极大似然辨识

设总体X是离散型随机变量，其概率函数为 $p(x;\theta)$，其中 θ 是未知参数。设 $X_1X_2\cdots X_n$ 为取自总体X的样本。$X_1X_2\cdots X_n$ 的联合概率函数为 $\prod_{i=1}^{n} p(X_i,\theta)$。这里，$\theta$ 是常量，$X_1X_2\cdots X_n$ 是变量。

如果样本取值 $x_1x_2\ldots x_n$，则事件 $\{X_1=x_1,\cdots,X_n=x_n\}$ 发生的概率为 $\prod_{i=1}^{n} p(x_i,\theta)$。这一概率随 θ 的值变化而变化。从直观上来看，既然样本值 $x_1x_2\ldots x_n$ 已经出现了，它们出现的概率相对来说应比较大，应使其概率取比较大的值。取似然函数如下：

$$L(\theta)=L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n} p(x_i;\theta)$$

极大似然估计法就是在参数 θ 的可能取值范围内，选取使 $L(\theta)$ 达到最大的参数值 $\hat{\theta}$：

$$L(\theta) = L(x_1, x_2, \cdots, x_n; \hat{\theta}) = \max_{\theta\in\Theta} L(x_1, x_2, \cdots, x_n; \theta)$$

$$= \max_{\theta\in\Theta} \prod_{i=1}^{n} p(x_i; \theta)$$

因此，求参数 θ 的极大似然估计值的问题就是求似然函数最大值问题。这通过解方程 $dL(\theta)/d\theta = 0$ 来得到。因为 $\ln L(\theta)$ 和 $L(\theta)$ 的增减性相同，所以它们在 θ 的同一值处取得最大值，称 $\ln L(\theta)$ 为对数似然函数。可以通过求解下列方程来得到极大似然解。

$$\frac{d \ln L(\theta)}{d\theta} = 0$$

例1：设某工序生产的产品的不合格率为p，抽n个产品作检验，发现有T个不合格，试求p的极大似然估计值。

分析：设X是抽查一个产品时的不合格品的个数，则X服从参数为p的两点分布。抽查n个产品，则得样本$X_1, X_2, \ldots X_n$，其观察值为$x_1, x_2 \ldots x_n$，假如样本有T个不合格，即表示$x_1, x_2 \ldots x_n$中有T个取值为1，有n-T个取值为0。基于此求参数p的极大似然估计值。

（1）写出似然函数 $$L(p)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$ 无限总体—二项分布

（2）对似然函数取对数，得到对数似然函数：

$$l(p)=\sum_{i=1}^{n}[x_i \ln p+(1-x_i)\ln(1-p)]$$

$$=n\ln(1-p)+\sum_{i=1}^{n} x_i[\ln p-\ln(1-p)]$$

（3）对似然函数求导，令其为零，得到似然估计值

$$\frac{dl(p)}{dp}=-\frac{n}{1-p}+\sum_{i=1}^{n} x_i\left(\frac{1}{p}+\frac{1}{1-p}\right)=-\frac{n}{1-p}+\frac{1}{p(1-p)}\sum_{i=1}^{n} x_i=0$$

$$\Rightarrow \hat{p}=\frac{1}{n}\sum_{i=1}^{n} x_i=\frac{T}{n}$$

例2：设某机床加工的轴的直径与图纸规定的中心尺寸的偏差服从 $N(\mu,\sigma^2)$，其中参数 μ,σ^2 未知。为了估计 μ,σ^2，从中随机抽取n=100根轴，测得其偏差为 $x_1, x_2 \ldots x_{100}$。试求 μ,σ^2 的极大似然估计。

分析：显然，该问题是求解含有多个（两个）未知参数的极大似然估计问题。通过建立关于未知参数 μ,σ^2 的似然方程组，从而进行求解。

$f(x_i;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

似然函数

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

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

↑与 μ 无关

$$\begin{cases}\dfrac{\partial l(\mu,\sigma^2)}{\partial\mu}=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)=0 \longrightarrow \sum_{i=1}^{n}x_i-n\mu=0\\[2ex] \dfrac{\partial l(\mu,\sigma^2)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2=0 \longrightarrow \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\end{cases}$$

注意是对 σ^2 求导

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x} \qquad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

例3：某电子管的使用寿命X(单位:小时)服从指数分布：

$$X: \quad p(x;\theta)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}}, & x>0\\ 0 \quad , & other\end{cases} \qquad (\theta>0)$$

今取得一组样本X_k数据如下，问如何估计θ？

16	29	50	68	100	130	140	270	280
340	410	450	520	620	190	210	800	1100

似然函数. $$L(\theta)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-\frac{x_i}{\theta}}=\theta^{-n}\cdot e^{-\frac{1}{\theta}\sum_{i=1}^{n}x_i}$$

$$\ln L=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i$$

$$\frac{d\ln L}{d\theta}=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i=0$$

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}=\frac{1}{18}\cdot 5723\approx 318$$

极大似然估计的法的运算步骤：

1、由总体分布导出样本的联合概率函数；

2、把样本联合概率函数中自变量看成已知常数，而把参数 θ 看作自变量，得到似然函数 $L(\theta)$；

3、求似然函数的最大值点（常转化为求对数似然函数的最大值点）；

4、在最大值点的表达式中，用样本值代入就得参数的极大似然估计值。

最后一步

下面利用极大似然原理，分析动态系统模型参数的极大似然估计问题。

考虑系统模型为线性差分方程：

$$y(k)=-a_1y(k-1)-\cdots-a_ny(k-n)+b_0u(k)+\cdots+b_nu(k-n)+\varepsilon(k)$$

其中 $\varepsilon(k)\sim N\left(0,\sigma^2\right)$ 为高斯白噪声，模型的估计问题可以表示成以下向量问题：

$$Y = \left[y(n+1) \;\; y(n+2) \;\; \cdots \;\; y(n+N) \right]^T$$

$$e = \left[\varepsilon(n+1) \;\; \varepsilon(n+2) \;\; \cdots \;\; \varepsilon(n+N) \right]^T$$

$$\theta = \left[a_1 \; a_2 \cdots \; a_n \;\; b_0 \; b_1 \cdots \; b_n \right]^T$$

$$\Phi = \begin{bmatrix} -y(n) & \cdots & -y(1) & u(n+1) & \cdots & u(1) \\ -y(n+1) & \cdots & -y(2) & u(n+2) & \cdots & u(2) \\ & \cdots & & \cdots & \cdots & \\ -y(n+N-1) & \cdots & -y(N) & u(n+N) & \cdots & u(N) \end{bmatrix}$$

$\varepsilon(n+1) \sim N(0,\sigma^2)$

$\varepsilon(n+2) \sim N(0,\sigma^2)$

$\varepsilon(n+N) \sim N(0,\sigma^2)$

$$Y = \Phi\theta + e \qquad \Rightarrow e = Y - \Phi\theta$$

输出序列的似然函数难以列出

而Φ、θ都是确定性的 ⇒ 列写Y的似然函数可以归结为列写e的似然函数

由于 $\{e(k)\}$ 是均值为零的高斯不相关序列，且与{u(k)}不相关，于是得到似然函数：

$$\frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \longleftarrow$$

n维正态分布的联合概率密度

向量e服从N维正态分布，令n=N

$\Sigma=\begin{pmatrix}\sigma^2 & & \\ & \ddots & \\ & & \sigma^2\end{pmatrix}=\sigma^2 I,\ |\Sigma|=(\sigma^2)^N$

$$L=P\left(e\middle|\theta,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-\frac{N}{2}}\exp\left\{-\frac{1}{2\sigma^2}\underbrace{\left(Y-\Phi\theta\right)^T\left(Y-\Phi\theta\right)}_{=\sum_{k=n+1}^{n+N}e^2(k)}\right\}$$

对应的$\underbrace{\text{负对数似然函数}}_{\text{最小化}}$为：

$$-\ln L=\underbrace{\frac{N}{2}\ln\sigma^2+\frac{N}{2}\ln 2\pi}_{\text{常数}}+\underbrace{\frac{1}{2\sigma^2}}_{\text{常系数}}\underbrace{\left(Y-\Phi\theta\right)^T\left(Y-\Phi\theta\right)}_{\text{最小化}}$$

根据极大似然原理，求上式对未知参数θ,σ^2求偏导数且令其为0，可得：

$$\hat{\theta}_{ML}=\left(\Phi^T\Phi\right)^{-1}\Phi^T Y$$

$$\hat{\sigma}^2_{ML}=\frac{1}{N}\left(Y-\Phi\hat{\theta}_{ML}\right)^T\left(Y-\Phi\hat{\theta}_{ML}\right)$$

这与最小二乘法的结果相同，这说明当噪声为高斯白噪声时，参数θ的极大似然估计和最小二乘估计是等价的。

在实际问题中，$\{e(k)\}$往往不是白噪声序列，而是相关噪声序列。下面讨论残差相关的情况下极大似然估计的求解。

数值解法

考虑模型为如下形式：

高斯噪声

$$A(z^{-1})y(k)=B(z^{-1})u(k)+\underbrace{C(z^{-1})}_{\text{新引入}}\varepsilon(k)$$

$$\underset{\text{白噪声}}{\varepsilon(k)} \xrightarrow{\text{表示定理}} \underset{\text{有色噪声(相关噪声)}}{\frac{C(z^{-1})}{A(z^{-1})}\varepsilon(k)}$$

$$A(z^{-1})=1+a_1z^{-1}+\cdots+a_nz^{-n}$$

$$B(z^{-1})=b_0+b_1z^{-1}+\cdots+b_nz^{-n}$$

$$C(z^{-1})=1+c_1z^{-1}+\cdots+c_nz^{-n}$$

上式可以改写为：

当 $c_1=c_2=\cdots=c_n=0$ 时退化为原来的简单情况

$$\varepsilon(k)=y(k)+\sum_{i=1}^{n}a_iy(k-i)-\sum_{i=0}^{n}b_iu(k-i)-\sum_{i=1}^{n}c_i\varepsilon(k-i)$$

列写Y的似然函数可以归结为列写高斯噪声e的似然函数

$$\varepsilon(k)=y(k)+\sum_{i=1}^{n}a_i y(k-i)-\sum_{i=0}^{n}b_i u(k-i)-\sum_{i=1}^{n}c_i\varepsilon(k-i)$$

在独立观测的前提下，得到输入输出数据{y(k)}和{u(k)}，测量N次，得到N值白噪声向量为：

$$\varepsilon=\left[\varepsilon(n+1)\ \ \varepsilon(n+2)\ \cdots\ \varepsilon(n+N)\right]^T,\quad \varepsilon\sim N\left(0,\sigma^2 I\right)$$

噪声的协方差阵为：

$$R=E\left\{\varepsilon\varepsilon^T\right\}=\sigma^2 I$$

令：$\theta=\left[a_1\ a_2\cdots\ a_n\ \ b_0\ \ b_1\cdots\ b_n\ \ c_1\ \ c_2\cdots\ c_n\right]^T$

$$Y=\begin{bmatrix} y(n+1) \\ y(n+2) \\ \vdots \\ y(n+N) \end{bmatrix}$$

向量形式的方程组可以写为：

$$Y=\Phi\theta+\varepsilon\Rightarrow\varepsilon=Y-\Phi\theta$$

$$\underline{\varepsilon=Y-\Phi(\theta)\theta}$$

$$\Phi=\left[\begin{array}{c:c:c} 关于y & 关于u & 关于\varepsilon \end{array}\right]$$

Φ是θ的非线性函数

含参数θ

要考

√非线性最小二乘 线性最小二乘: $\min_x \|Ax-b\|_2^2=\sum_{i=1}^{n}(a_i^Tx-b_i)^2$, 其中 $A=\begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix}$

→非线性函数 $f_i(x)$

$$\min_x f(x)=\frac{1}{2}\sum_{i=1}^{n} f_i^2(x)=\frac{1}{2}F(x)^T F(x)=\frac{1}{2}\|F(x)\|_2^2$$

$$F(x)=[f_1(x), f_2(x), \cdots, f_n(x)]^T$$

梯度 $\nabla f=J(x)^T F(x)$, 其中 $J(x)=\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix}$ $\begin{matrix} \to \frac{\partial f_1}{\partial x} \\ \\ \to \frac{\partial f_n}{\partial x} \end{matrix}$

负梯度方向 抖动大,收敛慢 => 牛顿下降方向:收敛快

$$\nabla^2 f \approx J(x)^T J(x)$$

$$x_{k+1}=x_k-[J(x_k)^T J(x_k)]^{-1}\nabla f(x_k)$$

只用到一阶导数信息

$$Y=\Phi\theta+\varepsilon\Rightarrow\varepsilon=Y-\Phi\theta$$

此时的联合概率密度为：

θ的非线性函数，可写成$\varepsilon_k(\theta)$

$$P\left(\varepsilon\middle|\theta,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-\frac{N}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{k=n+1}^{n+N}\varepsilon^2(k)\right\}$$

当θ是某个估计值时，把$\varepsilon(k)$改写为v(k)，则得到似然函数，并求对数得到：

$$\ln L=-\frac{N}{2}\ln 2\pi-\frac{N}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{k=n+1}^{n+N}v^2(k)$$

$$\frac{\partial\ln L}{\partial\hat{\sigma}^2}=0\qquad\Rightarrow\hat{\sigma}^2=\frac{1}{N}\sum_{k=n+1}^{n+N}v^2(k)$$

其中：

$$v(k)=y(k)+\sum_{i=1}^{n}\hat{a}_i y(k-i)-\sum_{i=0}^{n}\hat{b}_i u(k-i)-\sum_{i=1}^{n}\hat{c}_i v(k-i)$$

进一步得到：

$$\ln L = const - \frac{N}{2}\ln\frac{1}{N}\sum_{k=n+1}^{n+N} v^2(k)$$

根据极大似然原理，对数似然函数取极值，等价于：

$$\min \quad V(\hat{\theta}_{ML}) = \sum_{k=n+1}^{n+N} v^2(k)\Big|_{\hat{\theta}_{ML}}$$

式中$v(k)$满足约束条件。

综合以上分析，极大似然估计就是使得 $\min\ V(\hat{\theta}_{ML})$

因为$V(\theta)$ 是参数的非线性函数，只能通过迭代法求解. 这里介绍 Newton-Raphson法。

（1） 选定初始值$\hat{\theta}(0)$。对于$\hat{\theta}(0)$中的参数$a_1, a_2 \cdots a_n$，$b_0, b_1 \ldots b_n$，可按模型：

$$v(k) = \hat{A}(z^{-1})y(k) - \hat{B}(z^{-1})u(k)$$

用最小二乘法求得，对于$\hat{\theta}(0)$中的$c_0, c_1 \ldots c_n$可以先假定一些值。

例如$c_0 = c_1 = \cdots = c_n = 0$

(2) 计算预测误差 $v(k)=y(k)-\hat{y}(k) \qquad k=n+1,\cdots\cdots$

目标最优化 $J=\frac{1}{2}\sum_{k=n+1}^{n+N}v^2(k)$

(3) 计算J的梯度 $\partial J/\partial\hat{\theta}$ 和Hessian矩阵 $\partial^2 J/\partial\hat{\theta}\partial\theta^T$

$$\frac{\partial J}{\partial\hat{\theta}}=\sum_{k=n+1}^{n+N}v(k)\frac{\partial v(k)}{\partial\hat{\theta}}$$

其中：

$$\left.\begin{aligned}
\frac{\partial v(k)}{\partial\hat{a}_i}&=y(k-i)-\sum_{j=1}^{n}\hat{c}_j\frac{\partial v(k-j)}{\partial\hat{a}_i}\\
\frac{\partial v(k)}{\partial\hat{b}_i}&=-u(k-i)-\sum_{j=1}^{n}\hat{c}_j\frac{\partial v(k-j)}{\partial\hat{b}_i}\\
\frac{\partial v(k)}{\partial\hat{c}_i}&=-v(k-i)-\sum_{j=1}^{n}\hat{c}_i\frac{\partial v(k-j)}{\partial\hat{c}_j}
\end{aligned}\right\}\quad (*)$$

可以看出上面三个等式为差分方程，这些差分方程的初始条件为0，可以求解这些差分方程，分别求出v(k)关于 $\hat{a}_1,\cdots,\hat{a}_n,\hat{b}_0,\cdots,\hat{b}_n,\hat{c}_1,\cdots,\hat{c}_n$ 的全部偏导数。

再由向量 $\partial J/\partial\hat{\theta}$ 对参数向量$\hat{\theta}$ 求偏导数，得到

$$\frac{\partial^2 J}{\partial\hat{\theta}\partial\hat{\theta}^T}=\sum_{k=n+1}^{n+N}\frac{\partial v(k)}{\partial\hat{\theta}}\frac{\partial v(k)}{\partial\hat{\theta}^T}+\sum_{k=n+1}^{n+N}v(k)\frac{\partial^2 v(k)}{\partial\hat{\theta}\partial\hat{\theta}^T}$$

因为v(k)是个小量， $\sum_{k=n+1}^{n+N}v(k)\frac{\partial^2 v(k)}{\partial\hat{\theta}\partial\hat{\theta}^T}$ 可以忽略。

$$\frac{\partial^2 J}{\partial\hat{\theta}\partial\hat{\theta}^T}\approx\sum_{k=n+1}^{n+N}\frac{\partial v(k)}{\partial\hat{\theta}}\frac{\partial v(k)}{\partial\hat{\theta}^T}$$

(4) 按照Newton-Raphson法计算：

$$\hat{\theta}(k+1)=\hat{\theta}(k)-\left(\frac{\partial^2 J}{\partial\hat{\theta}\partial\hat{\theta}^T}\right)^{-1}\frac{\partial J}{\partial\hat{\theta}}\bigg|_{\hat{\theta}=\hat{\theta}(k)}$$

(5) 重复(2)至(4)的计算步骤，迭代求新的参数估计值 $\hat{\theta}(k+1)$，直至v(k)方差的相对误差小于某个正小数，所得到的参数估计值就是极大似然估计值。