目　录

写在前面

Life, thin and light-off time and time again
Frivolous tireless
生命，一次又一次轻薄过
轻狂不知疲倦

基础知识

作为第 **??** 章统计模型和第 **??** 章参数估计的知识准备，本章给出主要的知识点。第 0.1 节首先介绍指数族的一般形式，包含各成分的定义，特别介绍正态分布、二项分布和泊松分布情形下均值函数、联系函数和方差函数等特征量。第 0.2 节介绍线性模型下，设计矩阵保持正定时的最小二乘估计和加权最小二乘估计。第 0.3 节介绍极大似然估计的定义，相合性，以及在一定条件下的渐近正态性。第 0.4 节介绍平稳高斯过程的定义，均方连续性和可微性的定义，以及判断可微性的一个充要条件。第 0.5 介绍先验、后验分布和 Jeffreys 无信息先验分布。

0.1 指数族

一般地，随机变量 Y 的分布服从指数族，即形如

$$f_Y(y;\theta,\phi) = \exp\Big\{\big(y\theta - b(\theta)\big)/a(\phi) + c(y,\phi)\Big\} \tag{1}$$

其中，$a(\cdot), b(\cdot), c(\cdot)$ 是某些特定的函数。如果 ϕ 已知，这是一个含有典则参数 θ 的指数族模型，如果 ϕ 未知，它可能是含有两个参数的指数族。对于正态分布

$$\begin{aligned} f_Y(y;\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(y-\mu)^2}{2\sigma^2}\} \\ &= \exp\Big\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}\Big(y^2/\sigma^2 + \log(2\pi\sigma^2)\Big)\Big\} \end{aligned} \tag{2}$$

通过与 (1) 式对比，可知 $\theta = \mu$，$\phi = \sigma^2$，并且有

$$a(\phi) = \phi, \quad b(\theta) = \theta^2/2, \quad c(y,\phi) = -\frac{1}{2}\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}$$

记 $l(\theta,\phi;y) = \log f_Y(y;\theta,\phi)$ 为给定样本点 y 的情况下，关于 θ 和 ϕ 的对数似然函数。样本 Y 的均值和方差具有如下关系[?]

$$\mathsf{E}\Big(\frac{\partial l}{\partial\theta}\Big) = 0 \tag{3}$$

和

$$\mathsf{E}\Big(\frac{\partial^2 l}{\partial\theta^2}\Big) + \mathsf{E}\Big(\frac{\partial l}{\partial\theta}\Big)^2 = 0 \tag{4}$$

从 (1) 式知

$$l(\theta,\phi;y) = y\theta - b(\theta)/a(\phi) + c(y,\phi)$$

因此，

$$\begin{aligned}\frac{\partial l}{\partial \theta} &= y - b'(\theta)/a(\phi) \\ \frac{\partial^2 l}{\partial \theta^2} &= -b''(\theta)/a(\phi)\end{aligned} \tag{5}$$

从 (3) 式和 (5)，可以得出

$$0 = \mathsf{E}\Big(\frac{\partial l}{\partial \theta}\Big) = \Big\{\mu - b'(\theta)\Big\}/a(\phi)$$

所以

$$\mathsf{E}(Y) = \mu = b'(\theta)$$

根据 (4) 式和 (5) 式，可得

$$0 = -\frac{b''(\theta)}{a(\phi)} + \frac{\mathsf{Var}(Y)}{a^2(\phi)}$$

所以

$$\mathsf{Var}(Y) = b''(\theta)a(\phi)$$

可见，Y 的方差是两个函数的乘积，一个是 $b''(\theta)$，它仅仅依赖典则参数，叫做方差函数，方差函数可以看作是 μ 的函数，记作 $V(\mu)$。另一个是 $a(\phi)$，它独立于 θ，仅仅依赖 ϕ，函数 $a(\phi)$ 通常形如

$$a(\phi) = \phi/w$$

其中 ϕ 可由 σ^2 表示，故而也叫做发散参数（dispersion parameter），是一个与样本观察值相关的常数，w 是已知的权重，随样本观察值变化。对正态分布模型而言，w 的分量是 m 个相互独立的样本观察值的均值，有 $a(\phi) = \sigma^2/m$，所以，$w = m$。

根据 (1)式，正态、泊松和二项分布的特征见表 1，符号约定同 McCullagh 和 Nelder（1989 年）所著的《广义线性模型》。

表 1: 指数族内常见的一元分布的共同特征及符号表示

	正态分布	泊松分布	二项分布
记号	$\mathcal{N}(\mu,\sigma^2)$	Poisson(μ)	Binomial(m,p)
y 取值范围	$(-\infty,\infty)$	$0(1)\infty$	$0(1)m$
ϕ	$\phi=\sigma^2$	1	$1/m$
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1+e^\theta)$
$c(y;\theta)$	$-\frac{1}{2}\Big(\frac{y^2}{\phi}+\log(2\pi\phi)\Big)$	$-\log(y!)$	$\log\binom{m}{my}$
$\mu(\theta)=\mathsf{E}(Y;\theta)$	θ	$\exp(\theta)$	$e^\theta/(1+e^\theta)$
联系函数：$\theta(\mu)$	identity	log	logit

	正态分布	泊松分布	二项分布
方差函数：$V(\mu)$	1	μ	$\mu(1-\mu)$

0.2 最小二乘估计

考虑如下线性模型的最小二乘估计

$$\mathsf{E}\mathbf{Y}=\mathbf{X}\boldsymbol{\beta} \qquad \mathsf{Var}(\mathbf{Y})=\sigma^2\mathbf{I}_n \tag{6}$$

其中，$\mathbf{Y}$ 为 $n\times 1$ 维观测向量，$\mathbf{X}$ 为已知的 $n\times p(p\le n)$ 维设计矩阵，$\boldsymbol{\beta}$ 为 $p\times 1$ 维未知参数，σ^2 未知，$\mathbf{I}_n$ 为 n 阶单位阵。

定义 0.1 (最小二乘估计)**.** 在模型 (6) 中，如果

$$(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})=\min_{\beta}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \tag{7}$$

则称 $\hat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的最小二乘估计 (简称 LSE)[?]。

定理 0.1 (最小二乘估计)**.** *若模型 (6) 中的 $\mathbf{X}$ 是列满秩的矩阵，则 $\boldsymbol{\beta}$ 的最小二乘估计为*

$$\hat{\boldsymbol{\beta}}_{LS}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y},\quad \mathsf{Var}(\hat{\boldsymbol{\beta}}_{LS})=\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

σ^2 的最小二乘估计为

$$\hat{\sigma^2}_{LS}=(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{LS})^\top(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{LS})/(n-p)$$

若将模型 (6) 的条件 $\mathsf{Var}(\mathbf{Y})=\sigma^2\mathbf{I}_n$ 改为 $\mathsf{Var}(\mathbf{Y})=\sigma^2\mathbf{G}$，$\mathbf{G}(>0)$ 为已知正定阵，则 $\boldsymbol{\beta}$ 的最小二乘估计为

$$\tilde{\boldsymbol{\beta}}_{LS}=(\mathbf{X}^\top\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{G}^{-1}\mathbf{Y}$$

称 $\tilde{\boldsymbol{\beta}}_{LS}$ 为广义最小二乘估计，特别地，当 $\mathbf{G}=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$，$\sigma_i^2,i=1,\ldots,n$ 已知时，称 $\tilde{\boldsymbol{\beta}}_{LS}$ 为加权最小二乘估计[?]。

0.3 极大似然估计

定义 0.2 (极大似然估计)**.** 设 $p(\mathbf{x};\boldsymbol{\theta}),\boldsymbol{\theta}\in\boldsymbol{\Theta}$ 是 $(\mathbb{R}^n,\mathscr{P}_{\mathbb{R}^n})$ 上的一族联合密度函数，对给定的 $\mathbf{x}$，称

$$L(\boldsymbol{\theta};\mathbf{x})=kp(\mathbf{x};\boldsymbol{\theta})$$

为 $\boldsymbol{\theta}$ 的似然函数，其中 $k>0$ 是不依赖于 $\boldsymbol{\theta}$ 的量，常取 $k=1$。进一步，若存在 $(\mathbb{R}^n, \mathscr{P}_{\mathbb{R}^n})$ 到 $(\boldsymbol{\Theta}, \mathscr{P}_{\boldsymbol{\Theta}})$ 的统计量 $\hat{\boldsymbol{\theta}}(\mathbf{x})$ 使

$$L(\hat{\boldsymbol{\theta}}(\mathbf{x});\mathbf{x}) = \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\mathbf{x})$$

则 $\hat{\boldsymbol{\theta}}(\mathbf{x})$ 称为 $\boldsymbol{\theta}$ 的一个极大似然估计（简称 MLE）[?]。

概率密度函数很多可以写成具有指数函数的形式，如指数族，采用似然函数的对数通常更为简便。称

$$l(\boldsymbol{\theta},\mathbf{x}) = \ln L(\boldsymbol{\theta},\mathbf{x})$$

为 $\boldsymbol{\theta}$ 的对数似然函数。对数变换是严格单调的，所以 $l(\boldsymbol{\theta},\mathbf{x})$ 与 $L(\boldsymbol{\theta},\mathbf{x})$ 的极大值是等价的。当 MLE 存在时，寻找 MLE 的常用方法是求导数。如果 $\hat{\boldsymbol{\theta}}(\mathbf{x})$ 是 $\boldsymbol{\Theta}$ 的内点，则 $\hat{\boldsymbol{\theta}}(\mathbf{x})$ 是下列似然方程组

$$\partial l(\boldsymbol{\theta},\mathbf{x})/\partial \boldsymbol{\theta}_i = 0, \quad i=1,\ldots,m \tag{8}$$

的解。$p(\mathbf{x};\boldsymbol{\theta})$ 属于指数族时，似然方程组 (8) 的解唯一[?]。

定理 0.2 (相合性)**.** *设 $x_1,\ldots,x_n$ 是来自概率密度函数 $p(\mathbf{x};\boldsymbol{\theta})$ 的一个样本，叙述简单起见，考虑单参数情形，参数空间 $\boldsymbol{\Theta}$ 是一个开区间，$l(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \ln p(x_i;\boldsymbol{\theta})$。*

若 $\ln(p;\boldsymbol{\theta})$ 在 $\boldsymbol{\Theta}$ 上可微，且 $p(\mathbf{x};\boldsymbol{\theta})$ 是可识别的（即 $\forall \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2, \{\mathbf{x} : p(\mathbf{x};\boldsymbol{\theta}_1) \neq p(\mathbf{x};\boldsymbol{\theta}_2)\}$ 不是零测集），则似然方程 (8) 在 $n\to\infty$ 时，以概率 1 有解，且此解关于 $\boldsymbol{\theta}$ 是相合的[?]。

定理 0.3 (渐近正态性)**.** *假设 $\boldsymbol{\Theta}$ 为开区间，概率密度函数 $p(\mathbf{x};\boldsymbol{\theta}), \boldsymbol{\theta}\in\boldsymbol{\Theta}$ 满足:*

1. *在参数真值 $\boldsymbol{\theta}_0$ 的邻域内，$\partial \ln p/\partial\boldsymbol{\theta}, \partial^2 \ln p/\partial\boldsymbol{\theta}^2, \partial^3 \ln p/\partial\boldsymbol{\theta}^3$ 对所有的 $\mathbf{x}$ 都存在;*
2. *在参数真值 $\boldsymbol{\theta}_0$ 的邻域内，$|\partial^3 \ln p/\partial\boldsymbol{\theta}^3| \leq H(\mathbf{x})$，且 $\mathsf{E}H(\mathbf{x}) < \infty$;*
3. *在参数真值 $\boldsymbol{\theta}_0$ 处，$\mathsf{E}_{\boldsymbol{\theta}_0}\left[\frac{p'(\mathbf{x},\boldsymbol{\theta}_0)}{p(\mathbf{x},\boldsymbol{\theta}_0)}\right] = 0, \mathsf{E}_{\boldsymbol{\theta}_0}\left[\frac{p''(\mathbf{x},\boldsymbol{\theta}_0)}{p(\mathbf{x},\boldsymbol{\theta}_0)}\right] = 0, I(\boldsymbol{\theta}_0) = \mathsf{E}_{\boldsymbol{\theta}_0}\left[\frac{p'(\mathbf{x},\boldsymbol{\theta}_0)}{p(\mathbf{x},\boldsymbol{\theta}_0)}\right]^2 > 0$。*

其中，撇号表示对 $\boldsymbol{\theta}$ 的微分。记 $\hat{\boldsymbol{\theta}}_n$ 为 $n\to\infty$ 时，似然方程组的相合解，则 $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \longrightarrow \mathcal{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}))$[?]。

0.4 平稳高斯过程

一般地，空间高斯过程 $\mathcal{S} = \{S(x), x\in\mathbb{R}^2\}$ 必须满足条件：任意给定一组空间位置 $x_1, x_2, \ldots, x_n, \forall x_i \in \mathbb{R}^2$，每个位置上对应的随机变量 $S(x_i), i=1,2,\ldots,n$ 的联合

分布 $\mathcal{S}=\{S(x_1),S(x_2),\ldots,S(x_n)\}$ 是多元高斯分布，其由均值 $\mu(x)=\mathsf{E}[S(x)]$ 和协方差 $G_{ij}=\gamma(x_i,x_j)=\mathsf{Cov}\{S(x_i),S(x_j)\}$ 完全确定，即 $\mathcal{S}\sim\mathcal{N}(\mu_S,G)$。

平稳空间高斯过程需要空间高斯过程满足平稳性条件：其一，$\mu(x)=\mu,\forall x\in\mathbb{R}^2$，其二，自协方差函数 $\gamma(x_i,x_j)=\gamma(u),u=\|x_i-x_j\|$。可见均值 μ 是一个常数，而自协方差函数 $\gamma(x_i,x_j)$ 只与空间距离有关。

平稳高斯过程 $\mathcal{S}$ 的方差是一个常数，即 $\sigma^2=\gamma(0)$，然后可以定义自相关函数 $\rho(u)=\gamma(u)/\sigma^2$，并且 $\rho(u)$ 是关于空间距离 u 对称的，即 $\rho(u)=\rho(-u)$。因为对 $\forall u,\mathsf{Corr}\{S(x),S(x-u)\}=\mathsf{Corr}\{S(x-u),S(x)\}=\mathsf{Corr}\{S(x),S(x+u)\}$，这里的第二个等式是根据平稳性得来的，由协方差的定义不难验证。如果不特别说明，平稳就指上述协方差意义下的平稳，因为这种平稳性条件广泛应用于空间数据的统计建模。不失一般性，介绍一维空间下随机过程 $S(x)$ 的均方连续性和可微性定义。

定义 0.3 (连续性和可微性)**.** 随机过程 $S(x)$ 满足

$$\lim_{h\to0}\mathsf{E}\Big[\{S(x+h)-S(x)\}^2\Big]=0$$

则称 $S(x)$ 是均方连续（mean-square continuous）的。随机过程 $S(x)$ 满足

$$\lim_{h\to0}\mathsf{E}\Big[\{\frac{S(x+h)-S(x)}{h}-S'(x)\}^2\Big]=0$$

则称 $S(x)$ 是均方可微（mean-square differentiable）的，并且 $S'(x)$ 就是均方意义下的一阶导数。如果 $S'(x)$ 是均方可微的，则 $S(x)$ 是二次均方可微的，随机过程 $S(x)$ 的高阶均方可微性可类似定义[?]。Bartlett （1955 年)[?] 得到如下重要结论

定理 0.4 (平稳随机过程的可微性)**.** *自相关函数为 $\rho(u)$ 的平稳随机过程是 k 次均方可微的，当且仅当 $\rho(u)$ 在 $u=0$ 处是 $2k$ 次可微的。*

0.5 先验和后验分布

贝叶斯推断中，常涉及模型参数的先验、后验分布，以及一种特殊的无信息先验分布 — Jeffreys 先验，下面分别给出它们的概念定义[?]。

定义 0.4 (先验分布)**.** 参数空间 Θ 上的任一概率分布都称作先验分布（prior distribution)[?]。

定义 0.5 (后验分布)**.** 在获得样本 $\mathbf{Y}$ 后，模型参数 $\boldsymbol{\theta}$ 的后验分布（posterior distribution）就是在给定样本 $\mathbf{Y}$ 的条件下 $\boldsymbol{\theta}$ 的分布[?]。

定义 0.6 (Jeffreys 先验分布)**.** 设 $\mathbf{x}=(x_1,\ldots,x_n)$ 是来自密度函数 $p(x|\boldsymbol{\theta})$ 的一个样本，其中 $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)$ 是 p 维参数向量。在对 $\boldsymbol{\theta}$ 无任何先验信息可用时，Jeffreys（1961

年）利用变换群和 Harr 测度导出 $\boldsymbol{\theta}$ 的无信息先验分布可用 Fisher 信息阵的行列式的平方根表示。这种无信息先验分布常称为 Jeffreys 先验分布。其求取步骤如下：

1. 写出样本的对数似然函数 $l(\boldsymbol{\theta}|x)=\sum_{i=1}^{n}\ln p(x_i|\boldsymbol{\theta})$；
2. 算出参数 $\boldsymbol{\theta}$ 的 Fisher 信息阵

$$\mathbf{I}(\boldsymbol{\theta})=\mathsf{E}_{x|\theta}\Big(-\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\Big)_{i,j=1,\dots,p}$$

   在单参数场合，$\mathbf{I}(\theta)=\mathsf{E}_{x|\theta}\Big(-\frac{\partial^2 l}{\partial\theta^2}\Big)$；
3. $\boldsymbol{\theta}$ 的无信息先验密度函数为 $\pi(\boldsymbol{\theta})=[\det\mathbf{I}(\boldsymbol{\theta})]^{1/2}$，在单参数场合，$\pi(\boldsymbol{\theta})=[\mathbf{I}(\theta)]^{1/2}$[?]。

0.6 常用贝叶斯估计

定理 0.5 (平方损失). *在给定先验分布 $\pi(\boldsymbol{\theta})$ 和平方损失 $L(\boldsymbol{\theta},\boldsymbol{\delta})=(\boldsymbol{\delta}-\boldsymbol{\theta})^2$ 下，$\boldsymbol{\theta}$ 的贝叶斯估计 $\boldsymbol{\delta}^{\pi}(x)$ 为后验分布 $\pi(\boldsymbol{\theta}|x)$ 的均值，即 $\boldsymbol{\delta}^{\pi}(x)=\mathsf{E}(\boldsymbol{\theta}|x)$[?]。*

定理 0.6 (0 - 1 损失). *在给定先验分布 $\pi(\boldsymbol{\theta})$ 和 0 - 1 损失函数*

$$L(\boldsymbol{\theta},\boldsymbol{\delta})=\begin{cases}1, & |\boldsymbol{\delta}-\boldsymbol{\theta}|\le\epsilon\\ 0, & |\boldsymbol{\delta}-\boldsymbol{\theta}|>\epsilon\end{cases}$$

当 ϵ 较小时，$\boldsymbol{\theta}$ 的贝叶斯估计 $\boldsymbol{\delta}^{\pi}(x)$ 为后验分布 $\pi(\boldsymbol{\theta}|x)$ 的众数[?]。

定理 0.7 (绝对值损失). *在给定先验分布 $\pi(\boldsymbol{\theta})$ 和绝对损失函数 $L(\boldsymbol{\theta},\boldsymbol{\delta})=|\boldsymbol{\delta}-\boldsymbol{\theta}|$ 下，$\boldsymbol{\theta}$ 的贝叶斯估计 $\boldsymbol{\delta}^{\pi}(x)$ 为后验分布 $\pi(\boldsymbol{\theta}|x)$ 的中位数[?]。*

评价贝叶斯估计 $\boldsymbol{\delta}^{\pi}(x)$ 的精度常用后验均方误差

$$\mathsf{MSE}(\boldsymbol{\delta}^{\pi}|x)=\mathsf{E}_{\boldsymbol{\theta}|x}(\boldsymbol{\delta}^{\pi}-\boldsymbol{\theta})^2$$

表示，或用其平方根 $[\mathsf{MSE}(\boldsymbol{\delta}^{\pi}|x)]^{1/2}$ （称为标准误）表示。容易算得

$$\mathsf{MSE}(\boldsymbol{\delta}^{\pi}|x)=\mathsf{Var}(\boldsymbol{\delta}^{\pi}|x)+[\boldsymbol{\delta}^{\pi}(x)-\mathsf{E}(\boldsymbol{\theta}|x)]^2$$

可见，当贝叶斯估计 $\boldsymbol{\delta}^{\pi}(x)$ 为后验均值时，贝叶斯估计的精度就用 $\boldsymbol{\delta}^{\pi}$ 的后验方差 $\mathsf{Var}(\boldsymbol{\delta}^{\pi}|x)$ 表示，或用后验标准差 $[\mathsf{Var}(\boldsymbol{\delta}^{\pi}|x)]^{1/2}$ 表示[?]。

0.7 本章小结

本章第0.1节介绍了指数族的一般形式，指出基于样本点的对数似然函数和样本均值、样本方差的关系，以表格的形式列出了正态、泊松和二项分布的各个特征，为

第**??**章统计模型和第**??**章参数估计作铺垫。接着，第0.2节和第0.3节分别介绍了最小二乘估计和极大似然估计的定义、性质，给出了线性模型的最小二乘估计，极大似然估计的相合性和渐进正态性。第0.4节介绍了平稳高斯过程，给出了其均方连续性、可微性定义以及一个均方可微的判断定理，平稳高斯过程作为空间随机效应的实现，多次出现在后续章节中。第0.5节至第0.6节分别是与贝叶斯相关的概念定义。

Literature

Here is a review of existing methods.

Methods

We describe our methods in this chapter.

Applications

Some *significant* applications are demonstrated in this chapter.

0.8 Example one

0.9 Example two

Final Words

We have finished a nice book.