E10: The Bayesian Philosophy

So far we considered the parameter θ to be an unknown deterministic variable. In the Bayesian approach, we will take it to be a random variable. Consider the Bayesian MSE objective $B_{MSE}(\hat{\theta}) = \iint (\theta - \hat{\theta})^T (\theta - \hat{\theta}) p(x, \theta)\, dx\, d\theta$, in contrast with the classical $MSE(\hat{\theta}) = \int (\theta - \hat{\theta})^T (\theta - \hat{\theta}) p(x; \theta) dx$.

Using Bayes rule: $p(x, \theta) = p(\theta | x) p(x)$, we have

$B_{MSE}(\hat{\theta}) = \int \left[\int (\theta - \hat{\theta})^T (\theta - \hat{\theta}) p(\theta | x) d\theta \right] p(x) dx$. Since $p(x) \geq 0 \;\forall x$, if we can find a $\hat{\theta}$ that minimizes the term in brackets $\forall x$, that $\hat{\theta}$ also minimizes $B_{MSE}(\hat{\theta})$.

$$\frac{\partial}{\partial \hat{\theta}^T} \int (\theta - \hat{\theta})^T (\theta - \hat{\theta}) p(\theta | x) d\theta = -2 \int (\theta - \hat{\theta}) p(\theta | x) d\theta$$

$$= -2 \int \theta p(\theta | x) d\theta + 2 \hat{\theta} \underbrace{\int p(\theta | x) d\theta}_{=1} = 0$$

$$\Rightarrow \hat{\theta} = \int \theta p(\theta | x) d\theta = E\{\theta | x\}.$$

Once again using Bayes rule: $p(\theta | x) = \dfrac{p(x|\theta) p(\theta)}{p(x)}$

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

Then $\hat{\theta} = \dfrac{\int \theta p(x|\theta) p(\theta) d\theta}{\int p(x|\theta) p(\theta) d\theta}$.

Evaluating these integrals will become a significant challenge for some models.

Choosing a Prior PDF

Ex] Assume a situation where $p(A) = \frac{1}{\sqrt{2\pi\sigma_A^2}} e^{-\frac{(A-\mu_A)^2}{2\sigma_A^2}}$

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

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]} e^{-\frac{1}{2\sigma^2}(NA^2 - 2NA\bar{x})}$$

Then $p(A|x) = \frac{p(x|A)p(A)}{\int p(x|A)p(A)\,dA} = \dots = \frac{e^{-\frac{1}{2}\left[\frac{1}{\sigma^2}(NA^2-2NA\bar{x}) + \frac{1}{\sigma_A^2}(A-\mu_A)^2\right]}}{\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\frac{1}{\sigma^2}(NA^2-2NA\bar{x}) + \frac{1}{\sigma_A^2}(A-\mu_A)^2\right]} dA}$

$$= \frac{e^{-Q(A)/2}}{\int_{-\infty}^{\infty} e^{-Q(A)/2}\,dA} \quad \text{where } Q(A) = \frac{NA^2}{\sigma^2} - \frac{2NA\bar{x}}{\sigma^2} + \frac{A^2}{\sigma_A^2} - \frac{2\mu_A A}{\sigma_A^2} + \frac{\mu_A^2}{\sigma_A^2}$$

$$= \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)A^2 - 2\left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_A}{\sigma_A^2}\right)A + \frac{\mu_A^2}{\sigma_A^2}$$

Letting $\sigma_{A|x}^2 = \frac{1}{N/\sigma^2 + 1/\sigma_A^2}$ and

$$\mu_{A|x} = \left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_A}{\sigma_A^2}\right)\sigma_{A|x}^2$$

we can get $Q(A) = \frac{1}{\sigma_{A|x}^2}\left(A^2 - 2\mu_{A|x}A + \mu_{A|x}^2\right) - \frac{\mu_{A|x}^2}{\sigma_{A|x}^2} + \frac{\mu_A^2}{\sigma_A^2}$.

$$= \frac{1}{\sigma_{A|x}^2}(A-\mu_{A|x})^2 - \frac{\mu_{A|x}^2}{\sigma_{A|x}^2} + \frac{\mu_A^2}{\sigma_A^2}$$

$\therefore p(A|x) = \dots = \frac{1}{\sqrt{2\pi\sigma_{A|x}^2}} e^{-\frac{(A-\mu_{A|x})^2}{2\sigma_{A|x}^2}}$ The posterior pdf is Gaussian.

Then $\hat{A} = E[A|x] = \mu_{A|x} = \dots = \alpha\bar{x} + (1-\alpha)\mu_A$ where $\alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}$.

Note that $0 < \alpha < 1$ is a weighting factor for $\bar{x}$ & μ_A.

$N \to \infty \Rightarrow \alpha \to 1 \Rightarrow \hat{A} \to \bar{x}$; the effect of the prior diminishes.

For $N=1$, $\alpha = \frac{\sigma_A^2}{\sigma_A^2+\sigma^2}$ and sample x and μ_A will be weighed accordingly.

Back to B_{MSE} ...

$$B_{MSE}(\hat{\theta}) = \iint (\theta - \hat{\theta})^T(\theta - \hat{\theta})\, p(x,\theta)\, dx\, d\theta$$

$$= \iint (\theta - \hat{\theta})^T(\theta - \hat{\theta})\, p(\theta|x)\, p(x)\, d\theta\, dx$$

$\hat{\theta} = E[\theta|x]$

$$= \iint (\theta - E[\theta|x])^T(\theta - E[\theta|x])\, p(\theta|x)\, d\theta\ p(x)\, dx$$

$$= \int tr\left(Cov(\theta|x)\right) p(x)\, dx$$

$$= E_x\left[tr\left(Cov(\theta|x)\right)\right]$$

$\therefore$ Bayesian MSE is the average of the "total variance" of the posterior pdf over the data distribution.

Ex] Back to the example --- $var(A|x) = \sigma^2_{A|x} = \dfrac{1}{N/\sigma^2 + 1/\sigma_A^2}$.

$$B_{MSE}(\hat{A}) = \int \sigma^2_{A|x}\, p(x)\, dx = \sigma^2_{A|x} = \frac{1}{N/\sigma^2 + 1/\sigma_A^2} = \frac{\sigma^2}{N}\left(\frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/N}\right)$$

$\Rightarrow B_{MSE}(\hat{A}) < \dfrac{\sigma^2}{N}$ where $\dfrac{\sigma^2}{N}$ is the minimum MSE that one would achieve without any prior knowledge ($\sigma_A \to \infty$).

$\therefore$ The use of prior knowledge improved the estimator when used in the Bayesian sense.

Note: In the example above we saw that the Gaussian prior $p(A)$ was a conjugate prior for Gaussian $p(x|A)$.

Properties of the Gaussian PDF

Thm 10.2 Conditional PDF of Multivariate Gaussian

If x and y are jointly Gaussian, where x is $k\times 1$ and y is $l\times 1$, with mean vector $\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} E[x] \\ E[y] \end{bmatrix}$

and covariance matrix $C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}$, so that

$$p(x,y) = \frac{1}{(2\pi)^{\frac{k+l}{2}} |C|^{1/2}} e^{-\frac{1}{2}\left(\begin{bmatrix} x \\ y \end{bmatrix} - \mu\right)^T C^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \mu\right)}$$

then the conditional pdf $p(y|x)$ is also Gaussian

with $\mu_{y|x} \triangleq E[y|x] = \mu_y - C_{yx} C_{xx}^{-1}(x - \mu_x)$

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

Bayesian Linear Model

Let $x[n] = A + w[n]$ $n = 0, \ldots, (N-1)$ where $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$ and $w[n]$ is WGN independent of A. In vector form

$$\mathbf{x} = \mathbf{1} A + \mathbf{w}$$

This is a simple linear model where we now have a prior for A.

We will now consider the general linear model in this framework.

* The Bayesian general linear model:

Let $x = H\theta + w$ where x is $N\times 1$, H is known and $N\times p$, θ is $p\times 1$ and random with prior pdf $\mathcal{N}(\mu_\theta, C_\theta)$, and w is $N\times 1$ with pdf $\mathcal{N}(0, C_w)$ and independent of θ.

$$\text{Let } z = \begin{bmatrix} x \\ \theta \end{bmatrix} = \begin{bmatrix} H\theta + w \\ \theta \end{bmatrix} = \begin{bmatrix} H & I \\ I & 0 \end{bmatrix} \begin{bmatrix} \theta \\ w \end{bmatrix}$$

Since θ & w are independent and individually Gaussian, they are jointly Gaussian. Clearly z is then jointly Gaussian.

$\mu_x = E[x] = E[H\theta + w] = HE[\theta] + E[w] = H\mu_\theta$

$\mu_y = E[y] = E[\theta] = \mu_\theta$

$C_{xx} = E[(x-\mu_x)(x-\mu_x)^T] = \cdots = HC_\theta H^T + C_w$

$C_{yx} = E[(y-\mu_y)(x-\mu_x)^T] = \cdots = C_\theta H^T$

$C_{yy} = C_\theta$

<u>Thm 10-3</u> Posterior PDF for the Bayesian General Linear Model

If $x = H\theta + w$ ($x \sim N\times 1$, $H \sim N\times p$ & known, $w \sim N\times 1$, $\theta \sim p\times 1$) with $\theta \sim \mathcal{N}(\mu_\theta, C_\theta)$ and $w \sim \mathcal{N}(0, C_w)$ independent of θ, then $p(\theta|x)$ is Gaussian with

$$\mu_{\theta|x} = E[\theta|x] = \mu_\theta + C_\theta H^T(HC_\theta H^T + C_w)^{-1}(x - H\mu_\theta)$$

$$C_{\theta|x} = \mathrm{Cov}(\theta|x) = C_\theta - C_\theta H^T(HC_\theta H^T + C_w)^{-1}HC_\theta .$$

<u>Note</u>: Due to $C_w > 0$, we do not need H to be full rank for $(HC_\theta H^T + C_w)^{-1}$ to exist.

Ex | DC Level in WGN with Gaussian prior

$x[n] = A + w[n]$ for $n = 0, \dots, (N-1)$ with $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$

and $w[n]$ is WGN with variance σ^2 and indep. of A.

$$\mathbf{x} = \mathbf{1}A + \mathbf{w}$$

$\therefore$ $p(A|\mathbf{x})$ is Gaussian with

$$E[A|\mathbf{x}] = \mu_A + \sigma_A^2 \mathbf{1}^T (\mathbf{1}\sigma_A^2 \mathbf{1}^T + \sigma^2 I)^{-1} (\mathbf{x} - \mathbf{1}\mu_A)$$

Using Woodbury identity: $\left(I + \frac{\sigma_A^2}{\sigma^2} \mathbf{1}\mathbf{1}^T\right)^{-1} = I - \frac{\left(\frac{\sigma_A^2}{\sigma^2}\right) \mathbf{1}\mathbf{1}^T}{\left(1 + N\frac{\sigma_A^2}{\sigma^2}\right)}$

$$\Rightarrow E[A|\mathbf{x}] = \dots = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} (\bar{x} - \mu_A)$$

$$\mathrm{Var}(A|\mathbf{x}) = \sigma_A^2 - \sigma_A^2 \mathbf{1}^T (\mathbf{1}\sigma_A^2 \mathbf{1}^T + \sigma^2 I)^{-1} \mathbf{1}\sigma_A^2$$

$$= \dots = \left(\frac{\sigma^2}{N}\sigma_A^2\right) \Big/ \left(\sigma_A^2 + \frac{\sigma^2}{N}\right)$$

Alternative forms of $\mu_{\theta|x}$ and $C_{\theta|x}$ for $p(\theta|\mathbf{x})$:

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

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

$$\underbrace{C_{\theta|x}^{-1}}_{\text{posterior information}} = \underbrace{C_\theta^{-1}}_{\text{prior information}} + \underbrace{H^T C_w^{-1} H}_{\text{information of data}}$$

Nuisance Parameters

Typically, the model has more parameters than we are interested in estimating. These nuisance parameters increase the dimensionality of our parameter estimation problem if considered as deterministic unknowns. In the Bayesian approach, they can be "integrated out".

Suppose, we are interested in estimating parameters θ and some nuisance parameters α are present. Then

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

where $p(x|\theta) = \int p(x|\theta,\alpha)p(\alpha)\,d\alpha$.

Ex| Let $x \sim p(x|\theta,\sigma^2) \equiv \mathcal{N}(0, \sigma^2 C(\theta))$. Here σ^2 is a nuisance parameter and we want to estimate θ.

Assume that $p(\sigma^2) = \begin{cases} \frac{\lambda}{\sigma^4} e^{-\lambda/\sigma^2} & \text{if } \sigma^2 > 0 \\ 0 & \text{if } \sigma^2 \le 0 \end{cases}$ with $\lambda > 0$, $\sigma^2 \perp\!\!\!\perp \theta$

(inverted gamma) $\quad$ independent

Then $p(x|\theta) = \int p(x|\theta,\sigma^2)p(\sigma^2)\,d\sigma^2$

$$= \int_0^\infty (2\pi)^{-N/2} |\sigma^2 C(\theta)|^{-1/2} e^{-\frac{1}{2}x^T(\sigma^2 C(\theta))^{-1}x} \frac{\lambda e^{-\lambda/\sigma^2}}{\sigma^4}\,d\sigma^2$$

$\xi \triangleq \sigma^{-2}$

$$= \int_0^\infty \lambda(2\pi)^{-N/2} |C(\theta)|^{-1/2} \xi^{N/2} e^{-(\lambda + \frac{1}{2}x^T C^{-1}(\theta)x)\xi}\,d\xi$$

$$\int_0^\infty x^{m-1} e^{-ax}\,dx = a^{-m}\Gamma(m) \quad \text{for } a>0 \text{ and } m>0,$$

$$\theta) = \lambda\Gamma\left(\frac{N}{2}+1\right)(2\pi)^{-N/2}\,|C(\theta)|^{-1/2}\left(\lambda + \frac{1}{2}x^T C^{-1}(\theta)x\right)^{-\left(\frac{N}{2}+1\right)}.$$

$$p(\theta|x) \propto p(x|\theta)p(\theta).$$

proportional to

n Estimation for Deterministic Parameters

apply the Bayesian estimation framework to
n of estimating a deterministic θ, we
d up with a MMSE estimator, which will
ll on average (as different deterministic θ
e encountered. For the particular θ value,
s a risk of getting poor performance.

a previous example we had $p(A) = \mathcal{N}(\mu_A, \sigma_A^2)$
$x[n] = A + w[n]$ where n is WGN with σ^2 variance.
yesian estimator of A was found to be

$$\alpha\bar{x} + (1-\alpha)\mu_A \quad \text{where } \alpha = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/N} \text{ with } 0<\alpha<1.$$

s deterministic, then

$$(\hat{A}) = \mathrm{var}(\hat{A}) + b^2(\hat{A}) = \alpha^2\,\mathrm{var}(\bar{x}) + [\overbrace{\alpha A + (1-\alpha)\mu_A}^{E[\hat{A}]} - A]^2$$

$$= \underbrace{\alpha^2\frac{\sigma^2}{N}}_{\Rightarrow \mathrm{var}(\hat{A}) = \alpha^2\frac{\sigma^2}{N}} + \underbrace{(1-\alpha)^2(A-\mu_A)^2}_{\Rightarrow b(\hat{A}) = (1-\alpha)(A-\mu_A)}.$$

* Since $0<\alpha<1$ using the Bayesian approach reduced the variance but may increase bias depending on μ_A!

μ_A is close to the true deterministic A, then the
yesian estimator could have smaller MSE than the
u estimator. Otherwise, its MSE could be significantly
eater. However, on average

$$B_{MSE}(\hat{A}) = E_A[MSE(\hat{A})]$$
$$= \alpha^2 \frac{\sigma^2}{N} + (1-\alpha)^2 E_A[(A-\mu_A)^2]$$
$$= \alpha^2 \frac{\sigma^2}{N} + (1-\alpha)^2 \sigma_A^2$$
$$= \frac{\sigma^2}{N} \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/N}$$
$$< \frac{\sigma^2}{N} = B_{MSE}(\bar{x})$$

, using the Bayesian approach and assuming a prior
θ, corresponds to making a trade-off between
s and variance to reduce, on average, overall MSE.
In practice, if we are not certain about the
ge of values θ may take, we would use a "flat"
or. In the example, this means $\sigma_A^2 \to \infty$ as our
ertainty gets higher prior to observing data.

---

sted Problems: **2**, 3, 4, 10, 14, 15, 16