则 Fisher 信息矩阵为

$$J(\boldsymbol{\theta}) = -\mathrm{E}\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)$$

Jeffreys 先验

Jeffreys 先验：与 Fisher 信息平方根成比例，即

$$\pi(\boldsymbol{\theta}) \propto |J(\boldsymbol{\theta})|^{1/2}$$



- Good news: 是应用最广泛的无信息先验分布之一。
  - 它仅依赖于似然函数的形式，与数据无关
  - Jeffreys 具有参数不变性
- Bad news:
  - 有时 Fisher 信息不存在（如 Cauchy 分布）
  - 多参数时推导不易
  - 可能 improper，应用时要额外小心

0-1 分布的 Jefreys 先验

设 $x_1, x_2, \ldots, x_n \overset{iid}{\sim} \mathrm{B}(1,p)$,$x = \sum x_i$，似然函数对数为

$$l(p) = \log L(p) = x \log p + (n-x)\log(1-p) + constant$$

$$\frac{\partial^2 l(p)}{\partial p^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}$$

由于 $E(X) = np$, 关于 p 的 Jeffreys 先验为

$$\pi(p) \propto p^{-1/2}(1-p)^{-1/2}$$

- 这是什么分布？

正态分布的 Jeffreys 先验

设 $x_1, x_2, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$，$\boldsymbol{\theta} = (\mu, \sigma)^T$ 未知，则似然函数为

$$\begin{aligned} L(\mu, \sigma^2|\boldsymbol{x}) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2\right), \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(vs^2 + n(\mu - \hat{\mu})^2\right), \end{aligned}$$

$$l(\mu, \sigma^2|\boldsymbol{x}) = c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(vs^2 + n(\mu - \hat{\mu})^2),$$

其中 $\hat{\mu} = \bar{x},\ v = n - 1,\ s^2 = \frac{1}{n-1}\sum(x_i - \hat{\mu})^2$，因此 Fisher 信息矩阵为

$$J(\boldsymbol{\theta}) = -\mathrm{E}\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^2) \end{pmatrix}$$

因此 Jeffreys 先验为：

$$\pi(\boldsymbol{\theta}) \propto |J(\boldsymbol{\theta})|^{1/2} \propto 1/\sigma^2$$

参照先验（Reference priors）

- 参照先验（Reference priors）：先验信息最小（数据信息最大）的先验分布
- Berger, James O., José M. Bernardo, and Dongchu Sun. "The formal definition of reference priors." The Annals of Statistics (2009): 905-938.
- 使似然函数与后验分布之间的距离最小（Kullback-Leibler distance）
- Locally uniform prior

经验贝叶斯分析

- 经验贝叶斯 (Empirical Bayesian approach): 利用样本信息确定先验分布的超参数
- 始于 Robbins(1951,1955,1961)
- 分为两类 (Morris,1983)
  - 参数经验贝叶斯 (PEB)
  - 非参数经验贝叶斯 (NPEB)

正态分布的经验贝叶斯

假设 $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, 则似然函数

$$\begin{aligned} L(\mu, \sigma^2|\boldsymbol{x}) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(vs^2+n(\mu-\hat{\mu})^2\right) \end{aligned}$$

其中 $\hat{\mu} = \bar{x},\ v = n-1,\ s^2 = \frac{1}{n-1}\sum(x_i-\hat{\mu})^2$, 因此

$$\begin{aligned} \mu|\sigma^2 &\sim N(\hat{\mu}, \frac{\sigma^2}{n}) \\ \sigma^2 &\sim \text{Inv}-\chi^2(v, s^2) \end{aligned}$$

以上两个分布就作为先验分布

2 共轭先验 (Conjugate priors)

共轭先验的概念

定义 1 (共轭先验). *如果先验分布与后验分布属于相同的分布族，则称该先验分布为共轭先验 (Conjugate priors)*

- 共轭先验数学上简单，但不一定能反映真实信息
  - 计算方便
  - 易于理解
  - 限于参数分布族
- 可用于真实先验信息的近似

常见共轭先验

1. Binomial: $x \sim \text{B}(n, \theta)$,
   $\theta \sim \text{Beta}(p, q)$, $\theta|x \sim \text{Beta}(p+x, q+n-x)$
2. Poisson:$x_1, x_2, \ldots, x_n \sim \text{Pois}(\theta)$,
   $\theta \sim \text{Gamma}(p, q)$, $\theta|x \sim \text{Gamma}(p+\sum x_i, q+n)$
3. Normal(variance known): $x_1, x_2, \ldots, x_n \sim \text{N}(\theta, \sigma^2 = 1/\tau)$,
   $\theta \sim \text{N}(\mu_0, \sigma_0^2 = 1/\tau_0)$, $\theta|x \sim \text{N}(\frac{\mu_0\tau_0+n\bar{x}\tau}{\tau_0+n\tau}, \frac{1}{\tau_0+n\tau})$
4. Gamma:$x_1, x_2, \ldots, x_n \sim \text{Gamma}(k, \theta)$,
   $\theta \sim \text{Gamma}(p, q)$, $\theta|x \sim \text{Gamma}(p+nk, q+\sum x_i)$
5. Nagative Binomial: $x \sim \text{NB}(r, \theta)$,
   $\theta \sim \text{Beta}(p, q)$, $\theta|x \sim \text{Beta}(p+r, q+x-r)$

6. Normal(variance unknown): $x_1, x_2, \ldots, x_n \sim \mathrm{N}(\mu, \theta = \sigma^2)$，
   $\theta \sim \mathrm{Scaled-Inv}{-}\chi^2(\nu_0, \sigma_0^2)$，
   $\theta|x \sim \mathrm{Scaled-Inv}{-}\chi^2(\nu_0 + n, \nu_0\sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2)$

3 有信息先验 (Informative Prior)

指数先验（Power Priors）

Power Priors: 把过去的数据以一定权重引入到当前模型的先验分布（Ibrahim and Chen，2000）。
假设 $\boldsymbol{x}_0$ 表示过去的数据，$\boldsymbol{x}$ 表示现在的数据，则先验分布可指定为：

$$\pi(\theta|\boldsymbol{x}_0, a_0) \propto \pi(\theta)[L(\theta|\boldsymbol{x}_0)]^{a_0}$$

其中 $\pi(\theta)$ 是不考虑过去数据而给定的先验分布（比如无信息先验），$L(\theta|\boldsymbol{x}_0)$ 是历史数据的似然函数，$a_0 \in [0,1]$ 是一个常数，表示对历史数据的信任强度。

针对以上先验，模型的后验分布为

$$p(\theta|\boldsymbol{x}) \propto \pi(\theta|\boldsymbol{x}_0, a_0)L(\theta|\boldsymbol{x})$$

其中 $L(\theta|\boldsymbol{x})$ 为针对现有数据的似然函数。

为了避免指定 a_0 的任意性，也可以给定一个分布 $p(a_0)$，从而

$$\pi(\theta|\boldsymbol{x}_0) = \int_0^1 \pi(\theta)[L(\theta|\boldsymbol{x}_0)]^{a_0} p(a_0)\mathrm{d}a_0$$

导出先验（Elicited Priors）

挑战：把“专家意见”转换为“先验分布”

办法：

1. 询问若干专家，给出一些分位数的值 Z_α（如 $\alpha = 0.1$，0.25，0.5，0.75，0.9）
   - 如：25% 的深圳人收入不超过 6000 元，即 $Z_{0.25} = 6000$
2. 假设先验分布为正态分布 $\theta \sim \mathrm{N}(\mu, \sigma^2)$，则 α 分位数与标准正态分布函数的关系为
   $$Z_\alpha = \mu + \Phi^{-1}(\alpha)\sigma$$
3. 根据专家给出的数据，进行简单线性回归 ($y = Z_\alpha, x = \Phi^{-1}(\alpha)$)，可以估计出 μ, σ

如何导出先验分布？

某金融资产每月收益率服从正态分布 $N(\mu, \sigma^2)$（σ^2 为已知），现在需要确定参数的先验分布 $\mu \sim N(\mu_0, \sigma_0^2)$。

根据平时观察，一般人的平均收益率在 3% 左右，大概有四分之一的投资者平均收益不到 1%。

先验分布的导出：

- 取 $\mu_0 = 3\%$；
- 又根据 $P(\mu \le 1) = 0.25$，即 $\Phi[(1-3)/\sigma_0] = 0.25, -2/\sigma_0 = -0.67(\mathrm{qnorm}(0.25)), \sigma_0 = 2.99 \approx 3.$
- 因此选取先验分布为：$\mu \sim N(3, 3^2)$。

最大熵先验

- 目的：仅有部分先验信息，除此之外希望尽可能采用无信息先验
- **熵 (Entropy)：分布中固有不确定性总量的一种度量**。
- **最大熵先验**：在分布满足给定约束（已知先验）条件下，使熵最大化的先验
- Bad news: 连续型有困难
- Good news: 允许加入分布尾部要求（用分位数）

定义 2 (熵). *设 Θ 是离散的，$\pi(\theta)$ 是 Θ 上的 pdf，则分布 π 的熵定义为*

$$\text{Entropy} = -\sum_{\theta\in\Theta} \pi(\theta_i)\log\pi(\theta_i)$$

混合共轭先验

设 $\pi_1(\theta), \pi_2(\theta), \ldots, \pi_m(\theta)$ 为 θ 的 m 个共轭先验,其相应后验分布为 $p_1(\theta|x), p_2(\theta|x), \ldots, p_m(\theta|x)$。
现在考虑一族混合先验分布 $\pi(\theta) = \sum_{i=1}^m w_i\pi_i(\theta)$
则混合后验分布为

$$\begin{aligned} p(\theta|x) &= \pi(\theta)f(x|\theta) = \sum_{i=1}^m w_i\pi_i(\theta)f(x|\theta) \\ &\propto \sum_{i=1}^m w_i^* p_i(\theta|x) \end{aligned}$$

即为混合分布族 (Mixture family)

- 男生与女生成绩明显不同，全班同学的成绩就是两者的混合分布
- 抛一枚硬币，一般出现正面的概率都是 0.5。但是，如果那枚硬币的边缘有破损，则出现**正面**的概率就不是 0.5，也许 0.3，也许 0.7。此时先验分布就应该出现“双峰”

例：混合 beta

假设 $X|\theta \sim \text{Bin}(n,\theta)$, 考虑混合先验：

$$\theta \sim C_1 Beta(a_1, b_1) + C_2\text{Beta}(a_2, b_2)$$

则后验分布为：

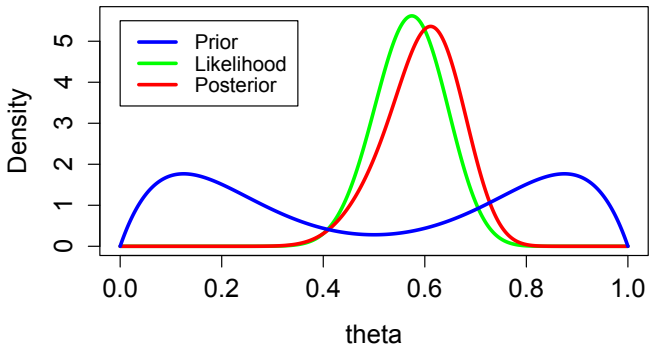
$$\begin{aligned} p(\theta|x) &\propto \pi(\theta)f(x|\theta) \\ &\propto [C_1\theta^{a_1-1}(1-\theta)^{b_1-1} + C_2\theta^{a_2-1}(1-\theta)^{b_2-1}]\theta^x(1-\theta)^{n-x} \\ &= C_1\theta^{a_1+x-1}(1-\theta)^{b_1+n-x-1} + C_2\theta^{a_2+x-1}(1-\theta)^{b_2+n-x-1} \end{aligned}$$

因此 $\theta|x \sim C_1^*\text{Beta}(a_1+x, b_1+n-x) + C_2^*\text{Beta}(a_2+x, b_2+n-x)$

例：混合 beta

总体 $X \sim \text{Bin}(n, \theta)$, $n = 47$, $x = 27$, 混合先验: $\theta \sim .5\text{Beta}(8, 2) + .5\text{Beta}(2, 8)$，则后验分布

$$\theta|x \sim C_1\text{Beta}(35, 22) + C_2\text{Beta}(29, 28)$$



Summary：先验分布的确定

1. 贝叶斯分析：先验信息要用先验分布表示
2. 先验分布的类型
   - 无信息先验（uninformative）
     - proper vs improper
     - 均匀先验（uniform）
     - Jeffreys prior：具有参数不变性的无信息先验
     - 参照先验（Reference priors）
     - 经验贝叶斯 (Empirical Bayesian)
   - 共轭先验（Conjugate prior）：先验与后验分布属于同一个分布族
     - 指数族
   - 有信息先验 (Informative Prior)
     - 指数先验（Power Priors）
     - 导出先验（Elicited Priors）
     - 最大熵先验 (Entropy)
     - 混合先验 (Mixture prior)