第2章：估计理论基础

统计推断：估计和假设检验
估计理论：经典估计，贝叶斯估计
估计方法：点估计和区间估计

估计器

$$\hat{\theta} = g(x(0), x(1), \cdots x(N-1))$$

a. $\hat{\theta}$ 是一个随机变量

b. 估计器设计依赖于样本数据的概率密度函数(PDF)的假设

例

$$x(n) = A + W(n) \quad n=0,1,2,\ldots N-1$$

$W(n)$ 是零均值白噪声，估计A

一个直观的估计器为 $\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x(n)$

a. 这个估计器怎样接近于真实值A.

b. 有没有更好的估计器，怎样设计好的估计器？

注：对于类似问题，可以由传统估计理论设计方法
也可以通过机器学习设计方法
本课程注重前者，针对不同应用对象，两者各有优劣。

估计器性能的其他描述

- 无偏估计：若估计器满足：$E(\hat{\theta})=\theta$ 则称为无偏估计，

  上例 $\hat{A}=\frac{1}{N}\sum_{n=0}^{H-1}x(n)$ 是无偏的。

  渐进无偏估计器：$\lim_{N\to\infty} E(\hat{\theta})=\theta$ 。

- 有偏估计：对于不满足无偏估计的估计器，

  定义：$b(\theta)=E(\hat{\theta})-\theta$ 为估计的偏。

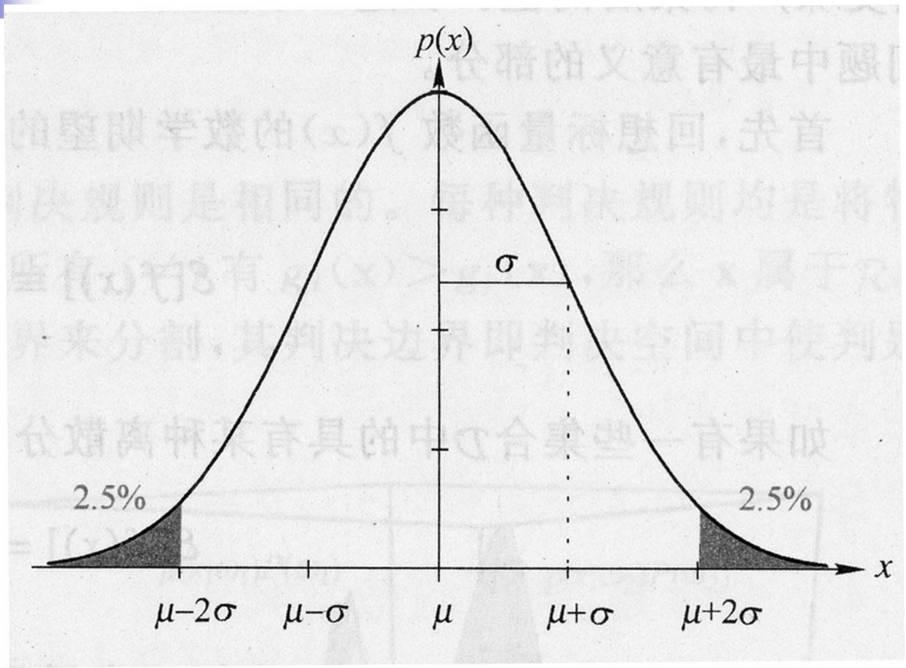
- 最小均方准则：$mse(\hat{\theta})=E[(\hat{\theta}-\theta)^2]=\mathrm{var}(\hat{\theta})+b^2(\theta)$ 。

  在假设待估计量是确定性变量时，在很多情况下，最小均方估计是不可实现

- 一致估计：随着 N 增加，估计得到改善，N 趋于无穷时，估计器的均方误差趋于零。相当于估计的偏和估计误差均趋于零。对于无偏估计，估计的方差趋于零。

- 最小方差无偏估计器（MVU）：设计一个无偏估计器，令估计方差 $\mathrm{var}(\hat{\theta})$ 最小

估计器性能的方差描述



$$P\left\{A-2\sigma \le \hat{A} \le A+2\sigma\right\}$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\int_{A-2\sigma}^{A+2\sigma} e^{-\frac{(x-A)^2}{2\sigma^2}}dx = G(2)-G(-2)\approx 0.95$$

几个常用估计量

均值估计 $$\hat{\mu}_x = \frac{1}{N}\sum_{n=0}^{N-1} x(n)$$

方差估计 $$\hat{\sigma}_x^2 = \frac{1}{N}\sum_{n=0}^{N-1}\left(x(n)-\hat{\mu}_x\right)^2$$

自相关估计 $$\hat{r}_x(l) = \begin{cases} \dfrac{1}{N}\sum_{n=l}^{N-1} x(n)x^*(n-l) & 0 \le l \le N-1 \\ \hat{r}_x^*(-l) & -(N-1) \le l < 0 \\ 0 & |l| \ge N \end{cases}$$

互相关估计

$$\hat{r}_{xy}(l) = \frac{1}{N}\sum_{n=l}^{N-1} x(n)y^*(n-l) \qquad 0 \le l \le N-1$$

$$\hat{r}_{yx}(l) = \frac{1}{N}\sum_{n=l}^{N-1} y(n)x^*(n-l) \qquad 0 \le l \le N-1$$

$$\hat{r}_{xy}(l) = \hat{r}_{yx}^*(-l) \qquad -N+1 \le l < 0$$

$$\hat{r}_{yx}(l) = \hat{r}_{xy}^*(-l) \qquad -N+1 \le l < 0$$

Cramer-Rao下界

最小方差无偏估计器（MVU），它的最好估计性能.

定理一：假设 PDF $p(\mathbf{x};\theta)$ 满足规则性条件：$E\left[\dfrac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right]=0$

则任意无偏估计器的方差满足：$\operatorname{var}(\hat{\theta}) \geq \dfrac{1}{-E\left[\dfrac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right]}$

注：本讲义对经典估计部分如下表示等价

$$p_x(\mathbf{x}|\theta)=p_x(\mathbf{x};\theta)$$

（定理一继续）

当且仅当 $\dfrac{\partial \ln p(\boldsymbol{x};\theta)}{\partial\theta} = I(\theta)\cdot(g(\boldsymbol{x})-\theta)$ 时，一个估计器可以达到下界值

这里 $I(\theta) = -E\left[\dfrac{\partial^2 \ln p(\boldsymbol{x};\theta)}{\partial\theta^2}\right]$，$g$ 是一个函数，

估计器 $\hat{\theta} = g(\boldsymbol{x})$ 是 MVU 估计器，且满足 $\operatorname{var}(\hat{\theta}) = \dfrac{1}{I(\theta)}$，

$I(\theta)$ 称为 Fisher 信息函数。

定理2.2.2 克拉美-罗下界(Cramer-Rao)的向量情况，假设PDF满足规则条件

无偏估计器$\hat{\boldsymbol{\theta}}$的协方差矩阵满足：

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathbf{I}^{-1}(\boldsymbol{\theta}) \succ= \mathbf{0}$$

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = -E\left[\frac{\partial^2 \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]$$

$\mathbf{I}(\boldsymbol{\theta})$ 为 Fisher 信息矩阵

当下界可达时，估计器第i个分量$\hat{\theta}_i$的方差为

$$Var(\hat{\theta}_i) = \left[\mathbf{C}_{\hat{\boldsymbol{\theta}}}\right]_{ii} = \left[\mathbf{I}^{-1}(\theta)\right]_{ii}$$

例：由一组观测值 $x(n)=A+w(n)$ n=0,1,2,…N-0,

$w(n)$ 是 WGN 且方差为 σ^2，均值为 0，估计未知量 A.

由于 w(n)是 WGN (高斯白噪声), 故有:

$$p(\boldsymbol{x};A)=\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x(n)-A)^2\right]$$

$$\frac{\partial\ln p(\boldsymbol{x};A)}{\partial A}=\frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}(x(n)-A)\right]$$

$$\frac{\partial^2\ln p(\boldsymbol{x};A)}{\partial A^2}=-\frac{N}{\sigma^2},\quad I(A)=-E\left[\frac{\partial^2\ln p(\boldsymbol{x};A)}{\partial A^2}\right]=\frac{N}{\sigma^2}$$

再由
$$\frac{\partial \ln p(\mathbf{x};A)}{\partial A} = \frac{1}{\sigma^2}\left(\sum_{n=0}^{N-1}(x(n)-A)\right)$$
$$= \frac{N}{\sigma^2}\left[\frac{1}{N}\sum_{n=0}^{N-1}x(n) - A\right] = I(A)(g(\mathbf{x}) - A)$$

故由定理 1 得

$$\hat{A} = g(\mathbf{x}) = \frac{1}{N}\sum_{n=0}^{N-1}x(n) \qquad Var(\hat{A}) = \frac{1}{I(A)} = \frac{\sigma^2}{N}$$

最大似然估计（MLE）

最实用的估计器，有良好的渐近特性

将 PDF $p(\mathbf{x};\theta)$ 中的 $\mathbf{x}$ 固定，考虑 θ 变化的影响，

这时将 $p(\mathbf{x};\theta)$ 称为似然函数（Likelihood）。

或将 $\ln p(\mathbf{x};\theta)$ 称为对数似然函数，

当 $\theta=\hat{\theta}$ 时，（对数）似然函数最大，得到估计器 $\hat{\theta}=g(\mathbf{x})$

$$\left.\frac{\partial p(\mathbf{x};\theta)}{\partial\theta}\right|_{\hat{\theta}}=0 \quad 或 \quad \left.\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right|_{\hat{\theta}}=0$$

例 $x(n) = A + w(n)$

n=0,1,2,…,N-1. $w(n)$为 WGN，方差为σ^2，估计 A

$$p(\mathbf{x};A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x(n)-A)^2\right]$$

$$\frac{\partial \ln p(\mathbf{x};A)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x(n)-A)\bigg|_{A=\hat{A}} = 0$$

得 $\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x(n)$

例 2.3.3 设信号 $x(n)$ 仅取 1 和 0 两个值

利用计数器得到观测量 $y=\sum_{n=0}^{N-1}x(n)$

$x(n)$ 取 1 的概率 π 利用 y 求 π 的 MLE

$x(n)$ 服从伯努力分布（Bernoulli)

$$p(x|\pi)=\pi^{x}(1-\pi)^{1-x}$$

y 满足二项分布（Binomial） 设 y 已确定，似然函数

$$L(\pi|y,N)=p(y|\pi,N)=\binom{N}{y}\pi^{y}(1-\pi)^{N-y}$$

解如下方程

$$\frac{\partial\ln L(\pi|y,N)}{\partial\pi}=\frac{\partial\left\{\ln\binom{N}{y}+y\ln\pi+(N-y)\ln(1-\pi)\right\}}{\partial\pi}$$

$$=\frac{y}{\pi}-\frac{N-y}{1-\pi}=0$$

解得 $\pi=y/N$

当θ下界可达时，MLE 得到 MVU 估计器。由

$$\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial \theta} = I(\theta)(g(\boldsymbol{x})-\theta)\Big|_{\theta=\hat{\theta}} = 0$$

$$\hat{\theta} = g(\boldsymbol{x})$$ MVU就是MLE

概率密度函数估计问题

对于向量信号 $\boldsymbol{x}(n)=[x_1(n),x_2(n),\cdots,x_M(n)]^T$

概率密度函数表示为

$$p_x(\boldsymbol{x}|\boldsymbol{\mu}_{\mathbf{x}},\mathbf{C}_{\mathbf{xx}})=\frac{1}{(2\pi)^{M/2}\det^{1/2}(\mathbf{C}_{\mathbf{xx}})}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathbf{x}})^T\mathbf{C}_{\mathbf{xx}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathbf{x}})\right)$$

.自变量向量 $\boldsymbol{x}=[x_1,x_2,\cdots,x_M]^T$

得到 N 个独立样本 $\boldsymbol{x}(n_0),\boldsymbol{x}(n_1),\cdots,\boldsymbol{x}(n_{N-1})$

估计 $\boldsymbol{\mu}_{\mathbf{x}}$，$\mathbf{C}_{\mathbf{xx}}$

似然函数为

$$L(\boldsymbol{\mu}_{\mathbf{x}},\mathbf{C}_{\mathbf{xx}}|X)=\prod_{i=0}^{N-1}p_x(x(n_i)|\boldsymbol{\mu}_{\mathbf{x}},\mathbf{C}_{\mathbf{xx}})$$

这里 $X=[x(n_0),\quad x(n_1),\quad \cdots,\quad x(n_{N-1})]$

$$\log L(\boldsymbol{\mu}_{\mathbf{x}},\mathbf{C}_{\mathbf{xx}}|X)$$

$$=-\frac{NM}{2}\log 2\pi-\frac{N}{2}\log|\mathbf{C}_{\mathbf{xx}}|-\frac{1}{2}\sum_{i=0}^{N-1}(x(n_i)-\boldsymbol{\mu}_{\mathbf{x}})^T\mathbf{C}_{\mathbf{xx}}^{-1}(x(n_i)-\boldsymbol{\mu}_{\mathbf{x}})$$

对 $\boldsymbol{\mu}_{\mathbf{x}},\mathbf{C}_{\mathbf{xx}}$ 求极大值 得到

$$\hat{\boldsymbol{\mu}}_{\mathbf{x}}=\frac{1}{N}\sum_{i=0}^{N-1}x(n_i)$$

$$\hat{\mathbf{C}}_{\mathbf{xx}}=\frac{1}{N}\sum_{i=0}^{N-1}(x(n_i)-\hat{\boldsymbol{\mu}}_{\mathbf{x}})(x(n_i)-\hat{\boldsymbol{\mu}}_{\mathbf{x}})^T$$

MLE渐近特性

如果 PDF $p(x;\boldsymbol{\theta})$ 满足规则性条件，未知参数$\boldsymbol{\theta}$的 MLE 估计渐近于如下分布：

$$\hat{\theta} \sim N(\theta, I^{-1}(\theta))$$

这里$I(\theta)$是 Fisher 信息矩阵，且在θ的真值处取值，

<u>MLE逼近于一个无偏的，最小方差可达的MVU估计器</u>

<u>对于一般的**PDF, MLE**可以通过迭代计算</u>

2.4 Bayesian估计

与经典的确定性参数的估计方法不同.

Bayesian 估计假设所估计的参数 θ 是一个随机变量.

估计的是它的一次实现的值

参数 θ 本身满足概率密度函数，$\theta \sim p(\theta)$，

观测数据与参数 θ 的联合概率密度函数为

$$p(\boldsymbol{x},\theta) = p(\theta)p(\boldsymbol{x} \mid \theta) = p(\theta \mid \boldsymbol{x})p(\boldsymbol{x})$$

最小均方Bayesian估计

这里θ是随机量，它是概率空间的一个分量，

均方误差定 义为

$$mse(\hat{\theta}) = \iint (\theta - \hat{\theta})^2 p(\mathbf{x}, \theta) d\mathbf{x} d\theta$$

$$= \int [\int (\theta - \hat{\theta})^2 p(\theta \mid \mathbf{x}) d\theta] p(\mathbf{x}) d\mathbf{x}$$

上式两边对$\hat{\theta}$求导，并交换积分和求导顺序，得

$$\frac{\partial mse(\hat{\theta})}{\partial \hat{\theta}} = \int \left[\frac{\partial}{\partial \hat{\theta}} \int (\hat{\theta} - \theta)^2 p(\theta | \mathbf{x}) d\theta \right] p(\mathbf{x}) d\mathbf{x}$$

为求最小均方估计$\hat{\theta}$，只需令上式为 0：

欲使$\dfrac{\partial mse(\hat{\theta})}{\partial \hat{\theta}}$为零，只需：

$$\frac{\partial}{\partial \hat{\theta}} \int (\theta - \hat{\theta})^2 p(\theta \mid \boldsymbol{x}) d\theta = 0$$

得： $-2\int (\theta - \hat{\theta}) p(\theta \mid \boldsymbol{x}) d\theta = 0$

$$\hat{\theta} = \int \theta p(\theta \mid \boldsymbol{x}) d\theta = E(\theta \mid \boldsymbol{x})$$

这是最小均方误差（MMSE）**Bayesian** 估计器。

在实际中，$p(\boldsymbol{x}|\theta)$ 是更容易获得的.

在计算时，经常利用关系式

$$p(\theta|\boldsymbol{x})=\frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})}=\frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

最小均方误差为

$$Bmse(\hat{\theta})=\iint(\theta-E(\theta|x))^2\,p(\mathbf{x},\theta)d\mathbf{x}d\theta=C_{\theta|x}$$

矢量情况

$$\hat{\boldsymbol{\theta}} = E(\hat{\boldsymbol{\theta}} \mid \boldsymbol{x}) = \begin{pmatrix} E(\theta_1 \mid \boldsymbol{x}) \\ E(\theta_2 \mid \boldsymbol{x}) \\ \vdots \\ E(\theta_{N-1} \mid \boldsymbol{x}) \end{pmatrix}$$

例

设观测值 $x(n)=A+w(n)$，$n=0,1,2,\cdots,N-1$

$w(n)$ 为 WGN，方差为 1

已知 A 满足高斯分布，$p_A(a)=\dfrac{1}{\sqrt{2\pi}}e^{-a^2/2}$

求参数 A 的 Bayesian 估计

解：先求 $p(\boldsymbol{x}\mid A)$，显然

$$p(\boldsymbol{x}|A)=\frac{1}{(2\pi)^{\frac{N}{2}}}\exp\left[-\frac{1}{2}\sum_{n=0}^{N-1}(x(n)-A)^2\right]$$

$$p(A|\mathbf{x}) = \frac{p(\mathbf{x}|A)p_A(A)}{\int p(\mathbf{x}|A)p_A(A)dA}$$

$$= \frac{\dfrac{1}{(2\pi)^{\frac{N}{2}}}\exp\left[-\dfrac{1}{2}\sum_{n=0}^{N-1}(x(n)-A)^2\right]\dfrac{1}{\sqrt{2\pi}}e^{-A^2/2}}{\int_{-\infty}^{+\infty}\dfrac{1}{(2\pi)^{\frac{N}{2}}}\exp\left[-\dfrac{1}{2}\sum_{n=0}^{N-1}(x(n)-A)^2\right]\dfrac{1}{\sqrt{2\pi}}e^{-A^2/2}dA}$$

$$= \frac{(N+1)^{1/2}}{(2\pi)^{1/2}}\exp\left[-\frac{1}{2}(N+1)(A-\frac{N\overline{X}}{N+1})^2\right]$$

因此 $\hat{A} = E(\theta|\mathbf{x}) = \int \theta p(\theta|\mathbf{x})d\theta = \dfrac{N\overline{X}}{N+1} = \dfrac{1}{N+1}\sum_{n=0}^{N-1}x(n)$

随机参数估计的Cramer－Rao界

$$Var\left(\hat{\theta}\right) \geq -\frac{\left(1+\frac{\partial b(\theta)}{\partial \theta}\right)^2}{E\left[\frac{\partial^2 \ln p(\boldsymbol{x}|\theta)}{\partial \theta^2}+\frac{\partial^2 \ln p(\theta)}{\partial \theta^2}\right]}$$

·高斯情况

如果 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 是联合高斯，$\boldsymbol{x}$ 是 $k\times 1$，$\boldsymbol{y}$ 是 $\ell\times 1$ 矢量，

均值矢量为 $[E(\boldsymbol{x}),E(\boldsymbol{y})]^T$，

分块协方差矩阵 $C=\begin{pmatrix} C_{xx}, & C_{xy} \\ C_{yx}, & C_{yy} \end{pmatrix}$

$$p(x,y)=\frac{1}{(2\pi)^{\frac{k+l}{2}}[\det(C)]^{\frac{1}{2}}}\exp\left\{-\left[\frac{1}{2}\begin{bmatrix}\boldsymbol{x}-E(\boldsymbol{x})\\ \boldsymbol{y}-E(\boldsymbol{y})\end{bmatrix}^T C^{-1}\begin{bmatrix}\boldsymbol{x}-E(\boldsymbol{x})\\ \boldsymbol{y}-E(\boldsymbol{y})\end{bmatrix}\right]\right\}$$

则条件 PDF： $p(\mathbf{y}\,|\,\mathbf{x})$ 也是高斯的，且有：

$$E(\mathbf{y}\,|\,\mathbf{x}) = E(\mathbf{y}) + C_{yx}C_{xx}^{-1}(\mathbf{x} - E(\mathbf{x}))$$

$$C_{y|x} = C_{yy} - C_{yx}C_{xx}^{-1}C_{xy}$$

这里若 y 是待估计参数 θ， $\mathbf{x}$ 是数据矢量，

Baysian估计 $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}\,|\,\mathbf{x}) = E(\boldsymbol{\theta}) + C_{\theta x}C_{x}^{-1}(\mathbf{x} - E(\mathbf{x}))$

零均值情况 $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}\,|\,\mathbf{x}) = R_{\theta x}R_{x}^{-1}\mathbf{x}$

高斯情况有两个特点：

1.有闭式解 2. 估计值是观测矢量的线性函数.(线性估计)

一般Bayesian估计

设 $\varepsilon = \theta - \hat{\theta}$ 表示估计误差，令 $C(\varepsilon)$ 为消耗函数，

定义：$R = E[C(\varepsilon)]$ 为 Beyes 风险函数。

令 Beyes 风险函数最小，得到各种贝叶斯估计。

1．Bayesian MSE　$C(\varepsilon) = \varepsilon^2$　方差最小准则的 Bayesian 估计

2．绝对误差准则　$C(\varepsilon) = |\varepsilon|$

3．"命中或错过"准则 "Hit-or-Miss"

$$C(\varepsilon) = \begin{cases} 0 & |\varepsilon| < \delta \\ 1 & |\varepsilon| > \delta \end{cases}$$

2． 的解是后验中值估计，

$\hat{\theta}$满足： $\int_{-\infty}^{\hat{\theta}} p(\theta \mid \boldsymbol{x}) d\theta = \int_{\hat{\theta}}^{+\infty} p(\theta \mid \boldsymbol{x}) d\theta$

3． 的解是最大后验概率（MAP）估计器

$$\hat{\theta} = \arg\max \ p(\theta \mid \boldsymbol{x})$$

再由： $p(\theta \mid \boldsymbol{x}) = \dfrac{p(\boldsymbol{x} \mid \theta) p(\theta)}{p(\boldsymbol{x})}$

得 $\hat{\theta} = \arg\max p(\boldsymbol{x} \mid \theta) p(\theta)$

或 $\hat{\theta} = \arg\max[\ln p(\boldsymbol{x} \mid \theta) + \ln p(\theta)]$

例 设观测值 $x(n)=A+w(n)$，$n=0,1,2,\cdots,N-1$，

$w(n)$ 为 WGN，方差为 σ_w^2，且已知 A 满足高斯分布，

$$p_A(A)=\frac{1}{\sqrt{2\pi\sigma_A^2}}e^{\frac{-(A-\mu_A)^2}{2\sigma_A^2}},$$

求参数 A 的 MAP Bayesian 估计。

解

$$p(\mathbf{x}|A)=\frac{1}{(2\pi\sigma_w^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma_w^2}\sum_{n=0}^{N-1}(x(n)-A)^2\right]$$

$$p(\mathbf{x}|A)P_A(A)=\frac{1}{(2\pi\sigma_w^2)^{\frac{N}{2}}}\frac{1}{(2\pi\sigma_A^2)^{\frac{1}{2}}}\exp\left[-\frac{1}{2\sigma_w^2}\sum_{n=0}^{N-1}(x(n)-A)^2\right]\exp\left[-\frac{1}{2\sigma_A^2}(A-\mu_A)^2\right]$$

上式两边取对数，并求最大值点，解得 A 的 MAP 估计为

$$\hat{A}_{map} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_w^2 / N} \frac{1}{N} \sum_{n=0}^{N-1} x(n) + \frac{\sigma_w^2 / N}{\sigma_A^2 + \sigma_w^2 / N} \mu_A$$

当 $N \to \infty$时，$\hat{A}_{map} \to \frac{1}{N} \sum_{n=0}^{N-1} x(n)$

2.5 线性贝叶斯估计器

由数据集 $\{x(0), x(1), \cdots, x(N-1)\}$ 估计标量参数 θ

如果将估计限制在一个线性估计器

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x(n) + a_N$$

选择系数集 $\{a_n, n = 0, \cdots N\}$，使 Bayesian MSE 最小，即

$$\min_{\{a_n\}} \left\{mse(\hat{\theta})\right\} = E[(\theta - \hat{\theta})^2] = \iint (\theta - \hat{\theta})^2 p(\mathbf{x}, \theta) dx d\theta$$

线性估计器

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x(n) + a_N$$

$$= E(\theta) + C_{x\theta}^T C_{xx}^{-1}(\boldsymbol{x} - E(\boldsymbol{x}))$$

若有 $E(\theta) = 0$， $E(x(n)) = 0$，则.

$$\begin{cases} \boldsymbol{a} = R_{xx}^{-1}\boldsymbol{r}_{x\theta} \\ \hat{\theta} = \boldsymbol{r}_{x\theta}^T R_{xx}^{-1}\boldsymbol{x} \\ Bmse(\hat{\theta}) = \sigma_\theta^2 - \boldsymbol{r}_{x\theta}^T R_{xx}^{-1}\boldsymbol{r}_{x\theta} \end{cases}$$

- 这组关系式可以直接联系到Wiener滤波问题
- 线性Bayesian估计与高斯分布下的Bayesian估计是一致的，在高斯分布下，线性估计可达最优

附注

- 高斯分布在随机信号处理中有特殊的地位，高斯分布时，线性估计是真正的最优估计，因此，对于高斯分布，线性系统是适用的。这些概念推广到时间波形的估计也是适用的。
- 对于非高斯分布，真正的最优系统是非线性系统，当线性系统的简化实现不能满足性能要求时，人们要回到贝叶斯估计的源头，寻找对“后验期望”或“最大后验概率”的更加精确的逼近实现。

附注（续）

- 建立在均方误差最小和“Hit－Miss”准则下的“后验期望”和“最大后验概率”方法只是贝叶斯理论的两个常用准则下的结果，在不同应用中，利用对专门问题的研究，提出新的风险函数和误差准则，发展新的贝叶斯估计方法，仍是将统计方法用于各种新问题时，一个有价值的研究方向。

2.6 最小二乘估计

最小二乘问题（LS）

通过解方程组引入最小二乘问题，

设有一线性方程组

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

当方程组数目 M 不等于变量个数 N 时，求最小二乘解

M>N 时：找到一个 $\boldsymbol{x}$，使如下二乘误差最小

$$\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 = (\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^H(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) \stackrel{\triangle}{=} J(\boldsymbol{x})$$

称为最小二乘解。

$M>N$:最小二乘的过确定问题。

令 $$\frac{\partial J(\boldsymbol{x})}{\partial \boldsymbol{x}^*}=\frac{\partial\left(\boldsymbol{b}^H\boldsymbol{b}-\boldsymbol{b}^H A\boldsymbol{x}-\boldsymbol{x}^H A^H\boldsymbol{b}+\boldsymbol{x}^H A^H A\boldsymbol{x}\right)}{\partial \boldsymbol{x}^*}$$

$$=-2A^H\boldsymbol{b}+2A^H A\boldsymbol{x}=\mathbf{0}$$

如果 $A^H A$ 可逆，得到:

$$\boldsymbol{x}=(A^H A)^{-1}A^H\boldsymbol{b}$$

在 $M<N$ 时，方程组有若干解，取使 $\boldsymbol{x}^H\boldsymbol{x}$ 最小的解。

$$\boldsymbol{x}=A^H(AA^H)^{-1}\boldsymbol{b}$$

$M<N$:最小二乘的欠确定问题。

线性模型估计问题

线性模型： $\boldsymbol{s} = \mathbf{A}\boldsymbol{\theta}$ 估计 $\boldsymbol{\theta}$

允许模型存在误差，且误差平方和最小：

$$\boldsymbol{s} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{e}$$

解为：

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\boldsymbol{s}$$

误差和：

$$\begin{aligned}
J_{\min} &= J(\hat{\boldsymbol{\theta}}) = \boldsymbol{e}^T\boldsymbol{e} \\
&= \left[\boldsymbol{s} - \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\boldsymbol{s}\right]^T\left[\boldsymbol{s} - \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\boldsymbol{s}\right] \\
&= \boldsymbol{s}^T\left[\mathbf{I} - \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\right]^T\boldsymbol{s} \\
&= \boldsymbol{s}^T\left(\boldsymbol{s} - \mathbf{A}\hat{\boldsymbol{\theta}}\right) = \boldsymbol{s}^T\boldsymbol{s} - \boldsymbol{s}^T\mathbf{A}\hat{\boldsymbol{\theta}}
\end{aligned}$$

例 假设有一个多变量系统，输出为

$$s(n) = f\left(v_0(n), v_1(n), \cdots v_{M-1}(n)\right)$$

用线性模型逼近

$$s(n) = \theta_0 v_0(n) + \theta_1 v_1(n) + \cdots + \theta_{M-1} v_{M-1}(n)$$

观测到时刻 $n=0$ 至 $n=N-1$ 的 $s(n)$ 和 $v_1(n), v_2(n), \cdots v_M(n)$

$$\mathbf{A} = \begin{bmatrix} v_0(0) & v_1(0) & \cdots & v_{M-1}(0) \\ v_0(1) & v_1(1) & \cdots & v_{M-1}(1) \\ \vdots & \vdots & \cdots & \vdots \\ v_0(N-1) & v_1(N-1) & \cdots & v_{M-1}(N-1) \end{bmatrix}$$

正则化最小二乘估计

$$J(\boldsymbol{\theta})=\sum_{n=0}^{N-1}e^2(n)+\lambda\|\boldsymbol{\theta}\|^2=\|\mathbf{A}\boldsymbol{\theta}-\mathbf{s}\|^2+\lambda\|\boldsymbol{\theta}\|^2$$

$$=(\mathbf{A}\boldsymbol{\theta}-\mathbf{s})^H(\mathbf{A}\boldsymbol{\theta}-\mathbf{s})+\lambda\boldsymbol{\theta}^T\boldsymbol{\theta}$$

正则化LS解为

$$\hat{\boldsymbol{\theta}}=\left(\mathbf{A}^H\mathbf{A}+\lambda\mathbf{I}\right)^{-1}\mathbf{A}^H\boldsymbol{s}$$