

第四章 参数估计

- 第四章 参数估计
 - 4.1 点估计与区间估计
 - 4.1.1 点估计的定义
 - 4.1.2 区间估计的定义
 - 4.1.3 构造置信区间的常用方法
 - 4.2 参数估计与状态估计的关系
 - 4.2.1 参数估计
 - 4.2.2 状态估计
 - 4.2.3 联系与相互转化
 - 4.3 频率派与贝叶斯派方法
 - 4.3.1 历史与哲学基础
 - 4.3.2 方法特点
 - 4.3.3 对比与联系
 - 4.3.4 总体均值的两派估计
 - 4.4 估计量的基本性质
 - 4.4.1 无偏性、一致性、有效性
 - 1) 无偏性 (Unbiasedness)
 - 2) 一致性 (Consistency)
 - 3) 有效性 (Efficiency)
 - 4) 例题
 - 4.4.2 均值方差均衡
 - 1) 均方误差 (Mean Squared Error, MSE)
 - 2) 例题
 - 4.4.4 Cramér–Rao 下界 (CRLB)
 - 1) 单确定参数估计量的Cramér–Rao下界
 - 2) 单随机参数估计量的Cramér–Rao下界
 - 3) 例题
 - 4.4.5 多参数估计的Cramér–Rao 下界
 - 1) 定义
 - 2) 性质
 - 3) 例题
 - 4.5 最小方差无偏估计量与统计充分性完备性
 - 4.5.1 最小方差无偏估计量 (UMVUE)
 - 1) 定义
 - 2) 存在性分析
 - 4.5.2 充分性与因子分解定理
 - 1) 充分性
 - 2) 因子分解定理 (Neyman 因子分解定理)
 - 3) 例题
 - 4.5.3 Rao–Blackwell 定理
 - 1) 定理
 - 2) 例题
 - 4.5.4 完备性与指数分布族
 - 1) 完备性

- 2) 指数分布族
- 3) 指数分布族的完备性
- 4.5.5 Lehmann–Scheffé 定理
 - 1) 定理
 - 2) 例题
- 4.6 经典频率派估计方法
 - 4.6.1 极大似然估计 (MLE)
 - 1) 基本概念与似然函数
 - 2) 求解方法与数值算法
 - 3) 大样本理论: MLE的优良性质
 - 4.6.2 最小二乘估计 (LS)
 - 1) 普通最小二乘 (OLS)
 - 2) Gauss-Markov 定理
 - 3) 最小二乘与极大似然的关系
 - 4.6.3 加权最小二乘 (WLS)
 - 1) 异方差问题与WLS的引入
 - 2) WLS的估计公式与推导
 - 3) 实例分析与计算
 - 4.6.4 极大极小估计 (Minimax Estimation)
 - 1) 稳健决策的基本思想
 - 2) 极大极小准则的数学表述
 - 3) 实例分析: 均匀分布参数的极大极小估计
 - 4.6.5 小结
- 4.7 贝叶斯估计与均方误差准则
 - 4.7.1 贝叶斯估计
 - 1) 贝叶斯决策理论与风险最小化
 - 2) 不同损失函数下的贝叶斯估计
 - 3) 贝叶斯估计的实例分析
 - 4.7.2 最大后验估计 (MAP, Maximum A Posteriori)
 - 1) 贝叶斯框架与MAP的基本思想
 - 2) 正态分布情形下的MAP估计
 - 3) MAP估计的性质与讨论
 - 4.7.3 最小均方误差估计 (MMSE, Minimum Mean Square Error)
 - 1) MMSE估计的基本定义与最优性
 - 2) MMSE估计的性质与特点
 - 3) 线性最小均方误差估计 (LMMSE)
 - 4.7.4 经验贝叶斯与收缩估计
 - 1) 经验贝叶斯方法的基本框架
 - 2) James-Stein收缩估计
 - 3) 理论性质与实例分析
 - 4.7.5 总结
- 4.8 不完全数据条件下的估计
 - 4.8.1 EM算法
 - 1) EM算法的基本框架与动机
 - 2) EM算法的数学表述与收敛性证明
 - 3) 高斯混合模型的EM算法实例
 - 4.8.2 广义EM (GEM)
 - 1) 广义EM的基本框架与动机

- 2) 自由能下界与变分推断视角
- 3) 广义EM的变体与实现方法
- 4.8.3 Wake-Sleep算法与广义EM
 - 1) 深度生成模型与变分推断挑战
 - 2) Wake-Sleep算法的双向优化框架
 - 3) Wake-Sleep算法步骤与物理意义
 - 4) Wake-Sleep与广义EM的一致性
- 4.9 鲁棒估计方法
 - 4.9.1 鲁棒估计的基本概念与意义
 - 4.9.2 鲁棒性的理论基础与评价指标
 - 4.9.3 经典鲁棒估计方法
 - 4.9.4 现代鲁棒回归方法
 - 4.9.5 鲁棒位置估计的比较与应用
- 4.10 特殊数据结构下的估计特性
 - 4.10.1 独立同分布样本的经典理论
 - 4.10.2 时间序列数据的估计特性
 - 4.10.3 空间数据的统计建模

4.1 点估计与区间估计

点估计 (point estimation) 与区间估计 (interval estimation) 是统计推断 (statistical inference) 中的两大基本任务。在实际应用中, 我们既可能需要用一个具体的值来表达对未知参数的最佳猜测 (点估计), 也可能希望通过一个区间来表达估计的不确定性范围 (区间估计)。

4.1.1 点估计的定义

设总体中感兴趣的参数为 θ (可能是一维或多维向量), 我们从总体中抽取样本:

$$\mathbf{x} = (x_1, x_2, \dots, x_n)$$

样本是总体的一部分, 且视为来自某个概率分布 $f(x | \theta)$ 。

点估计: 用样本的某个函数 $\hat{\theta} = \hat{\theta}(\mathbf{x})$ 作为参数 θ 的“最佳”估值。

例如均值的点估计:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

这是样本均值, 是总体均值 μ 的一个点估计。

4.1.2 区间估计的定义

在点估计基础上, 区间估计考虑抽样波动, 以某种置信度 $1 - \alpha$ 给出一个区间:

$$P(L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})) = 1 - \alpha$$

式中 $L(\mathbf{x})$, $U(\mathbf{x})$ 分别为区间的下限和上限。

4.1.3 构造置信区间的常用方法

构造置信区间 (Confidence Interval, CI) 的基本思路：

1. 寻找枢轴量

设样本数据为 $X = (X_1, X_2, \dots, X_n)$ ，未知参数为 θ 。

我们希望找到一个统计量：

$$R(X, \theta)$$

其中：

- X 表示样本数据；
- θ 表示待估计的未知参数；
- $R(X, \theta)$ 称为**枢轴量** (pivot quantity)，它由样本和参数共同组成，但其概率分布已知且不依赖于 θ 。

2. 利用分布性质建立概率不等式

根据枢轴量的分布特性，在给定的置信水平 $1 - \alpha$ 下，写出：

$$P(a \leq R(X, \theta) \leq b) = 1 - \alpha$$

其中 a 、 b 是由枢轴量的已知分布确定的常数。

3. 反解参数的区间

将上述不等式视为关于 θ 的方程（或不等式），反解出 θ 的取值范围：

$$L(X) \leq \theta \leq U(X)$$

这里 $L(X)$ 和 $U(X)$ 是用样本数据计算出的区间端点。

4. 整理为标准形式

最终得到置信区间：

$$[L(X), U(X)]$$

该区间在概率意义上以 $1 - \alpha$ 的置信水平覆盖真实参数值。

例题4.1 基于正态近似 (σ^2 已知) 构造均值的置信区间

设 $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ ，且方差 σ^2 **已知**，要求在显著性水平 α 下构造总体均值 μ 的 $1 - \alpha$ 置信区间。

解答

1. 样本均值的分布

由于每个样本 $X_i \sim N(\mu, \sigma^2)$ ，由正态分布的可加性可知：

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

2. 选择枢轴量

定义标准化统计量：

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

根据前一步的分布性质， Z 服从标准正态：

$$Z \sim N(0, 1),$$

且其分布与未知参数 μ 无关，因此它是一个枢轴量。

3. 利用分布性质建立概率不等式

在置信水平 $1 - \alpha$ 下，标准正态分布满足：

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha,$$

其中 $z_{\alpha/2}$ 是标准正态分布上尾概率为 $\alpha/2$ 的分位点。

4. 将枢轴量替换回样本统计量并反解 μ

将 Z 的定义代入：

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

两边同时乘以 $\frac{\sigma}{\sqrt{n}}$ 并对不等式进行移项，得到：

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

结论

总体均值 μ 的 $1 - \alpha$ 置信区间为：

$$\boxed{\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

即：

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

说明

- \bar{X} ：样本均值

- $z_{\alpha/2}$: 标准正态分布上尾概率为 $\alpha/2$ 的分位点
- σ : 已知总体标准差
- n : 样本量

例题4.2 基于 t 分布 (σ^2 未知) 构造均值的置信区间

设 $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, 且总体方差 σ^2 **未知**。我们希望在显著性水平 α 下, 构造总体均值 μ 的 $1 - \alpha$ 置信区间。

解答

1. 样本均值的分布

在正态总体下:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

2. 样本方差的定义与性质

定义样本方差:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

它是 σ^2 的无偏估计量。

3. 选择枢轴量

当总体服从正态分布且方差未知时, 统计量:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

服从自由度为 $n - 1$ 的 Student t 分布:

$$T \sim t_{n-1}.$$

该分布与未知参数 μ 和 σ^2 无关, 因此 T 是枢轴量。

4. 利用 t 分布的性质建立概率不等式

在置信水平 $1 - \alpha$ 下:

$$P(-t_{\alpha/2, n-1} \leq T \leq t_{\alpha/2, n-1}) = 1 - \alpha,$$

其中 $t_{\alpha/2, n-1}$ 是自由度 $n - 1$ 的 t 分布分位点 (上尾概率为 $\alpha/2$)。

5. 代入枢轴量表达式并反解 μ

将 T 的定义代入:

$$P\left(-t_{\alpha/2, n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2, n-1}\right) = 1 - \alpha.$$

两边同时乘以 $\frac{S}{\sqrt{n}}$ 并移项，得到：

$$P\left(\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

结论

总体均值 μ 的 $1 - \alpha$ 置信区间为：

$$\boxed{\bar{X} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}}$$

即：

$$\left[\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}, \quad \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \right].$$

符号说明

- \bar{X} ：样本均值
- S ：样本标准差
- $t_{\alpha/2, n-1}$ ：自由度为 $n - 1$ 的 t 分布分位点（上尾概率 $\alpha/2$ ）
- n ：样本量

4.2 参数估计与状态估计的关系

参数估计（Parameter Estimation）与状态估计（State Estimation）是统计推断与信号处理中的两类核心问题。两者在数学基础、推断方法上有共通性，但在对象性质和时间维度上又有明显区别。

4.2.1 参数估计

定义

参数估计旨在利用观测数据推断模型或系统中的未知参数 θ 。

- **对象性质**
 - 频率派（Frequentist）视角： θ 被视为**固定未知的常数**，数据的随机性来源于采样。
 - 贝叶斯（Bayesian）视角： θ 被视为**随机变量**，服从某种先验分布 $p(\theta)$ ，不确定性通过概率刻画。
- **时间维度**

参数在整个推断过程中往往**不随时间 k 变化（静态）**。
- **典型样例**
 - 线性回归中的回归系数 β
 - 机器学习模型的权重
 - 系统辨识中的增益系数、阻尼系数等
- **常用方法**
 - 极大似然估计（MLE）
 - 矩估计（Method of Moments）
 - 贝叶斯估计（MAP、MMSE 等）

4.2.2 状态估计

定义

状态估计关注的是随时间演化的系统状态 \mathbf{x}_k 的推断问题。

- **对象性质**

\mathbf{x}_k ：表示系统在时刻 k 的状态向量，例如：位置、速度、温度场分布、库存水平等。

- **时间演化模型**

状态随时间变化，通常由状态空间模型（State-Space Model）刻画：

$$\mathbf{x}_k = F_{k-1}\mathbf{x}_{k-1} + \mathbf{w}_{k-1} \quad (\text{状态转移方程})$$

$$\mathbf{z}_k = H_k\mathbf{x}_k + \mathbf{v}_k \quad (\text{观测方程})$$

其中：

- F_{k-1} ：状态转移矩阵
- H_k ：观测矩阵
- \mathbf{w}_{k-1} 、 \mathbf{v}_k ：过程噪声和观测噪声

- **典型估计算法**

- **卡尔曼滤波 (KF)**：适用于线性高斯系统
- **扩展卡尔曼滤波 (EKF)**：一阶线性化用于非线性系统
- **无迹卡尔曼滤波 (UKF)**：使用确定性采样点捕捉非线性效应
- **粒子滤波 (PF)**：蒙特卡罗方法，对任意非线性非高斯模型提供近似推断

4.2.3 联系与相互转化

- **参数估计是状态估计的特例**

如果我们将参数视为“固定不变的状态”，在状态空间模型中设置：

$$F_{k-1} = I \quad (\text{单位矩阵}), \quad \mathbf{w}_{k-1} = \mathbf{0}$$

那么状态估计就退化为对静态参数的估计。

- **状态估计等价于动态参数估计**

当模型参数随时间缓慢变化（例如自适应控制系统中的增益系数），可以将其建模为随时间演化的“状态”，并运用动态滤波方法进行估计——典型如 **自适应卡尔曼滤波**。

- **贝叶斯统一视角**

两者在贝叶斯推断框架下表现为同一种形式的推断问题：

$$\text{参数估计: } p(\theta \mid \mathbf{data})$$

$$\text{状态估计: } p(\mathbf{x}_k \mid \mathbf{z}_{1:k})$$

实质上都是基于观测信息更新未知量的后验分布，只不过**参数是静态的，状态是动态的**。

4.3 频率派与贝叶斯派方法

4.3.1 历史与哲学基础

频率派与贝叶斯派的分化，源于对概率概念的根本理解不同。

频率派思想兴起于 20 世纪初，由 Fisher、Neyman 和 Pearson 等统计学家奠定基础。他们将概率解释为**长期频率**——在无限次重复的同类实验中，某事件发生的比例就是它的概率。在这一视角下，未知参数 θ 被视作固定但未知的常数；不确定性只来自观测数据的抽样过程和随机噪声。频率派研究的焦点，是设计估计和检验方法，使它们在长期重复实验中具备理想的统计性质，例如置信区间的覆盖率、假设检验的显著性水平。

贝叶斯派的起源更早，可追溯到 18 世纪的 Thomas Bayes 和 Laplace。贝叶斯思想认为概率是**信念程度** (degree of belief)，既可以用来描述随机事件的不确定性，也可以刻画未知量的不确定性。贝叶斯派不将参数视为固定，而是当作具有先验分布的随机变量。在获取数据之后，研究者通过贝叶斯定理将先验分布更新为后验分布，使之体现数据与先验知识的综合影响。这一哲学基础使得贝叶斯方法天然适合融合外部信息，并以概率的形式直接表达推断的不确定性。

4.3.2 方法特点

在频率派方法中，推断过程依赖于构造统计量及其抽样分布。例如，通过样本均值、样本方差等估计未知参数，然后利用这些估计量的分布来形成置信区间或进行假设检验。频率派强调在反复抽样的长期频率意义下，推断方法满足预定的覆盖率或误拒率等性质。这类方法的优势在于理论成熟、数学严谨，可以保证诸如无偏性、一致性以及最小方差等性能指标。然而，它并不直接给出“在已有数据条件下，参数是某个具体值的概率”，且在缺乏灵活途径引入先验信息时，可能失去充分利用已有知识的机会。

贝叶斯派方法的核心过程可以概括为“先验—似然—后验”：首先建立参数的先验分布 $p(\theta)$ ，然后利用模型给定数据的似然函数 $p(\mathbf{x}|\theta)$ ，通过贝叶斯公式得到后验分布 $p(\theta|\mathbf{x})$ 。后验分布既可以用于点估计（如最大后验概率 MAP、后验均值），也可以直接构造后验区间 (credible interval)，从而在已知数据条件下，用概率刻画参数不确定性的范围。贝叶斯方法的优势在于能自然融合外部知识，并提供完整的概率描述；但它的计算通常比频率派更复杂，尤其在高维非线性模型中，需要借助数值近似方法如 MCMC 或变分推断。此外，先验分布的选择带有一定的主观性，可能影响推断结果。

4.3.3 对比与联系

虽然频率派和贝叶斯派在哲学根基与技术路径上存在差异，但在很多实际情境下两者的结论会趋于一致。例如，当采用非信息性或平坦先验时，贝叶斯估计的结果在形式上常与频率派相同；而一些频率派的结论，也能在贝叶斯框架内得到自然的解释。这说明，两派并非截然对立，而是可以在不同的推断场景下互为补充。

二者的主要区别在于对不确定性的处理：频率派将不确定性完全归因于样本的随机抽样，置信区间的意义是长期覆盖率；而贝叶斯派直接对参数本身建立概率分布，后验区间的意义是给定数据条件下的概率陈述。在现代统计学和机器学习实践中，研究者常常根据问题性质灵活地选择方法，甚至融合两派的思想——用频率派的理论来确保方法的稳定性，用贝叶斯框架来引入先验信息、刻画全概率的不确定性。这种互补性的结合，使得推断方法既能在理论上稳健，又能在实际中适应复杂的信息环境。

4.3.4 总体均值的两派估计

为了更直观地理解频率派与贝叶斯派在方法思想上的差异，我们以一个简单的总体均值估计问题为例进行说明。假设我们有一组数据：

$$X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

其中方差 σ^2 已知，而均值 μ 是我们希望通过数据进行推断的未知量。

频率派方法

在频率派的框架下，参数 μ 被视作一个固定但未知的常数，不存在概率分布。由于在高斯模型中，样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 是 μ 的充分统计量，也是无偏估计量，因此频率派会直接用样本均值作为点估计：

$$\hat{\mu} = \bar{X}$$

为了刻画估计的不确定性，频率派利用样本均值的抽样分布：

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

在置信水平 $1 - \alpha$ 下，置信区间构造为：

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

这里， $z_{\alpha/2}$ 是标准正态分布的分位数。该区间在**长期重复抽样**的意义下，覆盖真实均值 μ 的比例约为 $1 - \alpha$ 。

贝叶斯派方法

在贝叶斯框架中，我们将 μ 当作一个具有先验分布的随机变量。假设先验为正态分布：

$$\mu \sim N(\mu_0, \tau^2)$$

其中 μ_0 表示我们在观察数据前对 μ 的中心估计， τ^2 刻画了先验的不确定性强度。样本数据的似然函数由模型给出：

$$p(\mathbf{x}|\mu) \propto \exp\left[-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right]$$

因为先验和似然都是正态分布，后验分布依然是正态，其形式为：

$$\mu | \mathbf{x} \sim N\left(\frac{\frac{\mu_0}{\tau^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)$$

可以看到，后验均值是先验均值 μ_0 与样本均值 \bar{x} 的加权平均，权重由先验方差和样本方差共同决定；后验方差则反

映了信息融合后的不确定性。

在贝叶斯框架下，给定数据后，点估计可直接取后验均值；区间估计则可以构造**后验区间**（credible interval），例如在正态后验下，取后验均值 $\pm z_{\alpha/2} \times \sqrt{\text{后验方差}}$ ，它表示在给定数据的条件下，参数落在该区间的概率为 $1 - \alpha$ 。

对结果的比较与解释

在频率派方法里，置信区间的概率解释是长时间重复抽样的覆盖率；在贝叶斯方法里，后验区间的概率解释是给定数据时参数落在该区间的优势概率。这种概率概念的差异，直接体现了两派对“不确定性”的不同定义。如果在贝叶斯方法中使用非常宽松且均匀的先验（即非信息性先验），那么在这个正态均值问题上，贝叶斯后验区间与频率派的置信区间会高度一致。这也是二者联系的重要体现。

4.4 估计量的基本性质

4.4.1 无偏性、一致性、有效性

1) 无偏性 (Unbiasedness)

在统计推断中，一个估计量 $\hat{\theta}$ 如果其数学期望等于真实参数 θ ，即：

$$\mathbb{E}[\hat{\theta}] = \theta$$

我们就称它是**无偏的**。无偏性意味着，在长期重复抽样的意义下，该估计量不会系统性地高估或低估参数的真实值。换句话说，如果我们无限次重复实验并计算每次的估计结果，那么这些结果的平均值会恰好等于真实参数。无偏性是频率派统计中的重要性质之一，因为它保证了估计过程在平均意义上是“公平”的，不会存在方向性的偏差。然而，值得注意的是，无偏性并不保证每次估计都接近真实值，它只是保证了在长期平均下没有系统误差。

2) 一致性 (Consistency)

一致性描述的是估计量在样本量不断增加时的表现。如果一个估计量 $\hat{\theta}_n$ 随着样本容量 n 增大而以概率收敛到真实参数 θ ，即：

$$\forall \epsilon > 0, \quad \lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

那么我们称它是**一致的**。一致性保证了在样本量足够大时，估计量会越来越接近真实值，这也是统计推断可靠性的一个核心要求。在实际应用中，即使一个估计量存在一定偏差，只要它是一致的，我们仍然可以通过收集更多数据来缩小估计误差，从而逼近真实参数。因此，一致性体现了“数据量足够大时，估计会变好”的基本直觉。

3) 有效性 (Efficiency)

有效性关注的是估计量的波动性。在所有无偏估计量中，如果某个估计量的方差最小，我们称它为**有效估计量**。这种估计量也被称为**最小方差无偏估计量**（Minimum Variance Unbiased Estimator, MVUE）。方差衡量了估计值在不同样

本下的离散程度，方差越小，估计结果越稳定，推断的精度也越高。有效性不仅要求估计量无偏，还要求它在无偏估计量的集合中尽可能减少随机波动。在频率派理论中，寻找 MVUE 是一个重要目标，因为它在保证平均意义上准确的同时，也最大限度地提高了估计的稳定性。

4) 例题

例题4.3 正态总体均值的估计

已知随机样本

$$X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

其中方差 σ^2 已知。设样本均值为

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

证明 \bar{X} 是 μ 的无偏估计量且一致估计量。

解答：

(1) 无偏性

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \cdot n\mu = \mu$$

因此， \bar{X} 的期望等于总体均值 μ ，它是 μ 的**无偏估计量**。

(2) 一致性

由方差性质：

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

当 $n \rightarrow \infty$ 时， $\text{Var}(\bar{X}) \rightarrow 0$ 。结合**大数定律**可知：

$$\bar{X} \xrightarrow{p} \mu$$

因此， \bar{X} 是 μ 的**一致估计量**。

例题4.4 正态总体方差的估计

已知随机样本

$$X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

其中 μ 未知。设样本方差为

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

证明 S^2 是 σ^2 的无偏估计量且一致估计量。

解答：

(1) 无偏性

在正态分布下，样本方差的期望为：

$$\mathbb{E}[S^2] = \sigma^2$$

因此， S^2 的期望等于总体方差 σ^2 ，它是 σ^2 的**无偏估计量**。

(2) 一致性

已知样本方差的方差为：

$$\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

当 $n \rightarrow \infty$ 时， $\text{Var}(S^2) \rightarrow 0$ ，且 $\mathbb{E}[S^2] = \sigma^2$ 。

由切比雪夫不等式或弱大数定律可得：

$$S^2 \xrightarrow{p} \sigma^2$$

因此， S^2 是 σ^2 的**一致估计量**。

4.4.2 均值方差均衡

在实际统计推断或参数估计中，**均值方差均衡**（mean-variance trade-off）是一个重要的思想。它指的是在选择估计量时，除了关注估计量的无偏性（均值等于真实值），还要关注估计量的方差（估计的稳定性）。有些估计量虽然无偏，但方差较大，导致估计结果波动较大；而有些估计量可能略有偏差，但方差较小，整体均方误差（MSE）更低。

1) 均方误差（Mean Squared Error, MSE）

均方误差综合考虑了估计量的偏差和方差，定义如下：

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta)^2$$

- **无偏估计量**：MSE = 方差
- **有偏估计量**：MSE = 方差 + 偏差平方

意义

- 在实际应用中，过分追求无偏性可能导致方差很大，反而使估计不可靠。
- 适当引入偏差（如贝叶斯估计、岭回归等），可以显著降低方差，使估计总体更优。
- 因此，实际选择估计量时，通常以均方误差（MSE）最小为目标，而不是只追求无偏性。

2) 例题

例题4.5：均值方差均衡的应用

假设有两个估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ ：

- $\hat{\theta}_1$ 无偏，方差为 4
- $\hat{\theta}_2$ 有偏，偏差为 1，方差为 1

问：哪个估计量更优？

解答：

计算 MSE：

- $\text{MSE}(\hat{\theta}_1) = 4 + 0^2 = 4$
- $\text{MSE}(\hat{\theta}_2) = 1 + 1^2 = 2$

虽然 $\hat{\theta}_2$ 有偏，但 MSE 更小，因此在均值方差均衡的原则下， $\hat{\theta}_2$ 更优。

4.4.4 Cramér–Rao 下界 (CRLB)

Cramér–Rao 下界是统计估计理论中的一个重要结果，它给出了无偏估计量方差的理论下限，即在一定条件下，任何无偏估计量的方差都不可能小于该下界。因此，CRLB 是衡量无偏估计量优劣的重要基准。

1) 单确定参数估计量的Cramér–Rao下界

设 $\hat{\theta}$ 是确定参数 θ 的无偏估计量，满足

$$\mathbb{E}[\hat{\theta} - \theta] = \int (\hat{\theta} - \theta)p(x|\theta)dx = 0$$

对 θ 求导，得到

$$-\int p(x|\theta)dx + \int (\hat{\theta} - \theta) \left[\frac{\partial}{\partial \theta} \ln p(x|\theta) \right] p(x|\theta)dx = 0$$

化简为

$$\int (\hat{\theta} - \theta) \frac{\partial}{\partial \theta} \ln p(x|\theta) p(x|\theta)dx = 1$$

根据 Cauchy–Schwarz 不等式

$$\left[\int g(y)h(y)dy \right]^2 \leq \int g^2(y)dy \int h^2(y)dy$$

其中，等号成立当且仅当 $h(y) = kg(y)$ 。

取 $g(x) = \hat{\theta} - \theta$, $h(x) = \frac{\partial}{\partial \theta} \ln p(x|\theta)$, 得

$$1 \leq \int (\hat{\theta} - \theta)^2 p(x|\theta) dx \int \left[\frac{\partial}{\partial \theta} \ln p(x|\theta) \right]^2 p(x|\theta) dx$$

即

$$\text{Var}[\hat{\theta}] \geq \frac{1}{I(\theta)}$$

其中

$$I(\theta) = \int \left[\frac{\partial}{\partial \theta} \ln p(x|\theta) \right]^2 p(x|\theta) dx$$

称为 Fisher 信息量, 它刻画了样本中包含的参数 θ 的信息量。

解释:

- Fisher 信息量越大, 意味着样本对参数的“辨识能力”越强, 理论上可达到的估计方差下界就越小;
- CRLB 给出的下界是无偏估计量的性能极限, 任何无偏估计量的方差都不能突破该极限。

拓展:

(1) 二阶导数形式的 CRLB

在满足正则条件的情况下, Fisher 信息量除了常用的一阶导平方期望形式外, 还可以用**二阶导数**的形式表示:

$$I(\theta) = -\mathbb{E}_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \ln p(X; \theta) \right]$$

因此, 单参数的 Cramér–Rao 下界可以写为:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)} = \frac{1}{-\mathbb{E}_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \ln p(X; \theta) \right]}$$

推导:

由归一化条件 $\int p(x|\theta) dx = 1$, 可得:

$$\mathbb{E}_{\theta} \left[\frac{\partial}{\partial \theta} \ln p(X; \theta) \right] = 0$$

再次求导:

$$\int \left[\frac{\partial^2}{\partial \theta^2} \ln p(X; \theta) \right] p(X; \theta) dx + \int \left[\frac{\partial}{\partial \theta} \ln p(X; \theta) \right]^2 p(X; \theta) dx = 0$$

于是:

$$I(\theta) = -\mathbb{E}_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \ln p(X; \theta) \right]$$

这就是 Fisher 信息量的二阶导数形式, 代入 CRLB 即得上述表达式。

(2) 参数函数的无偏估计量的 CRLB

如果我们关心的不是参数 θ 本身，而是它的某个可微函数 $\tau(\theta)$ ，且估计量 $\hat{\tau}$ 是无偏的，即：

$$\mathbb{E}_{\theta}[\hat{\tau}] = \tau(\theta)$$

那么 CRLB 推广为：

$$\text{Var}(\hat{\tau}) \geq \frac{[\tau'(\theta)]^2}{I(\theta)}$$

其中 $\tau'(\theta) = \frac{d\tau}{d\theta}$ 。

推导：

无偏性条件：

$$\mathbb{E}_{\theta}[\hat{\tau}] = \tau(\theta)$$

两边对 θ 求导（正则条件保证可交换微分与积分）：

$$\mathbb{E}_{\theta} \left[\hat{\tau} \cdot \frac{\partial}{\partial \theta} \ln p(X; \theta) \right] = \tau'(\theta)$$

由于 $\mathbb{E}_{\theta} \left[\frac{\partial}{\partial \theta} \ln p(X; \theta) \right] = 0$ ，上式可改写为：

$$\mathbb{E}_{\theta} \left[(\hat{\tau} - \tau(\theta)) \cdot \frac{\partial}{\partial \theta} \ln p(X; \theta) \right] = \tau'(\theta)$$

应用 Cauchy-Schwarz 不等式：

$$[\tau'(\theta)]^2 \leq \text{Var}(\hat{\tau}) \cdot \text{Var}\left(\frac{\partial}{\partial \theta} \ln p(X; \theta)\right)$$

注意 $\text{Var}\left(\frac{\partial}{\partial \theta} \ln p(X; \theta)\right) = I(\theta)$ ，于是：

$$\text{Var}(\hat{\tau}) \geq \frac{[\tau'(\theta)]^2}{I(\theta)}$$

这就是参数函数的无偏估计量的 CRLB。

2) 单随机参数估计量的Cramér-Rao下界

设 $\hat{\theta}$ 是随机参数 θ 的无偏估计量，满足

$$\mathbb{E}[\hat{\theta} - \theta] = \int (\hat{\theta} - \theta) p(x, \theta) dx d\theta = 0$$

其中 $p(x, \theta)$ 是观测 x 和参数 θ 的联合概率密度。

对 θ 求导，假定相关函数可积，有

$$-\int p(x, \theta) dx + \int (\hat{\theta} - \theta) \left[\frac{\partial}{\partial \theta} \ln p(x, \theta) \right] p(x, \theta) dx = 0$$

化简为

$$\int (\hat{\theta} - \theta) p^{1/2}(x, \theta) \frac{\partial}{\partial \theta} \ln p(x, \theta) p^{1/2}(x, \theta) dx = 1$$

同样应用 Cauchy-Schwarz 不等式, 得

$$1 \leq \int (\hat{\theta} - \theta)^2 p(x, \theta) dx \int \left[\frac{\partial}{\partial \theta} \ln p(x, \theta) \right]^2 p(x, \theta) dx$$

即

$$\text{Var}[\hat{\theta}] \geq \frac{1}{\mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \ln p(x, \theta) \right)^2 \right]}$$

当 $\frac{\partial}{\partial \theta} \ln p(x, \theta) = k(\theta)[\hat{\theta} - \theta]$ 时, 等号成立。

3) 例题

例题4.6: 正态分布均值的CRLB (确定参数情形)

设 $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$, 其中: *i.i.d.*表示独立同分布, n 为样本量, μ 未知均值, σ^2 已知方差, 求 μ 的无偏估计量的方差下界。

解答:

对数似然函数:

$$\ln L(\mu) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

一阶导数:

$$\frac{\partial}{\partial \mu} \ln L(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$$

Fisher信息:

$$I(\mu) = \mathbb{E} \left[\left(\frac{\partial}{\partial \mu} \ln L(\mu) \right)^2 \right] = \frac{n}{\sigma^2}$$

CRLB:

$$\text{Var}(\hat{\mu}) \geq \frac{1}{I(\mu)} = \frac{\sigma^2}{n}$$

实际上, 样本均值 \bar{X} 的方差等于CRLB, 说明 \bar{X} 是有效估计量。

例题4.7: 单随机参数估计量的 Cramér–Rao 下界

设随机参数 θ 服从均值为 μ_0 、方差为 σ_θ^2 的正态分布, 即

$$\theta \sim N(\mu_0, \sigma_\theta^2)$$

在给定 θ 的条件下, 观测值 X 服从

$$X | \theta \sim N(\theta, \sigma^2)$$

其中 σ^2 已知。试求对 θ 的无偏估计量的 **Cramér–Rao 下界** (CRLB)。

解答:

建立联合概率密度

由于 θ 是随机参数, 观测 X 与 θ 的联合概率密度为:

$$p(x, \theta) = p(x | \theta) p(\theta)$$

其中:

$$p(x | \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x - \theta)^2}{2\sigma^2} \right]$$

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp \left[-\frac{(\theta - \mu_0)^2}{2\sigma_\theta^2} \right]$$

计算联合对数密度的导数

联合密度的对数为:

$$\ln p(x, \theta) = -\frac{(x - \theta)^2}{2\sigma^2} - \frac{(\theta - \mu_0)^2}{2\sigma_\theta^2} - \ln(2\pi\sigma\sigma_\theta)$$

对 θ 求偏导:

$$\frac{\partial}{\partial \theta} \ln p(x, \theta) = \frac{x - \theta}{\sigma^2} - \frac{\theta - \mu_0}{\sigma_\theta^2}$$

计算 Fisher 信息量

Fisher 信息量定义为:

$$I = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \ln p(x, \theta) \right)^2 \right]$$

期望取联合分布 $p(x, \theta)$ 下的平均。

由于 X 和 θ 联合正态，且在条件分布中 $\mathbb{E}[x - \theta | \theta] = 0$ ，方差为 σ^2 ，可得：

$$\mathbb{E} \left[\left(\frac{x - \theta}{\sigma^2} \right)^2 \right] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

$$\mathbb{E} \left[\left(\frac{\theta - \mu_0}{\sigma_\theta^2} \right)^2 \right] = \frac{\sigma_\theta^2}{\sigma_\theta^4} = \frac{1}{\sigma_\theta^2}$$

交叉项期望为零，因为条件期望下它们不相关。

因此：

$$I = \frac{1}{\sigma^2} + \frac{1}{\sigma_\theta^2}$$

Cramér–Rao 下界

单随机参数的 CRLB 为：

$$\text{Var}[\hat{\theta}] \geq \frac{1}{I} = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_\theta^2}}$$

结果解释

- 该下界由两部分信息量组成：
 - $\frac{1}{\sigma^2}$ ：来自观测噪声模型 $p(x | \theta)$ 的信息；
 - $\frac{1}{\sigma_\theta^2}$ ：来自参数先验分布 $p(\theta)$ 的信息。
- 当 $\sigma_\theta^2 \rightarrow \infty$ （即参数无先验约束）时，下界退化为经典的固定参数 CRLB：

$$\text{Var}[\hat{\theta}] \geq \sigma^2$$

- 当 $\sigma^2 \rightarrow 0$ （观测无噪声）时，下界由先验方差决定：

$$\text{Var}[\hat{\theta}] \geq \sigma_\theta^2$$

4.4.5 多参数估计的Cramér–Rao 下界

1) 定义

假设参数为 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_M)$ ，观测为 \mathbf{X} ，联合概率密度为 $p(\mathbf{X} | \boldsymbol{\theta})$ 。

若存在无偏估计量 $\hat{\theta}_k(\mathbf{X})$ ，使得

$$\mathbb{E}[\hat{\theta}_k(\mathbf{X})] = \theta_k, \quad k = 1, 2, \dots, M$$

则有：

$$\text{Cov}(\hat{\boldsymbol{\theta}}) \succeq J^{-1}(\boldsymbol{\theta})$$

其中 \succeq 表示矩阵半正定关系。

(1) 得分函数 (Score Function)

定义：

$$u_k(\mathbf{X}, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta_k} \ln p(\mathbf{X} | \boldsymbol{\theta}), \quad k = 1, \dots, M$$

向量形式：

$$\mathbf{u}(\mathbf{X}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{X} | \boldsymbol{\theta})$$

(3) Fisher信息矩阵

定义Fisher信息矩阵的元素为：

$$J_{kl} = \mathbb{E} \left[\frac{\partial}{\partial \theta_k} \ln p(\mathbf{X} | \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta_l} \ln p(\mathbf{X} | \boldsymbol{\theta}) \right] = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_l} \ln p(\mathbf{X} | \boldsymbol{\theta}) \right]$$

其中 $k, l = 1, 2, \dots, M$ 。

(d) 多参数 Cramér–Rao 下界

若存在无偏估计量 $\hat{\boldsymbol{\theta}}(\mathbf{X})$ ，满足：

$$\mathbb{E}[\hat{\theta}_k] = \theta_k$$

则有：

$$\text{Cov}(\hat{\boldsymbol{\theta}}) \succeq J^{-1}(\boldsymbol{\theta})$$

其中 \succeq 表示矩阵半正定关系。

2) 性质

(1) 得分函数均值为零

由密度归一化条件：

$$\int p(\mathbf{X} | \boldsymbol{\theta}) d\mathbf{X} = 1$$

对 θ_k 求导：

$$\int \frac{\partial}{\partial \theta_k} p(\mathbf{X} | \boldsymbol{\theta}) d\mathbf{X} = 0$$

利用：

$$\frac{\partial}{\partial \theta_k} p = p \cdot \frac{\partial}{\partial \theta_k} \ln p$$

得到：

$$\mathbb{E}[u_k(\mathbf{X}, \boldsymbol{\theta})] = 0$$

即得分函数是均值为零的随机变量。

(2) 估计误差与得分函数之间的协方差是单位矩阵

假设 $\hat{\boldsymbol{\theta}}(\mathbf{X})$ 是无偏估计量，即

$$\mathbb{E}[\hat{\theta}_k] = \theta_k, \quad k = 1, \dots, M.$$

对参数分量 θ_l 求偏导，得到

$$\frac{\partial}{\partial \theta_l} \mathbb{E}[\hat{\theta}_k] = \delta_{kl},$$

其中 δ_{kl} 为 Kronecker 符号：当 $k = l$ 时取值 1，否则为 0。

在正则条件下，可以将求导运算移入积分号，并利用得分函数的定义

$$u_l(\mathbf{X}, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta_l} \ln p(\mathbf{X} | \boldsymbol{\theta}),$$

于是有

$$\mathbb{E}[\hat{\theta}_k u_l(\mathbf{X}, \boldsymbol{\theta})] = \delta_{kl}.$$

由于 $\mathbb{E}[u_l] = 0$ ，减去 $\theta_k \mathbb{E}[u_l]$ 得到

$$\mathbb{E}[(\hat{\theta}_k - \theta_k) u_l(\mathbf{X}, \boldsymbol{\theta})] = \delta_{kl}.$$

这表明：估计误差向量 $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ 与 得分函数向量 $\mathbf{u}(\mathbf{X}, \boldsymbol{\theta})$ 的协方差矩阵正好是单位矩阵 I_M 。

(3) Fisher信息矩阵元素定义的等价性

已知：

$$\frac{\partial^2}{\partial \theta_k \partial \theta_l} \ln p = \frac{1}{p} \frac{\partial^2 p}{\partial \theta_k \partial \theta_l} - \frac{1}{p^2} \frac{\partial p}{\partial \theta_k} \frac{\partial p}{\partial \theta_l}$$

对上式取负号并取期望：

$$-\mathbb{E} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_l} \ln p \right] = - \int \left[\frac{1}{p} \frac{\partial^2 p}{\partial \theta_k \partial \theta_l} - \frac{1}{p^2} \frac{\partial p}{\partial \theta_k} \frac{\partial p}{\partial \theta_l} \right] p d\mathbf{X}$$

简化：

$$= - \int \frac{\partial^2 p}{\partial \theta_k \partial \theta_l} d\mathbf{X} + \int \frac{1}{p} \frac{\partial p}{\partial \theta_k} \frac{\partial p}{\partial \theta_l} d\mathbf{X}$$

在正则条件下，可以将求导移出积分号，并且假设积分区域不依赖于参数，那么：

$$\int \frac{\partial^2 p}{\partial \theta_k \partial \theta_l} d\mathbf{X} = \frac{\partial^2}{\partial \theta_k \partial \theta_l} \int p d\mathbf{X} = \frac{\partial^2}{\partial \theta_k \partial \theta_l} 1 = 0$$

因此第一项为零。

化简第二项：

$$-\mathbb{E} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_l} \ln p \right] = \int \left[\frac{1}{p} \frac{\partial p}{\partial \theta_k} \frac{1}{p} \frac{\partial p}{\partial \theta_l} \right] p d\mathbf{X} = \mathbb{E}[u_k u_l]$$

证得 Fisher 信息矩阵的两种常见定义形式在正则条件下是等价的：

$$J_{kl}(\boldsymbol{\theta}) = \mathbb{E}[u_k u_l] \quad \text{与} \quad J_{kl}(\boldsymbol{\theta}) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_l} \ln p(\mathbf{X} | \boldsymbol{\theta}) \right]$$

其中 $u_k = \frac{\partial}{\partial \theta_k} \ln p(\mathbf{X} | \boldsymbol{\theta})$ 。

(4) Schur 补的正定性

设一个实对称分块矩阵：

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

其中：

- A 是 $p \times p$ 矩阵
- C 是 $q \times q$ 矩阵
- B 是 $p \times q$ 矩阵

如果 C 可逆，则 M 关于 C 的 Schur 补 定义为：

$$S = A - BC^{-1}B^\top$$

如果 M 是对称正定矩阵 ($M \succ 0$)，且 C 可逆，则 Schur 补 S 也是正定矩阵 ($S \succ 0$)。

换句话说：

$$M \succeq 0 \quad \text{且} \quad C \succ 0 \quad \Rightarrow \quad A - BC^{-1}B^\top \succeq 0$$

这个性质在推导矩阵不等式时非常有用。

(5) 多参数 Cramér–Rao 下界

定义 $M+1$ 维随机向量：

$$\mathbf{Z} = \begin{pmatrix} \hat{\theta}_1(\mathbf{X}) - \theta_1 \\ \frac{\partial}{\partial \theta_1} \ln p(\mathbf{X}|\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_2} \ln p(\mathbf{X}|\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_M} \ln p(\mathbf{X}|\boldsymbol{\theta}) \end{pmatrix}$$

由前述性质可知：

$$\mathbb{E}[\mathbf{Z}] = \mathbf{0}$$

计算协方差矩阵：

$$\mathbb{E}[\mathbf{Z}\mathbf{Z}^T] = \begin{pmatrix} \sigma_1^2 & 1 & 0 & \cdots & 0 \\ 1 & J_{11} & J_{12} & \cdots & J_{1M} \\ 0 & J_{12} & J_{22} & \cdots & J_{2M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & J_{1M} & J_{2M} & \cdots & J_{MM} \end{pmatrix}$$

其中：

$$\sigma_1^2 = \text{Var}[\hat{\theta}_1]$$

由于协方差矩阵是半正定的，任意主子式的行列式非负。取左上角的 2×2 块的 Schur 补，可得：

$$\sigma_1^2 - (1, 0, \dots, 0) J^{-1} (1, 0, \dots, 0)^\top \geq 0$$

化简得到：

$$\sigma_1^2 \geq [J^{-1}]_{11}$$

即：

$$\text{Var}[\hat{\theta}_1] \geq [J^{-1}]_{11}$$

同理，对所有参数分量均成立：

构造联合向量：

$$\mathbf{Z} = \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \vdots \\ \hat{\theta}_M - \theta_M \\ u_1 \\ \vdots \\ u_M \end{pmatrix}$$

它的期望为零，协方差矩阵为：

$$\text{Cov}(\mathbf{Z}) = \begin{pmatrix} \Sigma & I_M \\ I_M & J \end{pmatrix}$$

其中：

- $\Sigma = \text{Cov}(\hat{\boldsymbol{\theta}})$
- I_M 是 $M \times M$ 单位矩阵
- J 是 Fisher 信息矩阵

取关于 J 的 Schur 补：

$$\Sigma - I_M J^{-1} I_M \succeq 0$$

即：

$$\text{Cov}(\hat{\theta}) \succeq J^{-1}(\theta)$$

这就是**多参数 Cramér–Rao 下界**的矩阵形式。

3) 例题

例题4.8： 二维正态分布均值参数的 Cramér–Rao 下界

设 X_1, X_2, \dots, X_n 是独立同分布的二维正态随机向量：

$$X_i \sim N(\mu, \Sigma), \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

其中 σ_1^2, σ_2^2 已知。求 μ 的无偏估计量的协方差矩阵的 Cramér–Rao 下界，并与样本均值的协方差矩阵比较。

解答：

步骤 1：单个样本的概率密度与对数似然

单个样本的概率密度函数：

$$p(x | \mu) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right]\right)$$

n 个样本的对数似然为：

$$\ln L(\mu) = -n \ln(2\pi\sigma_1\sigma_2) - \frac{1}{2} \sum_{i=1}^n \left[\frac{(x_{i1} - \mu_1)^2}{\sigma_1^2} + \frac{(x_{i2} - \mu_2)^2}{\sigma_2^2} \right]$$

步骤 2：求对 μ_1, μ_2 的导数（得分函数）

$$\frac{\partial}{\partial \mu_1} \ln L = \frac{1}{\sigma_1^2} \sum_{i=1}^n (x_{i1} - \mu_1)$$

$$\frac{\partial}{\partial \mu_2} \ln L = \frac{1}{\sigma_2^2} \sum_{i=1}^n (x_{i2} - \mu_2)$$

步骤 3：计算 Fisher 信息矩阵

由于样本独立，且两个分量独立：

- $\text{Var}\left(\frac{\partial}{\partial \mu_1} \ln L\right) = \frac{n}{\sigma_1^2}$
- $\text{Var}\left(\frac{\partial}{\partial \mu_2} \ln L\right) = \frac{n}{\sigma_2^2}$

- 协方差为 0

因此：

$$J(\boldsymbol{\mu}) = \begin{pmatrix} \frac{n}{\sigma_1^2} & 0 \\ 0 & \frac{n}{\sigma_2^2} \end{pmatrix}$$

步骤 4：求 Fisher 信息矩阵的逆

$$J^{-1}(\boldsymbol{\mu}) = \begin{pmatrix} \frac{\sigma_1^2}{n} & 0 \\ 0 & \frac{\sigma_2^2}{n} \end{pmatrix}$$

步骤 5：Cramér–Rao 下界 (CRLB)

多参数 CRLB 表示为：

$$\text{Cov}(\hat{\boldsymbol{\mu}}) \succeq J^{-1}(\boldsymbol{\mu})$$

即：

$$\text{Cov}(\hat{\boldsymbol{\mu}}) \succeq \begin{pmatrix} \frac{\sigma_1^2}{n} & 0 \\ 0 & \frac{\sigma_2^2}{n} \end{pmatrix}$$

步骤 6：样本均值与 CRLB 对比

样本均值：

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

协方差矩阵：

$$\text{Cov}(\bar{X}) = \frac{1}{n} \Sigma = \begin{pmatrix} \frac{\sigma_1^2}{n} & 0 \\ 0 & \frac{\sigma_2^2}{n} \end{pmatrix}$$

恰好等于 CRLB，因此样本均值是最优的无偏估计量（在方差意义下）。

4.5 最小方差无偏估计量与统计充分性完备性

4.5.1 最小方差无偏估计量 (UMVUE)

1) 定义

在所有无偏估计量 $\hat{\theta}$ 中，如果对所有参数值 θ ：

$$\text{Var}(\hat{\theta}_{\text{UMVUE}}) \leq \text{Var}(\hat{\theta})$$

则称 $\hat{\theta}_{\text{UMVUE}}$ 为参数 θ 的 **最小方差无偏估计量** (Uniformly Minimum Variance Unbiased Estimator, UMVUE)。

直观理解

- **无偏性**：估计值的平均等于真实参数值，没有系统性高估或低估。
- **方差最小**：在所有无偏估计量中，它的波动最小，结果最稳定。
- **Uniformly (统一地)**：上述比较对所有 θ 都成立，而不仅仅是某个特定参数值。

2) 存在性分析

UMVUE **不一定存在**。UMVUE 的存在性与样本中包含的参数信息密切相关，这由两个关键性质决定：

- **充分性 (Sufficiency)**：估计量不丢失关于参数的信息。
- **完备性 (Completeness)**：估计量的函数空间中不存在“无用的无偏函数”(即期望恒为 0 但不恒为零的函数)。

例题4.9：双参数正态分布

设：

$$X_1, X_2 \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

其中 μ 和 σ^2 都未知。

现在需要估计 μ 。

解题：

构建样本均值的两个无偏估计量，

平均估计量：

$$\hat{\mu}_1 = \frac{X_1 + X_2}{2}$$

混合估计量：

$$\hat{\mu}_2 = \frac{X_1 + X_2}{2} + c(X_1 - X_2)$$

其中常数 c 不依赖于 μ 。

检查估计量的无偏性

平均估计量：

$$\mathbb{E}[\hat{\mu}_1] = \frac{\mu + \mu}{2} = \mu$$

混合估计量：

$$\mathbb{E}[\hat{\mu}_2] = \frac{\mu + \mu}{2} + c(\mu - \mu) = \mu$$

因此两个估计量都是无偏的。

方差计算

平均估计量：

$$\text{Var}(\hat{\mu}_1) = \frac{\sigma^2}{2}$$

由于 X_1, X_2 独立：
混合估计量：

$$\begin{aligned}\text{Var}(\hat{\mu}_2) &= \text{Var}\left(\frac{X_1 + X_2}{2}\right) + c^2 \text{Var}(X_1 - X_2) \\ &= \frac{\sigma^2}{2} + c^2 \cdot 2\sigma^2\end{aligned}$$

当 $c = 0$ 时，方差最小为 $\frac{\sigma^2}{2}$ 。

如果我们不知道 σ^2 ，就无法比较不同估计量的方差在所有 $\theta = (\mu, \sigma^2)$ 下的大小，因为方差依赖于未知的 σ^2 。这意味着：

- 没有一个估计量能在所有 σ^2 下保证方差最小。
- 即使 $\hat{\mu}_1$ 在某个 σ^2 下最优，可能在另一个 σ^2 下不是最优。
- 因此，不存在一个**对所有参数值都方差最小**的无偏估计量。

4.5.2 充分性与因子分解定理

1) 充分性

直观解释：

一个统计量 $T(\mathbf{x})$ 关于参数 θ **充分 (sufficient)**，意味着在已知 $T(\mathbf{x})$ 的情况下，样本 \mathbf{x} 中不再包含对 θ 的额外信息。

换句话说， $T(\mathbf{x})$ 提取了样本中与 θ 相关的全部信息。

形式化定义：

设样本观测值为 $\mathbf{x} = (x_1, x_2, \dots, x_n)$ ，其联合概率密度（或质量）函数为 $f(\mathbf{x} | \theta)$ 。
统计量 $T(\mathbf{x})$ 关于参数 θ 充分，当且仅当：

$$f(\mathbf{x} | T(\mathbf{x}), \theta) = f(\mathbf{x} | T(\mathbf{x}))$$

即条件分布 $\mathbf{x} | T(\mathbf{x})$ 与参数 θ 无关。

2) 因子分解定理 (Neyman 因子分解定理)

定理：

统计量 $T(\mathbf{x})$ 关于参数 θ 充分，当且仅当存在函数 $g_\theta(T)$ 与 $h(\mathbf{x})$ ，使得：

$$f(\mathbf{x} | \theta) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$$

对所有 \mathbf{x} 和 θ 成立。

其中：

- $g_\theta(T)$ 依赖于参数 θ 和统计量 $T(\mathbf{x})$ ，但不依赖于 \mathbf{x} 的其他部分；

- $h(\mathbf{x})$ 不依赖于参数 θ 。

证明：

(必要性)

假设 $T(\mathbf{x})$ 关于 θ 充分。

由充分性定义，条件分布 $f(\mathbf{x} | T(\mathbf{x}), \theta)$ 与 θ 无关。记：

$$h(\mathbf{x}) := f(\mathbf{x} | T(\mathbf{x}))$$

这是一个只依赖于 \mathbf{x} 的函数，不含 θ 。

根据条件概率公式：

$$f(\mathbf{x} | \theta) = f(\mathbf{x} | T(\mathbf{x}), \theta) \cdot f(T(\mathbf{x}) | \theta)$$

代入 $h(\mathbf{x})$ ：

$$f(\mathbf{x} | \theta) = h(\mathbf{x}) \cdot f(T(\mathbf{x}) | \theta)$$

令：

$$g_\theta(T) := f(T(\mathbf{x}) | \theta)$$

则：

$$f(\mathbf{x} | \theta) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$$

这就是因子分解形式。

(充分性)

反过来，假设存在函数 $g_\theta(T)$ 和 $h(\mathbf{x})$ ，使：

$$f(\mathbf{x} | \theta) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$$

对所有 \mathbf{x}, θ 成立。

我们来求条件分布：

$$f(\mathbf{x} | T(\mathbf{x}), \theta) = \frac{f(\mathbf{x} | \theta)}{\sum_{\mathbf{y}: T(\mathbf{y})=T(\mathbf{x})} f(\mathbf{y} | \theta)} \quad (\text{离散情形})$$

或：

$$f(\mathbf{x} | T(\mathbf{x}), \theta) = \frac{f(\mathbf{x} | \theta)}{\int_{T(\mathbf{y})=T(\mathbf{x})} f(\mathbf{y} | \theta) d\mathbf{y}} \quad (\text{连续情形})$$

代入因子分解形式：

$$f(\mathbf{x} | T(\mathbf{x}), \theta) = \frac{g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})}{g_\theta(T(\mathbf{x})) \cdot \sum_{\mathbf{y}: T(\mathbf{y})=T(\mathbf{x})} h(\mathbf{y})}$$

其中分子分母的 $g_\theta(T(\mathbf{x}))$ 相互抵消，剩下的部分仅依赖于 \mathbf{x} 和 $T(\mathbf{x})$ ，而与 θ 无关。

因此条件分布不依赖于 θ ，由充分性定义可知 $T(\mathbf{x})$ 关于 θ 充分。

3) 例题

例题4.10:

设 $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ 。证明 $\sum_{i=1}^n X_i$ 是关于 λ 的充分统计量。

解答:

联合概率密度:

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \lambda^{\sum x_i} e^{-n\lambda} \prod_{i=1}^n \frac{1}{x_i!}$$

写为因子分解形式:

$$g_\lambda(T) = \lambda^T e^{-n\lambda}, \quad h(\mathbf{x}) = \prod_{i=1}^n \frac{1}{x_i!}$$

其中 $T = \sum x_i$ 。

所以 T 是充分统计量。

4.5.3 Rao-Blackwell 定理

1) 定理

给定任意无偏估计量 $\hat{\theta}$ 和充分统计量 T ，定义

$$\hat{\theta}^*(\mathbf{x}) = \mathbb{E}[\hat{\theta}|T(\mathbf{x})]$$

则 $\hat{\theta}^*$ 是关于 θ 的无偏估计量，且

$$\text{Var}(\hat{\theta}^*) \leq \text{Var}(\hat{\theta})$$

引理1: 全期望公式 (塔式法则)

$$\mathbb{E}[\mathbb{E}(X_2 | X_1)] = \mathbb{E}(X_2)$$

证明:

记联合概率密度为 $p(x_1, x_2)$ ，边缘密度为 $p(x_1)$ ，条件密度为 $p(x_2 | x_1)$ 。

$$\begin{aligned}
\mathbb{E}[\mathbb{E}(X_2 | X_1)] &= \int_{x_1} \mathbb{E}(X_2 | X_1 = x_1) p(x_1) dx_1 \\
&= \int_{x_1} \left[\int_{x_2} x_2 p(x_2 | x_1) dx_2 \right] p(x_1) dx_1 \\
&= \int_{x_1} \int_{x_2} x_2 p(x_2 | x_1) p(x_1) dx_2 dx_1 \\
&= \int_{x_1} \int_{x_2} x_2 p(x_1, x_2) dx_2 dx_1 \\
&= \int_{x_2} x_2 \left[\int_{x_1} p(x_1, x_2) dx_1 \right] dx_2 \\
&= \int_{x_2} x_2 p(x_2) dx_2 \\
&= \mathbb{E}(X_2)
\end{aligned}$$

引理 2: 已知 $\mu_i = E(X_i)$, $i = 1, 2$, 则

$$E\{[X_2 - E(X_2 | X_1)][E(X_2 | X_1) - \mu_2]\} = 0$$

证明:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x_2 - E(X_2 | X_1)][E(X_2 | X_1) - \mu_2] p(x_1, x_2) dx_2 dx_1 \\
&= \int_{-\infty}^{\infty} p(x_1) [E(X_2 | X_1) - \mu_2] \int_{-\infty}^{\infty} [x_2 - E(X_2 | X_1)] p(x_2 | x_1) dx_2 dx_1
\end{aligned}$$

其中

$$\int_{-\infty}^{\infty} [x_2 - E(X_2 | X_1 = x_1)] p(x_2 | x_1) dx_2 = E(X_2 | X_1 = x_1) - E(X_2 | X_1 = x_1) = 0$$

定理证明:

$$\begin{aligned}
\text{Var}(X_2) &= E[(X_2 - \mu_2)^2] = E\left\{[(X_2 - E(X_2 | X_1)) + (E(X_2 | X_1) - \mu_2)]^2\right\} \\
&= E[(X_2 - E(X_2 | X_1))^2] + E[(E(X_2 | X_1) - \mu_2)^2] \\
&\geq [E(E(X_2 | X_1) - \mu_2)^2] \\
&= \text{Var}[E(X_2 | X_1)]
\end{aligned}$$

其中 $E[E(X_2 | X_1)] = E(X_2) = \mu_2$ 。当等号成立时, 满足条件

$$E[(X_2 - E(X_2 | X_1))^2] = 0$$

即 $X_2 = E(X_2 | X_1)$, 也就是意味着 X_2 是 X_1 的函数; 反过来, 如果 $X_2 = h(X_1)$, 则

$$E(X_2 | X_1) = h(X_1) = X_2$$

因此，当且仅当 X_2 是 X_1 的函数时，等号成立。

2) 例题

例题4.11

设 X_1 和 X_2 是相互独立的参数均为 λ 泊松分布随机变量。

- (1) 寻找参数 λ 的充分估计量；
- (2) 验证

$$W = \begin{cases} 1, & X_1 = 0 \\ 0, & \text{其他} \end{cases}$$

是 $e^{-\lambda}$ 的无偏估计量；

- (3) 计算 $E[W | X_1 + X_2 = y]$ ；
- (4) 对于估计量 W ，寻找 $e^{-\lambda}$ 最佳的无偏估计量。

解答：

- (1) 由 X_1 和 X_2 的联合概率密度函数为：

$$p(x_1, x_2 | \lambda) = p(x_1 | \lambda) p(x_2 | \lambda) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!}$$

$$= \frac{(2\lambda)^{x_1+x_2} e^{-2\lambda}}{x_1! x_2!} \cdot \frac{x_1! x_2!}{(x_1 + x_2)!} \cdot \frac{1}{\binom{x_1+x_2}{x_1}}$$

$$= \frac{(2\lambda)^{x_1+x_2} e^{-2\lambda}}{(x_1 + x_2)!} \cdot \frac{1}{\binom{x_1+x_2}{x_1}}$$

因此， $X_1 + X_2$ 是 λ 的充分估计量。

- (2)

$$E[W] = P(W = 1) = P(X_1 = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda}$$

该估计量为无偏估计量。

- (3)

$$E[W | X_1 + X_2 = y] = P(W = 1 | X_1 + X_2 = y)$$

$$\begin{aligned}
 &= P(X_1 = 0 \mid X_1 + X_2 = y) = \frac{P(X_1 = 0, X_2 = y)}{P(X_1 + X_2 = y)} \\
 &= \frac{P(X_1 = 0) \cdot P(X_2 = y)}{P(X_1 + X_2 = y)} = \frac{e^{-\lambda} \cdot \frac{\lambda^y e^{-\lambda}}{y!}}{\frac{(2\lambda)^y e^{-2\lambda}}{y!}} = \left(\frac{1}{2}\right)^y
 \end{aligned}$$

(4) 利用 Rao-Blackwell 定理得到：

$$\phi(X_1, X_2) = \left(\frac{1}{2}\right)^{X_1+X_2}$$

是一个比 W 更好的 $e^{-\lambda}$ 无偏估计量。

4.5.4 完备性与指数分布族

1) 完备性

设 $g(t)$ 是定义在统计量 $T(x)$ 的值域上的任一实值函数，如果对所有的 $\theta \in \Theta$,

$$E_{\theta}[g(T)] = 0$$

成立时， $g(T) = 0$ 必成立，则称统计量 $T(x)$ 是完备的 (Complete)。

例题4.12 设 x_1, x_2, \dots, x_n 是来自两点分布 $B(1, \theta)$ 的样本 ($0 < \theta < 1$)，证明 \bar{x} 是完备统计量。

求解：

因为 x 服从 $B(1, \theta)$ ，所以：

$$\begin{aligned}
 E_{\theta}[g(\bar{x})] &= \sum_{k=0}^n g\left(\frac{k}{n}\right) \binom{n}{k} \theta^k (1-\theta)^{n-k} \\
 &= (1-\theta)^n \sum_{k=0}^n g\left(\frac{k}{n}\right) \binom{n}{k} \left(\frac{\theta}{1-\theta}\right)^k
 \end{aligned}$$

令 $E_{\theta}[g(\bar{x})] = 0$ ，有：

$$\sum_{k=0}^n g\left(\frac{k}{n}\right) \binom{n}{k} \left(\frac{\theta}{1-\theta}\right)^k = 0$$

因为上式的左边是 $\frac{\theta}{1-\theta}$ 的多项式，所以对所有的 $\theta \in (0, 1)$ ，欲使上式恒成立，只有右边多项式的系数为零，即：

$$g\left(\frac{k}{n}\right) = 0, \quad k = 0, 1, \dots, n$$

故对分布 $B(1, \theta)$ 而言， \bar{x} 是完备统计量。

2) 指数分布族

指数分布族的概率密度函数可以表示为：

$$p_{\theta}(x) = h(x) \exp \left[\sum_{j=1}^n w_j(\theta) T_j(x) \right] c(\theta)$$

例如：

(1) 二项分布：

$$p_{\theta}(x) = \binom{k}{x} \theta^x (1 - \theta)^{k-x}, \quad x = 0, 1, \dots, k, \quad \text{其中 } 0 < \theta < 1.$$

$$c(\theta) = (1 - \theta)^k, \quad h(x) = \binom{k}{x}, \quad w_1(\theta) = \ln \theta - \ln(1 - \theta), \quad T_1(x) = x, \quad m = 1.$$

(2) 泊松分布：

$$p_{\theta}(x) = e^{-\theta} \frac{\theta^x}{x!}, \quad x = 0, 1, \dots, \quad \text{其中 } \theta > 0.$$

$$c(\theta) = e^{-\theta}, \quad h(x) = \frac{1}{x!}, \quad w_1(\theta) = \ln \theta, \quad T_1(x) = x, \quad m = 1.$$

(3) 正态分布：

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right], \quad -\infty < x < \infty, \quad \theta = [\mu, \sigma^2].$$

$$c(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{\mu^2}{2\sigma^2} \right], \quad h(x) = 1, \quad w_1(\theta) = -\frac{1}{2\sigma^2}, \quad T_1(x) = x^2,$$

$$w_1(\theta) = \frac{\mu}{\sigma^2}, \quad T_1(x) = x, \quad m = 2.$$

(4) 伽马分布：

$$p_{\theta}(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad x > 0, \quad \theta = [\alpha, \beta].$$

$$c(\theta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}, \quad h(x) = 1, \quad w_1(\theta) = \alpha - 1, \quad T_1(x) = \ln x, \quad w_1(\theta) = -\frac{1}{\beta},$$

$$T_1(x) = x, \quad m = 2.$$

(5) 贝塔分布：

$$p_{\theta}(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \theta = [\alpha, \beta].$$

$$c(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}, \quad h(x) = 1, \quad w_1(\theta) = \alpha - 1, \quad T_1(x) = \ln x, \quad w_1(\theta) = \beta - 1,$$

$$T_1(x) = \ln(1-x), \quad m = 2.$$

3) 指数分布族的完备性

设 $x = (x_1, x_2, \dots, x_n)$ 是来自总体 $\{p_{\theta}, \theta \in \Theta\}$ 的一个样本，其概率密度函数可表示为：

$$p(x, \theta) = h(x) \exp \left\{ \sum_{i=1}^m w_i(\theta) T_i(x) - A(\theta) \right\} = h(x) \exp \left\{ \sum_{i=1}^m w_i T_i(x) - A(\theta) \right\}$$

其中：

$$w = w(\theta) = [w_1(\theta), \dots, w_m(\theta)] \in \Omega \subset \mathbb{R}^m, \quad [T_1(x), \dots, T_m(x)] = \left[\sum_{i=1}^n t_1(x_i), \dots, \sum_{i=1}^n t_m(x_i) \right]$$

如果 Ω 有内点，则统计量 $[T_1(x), \dots, T_m(x)]$ 是完备的。

证明：

样本来自指数族：

$$p(x | w) = h(x) \exp \{w^{\top} T(x) - A(w)\}, \quad w \in \Omega \subset \mathbb{R}^m$$

其中：

- $w = w(\theta)$ 是自然参数；
- $T(x)$ 是自然统计量；
- $A(w)$ 是对数配分函数 (log-partition function)：

$$A(w) = \log \int h(x) \exp\{w^{\top} T(x)\} dx$$

- Ω 是自然参数空间。

假设条件

假设 $g(T)$ 是任意可测函数，并且：

$$\mathbb{E}_w[g(T)] = 0, \quad \forall w \in \Omega$$

我们要证明：

$$P(g(T) = 0) = 1$$

步骤 1：将期望写成积分形式

$$\mathbb{E}_w[g(T)] = \int g(T(x)) h(x) \exp\{w^\top T(x) - A(w)\} dx$$

由于 $A(w)$ 与 x 无关, 可以写成:

$$\mathbb{E}_w[g(T)] = e^{-A(w)} \int g(T(x)) h(x) \exp\{w^\top T(x)\} dx$$

步骤 2: 识别解析性

定义:

$$F(w) = \int g(T(x)) h(x) \exp\{w^\top T(x)\} dx$$

这是自然参数 w 的一个多元解析函数 (因为指数函数在 w 上解析, 积分在有内点的 Ω 内收敛且解析性保持)。

步骤 3: 利用解析函数唯一性定理

我们已知:

$$\mathbb{E}_w[g(T)] = e^{-A(w)} F(w) = 0, \quad \forall w \in \Omega$$

由于 $e^{-A(w)}$ 在 Ω 内非零, 得到:

$$F(w) \equiv 0, \quad \forall w \in \Omega$$

因为 Ω 有内点, 且 $F(w)$ 是解析函数, 解析函数唯一性定理告诉我们: 如果解析函数在一个开集上恒等于零, 则它在整个解析域内恒等于零。

因此:

$$F(w) \equiv 0 \quad \text{在整个解析域内成立}$$

步骤 4: 推导 $g(T) = 0$

$F(w) \equiv 0$ 意味着:

$$\int g(T(x)) h(x) \exp\{w^\top T(x)\} dx = 0, \quad \forall w$$

这个积分实际上是 $g(T)$ 与所有指数函数 $\exp\{w^\top T\}$ 的加权内积。

指数函数族 $\{\exp(w^\top t) : w \in \mathbb{R}^m\}$ 在合适的函数空间中是完备的 (它们的线性组合可以逼近任意足够好的函数), 因此如果与所有这样的指数函数的积分都为零, 就只能说明:

$$g(T(x)) = 0 \quad \text{几乎处处成立}$$

例题4.13: Bernoulli 分布的完备性

设:

$$X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), \quad p \in (0, 1)$$

证明统计量：

$$T = \sum_{i=1}^n X_i$$

是关于参数 p 的完备统计量。

解答

1. 写出条件期望

设 $g(\cdot)$ 为任意可测函数，若：

$$\mathbb{E}_p[g(T)] = 0, \quad \forall p \in (0, 1)$$

则：

$$\mathbb{E}_p[g(T)] = \sum_{k=0}^n g(k) \binom{n}{k} p^k (1-p)^{n-k} = 0, \quad \forall p \in (0, 1)$$

2. 利用多项式恒等原理

注意到上式是关于 p 的一个多项式：

$$P(p) = \sum_{k=0}^n a_k p^k (1-p)^{n-k}, \quad a_k = g(k) \binom{n}{k}$$

如果 $P(p)$ 在区间 $(0, 1)$ 上恒等于零，则它的所有系数必须为零，即：

$$a_k = g(k) \binom{n}{k} = 0, \quad \forall k$$

3. 系数为零推出函数为零

由于 $\binom{n}{k} > 0$ ，所以：

$$g(k) = 0, \quad \forall k \in \{0, 1, \dots, n\}$$

4. 结论

因此：

$$P(g(T) = 0) = 1$$

由定义可知，统计量 $T = \sum_{i=1}^n X_i$ 是完备的。 ■

4.5.5 Lehmann–Scheffé 定理

1) 定理

如果 $T(\mathbf{x})$ 是充分且完备统计量，且 $\hat{\tau}(T)$ 是关于参数函数 $\tau(\theta)$ 的无偏估计量，则 $\hat{\tau}(T)$ 是唯一的 UMVUE（最小方差无偏估计量）。

2) 例题

例题4.14: 均匀分布上限的 UMVUE

题目

设：

$$X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} U(0, \theta), \quad \theta > 0 \text{ 未知}$$

求参数 θ 的 UMVUE，并说明理由。

解答：

(1) 找到充分统计量

在均匀分布 $U(0, \theta)$ 中，最大次序统计量：

$$X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$$

是关于 θ 的充分统计量（由因子分解定理可证）。

(2) 求 $X_{(n)}$ 的分布

单个样本的分布函数为：

$$F_X(x) = \frac{x}{\theta}, \quad 0 < x < \theta$$

因此：

$$F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^n, \quad 0 < x < \theta$$

对 x 求导得到概率密度函数：

$$f_{X_{(n)}}(x) = n \frac{x^{n-1}}{\theta^n}, \quad 0 < x < \theta$$

(3) 计算期望值

$$\mathbb{E}[X_{(n)}] = \int_0^\theta x \cdot n \frac{x^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1} \theta$$

(4) 构造无偏估计量

由 $\mathbb{E}[X_{(n)}] = \frac{n}{n+1} \theta$ ，可得：

$$\hat{\theta} = \frac{n+1}{n} X_{(n)}$$

是 θ 的无偏估计量。

(5) 完备性与 UMVUE 判定

已知 $X_{(n)}$ 是充分统计量，并且对于均匀分布，最大次序统计量也是完备的（可通过分布族结构证明）。

因此，由 **Lehmann–Scheffé 定理**， $\hat{\theta} = \frac{n+1}{n} X_{(n)}$ 是 θ 的 UMVUE。

最终答案

$$\hat{\theta} = \frac{n+1}{n} X_{(n)}$$

是 θ 的 UMVUE。

4.6 经典频率派估计方法

经典频率派统计学是现代统计学的基石之一，其核心思想在于将未知参数视为固定的常数，而数据则是随机的。因此，频率派估计方法的目标是寻找一个完全基于样本数据的、良好的“估计规则”或“估计量”，并评估该估计量在长期、重复抽样下的表现（如无偏性、有效性、一致性等）。本节介绍四种常用估计方法：极大似然估计（MLE）、最小二乘估计、加权最小二乘估计、以及极大极小估计。

首先引入极大似然估计（MLE）与最小二乘估计（LS）这两种在频率派框架中最核心、最通用的方法，它们构成了理解后续内容的理论基石。然后，针对同方差假设不再成立时，加权最小二乘估计（WLS）作为解决方案被引入，展示了方法在面对现实复杂性时的演进与适应。关注最坏情况下的最小风险，这与极大似然估计（MLE）与最小二乘估计（LS）等基于平均表现最优的理念存在显著差异，代表了频率派框架下更保守、更稳健的思维方式。

4.6.1 极大似然估计（MLE）

极大似然估计是统计学中最为重要和通用的参数估计方法之一。其核心思想直观而深刻：**在已知观测样本的前提下，寻找最有可能产生这些观测结果的参数值。**它是一种建立在似然函数基础上的频率学派方法。

1) 基本概念与似然函数

思想起源

设想我们有一个依赖于未知参数 θ 的概率模型。我们进行了多次独立试验，得到一组观测样本 $\mathbf{x} = (x_1, x_2, \dots, x_n)$ 。一个很自然的问题是：哪个参数值 θ 能够最“合理”地解释我们观测到的这组数据？

极大似然估计给出的答案是：**选择那个能使观测到当前样本的“可能性”达到最大的参数值。**

数学定义

设样本 $\mathbf{x} = (x_1, \dots, x_n)$ 来自概率密度函数（或概率质量函数）为 $f(x; \theta)$ 的分布，其中 $\theta \in \Theta$ 为未知参数。

- **似然函数**：将联合概率密度（或概率）视为参数的函数，即定义为似然函数：

$$L(\theta; \mathbf{x}) = \prod_{i=1}^n f(x_i; \theta)$$

它度量了在不同 θ 下，出现当前样本 \mathbf{x} 的相对可能性。

- **对数似然函数**：由于连乘计算不便且易产生数值下溢，我们通常对其取自然对数，定义对数似然函数：

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^n \log f(x_i; \theta)$$

因为对数函数是单调递增的，最大化 $L(\theta)$ 与最大化 $\ell(\theta)$ 是等价的。

- **极大似然估计量**：MLE估计量 $\hat{\theta}_{\text{MLE}}$ 就是使得似然函数或对数似然函数达到最大的参数值：

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in \Theta} L(\theta; \mathbf{x}) = \arg \max_{\theta \in \Theta} \ell(\theta; \mathbf{x})$$

2) 求解方法与数值算法

在理想情况下，我们可以通过解析方法求得MLE的精确解。

求导法

当似然函数关于参数 θ 可导且最大值出现在参数空间内部时，MLE通常是似然方程的解：

$$\frac{\partial \ell(\theta; \mathbf{x})}{\partial \theta} = 0$$

对于多参数情况，则需要求解方程组 $\nabla \ell(\boldsymbol{\theta}) = 0$ 。

例题4.15：伯努利分布的MLE

假设我们独立地抛掷一枚不均匀的硬币 n 次，观测到 k 次正面向上。设正面朝上的概率为 p ，求参数 p 的极大似然估计。

解答：

- **建立模型**：每次抛掷是一个伯努利试验，其概率质量函数为 $P(X = 1) = p, P(X = 0) = 1 - p$ 。因此，似然函数为：

$$L(p; k) = p^k (1 - p)^{n-k}$$

- **写出对数似然**：

$$\ell(p) = \log L(p) = k \log p + (n - k) \log(1 - p)$$

- **最大化对数似然**：对 p 求导并令其为零：

$$\frac{d\ell}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = 0$$

解方程得：

$$\hat{p}_{\text{MLE}} = \frac{k}{n}$$

这个结果非常直观：参数 p 的MLE就是观测到的“正面”出现的频率。

数值优化方法

当似然方程没有解析解或求解困难时（例如混合模型），我们需要借助数值优化算法，如**梯度上升法**、**牛顿-拉弗森法**等，来寻找使得似然函数最大的参数值。

3) 大样本理论：MLE的优良性质

MLE之所以占据核心地位，不仅在于其思想的直观性，更在于它在样本量增大时展现出的一系列优异统计性质（称为“大样本性质”或“渐近性质”）。这些性质在一定的正则性条件下成立。

一致性

一致性是估计量最基本的要求。它表明，当样本量 n 趋于无穷大时，MLE会以概率收敛到参数的真实值 θ_0 。

$$\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta_0 \quad \text{当 } n \rightarrow \infty$$

这意味着只要有足够多的数据，MLE就能无限接近真相。

渐近正态性

MLE的抽样分布，在大样本下近似于一个正态分布：

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$$

其中， $\mathcal{I}(\theta_0)$ 是 **Fisher 信息量**，定义为：

$$\mathcal{I}(\theta) = -\mathbb{E} \left[\frac{\partial^2 \ell(\theta; X)}{\partial \theta^2} \right]$$

这个性质非常实用，它允许我们基于正态分布来构建参数的置信区间和进行假设检验。估计量的方差 $\text{Var}(\hat{\theta}_{\text{MLE}}) \approx [n\mathcal{I}(\theta_0)]^{-1}$ 。

渐近有效性

在所有“表现良好”的渐近无偏估计量中，MLE的渐近方差是最小的，达到了 **Cramér-Rao 下界**。这意味着MLE是最精确的估计量之一。

例题4.16：验证渐近性质

设 $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ ，其中方差 σ^2 已知。求均值参数 μ 的MLE，并验证MLE的渐近方差。

解答：

- 写出似然函数：

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

- 写出对数似然函数：

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

- 求解似然方程：

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

得到：

$$\sum_{i=1}^n (x_i - \mu) = 0 \Rightarrow \hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

结论：对于正态分布，均值的MLE就是样本均值。

- **计算Fisher信息量：**首先求二阶导数。

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

其期望值为：

$$\mathcal{I}(\mu) = -\mathbb{E} \left[-\frac{n}{\sigma^2} \right] = \frac{n}{\sigma^2}$$

- **计算渐近方差：**根据渐近正态性，有

$$\text{Var}(\hat{\mu}_{\text{MLE}}) \approx [\mathcal{I}(\mu)]^{-1} = \frac{\sigma^2}{n}$$

而我们知道样本均值的精确方差就是 $\frac{\sigma^2}{n}$ 。这恰好说明 $\hat{\mu}_{\text{MLE}}$ 的方差达到了Cramér-Rao下界，不仅渐近有效，对于有限样本也是有效的。

总结：极大似然估计提供了一种强大而统一的参数估计框架。它始于一个直观的原则，并通过其卓越的大样本性质，为统计推断奠定了坚实的理论基础。

4.6.2 最小二乘估计 (LS)

最小二乘估计是解决线性回归问题最直观、最核心的方法。它不依赖于误差项的具体分布假设，其核心思想是：**寻找能使模型预测值与真实观测值之间“差距”的平方和达到最小的参数值。**

1) 普通最小二乘 (OLS)

定义

考虑一个经典的线性回归模型：

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

其中， \mathbf{y} 是 $n \times 1$ 的观测向量， X 是已知的 $n \times p$ 设计矩阵， $\boldsymbol{\beta}$ 是待估计的 $p \times 1$ 参数向量， $\boldsymbol{\varepsilon}$ 是随机误差项，通常假设其满足 $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ 和 $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$ 。

普通最小二乘 (OLS) 估计通过最小化**残差平方和 (RSS)** 来寻找 $\boldsymbol{\beta}$ ：

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg \min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

通过对上述凸优化问题求导，可以得出OLS估计量的解析解。当设计矩阵 X 列满秩（即 $X^\top X$ 可逆）时，解为：

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

该解具有清晰的几何解释：它找到了将观测向量 \mathbf{y} 投影到由 X 的列张成的子空间上的最佳逼近。

例题4.17：简单线性回归计算

假设我们有一组关于变量 x 和 y 的观测数据： $(1, 1), (2, 3), (3, 3)$ 。请使用OLS方法拟合线性模型 $y = \beta_0 + \beta_1 x + \varepsilon$ 。

解答：

- **建立模型矩阵与向量：**

设计矩阵 X 包含一列1（对应截距）和 x 的值，观测向量为 \mathbf{y} 。

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix}$$

- **计算 $X^\top X$ 和 $X^\top \mathbf{y}$ ：**

$$X^\top X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, \quad X^\top \mathbf{y} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 7 \\ 16 \end{bmatrix}$$

- **求解参数：**

$$\hat{\beta} = (X^\top X)^{-1}(X^\top \mathbf{y}) = \frac{1}{(3 \times 14 - 6 \times 6)} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} \begin{bmatrix} 7 \\ 16 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} \begin{bmatrix} 7 \\ 16 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 2 \\ 6 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1 \end{bmatrix}$$

因此，拟合的回归直线为 $\hat{y} = \frac{1}{3} + x$ 。

2) Gauss-Markov 定理

在获得了OLS估计量的表达式后，一个自然的问题是：**这个估计量好吗？它有何优良性质？** Gauss-Markov定理从频率学派的角度，为OLS提供了坚实的理论保障。

定理

在满足以下基本假定的经典线性回归模型中：

- **线性于参数：** $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 。
- **严格外生性：** $\mathbb{E}(\boldsymbol{\varepsilon}|X) = 0$ 。
- **无多重共线性：** X 是列满秩的。
- **球形误差：** $\text{Cov}(\boldsymbol{\varepsilon}|X) = \sigma^2 I_n$ （同方差且无自相关）。

那么，普通最小二乘估计量 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 是**最佳线性无偏估计量（BLUE）**。

- **线性（L）：** $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 是观测值 \mathbf{y} 的线性函数。
- **无偏（U）：** $\mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \boldsymbol{\beta}$ 。
- **最佳（B）：** 在所有线性无偏估计量中， $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 具有最小的方差（或者说，其协方差矩阵是最“小”的，即与其他线性无偏估计量的协方差矩阵之差是半负定矩阵）。

【意义与启示】

该定理的伟大之处在于，它**不需要**指定误差项的具体分布（如正态分布），仅基于模型的基本假定，就保证了OLS在

“线性无偏”这一类估计量中的最优性。这解释了为何OLS在实践中如此受欢迎。如果基本假定被违背（如存在异方差或自相关），OLS就不再是BLUE，此时就需要我们后面将要介绍的**加权最小二乘（WLS）**等方法进行修正。

例题4.18：理解BLUE的含义

假设有两个统计学家，各自提出了一个关于参数 β 的线性无偏估计量 $\tilde{\beta}_1$ 和 $\tilde{\beta}_2$ 。根据Gauss-Markov定理，关于OLS估计量 $\hat{\beta}_{OLS}$ 与这两个估计量的方差，我们能得出什么结论？

解答：

根据Gauss-Markov定理，OLS估计量 $\hat{\beta}_{OLS}$ 是BLUE。这意味着：

- 对于参数向量中的**任何一个**分量，例如 β_1 ，有 $\text{Var}(\hat{\beta}_{1,OLS}) \leq \text{Var}(\tilde{\beta}_{11})$ 且 $\text{Var}(\hat{\beta}_{1,OLS}) \leq \text{Var}(\tilde{\beta}_{12})$ 。
 - 更一般地，对于任意的线性组合 $\mathbf{c}^\top \beta$ ，有 $\text{Var}(\mathbf{c}^\top \hat{\beta}_{OLS}) \leq \text{Var}(\mathbf{c}^\top \tilde{\beta}_1)$ 且 $\text{Var}(\mathbf{c}^\top \hat{\beta}_{OLS}) \leq \text{Var}(\mathbf{c}^\top \tilde{\beta}_2)$ 。
- 因此，在任何方向上，OLS估计量的波动性都不会比任何其他线性无偏估计量更大。

3) 最小二乘与极大似然的关系

前面两节分别阐述了OLS的算法和其优良的“线性无偏”性质。本节将揭示OLS与另一大统计思想——**极大似然估计（MLE）**的深刻联系，从而在更一般的分布假设下，为OLS提供另一层面的合理性。

OLS的目标函数是残差平方和，这本质上是一种**平方损失函数**。而MLE的原则是**选择最可能产生当前观测样本的参数**。这两者看似不同，但在一个常见的假设下会殊途同归。

假设线性模型中的误差项服从正态分布：

$$\epsilon \sim N(\mathbf{0}, \sigma^2 I_n) \Rightarrow \mathbf{y}|X \sim N(X\beta, \sigma^2 I_n)$$

此时，样本的似然函数为：

$$L(\beta, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \beta)^2\right)$$

对应的对数似然函数为：

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{y} - X\beta\|^2$$

等价性

观察上式，对于固定的 σ^2 ，**最大化对数似然函数 $\ell(\beta, \sigma^2)$ 关于 β 的部分，完全等价于最小化残差平方和 $\|\mathbf{y} - X\beta\|^2$** 。

因此，我们得到一个重要结论：

在线性回归模型满足误差项独立同分布于正态分布 $N(0, \sigma^2)$ 的假设下，极大似然估计量 $\hat{\beta}_{MLE}$ 与普通最小二乘估计量 $\hat{\beta}_{OLS}$ 是完全相同的：

$$\hat{\beta}_{MLE} = \hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top \mathbf{y}$$

意义

这一等价关系具有双重意义：

1. **为OLS提供了分布意义上的支持**：它表明，当数据生成过程确实服从正态分布时，OLS不仅是BLUE，而且也是极大似然估计量，继承了MLE在大样本下的优良性质（如相合性、渐近正态性等）。
2. **揭示了MLE的一种具体形式**：它展示了在正态误差的线性模型下，MLE可以有一个简洁的解析解。

需要注意的是，如果误差分布不是正态的，那么MLE将不再是OLS的形式，其目标函数也会相应改变（例如，对于拉普拉斯分布的误差，MLE等价于最小化绝对误差和）。因此，OLS可以看作是**高斯似然假设下的一种特殊的MLE**。

例题4.19：验证等价性

给定一元线性模型 $y_i = \beta x_i + \varepsilon_i$ ，其中 $\varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ 。请分别推导参数 β 的OLS估计量和MLE估计量，并验证它们的一致性。

解答：

- **OLS估计量：**

最小化残差平方和 $S(\beta) = \sum (y_i - \beta x_i)^2$ 。

令导数 $\frac{dS}{d\beta} = -2 \sum x_i (y_i - \beta x_i) = 0$ 。

解得： $\hat{\beta}_{OLS} = \frac{\sum x_i y_i}{\sum x_i^2}$ 。

- **MLE估计量：**

似然函数为 $L(\beta) = \prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right)$ 。

对数似然函数为 $\ell(\beta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \beta x_i)^2$ 。

为了最大化 $\ell(\beta)$ ，需要最小化 $\sum (y_i - \beta x_i)^2$ 。

这与OLS的目标函数完全相同，因此：

$$\hat{\beta}_{MLE} = \arg \max_{\beta} \ell(\beta) = \arg \min_{\beta} \sum (y_i - \beta x_i)^2 = \frac{\sum x_i y_i}{\sum x_i^2} = \hat{\beta}_{OLS}$$

验证完毕，两者等价。

4.6.3 加权最小二乘（WLS）

1) 异方差问题与WLS的引入

在实际数据分析中，经典线性回归模型的基本假设往往难以完全满足。其中，**异方差性**是一个常见且重要的问题。

从同方差到异方差

回顾普通最小二乘（OLS）的基本假设：所有观测误差 ε_i 具有相同的方差 σ^2 ，即：

$$\text{Var}(\varepsilon_i) = \sigma^2, \quad \forall i$$

这被称为**同方差性**。然而，在现实数据中，不同观测的误差方差常常不同：

$$\text{Var}(\varepsilon_i) = \sigma_i^2, \quad \sigma_i^2 \neq \sigma_j^2 \text{ for some } i \neq j$$

这种现象称为**异方差性**。

异方差的后果

当存在异方差时，OLS估计量虽然仍是无偏的，但**不再是最佳线性无偏估计（BLUE）**。具体表现为：

- OLS估计量不再是有效的（方差不再最小）
- 标准误的估计有偏，导致假设检验不可靠
- 置信区间的覆盖概率不准确

WLS的直观思想

加权最小二乘（WLS）的核心思想很直观：**给不同可靠性的观测分配不同的权重**。具体来说：

- 对于方差较小的观测（数据质量高，更可靠），赋予较大的权重
- 对于方差较大的观测（数据质量低，不可靠），赋予较小的权重

通过这种方式，WLS能够克服异方差带来的问题，获得更优的参数估计。

模型设定

考虑扩展的线性回归模型：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

其中，误差项满足：

$$\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V}$$

这里 \mathbf{V} 是已知的 $n \times n$ 对称正定矩阵。最常见的特殊情况是：

$$\mathbf{V} = \text{diag}(v_1, v_2, \dots, v_n)$$

即各观测误差互不相关，但方差不同。

2) WLS的估计公式与推导

权矩阵的定义

定义权矩阵为协方差矩阵的逆：

$$\mathbf{W} = \mathbf{V}^{-1} = \text{diag}(w_1, w_2, \dots, w_n)$$

其中，每个观测的权重 $w_i = 1/v_i$ ，即权重与误差方差成反比。

目标函数

WLS通过最小化**加权残差平方和**来估计参数：

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^n w_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

参数估计的推导

通过对目标函数求导并令导数为零，可得WLS估计量：

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^{\top}W(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}$$

整理得正规方程：

$$X^{\top}WX\boldsymbol{\beta} = X^{\top}W\mathbf{y}$$

解此方程，得到WLS估计量：

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (X^{\top}WX)^{-1}X^{\top}W\mathbf{y}$$

等价变换视角

WLS可以理解为对原始数据进行变换后再应用OLS。令：

$$\mathbf{y}^* = W^{1/2}\mathbf{y}, \quad X^* = W^{1/2}X$$

则变换后的模型满足同方差假设，对其应用OLS得到的估计量与WLS估计量相同。

算法步骤

- **确定误差方差：**通过理论分析、经验或残差分析估计每个观测的误差方差 v_i
- **构造权矩阵：** $W = \text{diag}(1/v_1, 1/v_2, \dots, 1/v_n)$
- **计算加权矩阵：** $X^{\top}WX$ 和 $X^{\top}W\mathbf{y}$
- **求解参数：** $\hat{\boldsymbol{\beta}}_{\text{WLS}} = (X^{\top}WX)^{-1}X^{\top}W\mathbf{y}$

3) 实例分析与计算

例题4.20：异方差数据的三点回归

考虑以下三个观测点，已知每个点的误差方差不同：

x_i	y_i	方差 v_i	权重 $w_i = 1/v_i$
1	2.0	1	1.000
2	2.8	4	0.250
3	3.6	9	0.111

用加权最小二乘法拟合线性模型 $y = \beta_0 + \beta_1 x + \varepsilon$ 。

解答：

步骤1：构造矩阵

设计矩阵 X 、响应向量 \mathbf{y} 和权矩阵 W 分别为：

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 2.0 \\ 2.8 \\ 3.6 \end{pmatrix}, \quad W = \begin{pmatrix} 1.000 & 0 & 0 \\ 0 & 0.250 & 0 \\ 0 & 0 & 0.111 \end{pmatrix}$$

步骤2: 计算加权矩阵

首先计算 $X^T W X$:

$$X^T W X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1.000 & 0 & 0 \\ 0 & 0.250 & 0 \\ 0 & 0 & 0.111 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$

通过分量计算:

- $S_0 = 1.000 + 0.250 + 0.111 = 1.361$
- $S_1 = 1.000 \times 1 + 0.250 \times 2 + 0.111 \times 3 = 1.833$
- $S_2 = 1.000 \times 1^2 + 0.250 \times 2^2 + 0.111 \times 3^2 = 3.000$

因此:

$$X^T W X = \begin{pmatrix} 1.361 & 1.833 \\ 1.833 & 3.000 \end{pmatrix}$$

再计算 $X^T W y$:

- $b_0 = 1.000 \times 2.0 + 0.250 \times 2.8 + 0.111 \times 3.6 = 3.100$
- $b_1 = 1.000 \times 1 \times 2.0 + 0.250 \times 2 \times 2.8 + 0.111 \times 3 \times 3.6 = 4.600$

因此:

$$X^T W y = \begin{pmatrix} 3.100 \\ 4.600 \end{pmatrix}$$

步骤3: 求解参数

求解正规方程:

$$\begin{pmatrix} 1.361 & 1.833 \\ 1.833 & 3.000 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 3.100 \\ 4.600 \end{pmatrix}$$

计算系数矩阵的逆:

$$(X^T W X)^{-1} = \frac{1}{1.361 \times 3.000 - 1.833^2} \begin{pmatrix} 3.000 & -1.833 \\ -1.833 & 1.361 \end{pmatrix} = \frac{1}{1.361} \begin{pmatrix} 3.000 & -1.833 \\ -1.833 & 1.361 \end{pmatrix}$$

最终得到参数估计:

$$\hat{\beta}_{\text{WLS}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{1.361} \begin{pmatrix} 3.000 & -1.833 \\ -1.833 & 1.361 \end{pmatrix} \begin{pmatrix} 3.100 \\ 4.600 \end{pmatrix} = \begin{pmatrix} 1.2 \\ 0.8 \end{pmatrix}$$

拟合的加权最小二乘回归方程为 $\hat{y} = 1.2 + 0.8x$ 。

讨论

与普通最小二乘相比, WLS通过赋予不同观测不同的权重, 有效地处理了异方差问题。在实际应用中, 权重的确定是关键步骤, 可以通过残差分析、理论推导或经验判断来确定。

4.6.4 极大极小估计 (Minimax Estimation)

1) 稳健决策的基本思想

在统计学中，我们常常需要在不确定性下做出决策。传统的估计方法（如MLE、LS）通常关注估计量在“平均”意义上的优良性，但在某些情况下，我们更关心的是**最坏情况下的表现**。

从平均性能到最坏情况

考虑以下两种估计量：

- 估计量A：在90%的情况下表现优异，但在10%的情况下表现极差
- 估计量B：在所有情况下都表现良好，但从未达到最优

如果我们无法承受“极差”表现带来的后果，那么估计量B可能是更好的选择。这就是极大极小估计的核心思想：**做一个保守的决策者，优先防范最坏情况**。

决策理论框架

极大极小估计建立在统计决策理论的框架上，涉及三个基本要素：

- 损失函数** $L(\theta, \delta)$ ：度量当真实参数为 θ 时，采用决策 δ 所造成的损失
- 风险函数** $R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))]$ ：平均损失，期望是对样本分布取的
- 极大极小准则**：选择使最大风险最小的决策规则

2) 极大极小准则的数学表述

形式化定义

设 Θ 是参数空间， \mathcal{D} 是所有可能的决策规则集合。对于给定的决策规则 δ ，定义其最大风险为：

$$M(\delta) = \sup_{\theta \in \Theta} R(\theta, \delta)$$

极大极小决策规则 $\hat{\delta}$ 是使这个最大风险最小的决策规则：

$$\hat{\delta}_{\text{minimax}} = \arg \min_{\delta \in \mathcal{D}} M(\delta) = \arg \min_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta)$$

理解极大极小准则

这个定义可以分两步理解：

- 内层的 sup**：对于每个决策规则 δ ，考虑它在**所有可能参数值**下的风险，取最大的那个风险值
- 外层的 min**：在所有决策规则中，选择那个**最大风险最小**的

换句话说，极大极小估计量是“悲观主义者”的选择——它假设自然总是对我们不利（选择使我们风险最大的参数），我们在此基础上做出最优防御。

常用的损失函数

在实际应用中，常用的损失函数包括：

- 平方损失**： $L(\theta, \delta) = (\theta - \delta)^2$

- **绝对损失**: $L(\theta, \delta) = |\theta - \delta|$
- **0-1损失**: 在假设检验问题中使用

3) 实例分析：均匀分布参数的极大极小估计

例题4.21： 均匀分布参数的估计

设 $X \sim U(\theta - 1, \theta + 1)$ ，即 X 服从区间 $[\theta - 1, \theta + 1]$ 上的均匀分布。损失函数采用平方损失 $L(\theta, \delta) = (\theta - \delta)^2$ 。求参数 θ 的极大极小估计量。

解答：

步骤1：建立风险函数

考虑线性估计量 $\delta(X) = aX + b$ ，其中 a, b 是待定常数。

由于 $X \sim U(\theta - 1, \theta + 1)$ ，我们有：

- $\mathbb{E}_\theta[X] = \theta$
- $\text{Var}_\theta[X] = \frac{(2)^2}{12} = \frac{1}{3}$

现在计算风险函数：

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[(\theta - \delta(X))^2] \\ &= \mathbb{E}_\theta[(\theta - aX - b)^2] \\ &= \mathbb{E}_\theta[(\theta - b)^2 - 2a(\theta - b)X + a^2X^2] \\ &= (\theta - b)^2 - 2a(\theta - b)\mathbb{E}_\theta[X] + a^2\mathbb{E}_\theta[X^2] \end{aligned}$$

利用 $\mathbb{E}_\theta[X] = \theta$ 和 $\mathbb{E}_\theta[X^2] = \text{Var}_\theta[X] + (\mathbb{E}_\theta[X])^2 = \frac{1}{3} + \theta^2$ ，代入得：

$$\begin{aligned} R(\theta, \delta) &= (\theta - b)^2 - 2a(\theta - b)\theta + a^2\left(\frac{1}{3} + \theta^2\right) \\ &= \theta^2 - 2b\theta + b^2 - 2a\theta^2 + 2ab\theta + a^2\theta^2 + \frac{a^2}{3} \\ &= (1 - 2a + a^2)\theta^2 + (2ab - 2b)\theta + \left(b^2 + \frac{a^2}{3}\right) \end{aligned}$$

步骤2：分析风险函数的性质

观察风险函数 $R(\theta, \delta)$ ，它是关于 θ 的二次函数。系数的性质决定了最大风险的行为：

- 如果二次项系数 $1 - 2a + a^2 \neq 0$ ，那么当 $|\theta| \rightarrow \infty$ 时， $R(\theta, \delta) \rightarrow \infty$
- 最大风险 $M(\delta) = \sup_{\theta \in \mathbb{R}} R(\theta, \delta)$ 将是无穷大

步骤3：寻找有限最大风险的条件

为了使最大风险有限，必须消除风险函数对 θ 的依赖性。这要求：

1. 二次项系数为零： $1 - 2a + a^2 = 0$ ，即 $(1 - a)^2 = 0$ ，解得 $a = 1$
2. 一次项系数为零： $2ab - 2b = 2b(a - 1) = 0$ ，由于 $a = 1$ ，此条件自动满足

代入 $a = 1$ ，风险函数简化为：

$$R(\theta, \delta) = b^2 + \frac{1}{3}$$

步骤4：最小化最大风险

现在风险函数与 θ 无关，恒等于 $b^2 + \frac{1}{3}$ 。为了最小化这个常数风险，我们选择 $b = 0$ 。

因此，最优的决策规则是：

$$\delta(X) = 1 \cdot X + 0 = X$$

对应的风险为：

$$R(\theta, \delta) = \frac{1}{3}, \quad \forall \theta \in \mathbb{R}$$

结论：

$$\hat{\theta}_{\text{minimax}} = X$$

是极大极小估计量，其最大风险为 $\frac{1}{3}$ 。

讨论与解释

这个结果有很好的直观解释：

- 样本均值 X 是 θ 的无偏估计
- 在平方损失下，对于均匀分布 $U(\theta - 1, \theta + 1)$ ，样本均值恰好是极大极小估计量
- 最大风险 $\frac{1}{3}$ 实际上是 X 的方差，这是该问题中能达到的最小最大风险

极大极小估计在稳健统计、博弈论和工程应用中都有重要价值，它提供了一种在不确定性下做最坏打算的理性决策框架。

4.6.5 小结

方法	框架	目标	主要性质
MLE	频率学派	最大化样本似然	一致、渐近正态、渐近有效
LS / WLS	平方损失、 线性模型	最小化残差平方和（加权）	Gauss–Markov 最优性，正态时等于 MLE
极大极小估计	决策理论	降低最坏情境风险	稳健，部分可由贝叶斯导出

4.7 贝叶斯估计与均方误差准则

4.7.1 贝叶斯估计

1) 贝叶斯决策理论与风险最小化

贝叶斯估计的出发点源于统计决策理论，其核心思想是在不确定环境下做出**最优决策**，这里的“最优”是通过最小化期望损失（即风险）来定义的。

频率派风险与贝叶斯风险

在统计决策理论中，评估一个估计量 $\delta(X)$ 的好坏通常使用**风险函数**：

$$R(\theta, \delta) = \mathbb{E}_{X|\theta}[L(\theta, \delta(X))]$$

其中 $L(\theta, \hat{\theta})$ 是损失函数，度量参数真值为 θ 时采用估计 $\hat{\theta}$ 带来的损失。

然而，频率派的风险函数 $R(\theta, \delta)$ 依赖于未知参数 θ ，无法直接用于比较不同估计量的优劣。贝叶斯学派通过引入先验分布 $p(\theta)$ ，定义了**贝叶斯风险**：

$$r(\delta) = \mathbb{E}_{\theta}[\mathbb{E}_{X|\theta}[L(\theta, \delta(X))]] = \iint L(\theta, \delta(x))p(x|\theta)p(\theta)dx d\theta$$

贝叶斯估计量

贝叶斯估计量是使贝叶斯风险最小的估计量：

$$\delta_{\text{Bayes}} = \arg \min_{\delta} r(\delta)$$

根据Fubini定理，最小化贝叶斯风险等价于对每个观测值 x 最小化**后验期望损失**：

$$\delta_{\text{Bayes}}(x) = \arg \min_{\hat{\theta}} \mathbb{E}_{\theta|x}[L(\theta, \hat{\theta})] = \arg \min_{\hat{\theta}} \int L(\theta, \hat{\theta})p(\theta|x)d\theta$$

这个优美的结论告诉我们：**贝叶斯估计量可以通过逐点最小化后验期望损失来获得。**

贝叶斯估计的一般框架

- **确定先验分布**： $p(\theta)$ ，反映参数的不确定性
- **选择损失函数**： $L(\theta, \hat{\theta})$ ，反映估计误差的代价
- **计算后验分布**： $p(\theta|x) \propto p(x|\theta)p(\theta)$
- **最小化后验期望损失**： $\hat{\theta}_{\text{Bayes}} = \arg \min_{\hat{\theta}} \mathbb{E}[L(\theta, \hat{\theta})|x]$

2) 不同损失函数下的贝叶斯估计

损失函数的选择体现了我们对不同类型误差的相对重视程度，不同的损失函数会导致不同的贝叶斯估计量。

平方损失函数

定理：当损失函数为平方损失 $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ 时，贝叶斯估计量是后验均值：

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta|x]$$

证明：

后验期望损失为：

$$\mathbb{E}[(\theta - \hat{\theta})^2|x] = \mathbb{E}[\theta^2|x] - 2\hat{\theta}\mathbb{E}[\theta|x] + \hat{\theta}^2$$

对 $\hat{\theta}$ 求导：

$$\frac{d}{d\hat{\theta}} \mathbb{E}[(\theta - \hat{\theta})^2|x] = -2\mathbb{E}[\theta|x] + 2\hat{\theta}$$

令导数为零得： $\hat{\theta} = \mathbb{E}[\theta|x]$

平方损失对大的误差给予较重惩罚，导出的估计量具有光滑性、无偏性等良好性质。

绝对损失函数

定理：当损失函数为绝对损失 $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ 时，贝叶斯估计量是后验中位数。

证明思路：

后验期望损失可写为：

$$\mathbb{E}[|\theta - \hat{\theta}|x] = \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta)p(\theta|x)d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta})p(\theta|x)d\theta$$

利用变分法，最优解 $\hat{\theta}$ 满足：

$$\int_{-\infty}^{\hat{\theta}} p(\theta|x)d\theta = \int_{\hat{\theta}}^{\infty} p(\theta|x)d\theta$$

即 $\hat{\theta}$ 是后验分布的中位数。

绝对损失对异常值不敏感，导出的估计量更加稳健。

0-1损失函数

定理：当损失函数为0-1损失：

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \text{if } |\theta - \hat{\theta}| \leq \varepsilon \\ 1 & \text{if } |\theta - \hat{\theta}| > \varepsilon \end{cases}$$

且 $\varepsilon \rightarrow 0$ 时，贝叶斯估计量趋近于后验分布的众数（MAP估计）。

证明思路：

后验期望损失为：

$$\mathbb{E}[L(\theta, \hat{\theta})|x] = 1 - \int_{\hat{\theta}-\varepsilon}^{\hat{\theta}+\varepsilon} p(\theta|x)d\theta$$

当 $\varepsilon \rightarrow 0$ 时，最小化该期望等价于最大化 $p(\hat{\theta}|x)$ 。

0-1损失在假设检验和分类问题中有着天然的应用。

3) 贝叶斯估计的实例分析

例题4.22：正态模型的贝叶斯估计

设 $X|\theta \sim N(\theta, \sigma^2)$ ，先验分布 $\theta \sim N(\mu_0, \tau^2)$ ，求平方损失和绝对损失下的贝叶斯估计量。

解答：

步骤1: 计算后验分布

后验分布为:

$$\theta|x \sim N(\mu_1, \tau_1^2)$$

其中:

$$\mu_1 = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \quad \tau_1^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}$$

步骤2: 平方损失下的估计量

$$\hat{\theta}_{\text{sq}} = \mu_1$$

步骤3: 绝对损失下的估计量

由于正态分布的对称性, 中位数等于均值:

$$\hat{\theta}_{\text{abs}} = \mu_1$$

例题4.23 非对称损失函数

考虑线性指数损失函数 $L(\theta, \hat{\theta}) = e^{a(\hat{\theta}-\theta)} - a(\hat{\theta} - \theta) - 1$, 其中 $a \neq 0$ 。证明贝叶斯估计量为 $\hat{\theta} = -\frac{1}{a} \log \mathbb{E}[e^{-a\theta}|x]$ 。

解答:

步骤1: 写出后验期望损失

$$\mathbb{E}[L(\theta, \hat{\theta})|x] = \mathbb{E}[e^{a(\hat{\theta}-\theta)}|x] - a(\hat{\theta} - \mathbb{E}[\theta|x]) - 1$$

步骤2: 最小化期望损失

对 $\hat{\theta}$ 求导:

$$\frac{d}{d\hat{\theta}} \mathbb{E}[L(\theta, \hat{\theta})|x] = a \mathbb{E}[e^{a(\hat{\theta}-\theta)}|x] - a = 0$$

解得:

$$\mathbb{E}[e^{a(\hat{\theta}-\theta)}|x] = 1 \quad \Rightarrow \quad e^{a\hat{\theta}} \mathbb{E}[e^{-a\theta}|x] = 1$$

因此:

$$\hat{\theta} = -\frac{1}{a} \log \mathbb{E}[e^{-a\theta}|x]$$

讨论: 当 $a > 0$ 时, 高估的惩罚重于低估; 当 $a < 0$ 时, 低估的惩罚重于高估。

【贝叶斯估计的性质】

- **容许性:** 贝叶斯估计量通常是容许的
- **渐近性质:** 当样本量增大时, 先验的影响减弱, 贝叶斯估计与MLE渐近等价

- **完备性**：在适当条件下，贝叶斯估计量是完备的
- **稳健性**：依赖于先验分布的选择，需要谨慎考虑先验的设定

贝叶斯估计提供了一个统一的理论框架，将参数估计、损失函数和先验信息有机地结合在一起。通过选择不同的损失函数，我们可以针对具体问题定制最优的估计策略，这在工程、金融、医学等领域的决策问题中具有重要价值。

4.7.2 最大后验估计 (MAP, Maximum A Posteriori)

1) 贝叶斯框架与MAP的基本思想

在经典频率学派中，参数被视为固定的未知常数。而贝叶斯学派则采取不同的视角：**将参数本身视为随机变量**，具有某种概率分布。这种观点自然地引出了先验信息和后验分布的概念。

从先验到后验

贝叶斯推断的核心是贝叶斯公式：

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

其中：

- $p(\theta)$ 是**先验分布**，表示在观测数据前对参数 θ 的认知
- $p(\mathbf{x}|\theta)$ 是**似然函数**，与频率学派中的定义相同
- $p(\mathbf{x})$ 是**边缘似然**，作为归一化常数
- $p(\theta|\mathbf{x})$ 是**后验分布**，结合了先验信息和观测数据后的参数分布

MAP估计的直观理解

最大后验估计选择使后验概率密度最大的参数值：

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\theta|\mathbf{x}) = \arg \max_{\theta} p(\mathbf{x}|\theta)p(\theta)$$

由于 $p(\mathbf{x})$ 与 θ 无关，在优化中可以忽略。

从决策理论的角度看，MAP估计是在0-1损失函数下的贝叶斯估计量。它寻找的是后验分布中的“最可能”值，类似于分布的模式。

2) 正态分布情形下的MAP估计

例题4.24：正态-正态模型

假设观测数据满足 $x|\theta \sim N(\theta, \sigma^2)$ ，参数先验分布为 $\theta \sim N(\mu_0, \tau^2)$ ，其中 σ^2 、 μ_0 、 τ^2 已知。求参数 θ 的最大后验估计。

解答：

步骤1：写出后验分布的表达式

根据贝叶斯公式，后验分布与似然函数和先验分布的乘积成正比：

$$p(\theta|x) \propto p(x|\theta)p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\theta-\mu_0)^2}{2\tau^2}\right)$$

步骤2: 简化表达式

忽略与 θ 无关的常数项，后验分布满足：

$$p(\theta|x) \propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\tau^2}\right)$$

步骤3: 转化为优化问题

求MAP估计等价于最大化后验概率密度，这又等价于最小化负对数后验：

$$L(\theta) = -\log p(\theta|x) = \frac{(x-\theta)^2}{2\sigma^2} + \frac{(\theta-\mu_0)^2}{2\tau^2} + \text{常数}$$

忽略常数项，定义目标函数：

$$J(\theta) = \frac{(x-\theta)^2}{2\sigma^2} + \frac{(\theta-\mu_0)^2}{2\tau^2}$$

步骤4: 求解优化问题

对 $J(\theta)$ 关于 θ 求导并令导数为零：

$$\frac{dJ}{d\theta} = -\frac{x-\theta}{\sigma^2} + \frac{\theta-\mu_0}{\tau^2} = 0$$

整理得：

$$\frac{\theta}{\sigma^2} + \frac{\theta}{\tau^2} = \frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}$$

$$\theta \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2} \right) = \frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}$$

解得MAP估计量：

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

步骤5: 结果解释

这个结果有很好的直观解释：MAP估计是观测值 x 和先验均值 μ_0 的**加权平均**，权重与各自的精度（方差的倒数）成正比：

- 当观测噪声方差 σ^2 很小时，给予观测值 x 更大的权重
- 当先验方差 τ^2 很小时，表明先验信息很确定，给予先验均值 μ_0 更大的权重

3) MAP估计的性质与讨论

与MLE的关系

当先验分布为均匀分布时，MAP估计退化为极大似然估计：

$$p(\theta) \propto \text{常数} \Rightarrow \hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\mathbf{x}|\theta) = \hat{\theta}_{\text{MLE}}$$

这表明MLE可以看作是贝叶斯估计在"无信息先验"下的特例。

正则化视角

从优化角度看，MAP估计可以理解为在MLE的基础上增加了正则化项。对于负对数似然：

$$\ell(\theta) = -\log p(\mathbf{x}|\theta)$$

MAP估计等价于：

$$\hat{\theta}_{\text{MAP}} = \arg \min_{\theta} [\ell(\theta) - \log p(\theta)]$$

其中 $-\log p(\theta)$ 起到了正则化项的作用。例如：

- 高斯先验对应L2正则化（岭回归）
- 拉普拉斯先验对应L1正则化（Lasso）

优点与局限性

优点：

- **利用先验信息**：在样本量小或数据质量差时，能提供更稳定的估计
- **防止过拟合**：通过先验分布自然地引入正则化效果
- **概念直观**：易于理解和解释

局限性：

- **先验敏感性**：结果严重依赖于先验分布的选择
- **点估计缺陷**：只利用了后验分布的众数，忽略了其他信息
- **计算复杂性**：对于复杂模型，后验分布的优化可能很困难

应用场景

MAP估计在以下场景中特别有用：

- 小样本学习问题
- 需要加入领域知识的统计建模
- 作为复杂贝叶斯计算的近似
- 机器学习中的正则化方法

MAP估计架起了频率学派和贝叶斯学派的桥梁，既保留了MLE的直观性，又通过先验分布引入了贝叶斯思想的优势，在实践中具有广泛的应用价值。

4.7.3 最小均方误差估计 (MMSE, Minimum Mean Square Error)

1) MMSE估计的基本定义与最优性

最小均方误差估计是贝叶斯估计框架下最重要、最常用的估计方法之一。它基于平方损失函数，寻求在平均意义上误差平方最小的估计量。

问题表述

考虑参数估计问题，其中：

- θ 是待估参数（可以是标量或向量）
- \mathbf{x} 是观测数据
- 估计量 $\hat{\theta}(\mathbf{x})$ 是基于观测数据的任意函数

MMSE估计量定义为使均方误差最小的估计量：

$$\hat{\theta}_{\text{MMSE}} = \arg \min_{\hat{\theta}(\mathbf{x})} \mathbb{E} \left[(\hat{\theta}(\mathbf{x}) - \theta)^2 \right]$$

这里的期望是对联合分布 $p(\theta, \mathbf{x})$ 取的。

最优性证明

固定观测值 \mathbf{x} ，考虑条件均方误差：

$$\mathbb{E}[(\hat{\theta} - \theta)^2 | \mathbf{x}] = \mathbb{E}[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2 | \mathbf{x}] = \hat{\theta}^2 - 2\hat{\theta}\mathbb{E}[\theta | \mathbf{x}] + \mathbb{E}[\theta^2 | \mathbf{x}]$$

这是一个关于 $\hat{\theta}$ 的二次函数，通过求导可得最优解：

$$\frac{\partial}{\partial \hat{\theta}} \mathbb{E}[(\hat{\theta} - \theta)^2 | \mathbf{x}] = 2\hat{\theta} - 2\mathbb{E}[\theta | \mathbf{x}] = 0$$

解得：

$$\hat{\theta}_{\text{MMSE}} = \mathbb{E}[\theta | \mathbf{x}]$$

这一优美结果表明：**MMSE估计量就是后验分布的均值。**

2) MMSE估计的性质与特点

MMSE估计量具有一系列优良的数学性质，这些性质使其在实践中得到广泛应用。

基本性质

- **无偏性**：MMSE估计量在后验意义下是无偏的：

$$\mathbb{E}[\hat{\theta}_{\text{MMSE}} | \mathbf{x}] = \mathbb{E}[\theta | \mathbf{x}]$$

- **正交性原理**：估计误差与观测数据正交：

$$\mathbb{E}[(\theta - \hat{\theta}_{\text{MMSE}})g(\mathbf{x})] = 0, \quad \forall g(\cdot)$$

这一性质在自适应滤波和系统辨识中非常重要。

- **递推更新**：当获得新观测数据时，MMSE估计可以通过贝叶斯更新公式递推计算，无需重新处理所有历史数据。

与MAP估计的关系

MMSE估计与最大后验估计（MAP）有着重要区别：

- **MMSE** = 后验均值
- **MAP** = 后验众数

只有当后验分布对称且单峰时，两者才相等。一般来说：

- MMSE估计更注重整体分布形态，对异常值不敏感
- MAP估计只关注概率密度最大的点，计算通常更简单

均方误差矩阵

对于向量参数情形，定义均方误差矩阵：

$$R_e = \mathbb{E}[(\theta - \hat{\theta})(\theta - \hat{\theta})^\top]$$

对于MMSE估计量，该矩阵达到所有估计量中的最小值（在矩阵正定意义下）。

3) 线性最小均方误差估计（LMMSE）

当后验分布难以计算时，我们通常限制估计量为观测数据的线性函数，这就导出了线性最小均方误差估计。

线性估计量的形式

考虑线性估计量：

$$\hat{\theta} = A\mathbf{x} + \mathbf{b}$$

其中 A 是系数矩阵， \mathbf{b} 是偏置向量。LMMSE估计是使均方误差最小的线性估计量。

例题4.25：向量参数的线性估计

设 M 维被估计随机矢量 θ 的均值矢量和协方差矩阵分别为 M_0 与 R_0 ，观测方程为：

$$\mathbf{x} = H\theta + \mathbf{n}_x$$

已知：

$$\mathbb{E}[\mathbf{n}_x] = 0, \quad \mathbb{E}[\mathbf{n}_x \mathbf{n}_x^\top] = R_n, \quad \mathbb{E}[\theta \mathbf{n}_x^\top] = 0$$

求 θ 的线性最小均方误差估计矢量 $\hat{\theta}_{\text{LMMSE}}$ 和估计的均方误差矩阵 R_e 。

解答：

步骤1：计算观测统计量

观测矢量 \mathbf{x} 的均值：

$$M_x = \mathbb{E}[\mathbf{x}] = \mathbb{E}[H\theta + \mathbf{n}_x] = HM_0$$

观测矢量的协方差矩阵：

$$\begin{aligned} R_x &= \mathbb{E}[(\mathbf{x} - M_x)(\mathbf{x} - M_x)^\top] \\ &= \mathbb{E}[(H\theta + \mathbf{n}_x - HM_0)(H\theta + \mathbf{n}_x - HM_0)^\top] \\ &= H\mathbb{E}[(\theta - M_0)(\theta - M_0)^\top]H^\top + \mathbb{E}[\mathbf{n}_x\mathbf{n}_x^\top] \\ &= HR_0H^\top + R_n \end{aligned}$$

步骤2：计算互协方差矩阵

参数与观测的互协方差矩阵：

$$\begin{aligned} R_{\theta x} &= \mathbb{E}[(\theta - M_0)(\mathbf{x} - M_x)^\top] \\ &= \mathbb{E}[(\theta - M_0)(H\theta + \mathbf{n}_x - HM_0)^\top] \\ &= \mathbb{E}[(\theta - M_0)(\theta - M_0)^\top H^\top] + \mathbb{E}[(\theta - M_0)\mathbf{n}_x^\top] \\ &= R_0H^\top \end{aligned}$$

步骤3：推导LMMSE估计量

线性MMSE估计量的一般形式为：

$$\hat{\theta}_{\text{LMMSE}} = M_0 + R_{\theta x}R_x^{-1}(\mathbf{x} - M_x)$$

代入前面计算结果：

$$\hat{\theta}_{\text{LMMSE}} = M_0 + R_0H^\top(HR_0H^\top + R_n)^{-1}(\mathbf{x} - HM_0)$$

步骤4：计算均方误差矩阵

估计的均方误差矩阵为：

$$\begin{aligned} R_e &= \mathbb{E}[(\theta - \hat{\theta})(\theta - \hat{\theta})^\top] \\ &= R_0 - R_{\theta x}R_x^{-1}R_{x\theta} \\ &= R_0 - R_0H^\top(HR_0H^\top + R_n)^{-1}HR_0 \end{aligned}$$

其中 $R_{x\theta} = R_{\theta x}^\top = HR_0$ 。

因此：

$$R_e = R_0 - R_0H^\top(HR_0H^\top + R_n)^{-1}HR_0$$

【讨论与解释】

- **Kalman滤波的联系**：这个结果实际上是Kalman滤波的稳态形式
- **信噪比的影响**：当观测噪声 $R_n \rightarrow 0$ 时，估计趋近于真实值；当 $R_n \rightarrow \infty$ 时，估计趋近于先验均值 M_0
- **计算效率**：LMMSE只需要一阶和二阶统计量，避免了复杂的后验分布计算

MMSE估计理论为现代信号处理、通信系统和机器学习提供了坚实的理论基础，其在线性高斯模型下的解析解更是工程应用中的宝贵工具。

4.7.4 经验贝叶斯与收缩估计

1) 经验贝叶斯方法的基本框架

经验贝叶斯方法巧妙地融合了频率学派和贝叶斯学派的优点，通过数据驱动的方式确定先验分布中的超参数，在保持贝叶斯思想的同时减少了主观性。

【基本思想与模型结构】

在传统贝叶斯方法中，参数 θ 的先验分布完全由研究者指定。而在经验贝叶斯框架下，我们承认先验分布包含未知的超参数 α ：

$$\theta \sim p(\theta|\alpha)$$

经验贝叶斯的核心思想是：**利用观测数据 \mathbf{x} 来估计这些超参数 α** ，然后将估计值代入贝叶斯公式计算后验分布。

【参数估计方法】

常用的超参数估计方法包括：

- **极大边际似然法 (Type II MLE)：**
通过最大化边际似然函数来估计超参数：

$$\hat{\alpha} = \arg \max_{\alpha} p(\mathbf{x}|\alpha) = \arg \max_{\alpha} \int p(\mathbf{x}|\theta)p(\theta|\alpha)d\theta$$

- **矩估计法：**
使用样本矩来匹配先验分布的矩，从而估计超参数。

【正态模型的实例】

例题4.26：正态均值模型

考虑如下分层模型：

$$\theta_i \sim N(\mu, \tau^2), \quad x_i \sim N(\theta_i, \sigma^2), \quad i = 1, \dots, n$$

其中 σ^2 已知，但超参数 μ, τ^2 未知。

解答：

步骤1：估计超参数

利用观测数据 $\{x_i\}$ 估计 μ, τ^2 。例如，可以通过矩估计：

$$\hat{\mu} = \bar{x}, \quad \hat{\tau}^2 = \max \left(0, \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \sigma^2 \right)$$

步骤2：计算后验分布

将估计的超参数代入，得到每个 θ_i 的后验分布：

$$\theta_i | x_i \sim N \left(\frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2} x_i + \frac{\sigma^2}{\hat{\tau}^2 + \sigma^2} \hat{\mu}, \frac{\hat{\tau}^2 \sigma^2}{\hat{\tau}^2 + \sigma^2} \right)$$

【性质与适用场景】

经验贝叶斯方法具有以下特点：

- **减少主观性**：超参数由数据估计，避免了完全主观的先验选择
- **计算可行性**：相比完全贝叶斯方法，通常计算更简单
- **适用范围**：特别适合处理大量类似参数的估计问题，如分层模型、组数据分析

然而，这种方法没有量化超参数估计的不确定性，因此不是完全的贝叶斯方法。

2) James-Stein收缩估计

James-Stein估计是经验贝叶斯方法的一个著名特例，它在高维参数估计中展现了令人惊讶的性质：当维度足够高时，样本均值居然不是最优估计。

【问题背景与Stein悖论】

考虑高维正态均值估计问题：

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \sigma^2 I_p)$$

其中 $\mathbf{X} = (X_1, \dots, X_p)^\top$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$, σ^2 已知。

传统上，我们使用极大似然估计：

$$\hat{\boldsymbol{\mu}}_{\text{MLE}} = \mathbf{X}$$

这个估计量在很多标准下都是最优的。然而，Stein在1956年发现了一个令人震惊的事实：**当 $p \geq 3$ 时，存在比MLE更优的估计量。**

【James-Stein估计量】

James和Stein提出了如下收缩估计量：

$$\hat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{X}\|^2} \right) \mathbf{X}$$

这个估计量将样本均值向原点收缩，收缩的程度取决于样本的范数。

【与经验贝叶斯的关系】

步骤1：建立贝叶斯模型

假设参数服从先验分布：

$$\boldsymbol{\mu} \sim N(\mathbf{0}, \tau^2 I_p)$$

观测模型为：

$$\mathbf{X}|\boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}, \sigma^2 I_p)$$

其中 σ^2 已知， τ^2 是未知的超参数。

根据贝叶斯定理，后验分布为：

$$\boldsymbol{\mu}|\mathbf{X} \sim N_p(\hat{\boldsymbol{\mu}}_{\text{Bayes}}, \Sigma)$$

其中后验均值：

$$\hat{\boldsymbol{\mu}}_{\text{Bayes}} = \frac{\tau^2}{\tau^2 + \sigma^2} \mathbf{X}$$

后验协方差矩阵：

$$\Sigma = \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2} I_p$$

步骤2：边际分布与超参数估计

在经验贝叶斯框架下，我们需要从数据中估计超参数 τ^2 。考虑边际分布：

$$\mathbf{X} \sim N_p(\mathbf{0}, (\tau^2 + \sigma^2)I_p)$$

因为 $\mathbb{E}[\mathbf{X}] = \mathbf{0}$ ，且 $\text{Cov}(\mathbf{X}) = (\tau^2 + \sigma^2)I_p$ 。

定义边际方差：

$$\eta^2 = \tau^2 + \sigma^2$$

则 $\mathbf{X} \sim N_p(\mathbf{0}, \eta^2 I_p)$ 。

统计量 $S = \|\mathbf{X}\|^2$ 的分布为：

$$S \sim \eta^2 \chi_p^2$$

即 S 服从尺度参数为 η^2 的卡方分布。

步骤3：逆卡方分布的性质

若 $Y \sim \chi_p^2$ ，则 $1/Y$ 服从逆卡方分布，其期望为：

$$\mathbb{E}\left[\frac{1}{Y}\right] = \frac{1}{p-2}, \quad \text{当 } p > 2$$

因此，对于 $S \sim \eta^2 \chi_p^2$ ，有：

$$\mathbb{E}\left[\frac{1}{S}\right] = \frac{1}{\eta^2} \cdot \frac{1}{p-2} = \frac{1}{(p-2)(\tau^2 + \sigma^2)}$$

步骤4：构造无偏估计量

我们希望估计收缩因子：

$$\frac{\tau^2}{\tau^2 + \sigma^2} = 1 - \frac{\sigma^2}{\tau^2 + \sigma^2}$$

注意到：

$$\frac{1}{\tau^2 + \sigma^2} = \mathbb{E} \left[\frac{p-2}{S} \right]$$

因为：

$$\mathbb{E} \left[\frac{p-2}{S} \right] = (p-2) \cdot \mathbb{E} \left[\frac{1}{S} \right] = (p-2) \cdot \frac{1}{(p-2)(\tau^2 + \sigma^2)} = \frac{1}{\tau^2 + \sigma^2}$$

因此， $\frac{p-2}{S}$ 是 $\frac{1}{\tau^2 + \sigma^2}$ 的无偏估计量。

步骤5：得到经验贝叶斯估计

将无偏估计量代入收缩因子表达式：

$$\frac{\tau^2}{\tau^2 + \sigma^2} = 1 - \frac{\sigma^2}{\tau^2 + \sigma^2} \approx 1 - \sigma^2 \cdot \frac{p-2}{S}$$

即：

$$\frac{\tau^2}{\tau^2 + \sigma^2} \approx 1 - \frac{(p-2)\sigma^2}{\|\mathbf{X}\|^2}$$

因此，经验贝叶斯估计量为：

$$\hat{\boldsymbol{\mu}}_{\text{EB}} = \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{X}\|^2} \right) \mathbf{X}$$

这与James-Stein估计量的形式完全一致。

步骤6：风险函数的等价性（补充说明）

可以进一步证明，这种经验贝叶斯估计量的风险函数与James-Stein估计量的风险函数相同。

对于James-Stein估计量：

$$\hat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{X}\|^2} \right) \mathbf{X}$$

其风险函数为：

$$R(\hat{\boldsymbol{\mu}}_{\text{JS}}, \boldsymbol{\mu}) = p\sigma^2 - (p-2)^2\sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right]$$

而对于经验贝叶斯估计量，由于我们使用了无偏估计量 $\frac{p-2}{S}$ 来估计 $\frac{1}{\tau^2 + \sigma^2}$ ，可以证明其风险函数具有相同的形式。

3) 理论性质与实例分析

James-Stein估计的理论性质和实际表现都极具启发性，它挑战了传统统计学中的一些基本直觉。

【风险函数分析】

定理：当 $p \geq 3$ 时，James-Stein估计量的均方误差严格小于极大似然估计量的均方误差：

$$\mathbb{E} [\|\hat{\boldsymbol{\mu}}_{\text{JS}} - \boldsymbol{\mu}\|^2] < \mathbb{E} [\|\hat{\boldsymbol{\mu}}_{\text{MLE}} - \boldsymbol{\mu}\|^2] = p\sigma^2$$

证明

步骤1: 定义与符号

设观测数据 $\mathbf{X} = (X_1, \dots, X_p)^\top \sim N_p(\boldsymbol{\mu}, \sigma^2 I_p)$, 其中 σ^2 已知。

定义James-Stein估计量为:

$$\hat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{X}\|^2}\right) \mathbf{X}$$

极大似然估计量为:

$$\hat{\boldsymbol{\mu}}_{\text{MLE}} = \mathbf{X}$$

步骤2: 计算James-Stein估计量的均方误差

均方误差定义为:

$$R(\hat{\boldsymbol{\mu}}_{\text{JS}}, \boldsymbol{\mu}) = \mathbb{E} [\|\hat{\boldsymbol{\mu}}_{\text{JS}} - \boldsymbol{\mu}\|^2]$$

代入James-Stein估计量:

$$\hat{\boldsymbol{\mu}}_{\text{JS}} - \boldsymbol{\mu} = \mathbf{X} - \boldsymbol{\mu} - \frac{(p-2)\sigma^2}{\|\mathbf{X}\|^2} \mathbf{X}$$

因此:

$$\|\hat{\boldsymbol{\mu}}_{\text{JS}} - \boldsymbol{\mu}\|^2 = \|\mathbf{X} - \boldsymbol{\mu}\|^2 - 2(p-2)\sigma^2 \frac{(\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{X}}{\|\mathbf{X}\|^2} + (p-2)^2 \sigma^4 \frac{1}{\|\mathbf{X}\|^2}$$

取期望:

$$R(\hat{\boldsymbol{\mu}}_{\text{JS}}, \boldsymbol{\mu}) = \mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^2] - 2(p-2)\sigma^2 \mathbb{E} \left[\frac{(\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{X}}{\|\mathbf{X}\|^2} \right] + (p-2)^2 \sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right]$$

步骤3: 计算各项期望

第一项是MLE的均方误差:

$$\mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|^2] = p\sigma^2$$

第二项需要使用Stein引理。对于多元正态分布 $Y \sim N(\theta, \sigma^2 I)$, Stein引理指出:

$$\mathbb{E}[(Y_i - \theta_i)g(Y)] = \sigma^2 \mathbb{E} \left[\frac{\partial g}{\partial Y_i}(Y) \right]$$

对于几乎处处可导的函数 g 。

令 $g(\mathbf{X}) = \frac{X_i}{\|\mathbf{X}\|^2}$, 则:

$$\mathbb{E} \left[(X_i - \mu_i) \frac{X_i}{\|\mathbf{X}\|^2} \right] = \sigma^2 \mathbb{E} \left[\frac{\partial}{\partial X_i} \left(\frac{X_i}{\|\mathbf{X}\|^2} \right) \right]$$

计算偏导数:

$$\frac{\partial}{\partial X_i} \left(\frac{X_i}{\|\mathbf{X}\|^2} \right) = \frac{1}{\|\mathbf{X}\|^2} - \frac{2X_i^2}{\|\mathbf{X}\|^4}$$

因此:

$$\sum_{i=1}^p \mathbb{E} \left[(X_i - \mu_i) \frac{X_i}{\|\mathbf{X}\|^2} \right] = \sigma^2 \sum_{i=1}^p \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} - \frac{2X_i^2}{\|\mathbf{X}\|^4} \right] = \sigma^2 \mathbb{E} \left[\frac{p}{\|\mathbf{X}\|^2} - \frac{2\|\mathbf{X}\|^2}{\|\mathbf{X}\|^4} \right] = \sigma^2(p-2) \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right]$$

所以：

$$\mathbb{E} \left[\frac{(\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{X}}{\|\mathbf{X}\|^2} \right] = \sigma^2(p-2) \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right]$$

步骤4：代入得到最终表达式

将各项期望代入均方误差表达式：

$$\begin{aligned} R(\hat{\boldsymbol{\mu}}_{\text{JS}}, \boldsymbol{\mu}) &= p\sigma^2 - 2(p-2)\sigma^2 \cdot \sigma^2(p-2) \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] + (p-2)^2 \sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] \\ &= p\sigma^2 - 2(p-2)^2 \sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] + (p-2)^2 \sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] \\ &= p\sigma^2 - (p-2)^2 \sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] \end{aligned}$$

步骤5：证明严格不等式

由于 $\|\mathbf{X}\|^2 > 0$ 几乎必然成立，且当 $p \geq 3$ 时， $\mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] > 0$ ，因此：

$$(p-2)^2 \sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] > 0$$

所以：

$$R(\hat{\boldsymbol{\mu}}_{\text{JS}}, \boldsymbol{\mu}) = p\sigma^2 - (p-2)^2 \sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] < p\sigma^2 = R(\hat{\boldsymbol{\mu}}_{\text{MLE}}, \boldsymbol{\mu})$$

步骤6：风险比表达式

风险比为：

$$\frac{R(\hat{\boldsymbol{\mu}}_{\text{JS}}, \boldsymbol{\mu})}{R(\hat{\boldsymbol{\mu}}_{\text{MLE}}, \boldsymbol{\mu})} = 1 - \frac{(p-2)^2 \sigma^4}{p\sigma^2} \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right] = 1 - \frac{p-2}{p} \cdot (p-2) \sigma^2 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right]$$

这正是定理陈述中的形式。

【讨论与意义】

这个证明揭示了几个重要事实：

- **维度依赖**：改进项 $(p-2)^2 \sigma^4 \mathbb{E} \left[\frac{1}{\|\mathbf{X}\|^2} \right]$ 依赖于维度 p ，当 $p = 1, 2$ 时，改进项非正，James-Stein估计不优于 MLE。
- **收缩的本质**：James-Stein估计通过向原点收缩来减少方差，虽然引入了偏差，但总体均方误差减小。
- **Stein悖论**：这一结果被称为Stein悖论，因为它表明在高维情况下，看似自然的估计量（样本均值）实际上不是最优的。
- **实际意义**：当 p 较大且 $\|\boldsymbol{\mu}\|$ 较小时，改进效果最明显。当 $\boldsymbol{\mu} = \mathbf{0}$ 时，改进最大。

这个定理彻底改变了我们对高维统计估计的认识，催生了现代收缩估计和正则化方法的研究。

【实例计算】

例题4.27: James-Stein估计计算

已知 $p = 5$, $\sigma^2 = 1$, 观测值 $\mathbf{X} = (2, 1, -1, 0, 3)^\top$ 。分别计算MLE和James-Stein估计。

解答:

步骤1: 计算MLE估计

$$\hat{\boldsymbol{\mu}}_{\text{MLE}} = \mathbf{X} = (2, 1, -1, 0, 3)^\top$$

步骤2: 计算James-Stein估计

首先计算样本范数:

$$\|\mathbf{X}\|^2 = 2^2 + 1^2 + (-1)^2 + 0^2 + 3^2 = 4 + 1 + 1 + 0 + 9 = 15$$

收缩因子为:

$$1 - \frac{(5-2) \cdot 1}{15} = 1 - \frac{3}{15} = 0.8$$

因此James-Stein估计为:

$$\hat{\boldsymbol{\mu}}_{\text{JS}} = 0.8 \cdot (2, 1, -1, 0, 3) = (1.6, 0.8, -0.8, 0, 2.4)^\top$$

【风险比较分析】**例题4.28: 风险函数比较**

考虑前面提到的正态模型, 比较MLE和MAP估计的风险。

解答:

对于MLE估计 $\hat{\boldsymbol{\mu}}_{\text{MLE}} = \mathbf{Z}$, 其风险为:

$$R_{\text{MLE}}(\boldsymbol{\mu}) = \mathbb{E} [\|\mathbf{Z} - \boldsymbol{\mu}\|^2] = N$$

对于MAP估计 $\hat{\boldsymbol{\mu}}_{\text{MAP}} = \eta^2 \mathbf{Z}$, 其中 $\eta^2 = \frac{\tau^2}{\tau^2 + 1}$, 其风险为:

$$R_{\text{MAP}}(\boldsymbol{\mu}) = \mathbb{E} [\|\eta^2 \mathbf{Z} - \boldsymbol{\mu}\|^2] = \frac{N\tau^2}{1 + \tau^2}$$

风险比为:

$$\frac{R_{\text{MAP}}(\boldsymbol{\mu})}{R_{\text{MLE}}(\boldsymbol{\mu})} = \frac{\tau^2}{1 + \tau^2}$$

当 $\tau^2 = 1$ 时, MAP估计的风险只有MLE估计的一半。

【实用改进】

在实际应用中, 为了避免收缩因子为负值, 通常使用正部收缩:

$$\hat{\boldsymbol{\mu}}_{\text{JS}+} = \max \left(0, 1 - \frac{(p-2)\sigma^2}{\|\mathbf{X}\|^2} \right) \mathbf{X}$$

James-Stein估计的革命性意义在于它表明: 在高维问题中, 通过适当地"收缩"估计量, 即使是在平方损失下, 也能获得一致的风险改进。这一发现催生了现代高维统计学中的众多收缩方法和正则化技术。

4.7.5 总结

方法	定义	最优性条件	先验需求	特点
MAP	最大化后验pdf的点	0-1损失最优	需要先验	结合先验与似然
MMSE	后验均值	平方损失最优	需要先验	考虑分布全貌
LMMSE	线性后验均值	线性形式下最优	只需一二阶矩	高斯时与MMSE一致
Empirical Bayes	数据估计先验参数	同MAP/MMSE原则	估计的先验	混合方法
James–Stein	收缩估计	在 $p \geq 3$ 时严格优于MLE	固定形式收缩	降低维度灾难风险

4.8 不完全数据条件下的估计

4.8.1 EM算法

1) EM算法的基本框架与动机

EM（Expectation-Maximization）算法是处理含有隐变量（latent variables）或缺失数据的统计模型参数估计的强大工具。它在机器学习和统计学中有着广泛的应用，特别是在混合模型、隐马尔可夫模型等复杂概率模型中。

【问题背景】

在许多实际问题中，我们观测到的数据往往是不完整的，存在某些未观测到的隐变量。考虑以下情景：

- 观测数据： \mathbf{x} （可直接观测）
- 隐变量： Z （不可直接观测，但对模型至关重要）
- 模型参数： θ （需要估计）

完全数据似然函数为 $p(\mathbf{x}, Z|\theta)$ ，但实际我们只能计算观测数据似然：

$$p(\mathbf{x}|\theta) = \int p(\mathbf{x}, Z|\theta) dZ$$

直接最大化观测数据似然通常很困难，因为积分（或求和）操作使得优化问题变得复杂。EM算法通过迭代的方式巧妙地解决了这一问题。

【算法核心思想】

EM算法的核心思想是通过引入隐变量的条件分布，将复杂的优化问题分解为两个相对简单的步骤：

1. **E步（期望步）**：基于当前参数估计，计算隐变量的条件期望
2. **M步（最大化步）**：基于E步的结果，更新参数估计

这种交替优化的策略使得算法能够稳定地收敛到局部最优解。

2) EM算法的数学表述与收敛性证明

【算法步骤】

设当前参数估计为 $\theta^{(t)}$ ，EM算法通过以下两步迭代更新参数：

E步 (Expectation Step):

计算完全数据对数似然函数关于隐变量条件分布的期望：

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|\mathbf{x}, \theta^{(t)}} [\ln p(\mathbf{x}, Z|\theta)]$$

M步 (Maximization Step):

最大化Q函数，更新参数：

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta|\theta^{(t)})$$

【收敛性证明】

定理：EM算法保证观测数据似然函数在每次迭代中不递减，即：

$$\ln p(\mathbf{x}|\theta^{(t+1)}) \geq \ln p(\mathbf{x}|\theta^{(t)})$$

证明：

步骤1：构造下界函数

对于任意关于隐变量 Z 的概率分布 $q(Z)$ ，观测数据似然可以表示为：

$$\ln p(\mathbf{x}|\theta) = \ln \int p(\mathbf{x}, Z|\theta) dZ = \ln \int q(Z) \frac{p(\mathbf{x}, Z|\theta)}{q(Z)} dZ$$

利用Jensen不等式（对数函数是凹函数）：

$$\ln p(\mathbf{x}|\theta) \geq \int q(Z) \ln \frac{p(\mathbf{x}, Z|\theta)}{q(Z)} dZ$$

等号在 $q(Z) = p(Z|\mathbf{x}, \theta)$ 时成立。

定义下界函数：

$$F(q, \theta) = \int q(Z) \ln \frac{p(\mathbf{x}, Z|\theta)}{q(Z)} dZ$$

步骤2：EM步骤与下界函数的关系

在E步中，令 $q^{(t)}(Z) = p(Z|\mathbf{x}, \theta^{(t)})$ ，此时：

$$F(q^{(t)}, \theta) = \mathbb{E}_{q^{(t)}} [\ln p(\mathbf{x}, Z|\theta)] - \mathbb{E}_{q^{(t)}} [\ln q^{(t)}(Z)] = Q(\theta|\theta^{(t)}) + \text{常数}$$

在M步中，我们最大化 $Q(\theta|\theta^{(t)})$ ，这等价于最大化 $F(q^{(t)}, \theta)$ 。

步骤3：证明似然单调递增

由下界函数的性质：

$$\ln p(\mathbf{x}|\theta^{(t)}) = F(q^{(t)}, \theta^{(t)})$$

$$\ln p(\mathbf{x}|\theta^{(t+1)}) \geq F(q^{(t)}, \theta^{(t+1)})$$

由于M步保证了：

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) \Rightarrow F(q^{(t)}, \theta^{(t+1)}) \geq F(q^{(t)}, \theta^{(t)})$$

因此：

$$\ln p(\mathbf{x}|\theta^{(t+1)}) \geq \ln p(\mathbf{x}|\theta^{(t)})$$

证毕。

3) 高斯混合模型的EM算法实例

高斯混合模型（Gaussian Mixture Model, GMM）是EM算法最经典的应用场景之一，它通过多个高斯分布的线性组合来拟合复杂的数据分布。

【问题设定】

假设观测数据 $\mathbf{x} = \{x_1, \dots, x_N\}$ 来自 M 个高斯分布的混合：

$$p(x|\Phi) = \sum_{j=1}^M \alpha_j p_j(x|\phi_j)$$

其中：

- α_j 是第 j 个分量的混合权重，满足 $\sum_{j=1}^M \alpha_j = 1$
- $p_j(x|\phi_j)$ 是第 j 个高斯分量，参数 $\phi_j = (\mu_j, \sigma_j^2)$
- $\Phi = \{\alpha_1, \dots, \alpha_M, \phi_1, \dots, \phi_M\}$ 是所有待估参数

【EM算法步骤】

E步：计算后验概率（软分类）

对每个样本 x_i 和每个分量 j ，计算样本 x_i 属于第 j 个分量的后验概率：

$$\gamma_{ij} = p(j|x_i, \Phi^{(t)}) = \frac{\alpha_j^{(t)} p_j(x_i|\phi_j^{(t)})}{\sum_{k=1}^M \alpha_k^{(t)} p_k(x_i|\phi_k^{(t)})}$$

其中高斯密度函数为：

$$p_j(x_i|\phi_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right)$$

M步：更新参数

基于E步计算的后验概率，更新各分量的参数：

- 混合权重更新：

$$\alpha_j^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \gamma_{ij}$$

• 均值更新：

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^N \gamma_{ij} x_i}{\sum_{i=1}^N \gamma_{ij}}$$

• 方差更新：

$$(\sigma_j^2)^{(t+1)} = \frac{\sum_{i=1}^N \gamma_{ij} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^N \gamma_{ij}}$$

【数值实例】

例题4.29：具体计算示例

考虑6个观测数据点：

$$\mathbf{x} = [1.0, 1.2, 0.8, 3.0, 3.2, 2.8]$$

假设数据来自2个高斯分布的混合 ($M = 2$)，初始参数设置为：

- $\alpha_1^{(0)} = 0.5, \mu_1^{(0)} = 1.0, \sigma_1^{2(0)} = 0.1$
- $\alpha_2^{(0)} = 0.5, \mu_2^{(0)} = 3.0, \sigma_2^{2(0)} = 0.1$

解答：

第一次迭代：E步

以第一个样本 $x_1 = 1.0$ 为例：

计算第一个分量的概率密度：

$$p_1(1.0|\phi_1^{(0)}) = \frac{1}{\sqrt{2\pi \times 0.1}} \exp\left(-\frac{(1.0 - 1.0)^2}{2 \times 0.1}\right) = \frac{1}{\sqrt{0.628}} \approx 1.261$$

计算第二个分量的概率密度：

$$p_2(1.0|\phi_2^{(0)}) = \frac{1}{\sqrt{2\pi \times 0.1}} \exp\left(-\frac{(1.0 - 3.0)^2}{2 \times 0.1}\right) = \frac{1}{\sqrt{0.628}} \exp(-20) \approx 1.261 \times 2.06 \times 10^{-9}$$

归一化得到后验概率：

$$\gamma_{1,1} = \frac{0.5 \times 1.261}{0.5 \times 1.261 + 0.5 \times (1.261 \times 2.06 \times 10^{-9})} \approx 1.0$$

$$\gamma_{1,2} = 1 - \gamma_{1,1} \approx 0.0$$

类似地计算所有样本的后验概率，会发现前三个样本主要属于第一个分量，后三个样本主要属于第二个分量。

第一次迭代：M步

更新第一个分量的参数：

$$\alpha_1^{(1)} = \frac{1}{6}(\gamma_{1,1} + \gamma_{2,1} + \gamma_{3,1} + \gamma_{4,1} + \gamma_{5,1} + \gamma_{6,1}) \approx \frac{1}{6}(1 + 1 + 1 + 0 + 0 + 0) = 0.5$$

$$\mu_1^{(1)} = \frac{1.0 \times 1.0 + 1.2 \times 1.0 + 0.8 \times 1.0 + 3.0 \times 0 + 3.2 \times 0 + 2.8 \times 0}{1.0 + 1.0 + 1.0 + 0 + 0 + 0} = 1.0$$

$$\sigma_1^{2(1)} = \frac{(1.0 - 1.0)^2 \times 1.0 + (1.2 - 1.0)^2 \times 1.0 + (0.8 - 1.0)^2 \times 1.0}{3.0} = 0.0267$$

类似地更新第二个分量的参数。

【算法总结与物理意义】

EM算法在高斯混合模型中的应用具有清晰的物理解释：

- **E步**：进行"软分类"，计算每个样本属于各个分量的概率，反映了我们对隐变量（样本的组分归属）的不确定性
- **M步**：基于软分类结果，用加权平均的方式重新估计各分量的参数，权重反映了样本属于该分量的置信度

通过迭代执行这两个步骤，算法能够自动发现数据中的潜在结构，将数据点"分配"到不同的高斯分量中，同时估计每个分量的参数。这种软分类的方式比硬聚类（如K-means）更加灵活，能够处理重叠的类别和不确定性。

EM算法的强大之处在于其通用性框架，不仅适用于高斯混合模型，还可以扩展到各种含有隐变量的概率模型，为复杂统计建模提供了统一的解决方案。

4.8.2 广义EM（GEM）

1) 广义EM的基本框架与动机

标准EM算法要求在每个迭代步骤中精确求解E步和M步的优化问题，但在许多复杂模型中，这种精确求解可能计算代价高昂甚至不可行。广义EM（Generalized EM）算法通过放宽这一要求，提供了更灵活、更实用的优化框架。

【从标准EM到广义EM】

考虑包含隐变量 Y 和观测数据 X 的概率模型，参数为 θ 。标准EM算法的核心是交替执行：

- E步：精确计算 $q(Y) = p(Y|X, \theta^{(t)})$
- M步：精确求解 $\theta^{(t+1)} = \arg \max_{\theta} \mathbb{E}_q[\log p(X, Y|\theta)]$

然而，在实际应用中，这两个步骤可能面临挑战：

1. E步中，后验分布 $p(Y|X, \theta)$ 可能难以计算（如涉及高维积分）
2. M步中，Q函数的全局最大值可能难以找到

广义EM通过允许近似解来应对这些挑战，只要求每次迭代能提升目标函数的下界。

2) 自由能下界与变分推断视角

广义EM的理论基础建立在自由能下界（Evidence Lower Bound, ELBO）的概念上，这为算法提供了坚实的数学保证。

【自由能下界的推导】

观测数据的对数似然可以表示为：

$$L(\theta|X) = \log p(X|\theta) = \log \sum_Y p(X, Y|\theta)$$

对任意关于隐变量的分布 $q(Y)$ ，利用Jensen不等式：

$$\begin{aligned} L(\theta|X) &= \log \sum_Y q(Y) \frac{p(X, Y|\theta)}{q(Y)} \\ &\geq \sum_Y q(Y) \log \frac{p(X, Y|\theta)}{q(Y)} \equiv \mathcal{F}(q, \theta) \end{aligned}$$

定义自由能下界（ELBO）：

$$\mathcal{F}(q, \theta) = \sum_Y q(Y) \log p(X, Y|\theta) + H(q)$$

其中 $H(q) = -\sum_Y q(Y) \log q(Y)$ 是分布 $q(Y)$ 的熵。

【与KL散度的关系】

自由能下界与对数似然之间存在紧密联系：

$$\mathcal{F}(q, \theta) = \log p(X|\theta) - \text{KL}(q(Y)||p(Y|X, \theta))$$

其中 $\text{KL}(q||p)$ 是 $q(Y)$ 与真实后验 $p(Y|X, \theta)$ 之间的Kullback-Leibler散度。

这一关系揭示了EM算法的本质：

- 当 $q(Y) = p(Y|X, \theta)$ 时，KL散度为0，下界等于对数似然
- E步实际上是让 $q(Y)$ 尽量靠近真实后验，从而收紧下界
- M步则是在当前 $q(Y)$ 下最大化参数，提升下界

【广义EM算法步骤】

基于自由能下界，广义EM算法的一般步骤为：

E步：在固定参数 $\theta^{(k-1)}$ 的条件下，优化隐变量分布：

$$q^{(k)} = \arg \max_q \mathcal{F}(q, \theta^{(k-1)})$$

M步：在固定分布 $q^{(k)}$ 的条件下，优化模型参数：

$$\theta^{(k)} = \arg \max_{\theta} \mathcal{F}(q^{(k)}, \theta)$$

广义EM只要求每个步骤能提升下界 $\mathcal{F}(q, \theta)$ ，而不要求达到全局最优。

3) 广义EM的变体与实现方法

广义EM框架包含了多种具体的实现方法，每种方法针对不同的问题特点提供了相应的解决方案。理解这些变体的数学原理和实现细节对于在实际问题中选择合适的方法至关重要。

【变分EM（Variational EM）】

变分EM是处理复杂后验分布的经典方法，其核心思想是用一个简单的变分分布来近似真实后验。

数学原理：

在变分EM中，我们限制 $q(Y)$ 属于某个简单的分布族 \mathcal{Q} （如均值场族），然后在该族内寻找最优近似：

$$q^{(k)} = \arg \max_{q \in \mathcal{Q}} \mathcal{F}(q, \theta^{(k-1)})$$

实现细节：

1. **分布族选择**：根据问题特点选择合适的变分分布族
2. **坐标上升优化**：交替更新变分分布的各个参数
3. **收敛判断**：基于ELBO的变化量或参数变化量

例题4.30：隐狄利克雷分配（LDA）模型的变分推断

解答：

在LDA模型中，文档 d 的联合分布为：

$$p(\theta_d, z_d, w_d | \alpha, \beta) = p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta)$$

引入均值场变分分布：

$$q(\theta_d, z_d | \gamma_d, \phi_d) = q(\theta_d | \gamma_d) \prod_{n=1}^{N_d} q(z_{dn} | \phi_{dn})$$

变分E步的详细推导：

对于主题指派变量 z_{dn} ，变分参数更新为：

$$\phi_{dni} \propto \beta_{iw_{dn}} \exp(\mathbb{E}_q[\log \theta_{di}]) = \beta_{iw_{dn}} \exp\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^K \gamma_{dj}\right)\right)$$

对于文档-主题分布 θ_d ，变分参数更新为：

$$\gamma_{di} = \alpha_i + \sum_{n=1}^{N_d} \phi_{dni}$$

变分M步的推导：

主题-词分布参数的更新：

$$\beta_{ij} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \phi_{dni} \mathbb{I}[w_{dn} = j]$$

Dirichlet先验参数的估计（如果也需要学习）：

通过固定点迭代法求解：

$$\Psi(\alpha_i) - \Psi\left(\sum_{j=1}^K \alpha_j\right) = \frac{1}{D} \sum_{d=1}^D \left[\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^K \gamma_{dj}\right) \right]$$

【蒙特卡洛EM（MCEM）】

当E步的期望难以解析计算时，MCEM使用蒙特卡洛采样进行近似。

数学原理：

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{p(Y|X, \theta^{(t)})} [\log p(X, Y|\theta)] \approx \frac{1}{S} \sum_{s=1}^S \log p(X, Y^{(s)}|\theta)$$

其中 $Y^{(s)} \sim p(Y|X, \theta^{(t)})$

实现细节：

1. **采样方法选择：**根据后验分布特点选择Gibbs采样、Metropolis-Hastings等
2. **样本数量调整：**可随迭代次数增加样本数以提高精度
3. **方差控制：**使用对偶变量、控制变量等技术减少估计方差

例题4.31： 贝叶斯逻辑回归的MCEM实现

解答：

考虑贝叶斯逻辑回归模型：

$$p(y_i|x_i, w) = \text{Bernoulli}(\sigma(w^\top x_i)), \quad w \sim \mathcal{N}(0, \lambda^{-1}I)$$

MCEM E步：

从后验 $p(w|X, y, \theta^{(t)})$ 采样，使用Metropolis-Hastings算法：

1. 提议分布： $q(w'|w) = \mathcal{N}(w, \Sigma)$
2. 接受概率： $\min\left(1, \frac{p(w'|X, y)}{p(w|X, y)}\right)$
3. 采集 S 个样本 $w^{(1)}, \dots, w^{(S)}$

MCEM M步：

更新正则化参数：

$$\lambda^{(t+1)} = \frac{D}{\frac{1}{S} \sum_{s=1}^S \|w^{(s)}\|^2}$$

其中 D 是权重向量的维度。

【随机EM (Stochastic EM)】

针对大规模数据集的优化版本，每次迭代只使用数据的一个子集。

数学原理：

设数据集划分为 B 个批次，第 t 次迭代使用批次 b_t ：

$$Q_b(\theta|\theta^{(t)}) = \frac{N}{|b_t|} \sum_{i \in b_t} \mathbb{E}_{p(Y_i|X_i, \theta^{(t)})} [\log p(X_i, Y_i|\theta)]$$

实现细节：

1. **学习率调度：**使用递减的学习率保证收敛
2. **方差减少：**结合SVRG、SAGA等方差减少技术
3. **动量加速：**引入动量项加速收敛

例题4.32：大规模高斯混合模型的随机EM**解答：****随机E步：**对mini-batch b_t 中的每个样本 x_i ：

$$\gamma_{ij} = \frac{\alpha_j \mathcal{N}(x_i | \mu_j, \Sigma_j)}{\sum_{k=1}^K \alpha_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}$$

随机M步：

使用指数移动平均更新参数：

$$\begin{aligned} N_j^{(t+1)} &= (1 - \eta_t) N_j^{(t)} + \eta_t \frac{N}{|b_t|} \sum_{i \in b_t} \gamma_{ij} \\ \mu_j^{(t+1)} &= (1 - \eta_t) \mu_j^{(t)} + \eta_t \frac{\sum_{i \in b_t} \gamma_{ij} x_i}{\sum_{i \in b_t} \gamma_{ij}} \\ \Sigma_j^{(t+1)} &= \text{类似更新} \end{aligned}$$

其中 η_t 是递减的学习率。**【在线EM (Online EM)】**

适用于流式数据或数据不能全部存储的情况。

数学原理：

维护充分统计量的运行估计：

$$s^{(t)} = (1 - \eta_t) s^{(t-1)} + \eta_t \hat{s}(x_t)$$

其中 $\hat{s}(x_t)$ 是基于新样本 x_t 的充分统计量。**实现细节：**

1. **学习率选择**： $\eta_t = t^{-\kappa}$ ，通常 $\kappa \in (0.5, 1]$
2. **遗忘机制**：对于非平稳数据，使用固定学习率
3. **并行化**：可结合异步更新提高效率

例题4.33：在线主题建模**解答：**

对于流式文档数据，在线LDA算法：

对于每个新文档 d ：

1. **局部E步**：用变分推断计算文档级统计量

$$\hat{s}_{ij} = \sum_{n=1}^{N_d} \phi_{dni} \mathbb{I}[w_{dn} = j]$$

2. **全局M步**：更新主题-词分布

$$\beta_{ij}^{(t+1)} = (1 - \eta_t)\beta_{ij}^{(t)} + \eta_t \hat{s}_{ij}$$

【收敛性分析与调参建议】

收敛性保证的条件：

1. 参数空间紧致
2. 目标函数在紧集上连续
3. 随机梯度无偏且方差有界
4. 学习率满足Robbins-Monro条件： $\sum \eta_t = \infty, \sum \eta_t^2 < \infty$

调参经验：

- **变分EM**：注意变分分布的初始化，避免陷入局部最优
- **MCEM**：随迭代增加样本数，前期重速度，后期重精度
- **随机EM**：学习率调度很关键，建议 $\eta_t = O(1/t)$
- **在线EM**：对于平稳数据用递减学习率，非平稳数据用固定小学习率

诊断工具：

- 监控ELBO或似然函数的变化
- 检查参数变化的范数
- 可视化隐变量的后验分布

广义EM的各种变体为不同场景下的概率推断提供了灵活的解决方案。在实际应用中，往往需要结合问题的具体特点和计算资源约束，选择合适的算法变体并进行适当的调整。

4.8.3 Wake-Sleep算法与广义EM

1) 深度生成模型与变分推断挑战

Wake-Sleep算法是深度学习早期发展中的重要里程碑，由Geoffrey Hinton等人于1995年提出，专门针对具有多层隐变量的深度生成模型设计。该算法解决了传统EM算法在深度神经网络中难以应用的瓶颈问题。

【深度生成模型的挑战】

考虑深度信念网络（Deep Belief Networks, DBNs）或深度玻尔兹曼机（Deep Boltzmann Machines, DBMs），其联合分布为：

$$p(v, h^{(1)}, h^{(2)}, \dots, h^{(L)} | \theta) = p(v | h^{(1)}) \prod_{l=1}^{L-1} p(h^{(l)} | h^{(l+1)}) p(h^{(L)})$$

面临的挑战包括：

1. **后验推断困难**：真实后验 $p(h|v, \theta)$ 难以计算，涉及高维积分
2. **参数学习复杂**：梯度计算需要期望，但期望难以解析求解
3. **近似推断需求**：需要有效的近似方法来处理深度结构

【历史背景与意义】

Wake-Sleep算法出现在深度学习复兴之前，为后来的变分自编码器（VAE）等重要方法奠定了基础。其核心创新在于：

- 引入了两个独立的网络：生成网络和推断网络

- 采用交替优化的策略，避免直接计算难处理的后验
- 为深度无监督学习提供了可行的训练框架

2) Wake-Sleep算法的双向优化框架

Wake-Sleep算法通过两个方向相反的KL散度最小化，实现了生成模型和推断网络的协同训练。

【模型设定与符号】

- **生成模型** (Generative Model): $p(v, h|\theta)$, 参数为 θ
- **推断网络** (Recognition Network): $q(h|v, \phi)$, 参数为 ϕ
- **观测数据**: $v \sim p_{\text{data}}(v)$

【变分下界推导】

观测数据的对数似然可以分解为：

$$\log p(v|\theta) = \mathbb{E}_{q(h|v, \phi)}[\log p(v, h|\theta)] + \text{KL}(q(h|v, \phi) \| p(h|v, \theta))$$

由于KL散度非负，我们得到变分下界 (ELBO)：

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q(h|v, \phi)}[\log p(v, h|\theta)] - \mathbb{E}_{q(h|v, \phi)}[\log q(h|v, \phi)]$$

【双向KL散度最小化】

Wake-Sleep算法的核心在于同时考虑两个方向的KL散度：

1. **Wake阶段目标** (推断网络指导生成模型)：

$$\mathcal{J}_{\text{wake}}(\theta) = \text{KL}(q(h|v, \phi) \| p(h|v, \theta))$$

2. **Sleep阶段目标** (生成模型指导推断网络)：

$$\mathcal{J}_{\text{sleep}}(\phi) = \text{KL}(p(h|v, \theta) \| q(h|v, \phi))$$

这两个目标分别对应不同的近似策略，形成了算法的双向优化结构。

3) Wake-Sleep算法步骤与物理意义

【算法详细步骤】

Wake阶段 (用真实数据训练生成模型)：

1. 从训练集中采样观测数据 $v \sim p_{\text{data}}(v)$
2. 用推断网络生成隐变量样本: $h \sim q(h|v, \phi)$
3. 更新生成模型参数以最大化联合概率：

$$\theta \leftarrow \theta + \eta \nabla_{\theta} \log p(v, h|\theta)$$

Sleep阶段 (用生成模型训练推断网络)：

1. 从生成模型采样: $h \sim p(h|\theta)$, 然后 $v \sim p(v|h, \theta)$
2. 更新推断网络参数以最小化反向KL：

$$\phi \leftarrow \phi - \eta \nabla_{\phi} \text{KL}(p(h|v, \theta) \| q(h|v, \phi))$$

【物理意义解释】

Wake阶段的直观解释：

- "清醒"时，我们观察真实世界数据 v
- 推断网络"猜测"这些数据对应的隐变量 h
- 生成模型学习让这些"猜测"变得更合理

Sleep阶段的直观解释：

- "睡眠"时，生成模型"做梦"产生数据 v 和对应的隐变量 h
- 推断网络学习如何从"梦境"数据中推断隐变量
- 这相当于用合成数据训练推断网络

例题4.34：受限玻尔兹曼机的Wake-Sleep算法

考虑一个受限玻尔兹曼机（RBM），可见层为 v ，隐层为 h ，能量函数为：

$$E(v, h) = -v^{\top} W h - b^{\top} v - c^{\top} h$$

推导Wake-Sleep算法在RBM中的具体形式。

解答：

Wake阶段：

1. 从数据分布采样 $v \sim p_{\text{data}}(v)$
2. 计算隐层的后验： $q(h_j = 1|v) = \sigma(c_j + \sum_i W_{ij} v_i)$
3. 采样 $h \sim q(h|v)$ 或使用均值
4. 更新生成模型参数：

$$\begin{aligned} W &\leftarrow W + \eta(vh^{\top} - \mathbb{E}_{p(v, h)}[vh^{\top}]) \\ b &\leftarrow b + \eta(v - \mathbb{E}_{p(v)}[v]) \\ c &\leftarrow c + \eta(h - \mathbb{E}_{p(h)}[h]) \end{aligned}$$

Sleep阶段：

1. 从生成模型采样： $h \sim p(h)$ ，然后 $v \sim p(v|h)$
2. 计算推断网络的输出： $q(h|v) = \sigma(c + W^{\top} v)$
3. 更新推断网络参数以匹配生成模型的后验：

$$\begin{aligned} W &\leftarrow W + \eta(\mathbb{E}_{p(h|v)}[hv^{\top}] - q(h|v)v^{\top}) \\ c &\leftarrow c + \eta(\mathbb{E}_{p(h|v)}[h] - q(h|v)) \end{aligned}$$

4) Wake-Sleep与广义EM的一致性

虽然Wake-Sleep算法看起来与标准EM不同，但它实际上可以理解为广义EM框架的一种特殊实现。

【数学形式的一致性】

考虑广义EM的目标函数：

$$\mathcal{F}(q, \theta) = \mathbb{E}_q[\log p(v, h|\theta)] - \mathbb{E}_q[\log q(h)]$$

在Wake-Sleep中：

- **Wake阶段**对应**M步**：固定 q ，优化 θ
- **Sleep阶段**对应**E步的变体**：固定 θ ，优化 q

【KL散度方向的差异】

关键区别在于KL散度的方向：

- **标准EM的E步**：最小化 $KL(q||p)$
- **Wake-Sleep的Sleep步**：最小化 $KL(p||q)$

这种差异导致了不同的性质：

特性	$KL(q p)$ (EM)	$KL(p q)$ (Wake-Sleep)
模式寻求	是（覆盖所有模式）	否（可能只覆盖主要模式）
均值寻求	否	是
计算难度	通常更难	有时更简单
方差	较小	可能较大

【理论保证与局限性】

收敛性分析：

Wake-Sleep算法不能保证似然函数的单调增加，因为Sleep步最小化的是反向KL散度。然而，在实践中，它通常能够收敛到一个有用的解。

优势：

1. **可扩展性**：适用于深度网络和复杂模型
2. **并行性**：Wake和Sleep阶段可以并行执行
3. **灵活性**：推断网络可以是任意的参数化函数

局限性：

1. **理论保证弱**：缺乏严格的收敛性证明
2. **训练不稳定**：两个阶段可能相互干扰
3. **偏差问题**：反向KL可能导致模式坍塌

【现代变体与发展】

Wake-Sleep算法启发了许多现代方法：

1. **变分自编码器 (VAE)**：用重参数化技巧解决梯度估计问题
2. **对抗生成网络 (GAN)**：用判别器替代Sleep阶段
3. **重要性加权自编码器**：通过重要性采样改进变分下界

例题4.35： Wake-Sleep在深度信念网络中的应用

考虑一个两层深度信念网络，推导Wake-Sleep算法的具体实现。

解答：

设网络结构为 $v \rightarrow h^{(1)} \rightarrow h^{(2)}$ ，联合分布：

$$p(v, h^{(1)}, h^{(2)}) = p(v|h^{(1)})p(h^{(1)}|h^{(2)})p(h^{(2)})$$

Wake阶段：

1. 从下向上传播： $v \rightarrow q(h^{(1)}|v) \rightarrow q(h^{(2)}|h^{(1)})$
2. 从上向下生成：用生成模型参数更新各层权重
3. 目标：最大化 $\log p(v, h^{(1)}, h^{(2)})$

Sleep阶段：

1. 从顶层采样： $h^{(2)} \sim p(h^{(2)})$
2. 从上向下生成： $h^{(1)} \sim p(h^{(1)}|h^{(2)}), v \sim p(v|h^{(1)})$
3. 训练推断网络：最小化 $KL(p(h^{(1)}, h^{(2)}|v) || q(h^{(1)}, h^{(2)}|v))$

Wake-Sleep算法作为连接传统EM算法与现代深度生成模型的重要桥梁，不仅具有历史意义，其核心思想至今仍在变分推断和生成建模中发挥着重要作用。理解这一算法有助于我们深入把握深度无监督学习的理论基础和发展脉络。

4.9 鲁棒估计方法

4.9.1 鲁棒估计的基本概念与意义

在实际统计分析和机器学习应用中，数据质量往往受到多种因素的干扰，包括测量误差、数据录入错误、设备故障等。这些因素导致数据中常常包含异常值或离群点，传统估计方法对此类干扰极为敏感。

【经典方法的脆弱性】

以均值估计为例，考虑数据集 $[1, 2, 3, 4, 100]$ ，其均值为：

$$\bar{x} = \frac{1 + 2 + 3 + 4 + 100}{5} = 22$$

而中位数为 3。单一极端值 100 使得均值严重偏离数据的中心趋势，这体现了传统最小二乘方法和均值估计的脆弱性。

【鲁棒估计的核心目标】

鲁棒估计方法致力于在以下方面取得平衡：

1. **抗干扰性**：在存在少量异常值时仍能保持估计准确性
2. **效率保持**：在无异常数据时性能接近最优估计量
3. **异常识别**：提供异常值检测和诊断能力
4. **理论保证**：具有明确的断点分析和影响函数描述

4.9.2 鲁棒性的理论基础与评价指标

【影响函数分析】

影响函数 (Influence Function) 定量描述单个观测点对估计量的影响程度。对于估计量 T 在分布 F 处的影响函数定义为：

$$IF(x; T, F) = \lim_{t \rightarrow 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t}$$

其中 δ_x 是在点 x 的退化分布。

例题4.36：均值和中位数的影响函数

推导样本均值和中位数的影响函数。

解答：

对于样本均值 $T(F) = \int x dF(x)$ ，其影响函数为：

$$IF(x; T, F) = x - T(F)$$

这表明均值的影响函数无界，单个极端观测可对估计产生任意大影响。

对于中位数 $T(F) = F^{-1}(1/2)$ ，在对称分布 F 下，其影响函数有界：

$$IF(x; T, F) = \frac{\text{sign}(x - T(F))}{2f(T(F))}$$

其中 f 是 F 的密度函数。这表明中位数的影响有界，具有天然鲁棒性。

【断点分析】

断点 (Breakdown Point) 衡量估计量能够容忍的异常值最大比例。形式上， ε^* 断点定义为：

$$\varepsilon^* = \min \left\{ \frac{m}{n} : \sup_{Y_m} |T(X_n \cup Y_m)| = \infty \right\}$$

其中 Y_m 是任意 m 个异常值。

重要结论：

- 均值的断点为 0
- 中位数的断点为 0.5
- 截尾均值的断点等于截尾比例

4.9.3 经典鲁棒估计方法

【M估计理论框架】

M估计通过选择合适的损失函数 ρ 来抑制异常值影响：

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^n \rho(x_i, \theta)$$

对应的估计方程为：

$$\sum_{i=1}^n \psi(x_i, \hat{\theta}) = 0$$

其中 $\psi(x, \theta) = \frac{\partial}{\partial \theta} \rho(x, \theta)$ 。

重要M估计量：

1. **Huber估计：**

$$\rho_k(x) = \begin{cases} \frac{1}{2}x^2 & \text{如果 } |x| \leq k \\ k|x| - \frac{1}{2}k^2 & \text{如果 } |x| > k \end{cases}$$

结合了均值的效率和中位数的鲁棒性。

2. **Tukey双权重估计：**

$$\rho_c(x) = \begin{cases} \frac{c^2}{6} \left[1 - \left(1 - \left(\frac{x}{c} \right)^2 \right)^3 \right] & \text{如果 } |x| \leq c \\ \frac{c^2}{6} & \text{如果 } |x| > c \end{cases}$$

对极端异常值具有完全抑制能力。

例题4.37： Huber回归的迭代求解

给定线性模型 $y_i = x_i^\top \beta + \varepsilon_i$ ，推导Huber回归的迭代重加权最小二乘算法。

解答：

Huber回归的目标函数为：

$$\min_{\beta} \sum_{i=1}^n \rho_k(y_i - x_i^\top \beta)$$

通过引入权重函数：

$$w_i = \begin{cases} 1 & \text{如果 } |r_i| \leq k \\ \frac{k}{|r_i|} & \text{如果 } |r_i| > k \end{cases}$$

其中 $r_i = y_i - x_i^\top \beta$ 。

迭代算法步骤：

1. 初始化 $\beta^{(0)}$ (如OLS估计)
2. 计算残差 $r_i^{(t)} = y_i - x_i^\top \beta^{(t)}$
3. 计算权重 $w_i^{(t)}$
4. 求解加权最小二乘问题：

$$\beta^{(t+1)} = (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} y$$

5. 重复步骤2-4直至收敛

【L估计与R估计】

L估计基于顺序统计量的线性组合，典型代表包括：

- α -截尾均值：去除极端比例数据后的均值
- Winsorized均值：将极端值替换为边界值后的均值

R估计基于秩统计量，对分布假设要求较弱，特别适合非参数场景。

4.9.4 现代鲁棒回归方法

【最小绝对偏差回归】

LAD回归通过最小化绝对误差来获得鲁棒估计：

$$\min_{\beta} \sum_{i=1}^n |y_i - x_i^{\top} \beta|$$

理论性质：

- 断点达到 0.5
- 解不唯一时可能形成解集
- 可通过线性规划有效求解

例题4.38： LAD回归的线性规划形式

将LAD回归转化为线性规划问题。

解答：

引入辅助变量 $u_i, v_i \geq 0$ ，使得：

$$y_i - x_i^{\top} \beta = u_i - v_i, \quad |y_i - x_i^{\top} \beta| = u_i + v_i$$

线性规划形式：

$$\min_{\beta, u, v} \sum_{i=1}^n (u_i + v_i)$$

满足 $x_i^{\top} \beta + u_i - v_i = y_i, \quad u_i, v_i \geq 0$

为什么定义 $y_i - x_i^{\top} \beta = u_i - v_i$ 和 $|y_i - x_i^{\top} \beta| = u_i + v_i$ ？

- 对于 $u_i - v_i$ ：

任何实数 $r = y_i - x_i^{\top} \beta$ 都可以表示为两个非负变量的差，即 $r = u_i - v_i$ 。这是因为：

- 如果 $r \geq 0$ ，可以设 $u_i = r, v_i = 0$ 。
 - 如果 $r < 0$ ，可以设 $u_i = 0, v_i = -r$ 。
- 这样， $u_i - v_i$ 总能等于 r ，即 $y_i - x_i^{\top} \beta$ 。

- 对于 $u_i + v_i$ ：

在定义 $u_i - v_i = y_i - x_i^{\top} \beta$ 后，绝对值 $|y_i - x_i^{\top} \beta|$ 自然等于 $u_i + v_i$ 。因为：

- 如果 $r \geq 0$ ，有 $u_i = r, v_i = 0$ ，所以 $u_i + v_i = r = |r|$ 。
- 如果 $r < 0$ ，有 $u_i = 0, v_i = -r$ ，所以 $u_i + v_i = -r = |r|$ 。

此外，在线性规划中最小化 $\sum (u_i + v_i)$ 时，最优解会迫使对于每个 i ，要么 $u_i = 0$ 要么 $v_i = 0$ （如果两者都为正，可以通过减少其中一个来降低目标函数值），从而保证 $u_i + v_i = |y_i - x_i^{\top} \beta|$ 。

【RANSAC算法】

随机抽样一致性算法特别适用于高污染率数据：

算法步骤：

- 1. **随机采样**：从数据中随机选择最小样本集
- 2. **模型估计**：用最小样本集估计模型参数
- 3. **一致性评估**：计算所有数据与模型的吻合度
- 4. **模型选择**：选择具有最多内点的模型
- 5. **重估计**：用所有内点重新估计模型参数

例题4.39： RANSAC直线拟合

用RANSAC拟合包含50%异常值的直线数据。

解答：

假设数据点 $\{(x_i, y_i)\}_{i=1}^n$ ，其中约50%为异常值。

- 1. 随机选择2个点计算直线参数
- 2. 设置残差阈值 τ ，统计满足 $|y_i - (ax_i + b)| < \tau$ 的内点数量
- 3. 重复 N 次，选择内点最多的模型
- 4. 最终用所有内点通过最小二乘重新估计直线

所需采样次数 N 由下式确定：

$$N = \frac{\log(1 - p)}{\log(1 - (1 - \varepsilon)^m)}$$

其中 p 是置信水平， ε 是异常值比例， m 是最小样本量。

4.9.5 鲁棒位置估计的比较与应用

【多种位置估计量的比较】

估计量	断点	效率(正态)	计算复杂度	适用场景
均值	0	1.00	$O(n)$	纯净数据
中位数	0.5	0.64	$O(n)$	高污染数据
Huber估计	0.5	0.95	$O(n)$	平衡场景
10%截尾均值	0.1	0.97	$O(n \log n)$	轻度污染
25%截尾均值	0.25	0.83	$O(n \log n)$	中度污染

【实际应用建议】

- 1. **数据探索阶段**：计算多种鲁棒统计量进行比较
- 2. **模型选择**：根据预期的异常值比例选择合适方法
- 3. **参数调优**：对Huber估计的 k 值或截尾比例进行交叉验证
- 4. **结果验证**：结合残差分析和影响诊断

例题4.40： 综合鲁棒性分析

对数据集 $[2, 3, 4, 5, 6, 7, 8, 9, 10, 100]$ 进行全面的鲁棒性分析。

解答：

基础统计量：• **均值计算**

$$\text{公式: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

计算过程：

数据：2, 3, 4, 5, 6, 7, 8, 9, 10, 100

总和 = 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 100 = 154

均值 = 154 ÷ 10 = 15.4

• **中位数计算**

公式： 排序后中间位置的值（n为奇数取中间，n为偶数取中间两个的平均值）

计算过程：

数据已排序：2, 3, 4, 5, 6, 7, 8, 9, 10, 100

n = 10（偶数），中位数位置 = (10/2)和(10/2+1)的平均值

中位数 = (6 + 7) ÷ 2 = 13 ÷ 2 = 6.5

• **10%截尾均值计算**

公式： 去除两端各10%的数据后计算均值

计算过程：

- 数据总数：10个
- 去除数量：10 × 10% = 1个（两端各去1个，共去2个）
- 去除最小值2和最大值100
- 剩余数据：3, 4, 5, 6, 7, 8, 9, 10
- 剩余数据总和 = 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 = 52
- 剩余数据个数 = 8
- 截尾均值 = 52 ÷ 8 = 6.5
- **Huber估计 (k=1.345)**

Huber损失函数：

$$\rho_k(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq k \\ k|r| - \frac{1}{2}k^2 & \text{if } |r| > k \end{cases}$$

计算过程（迭代加权最小二乘）：

1. **初始估计：** 用中位数6.5作为初始值
2. **计算残差：** $r_i = x_i - 6.5$
[-4.5, -3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5, 93.5]
3. **计算权重：**
 $w_i = \min(1, k/|r_i|)$
[0.299, 0.384, 0.538, 0.897, 1, 1, 0.897, 0.538, 0.384, 0.014]
4. **加权均值：**
 $\hat{\mu} = \frac{\sum w_i x_i}{\sum w_i} = \frac{38.68}{5.95} \approx 6.5$
5. **迭代收敛**到最终值6.8

统计量总结：

- **均值：** 15.4
- **中位数：** 6.5

- **10%截尾均值**: 6.5 (按标准计算)
- **Huber估计**: 6.8

鲁棒性分析:

- 均值受异常值100严重影响 (15.4)
- 中位数和截尾均值对异常值不敏感 (6.5)
- Huber估计在保持效率的同时提供了良好的鲁棒性 (6.8)

推荐方案:

对于此数据集, 推荐使用25%截尾均值或Huber估计, 在保持效率的同时有效抑制极端值影响。

4.10 特殊数据结构下的估计特性

4.10.1 独立同分布样本的经典理论

独立同分布 (i.i.d.) 假设构成统计推断的理论基石。在此理想框架下, 众多估计量展现出优良的渐近性质, 为统计推断提供了坚实的理论基础。

大样本理论框架

设 $X_1, \dots, X_n \sim F_\theta$ 为独立同分布样本, 考虑估计量 $T_n = T(X_1, \dots, X_n)$, 其核心理论性质包括:

1. **相合性**: $T_n \xrightarrow{P} \theta$, 确保估计量随样本量增加而收敛于真实参数
2. **渐近正态性**: $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, V(\theta))$, 为置信区间构造提供依据
3. **效率界限**: Cramér-Rao下界给出无偏估计量的方差下限

重要估计量的精确分布

在特定分布假设下, 可获得估计量的精确有限样本分布, 为小样本推断奠定基础。

例题4.41: 正态分布的精确推断

设 $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, 推导样本均值和样本方差的精确分布。

解答:

样本均值的精确分布:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

样本方差的精确分布:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

重要性质: \bar{X} 与 S^2 相互独立。这些精确分布在有限样本情形下为统计推断提供了严格的理论保障。

4.10.2 时间序列数据的估计特性

时间序列数据普遍存在于经济、金融、气象等领域，其时间相关性使得传统i.i.d.假设下的统计理论不再适用，需要发展专门的建模和估计方法。

自回归模型的理论性质

考虑一阶自回归模型AR(1)：

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2)$$

普通最小二乘估计量为：

$$\hat{\phi} = \frac{\sum_{t=2}^T y_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2}$$

该估计量具有以下理论性质：

- **有限样本偏差：** $\mathbb{E}[\hat{\phi}] = \phi - \frac{1+3\phi}{T} + O(T^{-2})$
- **渐近分布：** $\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{d} N(0, 1 - \phi^2)$
- **相合性：** 当 $|\phi| < 1$ 时， $\hat{\phi} \xrightarrow{P} \phi$

异方差与自相关一致估计

在时间序列回归中，误差项往往存在异方差和自相关，此时标准误的估计需要修正。Newey-West提出如下HAC估计量：

$$\hat{V} = \hat{\Gamma}_0 + \sum_{j=1}^m w(j, m)(\hat{\Gamma}_j + \hat{\Gamma}_j^\top)$$

其中 $\hat{\Gamma}_j = \frac{1}{T} \sum_{t=j+1}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-j}$ ， $w(j, m)$ 为权重函数（常用Bartlett权重： $w(j, m) = 1 - \frac{j}{m+1}$ ）。

例题4.42：宏观经济模型的HAC估计

估计消费函数 $C_t = \alpha + \beta Y_t + \varepsilon_t$ 的标准误，考虑误差项可能存在自相关。

解答：

步骤1：初步OLS估计

用OLS估计原模型，得到残差序列 $\hat{\varepsilon}_t$

步骤2：自相关分析

计算样本自相关函数：

$$\hat{\rho}_j = \frac{\sum_{t=j+1}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-j}}{\sum_{t=1}^T \hat{\varepsilon}_t^2}$$

步骤3：带宽选择

按经验法则选择截断参数：

$$m = \lfloor 4(T/100)^{2/9} \rfloor$$

步骤4: Newey-West方差估计

计算HAC方差估计:

$$\hat{V}_{NW} = (X^\top X)^{-1} \left(\sum_{j=-m}^m w(j, m) \hat{\Gamma}_j \right) (X^\top X)^{-1}$$

该方法保证了在存在未知形式的异方差和自相关时, 标准误估计仍具有一致性。

4.10.3 空间数据的统计建模

空间数据普遍存在于环境科学、地理学、流行病学等领域, 其位置依赖性使得传统独立假设失效, 需要专门的空间统计方法。

空间自相关检验

Moran's I统计量是检验空间自相关性的经典方法:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

其中 w_{ij} 为空间权重矩阵, 度量位置 i 与 j 的空间邻近程度。

克里金插值理论体系

克里金法是最优空间预测的代表性方法, 其核心思想是通过变差函数建模空间相关性。普通克里金系统表示为:

$$\begin{bmatrix} \gamma(h_{11}) & \cdots & \gamma(h_{1n}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(h_{n1}) & \cdots & \gamma(h_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(h_{10}) \\ \vdots \\ \gamma(h_{n0}) \\ 1 \end{bmatrix}$$

其中 $\gamma(h)$ 是变差函数, μ 是拉格朗日乘子, λ_i 为克里金权重。

例题4.43: 环境监测的空间预测

某区域有10个空气质量监测站, 需要预测未监测位置的污染水平。

解答:

步骤1: 变差函数建模与估计

计算经验变差函数:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{|s_i - s_j| \approx h} (z(s_i) - z(s_j))^2$$

拟合理论变差模型 (以球状模型为例):

$$\gamma(h) = \begin{cases} c_0 + c \left(\frac{3h}{2a} - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right) & \text{若 } h \leq a \\ c_0 + c & \text{若 } h > a \end{cases}$$

其中 c_0 为块金效应, c 为部分基台值, a 为变程。

步骤2: 克里金预测系统求解

对于目标位置 s_0 , 求解克里金权重向量 λ , 预测值为:

$$\hat{z}(s_0) = \sum_{i=1}^{10} \lambda_i z(s_i)$$

同时可获得预测不确定性度量:

$$\sigma^2(s_0) = \sum_{i=1}^{10} \lambda_i \gamma(|s_i - s_0|) + \mu$$

现代发展趋势

1. **贝叶斯空间模型**: 通过层次建模结合先验信息, 处理复杂空间结构
2. **时空联合建模**: 同时刻画空间相关性和时间动态性
3. **大数据计算方法**: 发展面向海量空间数据的近似算法和分布式计算

特殊数据结构下的统计理论为各类实际应用提供了重要指导, 理解这些框架下的估计特性有助于针对性地选择合适方法, 有效处理现实世界中的复杂数据问题。