拟合概率模型

这一章讨论数据$\{x_i\}_{i=1}^{I}$的拟合概率模型。由于拟合时需要学习模型的参数$\boldsymbol{\theta}$，因此这一过程称为学习，本章也讨论计算新数据$\boldsymbol{x}^*$在最终模型下的概率。这称作评估**预测分布**。主要分析三个方法：**最大似然法、最大后验法、贝叶斯方法。**

最大似然法

最大似然估计（Maximum Likelihood Estimation, MLE）是统计学中常用的参数估计方法，其核心思想是：**在给定观测数据的情况下，寻找一组模型参数，使得该数据出现的概率（或概率密度）最大化**。换句话说，MLE试图找到最可能生成观测数据的参数值。

1. 基本概念

- **似然函数（Likelihood Function）**：给定参数 θ 和观测数据 D，似然函数 $L(\theta; D)$ 表示的是在参数 θ 下数据 D 出现的概率（离散分布）或概率密度（连续分布）。
  例如，若数据独立同分布（i.i.d.），似然函数可表示为：

  $$L(\theta; D) = p(\boldsymbol{x}_{1\ldots I}; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

  其中 $p(x_i; \theta)$ 是单个样本的概率（或概率密度函数）。

  参数的最大似然估计是：

  $$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} [p(\boldsymbol{x}_{\boldsymbol{1}\ldots \boldsymbol{I}} | \boldsymbol{\theta})] = \max_{\boldsymbol{\theta}} \left[\prod_{i=1}^{I} p(x_i | \boldsymbol{\theta}) \right]$$

- **对数似然函数（Log-Likelihood）**：为简化计算（尤其是连乘运算），通常对似然函数取对数：

  $$\ell(\theta; D) = \log L(\theta; D) = \sum_{i=1}^{n} \log p(x_i; \theta)$$

  对数函数是单调递增的，因此最大化对数似然等价于最大化原似然函数。

2. 求解步骤

1. **定义模型**
   假设数据服从某个分布（如高斯分布、二项分布等），并确定参数 θ。例如：抛硬币实验中，假设正面向上的概率为 θ，则单次试验的概率为 $p(x; \theta) = \theta^x (1-\theta)^{1-x}$，其中 $x \in \{0, 1\}$。
2. **写出似然函数**：对于观测数据 $D = \{x_1, x_2, \ldots, x_n\}$，似然函数为：

   $$L(\theta; D) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$$

3. **取对数**：对数似然函数为：

   $$\ell(\theta; D) = \sum_{i=1}^{n} [x_i \log \theta + (1-x_i) \log(1-\theta)]$$

4. **求导并解方程**：对 θ 求导并令导数为0，找到极值点：

$$\frac{d\ell}{d\theta} = \sum_{i=1}^{n} \left(\frac{x_i}{\theta} - \frac{1 - x_i}{1 - \theta} \right) = 0$$

解得：

$$\hat{\theta}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

即正面向上的频率。

5. **验证极值**：通常通过二阶导数或边界条件验证是否为最大值。

3. 高斯分布的MLE示例

假设数据服从正态分布 $N(\mu, \sigma^2)$，参数 $\theta = (\mu, \sigma^2)$。

- 似然函数：

$$L(\mu, \sigma^2; D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

- 对数似然函数：

$$\ell(\mu, \sigma^2; D) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

- 求导并解得：

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

注意：MLE估计的方差是有偏的（分母为 n），而样本方差通常用 $n-1$ 来无偏估计。

4. MLE的性质

- **一致性**：当样本量足够大时，MLE估计值收敛到真实参数。
- **渐近正态性**：MLE估计值在大样本下近似服从正态分布。
- **渐近有效性**：MLE在大样本下能达到最小方差下界（Cramér-Rao下界）。
- **可能有偏**：小样本时可能有偏（如高斯方差估计）。

5. 局限性

- **对模型假设敏感**：若假设的分布形式错误（如数据实际服从重尾分布却假设高斯分布），结果可能偏差较大。
- **小样本偏差**：如方差估计的偏差问题。
- **计算复杂性**：某些模型（如隐变量模型）的MLE需要EM算法等迭代方法。
- **过拟合风险**：在高维数据或复杂模型中，可能过度适应训练数据。

最大后验法

最大后验估计（Maximum A Posteriori Estimation, MAP）是贝叶斯统计框架中的一种参数估计方法，其核心思想是：**在给定观测数据的条件下，寻找后验概率最大的参数值**。与最大似然估计（MLE）不同，MAP不仅依赖于数据本身，还引入了参数的先验分布（Prior Distribution），从而结合了先验知识与数据信息。

1. 贝叶斯定理基础

MAP 的核心是贝叶斯定理（Bayes' Theorem）：

$$p(\theta \mid D)=\frac{p(D \mid \theta) p(\theta)}{p(D)}$$

其中：

- $p(\theta \mid D)$：后验分布（Posterior），即给定数据 D 时参数 θ 的概率分布。
- $p(D \mid \theta)$：似然函数（Likelihood），即在参数 θ 下数据 D 出现的概率。
- $p(\theta)$：先验分布（Prior），即参数 θ 的先验知识（在观测数据前的信念）。
- $p(D)$：证据（Evidence），是数据 D 的边缘概率，与参数 θ 无关，可视为归一化常数。

MAP 的目标是找到使后验概率 $p(\theta \mid D)$ 最大的参数 θ：

$$\hat{\theta}_{\mathrm{MAP}}=\arg \max _\theta p(\theta \mid D)$$

由于 $p(D)$ 是常数，可以忽略，因此等价于：

$$\hat{\theta}_{\mathrm{MAP}}=\arg \max _\theta[p(D \mid \theta) p(\theta)]$$

2. MAP 与 MLE 的关系

- **MLE 的目标**：最大化似然函数 $p(D \mid \theta)$。
- **MAP 的目标**：最大化后验概率 $p(\theta \mid D)$，即同时考虑似然 $p(D \mid \theta)$ 和先验 $p(\theta)$。

如果先验 $p(\theta)$ 是均匀分布（即无信息先验，Uniform Prior），则 MAP 等价于 MLE：

$$\hat{\theta}_{\mathrm{MAP}}=\arg \max _\theta[p(D \mid \theta) \cdot \operatorname{Uniform}(\theta)]=\arg \max _\theta p(D \mid \theta)=\hat{\theta}_{\mathrm{MLE}}$$

3. 数学推导

假设数据独立同分布（i.i.d.），且参数 θ 的先验分布为 $p(\theta)$，则 MAP 的优化目标为：

$$\hat{\theta}_{\mathrm{MAP}}=\arg \max _\theta\left[\prod_{i=1}^n p\left(x_i \mid \theta\right) \cdot p(\theta)\right]$$

取对数后转化为：

$$\hat{\theta}_{\mathrm{MAP}}=\arg \max _\theta\left[\sum_{i=1}^n \log p\left(x_i \mid \theta\right)+\log p(\theta)\right]$$

即：

$$\hat{\theta}_{\mathrm{MAP}}=\arg \max _\theta[\ell(\theta ; D)+\log p(\theta)]$$

其中 $\ell(\theta ; D)$ 是对数似然函数。

4. 示例：抛硬币实验

假设我们通过抛硬币实验估计正面向上的概率 $\theta \in[0,1]$，数据 $D=\left\{x_1, x_2, \ldots, x_n\right\}$，其中 $x_i \in\{0,1\}$（1表示正面，0表示反面）。

(1) MLE 解法

- 似然函数：$p(D \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$
- 对数似然：$\ell(\theta; D) = \sum_{i=1}^{n} [x_i \log\theta + (1-x_i)\log(1-\theta)]$
- MLE 解：$\hat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$

(2) MAP 解法

- 先验分布：假设 $\theta \sim \text{Beta}(\alpha, \beta)$，其概率密度函数为：

$$p(\theta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- 对数先验：$\log p(\theta) = (\alpha-1)\log\theta + (\beta-1)\log(1-\theta) + \text{常数}$
- MAP 优化目标：

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta}\left[\sum_{i=1}^{n}\ell(\theta; D) + \log p(\theta)\right] \\ &= \arg\max_{\theta}\left[\sum_{i=1}^{n}(x_i\log\theta + (1-x_i)\log(1-\theta)) + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)\right]\end{aligned}$$

- 求导并令导数为0：

$$\frac{d}{d\theta}\left[\sum_{i=1}^{n}x_i\log\theta + \sum_{i=1}^{n}(1-x_i)\log(1-\theta) + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)\right] = 0$$

  解得：

$$\hat{\theta}_{\text{MAP}} = \frac{\sum_{i=1}^{n}x_i + \alpha - 1}{n + \alpha + \beta - 2}$$

(3) 结果分析

- 当 $\alpha = \beta = 1$（均匀先验），MAP 退化为 MLE：

$$\hat{\theta}_{\text{MAP}} = \frac{\sum_{i=1}^{n}x_i}{n}$$

- 当 $\alpha = 2, \beta = 2$（偏好中间值的先验），MAP 估计值会向 0.5 偏移：

$$\hat{\theta}_{\text{MAP}} = \frac{\sum_{i=1}^{n}x_i + 1}{n+2}$$

5. 高斯分布的 MAP 估计

假设数据 $D = \{x_1, x_2, \ldots, x_n\}$ 服从高斯分布 $\mathcal{N}(\mu, \sigma^2)$，且已知方差 σ^2，需要估计均值 μ。假设 μ 的先验为 $\mathcal{N}(\mu_0, \tau^2)$。

(1) 后验分布推导

根据贝叶斯定理，后验分布 $p(\mu \mid D)$ 也是高斯分布，其均值和方差为：

$$\mu_{\text{post}} = \frac{\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \quad \sigma^2_{\text{post}} = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

其中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$。

(2) MAP 估计

由于高斯分布的峰值（均值）即 MAP 估计值：

$$\hat{\mu}_{\text{MAP}} = \mu_{\text{post}} = \frac{\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

这表明 MAP 估计是数据均值 $\bar{x}$ 和先验均值 μ_0 的加权平均，权重取决于数据量 n 和先验精度（方差倒数）。

6. MAP 与正则化的关系

在机器学习中，MAP 可以看作是对 MLE 的正则化扩展：

- **先验分布对应正则项**：例如，高斯先验 $\mathcal{N}(0, \tau^2)$ 相当于 L2 正则化（Ridge Regularization）。
- **稀疏先验**：Laplace 先验（双指数分布）对应 L1 正则化（Lasso Regularization）。

例如，线性回归的 MAP 解：

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \left[\sum_{i=1}^{n} (y_i - \theta^T x_i)^2 + \lambda \|\theta\|^2 \right]$$

其中 λ 是正则化系数，对应先验分布的精度。

7. MAP 的优缺点

优点

- **结合先验知识**：在数据量小或噪声大时，先验可以提供稳定性。
- **缓解过拟合**：通过引入先验（如稀疏先验），可防止模型过度复杂化。
- **贝叶斯框架**：与贝叶斯推断无缝衔接，支持不确定性量化。

缺点

- **先验选择敏感**：结果依赖于先验分布的合理性，不当的先验可能导致偏差。
- **计算复杂性**：非共轭先验可能导致后验分布难以解析求解，需依赖近似方法（如变分推断、MCMC）。
- **点估计局限性**：MAP 仅提供单个最优参数值，无法完全反映参数的不确定性（相比贝叶斯后验采样）。

贝叶斯方法

贝叶斯方法（Bayesian Inference）是统计学中一种基于**贝叶斯定理**的参数估计和推断方法，其核心思想是：**将参数视为随机变量，通过观测数据不断更新参数的概率分布**。与最大似然估计（MLE）和最大后验估计（MAP）不同，贝叶斯方法不仅给出参数的点估计，还提供参数的完整概率分布（后验分布），从而量化参数的不确定性。

1. 贝叶斯定理的核心

贝叶斯定理的数学形式为：

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

其中：

- $p(\theta \mid D)$：**后验分布**（Posterior），即给定数据 D 时参数 θ 的概率分布。
- $p(D \mid \theta)$：**似然函数**（Likelihood），即在参数 θ 下数据 D 出现的概率。
- $p(\theta)$：**先验分布**（Prior），即参数 θ 的先验知识（在观测数据前的信念）。
- $p(D)$：**证据**（Evidence），是数据 D 的边缘概率，可视为归一化常数：

$$p(D) = \int p(D \mid \theta)p(\theta)d\theta$$

贝叶斯方法的核心目标是**通过数据 D 更新先验分布 $p(\theta)$，得到后验分布 $p(\theta \mid D)$**。

2. 贝叶斯推断的步骤

1. **选择先验分布 $p(\theta)$**
   根据领域知识或无信息先验（如均匀分布）定义参数的初始分布。
2. **定义似然函数 $p(D \mid \theta)$**
   假设数据服从某个概率模型（如高斯分布、伯努利分布等）。
3. **计算后验分布 $p(\theta \mid D)$**
   利用贝叶斯定理结合先验和似然，求解后验分布：

   $$p(\theta \mid D) \propto p(D \mid \theta)p(\theta)$$

   注意：实际计算中通常忽略归一化常数 $p(D)$，因为可以通过采样或解析方法处理。
4. **推断与预测**
   - **点估计**：从后验分布中提取均值、中位数或最大后验（MAP）作为参数的估计值。
   - **不确定性量化**：通过后验分布的方差、置信区间（可信区间）描述参数的不确定性。
   - **预测分布**：对新数据 x_{new} 的预测通过积分后验分布获得：

$$\begin{aligned} p(x_{\text{new}} \mid D) &= \int p(x_{\text{new}}, \theta \mid D)d\theta \\ &= \int p(x_{\text{new}} \mid \theta, D) \cdot p(\theta \mid D)d\theta \\ &= \int p(x_{\text{new}} \mid \theta) \cdot p(\theta \mid D)d\theta \end{aligned}$$

3. 非共轭先验与近似方法

当先验和似然不满足共轭关系时，后验分布通常无法解析求解，需借助近似方法：

(1) 变分推断（Variational Inference）

- **思想**：将后验分布近似为一个简单的分布族（如高斯分布），通过优化参数使近似分布与真实后验的 KL 散度最小。
- **步骤**：
  1. 假设近似分布 $q(\theta) \in \mathcal{Q}$。
  2. 最小化 KL 散度：

     $$q^*(\theta) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\theta) \parallel p(\theta \mid D))$$

  3. 转化为最大化证据下界（ELBO）：

     $$\log p(D) = \mathrm{KL}(q(\theta) \parallel p(\theta \mid D)) + \mathbb{E}_q[\log p(D, \theta)] - \mathbb{E}_q[\log q(\theta)]$$

(2) 马尔可夫链蒙特卡洛（MCMC）

- **思想**：通过构造马尔可夫链生成后验分布的样本，用于近似后验分布。
- **常用算法**：
  - Metropolis-Hastings 算法
  - Gibbs 采样
- **优点**：无需归一化常数 $p(D)$，适用于复杂模型。

4. 贝叶斯预测与模型比较

(1) 预测分布

贝叶斯方法通过积分后验分布进行预测：

$$p(x_{\text{new}} \mid D) = \int p(x_{\text{new}} \mid \theta) p(\theta \mid D) d\theta$$

这反映了参数不确定性对预测的影响。

(2) 模型比较

贝叶斯框架可通过**边缘似然**（Marginal Likelihood）比较不同模型 $M_1, M_2, \ldots$：

$$p(D \mid M_i) = \int p(D \mid \theta, M_i) p(\theta \mid M_i) d\theta$$

模型的后验概率为：

$$p(M_i \mid D) \propto p(D \mid M_i) p(M_i)$$

其中 $p(M_i)$ 是模型的先验概率。

5. 贝叶斯方法的优缺点

优点

- **不确定性量化**：提供参数的完整分布，而非单一估计值。
- **结合先验知识**：通过先验分布融入领域知识或专家经验。
- **灵活建模**：适用于复杂模型（如层次模型、非参数模型）。
- **自动防止过拟合**：先验可作为正则化项，约束参数空间。

缺点

- **计算复杂性**：非共轭模型需依赖近似方法（如 MCMC、变分推断），计算成本高。
- **先验敏感性**：结果依赖于先验选择的合理性。
- **解释性挑战**：后验分布可能难以直观解释，尤其是高维参数。

举例：一元正态分布

我们以**一元正态分布**（Normal Distribution）为例，详细说明最大似然估计（MLE）、最大后验估计（MAP）和贝叶斯方法（Bayesian Inference）如何用于参数拟合。假设数据 $D = \{x_1, x_2, \ldots, x_n\}$ 独立同分布（i.i.d.），服从正态分布 $\mathcal{N}(\mu, \sigma^2)$，目标是估计均值 μ 和方差 σ^2。

1. 最大似然估计（MLE）

(1) 似然函数

正态分布的概率密度函数为：

$$p(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

似然函数为：

$$L(\mu, \sigma^2; D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

取对数似然函数：

$$\ell(\mu, \sigma^2; D) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

(2) 求解步骤

1. **对 μ 求导并令导数为0**：

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2. **对 σ^2 求导并令导数为0**：

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \Rightarrow \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{\text{MLE}})^2$$

(3) 结果

- **均值**：$\hat{\mu}_{\text{MLE}} = \bar{x}$（样本均值）。
- **方差**：$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$（有偏估计，分母为 n）。

(4) 特点

- **无先验**：纯数据驱动，不引入任何先验知识。
- **有偏性**：MLE 的方差估计是有偏的（真实方差的无偏估计分母为 $n-1$）。
- **计算简单**：解析解直接通过样本统计量得到。

2. 最大后验估计（MAP）

(1) 先验选择

假设：

- 均值 $\mu \sim \mathcal{N}(\mu_0, \tau^2)$（已知方差 σ^2）。
- 方差 σ^2 已知（简化推导）。

(2) 后验分布

根据贝叶斯定理：

$$p(\mu \mid D) \propto p(D \mid \mu) p(\mu)$$

- 似然函数：$p(D \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$
- 先验分布：$p(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\tau^2}\right)$

对数后验：

$$\log p(\mu \mid D) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{2\tau^2} (\mu - \mu_0)^2$$

(3) 求解 MAP

对 μ 求导并令导数为0：

$$\frac{d}{d\mu} \left[-\frac{n}{2\sigma^2} (\bar{x} - \mu)^2 - \frac{1}{2\tau^2} (\mu - \mu_0)^2 \right] = 0$$

解得：

$$\hat{\mu}_{\text{MAP}} = \frac{\frac{n}{\sigma^2} \bar{x} + \frac{1}{\tau^2} \mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

(4) 结果

- **均值**：$\hat{\mu}_{\text{MAP}}$ 是数据均值 $\bar{x}$ 和先验均值 μ_0 的加权平均，权重由数据量 n 和先验精度（方差倒数）决定。
- **方差**：假设已知，无需估计。

(5) 特点

- **引入先验**：通过先验 μ_0 和 τ^2 调节估计值，防止过拟合。
- **正则化效应**：先验相当于对 μ 的 L2 正则化（Ridge Regularization）。
- **解析解**：在共轭先验下可解析求解。

3. 贝叶斯方法（Bayesian Inference）

(1) 共轭先验选择

假设：

- 均值 $\mu \sim \mathcal{N}(\mu_0, \tau^2)$
- 方差 $\sigma^2 \sim \text{Inverse-Gamma}(\alpha, \beta)$（逆伽马分布）

联合先验为：

$$p(\mu, \sigma^2) = p(\mu \mid \sigma^2) p(\sigma^2)$$

(2) 后验分布

后验分布为：

$$p(\mu, \sigma^2 \mid D) \propto p(D \mid \mu, \sigma^2) p(\mu \mid \sigma^2) p(\sigma^2)$$

在正态-逆伽马先验下，后验分布仍为正态-逆伽马分布：

- $\mu \mid \sigma^2, D \sim \mathcal{N}(\mu_n, \sigma^2/\kappa_n)$
- $\sigma^2 \mid D \sim \text{Inverse-Gamma}(\alpha_n, \beta_n)$

其中：

- $\mu_n = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_0 + n}$
- $\kappa_n = \kappa_0 + n$
- $\alpha_n = \alpha_0 + \frac{n}{2}$
- $\beta_n = \beta_0 + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{\kappa_0 n (\bar{x} - \mu_0)^2}{2(\kappa_0 + n)}$

(3) 预测分布

新数据 x_{new} 的预测分布为：

$$p(x_{\text{new}} \mid D) = \int \int p(x_{\text{new}} \mid \mu, \sigma^2) p(\mu, \sigma^2 \mid D) \, d\mu \, d\sigma^2$$

在正态-逆伽马后验下，预测分布为学生t分布（Student-t Distribution）：

$$x_{\text{new}} \mid D \sim t_{2\alpha_n} \left(\mu_n, \frac{\beta_n}{\alpha_n \kappa_n} \right)$$

(4) 特点

- **参数分布**：提供 μ 和 σ^2 的完整后验分布，而非点估计。
- **不确定性量化**：通过后验分布的方差和置信区间描述参数的不确定性。
- **计算复杂性**：需处理多维积分，通常依赖MCMC或变分推断，除非使用共轭先验。

4. 三种方法对比总结

方法	是否使用先验	参数估计形式	计算复杂度	是否量化不确定性
MLE	否	点估计	低	否
MAP	是	点估计	中	否
贝叶斯方法	是	后验分布	高	是

关键区别

1. **MLE**：纯数据驱动，无偏但方差估计有偏，适用于数据量大且无先验信息的场景。
2. **MAP**：结合先验知识，防止过拟合，相当于正则化，但仅提供点估计。
3. **贝叶斯方法**：提供参数的完整后验分布，量化不确定性，适用于小样本或需要风险评估的场景。

5. 实际应用示例

假设观测数据 $D = \{1.2, 1.5, 1.8, 2.0, 2.3\}$，真实分布为 $\mathcal{N}(2, 0.1^2)$。

- **MLE**：$\hat{\mu}_{\text{MLE}} = 1.76$，$\hat{\sigma}^2_{\text{MLE}} = 0.098$。
- **MAP**（假设 $\mu_0 = 1.5, \tau^2 = 0.1$）：$\hat{\mu}_{\text{MAP}} \approx 1.73$。

- **贝叶斯方法**（假设正态-逆伽马先验）：后验分布 $\mu \sim \mathcal{N}(1.73, 0.05^2)$，预测分布为t分布。

通过对比可见，MAP 的估计值受先验影响，而贝叶斯方法提供更丰富的分布信息，适用于需要概率预测的场景（如金融风险评估）。