

# Estimación y Predicción en Series Temporales

Estimador maximo a posteriori (MAP)

Departamento de Procesamiento de Señales

Instituto de Ingeniería Eléctrica  
Facultad de Ingeniería

2022

- 1 Repaso Estimadores de máxima verosimilitud (MLE)
- 2 Metodología Bayesiana
- 3 Estimador Bayesiano de mínimo error cuadrático medio
- 4 Estimador Máximo a Posteriori (MAP)

Estimador de máxima verosimilitud (MLE)

# Estimador de Máxima Verosimilitud (MLE)

- En el problema de estimación de parámetros, una alternativa al estimador MVU es el Estimador de Máxima Verosimilitud (*Maximum Likelihood Estimator – MLE*) .
- Es la herramienta más popular para obtener estimadores prácticos ya que puede ser utilizado en problemas de estimación complejos, o en problemas donde el MVU no existe o no puede encontrarse.
- Tiene características asintóticas deseables:
  - es asintóticamente eficiente
  - es consistente
  - es invariante a re-parametrizaciones
- En muchos casos no puede encontrarse una fórmula cerrada para el MLE y se deben utilizar métodos numéricos.

# MLE: Descripción intuitiva

- Se observa un conjunto de datos  $\{x[0], x[1], \dots, x[N-1]\}$  que dependen de cierto parámetro desconocido  $\theta$  que se quiere estimar.
- Especificación del Modelo: los datos son generados por un proceso aleatorio caracterizado por cierta PDF:

$$p(\mathbf{x}; \theta), \quad \text{donde } \theta \in [a, b].$$

- Al variar el parámetro desconocido, se cambia la PDF que modela la generación de datos.
- El modelo es definido como una familia de PDFs indexada por el parámetro desconocido.
- Para estimar el parámetro desconocido, la idea es encontrar la PDF de la familia que **maximiza la probabilidad de haber generado los datos observados**.

**Estimador de Máxima Verosimilitud (MLE)** Si se asume el modelo de generación de datos dada por la PDF  $p(\mathbf{x}; \theta)$ , es decir

$$\mathbf{x} \sim p(\mathbf{x}; \theta)$$

y se observa  $\mathbf{x}_0$ , entonces el estimador de **máxima verosimilitud (MLE)** es

$$\hat{\theta}_{\text{MLE}}(\mathbf{x}_0) = \arg \max_{\theta \in \mathcal{D}_{\theta}} \log p(\mathbf{x} = \mathbf{x}_0; \theta).$$

- El MLE se define como el valor de  $\theta$  que maximiza el logaritmo de  $p(\mathbf{x}; \theta)$  fijando  $\mathbf{x}$ , es decir el valor que maximiza la func. de verosimilitud logarítmica.
- Observar que el tomar el logaritmo no afecta la posición del máximo de la función (logaritmo es función monótona creciente)

**Teorema.**

Si existe un estimador eficiente, el método de máxima verosimilitud permite encontrarlo.

*Demostración:*

- Por el teorema de Cramér-Rao, si existe un estimador eficiente, existen las funciones  $g(\mathbf{x})$  y  $I(\theta)$  tal que

$$\frac{\partial \log p(\mathbf{x}; \theta)}{\partial \theta} = I(\theta) (g(\mathbf{x}) - \theta).$$

El estimador eficiente es  $\hat{\theta}_{crlb} = g(\mathbf{x})$  con varianza  $I^{-1}(\theta)$ .

- Como el MLE es el valor de  $\theta$  que maximiza la función de verosimilitud logarítmica se tiene que

$$\left. \frac{\partial \log p(\mathbf{x}; \theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_{MLE}} = I(\hat{\theta}_{MLE}) (g(\mathbf{x}) - \hat{\theta}_{MLE}) = 0,$$

y por lo tanto

$$\hat{\theta}_{MLE} = g(\mathbf{x}) = \hat{\theta}_{crlb}.$$

## Teorema. Propiedades asintóticas del MLE.

Si la PDF  $p(\mathbf{x}; \theta)$  de los datos satisface ciertas condiciones de regularidad, el MLE del parámetro desconocido  $\theta$  es asintóticamente distribuido como

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

- $I(\theta)$  es la información de Fisher evaluada en el valor verdadero del parámetro desconocido.
- Las condiciones de regularidad son: (i) Existencia de las derivadas primera y segunda de la función de verosimilitud; (ii) condición de regularidad necesaria para teorema de CRLB.
- El MLE es asintóticamente eficiente y por lo tanto **asintóticamente óptimo**.

(Prueba. Ver Kay [1993], apéndice 7B.)

## Observaciones

- La expresión analítica de la PDF verdadera (no asintótica) del MLE en general imposible de derivar.
- En la práctica, no se sabe cuan grande debe ser  $N$  para estar cerca del comportamiento asintótico. Se suelen utilizar simulaciones numéricas para estudiar el desempeño.



## Teorema: Propiedad de invarianza del MLE

El MLE del parámetro  $\alpha = g(\theta)$ , donde la PDF  $p(\mathbf{x}, \alpha)$  está parametrizada por  $\theta$ , está dado por

$$\hat{\alpha} = g(\hat{\theta}),$$

donde  $\hat{\theta}$  es el MLE de  $\theta$ .

- el MLE de  $\theta$  se obtiene maximizando  $p(\mathbf{x}; \theta)$ .
- Si  $g$  no es una función biyectiva,  $\hat{\alpha}$  maximiza la función de verosimilitud modificada  $\bar{p}(\mathbf{x}; \alpha)$ , definida como

$$\bar{p}(\mathbf{x}; \alpha) = \max_{\{\theta: \alpha = g(\theta)\}} p(\mathbf{x}; \theta).$$

# Extensión a vector de parámetros

- Análogamente al caso escalar, el MLE para un vector de parámetros  $\theta$  es el valor que maximiza la función de verosimilitud  $p(\mathbf{x}; \theta)$  sobre todo el rango válido de  $\theta$ .
- Asumiendo que la función de verosimilitud es diferenciable, el MLE se encuentra como

$$\frac{\partial \log p(\mathbf{x}; \theta)}{\partial \theta} = \mathbf{0}.$$

- En caso de existir múltiples soluciones, el MLE es aquella que maximiza la función de verosimilitud, es decir aquella que produce el máximo global.

## Teorema: Propiedades asintóticas del MLE

Si la PDF  $p(\mathbf{x}; \boldsymbol{\theta})$  de los datos  $\mathbf{x}$  satisface ciertas condiciones de regularidad, el MLE del parámetro desconocido  $\boldsymbol{\theta}$  es asintóticamente distribuido como,

$$\hat{\boldsymbol{\theta}} \stackrel{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})),$$

- $\mathbf{I}(\boldsymbol{\theta})$  es la matriz de información de Fisher evaluada en el valor verdadero del parámetro desconocido.
- Las condiciones de regularidad son:
  - Existencia de las derivadas de primer y segundo orden de la función de verosimilitud.
  - Además se requiere la condición de regularidad (idem a CRLB),

$$\mathbb{E} \left[ \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] = \mathbf{0} \quad \forall \boldsymbol{\theta}.$$

## Estimación Bayesiana y Máximo a Posteriori (MAP)

# Enfoque Clásico (frecuentista) *versus* Bayesiano

- **Enfoque clásico o frecuentista:**  $\theta$  parámetro desconocido es una variable determinística desconocida. Sólo los datos son “aleatorios”
- **Enfoque Bayesiano:** considera incertidumbre sobre  $\theta$ : el parámetro desconocido es también considerado una variable aleatoria

## Motivación enfoque Bayesiano:

- Posibilita incorporar conocimiento previo sobre  $\theta$ , por ejemplo mediante un *prior* o distribución a priori  $p(\theta)$ .
- Si el prior es razonable, puede conducir a estimaciones más precisas.
- Útil cuando no podemos encontrar un MVU. Por ejemplo, si la varianza de un estimador insesgado no es uniformemente menor que la de todos los estimadores insesgados (e.g., no se cumple en  $\forall \theta$ ).
- Podría cumplirse que existe un estimador que minimiza el MSE para la **mayoría** de los valores de  $\theta$ .
- Si asignamos una RDE  $\pi(\theta)$ , es posible encontrar un estimador

# Motivación

Mostraremos con un ejemplo como el conocimiento a priori conduce a estimadores más exactos. **Ejemplo: Nivel de DC en WGN** Los

datos observados son

$$x[n] = A + w[n] \quad \text{con } n = 0, 1, \dots, N-1 \text{ y } w[n] \sim \mathcal{N}(0, \sigma^2) \quad \forall n,$$

- Se vio en clases anteriores que el MVU era  $\bar{x} = \sum_{n=0}^{N-1} x[n]$  (media muestral).
- Supongamos que sabemos que  $-A_0 \leq A \leq A_0$  con  $A_0 < \infty$  (conocido).
- Observar que  $\bar{x}$  puede conducir a estimaciones fuera de  $[-A_0, A_0]$  (debido al ruido). Claramente  $\bar{x}$  no será el mejor estimador en este caso.
- Postulemos el estimador media muestral truncada.

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

## Ejemplo: Estimación de nivel de DC en WGN

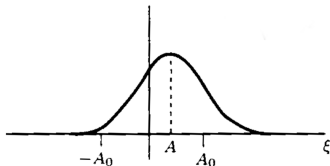
- ¿Cómo es la PDF de  $\hat{A}$  (media muestral),  $p_{\hat{A}}(\xi; A)$ ?

$$p_{\hat{A}}(\xi; A) = \frac{1}{\sqrt{2\pi\sigma^2/N}} \exp \left[ -\frac{1}{2\sigma^2/N} (\xi - A)^2 \right]$$

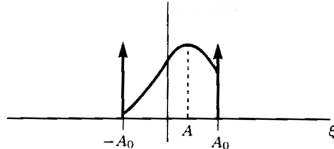
- ¿Cómo es la PDF de  $\check{A}$  (media muestral truncada),  $p_{\check{A}}(\xi; A)$ ?

$$\begin{aligned} p_{\check{A}}(\xi; A) = & \Pr(\bar{x} < -A_0) \delta(\xi + A_0) \\ & + p_{\hat{A}}(\xi; A) [H(\xi + A_0) - H(\xi - A_0)] \\ & + \Pr(\bar{x} > A_0) \delta(\xi - A_0). \end{aligned}$$

PDF media muestral



PDF media truncada



**Ejemplo:** Estimación de nivel de DC en WGN Comparación  $\hat{A}$  contra  $\check{A}$ .

- Estimador  $\check{A}$  tiene sesgo ( $\hat{A}$  insesgado)
- ¿Qué sucede con el MSE? ¿Cuál de los dos estimadores tiene menos MSE?

Sea  $A \in [-A_0, A_0]$ , entonces,

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

- Media muestral truncada es **mejor** estimador en cuanto al MSE.

**Observación.**  $\hat{A}$  sigue siendo el MVU, pero introduciendo un sesgo podemos reducir la varianza del estimador y por lo tanto reducir el MSE.



# MSE clásico y MSE Bayesiano (BMSE)

## MSE Clásico

- Si  $\theta$  es un parámetro desconocido, dado un estimador  $\hat{\theta}(\mathbf{x})$  definimos el MSE como

$$\text{mse}(\hat{\theta}) = \mathbb{E}_{\mathbf{x}} \left[ (\hat{\theta} - \theta)^2 \right] = \int_{\mathbf{x}} \left( \hat{\theta}(\mathbf{x}) - \theta \right)^2 p(\mathbf{x}; \theta) d\mathbf{x}.$$

- El resultado de optimizar el MSE puede depender de  $\theta$ .

## MSE Bayesiano

- Si contamos con conocimiento previo de  $\theta$ , por ejemplo que no puede ser negativo, o debe estar cerca de cierto valor (restricciones físicas).
- Si  $\theta$  es una variable aleatoria (realización desconocida) que sigue una distribución  $p(\theta)$ , dado un estimador  $\hat{\theta}(\mathbf{x})$  definimos el BMSE como

$$\text{bmse}(\hat{\theta}) = \mathbb{E}_{\mathbf{x}, \theta} \left[ (\hat{\theta} - \theta)^2 \right] = \int_{\theta} \int_{\mathbf{x}} \left( \hat{\theta}(\mathbf{x}) - \theta \right)^2 p(\mathbf{x}, \theta) d\mathbf{x} d\theta.$$

- En este caso el resultado no depende de  $\theta$  (promedio).

# MSE Bayesiano

Estimador *Minimum Mean Square Error*

**Objetivo:** Encontrar  $\hat{\theta} = \hat{\theta}(\mathbf{x})$  que minimice el BMSE (estimador MMSE).

$$\begin{aligned}\text{bmse}(\hat{\theta}) &= \mathbb{E}_{\mathbf{x}, \theta} \left[ (\hat{\theta} - \theta)^2 \right] \\ &= \int_{\theta} \int_{\mathbf{x}} \left( \hat{\theta}(\mathbf{x}) - \theta \right)^2 p(\mathbf{x}, \theta) d\mathbf{x} d\theta \\ &= \int_{\mathbf{x}} p(\mathbf{x}) \int_{\theta} \left( \hat{\theta}(\mathbf{x}) - \theta \right)^2 p(\theta | \mathbf{x}) d\theta d\mathbf{x}\end{aligned}$$

Como  $p(\mathbf{x}) \geq 0, \forall \mathbf{x}$ , minimizar  $\text{bmse}(\hat{\theta})$  equivale a minimizar

$$\begin{aligned}\forall \mathbf{x}, \int_{\theta} (\hat{\theta} - \theta)^2 p(\theta | \mathbf{x}) d\theta, \\ \frac{\partial}{\partial \hat{\theta}} \int_{\theta} (\hat{\theta} - \theta)^2 p(\theta | \mathbf{x}) d\theta &= 2 \int_{\theta} (\hat{\theta} - \theta) p(\theta | \mathbf{x}) d\theta \\ &= 2\hat{\theta} \int_{\theta} p(\theta | \mathbf{x}) d\theta - 2 \int_{\theta} \theta p(\theta | \mathbf{x}) d\theta \\ &= 2\hat{\theta} \underbrace{\int_{\theta} p(\theta | \mathbf{x}) d\theta}_{=1} - 2 \underbrace{\int_{\theta} \theta p(\theta | \mathbf{x}) d\theta}_{\mathbb{E}[\theta | \mathbf{x}]} = 0\end{aligned}$$

Con lo cual

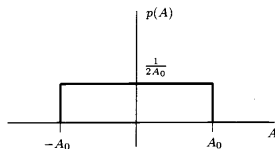
$$\hat{\theta}_{\text{mmse}} = \mathbb{E}[\theta | \mathbf{x}], \quad \forall \mathbf{x}. \quad (\text{media del posterior } p(\theta | \mathbf{x}))$$

# BMSE: Observaciones

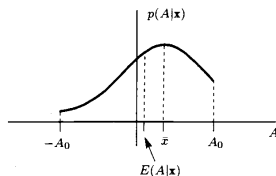
- El estimador MMSE (*minimum mean square error*) basado en BMSE se define como aquél que minimiza el promedio del MSE sobre todas las realizaciones posibles del parámetro desconocido  $\theta$ .
- Este estimador resulta ser la media de la distribución a posteriori  $p(\theta|\mathbf{x})$ ,

$$\hat{\theta}_{\text{mmse}} = \mathbb{E}[\theta|\mathbf{x}], \quad \forall \mathbf{x}.$$

- Intuición.** Antes de observar los datos la distribución de  $\theta$  es  $p(\theta)$ . Una vez observados los datos el parámetro  $\theta$  se distribuye según  $p(\theta|\mathbf{x})$  (posterior).



(a) Prior PDF



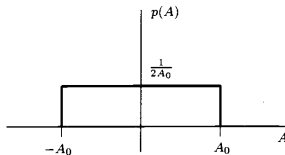
(b) Posterior PDF

**Ejemplo:** Estimación de nivel de DC en WGN Los datos observados son

$$x[n] = A + w[n] \quad \text{con } n = 0, 1, \dots, N-1 \text{ y } w[n] \sim \mathcal{N}(0, \sigma^2) \quad \forall n,$$

- Supongamos que sabemos que  $-A_0 \leq A \leq A_0$  con  $A_0 < \infty$  (conocido).
- Si no sabemos nada más sobre  $A_0$ , lo razonable es asumir que dentro del intervalo  $[-A_0, A_0]$  no hay ningún valor preferido, es decir  $A \sim \mathcal{U}([-A_0, A_0])$ .

$$p(A) = \begin{cases} \frac{1}{2A_0} & |A| \leq A_0 \\ 0 & |A| > A_0 \end{cases}$$



# Estimador MMSE

## Ejemplo: Estimación de nivel de DC en WGN

- Para determinar el estimador MMSE necesitamos la *posterior*  $p(A|\mathbf{x})$

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

- **Obs.** Denominador no depende de  $A$  (constante para normalizar, integral 1)
- Para especificar  $p(\mathbf{x}|A)$  necesitamos asumir  $A$  y  $w[n]$  son VAs independientes.

$$\begin{aligned} p_{\mathbf{x}}(x[n]|A) &= p_w(x[n] - A|A) \\ &= p_w(x[n] - A) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} (x[n] - A)^2 \right]. \end{aligned}$$

Con lo cual

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

- Notar que  $p(\mathbf{x}|A)$  tiene exactamente la misma forma que  $p(\mathbf{x}; A)$ . Sin embargo las dos expresiones tienen significados distintos:
  - $p(\mathbf{x}|A)$  PDF condicional dado  $A$  (separador “|”)

# Estimador MMSE

## Ejemplo: Estimación de nivel de DC en WGN

$$p(A|\mathbf{x}) = \begin{cases} \frac{\frac{1}{2A_0(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right]}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right] dA} & |A| \leq A_0 \\ 0 & |A| > A_0 \end{cases}$$

Observando que

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

se tiene

$$p(A|\mathbf{x}) = \begin{cases} \frac{1}{c\sqrt{2\pi\frac{\sigma^2}{N}}} \exp\left[-\frac{1}{2\frac{\sigma^2}{N}} (A - \bar{x})^2\right] & |A| \leq A_0 \\ 0 & |A| > A_0. \end{cases}$$

Donde la constante  $c$  es un factor de normalización (no depende de  $A$ )

$$c = \int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{N}}} \exp\left[-\frac{1}{2\frac{\sigma^2}{N}} (A - \bar{x})^2\right] dA.$$

## Ejemplo: Estimación de nivel de DC en WGN

El posterior  $p(A|\mathbf{x})$  es una Gaussiana truncada. El estimador MMSE es por lo tanto

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

### Observaciones:

- No puede evaluarse en forma cerrada (si, en forma numérica)
- Estimador es función de  $\bar{x}$ ,  $A_0$ ,  $\sigma^2$
- Si  $A_0 \gg \sqrt{\sigma^2/N}$ , entonces no hay casi truncamiento, es decir  $\hat{A} \rightarrow \bar{x}$ .
- Conocimiento a priori hace que  $|\hat{A}| < |\bar{x}|$  (sesgo hacia cero).
- Si  $N \gg 1$  (muchos datos), entonces  $\sigma^2/N \rightarrow 0$ , con lo cual el estimador MMSE se basa cada vez menos en el conocimiento a priori y confía más en los datos. Es decir si  $N \rightarrow \infty$  entonces  $\hat{A} \rightarrow \bar{x}$ .

# Metodología Bayesiana: Resumen

- Se asume que parámetro a estimar es V.A.  $\theta \sim p(\theta)$  (realización desconocida).
- *Prior*  $p(\theta)$  representa conocimiento que tenemos *a priori* sobre el parámetro.
- Luego de que los datos  $\mathbf{x}$  son observados, el estado de conocimiento del parámetro  $\theta$  se resume por la posterior PDF  $p(\theta|\mathbf{x})$ .
- Un estimador óptimo en términos de MSE, es aquel que minimiza el *Bayesian MSE* (MSE promedio en todo  $\theta$ ). Este estimador es

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

- Estimador MMSE depende de los datos y del conocimiento a priori sobre  $\theta$ .
- Si el conocimiento a priori es débil respecto a los datos, entonces el estimador ignora el conocimiento a priori.
- Si el conocimiento a priori es fuerte respecto a los datos, entonces el estimador tiene un sesgo hacia el conocimiento a priori.
- La elección del prior  $p(\theta)$  es crítica en estimación Bayesiana. Una mala elección del prior puede llevar a un desempeño pobre.
- Si no tenemos información sobre  $\theta$ , en general, es equivalente utilizar



# Elección del prior PDF

- El prior sobre el parámetro juega un rol clave.
- Estimador MMSE siempre existe (recordar: el MVU puede no existir)
- La cuestión práctica es si se puede calcular de forma cerrada. Para ello se necesita calcular

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

- Requiere calcular integral en el espacio del parámetro (dimensión p).
- Luego, calcular el estimador MMSE requiere calcular la media.
- En general esto no es posible hacer de forma cerrada.
- A continuación veremos un ejemplo en el que SI se puede.

# Modelo Lineal Bayesiano (y Gaussiano)

Se consideran las observaciones según el siguiente modelo

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \quad (\text{Modelo Lineal Bayesiano Gaussiano})$$

- $\mathbf{x} \in \mathbb{R}^N$  observaciones
- $\mathbf{w} \in \mathbb{R}^N$  ruido
- $\boldsymbol{\theta} \in \mathbb{R}^p$  vector de  $p$  parámetros
- $\mathbf{H} \in \mathbb{R}^{N \times p}$  sistema

Se asume que

- $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_w)$
- $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{u}_\theta, \mathbf{C}_\theta)$
- $\boldsymbol{\theta}$  y  $\mathbf{w}$  independientes.

**Objetivo:** Encontrar  $\hat{\boldsymbol{\theta}}$ , estimador que minimiza el BMSE (*MMSE estimator*)

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

# Modelo Lineal Bayesiano (y Gaussiano)

Consideremos:  $\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{H} & \mathbf{I}_N \\ \mathbf{I}_p & \mathbf{0}_{p \times N} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{w} \end{bmatrix}.$

Observar que  $\boldsymbol{\theta}, \mathbf{w}$  son VAs independientes y Gaussianas, con lo cual  $\begin{bmatrix} \boldsymbol{\theta} & \mathbf{w} \end{bmatrix}^T$  es un vector conjuntamente Gaussiano.

- Transformación lineal de vector Gaussiano es Gaussiano:  $\mathbf{z}$  es Gaussiano.
- Vamos a calcular la media y matriz de covarianza de  $\mathbf{z}$ .

$$\mathbb{E}[\mathbf{z}] = \mathbb{E} \begin{bmatrix} \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{H}\mathbb{E}[\boldsymbol{\theta}] + \mathbb{E}[\mathbf{w}] \\ \mathbb{E}[\boldsymbol{\theta}] \end{bmatrix} = \begin{bmatrix} \mathbf{H}\mathbf{u}_{\boldsymbol{\theta}} \\ \mathbf{u}_{\boldsymbol{\theta}} \end{bmatrix}$$

**Observación.** Si  $\mathbf{u}$  es un V.A., y  $\mathbf{A}$  una matriz, entonces  $\mathbf{b} = \mathbf{A}\mathbf{u}$  cumple que

$$\mathbf{C}_{\mathbf{b}\mathbf{b}} = \mathbb{E} \left[ (\mathbf{b} - \mathbb{E}[\mathbf{b}])(\mathbf{b} - \mathbb{E}[\mathbf{b}])^T \right] = \mathbb{E} \left[ \mathbf{A}(\mathbf{u} - \mathbb{E}[\mathbf{u}])(\mathbf{u} - \mathbb{E}[\mathbf{u}])^T \mathbf{A}^T \right] = \mathbf{A}\mathbf{C}_{\mathbf{u}\mathbf{u}}\mathbf{A}^T.$$

# Modelo Lineal Bayesiano (y Gaussiano)

Consideremos: 
$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{H} & \mathbf{I}_N \\ \mathbf{I}_p & \mathbf{0}_{p \times N} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{w} \end{bmatrix}.$$

$$\mathbb{E}[\mathbf{z}] = \mathbb{E} \begin{bmatrix} \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{H}\mathbb{E}[\boldsymbol{\theta}] + \mathbb{E}[\mathbf{w}] \\ \mathbb{E}[\boldsymbol{\theta}] \end{bmatrix} = \begin{bmatrix} \mathbf{H}\mathbf{u}_\theta \\ \mathbf{u}_\theta \end{bmatrix}$$

**Observación.** Si  $\mathbf{u}$  es un V.A., y  $\mathbf{A}$  una matriz, entonces  $\mathbf{b} = \mathbf{A}\mathbf{u}$  cumple que

$$\mathbf{C}_{\mathbf{b}\mathbf{b}} = \mathbb{E} \left[ (\mathbf{b} - \mathbb{E}[\mathbf{b}])(\mathbf{b} - \mathbb{E}[\mathbf{b}])^T \right] = \mathbb{E} \left[ \mathbf{A}(\mathbf{u} - \mathbb{E}[\mathbf{u}])(\mathbf{u} - \mathbb{E}[\mathbf{u}])^T \mathbf{A}^T \right] = \mathbf{A}\mathbf{C}_{\mathbf{u}\mathbf{u}}\mathbf{A}^T.$$

$$\begin{aligned} \mathbf{C}_{\mathbf{z}} = \mathbf{C}_{[\mathbf{x}, \boldsymbol{\theta}]} &= \begin{bmatrix} \mathbf{C}_{\mathbf{x}} & \mathbf{C}_{\mathbf{x}\boldsymbol{\theta}} \\ \mathbf{C}_{\boldsymbol{\theta}\mathbf{x}} & \mathbf{C}_{\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} \mathbf{H} & \mathbf{I}_N \\ \mathbf{I}_p & \mathbf{0}_{p \times N} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{\boldsymbol{\theta}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathbf{w}} \end{bmatrix} \begin{bmatrix} \mathbf{H}^T & \mathbf{I}_N \\ \mathbf{I}_p & \mathbf{0}_{p \times N} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T + \mathbf{H}\mathbf{C}_{\mathbf{w}} & \mathbf{H}\mathbf{C}_{\boldsymbol{\theta}} \\ \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T & \mathbf{C}_{\boldsymbol{\theta}} \end{bmatrix}. \end{aligned}$$

Luego

$$\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N} \left( [\mathbf{H}\mathbf{u}_\theta, \mathbf{u}_\theta]^T, \mathbf{C}_{\mathbf{z}} \right).$$

# Modelo Lineal Bayesiano (y Gaussiano)

Dado  $\mathbf{z} = [\boldsymbol{\theta}, \mathbf{x}]$ , se tiene que (Bayes)

$$p(\mathbf{z}) = p(\boldsymbol{\theta}, \mathbf{x}) = p(\boldsymbol{\theta}|\mathbf{x})p(\mathbf{x}),$$

donde  $p(\boldsymbol{\theta}, \mathbf{x})$  es Gaussiana multivariada.

**Ejercicio.** Si  $p(\boldsymbol{\theta}, \mathbf{x})$  es Gaussiana, entonces  $p(\boldsymbol{\theta}|\mathbf{x})$  es Gaussiana. Por lo tanto,

$$p(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{N}(\mathbf{u}_{\boldsymbol{\theta}|\mathbf{x}}, \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}}),$$

donde debemos especificar su primer y segundo momento  $\mathbf{u}_{\boldsymbol{\theta}|\mathbf{x}}, \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}}$ .

**Ejercicio.** Mostrar que

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

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

Luego, gracias a la identidad de Woodbury

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1},$$

se tiene que

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

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

# Modelo Lineal Bayesiano (y Gaussiano)

## Ejemplo: Nivel de DC en WGN - Prior Gaussiano

Se consideran las observaciones del nivel de continua en WGN,

$$x[n] = A + w[n], \quad \text{con } n = 0, 1, \dots, N-1 \text{ y } w[n] \sim \mathcal{N}(0, \sigma^2) \quad \forall n,$$

donde se quiere estimar  $A$ . Se asume el prior  $p(A) \sim \mathcal{N}(\mu_A, \sigma_A^2)$ .

### Objetivo:

- Calcular  $\hat{A} = \mathbb{E}[A|\mathbf{x}]$  y  $\text{var}(A|\mathbf{x})$
- Mostrar que  $\text{bmse}(\hat{A}) < \text{mse}(\bar{x})$ .

Podemos aplicar el resultado anterior (Modelo Lineal Bayesiano (y Gaussiano)) con

$$\mathbf{x} = \mathbf{1}A + \mathbf{w}, \quad \text{con } \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \text{ y con } \theta = A, \mathbf{H} = \mathbf{1}^T = [1, 1, \dots, 1].$$

Se tiene que

$$\begin{aligned} \text{var}(A|\mathbf{x}) &= \left( \frac{1}{\sigma_A^2} + \frac{\mathbf{1}^T \mathbf{1}}{\sigma^2} \right)^{-1} = \left( \frac{1}{\sigma_A^2} + \frac{N}{\sigma^2} \right)^{-1} = \frac{\sigma_A^2 \frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}} \\ \mathbb{E}[A|\mathbf{x}] &= \mu_A + \left( \frac{\sigma_A^2 \frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}} \right) \frac{1}{\sigma^2} \mathbf{1}^T (\mathbf{x} - \mathbf{1}\mu_A) = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} (\bar{\mathbf{x}} - \mu_A) \\ &= \alpha \bar{x} + (1 - \alpha) \mu_A, \quad \text{con } \alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}. \end{aligned}$$

# Modelo Lineal Bayesiano (y Gaussiano)

## Ejemplo: Nivel de DC en WGN - *Prior* Gaussiano

$$\hat{A} = \mathbb{E}[A|\mathbf{x}], \quad \text{var}(A | \mathbf{x}) = \frac{\sigma_A^2 \frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}}$$

Por un lado tenemos

$$\begin{aligned} \text{bmse}(\hat{A}) &= \mathbb{E}_{\mathbf{x}, A}[(\hat{A} - A)^2] \\ &= \mathbb{E}_{\mathbf{x}, A}[(\mathbb{E}(A | \mathbf{x}) - A)^2] \\ &= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{A|\mathbf{x}}[(\mathbb{E}(A | \mathbf{x}) - A)^2] \\ &= \mathbb{E}_{\mathbf{x}}[\text{var}(A | \mathbf{x})] \\ &= \mathbb{E}_{\mathbf{x}}\left[\frac{\sigma_A^2 \frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}}\right] \\ &= \left(\frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\right) \frac{\sigma^2}{N} \\ &< \frac{\sigma^2}{N} = \text{mse}(\bar{\mathbf{x}}). \end{aligned}$$

- “En promedio”  $\hat{A}$  es mejor estimador que  $\bar{\mathbf{x}}$ .

# Modelo Lineal Bayesiano (y Gaussiano)

## Ejemplo: Nivel de DC en WGN - *Prior Gaussiano*

$$\hat{A} = \mathbb{E}[A|\mathbf{x}] = \alpha \bar{x} + (1 - \alpha)\mu_A, \quad \text{con } \alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}.$$

Por un lado tenemos

$$\begin{aligned} \text{bmse}(\hat{A}) &= \mathbb{E}_{\mathbf{x}, A}[(\hat{A}(\mathbf{x}) - A)^2] \\ &= \mathbb{E}_{\mathbf{x}, A}[(\alpha \bar{x} + (1 - \alpha)\mu_A - A)^2] \\ &= \mathbb{E}_{\mathbf{x}, A}[(\alpha(\bar{x} - A) + (1 - \alpha)(\mu_A - A))^2] \\ &= \mathbb{E}_A \mathbb{E}_{\mathbf{x}|A}[(\alpha(\bar{x} - A) + (1 - \alpha)(\mu_A - A))^2] \\ &= \mathbb{E}_A \mathbb{E}_{\mathbf{x}|A}[\alpha^2(\bar{x} - A)^2 + 2\alpha(\bar{x} - A)(1 - \alpha)(\mu_A - A) + (1 - \alpha)^2(\mu_A - A)^2] \\ &= \alpha^2 \frac{\sigma^2}{N} + (1 - \alpha)^2 \mathbb{E}_A[(A - \mu_A)^2] \\ &= \alpha^2 \frac{\sigma^2}{N} + (1 - \alpha)^2 \sigma_A^2 \\ &= \frac{\sigma^2}{N} \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \\ &< \frac{\sigma^2}{N} = \text{mse}(\bar{x}). \end{aligned}$$

“En promedio”  $\hat{A}$  es mejor estimador que  $\bar{x}$ .



**Estimador “equivalente” Bayesiano del MLE:** en lugar de maximizar la verosimilitud, se maximiza la densidad a posteriori

$$\begin{aligned}\hat{\theta}_{MAP} &:= \arg \max_{\theta} p(\theta | \mathbf{x}) = \arg \max_{\theta} \frac{p(\mathbf{x} | \theta) p(\theta)}{p(\mathbf{x})} \\ &= \arg \max_{\theta} p(\mathbf{x} | \theta) p(\theta) \\ &= \arg \max_{\theta} \underbrace{\{\log p(\mathbf{x} | \theta)\}}_{\text{likelihood}} + \underbrace{\{\log p(\theta)\}}_{\text{prior}}\end{aligned}$$

**Observación.** Si  $p(\theta) \simeq \text{cte}$  (prior no informativo),  $\hat{\theta}_{MAP} \simeq \hat{\theta}_{ML}$

# Estimador MAP: Ejemplo lineal Gaussiano

$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ , con  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{w}})$ ,  $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{C}_{\boldsymbol{\theta}})$ ,  $\boldsymbol{\theta}$  y  $\mathbf{w}$  independientes.

**Objetivo:** Encontrar  $\hat{\boldsymbol{\theta}}_{MAP}$ .

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg \max_{\boldsymbol{\theta}} \left\{ \underbrace{-\frac{1}{2} \left[ (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \mathbf{C}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) \right]}_{f(\boldsymbol{\theta}, \mathbf{x})} \right\}$$

$$0 = \frac{\partial f}{\partial \boldsymbol{\theta}} = \frac{1}{2} \left[ 2\mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) + 2\mathbf{C}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) \right]$$

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

Con lo cual

$$\hat{\boldsymbol{\theta}}_{MAP} = \left( \mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1} \right)^{-1} \left( \mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{x} + \mathbf{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}} \right)$$

**Ejercicio.** Mostrar que

$$\hat{\boldsymbol{\theta}}_{MAP} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + \left( \mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1} \mathbf{H} + \mathbf{C}_{\boldsymbol{\theta}}^{-1} \right)^{-1} \mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_{\boldsymbol{\theta}})$$

Es decir, en este caso el estimador MAP coincide con el estimador MMSE.

- **Kay, S. M.** (1993)  
*Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Capítulos 10 y 11.