Estimation Theory

Fredrik Rusek

Chapter 10 + brief info on

- Conjugate priors
- Jeffrey's prior
- Reference priors

Chapter 10 – Bayesian Estimation

Previous chapters: No assumptions were made on θ

Chapter 10+ : We assume a prior distribution for θ

Benefits:

Problems:

Chapter 10 – Bayesian Estimation

Previous chapters: No assumptions were made on θ

Chapter 10+ : We assume a prior distribution for θ

Benefits:

More info -> better estimation precision

Optimal estimator (MSE sense) always exists

Performance is measured by a single value

Problems:

Chapter 10 – Bayesian Estimation

Previous chapters: No assumptions were made on θ

Chapter 10+ : We assume a prior distribution for θ

Benefits:

More info -> better estimation precision

Optimal estimator (MSE sense) always exists

Performance is measured by a single value

Problems:

Problematic to choose a prior distribution

Erroneous prior -> worse performance

Chapter 10 – Bayesian Estimation

Consider the problem of producing a radar to monitor a country's airspace

- France is about 500 km in diameter
- A reasonable prior for the distance from CDG to the airplane is U[0,600km]

Erroneous prior -> worse performance

Chapter 10 – Bayesian Estimation

Consider the problem of producing a radar to monitor a country's airspace

- France is about 500 km in diameter
- A reasonable prior for the distance from CDG to the airplane is U[0,600km]
- However, Belgium is about 90 km in diameter, reasonable prior is U[0,100km]

Erroneous prior -> worse performance

Chapter 10 – Bayesian Estimation

France and Belgium cannot buy the same equipment to monitor their airspaces if the manufacturer is using Bayesian estimation. With classical estimation they can since the estimator is optimized for all distances

- France is about 500 km in diameter
- A reasonable prior for the distance from CDG to the airplane is U[0,600km]
- However, Belgium is about 90 km in diameter, reasonable prior is U[0,100km]

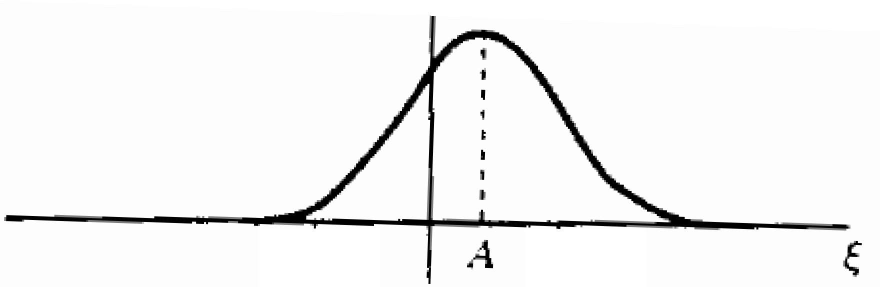
Erroneous prior -> worse performance

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again

- We know that the sample mean estimator $\hat{A} = \bar{x}$ is MVU
- We know that
  - Unbiased, i.e. $E(\bar{x})=A$
  - has variance $Var(\bar{x}) = \sigma^2/N$
- Therefore, the sample mean is distributed as

$$p_{\hat{A}}(\xi; A) = \frac{1}{\sqrt{2\pi\sigma^2/N}} e^{-\frac{1}{2\sigma^2/N}(\xi - A)^2}$$

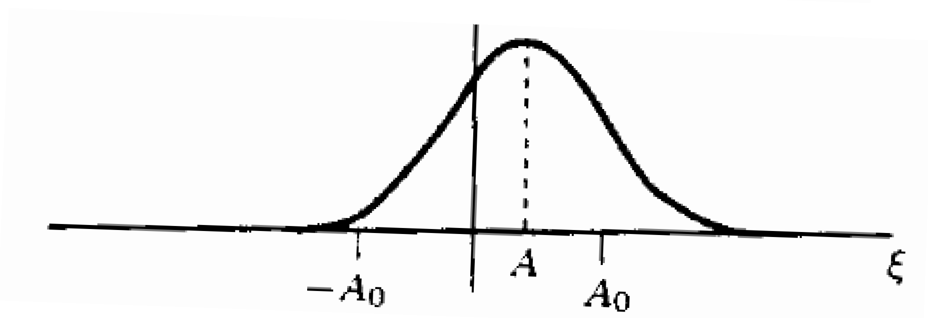


Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again

- We know that the sample mean estimator $\hat{A} = \bar{x}$ is MVU
- We know that
  - Unbiased, i.e. E($\bar{x}$)=A
  - has variance Var($\bar{x}$) = σ^2/N
- Therefore, the sample mean is distributed as

$$p_{\hat{A}}(\xi; A) = \frac{1}{\sqrt{2\pi\sigma^2/N}} e^{-\frac{1}{2\sigma^2/N}(\xi - A)^2}$$



Now assume that we know that

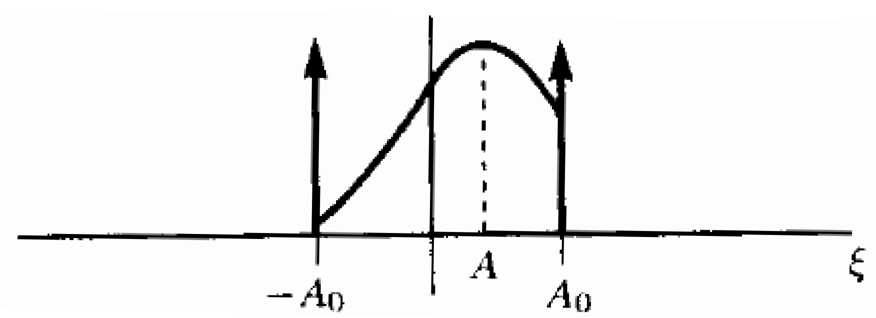
$$-A_0 \leq A \leq A_0$$

The MVU estimator fails to incorporate this info

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again

- We know that the sample mean estimator $\hat{A} = \bar{x}$ is MVU
- We know that
  - Unbiased, i.e. E($\bar{x}$)=A
  - has variance Var($\bar{x}$) = σ^2/N
- Therefore, the sample mean is distributed as



Define new estimator

$$\check{A} = \begin{cases} -A_0 & \bar{x} < -A_0 \\ \bar{x} & -A_0 \le \bar{x} \le A_0 \\ A_0 & \bar{x} > A_0 \end{cases}$$

Chapter 10 – Bayesian Estimation

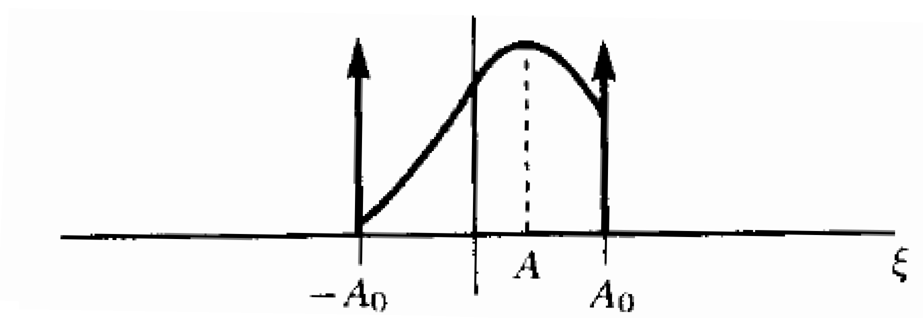
Consider DC-level estimation in white noise again

- We know that the sample mean estimator $\hat{A} = \bar{x}$ is MVU

PDF of new estimator is

$$
\begin{aligned}
p_{\check{A}}(\xi; A) \quad = \quad & \Pr\{\bar{x} \le -A_0\}\delta(\xi + A_0) \\
& + p_{\hat{A}}(\xi; A)[u(\xi + A_0) - u(\xi - A_0)] \\
& + \Pr\{\bar{x} \ge A_0\}\delta(\xi - A_0)
\end{aligned}
$$

Define new estimator



$$
\check{A} = \begin{cases} -A_0 & \bar{x} < -A_0 \\ \bar{x} & -A_0 \le \bar{x} \le A_0 \\ A_0 & \bar{x} > A_0 \end{cases}
$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again

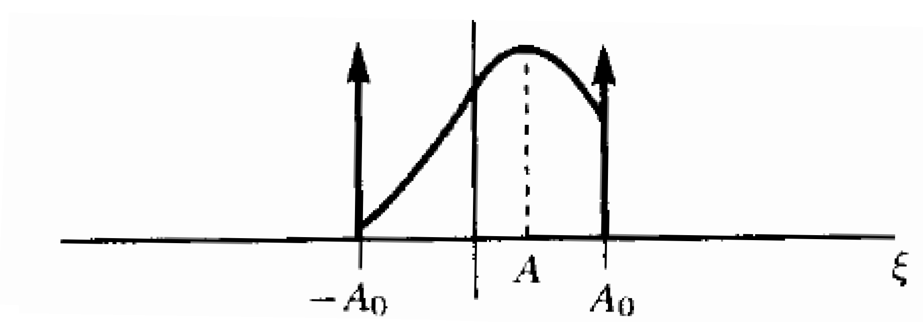
- We know that the sample mean estimator $\hat{A} = \bar{x}$ is MVU

PD

Checkpoint

Can we say that the new estimator is MVU given the prior info?

$$+\, p_{\hat{A}}(\xi; A)[u(\xi + A_0) - u(\xi - A_0)]$$
$$+\, \Pr\{\bar{x} \geq A_0\}\delta(\xi - A_0)$$

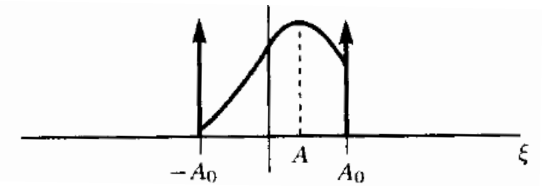


Define new estimator

$$\check{A} = \begin{cases} -A_0 & \bar{x} < -A_0 \\ \bar{x} & -A_0 \leq \bar{x} \leq A_0 \\ A_0 & \bar{x} > A_0 \end{cases}$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again

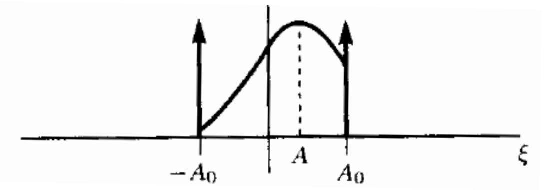


MSE of new estimator (depends on A)

$$\mathrm{mse}(\check{A}) =$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again



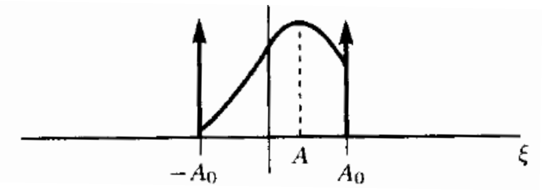
MSE of new estimator

$(-A_0-A)^2$ with probability $p(\bar{x} < -A_0) = \int_{-\infty}^{-A_0} p_{\hat{A}}(\xi; A)\, d\xi$

$$\mathrm{mse}(\check{A}) = \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again



MSE of new estimator

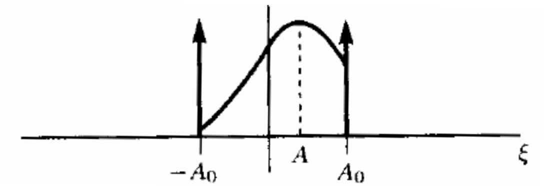
$(-A_0-A)^2$ with probability $p(\bar{x} < -A_0) = \int_{-\infty}^{-A_0} p_{\hat{A}}(\xi; A)\, d\xi$

$(-\xi-A)^2$ with probability $p(\bar{x} = \xi) = p_{\hat{A}}(\xi; A) \qquad -A_0 \leq \xi \leq A_0$

$$\mathrm{mse}(\check{A}) = \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again



MSE of new estimator

$(-A_0-A)^2$ with probability $p(\bar{x} < -A_0) = \int_{-\infty}^{-A_0} p_{\hat{A}}(\xi;A)\,d\xi$

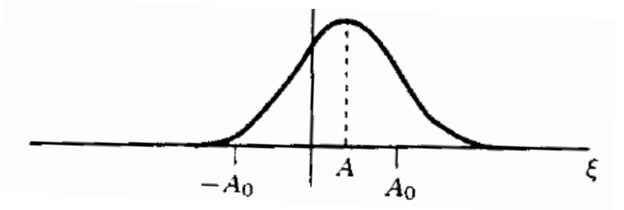
$(-\xi-A)^2$ with probability $p(\bar{x} = \xi) = p_{\hat{A}}(\xi;A) \qquad -A_0 \leq \xi \leq A_0$

$(A_0-A)^2$ with probability $p(\bar{x} \geq A_0) = \int_{A_0}^{\infty} p_{\hat{A}}(\xi;A)\,d\xi$

$$\boxed{\begin{aligned}\mathrm{mse}(\check{A}) = &\int_{-\infty}^{-A_0} (-A_0-A)^2 p_{\hat{A}}(\xi;A)\,d\xi + \int_{-A_0}^{A_0} (\xi-A)^2 p_{\hat{A}}(\xi;A)\,d\xi \\ &+ \int_{A_0}^{\infty} (A_0-A)^2 p_{\hat{A}}(\xi;A)\,d\xi\end{aligned}}$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again



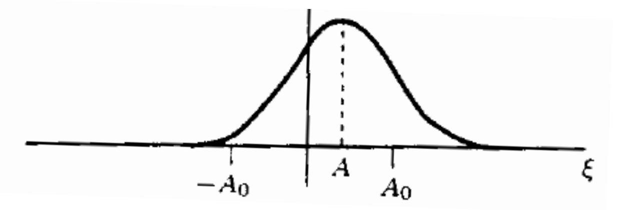
MSE of MVU estimator

$$\mathrm{mse}(\hat{A}) \quad =$$

$$\mathrm{mse}(\check{A}) = \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} (A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again



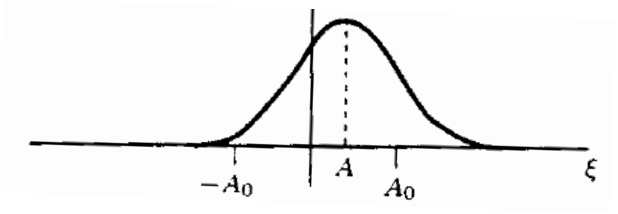
MSE of MVU estimator

$$\mathrm{mse}(\hat{A}) = \int_{-\infty}^{-A_0} p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} p_{\hat{A}}(\xi; A)\, d\xi$$

$$\mathrm{mse}(\check{A}) = \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} (A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again



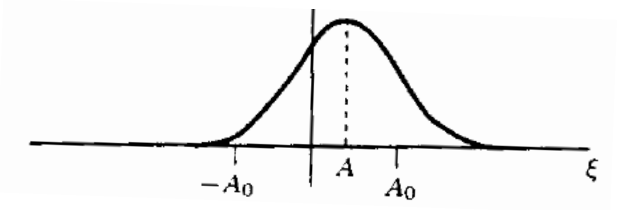
MSE of MVU estimator

$$\text{mse}(\hat{A}) = \int_{-\infty}^{-A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

$$\text{mse}(\check{A}) = \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} (A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again



MSE of MVU estimator

Bigger than

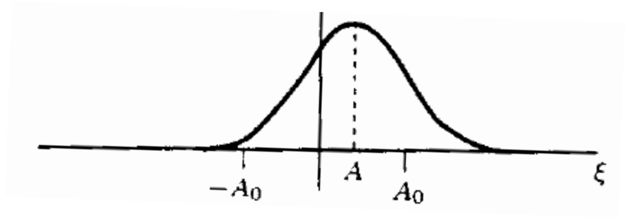
$$\operatorname{mse}(\hat{A}) = \int_{-\infty}^{-A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

$$\operatorname{mse}(\check{A}) = \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} (A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

Chapter 10 – Bayesian Estimation

Consider DC-level estimation in white noise again

$$\mathrm{mse}(\hat{A}) \;>\; \mathrm{mse}(\breve{A})$$



$$\mathrm{mse}(\hat{A}) = \int_{-\infty}^{-A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

$$\mathrm{mse}(\breve{A}) = \int_{-\infty}^{-A_0} (-A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{-A_0}^{A_0} (\xi - A)^2 p_{\hat{A}}(\xi; A)\, d\xi + \int_{A_0}^{\infty} (A_0 - A)^2 p_{\hat{A}}(\xi; A)\, d\xi$$

Chapter 10 – Bayesian Estimation

Lessons learned

- With prior information available, the MVU estimator can produce values that we know cannot be true

- With prior information, we can find an estimator that reduces the MSE compared with the MVU estimator for any A

Chapter 10 – Bayesian Estimation

Bayesian MSE

- In classical estimation, the MSE is defined as

$$\mathrm{mse}(\hat{A}) = \int (\hat{A} - A)^2 p(\mathbf{x}; A)\, d\mathbf{x}$$

which is a function of A

- In a Bayesian estimation, we can average this over the distribution of A

$$\boxed{\mathrm{Bmse}(\hat{A}) = E[(A - \hat{A})^2]}$$

Expectation is over both x and A (since A is random)

Chapter 10 – Bayesian Estimation

Bayesian MSE

- In classical estimation, the MSE is defined as

$$\mathrm{mse}(\hat{A}) = \int (\hat{A} - A)^2 p(\mathbf{x}; A)\, d\mathbf{x}$$

which is a function of A

- In a Bayesian estimation, we can average this over the distribution of A

$$\boxed{\begin{aligned}\mathrm{Bmse}(\hat{A}) = E[(A - \hat{A})^2] &= \int\int (A - \hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\ &= \int \mathrm{mse}(\hat{A})\, dA\end{aligned}}$$

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$\text{Bmse}(\hat{A}) = E[(A-\hat{A})^2] = \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA.$$

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$\text{Bmse}(\hat{A}) = E[(A-\hat{A})^2] = \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA$$

$$= \int \left[\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\right] p(\mathbf{x})\, d\mathbf{x}$$

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$\mathrm{Bmse}(\hat{A}) = E[(A-\hat{A})^2] = \iint (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA$$

$$= \int \underbrace{\left[\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\right]}_{} p(\mathbf{x})\, d\mathbf{x}$$

To minimize the BMSE, we should minimize this

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$\text{Bmse}(\hat{A}) = E[(A-\hat{A})^2] = \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA$$

$$= \int \underbrace{\left[\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\right]}_{} p(\mathbf{x})\, d\mathbf{x}$$

To minimize the BMSE, we should minimize this
The optimizer is not a function of A, since we integrate over A

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$
\begin{aligned}
\mathrm{Bmse}(\hat{A}) = E[(A-\hat{A})^2] &= \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\
&= \int \left[\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA \right] p(\mathbf{x})\, d\mathbf{x}
\end{aligned}
$$

$$
\frac{\partial}{\partial \hat{A}} \int (A-\hat{A})^2 p(A|\mathbf{x})\, dA \;=\; \int \frac{\partial}{\partial \hat{A}} (A-\hat{A})^2 p(A|\mathbf{x})\, dA
$$

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$\begin{aligned}
\mathrm{Bmse}(\hat{A}) = E[(A-\hat{A})^2] &= \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\
&= \int \underbrace{\left[\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\right]}\, p(\mathbf{x})\, d\mathbf{x}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \hat{A}} \int (A-\hat{A})^2 p(A|\mathbf{x})\, dA &= \int \frac{\partial}{\partial \hat{A}} (A-\hat{A})^2 p(A|\mathbf{x})\, dA \\
&= \int -2(A-\hat{A}) p(A|\mathbf{x})\, dA
\end{aligned}$$

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$\mathrm{Bmse}(\hat{A}) = E[(A-\hat{A})^2] = \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA$$

$$= \int \left[\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\right] p(\mathbf{x})\, d\mathbf{x}$$

$$\begin{aligned}
\frac{\partial}{\partial \hat{A}} \int (A-\hat{A})^2 p(A|\mathbf{x})\, dA &= \int \frac{\partial}{\partial \hat{A}} (A-\hat{A})^2 p(A|\mathbf{x})\, dA \\
&= \int -2(A-\hat{A}) p(A|\mathbf{x})\, dA \\
&= -2 \int A p(A|\mathbf{x})\, dA + 2\hat{A} \underbrace{\int p(A|\mathbf{x})\, dA}_{=1}
\end{aligned}$$

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$\begin{aligned}\mathrm{Bmse}(\hat{A}) = E[(A-\hat{A})^2] &= \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\ &= \int \underbrace{\left[\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\right]}\, p(\mathbf{x})\, d\mathbf{x}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \hat{A}} \int (A-\hat{A})^2 p(A|\mathbf{x})\, dA &= \int \frac{\partial}{\partial \hat{A}} (A-\hat{A})^2 p(A|\mathbf{x})\, dA \\ &= \int -2(A-\hat{A}) p(A|\mathbf{x})\, dA \\ &= -2\int A p(A|\mathbf{x})\, dA + 2\hat{A} \underbrace{\int p(A|\mathbf{x})\, dA}_{=1}\end{aligned}$$

$$\boxed{\hat{A} = \int A p(A|\mathbf{x})\, dA = E(A|\mathbf{x})}$$

Chapter 10 – Bayesian Estimation

Derivation of optimal Bayesian estimator

$$\mathrm{Bmse}(\hat{A}) = E[(A - \hat{A})^2]$$

The mean of the posterior distribution minimizes the BMSE
This is the MMSE estimator

$$\boxed{\hat{A} = \int A p(A|\mathbf{x})\, dA = E(A|\mathbf{x})}$$

Chapter 10 – Bayesian Estimation

Baye's rule allows us to evaluate the posterior distribution

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

From which we can evaluate the MMSE estimator

$$\boxed{\hat{A} = \int A p(A|\mathbf{x})\,dA = E(A|\mathbf{x})}$$

Chapter 10 – Bayesian Estimation

Baye's rule allows us to evaluate the posterior distribution

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

From which we can evaluate the MMSE estimator

Usually, one gets stuck either here or here

$$
\boxed{\hat{A} = \int A p(A|\mathbf{x})\,dA = E(A|\mathbf{x})}
$$

Chapter 10 – Bayesian Estimation

Baye's rule allows us to evaluate the posterior distribution

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

From which we can evaluate the MMSE estimator

Usually, one gets stuck either here or here

$$\boxed{\hat{A} = \int A p(A|\mathbf{x})\,dA = E(A|\mathbf{x})}$$

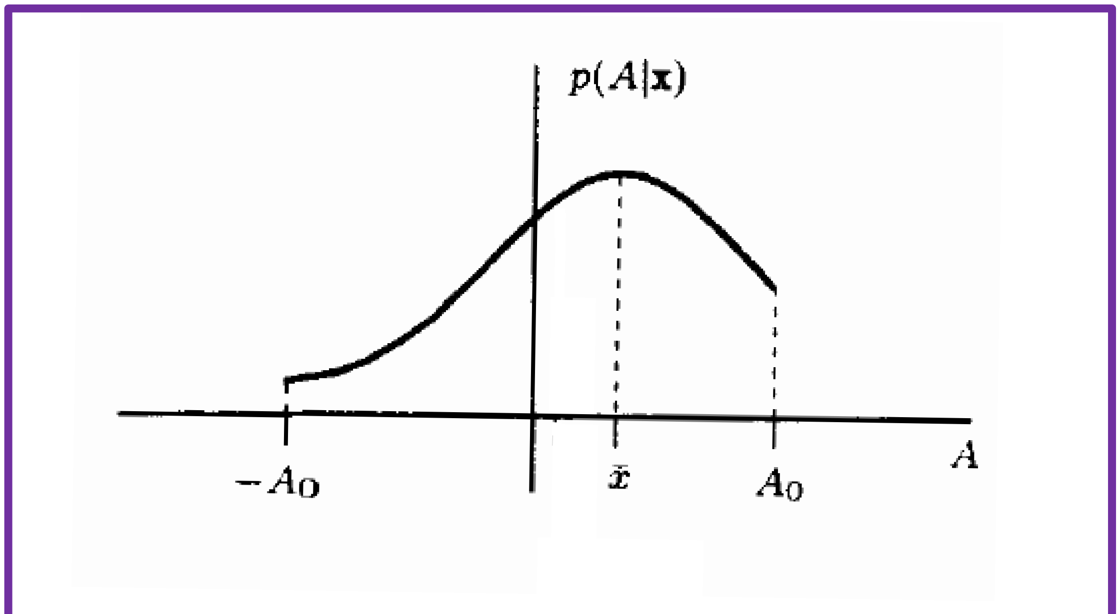
In DC level in WGN with uniform prior we get stuck at both places

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

Chapter 10 – Bayesian Estimation

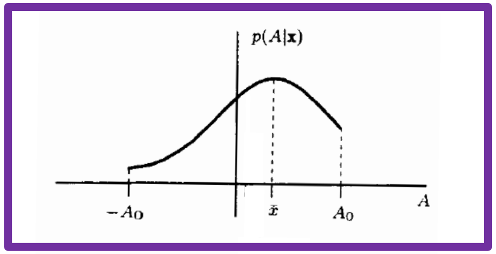
Rationale of MMSE estimator

DC-level, uniform prior $[-A_0, A_0]$



Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator

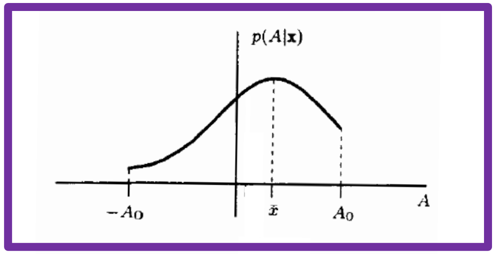


DC-level, uniform prior [-A_0,A_0]

Checkpoint: Why is the posterior maximized at the sample mean ?

Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator



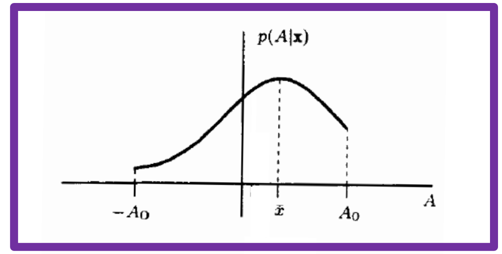
DC-level, uniform prior $[-A_0,A_0]$

Checkpoint: Why is the posterior maximized at the sample mean ?

1. **Sample mean is efficient**

Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator



DC-level, uniform prior $[-A_0, A_0]$

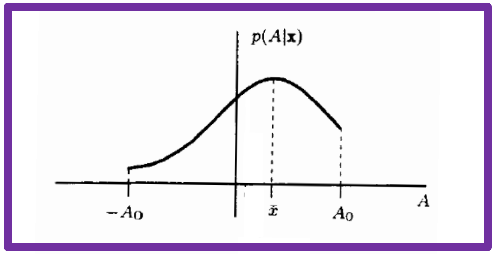
Checkpoint: Why is the posterior maximized at the sample mean ?

1. **Sample mean is efficient**
2. **Therefore, the sample mean coincides with the MLE**

Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator

DC-level, uniform prior $[-A_0,A_0]$



Checkpoint: Why is the posterior maximized at the sample mean ?

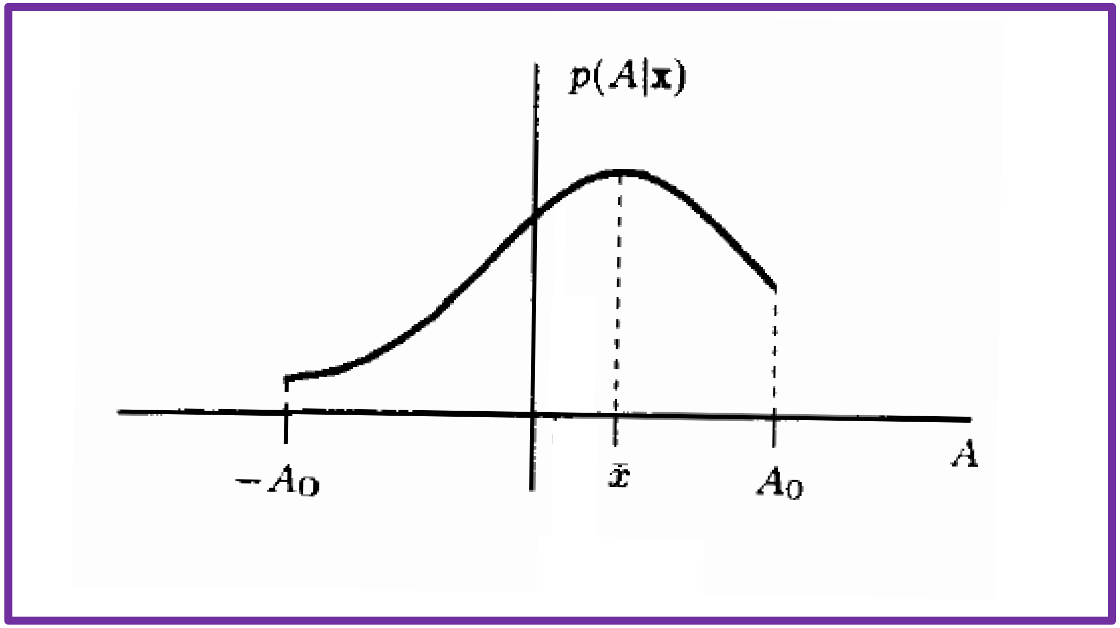
1. **Sample mean is efficient**
2. **Therefore, the sample mean coincides with the MLE**
3. **But due to the uniform prior**

$$\arg\max_A p(A|x) = \arg\max_A p(x|A)p(A) = \arg\max_A p(x|A) = \text{MLE}$$

Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator

DC-level, uniform prior [$-A_0$,A_0]

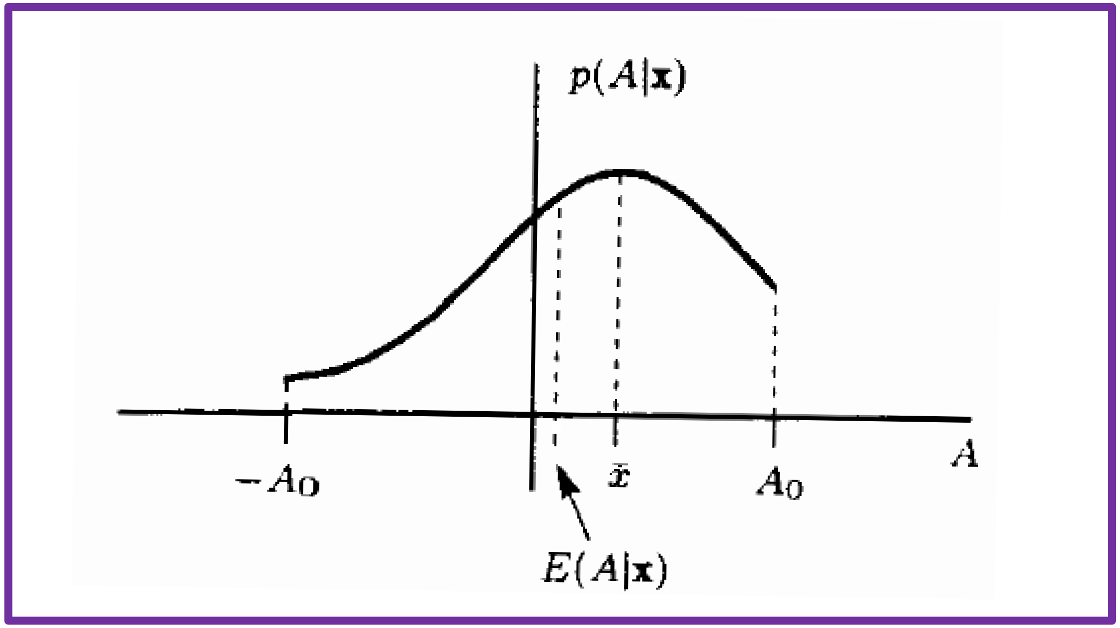


Where is the posterior mean?

Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator

DC-level, uniform prior $[-A_0,A_0]$

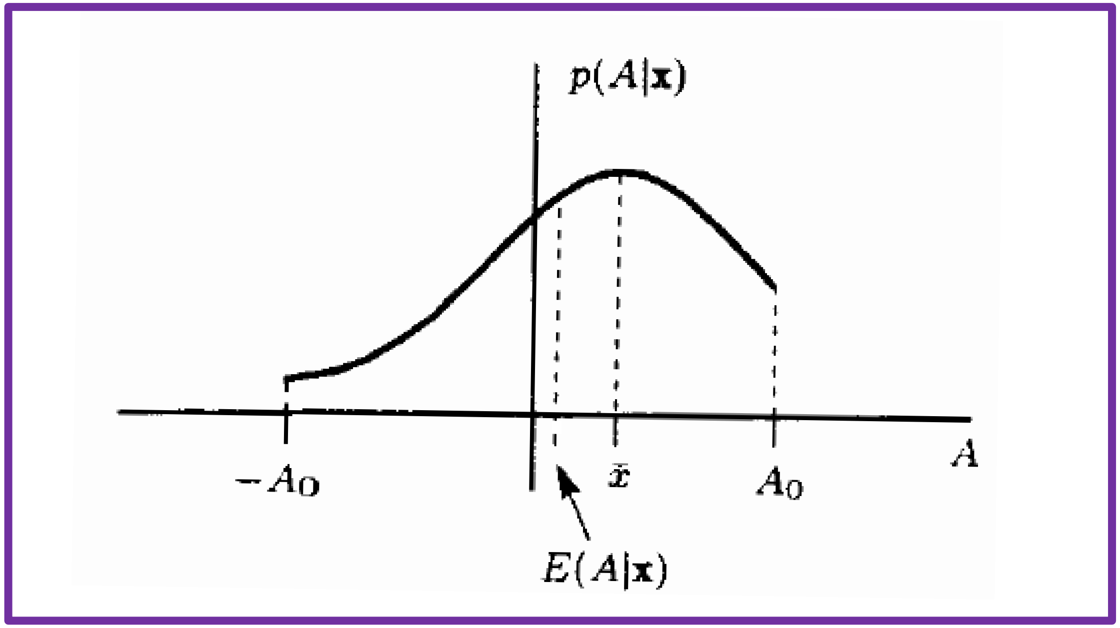


Why does the MMSE shrink the sample mean towards zero ?

Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator

DC-level, uniform prior [-A_0,A_0]



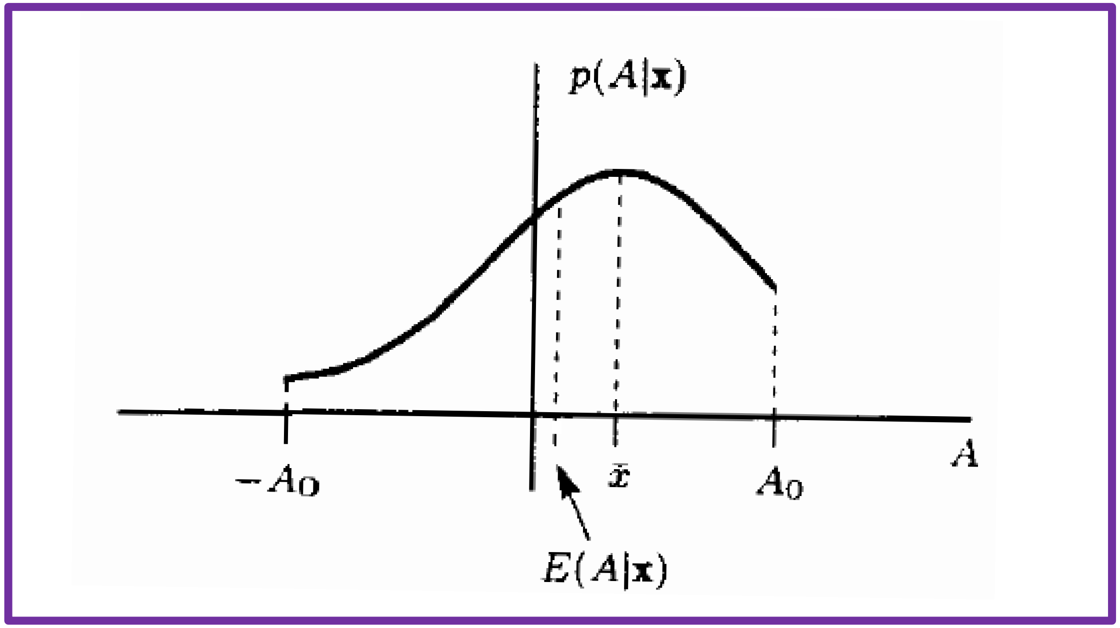
Because in the absence of data, the MMSE estimator is $\hat{A} = E(A) = 0$

Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator

The MMSE estimator compromises between the information in the prior (A = zero) and the information in the data (A = sample mean)

Estimator is biased towards the prior mean



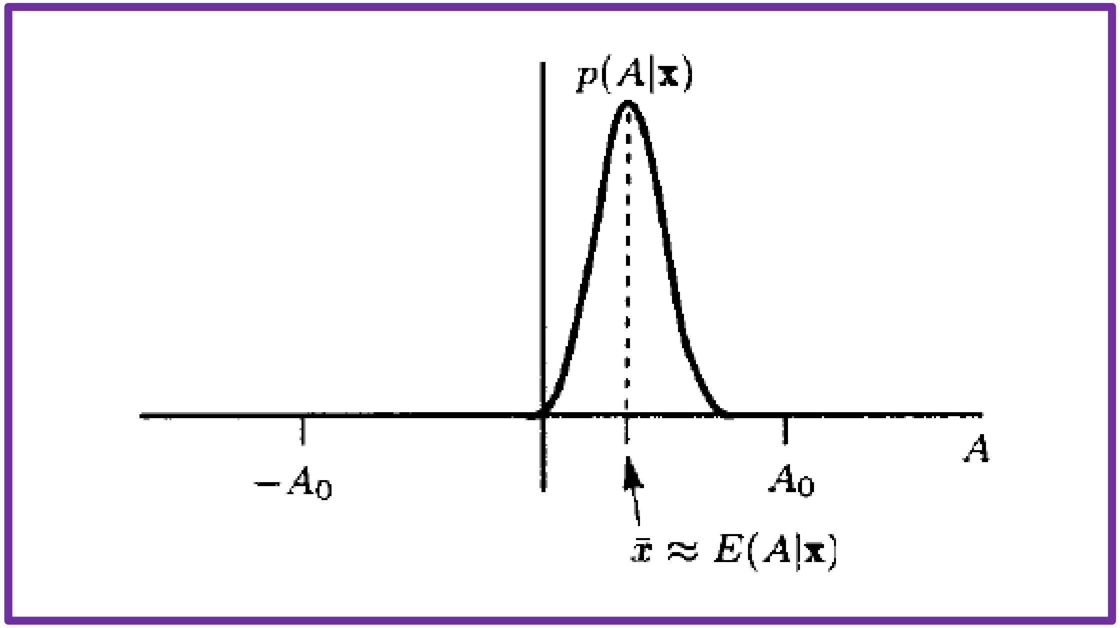
$$\hat{A} = E(A) = 0$$

Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator

With lots of data and/or good SNR, the MMSE ignores the prior

Bias of estimator is low

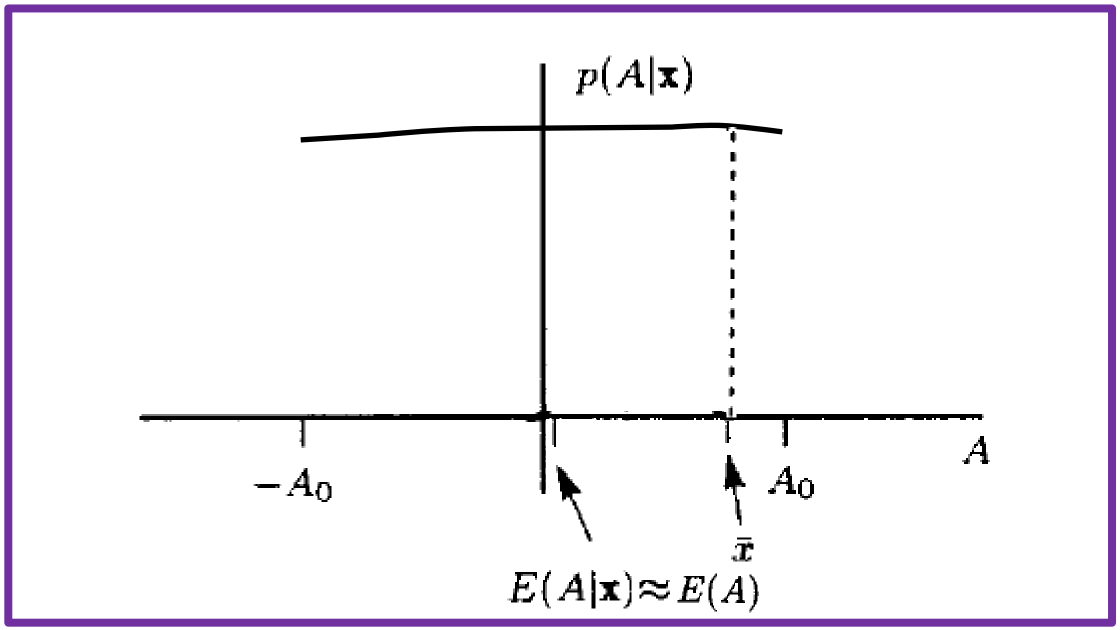


Chapter 10 – Bayesian Estimation

Rationale of MMSE estimator

With a very bad SNR, the MMSE basically ignores the data

Bias of estimator is high



Chapter 10 – Bayesian Estimation

Bernstein - von Mises Theorem (A.k.a. Bayesian central limit theorem)

Given a prior $p(\theta)$ and IID observations $X_1,\ldots,X_N$

Given some mild regularity conditions

$$p(\theta|\mathbf{X}) \to \mathcal{N}(\theta, I^{-1}(\theta)), \quad N \to \infty$$

Or in other words,

- The prior is not important asymptotically
- Posterior coincides with MLE estimate

Chapter 10 – Bayesian Estimation

Remaining problems

- **Select a prior that models the nature of the parameter well**
- **Make sure that the prior allows for computations of**

$$\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta} \quad \text{and} \quad E(\theta|\mathbf{x})$$

The second problem is, e.g., solved with Gaussian priors

Chapter 10 – Bayesian Estimation

Conditional Gaussian distribution

Theorem 10.2 (Conditional PDF of Multivariate Gaussian) *If* x *and* θ *are jointly Gaussian, where* $\mathbf{x}$ *is* $k \times 1$ *and* θ *is* $l \times 1$, *with mean vector* $[E(x)^T \, E(\theta)^T]^T$ *and partitioned covariance matrix*

$$\mathbf{C} = \begin{bmatrix} C_{xx} & C_{x\theta} \\ C_{\theta x} & C_{\theta\theta} \end{bmatrix} = \begin{bmatrix} k \times k & k \times l \\ l \times k & l \times l \end{bmatrix} \tag{10.23}$$

so that

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

then the conditional PDF $p(\theta|x)$ *is also Gaussian and*

$$\begin{aligned} E(\theta|x) &= E(\theta) + \mathbf{C}_{\theta y}\mathbf{C}_{xx}^{-1}(x - E(x)) \\ \mathbf{C}_{\theta|x} &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}. \end{aligned}$$

Note that conditional variance does not depend on x

Chapter 10 – Bayesian Estimation

"Bayesian" Linear model

Consider a model according to **x=Hθ+w, w** ~ $\mathcal{N}(\mathbf{0}, \mathbf{C}_w)$

Chapter 10 – Bayesian Estimation

”Bayesian” Linear model

Consider a model according to **x=Hθ+w, w** ~ $\mathcal{N}(\mathbf{0}, \mathbf{C}_w)$

Assign a normal prior to **θ:** $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$

We are interested in computing p(**θ|x**), and in particular E(**θ|x**)

Chapter 10 – Bayesian Estimation

"Bayesian" Linear model

Consider a model according to **x=Hθ+w, w** ~ $\mathcal{N}(\mathbf{0}, \mathbf{C}_w)$

Assign a normal prior to **θ:** $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$

We are interested in computing p(**θ|x**), and in particular E(**θ|x**)

Make observation that **x** and **θ** are jointly Gaussian **Theorem 2 Applies**

Chapter 10 – Bayesian Estimation

Conditional Gaussian distribution

Theorem 10.2 (Conditional PDF of Multivariate Gaussian) *If* x *and* θ *are jointly Gaussian, where* $\mathbf{x}$ *is* $k \times 1$ *and* θ *is* $l \times 1$, *with mean vector* $[E(\mathrm{x})^T \; E(\theta)^T]^T$ *and partitioned covariance matrix*

$$\mathbf{C} = \begin{bmatrix} C_{xx} & C_{x\theta} \\ C_{\theta x} & C_{\theta\theta} \end{bmatrix} = \begin{bmatrix} k \times k & k \times l \\ l \times k & l \times l \end{bmatrix} \tag{10.23}$$

so that

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

then the conditional PDF $p(\theta|x)$ *is also Gaussian and*

$$\begin{aligned} E(\theta|x) &= E(\theta) + \mathbf{C}_{\theta y}\mathbf{C}_{xx}^{-1}(x - E(x)) \\ \mathbf{C}_{\theta|x} &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}. \end{aligned}$$

Note that conditional variance does not depend on x

Chapter 10 – Bayesian Estimation

then the conditional PDF $p(\theta|x)$ *is also Gaussian and*

$$
\begin{aligned}
E(\theta|x) &= E(\theta) + \mathbf{C}_{\theta y}\mathbf{C}_{xx}^{-1}(x - E(x)) \\
\mathbf{C}_{\theta|x} &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}\,.
\end{aligned}
$$

x=Hθ+w

We need to compute

E(**θ**), but this is $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ by assumption $\mathcal{N}(\boldsymbol{\mu}_{\theta}, \mathbf{C}_{\theta})$

Chapter 10 – Bayesian Estimation

then the conditional PDF $p(\theta|x)$ *is also Gaussian and*

$$\begin{aligned} E(\theta|x) &= E(\theta) + \mathbf{C}_{\theta y}\mathbf{C}_{xx}^{-1}(x - E(x)) \\ \mathbf{C}_{\theta|x} &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}. \end{aligned}$$

x=Hθ+w

We need to compute

E(**θ**), but this is $\boldsymbol{\mu}_\theta$ by assumption $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$

E(**x**), but this is easy: $\mathbf{H}\boldsymbol{\mu}_\theta$

Chapter 10 – Bayesian Estimation

then the conditional PDF $p(\theta|x)$ *is also Gaussian and*

$$
\begin{aligned}
E(\theta|x) &= E(\theta) + \mathbf{C}_{\theta y}\mathbf{C}_{xx}^{-1}(x - E(x)) \\
\mathbf{C}_{\theta|x} &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}.
\end{aligned}
$$

x=Hθ+w

We need to compute

E(**θ**), but this is $\boldsymbol{\mu}_\theta$ by assumption $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$

E(**x**), but this is easy: $\mathbf{H}\boldsymbol{\mu}_\theta$

$\mathbf{C}_{\theta\theta}$, but this is $\mathbf{C}_\theta$ by assumption

Chapter 10 – Bayesian Estimation

then the conditional PDF $p(\theta|x)$ *is also Gaussian and*

$$\begin{aligned} E(\theta|x) &= E(\theta) + \mathbf{C}_{\theta y}\mathbf{C}_{xx}^{-1}(x - E(x)) \\ \mathbf{C}_{\theta|x} &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}. \end{aligned}$$

x=Hθ+w

We need to compute

E(**θ**), but this is $\boldsymbol{\mu}_\theta$ by assumption $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$

E(**x**), but this is easy: $\mathbf{H}\boldsymbol{\mu}_\theta$

$\mathbf{C}_{\theta\theta}$, but this is $\mathbf{C}_\theta$ by assumption

$\mathbf{C}_{xx}$,easy as well: $\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T+\mathbf{C}_w$

Chapter 10 – Bayesian Estimation

then the conditional PDF $p(\theta|x)$ *is also Gaussian and*

$$
\begin{aligned}
E(\theta|x) &= E(\theta) + \mathbf{C}_{\theta y}\mathbf{C}_{xx}^{-1}(x - E(x)) \\
\mathbf{C}_{\theta|x} &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}\,.
\end{aligned}
$$

x=Hθ+w

We need to compute

E(**θ**), but this is $\boldsymbol{\mu}_\theta$ by assumption $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$

E(**x**), but this is easy: $\mathbf{H}\boldsymbol{\mu}_\theta$

$\mathbf{C}_{\theta\theta}$, but this is $\mathbf{C}_\theta$ by assumption

$\mathbf{C}_{xx}$,,easy as well: $\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T + \mathbf{C}_w$

$\mathbf{C}_{\theta x}$, which is $\mathbf{C}_\theta\mathbf{H}^T$

Chapter 10 – Bayesian Estimation

Collect everything to get

Theorem 10.3 (Posterior PDF for the Bayesian General Linear Model) *If the observed data* $\mathbf{x}$ *can be modeled as*

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{10.27}$$

where $\mathbf{x}$ *is an* $N \times 1$ *data vector,* $\mathbf{H}$ *is a known* $N \times p$ *matrix,* $\boldsymbol{\theta}$ *is a* $p \times 1$ *random vector with prior PDF* $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$, *and* $\mathbf{w}$ *is an* $N \times 1$ *noise vector with PDF* $\mathcal{N}(\mathbf{0}, \mathbf{C}_w)$ *and independent of* $\boldsymbol{\theta}$, *then the posterior PDF* $p(\boldsymbol{\theta}|\mathbf{x})$ *is Gaussian with mean*

$$E(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + \mathbf{C}_\theta\mathbf{H}^T(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T + \mathbf{C}_w)^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta) \tag{10.28}$$

and covariance

$$\mathbf{C}_{\theta|x} = \mathbf{C}_\theta - \mathbf{C}_\theta\mathbf{H}^T(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^T + \mathbf{C}_w)^{-1}\mathbf{H}\mathbf{C}_\theta. \tag{10.29}$$

Note:

1. **H can be of reduced rank**
2. **Estimator is linear in observation x**

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear) | **Noise covariance** | **Prior**

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

$C_w = \mathbf{I}\sigma^2$

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$E(A|\mathbf{x}) =$ **Apply Theorem 10.3**

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)

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

Noise covariance

$$\mathbf{C_w} = \mathbf{I}\sigma^2$$

Prior

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$$E(A|\mathbf{x}) = \quad \ldots\ldots\ldots \quad = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}(\bar{x} - \mu_A)$$

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)	Noise covariance	Prior
$\mathbf{x} = \mathbf{1}A + \mathbf{w}$	$\mathbf{C_w}=\mathbf{I}\sigma^2$	$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$

$$E(A|\mathbf{x}) = \quad \ldots\ldots\ldots \quad = \quad \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}(\bar{x} - \mu_A)$$

Estimator with no data

Correction term accounting for observed data

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)

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

Noise covariance

$$\mathbf{C_w} = \mathbf{I}\sigma^2$$

Prior

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$$E(A|\mathbf{x}) = \quad \ldots\ldots\ldots \quad = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \underbrace{(\bar{x} - \mu_A)}$$

If sample mean agrees with prior mean, then there is no update

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)

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

Noise covariance

$$\mathbf{C_w} = \mathbf{I}\sigma^2$$

Prior

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$$E(A|\mathbf{x}) = \quad \ldots\ldots\ldots \quad = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \underbrace{(\bar{x} - \mu_A)}$$

If N is very large, or σ^2 very small, estimator is the sample mean

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)	Noise covariance	Prior
$\mathbf{x} = \mathbf{1}A + \mathbf{w}$	$\mathbf{C_w}=\mathbf{I}\sigma^2$	$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$

$$E(A|\mathbf{x}) = \quad \ldots\ldots\ldots \quad = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \underbrace{(\bar{x} - \mu_A)}$$

If prior is very "un-informative" $\sigma^2_A \approx \infty$, estimator is the sample mean

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)

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

Noise covariance

$$\mathbf{C_w}=\mathbf{I}\sigma^2$$

Prior

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$$E(A|\mathbf{x}) = \quad \ldots\ldots\ldots \quad = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}(\bar{x} - \mu_A)$$

If prior distribution is very precise $\sigma^2_A \approx 0$, then $E(A|\mathbf{x}) = \mu_A$

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)

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

Noise covariance

$$\mathbf{C_w} = \mathbf{I}\sigma^2$$

Prior

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$$\mathrm{var}(A|\mathbf{x}) = \frac{\sigma^2}{N}\left(\frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\right)$$

This is not the Bmse !

$$\begin{aligned}\mathrm{Bmse}(\hat{A}) &= \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\ &= \int\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\, p(\mathbf{x})\, d\mathbf{x}\end{aligned}$$

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)

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

Noise covariance

$$\mathbf{C_w} = \mathbf{I}\sigma^2$$

Prior

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$$\mathrm{var}(A|\mathbf{x}) = \frac{\sigma^2}{N}\left(\frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\right)$$

This is not the Bmse !

$$\begin{aligned}
\mathrm{Bmse}(\hat{A}) &= \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\
&= \int\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\, p(\mathbf{x})\, d\mathbf{x} \\
&= \int\int [A - E(A|\mathbf{x})]^2 p(A|\mathbf{x})\, dA\, p(\mathbf{x})\, d\mathbf{x} \\
&= \int \mathrm{var}(A|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}. =
\end{aligned}$$

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)

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

Noise covariance

$$\mathbf{C_w} = \mathbf{I}\sigma^2$$

Prior

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$$\mathrm{Bmse}(\hat{A}) = \frac{\sigma^2}{N}\left(\frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\right)$$

This is not the Bmse in general, but happens to be in this case…..!

$$\begin{aligned}
\mathrm{Bmse}(\hat{A}) &= \iint (A - \hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\
&= \iint (A - \hat{A})^2 p(A|\mathbf{x})\, dA\, p(\mathbf{x})\, d\mathbf{x} \\
&= \iint [A - E(A|\mathbf{x})]^2 p(A|\mathbf{x})\, dA\, p(\mathbf{x})\, d\mathbf{x} \\
&= \int \mathrm{var}(A|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}. = \mathrm{var}(A|\mathbf{x})
\end{aligned}$$

Chapter 10 – Bayesian Estimation

DC-level estimation with Gaussian prior. x[n]=A+w[n]

Signal model (linear)	Noise covariance	Prior
$\mathbf{x} = \mathbf{1}A + \mathbf{w}$	$\mathbf{C_w} = \mathbf{I}\sigma^2$	$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$

$$\mathrm{Bmse}(\hat{A}) = \frac{\sigma^2}{N}\left(\frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\right)$$

With no prior info $\sigma^2_A \approx \infty$
BMSE = σ^2/N
With prior info $\sigma^2_A < \infty$
BMSE < σ^2/N

This is not the Bmse in general, but happens to be in this case…..!

$$\begin{aligned}
\mathrm{Bmse}(\hat{A}) &= \int\int (A-\hat{A})^2 p(\mathbf{x}, A)\, d\mathbf{x}\, dA \\
&= \int\int (A-\hat{A})^2 p(A|\mathbf{x})\, dA\, p(\mathbf{x})\, d\mathbf{x} \\
&= \int\int [A - E(A|\mathbf{x})]^2 p(A|\mathbf{x})\, dA\, p(\mathbf{x})\, d\mathbf{x} \\
&= \int \mathrm{var}(A|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}. = \mathrm{var}(A|\mathbf{x})
\end{aligned}$$

Chapter 10 – Bayesian Estimation

Reproducing property and conjugate priors

Notice what happened

Prior $p(\theta)$ = Uniform -> $p(x|\theta)$ -> Posterior $p(\theta|x)$ = not Uniform

Prior $p(\theta)$ = Gaussian -> $p(x|\theta)$ -> Posterior $p(\theta|x)$ = Gaussian

Second case is <u>much</u> easier to work with as we do not have to compute the pdf of the posterior, only its parameters. Reproducing property

Chapter 10 – Bayesian Estimation

Reproducing property and conjugate priors

Conjugate prior

For a conditional pdf $\mathbf{p(x|\theta)}$, a prior $\mathbf{p(\theta)}$ with the property that the posterior $\mathbf{p(\theta|x)}$ has the same form as $\mathbf{p(\theta)}$ is said to be a conjugate prior

Very desirable property for analytically establishing the MMSE estimator

Long tables of conjugate priors exist

Conditional pdfs from the exponential family have conjugate priors

Chapter 10 – Bayesian Estimation

Jeffrey's prior

DC-level in white noise

If we know that $1 < A < 2$, then it is reasonable with a uniform prior U[1,2]

Chapter 10 – Bayesian Estimation

Jeffrey's prior

DC-level in white noise

If we know that $1 < A < 2$, then it is reasonable with a uniform prior U[1,2]

But, one could also say that it is the power that should be uniform, so $p(A^2)$ is U[1,4]

Chapter 10 – Bayesian Estimation

Jeffrey's prior

DC-level in white noise

If we know that $1 < A < 2$, then it is reasonable with a uniform prior U[1,2]

But, one could also say that it is the power that should be uniform, so $p(A^2)$ is U[1,4]

Maybe power in dBs is uniform, so $p(\log A^2)$ is U(0,0.6)

Maybe there is some other parametrization of space that makes sense?

Chapter 10 – Bayesian Estimation

Jeffrey's prior

Now note what happens

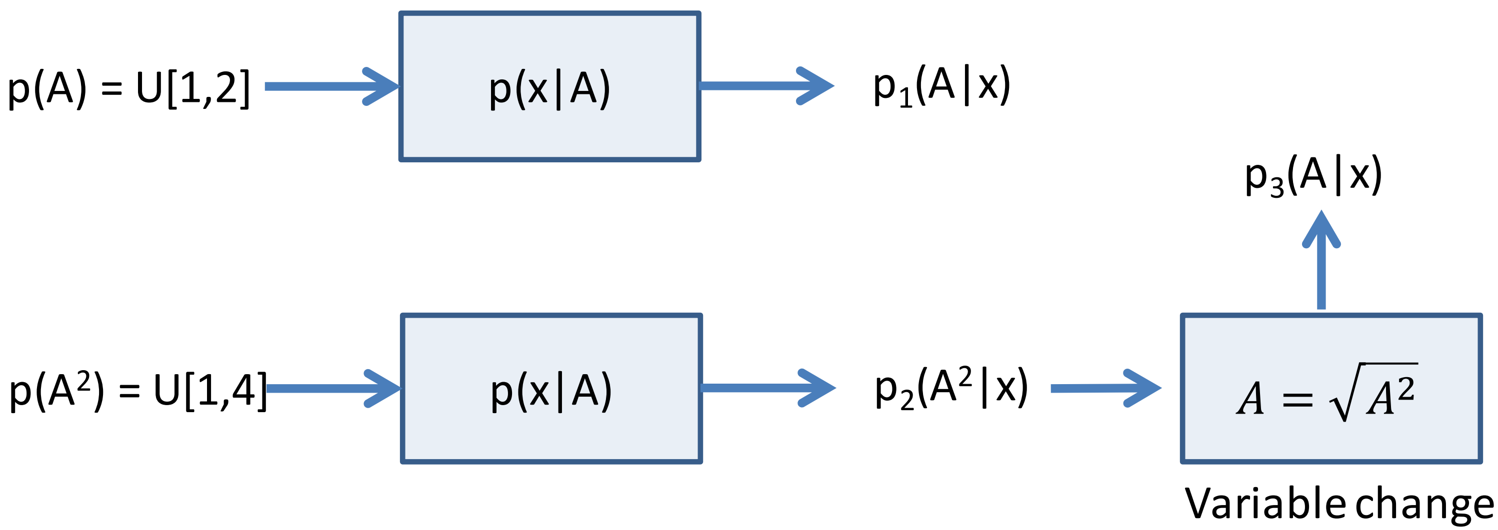
$$p(A) = U[1,2] \rightarrow \boxed{p(x|A)} \rightarrow p_1(A|x)$$

$$p(A^2) = U[1,4] \rightarrow \boxed{p(x|A)} \rightarrow p_2(A^2|x)$$

Chapter 10 – Bayesian Estimation

Jeffrey's prior

Now note what happens

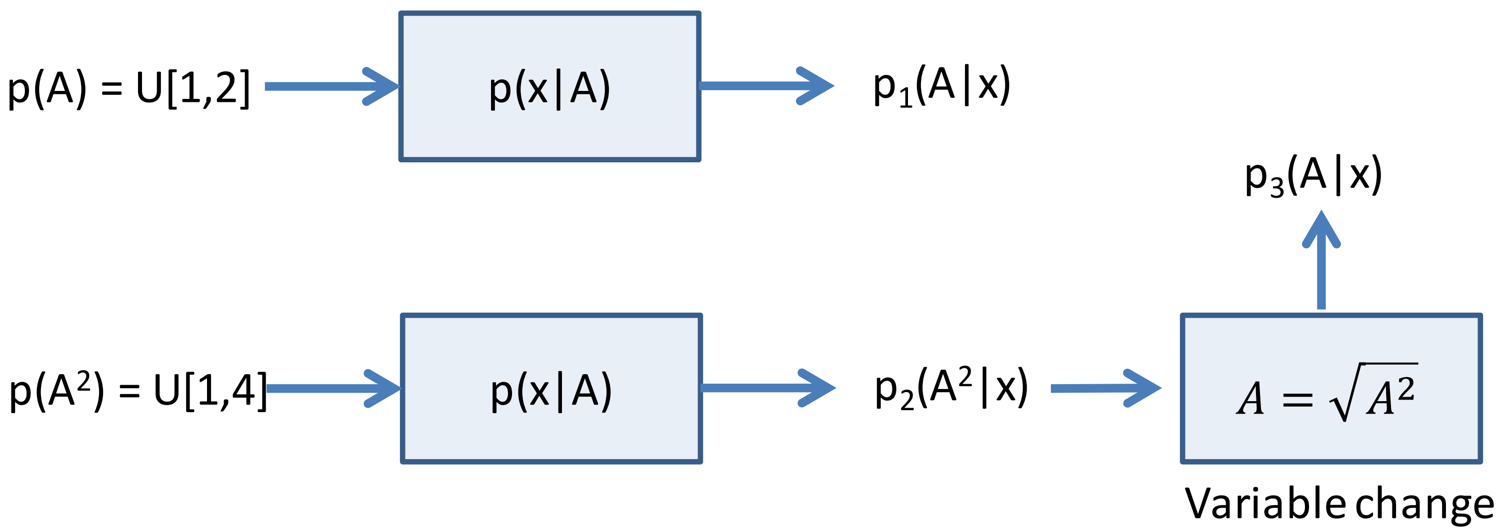


Do $p_1(A|x)$ and $p_3(A|x)$ match? That is, does the two different priors represent the same thing?

Chapter 10 – Bayesian Estimation

Jeffrey's prior

Now note what happens



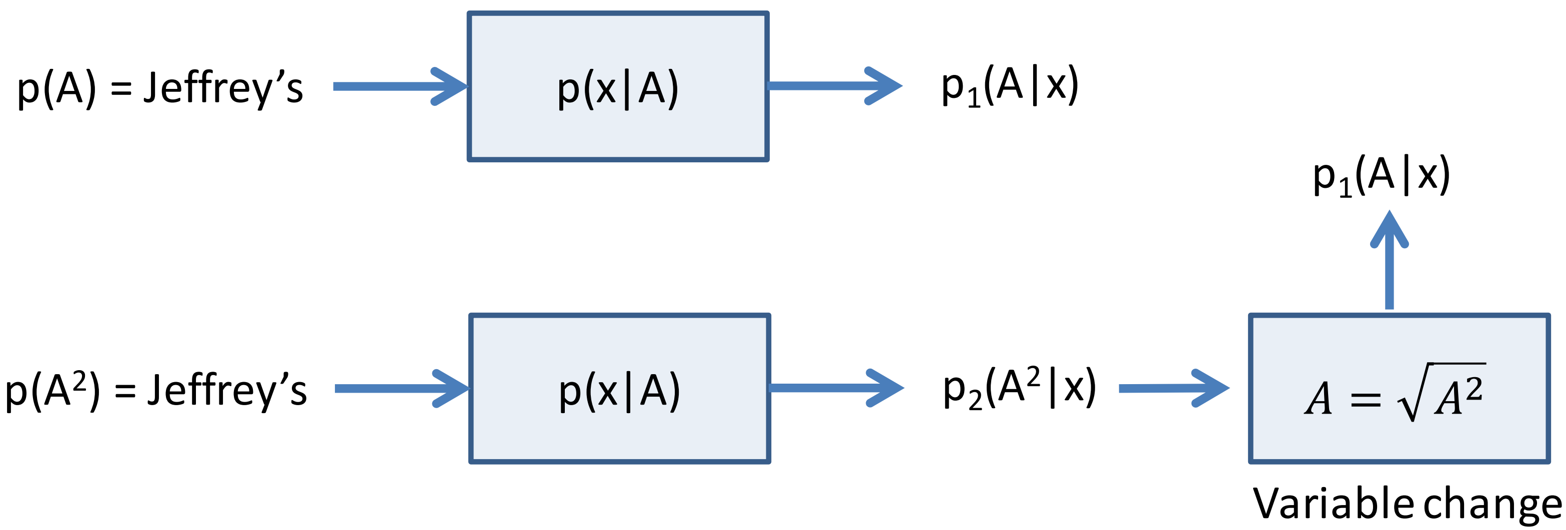
Do $p_1(A|x)$ and $p_3(A|x)$ match? That is, does the two different priors represent the same thing?

NO!!!

Chapter 10 – Bayesian Estimation

Jeffrey’s prior

Jeffrey’s prior is invariant under re-parameterization of space

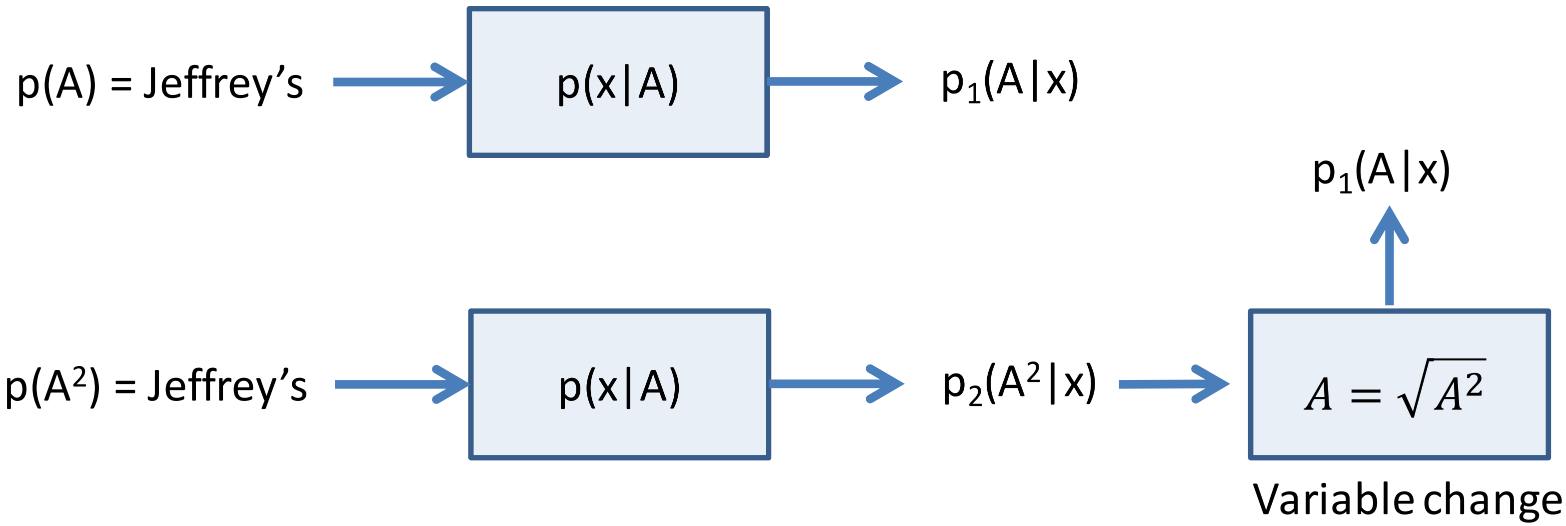


With Jeffrey’s prior it does not matter what parameterization we use, the results are invariant

Chapter 10 – Bayesian Estimation

Jeffrey's prior

Jeffrey's prior is invariant under re-parameterization of space



Jeffrey's prior:

$$p(\theta) \propto \sqrt{I(\theta)}$$

Chapter 10 – Bayesian Estimation

Jeffrey's prior

Jeffrey's prior is invariant under re-parameterization of space

Nobody said that Jeffrey's is alwyas representing the nature of the Parameter well, but it is a decent choice if the scale of the parameter is not known/understood

Variable change

Jeffrey's prior:

$$p(\theta) \propto \sqrt{I(\theta)}$$

Chapter 10 – Bayesian Estimation

Reference priors

Consider the situation once again

Prior $p(A)$ → Observe x → posterior $p(A|x)$

In order for the data observation to be meaningful, it should provide information of A

Chapter 10 – Bayesian Estimation

Reference priors

Consider the situation once again

Prior $p(A)$ → $\boxed{\text{Observe } x}$ → posterior $p(A|x)$

In order for the data observation to be meaningful, it should provide information of A

The prior should not dominate, it should be ”un-informative”

Chapter 10 – Bayesian Estimation

Reference priors

Consider the situation once again

Prior $p(A)$ → Observe x → posterior $p(A|x)$

In order for the data observation to be meaningful, it should provide information of A

The prior should not dominate, it should be "un-informative"

The posterior and the prior should be far from each other

Chapter 10 – Bayesian Estimation

Reference priors

Consider the situation once again

Prior $p(A)$ → Observe x → posterior $p(A|x)$

This is measured by the Kullback Leibler distance

$$\int p(A|x) log \frac{p(A|x)}{p(A)} dA$$

The posterior and the prior should be far from each other

Chapter 10 – Bayesian Estimation

Reference priors

Consider the situation once again

Prior p(A) → Observe x → posterior p(A|x)

This is measured by the Kullback Leibler distance

$$\int p(A|x) log \frac{p(A|x)}{p(A)} dA$$

Take expectation over x

$$\int p(x) \int p(A|x) log \frac{p(A|x)}{p(A)} dA \, dx$$

Chapter 10 – Bayesian Estimation

Reference priors

Consider the situation once again

Prior p(A) → Observe x → posterior p(A|x)

This is measured by the Kullback Leibler distance

$$\int p(A|x) log \frac{p(A|x)}{p(A)} dA$$

THIS IS MUTUAL INFORMATION BETWEEN A AND X

$$I(A;X) = \int p(x) \int p(A|x) log \frac{p(A|x)}{p(A)} dA\, dx$$

Chapter 10 – Bayesian Estimation

Reference priors

Consider the situation once again

$$\text{Prior } p(A) \longrightarrow \boxed{\text{Observe } x} \longrightarrow \text{posterior } p(A|x)$$

Reference prior

$$\boxed{p^*(A) = \max_{p(A)} I(A;X)}$$

Maximizes the contribution from the observed data. Provides the least information possible

Chapter 10 – Bayesian Estimation

Section 10.8 Bayesian estimators for deterministic parameters

If no MVU estimator exists, or is very hard to find, we can apply an MMSE estimator To deterministic parameters

Recall the form of the Bayesian estimator for DC-levels in WGN

$$E(A|\mathbf{x}) = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}(\bar{x} - \mu_A) = \alpha\bar{x} + (1-\alpha)\mu_A \qquad \alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}$$

Compute the MSE for a given value of A

Chapter 10 – Bayesian Estimation

Section 10.8 Bayesian estimators for deterministic parameters

If no MVU estimator exists, or is very hard to find, we can apply an MMSE estimator To deterministic parameters

Recall the form of the Bayesian estimator for DC-levels in WGN

$$E(A|\mathbf{x}) = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}(\bar{x} - \mu_A) = \alpha\bar{x} + (1-\alpha)\mu_A$$

$\alpha<1$

Compute the MSE for a given value of A

$$\begin{aligned}\text{mse}(\hat{A}) &= \text{var}(\hat{A}) + b^2(\hat{A}) \\ &= \alpha^2\text{var}(\bar{x}) + [\alpha A + (1-\alpha)\mu_A - A]^2 \\ &= \alpha^2\frac{\sigma^2}{N} + (1-\alpha)^2(A-\mu_A)^2.\end{aligned}$$

Chapter 10 – Bayesian Estimation

Section 10.8 Bayesian estimators for deterministic parameters

If no MVU estimator exists, or is very hard to find, we can apply an MMSE estimator To deterministic parameters

Recall the form of the Bayesian estimator for DC-levels in WGN

$$E(A|\mathbf{x}) = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}(\bar{x} - \mu_A) = \alpha\bar{x} + (1-\alpha)\mu_A$$

α<1

Compute the MSE for a given value of A

$$\begin{aligned}\operatorname{mse}(\hat{A}) &= \operatorname{var}(\hat{A}) + b^2(\hat{A}) \\ &= \alpha^2 \operatorname{var}(\bar{x}) + [\alpha A + (1-\alpha)\mu_A - A]^2 \\ &= \alpha^2 \frac{\sigma^2}{N} + (1-\alpha)^2 (A - \mu_A)^2.\end{aligned}$$

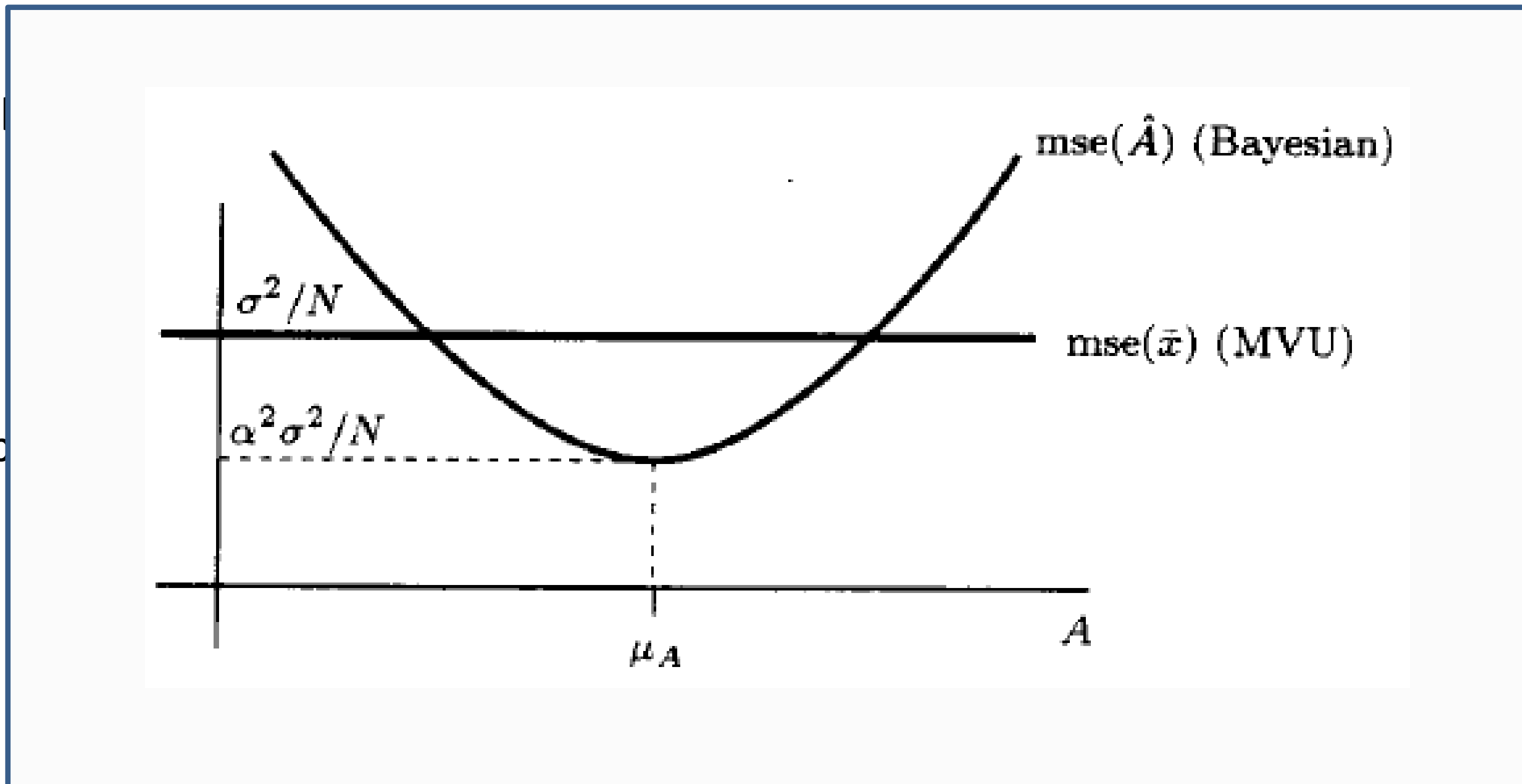
Variance smaller than classical estimator **Large bias for large A**

Chapter 10 – Bayesian Estimation

Section 10.8 Bayesian estimators for deterministic parameters

Recal

Comp



MSE for Bayesian is smaller for A close to the prior mean, but larger far away

Chapter 10 – Bayesian Estimation

Section 10.8 Bayesian estimators for deterministic parameters

However, the BMSE is smaller

$$\mathrm{Bmse}(\hat{A}) = E_A[\mathrm{mse}(\hat{A})]$$

$$\mathrm{mse}(\hat{A}) = \mathrm{var}(\hat{A}) + b^2(\hat{A})$$

$$\begin{aligned} &= \alpha^2 \mathrm{var}(\bar{x}) + [\alpha A + (1-\alpha)\mu_A - A]^2 \\ &= \alpha^2 \frac{\sigma^2}{N} + (1-\alpha)^2 (A - \mu_A)^2. \end{aligned}$$

Chapter 10 – Bayesian Estimation

Section 10.8 Bayesian estimators for deterministic parameters

However, the BMSE is smaller

$$\begin{aligned}\mathrm{Bmse}(\hat{A}) &= E_A[\mathrm{mse}(\hat{A})] \\ &= \alpha^2\frac{\sigma^2}{N} + (1-\alpha)^2 E_A[(A-\mu_A)^2]\end{aligned}$$

$$\begin{aligned}\mathrm{mse}(\hat{A}) &= \mathrm{var}(\hat{A}) + b^2(\hat{A}) \\ &= \alpha^2\mathrm{var}(\bar{x}) + [\alpha A + (1-\alpha)\mu_A - A]^2 \\ &= \alpha^2\frac{\sigma^2}{N} + (1-\alpha)^2(A-\mu_A)^2.\end{aligned}$$

Chapter 10 – Bayesian Estimation

Section 10.8 Bayesian estimators for deterministic parameters

However, the BMSE is smaller

$$\begin{aligned}\mathrm{Bmse}(\hat{A}) &= E_A[\mathrm{mse}(\hat{A})]\\ &= \alpha^2\frac{\sigma^2}{N} + (1-\alpha)^2 E_A[(A-\mu_A)^2]\\ &= \alpha^2\frac{\sigma^2}{N} + (1-\alpha)^2\sigma_A^2\end{aligned}$$

$$\mathrm{mse}(\hat{A}) = \mathrm{var}(\hat{A}) + b^2(\hat{A})$$

$$\begin{aligned}&= \alpha^2\mathrm{var}(\bar{x}) + [\alpha A + (1-\alpha)\mu_A - A]^2\\ &= \alpha^2\frac{\sigma^2}{N} + (1-\alpha)^2(A-\mu_A)^2.\end{aligned}$$

Chapter 10 – Bayesian Estimation

Section 10.8 Bayesian estimators for deterministic parameters

However, the BMSE is smaller

$$\begin{aligned}
\mathrm{Bmse}(\hat{A}) &= E_A[\mathrm{mse}(\hat{A})] \\
&= \alpha^2 \frac{\sigma^2}{N} + (1-\alpha)^2 E_A[(A-\mu_A)^2] \\
&= \alpha^2 \frac{\sigma^2}{N} + (1-\alpha)^2 \sigma_A^2 \\
&= \frac{\sigma^2}{N} \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \\
&< \frac{\sigma^2}{N} = \mathrm{Bmse}(\bar{x}),
\end{aligned}$$

$$\alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}$$

$$\mathrm{mse}(\hat{A}) = \mathrm{var}(\hat{A}) + b^2(\hat{A})$$

$$\begin{aligned}
&= \alpha^2 \mathrm{var}(\bar{x}) + [\alpha A + (1-\alpha)\mu_A - A]^2 \\
&= \alpha^2 \frac{\sigma^2}{N} + (1-\alpha)^2 (A-\mu_A)^2.
\end{aligned}$$