E11: General Bayesian Estimators

Previously, we derived the Bayesian estimator by minimizing $E_{p(x,\theta)}[\|\theta-\hat{\theta}\|_2^2]$. Let $\varepsilon = \theta - \hat{\theta}$ denote the estimation error for a particular (x, θ) pair. Let $C(\varepsilon) = \|\varepsilon\|_2^2$ be the cost function, then the Bayes risk we used was $\mathcal{R} = E[C(\varepsilon)]$.

In practice, we could choose to use other cost functions (instead of the quadratic one). For instance

Absolute error cost: $C(\varepsilon) = \|\varepsilon\|_1$

Hit-or-Miss cost: $C(\varepsilon) = \begin{cases} 0, & \|\varepsilon\| < \delta \\ 1, & \|\varepsilon\| \geq \delta \end{cases}$

In general,
$$\mathcal{R} = E[C(\varepsilon)]$$
$$= \iint C(\theta - \hat{\theta})\, p(x, \theta)\, dx\, d\theta$$
$$= \int \left[\int C(\theta - \hat{\theta})\, p(\theta | x)\, d\theta \right] p(x)\, dx$$

As we did for quadratic cost, if we can find an estimator $\hat{\theta}$ that minimizes the inner integral $\forall x$, then that is a global optimizer.

For simplicity, let's consider the scalar θ case. Then given x, $\hat{\theta}$ is a scalar.

Absolute Error: $\varrho(\hat{\theta}) = \int |\theta - \hat{\theta}|\, p(\theta|x)\, d\theta$

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

Recall Leibnitz's rule:

$$\frac{\partial}{\partial u} \int_{\phi_1(u)}^{\phi_2(u)} h(u,v)\, dv = \int_{\phi_1(u)}^{\phi_2(u)} \frac{\partial h(u,v)}{\partial u}\, dv + \frac{\partial \phi_2(u)}{\partial u} h(u, \phi_2(u)) - \frac{d\phi_1(u)}{du} h(u, \phi_1(u))$$

Here, $h(\hat{\theta}; \theta) = (\hat{\theta} - \theta)\, p(\theta|x)$:

$$h(u, \phi_2(u)) = h(\hat{\theta}, \hat{\theta}) = (\hat{\theta} - \hat{\theta})\, p(\hat{\theta}|x) = 0$$

$$\frac{d\phi_1(u)}{du} = 0 \quad \text{(since lower limit is } -\infty\text{)}$$

Then $$\frac{d\varrho(\hat{\theta})}{d\hat{\theta}} = \int_{-\infty}^{\hat{\theta}} p(\theta|x)\, d\theta - \int_{\hat{\theta}}^{\infty} p(\theta|x)\, d\theta = 0$$

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

$\Leftrightarrow$ $\hat{\theta}$ is the <u>median</u> of the posterior pdf.

Hit-or-Miss Error: In the scalar θ case, we have

$C(\varepsilon) = 1$ for $|\varepsilon| \geq \delta$ and ~~$\leq$~~ 0 otherwise:

$$\varrho(\hat{\theta}) = \int_{-\infty}^{\hat{\theta} - \delta} 1 \cdot p(\theta|x)\, d\theta + \int_{\hat{\theta} + \delta}^{\infty} 1 \cdot p(\theta|x)\, d\theta$$

since $\int_{-\infty}^{\infty} p(\theta|x)\, d\theta = 1$

$$= 1 - \int_{\hat{\theta} - \delta}^{\hat{\theta} + \delta} p(\theta|x)\, d\theta$$

$$\hat{\theta}_{HM} = \underset{\hat{\theta}}{\operatorname{argmin}}\ \rho_{HM}(\hat{\theta}) = \underset{\hat{\theta}}{\operatorname{argmax}} \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|x)\,d\theta$$

When δ is selected to be arbitrarily small, the optimizer is found to be the global maximizer of the posterior. This estimator is called the Maximum a Posteriori (MAP) estimator.

$$\hat{\theta}_{MAP} = \underset{\hat{\theta}}{\operatorname{argmin}}\ \rho_{HitMiss}(\hat{\theta}) = \underset{\hat{\theta}}{\operatorname{argmax}} \lim_{\delta \to 0^+} \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|x)\,d\theta$$

Summary: In Bayesian estimation

Minimum MSE $\Rightarrow$ Mean of the posterior pdf

Minimum Abs. Error $\Rightarrow$ Median of the posterior pdf

Minimum HitMiss Error $\Rightarrow$ Mode of the posterior pdf

For some posterior pdfs, all three estimators are identical.

Ex) Gaussian Posterior

$$p(\theta|x) = \frac{1}{\sqrt{2\pi\sigma_{\theta|x}^2}}\, e^{-\frac{1}{2\sigma_{\theta|x}^2}(\theta-\mu_{\theta|x})^2}$$

Minimum MSE Estimators

The MMSE estimator was determined to be $E[\theta|x]$, and is also called the conditional mean estimator.

In the vector parameter case, for this estimator, the minimum Bayesian MSE is

$$B_{MSE}(\hat{\theta}_i) = \int [C_{\theta|x}]_{ii}\, p(x)\, dx \quad i = 1, 2, \ldots, p$$

where $C_{\theta|x} = E_{\theta|x}\left[(\theta - E[\theta|x])(\theta - E[\theta|x])^T\right]$

Ex) Bayesian Fourier Analysis

$x[n] = a\cos(2\pi f_0 n) + b\sin(2\pi f_0 n) + w[n]$, $n = 0, 1, \ldots, (N-1)$

where f_0 is a multiple of $1/N$ except 0 or $1/2$, and $w[n]$ is WGN with variance σ^2. Let $\theta = \begin{bmatrix} a \\ b \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \sigma_\theta^2 I)$.

θ and $w[n]$ are independent. $\|\theta\|_2 \sim$ Rayleigh

(This is a Rayleigh fading sinusoid, which is used to model a sinusoid that propagated through a dispersive medium.)

$$x = H\theta + w \quad \text{with} \quad H = \begin{bmatrix} 1 & 0 \\ \cos(2\pi f_0) & \sin(2\pi f_0) \\ \vdots & \vdots \\ \cos(2\pi f_0(N-1)) & \sin(2\pi f_0(N-1)) \end{bmatrix}$$

This is the Bayesian linear model and

$$\mu_\theta = 0, \quad C_\theta = \sigma_\theta^2 I, \quad C_w = \sigma^2 I$$

$$\Rightarrow \hat{\theta} = E[\theta|x] = \sigma_\theta^2 H^T (H\sigma_\theta^2 H^T + \sigma^2 I)^{-1} x$$

$$C_{\theta|x} = \sigma_\theta^2 I - \sigma_\theta^2 H^T (H\sigma_\theta^2 H^T + \sigma^2 I)^{-1} H\sigma_\theta^2$$

since $H^T H = \frac{N}{2} I$

$$\text{OR} \Rightarrow \hat{\theta} = \frac{\sigma^{-2}}{\sigma_\theta^{-2} + \frac{N}{2}\sigma^{-2}} H^T x \Rightarrow \hat{a} = \frac{2\sigma_\theta^2}{N + 2\sigma^2} \sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)$$

$$\hat{b} = \frac{2\sigma_\theta^2}{N + 2\sigma^2} \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)$$

Also $C_{\theta|x} = \frac{\sigma_\theta^2}{N+2\sigma^2} I$, so $B_{MSE}(\hat{a}) = BMSE(\hat{b}) = \frac{1}{(1/\sigma_\theta^2) + \frac{1}{(2\sigma^2/N)}}$.

The MMSE estimator commutes over affine transformations. Let $\alpha = A\theta + b$ where $A \sim r \times p$ and $b \sim r \times 1$ are known. Then $\hat{\alpha}_{MMSE} = E[\alpha|x] = E[A\theta + b|x] = AE[\theta|x] + b = A\hat{\theta}_{MMSE} + b$.

The MMSE estimator also has a form of additivity property for two independent data sets. Consider θ, x_1, x_2 where x_1 and x_2 are two independent data sets. Let $x \triangleq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ be the combined data set.

Then $$C_{xx}^{-1} = \begin{bmatrix} C_{x_1x_1} & 0 \\ \underbrace{0}_{\text{0's due to independence}} & C_{x_2x_2} \end{bmatrix}^{-1} = \begin{bmatrix} C_{x_1x_1}^{-1} & 0 \\ 0 & C_{x_2x_2}^{-1} \end{bmatrix}.$$

Also $C_{\theta x} = E\left[\theta \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T\right] = [C_{\theta x_1} \ \ C_{\theta x_2}]$. Consider $\hat{\theta} = E[\theta|x_1, x_2]$,

$$\begin{aligned}
\hat{\theta} = E[\theta|x] &= E[\theta] + C_{\theta x} C_{xx}^{-1}(x - E[x]) \\
&= E[\theta] + [C_{\theta x_1} \ \ C_{\theta x_2}] \begin{bmatrix} C_{x_1x_1}^{-1} & 0 \\ 0 & C_{x_2x_2}^{-1} \end{bmatrix} \begin{bmatrix} x - E[x_1] \\ x - E[x_2] \end{bmatrix} \\
&= \underbrace{E[\theta]}_{\text{Prior Estimate}} + \underbrace{C_{\theta x_1} C_{x_1x_1}^{-1}(x_1 - E[x_1])}_{\text{Estimate Contribution from Dataset \#1}} + \underbrace{C_{\theta x_2} C_{x_2x_2}^{-1}(x_2 - E[x_2])}_{\text{Estimate Contribution from Dataset \#2}}
\end{aligned}$$

MAP Estimators

$$\hat{\theta} = \arg\max_{\theta} p(\theta|x) = \arg\max_{\theta} \frac{p(x|\theta)p(\theta)}{p(x)}$$

$$= \arg\max_{\theta} p(x|\theta)p(\theta)$$ since $p(x)$ does not depend on θ.

$$= \arg\max_{\theta} \left[\ln p(x|\theta) + \ln p(\theta) \right]$$

Ex] Exponential pdf

Assume $p(x[n]|\theta) = \begin{cases} \theta e^{-\theta x[n]} & \text{if } x[n] \geq 0 \\ 0 & \text{if } x[n] < 0 \end{cases}$ and

all $x[n]$'s are conditionally iid. Then

$$p(\mathbf{x}|\theta) = \prod_{n=0}^{N-1} p(x[n]|\theta).$$

Assume $p(\theta) = \begin{cases} \lambda e^{-\lambda\theta} & \text{if } \theta \geq 0 \\ 0 & \text{if } \theta < 0 \end{cases}$. Then

$$g(\theta) = \ln p(x|\theta) + \ln p(\theta)$$

$$= \ln \left[\theta^N e^{-\theta \sum_{n=0}^{N-1} x[n]} \right] + \ln \left[\lambda e^{-\lambda\theta} \right]$$

$$= N \ln \theta - N\theta\bar{x} + \ln\lambda - \lambda\theta \quad \text{for } \theta \geq 0$$
$$(-\infty \text{ for } \theta < 0)$$

$$\frac{d g(\theta)}{d\theta} = \left. \frac{N}{\theta} - N\bar{x} - \lambda \right|_{\theta = \hat{\theta}_{MAP}} = 0 \Rightarrow \hat{\theta}_{MAP} = \frac{1}{\bar{x} + (\lambda/N)}$$

As $N \to \infty$, $\hat{\theta} \to 1/\bar{x}$. As $\lambda \to 0$ (prior $\to$ uniform), $\hat{\theta} \to 1/\bar{x}$ as well.

Note that as $p(\theta) \to$ uniform, $\hat{\theta}_{MAP} \to \hat{\theta}_{MLE}$, in general.

The posterior can be found to be

$$p(A|x) = \begin{cases} \dfrac{\frac{1}{\sqrt{2\pi\sigma^2/N}} e^{-\frac{1}{2\sigma^2/N}(A-\bar{x})^2}}{\int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi\sigma^2/N}} e^{-\frac{1}{2\sigma^2/N}(A-\bar{x})^2} dA} & \text{for } |A| \le A_0 \\ 0 & \text{for } |A| > A_0 \end{cases}$$

uniform in $[-A_0, A_0]$

$\hat{A}_{MAP} = \underset{A}{\operatorname{argmax}} \ln p(A|x)$. Clearly the denominator of the first component does not depend on A. The numerator is maximized at $\bar{x}$ $\Rightarrow$ $\hat{A}_{MAP} = \begin{cases} -A_0 & \text{if } \bar{x} < -A_0 \\ \bar{x} & \text{if } \bar{x} \in [-A_0, A_0] \\ A_0 & \text{if } \bar{x} > A_0 \end{cases}$

In this case, the MMSE estimator is hard to obtain analytically due to the normalizing integral expression in the denominator.

MAP Estimator for vector Parameters

Let $\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$ and consider the estimation of θ_1.

Then $p(\theta_1|x) = \int \cdots \int p(\theta|x)\, d\theta_2 \cdots d\theta_p$ and $\hat{\theta}_1 = \underset{\theta_1}{\operatorname{argmax}}\, p(\theta_1|x)$ would minimize the average hit-miss risk function $\mathcal{R}_1 = E[C(\theta_1 - \hat{\theta}_1)]$ where the expectation is over $p(x, \theta_1)$. Note that

$$p(\theta_1|x) = \frac{p(x|\theta_1)p(\theta_1)}{p(x)} = \frac{\int \cdots \int p(x, \theta_2 \cdots \theta_p | \theta_1)\, d\theta_2 \cdots d\theta_p \int \cdots \int p(\theta)\, d\theta_2 \cdots d\theta_p}{p(x)}$$

With this extension to vector case, the primary advantage

Instead, if we expand the Hit/Miss cost function to use a vector norm (as shown originally), we get

$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|x)$ for $\theta \in \mathbb{R}^p$ become a multidimensional optimization problem. The two extensions are not equivalent in general.

Ex) DC level in WGN, unknown variance

$x[n] = A + w[n] \quad n = 0, 1, \ldots, (N-1) \quad \theta = \begin{bmatrix} A \\ \sigma^2 \end{bmatrix}$ $w[n] \sim$ WGN

Then $p(x|A,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2}$

Assume $p(A,\sigma^2) = p(A|\sigma^2)p(\sigma^2)$

$$= \frac{1}{\sqrt{2\pi\alpha\sigma^2}} e^{-\frac{1}{2\alpha\sigma^2}(A-\mu_A)^2} \quad \frac{\lambda e^{-\lambda/\sigma^2}}{\sigma^4}$$

In the prior $\sigma^2_{A|\sigma^2} = \alpha\sigma^2$, so this prior scales the variance of A given σ^2 proportional to σ^2.

$$g(A,\sigma^2) = p(A|A,\sigma^2)\,p(A,\sigma^2) = p(x|A,\sigma^2)\,p(A|\sigma^2)p(\sigma^2)$$

Let $h(A;\sigma^2) = p(\cancel{x}|A,\sigma^2)\,p(A|\sigma^2)$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}\sqrt{2\pi\sigma_A^2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]} \quad e^{-\frac{1}{2}Q(A)}$$

where $Q(A) = \frac{1}{\sigma^2_{A|x}}(A-\mu_{A|x})^2 - \frac{\mu^2_{A|x}}{\sigma^2_{A|x}} + \frac{\mu_A^2}{\sigma_A^2}$

with $\mu_{A|x} = \dfrac{N\bar{x}/\sigma^2 + \mu_A/\sigma_A^2}{N/\sigma^2 + 1/\sigma_A^2}$ and $\sigma^2_{A|x} = \dfrac{1}{N/\sigma^2 + 1/\sigma_A^2}$.

$$\hat{A}(\sigma^2) = \arg\max_A h(A;\sigma^2) = \arg\min_A Q(A) = \mu_{A|x}.$$

Since $\sigma_A^2 = \alpha\sigma^2$, we have:

$$\hat{A}(\sigma^2) = \frac{N\bar{x} + \mu_A/\alpha}{N + 1/\alpha} \quad \text{(does not depend on } \sigma^2\text{)}$$

$$\Rightarrow \hat{A}_{MAP} = \frac{\alpha N\bar{x} + \mu_A}{\alpha N + 1}$$

$$\hat{\sigma}^2_{MAP} = \arg\max_{\sigma^2} g(\hat{A}_{MAP}, \sigma^2) = \arg\max_{\sigma^2} h(\hat{A}_{MAP};\sigma^2)\, p(\sigma^2).$$

$$h(\hat{A},\sigma^2) = (2\pi\sigma^2)^{-N/2}(2\pi\sigma_A^2)^{-1/2} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]}\, e^{-\frac{1}{2}Q(\hat{A})}$$

where $Q(\hat{A}) = \dfrac{\mu_A^2}{\sigma_A^2} - \dfrac{\mu^2_{A|x}}{\sigma^2_{A|x}}$. With $\sigma_A^2 = \alpha\sigma^2$,

$$Q(\hat{A}) = \frac{1}{\sigma^2}\left[\frac{\mu_A^2}{\alpha} - \hat{A}^2\left(N + \frac{1}{\alpha}\right)\right] \triangleq \frac{\xi}{\sigma^2}. \text{ Then,}$$

$$g(\hat{A},\sigma^2) = (2\pi\sigma^2)^{-N/2}(2\pi\alpha\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n] - \frac{\xi}{2\sigma^2}}\, \frac{\lambda e^{-\lambda/\sigma^2}}{\sigma^4}$$

$$= \frac{c}{(\sigma^2)^{(N+5)/2}} e^{-a/\sigma^2} \quad \text{where } c \text{ is a constant, } a = \frac{1}{2}\sum_{n=0}^{N-1} x^2[n] + \frac{\xi}{2} + \lambda$$

$$\frac{\partial \ln g(\hat{A},\sigma^2)}{\partial\sigma^2} = -\frac{(N+5)/2}{\sigma^2} + \frac{a}{\sigma^4}\Big|_{\sigma^2 = \hat{\sigma}^2_{MAP}} = 0 \quad \text{which, after some work yields}$$

$$\Rightarrow \hat{\sigma}^2_{MAP} = \frac{2a}{N+5} = \ldots = \frac{1}{N+5}\sum_{n=0}^{N-1} x^2[n] - \frac{N\hat{A}^2}{N+5} + \frac{(\mu_A^2 - \hat{A}^2)}{(N+5)\alpha} + \frac{2\lambda}{N+5}.$$

If $N \to \infty$, then $\hat{A}_{MAP} \to \bar{x}$ and $\hat{\sigma}^2_{MAP} \to \frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\bar{x})^2$.
The data dominates the prior... This is true, in general, for MAP estimators, if as $N \to \infty$ we are getting "independent" evidence from each new sample.

Ex) Exponential pdf

Assume that $p(x[n]|\theta) = \begin{cases} \theta e^{-\theta x[n]} & \text{if } x[n] \geq 0 \\ 0 & \text{if } x[n] < 0 \end{cases}$

and $p(\theta) = \begin{cases} \lambda e^{-\lambda\theta} & \text{if } \theta \geq 0 \\ 0 & \text{if } \theta < 0. \end{cases}$ We have iid $\{x[0], \dots, x[N-1]\}$.

$\hat{\theta}_{MAP} = \dfrac{1}{\bar{x}+\lambda/N}$ was found earlier. Suppose we now want to estimate $\alpha = 1/\theta$. Recall that the MLE was "invariant" under invertible parameter transformations. So we might think that $\hat{\alpha}_{MAP}$ will be $1/\hat{\theta}_{MAP}$.

This is NOT the case

$$p(x[n]|\alpha) = \begin{cases} \frac{1}{\alpha} e^{-x[n]/\alpha} & \text{if } x[n] \geq 0 \\ 0 & \text{if } x[n] < 0 \end{cases}$$

Recall the fundamental theorem of probability

$$p_\alpha(\alpha) = \frac{p_\theta(\theta(\alpha))}{|d\alpha/d\theta|} = \begin{cases} \lambda e^{-\lambda/\alpha}/\alpha^2 & \text{if } \alpha \geq 0 \\ 0 & \text{if } \alpha < 0. \end{cases}$$

Then $g(\alpha) = \ln p(x|\alpha) + \ln p(\alpha)$

$$= \ln\left\{ \left(\frac{1}{\alpha}\right)^N e^{-\frac{1}{\alpha}\sum_{n=0}^{N-1} x[n]} \right\} + \ln \frac{\lambda e^{-\lambda/\alpha}}{\alpha^2}$$

$$= -N \ln \alpha - N \frac{\bar{x}}{\alpha} + \ln \lambda - \frac{\lambda}{\alpha} - 2 \ln \alpha$$

$$= -(N+2) \ln \alpha - \frac{N\bar{x} + \lambda}{\alpha} + \ln \lambda$$

$$\frac{dg}{d\alpha} = -\frac{N+2}{\alpha} + \frac{N\bar{x}+\lambda}{\alpha^2} \Bigg|_{\alpha = \hat{\alpha}_{MAP}} = 0 \Rightarrow \hat{\alpha}_{MAP} = \frac{N\bar{x}+\lambda}{N+2}$$

$$= \frac{N}{N+2}\bar{x} + \frac{\lambda}{N+2}$$

In contrast we had $1/\hat{\theta}_{MAP} = \bar{x} + \frac{\lambda}{N}$

$\therefore$ MAP estimator does <u>NOT</u> commute over nonlinear transformations, although it does over linear ones.

Performance Description

In general, given θ we observe data and estimate the parameter as $\hat{\theta}$ as a function of the data. Since the data is random, depending on the estimation method we obtain some $p(\hat{\theta}|\theta)$. If we let $\varepsilon = \theta - \hat{\theta}$, then the distribution (pdf) of ε is informative about the quality of an estimator. The Bayesian MSE is a measure of how wide spread ε is and how much it deviates from θ {E(var + bias^2)}, so it makes sense for Gaussian $p(\hat{\theta}|\theta)$.

- $\hat{\theta} = E[\theta|x]$, we have $\mathcal{E} = \theta - E[\theta|x]$.

$E_{x,\theta}[\mathcal{E}] = E_{x,\theta}[\theta - E[\theta|x]] = E_x[E_{\theta|x}(\theta) - E_{\theta|x}(\theta|x)]$

$= E_x[E[\theta|x] - E[\theta|x]] = 0$

$var(\mathcal{E}) = E_{x,\theta}[\mathcal{E}^2]$ since 0-mean

$= E_{x,\theta}[(\theta - \hat{\theta})^2] = B_{MSE}(\hat{\theta})$.

If $\mathcal{E}$ is Gaussian, then $\mathcal{E} \sim \mathcal{N}(0, B_{MSE}(\hat{\theta}))$.

DC level in WGN, Gaussian prior pdf

$$\hat{A} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \bar{x} + \frac{\sigma^2/N}{\sigma_A^2 + \sigma^2/N} \mu_A \quad \text{and} \quad B_{MSE} = \frac{1}{N/\sigma^2 + 1/\sigma_A^2}$$

r, $\mathcal{E} = A - \hat{A} \sim \mathcal{N}\left(0, \frac{1}{N/\sigma^2 + 1/\sigma_A^2}\right)$ since $\hat{A}$ depends

nearly on x, and x and A are jointly Gaussian.

As $N \to \infty$, the variance of $\mathcal{E} \to 0$, so the estimator

consistent in the Bayesian sense.

For vector θ, let $\mathcal{E} = \theta - \hat{\theta}$ where $\hat{\theta} = E[\theta|x]$.

$E_{x,\theta}[\mathcal{E}\mathcal{E}^T] = E_{x,\theta}[(\theta - E(\theta|x))(\theta - E(\theta|x))^T]$ The Bayesian MSE matrix

$[E[\theta|x]]_i = E[\theta_i|x] = \int \theta_i p(\theta|x) d\theta = \int \theta_i p(\theta_i|x) d\theta_i$

nds only on x. Hence $E[\theta|x]$ is a valid estimator.

t depends only on the data.

For the Bayesian linear model, we have

$$M_{\hat\theta} = E_{x,\theta}\left[(\theta - E(\theta|x))(\theta - E(\theta|x))^T\right]$$
$$= E_x\, E_{\theta|x}\left[(\theta - E(\theta|x))(\theta - E(\theta|x))^T\right]$$
$$= E_x(C_{\theta|x}) \quad \text{if } \theta \text{ and } x \text{ are jointly Gaussian}$$
$$= C_{\theta\theta} - C_{\theta x} C_{xx}^{-1} C_{x\theta} \quad \text{for the Bayesian linear model}$$
$$= C_\theta - C_\theta H^T (H C_\theta H^T + C_w)^{-1} H C_\theta \quad \text{woodbury identity}$$
$$= (C_\theta^{-1} + H^T C_w^{-1} H)^{-1}$$

Since for the Bayesian linear model $\varepsilon = \theta - \hat\theta$ is Gaussian, from $\varepsilon = \theta - \mu_\theta - C_\theta H^T (H C_\theta H^T + C_w)^{-1}(x - H\mu_\theta)$, we have $\varepsilon \sim \mathcal{N}(0, M_{\hat\theta})$.

Ex) Bayesian Fourier Analysis

Recall that $C_\theta = \sigma_\theta^2 I$, $C_w = \sigma^2 I$, $H^T H = \frac{N}{2} I$. Then

$M_{\hat\theta} = (C_\theta^{-1} + H^T C_w^{-1} H)^{-1} = \left(\frac{1}{\sigma_\theta^2} + \frac{N}{2\sigma^2}\right)^{-1} I$. Therefore,

$\varepsilon = \begin{bmatrix} \varepsilon_{\hat a} \\ \varepsilon_{\hat b} \end{bmatrix} \sim \mathcal{N}(0, M_{\hat\theta})$ where

$$M_{\hat\theta} = \left(\sigma_\theta^{-2} + \frac{N}{2\sigma^2}\right)^{-1} I.$$

Let $P = \Pr\{\varepsilon^T M_{\hat\theta}^{-1} \varepsilon \le c^2\}$. Since $u = \varepsilon^T M_{\hat\theta}^{-1}\varepsilon$ is a χ_2^2 random variable with pdf $p(u) = \begin{cases} \frac{1}{2} e^{-u/2} & \text{if } u \ge 0 \\ 0 & \text{if } u < 0 \end{cases}$,

$P = \int_0^{c^2} \frac{1}{2} e^{-u/2}\, du = 1 - e^{-c^2/2}$ is the probability that the error ε will be inside the ellipse $\varepsilon^T M_{\hat\theta}^{-1} \varepsilon \le c^2$ (circle in this case).

Thm 11.1. Performance of the MMSE Estimator for the Bayesian Linear Model.

If x can be modeled by the Bayesian linear model, the MMSE estimator is

$$\hat{\theta} = \mu_\theta + C_\theta H^T (H C_\theta H^T + C_w)^{-1}(x - H\mu_\theta)$$

$$= \mu_\theta + (C_\theta^{-1} + H^T C_w^{-1} H)^{-1} H^T C_w^{-1} (x - H\mu_\theta).$$

The performance of the estimator is, with $\varepsilon = \theta - \hat{\theta}$,

$$C_\varepsilon = E_{x,\theta}(\varepsilon \varepsilon^T)$$

$$= C_\theta - C_\theta H^T (H C_\theta H^T + C_w)^{-1} H C_\theta$$

$$= (C_\theta^{-1} + H^T C_w^{-1} H)^{-1}.$$

The error covariance matrix is also the minimum MSE matrix $M_{\hat{\theta}}$ where $[M_{\hat{\theta}}]_{ii} = [C_\varepsilon]_{ii} = B_{MSE}(\hat{\theta}_i)$.

Ex) Deconvolution

Let $x[n] = \sum_{m=0}^{n_s-1} h[n-m]\, s[m] + w[n] \qquad n = 0, 1, \dots, (N-1)$

Assume that $s[n]$ is a Gaussian process, $w[n]$ is WGN, and $h[n]$ is known. Then, we have $x = H\theta + w$, where

$$H = \begin{bmatrix} h[0] & 0 & \cdots & 0 \\ h[1] & h[0] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h[N-1] & h[N-2] & \cdots & h[N-n_s] \end{bmatrix}, \quad \theta = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[n_s-1] \end{bmatrix}$$

$s \sim N(0, C_s)$ where $[C_s]_{ij} = r_{ss}[i-j]$ and

$$\hat{s}_{MMSE} = C_s H^T (H C_s H^T + \sigma^2 I)^{-1} x$$

If $h[n]=\delta[n]$, then $H=I$ and $x=\theta+w$.

In that case $\hat{\theta}_{MMSE}=(H^T H)^{-1}H^T x = x$

In the Bayesian case for $H=I$, we have

$$\hat{s}_{BMSE} = C_s(C_s+\sigma^2 I)^{-1}x$$

Let $A = C_s(C_s+\sigma^2 I)^{-1}$ (this is called the Wiener filter).

For scalar $s[0]$ based on $x[0]$, we have

$$\hat{s}[0] = \frac{r_{ss}[0]}{r_{ss}[0]+\sigma^2}\, x[0]$$

For high SNR ($r_{ss}[0]/\sigma^2 \to \infty$), we have $\hat{s}[0] \to x[0]$.

For low SNR, $\hat{s}[0] \to 0$.

If $s[n] = -a[1]s[n-1]+u[n]$ (AR(1) process)

where $u[n]$ is WGN with variance σ_u^2, then

$$r_{ss}[k] = \frac{\sigma_u^2}{1-a^2[1]}(-a[1])^{|k|}$$

and $\hat{s} = C_s(C_s+\sigma^2 I)^{-1}x$ where $C_s = \begin{bmatrix} r_{ss}[0] & r_{ss}[1] \\ r_{ss}[1] & r_{ss}[0] \end{bmatrix}$.

Suggested Problems: 2, 8, 9, 11, 18