
中国石油大学
CHINA UNIVERSITY OF PETROLEUM

# 随机过程作业

题　　目　Bayes的极限分布与极大似然

学　　院　信息科学与工程学院

专　　业　控制科学与工程

姓名学号　孟庆鑫 2018312032

刘芳正 2019310704

韩亚楠 2019310708

任课教师　严彦文

# *Bayes* 的极限分布与极大似然

孟庆鑫　　刘芳正　　韩亚楠
*2018312032*　　*2019310704*　　*2019310708*

**摘　要**: 类比于 *Markov* 链的状态转移算子思路, 本文构造了序贯 *Bayes* 分布列, 以不太严谨的方式证明了 *Bayes* 的极限分布与极大似然的关系: 对于存在极大似然估计的似然函数、任意有界正先验分布下, *Bayes* 极限分布弱收敛于取值在极大似然估计值的 *Dirac* $\delta$ 分布. 共轭先验方法可用于构造特殊的显式序贯 *Bayes* 分布列.

## 1　基本定义

一个 Markov 链 $\mathcal{M}$ 由其初始分布 $\pi(0)$ 和状态转移矩阵 $P$ 完全确定. 若 $\mathcal{M}$ 存在极限分布 $\lim_{n\to\infty}\pi(n)=\pi$, 则 $\pi$ 满足

$$\pi P=\pi, \tag{1}$$

$\pi$ 是平稳分布, 刚好是 $P$ 的特征值 1 对应的特征向量. 一个 Markov 链, 不论初始分布 $\pi(0)$ 是什么, 如果能达到平稳分布, 那么平稳分布由状态转移矩阵 $P$ 确定.

记 $\Theta$ 为分布参数, $\mathcal{D}$ 为样本对应的随机变量, $\mathbb{P}(\Theta)$ 是 $\Theta$ 的先验分布, 根据 Bayes 公式

$$\underbrace{\mathbb{P}(\Theta|\mathcal{D})}_{\text{后验分布}}=\frac{\mathbb{P}(\mathcal{D}|\Theta)\,\mathbb{P}(\Theta)}{\mathbb{P}(\mathcal{D})}\propto\underbrace{\mathbb{P}(\mathcal{D}|\Theta)}_{\text{似然函数}}\,\underbrace{\mathbb{P}(\Theta)}_{\text{先验分布}}, \tag{2}$$

若 $\mathbb{P}(\Theta|\mathcal{D}),\mathbb{P}(\Theta)$ 有相同的形式, 则称 $\mathbb{P}(\Theta)$ 是 $\mathbb{P}(\mathcal{D}|\Theta)$ 关于 $\Theta$ 的共轭先验分布 [1].

## 2　Bayes 的极限分布

将 Markov 链的状态转移矩阵 $P$ 看作算子, 等式 (1) 的 $\pi$ 即为算子 $P$ 的不动点. 一个可达平稳分布的 Markov 链可以看作算子 $P$ 连续作用下趋近不动点的过程.

同样的, 等式 (2) 的先验分布 $\mathbb{P}(\Theta)$ 可以看作是来自于其它数据集 $\mathcal{D}'$ 上的经验, 即 $\mathbb{P}(\Theta) = \mathbb{P}(\Theta|\mathcal{D}')$, 为使先验尽可能符合当前样本, 取 $\mathcal{D}' = \mathcal{D}$, 如果把

$$\frac{(\bullet)\,\mathbb{P}(\mathcal{D}|\Theta)}{\oint_{\Omega_\Theta}(\bullet)\,\mathbb{P}(\mathcal{D}|\Theta)}$$

看作算子 $D$, 这就与 Markov 链很相似了: 算子 $D$ 连续作用在先验分布上, 极限分布趋近于分布空间的某个不动点. 下面通过例子来说明这一观点.

## 2.1 二项分布的例子

考虑 $n$ 次独立重复试验并观测到一组样本, 事件出现的次数记为随机变量 $X$, $X \sim B(n,\theta)$,

$$\mathbb{P}(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},$$

将 $\theta$ 也视为随机变量, $\theta \in (0,1)$, 但 $\theta$ 的分布不清楚, 我们先假设 $\theta \sim \text{Uniform}(0,1)$,

$$\mathbb{P}(\theta|x) = \frac{\mathbb{P}(x|\theta)\,\mathbb{P}(\theta)}{\int_0^1 \mathbb{P}(x|\theta)\,\mathbb{P}(\theta)\,\mathrm{d}\theta} = \frac{\Gamma(n+2)}{\Gamma(x+1)\,\Gamma(n-x+1)}\theta^{(x+1-1)}(1-\theta)^{(n-x+1)-1},$$

即 $\theta|x \sim \text{Beta}(x+1, n-x+1)$. 注意到 $\text{Uniform}(\theta|0,1) = \text{Beta}(\theta|1,1)$, 实际上我们选取均匀分布恰好是共轭先验的一个特例. 为此, 重新假设 $\theta \sim \text{Beta}(\alpha,\beta)$, 同理可得 $\theta|x \sim \text{Beta}(x+\alpha, n-x+\beta)$. 再将此后验分布作为先验分布, 如此, 可构造分布列 $\left\{\text{Beta}(kx+\alpha, k(n-x)+\beta)\right\}_{k=0}^{+\infty}$, 并有 $k \to +\infty$,

$$\text{Beta}(kx+\alpha, k(n-x)+\beta) \rightharpoonup \delta\left(\frac{x}{n}\right), \tag{3}$$

$\delta$ 是 Dirac 函数. (3) 式说明 $\forall \varepsilon > 0$,

$$\mathbb{P}\left(\left|\theta - \frac{x}{n}\right| \leqslant \varepsilon\right) = 1,$$

因此可确定 $\hat{\theta} = \dfrac{x}{n}$.

上面的讨论是在共轭先验构成的分布列上做的, 我们再重新假设 $\theta \sim \mathbb{P}(\theta)$, 其中 $\mathbb{P}(\theta)$ 满足

$$\forall\,\theta \in (0,1),\, \mathbb{P}(\theta) > 0,\, |\mathbb{P}(\theta)| < \infty,\, \text{且} \int_0^1 \mathbb{P}(\theta)\,\mathrm{d}\theta = 1, \tag{4}$$

仍有 $k \to +\infty$,

$$\frac{1}{Z_k(x)}\mathbb{P}(\theta)\left(\theta^x(1-\theta)^{n-x}\right)^k \rightharpoonup \delta\left(\frac{x}{n}\right),$$

其中 $Z_k(x)$ 是归一化系数. 这个结论说明, 不论先验分布如何选, 只要满足 (4) 式, 如此构造的序贯贝叶斯的极限分布都是 $\delta\left(\frac{x}{n}\right)$, 即参数 $\theta$ 在极限分布上只能取 $\frac{x}{n}$, 这个值恰好是 $\theta$ 的极大似然估计. 这里需要注意, 有界分布序列收敛到了无界可积分布 Dirac $\delta$.

## 2.2 Poisson 分布的例子

假设有一组观测样本 $n_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, $i = 1, 2, \cdots, m$, 记 $\boldsymbol{n} = (n_1, n_2, \cdots, n_m)$, 即

$$\mathbb{P}(\boldsymbol{n}|\lambda) = \prod_{i=1}^{m} \frac{\mathrm{e}^{-\lambda}\lambda^{n_i}}{\Gamma(n_i+1)},$$

选取参数 $\lambda$ 的共轭先验 [2]

$$\text{Gamma}(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\lambda}\lambda^{\alpha-1},$$

可以得到后验分布 $\mathbb{P}(\lambda|\boldsymbol{n},\alpha,\beta) = \text{Gamma}(\lambda|\alpha+\sum\boldsymbol{n}, \beta+m)$, 再将此后验分布作为先验分布, 如此, 可构造分布列 $\left\{\text{Gamma}(\lambda|\alpha+k\sum\boldsymbol{n}, \beta+km)\right\}_{k=0}^{+\infty}$, 并有 $k \to +\infty$,

$$\text{Gamma}\left(\lambda|\alpha+k\sum\boldsymbol{n}, \beta+km\right) \rightharpoonup \delta\left(\overline{\boldsymbol{n}}\right).$$

选取先验 $\lambda \sim \mathbb{P}(\lambda)$ 满足

$$\forall\, \lambda \in \mathbb{R}^+, \mathbb{P}(\lambda) > 0,\ |\mathbb{P}(\lambda)| < \infty, \text{ 且 } \int_{\mathbb{R}^+} \mathbb{P}(\lambda)\,\mathrm{d}\lambda = 1,$$

仍有 $k \to +\infty$,

$$\frac{1}{Z_k(\boldsymbol{n})}\mathbb{P}(\lambda)\left(\mathrm{e}^{-\lambda}\lambda^{\overline{\boldsymbol{n}}}\right)^{mk} \rightharpoonup \delta\left(\overline{\boldsymbol{n}}\right),$$

其中 $Z_k(\boldsymbol{n})$ 是归一化系数. 再次注意到, $\overline{\boldsymbol{n}}$ 恰好是 $\lambda$ 的极大似然估计.

## 2.3 正态分布的例子

假设一组观测样本 $x_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, $i = 1, 2, \cdots, n$, 其中 $\mu$ 是未知的, $\sigma^2$ 是已知的, 记 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, 则似然为

$$\mathbb{P}(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2\sigma^2}(x_i-\mu)^2},$$

假设参数 $\mu$ 的共轭先验为 $\mathcal{N}\left(\mu|\mu_k, \sigma_k^2\right)$[3], 则可计算后验分布为 $\mathcal{N}\left(\mu|\mu_{k+1}, \sigma_{k+1}^2\right)$, 满足递推关系

$$\frac{1}{\sigma_{k+1}^2} = \frac{1}{\sigma_k^2} + \frac{n}{\sigma^2}, \quad \frac{\mu_{k+1}}{\sigma_{k+1}^2} = \frac{\mu_k}{\sigma_k^2} + \frac{n\overline{\boldsymbol{x}}}{\sigma^2},$$

从而有

$$\frac{1}{\sigma_k^2}=\frac{1}{\sigma_0^2}+\frac{kn}{\sigma^2},\quad \frac{\mu_k}{\sigma_k^2}=\frac{\mu_0}{\sigma_0^2}+\frac{kn\overline{x}}{\sigma^2},$$

因此

$$\mathcal{N}\left(\mu|\mu_k,\sigma_k^2\right)=\mathcal{N}\left(\mu\left|\frac{\mu_0\sigma^2+kn\overline{x}\sigma_0^2}{\sigma^2+kn\sigma_0^2},\left(\frac{1}{\sigma_0^2}+\frac{kn}{\sigma^2}\right)^{-1}\right.\right)\underset{k\to+\infty}{\rightharpoonup}\delta\left(\overline{x}\right).$$

对 $\mu$ 的任意先验不再特别讨论, 而是统一的放在指数家族分布里说明.

## 2.4 指数家族的例子

考虑指数家族分布 [1]

$$\mathbb{P}\left(X|\eta\right)=h\left(X\right)\mathrm{e}^{\eta^TT(X)-A(\eta)}$$

及一组观测样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$, 似然函数为

$$\mathbb{P}\left(\boldsymbol{x}|\eta\right)=\mathrm{e}^{\sum_{i=1}^n\eta^TT(x_i)-nA(\eta)}\prod_{i=1}^nh\left(x_i\right),\tag{5}$$

选取先验 $\eta\sim\mathbb{P}\left(\eta\right)$ 满足

$$\forall\,\eta\in\Omega_\eta,\mathbb{P}\left(\eta\right)>0,\ \left|\mathbb{P}\left(\eta\right)\right|<\infty,\ \text{且}\int_{\Omega_\eta}\mathbb{P}\left(\eta\right)\,\mathrm{d}\eta=1,$$

注意 Bayes 分布列

$$\left\{\frac{1}{Z_k\left(\boldsymbol{x}\right)}\mathbb{P}\left(\eta\right)\left(\mathrm{e}^{\eta^T\overline{T(x)}-A(\eta)}\right)^{nk}\right\}_{k=0}^{\infty},$$

记 $f\left(\eta\right)=\mathrm{e}^{\eta^T\overline{T(x)}-A(\eta)}$ 的最大值在 $\eta^\star$ 处取得, 一方面, $\eta^\star$ 是 (5) 式的极大似然估计, 一方面, 考虑 $f\in C\left(\Omega_\eta\right)$, $\exists\zeta>0$, 使得 $\forall\,\varepsilon>0,\varepsilon<\zeta,\forall\,\xi\in B\left(\eta^\star,\varepsilon\right),\forall\,\eta\notin B\left(\eta^\star,\varepsilon\right)$, $f\left(\xi\right)>f\left(\eta\right)$, 则

$$\lim_{k\to+\infty}\mathbb{P}\left(\frac{\mathbb{P}\left(\eta\right)f^{nk}\left(\eta\right)}{Z_k\left(\boldsymbol{x}\right)}\in B\left(\eta^\star,\,\varepsilon\right)\right)=\lim_{k\to+\infty}\int_{B(\eta^\star,\varepsilon)}\frac{\mathbb{P}\left(\eta\right)f^{nk}\left(\eta\right)}{\displaystyle\int_{\Omega_\eta}\mathbb{P}\left(\eta\right)f^{nk}\left(\eta\right)\,\mathrm{d}\eta}\,\mathrm{d}\eta$$

$$\overset{\exists\,\vartheta\in B(\eta^\star,\varepsilon)}{=\!=\!=}\lim_{k\to+\infty}\frac{\displaystyle\int_{B(\eta^\star,\varepsilon)}\mathbb{P}\left(\eta\right)\,\mathrm{d}\eta}{\displaystyle\int_{\Omega_\eta\setminus B(\eta^\star,\varepsilon)}\mathbb{P}\left(\eta\right)\left(\frac{f\left(\eta\right)}{f\left(\vartheta\right)}\right)^{nk}\,\mathrm{d}\eta+\int_{B(\eta^\star,\varepsilon)}\mathbb{P}\left(\eta\right)\,\mathrm{d}\eta}=1,$$

这说明 $k\to+\infty$,

$$\frac{1}{Z_k\left(\boldsymbol{x}\right)}\mathbb{P}\left(\eta\right)f^{nk}\left(\eta\right)\rightharpoonup\delta\left(\eta^\star\right),$$

于是得到结论: 指数家族分布的 Bayes 极限分布弱收敛于 Dirac $\delta\left(\eta^{\star}\right)$ 分布, $\eta^{\star}$ 恰好是极大似然估计值.

## 3 结论

将 Bayes 后验分布当作先验分布, 构造了类似 Markov 链的分布列, 对于存在极大似然估计的似然函数, 任意有界正先验分布, 其 Bayes 极限分布弱收敛于 Dirac $\delta(\theta)$ 分布, 其中 $\theta$ 恰好是唯一的极大似然估计值, 即 $\forall \Theta \in \Omega_{\Theta}, \mathbb{P}(\Theta)>0,|\mathbb{P}(\Theta)|<\infty$, 且 $\oint_{\Omega_{\Theta}} \mathbb{P}(\Theta)=1, \exists ! \theta=\arg \max _{\Theta} \mathbb{P}(\mathcal{D} | \Theta)$, 记 $f(\Theta)=\mathbb{P}(\mathcal{D} | \Theta), f \in C\left(\Omega_{\Theta}\right)$, $\exists \zeta>0$, 使得 $\forall \varepsilon>0, \varepsilon<\zeta, \forall \xi \in B(\theta, \varepsilon), \forall \eta \notin B(\theta, \varepsilon), f(\xi)>f(\eta)$, 则有

$$\left(\frac{(\bullet) \mathbb{P}(\mathcal{D} | \Theta)}{\oint_{\Omega_{\Theta}}(\bullet) \mathbb{P}(\mathcal{D} | \Theta)}\right)^{k} \mathbb{P}(\Theta) \underset{k \to+\infty}{\rightharpoonup} \delta(\theta),$$

且有推论

$$\arg \max _{\Theta}\left(\frac{(\bullet) \mathbb{P}(\mathcal{D} | \Theta)}{\oint_{\Omega_{\Theta}}(\bullet) \mathbb{P}(\mathcal{D} | \Theta)}\right)^{k} \mathbb{P}(\Theta) \underset{k \to+\infty}{\longrightarrow} \theta.$$

证明过程类似于指数家族分布例子的证明,

$$\lim _{k \to+\infty} \mathbb{P}\left(\left(\frac{(\bullet) \mathbb{P}(\mathcal{D} | \Theta)}{\oint_{\Omega_{\Theta}}(\bullet) \mathbb{P}(\mathcal{D} | \Theta)}\right)^{k} \mathbb{P}(\Theta) \in B(\theta, \varepsilon)\right)=\lim _{k \to+\infty} \mathbb{P}\left(\frac{\mathbb{P}(\Theta) f^{k}(\Theta)}{Z_{k}(\mathcal{D})} \in B(\theta, \varepsilon)\right)=1,$$

以及 $\exists K$, 使 $k>K$,

$$\arg \max _{\Theta}\left(\frac{(\bullet) \mathbb{P}(\mathcal{D} | \Theta)}{\oint_{\Omega_{\Theta}}(\bullet) \mathbb{P}(\mathcal{D} | \Theta)}\right)^{k} \mathbb{P}(\Theta) \in B(\theta, \varepsilon),$$

再由 $\varepsilon$ 的任意性可得结论.

本文基于以下想法: 当 Markov 链到达极限分布的时候, 极限分布就是平稳分布, 并只由状态转移矩阵决定, 亦可看作状态转移矩阵的特征, 与初始分布无关; 从 Bayes 估计的角度看, 先验经过似然算子反复作用到达极限分布, 那么极限分布也由似然算子决定, 亦可看作似然算子的特征, 与初始先验无关, 似然由样本生成, 这意味着经过无穷次数据作用得到的后验, 将“抹去”先验的痕迹.

从结论来看, 任何 Dirac $\delta$ 分布都是似然算子的不动点, 这意味着初始先验分布不能选择 Dirac $\delta$ 分布, 选取有界正先验的意义在于, 先验是依赖有限次数据作用后得到, 并且对随机变量的任何取值均有概率, 而 Dirac $\delta$ 分布意味着经历了无限的数据作用, 变量已不再随机.

本文在举例中使用了共轭先验, 该方法可用于构造特殊的显式序贯 Bayes 分布列, 尤其是正态分布的例子, 可以直接看出极限分布. 对于非共轭先验分布, 先验后验之间一般不具备良好的解析性质.

# References


[1] C. Bishop and S. ligne, *Pattern Recognition and Machine Learning*, vol. 1. 01 2006.

[2] "Gamma distribution." `https://en.wikipedia.org/wiki/Gamma_distribution`.

[3] K. Murphy, "Conjugate bayesian analysis of the gaussian distribution," pp. 445–470, 11 2007.