

EE6617: 06 Bayesian Estimators

Major references:

- *Fundamentals of Statistical Processing, Volume I: Estimation Theory*, 1/E, by Steven M. Kay, Prentice Hall, 1993
- *Statistical Signal Processing*, 1/E, by Louis L. Scharf, Prentice Hall, 1991

We now depart from the classical approach to statistical estimation in which the parameter θ is assumed to be a deterministic but unknown constant.

If we knew it was a **random variable** and we knew the **prior distribution**, we can leverage this knowledge to **better estimate θ** .

Motivating Example: Prior Knowledge Improves Estimation Performance

Suppose we observe $\mathbf{x} = \mathbf{1}A + \mathbf{w}$ with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. We already know the MVU estimator is

$$\hat{A}_{MVU} = \frac{1}{N} \sum_{n=0}^{N-1} x[n].$$

and

$$\hat{A}_{MVU} \sim \mathcal{N}(A, \sigma^2/N).$$

This estimator assumes nothing about A .

Now, suppose we know **A is uniformly distributed** in $[-A_0, A_0]$. Let $\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$ and a new estimator be

$$\hat{A} = \begin{cases} -A_0 & \bar{x} < -A_0 \\ \bar{x} & -A_0 \leq \bar{x} < A_0 \\ A_0 & \bar{x} \geq A_0 \end{cases}.$$

Then,

$$\begin{aligned} mse(\hat{A}_{MVU}) &= E \left[\left| \hat{A}_{MVU} - A \right|^2 | A \right] \\ &= \int_{-\infty}^{\infty} (\bar{x} - A)^2 p(\bar{x}; A) d\bar{x} \\ &= \int_{-A_0}^{A_0} (\bar{x} - A)^2 p(\bar{x}; A) d\bar{x} + \int_{-\infty}^{-A_0} (\bar{x} - A)^2 p(\bar{x}; A) d\bar{x} + \int_{A_0}^{\infty} (\bar{x} - A)^2 p(\bar{x}; A) d\bar{x} \\ &\geq \int_{-A_0}^{A_0} (\bar{x} - A)^2 p(\bar{x}; A) d\bar{x} + (-A_0 - A)^2 \Pr(\bar{x} < -A_0; A) \\ &\quad + (A_0 - A)^2 \Pr(\bar{x} > A_0; A) \\ &= E \left[\left| \hat{A} - A \right|^2 | A \right] = mse(\hat{A}). \end{aligned}$$

We see that the **mean square error is reduced by using the prior distribution**.

We call

$$\begin{aligned} mse(\hat{A}) &= E_{\mathbf{x}} \left[\left| \hat{A} - A \right|^2 | A \right] \Leftarrow \text{Non-Bayesian} \\ B_{mse}(\hat{A}) &= E_{\mathbf{x}, A} \left[\left| \hat{A} - A \right|^2 \right] \Leftarrow \text{Bayesian.} \end{aligned}$$

Notice that

$$\begin{aligned} B_{mse}(\hat{A}) &= E_A \left[mse(\hat{A}) \right] . \\ &= \underbrace{\int \int \left(A - \hat{A} \right)^2 p(\mathbf{x}; A) d\mathbf{x} dA}_{mse(\hat{A})} \end{aligned} \quad (1)$$

and when we say it is Bayesian, it means we treat $p(\mathbf{x}; A)$ as the joint PDF between $\mathbf{x}$ and A .

In the **Bayesian approach**, we want to **minimize average mean square error**

$$\hat{A}_{mmse} = \underset{\hat{A}}{\operatorname{argmin}} B_{mse}(\hat{A}) .$$

The $\hat{A}_{mmse}$ is called minimum mean square error **(MMSE) estimator**.

By **Bayes' theorem**

$$p(\mathbf{x}; A) = p(A|\mathbf{x})p(\mathbf{x}),$$

$$\begin{aligned} \min_{\hat{A}} B_{mse}(\hat{A}) &= \min_{\hat{A}} \int \int \left(A - \hat{A} \right)^2 p(\mathbf{x}; A) d\mathbf{x} dA \\ &= \int \min_{\hat{A}} \left(\int \left(A - \hat{A} \right)^2 p(A|\mathbf{x}) dA \right) p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Since $\hat{A}$ is a function of $\mathbf{x}$, we can minimize $B_{mse}(\hat{A})$ by **minimizing**

$$\int \left(A - \hat{A} \right)^2 p(A|\mathbf{x}) dA \quad \text{for a given } \mathbf{x}.$$

Notice that

$$\begin{aligned} \frac{\partial}{\partial \hat{A}} \int \left(A - \hat{A} \right)^2 p(A|\mathbf{x}) dA &= \int -2(A - \hat{A}) p(A|\mathbf{x}) dA \\ &= -2 \int A p(A|\mathbf{x}) dA + 2\hat{A} \int p(A|\mathbf{x}) dA \\ \frac{\partial^2}{\partial^2 \hat{A}} \int \left(A - \hat{A} \right)^2 p(A|\mathbf{x}) dA &= 2 > 0 \Rightarrow \text{convex.} \end{aligned}$$

Notice $\int p(A|\mathbf{x})dA = 1$. Solving for $\frac{\partial}{\partial \hat{A}} \int (A - \hat{A})^2 p(A|\mathbf{x})dA = 0$ yields

$$\hat{A}_{mmse} = \int A p(A|\mathbf{x})dA = E[A|\mathbf{x}].$$

$\Rightarrow$ **The optimal estimator is the conditional mean of A given $\mathbf{x}$, i.e., $E[A|\mathbf{x}]$.**

In determining the MMSE estimator we first require the **posterior PDF**. We can use Bayes's rule to determine it as

$$\begin{aligned} p(A|\mathbf{x}) &= \frac{p(\mathbf{x}|A)p(A)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|A)p(A)}{\int p(\mathbf{x}|A)p(A)dA}. \end{aligned}$$

We continue our example.

$$\begin{aligned} p(A) &= \begin{cases} \frac{1}{2A_0}, & |A| < A_0 \\ 0, & \text{elsewhere} \end{cases} \\ p(\mathbf{x}|A) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right) \\ p(\mathbf{x}) &= \int_{-A_0}^{A_0} \frac{1}{2A_0 (2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right) dA \end{aligned} \tag{2}$$

Note that

$$\sum_{n=0}^{N-1} (x[n] - A)^2 = \sum_{n=0}^{N-1} x^2[n] - 2A \underbrace{\sum_{n=0}^{N-1} x[n]}_{=N\bar{x}} + NA^2 = N(A - \bar{x})^2 + \sum_{n=0}^{N-1} x^2[n] - N\bar{x}^2$$

So, the posteriori PDF is

$$\begin{aligned} p(A|\mathbf{x}) = \frac{p(\mathbf{x}|A)p(A)}{p(\mathbf{x})} &= \begin{cases} \frac{\frac{1}{2A_0(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} N(A-\bar{x})^2\right)}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} N(A-\bar{x})^2\right) dA}, & |A| < A_0 \\ 0, & \text{elsewhere} \end{cases} \\ &= \begin{cases} \frac{1}{c\sqrt{2\pi\sigma^2/N}} \exp\left(-\frac{1}{2\sigma^2/N} (A - \bar{x})^2\right), & |A| < A_0 \\ 0, & \text{elsewhere} \end{cases} \\ &\text{, where } c = \int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi\sigma^2/N}} \exp\left(-\frac{1}{2\sigma^2/N} (A - \bar{x})^2\right) dA. \end{aligned}$$

Thus, the MMSE estimator is

$$\begin{aligned}\hat{A}_{mmse} &= E[A|\mathbf{x}] = \int A p(A|\mathbf{x}) dA \\ &= \frac{\int_{-A_0}^{A_0} \frac{A}{\sqrt{2\pi\sigma^2/N}} \exp\left(-\frac{1}{2\sigma^2/N} (A - \bar{x})^2\right) dA}{\int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi\sigma^2/N}} \exp\left(-\frac{1}{2\sigma^2/N} (A - \bar{x})^2\right) dA}.\end{aligned}$$

Unfortunately, this can not be further simplified easily. But, $\hat{A}_{mmse} \rightarrow \bar{x}$ as $A_0 \rightarrow \infty$ or $N \rightarrow \infty$ or $\sigma^2 \rightarrow 0$. ■

Generally speaking, the estimator that minimizes $B_{mse}(\hat{\theta})$ is

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

This is known as the **minimum mean square error (MMSE) estimator**. While $p(\theta|\mathbf{x})$ might be difficult to deal with during integration (as seen from the example), many various cases can be solved exactly or you can use numerical integration.

1 Gaussian MMSE: Scalar Case

Consider $\mathbf{x} = \mathbf{1}A + \mathbf{w}$ where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ and $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$.

In deriving MMSE, we want to first find out $p(A|\mathbf{x})$. Normally, we do not like to work directly with $p(A|\mathbf{x})$. Instead use

$$p(A|\mathbf{x}) = \frac{p(\mathbf{x}|A)p(A)}{p(\mathbf{x})}. \quad (3)$$

So,

$$\begin{aligned}p(A) &= \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{1}{2\sigma_A^2} (A - \mu_A)^2\right) \\ p(\mathbf{x}|A) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n]\right) \exp\left(-\frac{1}{2\sigma^2} (NA^2 - 2NA\bar{x})\right) \\ p(\mathbf{x}|A)p(A) &= \frac{1}{(2\pi\sigma^2)^{N/2} \sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n]\right) \exp\left(-\frac{1}{2\sigma^2} (NA^2 - 2NA\bar{x})\right) \exp\left(-\frac{1}{2\sigma_A^2} (A - \mu_A)^2\right) \\ p(\mathbf{x}) &= \int p(\mathbf{x}|A)p(A) dA \\ &= \frac{1}{(2\pi\sigma^2)^{N/2} \sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n]\right) \int \exp\left(-\frac{1}{2\sigma^2} (NA^2 - 2NA\bar{x}) - \frac{1}{2\sigma_A^2} (A - \mu_A)^2\right) dA\end{aligned}$$

Then, from (3)

$$p(A|\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}\left(\frac{NA^2-2NA\bar{\mathbf{x}}}{\sigma^2} + \frac{(A-\mu_A)^2}{\sigma_A^2}\right)\right)}{\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{NA^2-2NA\bar{\mathbf{x}}}{\sigma^2} + \frac{(A-\mu_A)^2}{\sigma_A^2}\right)\right) dA}$$

Let

$$\begin{aligned} f(A) &= \frac{NA^2 - 2NA\bar{\mathbf{x}}}{\sigma^2} + \frac{(A - \mu_A)^2}{\sigma_A^2} \\ &= \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right) A^2 - 2\left(\frac{N\bar{\mathbf{x}}}{\sigma^2} + \frac{\mu_A}{\sigma_A^2}\right) A + \frac{\mu_A^2}{\sigma_A^2} \\ &= \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right) \left(A^2 - 2\frac{\left(\frac{N\bar{\mathbf{x}}}{\sigma^2} + \frac{\mu_A}{\sigma_A^2}\right)}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)} A\right) + \frac{\mu_A^2}{\sigma_A^2} \\ &= \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right) \left(A^2 - 2\frac{\left(\frac{N\bar{\mathbf{x}}}{\sigma^2} + \frac{\mu_A}{\sigma_A^2}\right)}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)} A + \frac{\left(\frac{N\bar{\mathbf{x}}}{\sigma^2} + \frac{\mu_A}{\sigma_A^2}\right)^2}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)^2}\right) + \frac{\mu_A^2}{\sigma_A^2} - \frac{\left(\frac{N\bar{\mathbf{x}}}{\sigma^2} + \frac{\mu_A}{\sigma_A^2}\right)^2}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)} \\ &= \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right) \left(A - \frac{\left(\frac{N}{\sigma^2}\bar{\mathbf{x}} + \frac{1}{\sigma_A^2}\mu_A\right)}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)}\right)^2 + \frac{\mu_A^2}{\sigma_A^2} - \frac{\left(\frac{N}{\sigma^2}\bar{\mathbf{x}} + \frac{1}{\sigma_A^2}\mu_A\right)^2}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)} \end{aligned}$$

We define

$$\begin{aligned} \sigma_{A|\mathbf{x}}^2 &= \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \\ \mu_{A|\mathbf{x}} &= \frac{\left(\frac{N}{\sigma^2}\bar{\mathbf{x}} + \frac{1}{\sigma_A^2}\mu_A\right)}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}\right)} = \left(\frac{N}{\sigma^2}\bar{\mathbf{x}} + \frac{1}{\sigma_A^2}\mu_A\right) \sigma_{A|\mathbf{x}}^2. \end{aligned}$$

Now, we can rewrite $f(A)$ as

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

So,

$$\begin{aligned}
p(A|\mathbf{x}) &= \frac{\exp\left(-\frac{1}{2}f(A)\right)}{\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}f(A)\right) dA} \\
&= \frac{\exp\left(-\frac{(A-\mu_{A|\mathbf{x}})^2}{2\sigma_{A|\mathbf{x}}^2}\right)}{\int_{-\infty}^{\infty} \exp\left(-\frac{(A-\mu_{A|\mathbf{x}})^2}{2\sigma_{A|\mathbf{x}}^2}\right) dA} \\
&= \frac{1}{\sqrt{2\pi\sigma_{A|\mathbf{x}}^2}} \exp\left(-\frac{(A-\mu_{A|\mathbf{x}})^2}{2\sigma_{A|\mathbf{x}}^2}\right)
\end{aligned} \tag{4}$$

where the last equality follows from $\frac{1}{\sqrt{2\pi\sigma_{A|\mathbf{x}}^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(A-\mu_{A|\mathbf{x}})^2}{2\sigma_{A|\mathbf{x}}^2}\right) dA = 1$.

The (4) implies that

$$A|\mathbf{x} \sim \mathcal{N}(\mu_{A|\mathbf{x}}, \sigma_{A|\mathbf{x}}^2)$$

Thus, the MMSE $\hat{A}$ for Gaussian case is

$$\begin{aligned}
\hat{A} = E[A|\mathbf{x}] &= \mu_{A|\mathbf{x}} \\
&= \frac{\frac{N}{\sigma^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \bar{\mathbf{x}} + \frac{\frac{1}{\sigma_A^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \mu_A \\
&= \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \bar{\mathbf{x}} + \frac{\frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}} \mu_A \\
&= \underbrace{\alpha \bar{\mathbf{x}}}_{\text{MVU estimator}} + (1 - \alpha) \underbrace{\mu_A}_{E[A]}
\end{aligned}$$

where

$$\alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} = \frac{1}{1 + \frac{\sigma^2}{N\sigma_A^2}}. \tag{5}$$

Can think $\frac{\sigma^2}{N\sigma_A^2} = \frac{1}{\frac{\sigma_A^2}{\sigma^2/N}}$ as $\frac{1}{SNR}$.

$$\begin{aligned}
\text{As } \frac{\sigma^2}{N\sigma_A^2} &\rightarrow 0 \text{ or } SNR \rightarrow \infty, \quad \hat{A} \rightarrow \bar{\mathbf{x}} \\
\frac{\sigma^2}{N\sigma_A^2} &\rightarrow \infty \text{ or } SNR \rightarrow 0, \quad \hat{A} \rightarrow \mu_A.
\end{aligned}$$

The MMSE estimator makes a **tradeoff** by weighting the **mean of A** and the **MVU estimator of A** . As well,

$$\text{var}(A|\mathbf{x}) = \sigma_{A|\mathbf{x}}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} = \frac{\frac{\sigma^2}{N}}{1 + \frac{\sigma^2}{N\sigma_A^2}} = \underbrace{\frac{\sigma^2}{N}}_{\text{variance of MVU estimator}} \alpha.$$

Thus, as $\frac{\sigma^2}{N\sigma_A^2} \rightarrow 0$ or $SNR \rightarrow \infty$, $\text{var}(A|\mathbf{x})$ approaches the variance of the MVU estimator. In fact,

$$\begin{aligned} B_{mse}(\hat{A}) &= \int \int (A - \hat{A})^2 p(A|\mathbf{x}) dA p(\mathbf{x}) d\mathbf{x} \\ &= \int \int (A - E[A|\mathbf{x}])^2 p(A|\mathbf{x}) dA p(\mathbf{x}) d\mathbf{x} \\ &= \int \text{var}(A|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= \int \sigma_{A|\mathbf{x}}^2 p(\mathbf{x}) d\mathbf{x} \\ &= \frac{\sigma^2}{N} \alpha \end{aligned}$$

Since $0 \leq \alpha \leq 1$,

$$B_{mse}(\hat{A}) = \frac{\sigma^2}{N} \alpha \leq \frac{\sigma^2}{N} = mse(\hat{A}_{mvu}).$$

$\Rightarrow$ we see an **improvement in average variance by a factor of α** when using MMSE estimator.

Gaussian MMSE estimator is very useful because closed form solution exists. Also, it is often a decent model for many real world prior distributions:

- Communication sources
- Image sources
- Various kinds of noise
- Channel, etc.

2 Gaussian Properties

Let X, θ be jointly Gaussian.

$$p(X, \theta) = \frac{1}{2\pi\sqrt{\det(\mathbf{C})}} \exp \left(-\frac{1}{2} \begin{bmatrix} X - E[X] \\ \theta - E[\theta] \end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix} X - E[X] \\ \theta - E[\theta] \end{bmatrix} \right)$$

where

$$\begin{aligned}\mathbf{C} &= E \begin{bmatrix} (X - E[X])^2 & (X - E[X])(\theta - E[\theta]) \\ (X - E[X])(\theta - E[\theta]) & (\theta - E[\theta])^2 \end{bmatrix} \\ &= \begin{bmatrix} \text{var}(X) & \text{cov}(X, \theta) \\ \text{cov}(X, \theta) & \text{var}(\theta) \end{bmatrix}.\end{aligned}$$

Theorem 1 *The **conditional PDF** $p(\theta|X)$ is also a **Gaussian** distribution with*

$$\begin{aligned}E[\theta|X] &= E[\theta] + \frac{\text{cov}(X, \theta)}{\text{var}(X)} (X - E[X]) \\ \text{var}(\theta|X) &= \text{var}(\theta) - \frac{\text{cov}(X, \theta)^2}{\text{var}(X)} \\ p(\theta|X) &= \mathcal{N}(E[\theta|X], \text{var}(\theta|X))\end{aligned}$$

Remark 1 *Observing X adjusts the mean and variance of θ . **Remember that conditioning reduces uncertainty of a random variable.** Note also that*

$$\begin{aligned}\text{var}(\theta|X) &= \text{var}(\theta) \left(1 - \frac{\text{cov}(X, \theta)^2}{\text{var}(X) \text{var}(\theta)} \right) \\ &= \text{var}(\theta) (1 - \rho^2)\end{aligned}$$

where $\rho = \frac{\text{cov}(X, \theta)}{\sqrt{\text{var}(X) \text{var}(\theta)}}$ is the correlation coefficient between X and θ , $0 \leq \rho \leq 1$.

We can compute

$$\begin{aligned}B_{mse}(\hat{\theta}) &= \int \text{var}(\theta|X) p(X) dX \\ &= \text{var}(\theta|X) \\ &= \text{var}(\theta) (1 - \rho^2).\end{aligned}$$

Higher the statistical dependency between X and θ , the better the estimator is.

3 Gaussian MMSE: Vector Case

Let $\mathbf{x} \in \mathbb{R}^{N \times 1}$ and $\boldsymbol{\theta} \in \mathbb{R}^{p \times 1}$ be jointly Gaussian

$$p(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{(N+p)/2} \sqrt{\det(\mathbf{C})}} \exp \left(-\frac{1}{2} \begin{bmatrix} X - E[X] \\ \boldsymbol{\theta} - E[\boldsymbol{\theta}] \end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix} X - E[X] \\ \boldsymbol{\theta} - E[\boldsymbol{\theta}] \end{bmatrix} \right)$$

where

$$\begin{aligned}\mathbf{C} &= E \begin{bmatrix} (\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T & (\mathbf{x} - E[\mathbf{x}])(\boldsymbol{\theta} - E[\boldsymbol{\theta}])^T \\ (\boldsymbol{\theta} - E[\boldsymbol{\theta}])(\mathbf{x} - E[\mathbf{x}])^T & (\boldsymbol{\theta} - E[\boldsymbol{\theta}])(\boldsymbol{\theta} - E[\boldsymbol{\theta}])^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{C}_{\mathbf{xx}} & \mathbf{C}_{\mathbf{x}\boldsymbol{\theta}} \\ \mathbf{C}_{\boldsymbol{\theta}\mathbf{x}} & \mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}} \end{bmatrix}.\end{aligned}$$

Theorem 2 The *conditional PDF* $p(\boldsymbol{\theta}|\mathbf{x})$ is also a Gaussian distribution with

$$\begin{aligned}E[\boldsymbol{\theta}|\mathbf{x}] &= E[\boldsymbol{\theta}] + \mathbf{C}_{\boldsymbol{\theta}\mathbf{x}}\mathbf{C}_{\mathbf{xx}}^{-1}(\mathbf{x} - E[\mathbf{x}]) \\ \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} &= \mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{C}_{\boldsymbol{\theta}\mathbf{x}}\mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{x}\boldsymbol{\theta}} \\ p(\boldsymbol{\theta}|\mathbf{x}) &= \mathcal{N}(E[\boldsymbol{\theta}|\mathbf{x}], \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}}).\end{aligned}$$

Recall that

$$\begin{aligned}\mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}} &= E[\boldsymbol{\theta}\boldsymbol{\theta}^T] - E[\boldsymbol{\theta}]E[\boldsymbol{\theta}^T] \\ \mathbf{C}_{\mathbf{x}\boldsymbol{\theta}} &= E[\mathbf{x}\boldsymbol{\theta}^T] - E[\mathbf{x}]E[\boldsymbol{\theta}^T] \\ \mathbf{C}_{\mathbf{xx}} &= E[\mathbf{x}\mathbf{x}^T] - E[\mathbf{x}]E[\mathbf{x}^T].\end{aligned}$$

4 Bayesian Linear Model: Gaussian MMSE

Now, we consider

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where $\mathbf{H}$ is known, $\boldsymbol{\theta} \sim \mathcal{N}(\mu_{\boldsymbol{\theta}}, \mathbf{C}_{\boldsymbol{\theta}})$, and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{w}})$. Since $\boldsymbol{\theta}$, $\mathbf{w}$ are Gaussian, $\mathbf{x}$ and $\boldsymbol{\theta}$ are jointly Gaussian.

$$\begin{aligned}E[\mathbf{x}] &= \mathbf{H}\mu_{\boldsymbol{\theta}} \\ \mathbf{C}_{\mathbf{xx}} &= E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T] \\ &= E[(\mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{H}\mu_{\boldsymbol{\theta}})(\mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{H}\mu_{\boldsymbol{\theta}})^T] \\ &= E[(\mathbf{H}(\boldsymbol{\theta} - \mu_{\boldsymbol{\theta}}) + \mathbf{w})(\mathbf{H}(\boldsymbol{\theta} - \mu_{\boldsymbol{\theta}}) + \mathbf{w})^T] \\ &= \mathbf{H}E[(\boldsymbol{\theta} - \mu_{\boldsymbol{\theta}})(\boldsymbol{\theta} - \mu_{\boldsymbol{\theta}})^T]\mathbf{H}^T + E[\mathbf{w}\mathbf{w}^T] \\ &= \mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T + \mathbf{C}_{\mathbf{w}}\end{aligned}$$

$$\begin{aligned}\mathbf{C}_{\boldsymbol{\theta}\mathbf{x}} &= E[(\boldsymbol{\theta} - E[\boldsymbol{\theta}])(\mathbf{x} - E[\mathbf{x}])^T] \\ &= E[(\boldsymbol{\theta} - \mu_{\boldsymbol{\theta}})(\mathbf{H}(\boldsymbol{\theta} - \mu_{\boldsymbol{\theta}}) + \mathbf{w})^T] \\ &= \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T.\end{aligned}$$

Thus, we can derive the **Gaussian linear MMSE (LMMSE)** as follows.

Theorem 3 If we observe $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ where $\mathbf{H}$ is known, $\boldsymbol{\theta} \sim \mathcal{N}(\mu_{\boldsymbol{\theta}}, \mathbf{C}_{\boldsymbol{\theta}})$, and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{w}})$,

$$\begin{aligned} E[\boldsymbol{\theta}|\mathbf{x}] &= \mu_{\boldsymbol{\theta}} + \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T (\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T + \mathbf{C}_{\mathbf{w}})^{-1} (\mathbf{x} - \mathbf{H}\mu_{\boldsymbol{\theta}}) \\ \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} &= \mathbf{C}_{\boldsymbol{\theta}} - \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T (\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T + \mathbf{C}_{\mathbf{w}})^{-1} \mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}. \end{aligned}$$

By using the matrix inversion lemma $(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}\mathbf{A}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\mathbf{D}\mathbf{A}^{-1}$, we can simplify

$$\mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} = (\mathbf{C}_{\boldsymbol{\theta}}^{-1} + \mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}$$

and

$$\begin{aligned} E[\boldsymbol{\theta}|\mathbf{x}] &= \mu_{\boldsymbol{\theta}} + (\mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1} \mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{H}\mu_{\boldsymbol{\theta}}) \\ &= \mu_{\boldsymbol{\theta}} + \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}}\mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{H}\mu_{\boldsymbol{\theta}}). \end{aligned}$$

The last equality follows from

$$\begin{aligned} \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T (\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T + \mathbf{C}_{\mathbf{w}})^{-1} &= \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T \left(\mathbf{C}_{\mathbf{w}}^{-1} - \mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} (\mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1} \mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1} \right) \\ &= \mathbf{C}_{\boldsymbol{\theta}} \left(\mathbf{I} - \mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} (\mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1} \right) \mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1} \\ &= \mathbf{C}_{\boldsymbol{\theta}} \left((\mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1}) - \mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} \right) (\mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1} \mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1} \\ &= (\mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1} \mathbf{H}^T\mathbf{C}_{\mathbf{w}}^{-1}. \end{aligned}$$

Example 2: DC Signal with Noise

Consider $\mathbf{x} = \mathbf{1}A + \mathbf{w}$ with $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$ and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Estimate A , given $\mathbf{x}$.

The optimal estimator that minimizes the Bayesian MSE is $\hat{A}_{mmse} = E[A|\mathbf{x}]$ and from theorem 3,

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

By using the matrix inversion lemma,

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

it is simplified to

$$\begin{aligned}
E[A|\mathbf{x}] &= \mu_A + \frac{\sigma_A^2}{\sigma^2} \mathbf{1}^T \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{N + \frac{\sigma^2}{\sigma_A^2}} \right) (\mathbf{x} - \mu_A \mathbf{1}) \\
&= \mu_A + \frac{\sigma_A^2}{\sigma^2} \left(\mathbf{1}^T - \frac{N\mathbf{1}^T}{N + \frac{\sigma^2}{\sigma_A^2}} \right) (\mathbf{x} - \mu_A \mathbf{1}) \\
&= \mu_A + \frac{\sigma_A^2}{\sigma^2} \underbrace{\left(1 - \frac{N}{N + \frac{\sigma^2}{\sigma_A^2}} \right)}_{\frac{\frac{\sigma^2}{\sigma_A^2}}{N + \frac{\sigma^2}{\sigma_A^2}}} (N\bar{\mathbf{x}} - N\mu_A) \\
&= \mu_A + \frac{N}{N + \frac{\sigma^2}{\sigma_A^2}} (\bar{\mathbf{x}} - \mu_A) \\
&= \mu_A + \underbrace{\frac{1}{1 + \frac{\sigma^2}{N\sigma_A^2}}}_{=\alpha} (\bar{\mathbf{x}} - \mu_A) \\
&= \alpha \bar{\mathbf{x}} + (1 - \alpha) \mu_A.
\end{aligned}$$

$$\begin{aligned}
\text{var}(A|\mathbf{x}) &= \sigma_A^2 - \sigma_A^2 \mathbf{1}^T (\mathbf{1}\sigma_A^2 \mathbf{1}^T + \sigma^2 \mathbf{I})^{-1} \mathbf{1}\sigma_A^2 \\
&= \sigma_A^2 - \frac{\sigma_A^2}{\sigma^2} \mathbf{1}^T \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{N + \frac{\sigma^2}{\sigma_A^2}} \right) \mathbf{1}\sigma_A^2 \\
&= \sigma_A^2 - \frac{\sigma_A^2}{\sigma^2} \left(1 - \frac{N}{N + \frac{\sigma^2}{\sigma_A^2}} \right) N\sigma_A^2 \\
&= \left(1 - \frac{N}{N + \frac{\sigma^2}{\sigma_A^2}} \right) \sigma_A^2 \\
&= \left(\frac{\frac{\sigma^2}{\sigma_A^2}}{N + \frac{\sigma^2}{\sigma_A^2}} \right) \sigma_A^2 \\
&= \frac{\frac{\sigma^2}{N}}{1 + \frac{\sigma^2}{N\sigma_A^2}} = \alpha \frac{\sigma^2}{N}
\end{aligned}$$

This is same as the Gaussian scalar case.

Example 3: Vector Signal with Noise

Observe $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ where $\mathbf{H} \in \mathbb{R}^{N \times p}$, $\boldsymbol{\theta} \in \mathbb{R}^{p \times 1}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$.

We have shown that $\boldsymbol{\theta}|\mathbf{x} \sim \mathcal{N}(\mu_{\boldsymbol{\theta}|\mathbf{x}}, \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}})$. Thus, the MMSE estimator of $\boldsymbol{\theta}$ is

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{mmse} = \mu_{\boldsymbol{\theta}|\mathbf{x}} &= \sigma_{\theta}^2 \mathbf{H}^T (\sigma_{\theta}^2 \mathbf{H} \mathbf{H}^T + \sigma^2 \mathbf{I})^{-1} \mathbf{x} \\
&= \sigma_{\theta}^2 \mathbf{H}^T \left(\frac{1}{\sigma^2} \mathbf{I} - \frac{1}{\sigma^2} \mathbf{H} \left(\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} + \frac{1}{\sigma_{\theta}^2} \mathbf{I} \right)^{-1} \frac{1}{\sigma^2} \mathbf{H}^T \right) \mathbf{x} \\
&= \sigma_{\theta}^2 \left(\frac{1}{\sigma^2} \mathbf{I} - \frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} \left(\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} + \frac{1}{\sigma_{\theta}^2} \mathbf{I} \right)^{-1} \frac{1}{\sigma^2} \right) \mathbf{H}^T \mathbf{x} \\
&= \sigma_{\theta}^2 \left(\frac{1}{\sigma^2} \left(\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} + \frac{1}{\sigma_{\theta}^2} \mathbf{I} \right) - \frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} \frac{1}{\sigma^2} \right) \left(\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} + \frac{1}{\sigma_{\theta}^2} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{x} \\
&= \frac{1}{\sigma^2} \left(\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} + \frac{1}{\sigma_{\theta}^2} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{x} \\
&= \left(\mathbf{H}^T \mathbf{H} + \frac{\sigma^2}{\sigma_{\theta}^2} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{x} = \left(\mathbf{H}^T \mathbf{H} + \frac{1}{SNR} \mathbf{I} \right)^{-1} \mathbf{H}^T \mathbf{x}.
\end{aligned}$$

$$\mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} = \left(\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} + \frac{1}{\sigma_{\theta}^2} \mathbf{I} \right)^{-1}.$$