

# Estimación y Predicción en Series Temporales

Estimador de máxima verosimilitud (MLE)

Departamento de Procesamiento de Señales

Instituto de Ingeniería Eléctrica  
Facultad de Ingeniería

2022

# Agenda

- 1 Repaso
- 2 Estimadores de máxima verosimilitud (MLE)

# Estimación de parámetros

## Planteo del Problema:

- Dadas  $N$  muestras de una señal discreta  $x[n]$  que depende de cierto parámetro  $\theta$  desconocido.
- Estimar  $\theta$  a partir de las  $N$  muestras  $x[0], x[1], \dots, x[N-1]$

Para ello se define un estimador de  $\theta$  que es función de los datos:

$$\hat{\theta} = g(x[0], x[1], \dots, x[N-1])$$

- $g$ : función a determinar
- $\hat{\theta}$ : estimador de  $\theta$

**Objetivo:** Encontrar función  $g$  de forma que  $\hat{\theta}$  sea buen estimador de  $\theta$ .

- Estimador  $\hat{\theta}$  debe ser cercano (en algún sentido a definir) al valor verdadero de  $\theta$ .
- El criterio de cercanía debe ser especificado teniendo en cuenta que  $\hat{\theta}$  es una Variable Aleatoria (función de V.As).

# Modelado de los datos

- Se dispone de un **conjunto de  $N$  datos**  $x[i] \in \mathbb{R}^n$ :

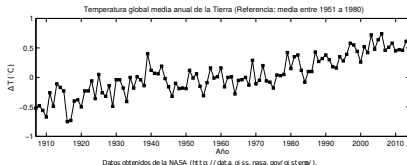
$$\mathcal{D} = \{x[0], x[1], \dots, x[N-1]\}$$

y un modelo que depende de un parámetro  $\theta$  desconocido.

- Debido a la complejidad del fenómeno a caracterizar, **modelamos los datos estadísticamente**, mediante la **función de densidad de probabilidad** o *pdf*,

$$p(x[0], x[1], \dots, x[N-1]; \theta)$$

- La PDF está parametrizada por el parámetro desconocido  $\theta$ , es decir define una familia de funciones.
- Puede interpretarse como que los datos son “aleatorios”



# Criterio de Mínima Varianza

- En la búsqueda de **estimadores óptimos** es necesario utilizar algún **criterio de optimalidad**.
- Uno natural es la minimización del **Error Cuadrático Medio** (*MSE, Mean Square Error*)

$$\text{MSE}(\hat{\theta}) = \mathbb{E} \left[ (\hat{\theta} - \theta)^2 \right].$$

- Análisis (descomposición) del error cuadrático medio:

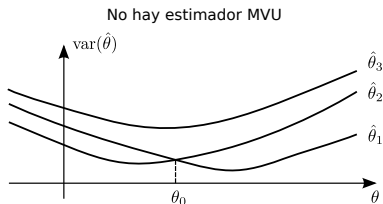
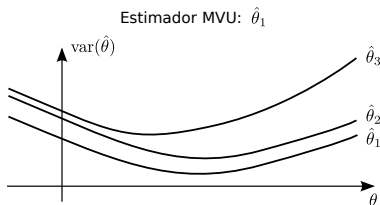
$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \mathbb{E} \left[ (\hat{\theta} - \theta)^2 \right] \\ &= \mathbb{E} \left\{ \left[ (\hat{\theta} - \mathbb{E}(\hat{\theta})) + (\mathbb{E}(\hat{\theta}) - \theta) \right]^2 \right\} \\ &= \mathbb{E} \left[ (\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 \right] + \underbrace{2 \mathbb{E} \left[ (\hat{\theta} - \mathbb{E}(\hat{\theta})) (\mathbb{E}(\hat{\theta}) - \theta) \right]}_{(\mathbb{E}(\hat{\theta}) - \theta) \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta})) = 0} + \mathbb{E} \left[ (\mathbb{E}(\hat{\theta}) - \theta)^2 \right] \\ &= \mathbb{E} \left[ (\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 \right] + (\mathbb{E}(\hat{\theta}) - \theta)^2 \\ &= \text{var}(\hat{\theta}) + b^2(\theta) \end{aligned}$$

- Descomposición sumamente útil *bias-variance*.

# Estimador insesgados de varianza mínima (MVU)

## Existencia de estimadores MVU.

- Se dice que **existe un estimador MVU** si hay un estimador de **menor varianza** que el resto de los posibles estimadores **para todo  $\theta$** .



- Dos ejemplos: izquierda (existe MVU); derecha (no existe MVU).
- El estimador MVU no tiene porqué existir.

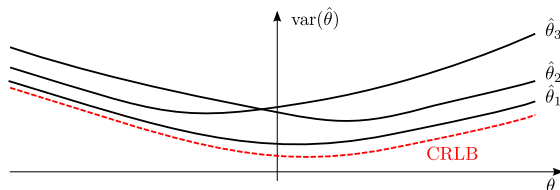
# Búsqueda de estimadores MVU

- Aún si existe un estimador MVU, puede no ser posible encontrarlo. No hay ninguna *receta infalible* para encontrar estimadores MVU.

## Enfoques de búsqueda de estimadores MVU:

### 1 Utilizando la cota de inferior de Cramér-Rao (CRLB, Cramér-Rao Lower Bound)

- Determinar la CRLB y ver si algún estimador la alcanza.
- CRLB determina un límite inferior en la varianza de cualquier estimador insesgado (Clase 3 / Capítulo 3 Kay)
- Si un estimador tiene varianza igual a la CRLB para todos los valores de  $\theta$ , es el estimador MVU.



# Cota Inferior de Cramér-Rao

## **Teorema: Cota Inferior de Cramér-Rao, parámetro escalar.**

Se asume que la PDF  $p(\mathbf{x}; \theta)$  satisface la condición de regularidad,

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

Entonces,

- 1 la varianza de todo estimador insesgado  $\hat{\theta}$  cumple que

$$\text{var}(\hat{\theta}) \geq \frac{1}{-\mathbb{E}_{\mathbf{x}} \left[ \frac{\partial^2 \log p(\mathbf{x}; \theta)}{\partial \theta^2} \right]},$$

donde la derivada se evalúa en el valor verdadero de  $\theta$ .

- 2 existe un estimador que alcanza la cota para todo  $\theta$  si y solo si

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

para alguna función  $I$  y  $g$ .

Este estimador, que es el MVU, es  $\hat{\theta} = g(\mathbf{x})$  y su varianza es  $\frac{1}{I(\theta)}$ .



# Estimador MVU en modelos lineales

## Teorema: Estimador MVU en modelos lineales.

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

- $\mathbf{x} \in \mathbb{R}^N$ : datos observados.
- $\boldsymbol{\theta} \in \mathbb{R}^p$ :  $p$  parámetros desconocidos.
- $\mathbf{H} \in \mathbb{R}^{N \times p}$ : matriz de observación con  $N > p$  y rango  $p$
- $\mathbf{w} \in \mathbb{R}^N$ : ruido en observación es **coloreado**, muestras no son independientes. Se asume  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ , con  $\mathbf{C}$  simétrica definida positiva.

Nota: El espacio  $\mathbb{R}$  puede ser remplazado por  $\mathbb{C}$ . Entonces,

Estimador MVU

$$\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x}) = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x}$$

Matriz de Covarianza

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta}) = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}$$

- Además, el estimador  $\hat{\boldsymbol{\theta}}$  es **eficiente** ya que alcanza la CRLB para todo  $\boldsymbol{\theta}$ .

## Enfoques de búsqueda de estimadores MVU:

### 2 **Buscar estadísticos suficientes y aplicar el teorema de Rao-Blackwell-Lehmann-Scheffé (RBLs)**

- Puede existir un estimador MVU que no alcance la CRLB.
- Clase 4 / Capítulo 5 (Kay)

### 3 **Restringir la clase de estimadores (e.g., lineales)**

- Restringir la clase de estimadores no sólo a los insesgados, sino también a los insesgados que sean lineales con los datos, y encontrar el MVU en esta clase.
- Este estimador no será óptimo, a menos que el estimador MVU sea lineal en ese problema en particular.
- Clase 5 / Capítulo 6 (Kay)

# Mejor Estimador Lineal Insesgado

*Best Linear Unbiased Estimator (BLUE)*

Se observa el conjunto de datos  $\{x[0], x[1], \dots, x[N-1]\}$  cuya PDF  $p(\mathbf{x}; \theta)$  depende del parámetro desconocido  $\theta$  que se quiere estimar, y de la cual se asumen conocidos el primer y segundo momento.

- Se dice que un estimador es **lineal** si se restringe a ser lineal con los datos,

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n].$$

- **Estimador BLUE:** estimador lineal, insesgado y tiene varianza mínima entre todos los estimadores lineales.
- Hay que determinar los coeficientes  $a_n$  para que el estimador cumpla estas condiciones.

# Búsqueda del BLUE

Para determinar el BLUE se impone que el estimador  $\hat{\theta}$  sea lineal e insesgado y se determinan los coeficientes  $a_n$  que minimizan la varianza.

- Estimador lineal

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] = \mathbf{a}^T \mathbf{x}$$

- Condición de estimador insesgado

$$\mathbb{E}(\hat{\theta}) = \sum_{n=0}^{N-1} a_n \mathbb{E}(x[n]) = \theta, \quad \forall \theta.$$

- Para satisfacer la condición de insesgado,  $\mathbb{E}(x[n])$  tiene que ser lineal con el parámetro desconocido  $\theta$ ,

$$\mathbb{E}(x[n]) = s[n]\theta,$$

con  $s[n]$  conocido.

**Nota.** Si esto no se cumple, es imposible satisfacer la condición de insesgado.

**Ejemplo.** si  $\mathbb{E}(x[n]) = \cos \theta$ , la condición de insesgado sería  $\sum_{n=0}^{N-1} a_n \cos \theta = \theta$ . No existen coeficientes  $a_n$  que cumplan esto para todo  $\theta$ .

# Búsqueda del BLUE

- Condición necesaria para ser insesgado, implica que el BLUE solo es aplicable en estimación de amplitud de señales conocidas en ruido,

$$x[n] = \theta s[n] + w[n].$$

- Se puede generalizar mediante transformaciones no lineales de los datos (e.g.  $y[n] = x[n]^2$  para estimar  $\sigma^2$  en WGN).
- Continuando con la condición de no sesgado,

$$\sum_{n=0}^{N-1} a_n \mathbb{E}(x[n]) = \theta$$

$$\sum_{n=0}^{N-1} a_n s[n] \theta = \theta$$

$$\sum_{n=0}^{N-1} a_n s[n] = 1$$

es decir,

$$\mathbf{a}^T \mathbf{s} = 1,$$

con  $\mathbf{s} = [s[0], s[1], \dots, s[N-1]]^T$ .

# Búsqueda del BLUE

- Por otro lado, la varianza del estimador es:

$$\begin{aligned}\text{var}(\hat{\theta}) &= \mathbb{E} \left[ \left( \hat{\theta} - \mathbb{E}(\hat{\theta}) \right)^2 \right] \\ &= \mathbb{E} \left[ \left( \mathbf{a}^T \mathbf{x} - \mathbb{E}(\mathbf{a}^T \mathbf{x}) \right)^2 \right] \\ &= \mathbb{E} \left[ \left( \mathbf{a}^T (\mathbf{x} - \mathbb{E}(\mathbf{x})) \right)^2 \right] \\ &= \mathbb{E} \left[ \mathbf{a}^T (\mathbf{x} - \mathbb{E}(\mathbf{x})) (\mathbf{x} - \mathbb{E}(\mathbf{x}))^T \mathbf{a} \right] \\ &= \mathbf{a}^T \mathbb{E} \left[ (\mathbf{x} - \mathbb{E}(\mathbf{x})) (\mathbf{x} - \mathbb{E}(\mathbf{x}))^T \right] \mathbf{a} \\ &= \mathbf{a}^T \mathbf{C} \mathbf{a}\end{aligned}$$

- Hay que encontrar  $\mathbf{a}$  de forma de minimizar la varianza manteniendo la restricción impuesta de estimador insesgado,

$$\mathbf{a}_* = \arg \min_{\mathbf{a}} \mathbf{a}^T \mathbf{C} \mathbf{a} \quad \text{sujeto a} \quad \mathbf{a}^T \mathbf{s} = 1$$

Se puede resolver utilizando multiplicadores de Lagrange.

# Búsqueda del BLUE

- El estimador BLUE es entonces

$$\hat{\theta} = \mathbf{a}_*^T \mathbf{x} = \frac{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{x}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}}$$

- Su varianza es,

$$\text{var}(\hat{\theta}) = \mathbf{a}_*^T \mathbf{C} \mathbf{a}_* = \frac{1}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}}$$

- Notar que el estimador es insesgado (usando que  $\mathbb{E}(x[n]) = s[n]\theta$ ),  $\forall \theta$ ,

$$\begin{aligned}\mathbb{E}(\hat{\theta}) &= \frac{\mathbf{s}^T \mathbf{C}^{-1} \mathbb{E}(\mathbf{x})}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}} \\ &= \frac{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s} \theta}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}} \\ &= \theta.\end{aligned}$$

- Observación: Para determinar el BLUE solo se requiere conocer:
  - $\mathbf{s}$  (media escalada) y  $\mathbf{C}$  (matrix de covarianza de  $\mathbf{x}$ )

Es decir, los dos primeros momentos de  $p(\mathbf{x})$  en lugar de la PDF completa.

# Estimador BLUE: Extensión a vector de parámetros

- El estimador BLUE en el caso vectorial queda,

$$\hat{\boldsymbol{\theta}} = \mathbf{a}_* \mathbf{x} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x}$$

y su matriz de covarianza es

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}.$$

## Observación

- El estimador BLUE es el mismo que el estimador MVU para el modelo lineal general

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}).$$

¿Por qué?



# Estimador BLUE: Extensión a vector de parámetros

**Teorema de Gauss-Markov.** Si los datos observados  $\mathbf{x}$  tienen la forma del modelo lineal general

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

donde

- $\mathbf{x}$ :  $N \times 1$  vector de observaciones
- $\mathbf{H}$ :  $N \times p$  matriz de observación conocida, con  $N \geq p$  y rango  $p$ .
- $\boldsymbol{\theta}$ :  $p \times 1$  vector de parámetros a estimar
- $\mathbf{w}$ :  $N \times 1$  vector de ruido con PDF arbitraria, media nula y covarianza  $\mathbf{C}$ .

Entonces, el estimador BLUE de  $\boldsymbol{\theta}$  es

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x},$$

y la varianza es

$$\text{var}(\hat{\theta}_i) = \left[ (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \right]_{ii}.$$

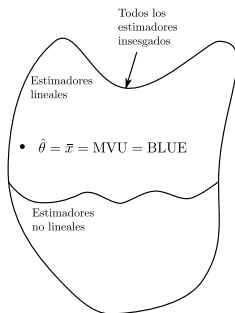
Además, la matriz de covarianza del estimador es

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}.$$

# Consideraciones sobre la optimalidad del BLUE

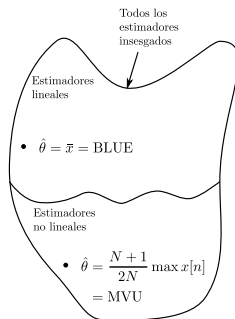
Nivel de DC en WGN

El BLUE es óptimo



Media en ruido uniforme

El BLUE es subóptimo



- Si el MVU pertenece a la clase de estimadores lineales, no se pierde desempeño con el BLUE.
- Si el MVU pertenece a la clase de estimadores no lineales, hay pérdida de desempeño (puede ser significativa).

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**.

# MLE: Descripción intuitiva

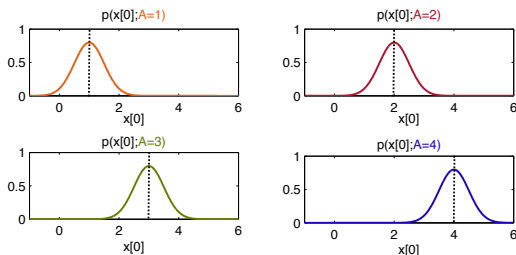
**Ejemplo:** Estimar nivel de DC en WGN (una única muestra)

$$x[0] = A + w[0], \quad \text{donde } w[0] \sim \mathcal{N}(0, \sigma^2)$$

- En este caso, la PDF de los datos es  $x[0] \sim \mathcal{N}(A, \sigma^2)$ ,

$$p(x[0]; A) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} (x[0] - A)^2 \right].$$

- Familia de PDFs indexadas por parámetro desconocido ( $\theta = A$ ):



# MLE: Descripción intuitiva

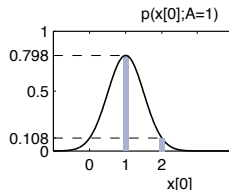
- Dado un valor del parámetro, la PDF correspondiente indica cómo se deberían distribuir los datos observados (probabilidad) si ese fuese el parámetro.
- En particular, dado  $A = A_0$ , la probabilidad de observar un valor de  $x[0]$  en un intervalo de tamaño  $\Delta$  centrado en  $x_0$  es:

$$\Pr \left( x[0] \in \left[ x_0 - \frac{\Delta}{2}, x_0 + \frac{\Delta}{2} \right] \right) = \int_{x_0 - \frac{\Delta}{2}}^{x_0 + \frac{\Delta}{2}} p(x; A = A_0) dx \\ \approx p(x[0] = x_0; A = A_0) \Delta.$$

- Por ejemplo, si  $A=1$  (con  $\sigma^2 = 1/4$ ),

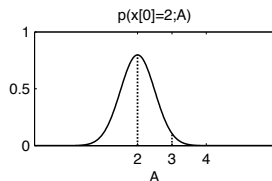
$$\Pr \left( x[0] \in \left[ 1 - \frac{\Delta}{2}, 1 + \frac{\Delta}{2} \right] \right) \approx 0.798\Delta$$

$$\Pr \left( x[0] \in \left[ 2 - \frac{\Delta}{2}, 2 + \frac{\Delta}{2} \right] \right) \approx 0.108\Delta$$



# MLE: Descripción intuitiva

- En la práctica los datos son los observados, así que vamos a resolver el problema inverso:
- Dados los datos observados y el modelo definido por la familia de PDFs, hay que encontrar la PDF que hace más probable haber producido esos datos observados.
- **Función de verosimilitud:** es la función dada por la PDF al variar el parámetro fijando el valor de los datos.
- Cuantifica que tan “verosímil” es cierto valor del parámetro desconocido luego de observados los datos (no es una probabilidad en  $A$  en el sentido matemático)
- Por ejemplo, la probabilidad de observar  $x[0] = 2$  para cada valor de  $A$  es aproximadamente  $p(x[0] = 2; A)\Delta$ .
- Si la observación fue  $x[0] = 2$ , inferir  $A = 3$  no sería razonable, ya que la probabilidad de observar  $x[0] = 2$  es muy pequeña.
- Es más probable que  $A \approx 2$ , ya que conduce a probabilidad alta de observar  $x[0] = 2$ .





# MLE: Descripción intuitiva

- Por lo tanto, si se observó  $x[0] = x_0$ , se elige como estimador  $\hat{A}$  el valor que maximiza  $p(x[0] = x_0; A)$ , la función de verosimilitud fijando los datos en  $x = x_0$ , sobre todo el dominio válido de  $A$ .

**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)

# Ejemplo I

**Ejemplo:** Nivel de DC en WGN (varianza y media relacionadas) Los datos observados son

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

donde  $A$  es desconocido. El parámetro desconocido se refleja en la **media** y en la **varianza**. Se quiere encontrar el estimador MVU.

## Determinación de la CRLB

- Par encontrar el MVU, una primera posibilidad es determinar la CRLB y ver si existe algún estimador cuya varianza la alcance.
- La PDF de los datos es,

$$\begin{aligned} p(\mathbf{x}; A) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi A}} \exp \left[ -\frac{1}{2A} (x[n] - A)^2 \right] \\ &= \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp \left[ -\frac{1}{2A} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \end{aligned}$$

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Determinación de la CRLB

- Tomando el logaritmo queda,

$$\log p(\mathbf{x}; A) = -\frac{N}{2} \log 2\pi - \frac{N}{2} \log A - \frac{1}{2A} \sum_{n=0}^{N-1} (x[n] - A)^2.$$

- Calculando la derivada primera de la función  $\log p(\mathbf{x}; A)$ ,

$$\begin{aligned} \frac{\partial \log p(\mathbf{x}; A)}{\partial A} &= -\frac{N}{2A} + \frac{1}{A} \sum_{n=0}^{N-1} (x[n] - A) + \frac{1}{2A^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \\ &\stackrel{?}{=} I(A) (g(\mathbf{x}) - A). \end{aligned}$$

- La derivada de la función de verosimilitud parece no poder factorizarse de la forma requerida (CRLB, estimador eficiente)
- No es obvio, pero se puede probar que no se puede factorizar de esa manera. Por lo tanto, no existe un estimador eficiente.

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Determinación de la CRLB

- Igualmente, podemos determinar la CRLB. Derivando nuevamente

$$\frac{\partial^2 \log p(\mathbf{x}; A)}{\partial A^2} = \frac{N}{2A^2} - \frac{N}{A} - \frac{2}{A^2} \sum_{n=0}^{N-1} (x[n] - A) - \frac{1}{A^3} \sum_{n=0}^{N-1} (x[n] - A)^2$$

- Tomando la esperanza en  $\mathbf{x}$  se llega a,

$$\begin{aligned} \mathbb{E} \left[ \frac{\partial^2 \log p(\mathbf{x}; A)}{\partial A^2} \right] &= \frac{N}{2A^2} - \frac{N}{A} - \frac{1}{A^3} NA \\ &= -\frac{N(A + \frac{1}{2})}{A^2} \end{aligned}$$

- Por lo tanto, la CRLB para este problema es

$$\text{var}(\hat{A}) \geq \frac{A^2}{N(A + \frac{1}{2})}.$$

# Ejemplo I

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Búsqueda del MVU mediante estadísticos suficientes de DC

- Se busca un estadístico suficiente de  $A$  en base a la factorización de Neyman-Fisher,

$$p(\mathbf{x}; A) = g(T(\mathbf{x}), A)h(\mathbf{x}).$$

- Observando que

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

la PDF de los datos se puede expresar como,

$$p(\mathbf{x}; A) = \underbrace{\frac{1}{(2\pi A)^{\frac{N}{2}}} \exp \left[ -\frac{1}{2} \left( \underbrace{\frac{1}{A} \sum_{n=0}^{N-1} x^2[n]}_{T(\mathbf{x})} + NA \right) \right]}_{g(T(\mathbf{x}), A)} \underbrace{\exp(N\bar{x})}_{h(\mathbf{x})}$$

- Se concluye que un estadístico suficiente de  $A$  es  $T(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n]$ .

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Búsqueda del MVU mediante estadísticos suficientes de DC

- Debemos verificar si el estadístico suficiente es completo. Para eso hay que buscar una única función  $g$  que lo haga insesgado,

$$\mathbb{E} \left[ g \left( \sum_{n=0}^{N-1} x^2[n] \right) \right] = A, \quad \forall A.$$

Dado que

$$\begin{aligned} \mathbb{E} \left[ \sum_{n=0}^{N-1} x^2[n] \right] &= N \mathbb{E} [x^2[n]] \\ &= N (\text{var} [x[n]] + \mathbb{E} [x[n]]^2) \\ &= N(A + A^2). \end{aligned}$$

no existe una forma obvia de elegir  $g$ .

- Esto agota las posibilidades de obtener un estimador MVU.

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Cálculo del MLE

- El MLE se define como

$$\hat{A}_{\text{MLE}} = \arg \max_{A>0} \log p(\mathbf{x}; A)$$

- Recordemos que la PDF de los datos es,

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

- PDF como función de A: se convierte en la *función de verosimilitud*.
- Para maximizar la función de verosimilitud logarítmica se diferencia e iguala a 0. Recordemos que la derivada es,

$$\frac{\partial \log p(\mathbf{x}; A)}{\partial A} = -\frac{N}{2A} + \frac{1}{A} \sum_{n=0}^{N-1} (x[n] - A) + \frac{1}{2A^2} \sum_{n=0}^{N-1} (x[n] - A)^2.$$

# Ejemplo I

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Cálculo del MLE

- Igualando a 0 y despejando se llega a que:

$$\hat{A}^2 + \hat{A} - \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] = 0.$$

- Resolviendo el polinomio de segundo grado en  $\hat{A}$  se obtienen las dos soluciones

$$\hat{A} = -\frac{1}{2} \pm \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}.$$

- Se elige la solución que produce estimadores positivos, en acuerdo a la restricción sobre  $A$ ,  $A > 0$ ,

$$\hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}.$$

- Falta verificar que la solución corresponde al máximo (no mínimo) [ejercicio]

$$\left. \frac{\partial^2 \log p(\mathbf{x}; A)}{\partial A^2} \right|_{A=\hat{A}} < 0.$$



## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Análisis del MLE: Sesgo

$$\begin{aligned}\mathbb{E}(\hat{A}) &= \mathbb{E} \left( -\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}} \right) \\ &\neq -\frac{1}{2} + \sqrt{\mathbb{E} \left( \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] \right) + \frac{1}{4}} \\ &= -\frac{1}{2} + \sqrt{A + A^2 + \frac{1}{4}} \\ &= -\frac{1}{2} + \sqrt{\left( A + \frac{1}{2} \right)^2} \\ &= A\end{aligned}$$

- El estimador tiene sesgo.

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Análisis del MLE: Comportamiento asintótico

- Si definimos

$$u = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n],$$

entonces, el MLE encontrado es una transformación  $g(u)$  no lineal de  $u$ ,

$$\hat{A} = g(u) = -\frac{1}{2} + \sqrt{u + \frac{1}{4}}.$$

- Cuando  $N \rightarrow \infty$ , por L.G.N,

$$\lim_{N \rightarrow \infty} u = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] = \mathbb{E}(x^2[n]) = A + A^2 = u_0.$$

- Si  $N$  es grande, los valores probables de  $u$  se encuentran en un intervalo pequeño en torno a su media  $u_0$ .

# Ejemplo I

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Análisis del MLE: Comportamiento asintótico

- En un intervalo pequeño en torno a  $u_0$  la función  $g(u)$  es aproximadamente lineal: **linealidad estadística de transformaciones**.
- Aproximando linealmente  $g(u)$  en torno a  $u_0$  tenemos que que:

$$g(u) \approx g(u_0) + g'(u_0)(u - u_0) \quad (1)$$

donde

$$g'(u) = \frac{\frac{1}{2}}{\sqrt{u + \frac{1}{4}}}.$$

- Teniendo en cuenta que  $u_0 = A + A^2$ ,

$$g(A^2 + A) = -\frac{1}{2} + \sqrt{A^2 + A + \frac{1}{4}} = A, \quad g'(A^2 + A) = \frac{\frac{1}{2}}{A + \frac{1}{2}}.$$

- Por lo que sustituyendo en (1), para  $N$  grande (asintótico),

$$\hat{A} \approx A + \frac{\frac{1}{2}}{A + \frac{1}{2}} \left[ \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] - (A + A^2) \right]. \quad (2)$$

# Ejemplo I

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Análisis del MLE: Comportamiento asintótico

- De (2) es

$$\mathbb{E}(\hat{A}) = A + \frac{\frac{1}{2}}{A + \frac{1}{2}} \left[ \mathbb{E} \left( \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] \right) - (A + A^2) \right] = A,$$

concluyendo que  $\hat{A}$  es asintóticamente insesgado.

- La varianza asintótica es,

$$\begin{aligned} \text{var}(\hat{A}) &= \left( \frac{\frac{1}{2}}{A + \frac{1}{2}} \right)^2 \text{var} \left( \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] \right) \\ &= \frac{\frac{1}{4}}{N(A + \frac{1}{2})^2} \text{var}(x^2[n]). \end{aligned} \quad (3)$$

- Si  $\zeta \sim \mathcal{N}(\mu, \sigma^2)$ , entonces

$$\text{var}(\zeta^2) = \mathbb{E}(\zeta^4) - \mathbb{E}^2(\zeta^2) = 4\mu^2\sigma^2 + 2\sigma^4.$$

- Como  $x[n] \sim \mathcal{N}(A, A)$ ,

$$\text{var}(x^2[n]) = 4A^3 + 2A^2 = 4A^2 \left( A + \frac{1}{2} \right). \quad (4)$$

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Análisis del MLE: Comportamiento asintótico

- Sustituyendo (4) en (3),

$$\begin{aligned}\text{var}(\hat{A}) &= \frac{\frac{1}{4}}{N \left(A + \frac{1}{2}\right)^2} 4A^2 \left(A + \frac{1}{2}\right) \\ &= \frac{A^2}{N \left(A + \frac{1}{2}\right)} \\ &= CRLB(A)\end{aligned}$$

concluyendo que  $\hat{A}$  alcanza la CRLB *asintóticamente*.

- Como el estimador es asintóticamente insesgado y alcanza asintóticamente la CRLB, se dice que **es asintóticamente eficiente**.
- Además, el estimador es **consistente**. Esto significa que se cumple que

$$\lim_{N \rightarrow \infty} \Pr \left\{ |\hat{A} - A| > \epsilon \right\} = 0, \quad \forall \epsilon > 0.$$

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

### Análisis del MLE: Comportamiento asintótico

- **PDF Gaussiana:** por el Teorema Central del Límite, la variable aleatoria  $u = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]$  es Gaussiana con  $N \rightarrow \infty$ , y como si  $N$  es grande  $\hat{A}$  es afín en  $u$  (ecuación (2)), también tiene PDF Gaussiana.

- Como el MLE es asintóticamente insesgado, alcanza asintóticamente la CRLB y tiene PDF Gaussiana. La PDF del MLE es

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

donde  $\stackrel{a}{\sim}$  significa *asintóticamente distribuido como*.

- Este resultado es general; implica la **optimalidad asintótica del MLE**.
- **N pequeño:** si bien el estimador es asintóticamente óptimo, no puede afirmarse nada sobre su desempeño si el conjunto de datos es pequeño. Es posible, de hecho, es probable que existan mejores estimadores.
- En ocasiones, el estimador MLE conduce al estimador eficiente para un conjunto de datos finito.

## Ejemplo II

### Ejemplo: Nivel de DC en WGN (varianza conocida)

Se observan  $N$  muestras dadas por

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

El parámetro a estimar es  $A$ . La varianza del ruido  $\sigma^2$  se asume conocida.

- La PDF de los datos es

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

- La función de verosimilitud logarítmica es

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

- Para encontrar el MLE, se debe derivar e igualar a cero (encontrar el máximo),

$$\frac{\partial \log p(\mathbf{x}; A)}{\partial A} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) = 0.$$

## Ejemplo: Nivel de DC en WGN (varianza conocida)

- El MLE queda

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

- Como la media muestral es el estimador eficiente de  $A$ , el MLE es eficiente en este caso.
- Este resultado es general y se formaliza en el siguiente teorema.



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

# Ejemplo I

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

Se quiere determinar el tamaño necesario de los datos para que se *cumplan* los resultados asintóticos.

- Previamente se encontró que

| Estimador MLE  | CRLB   | PDF asintótica  |
|--|--|---|
| $\hat{A}_{\text{MLE}} = -\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n]}$ | $\text{var} \hat{A} \geq \frac{A^2}{N(A + \frac{1}{2})}$ | $\hat{A}_{\text{MLE}} \stackrel{a}{\sim} \mathcal{N}\left(A, \frac{A^2}{N(A + \frac{1}{2})}\right)$ |

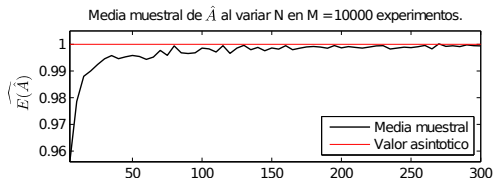
- Una estrategia podría ser encontrar la PDF exacta de  $\hat{A}$  y establecer para qué valor de  $N$  está cerca de la PDF asintótica.
- En principio es posible encontrar la PDF verdadera en este ejemplo, pero sería extremadamente tedioso.
- Si repetimos el experimento un número  $M$  de veces, es posible **estimar experimentalmente** la media y la varianza del estimador como

$$\widehat{\mathbb{E}}(\hat{A}) = \frac{1}{M} \sum_{i=1}^M \hat{A}_i \quad \widehat{\text{var}}(\hat{A}) = \frac{1}{M} \sum_{i=1}^M \left( \hat{A}_i - \widehat{\mathbb{E}}(\hat{A}) \right)^2$$

# Ejemplo I

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

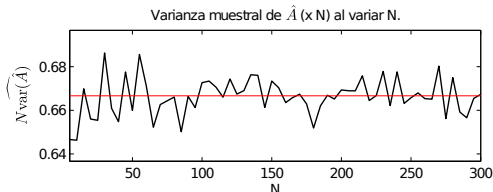
- Se generan  $N$  muestras de  $x[n]$  usando un valor de  $A = 1$  y se calcula  $\hat{A}$  (MLE).
- Se repite el experimento  $M = 10^4$  veces (sorteando siempre de manera independiente) y se calcula la media y la varianza muestral  $\widehat{\mathbb{E}}(\hat{A})$ ,  $\widehat{\text{var}}(\hat{A})$ .



**Valores asintóticos:**

$$\hat{A} = 1$$

$$N \text{var}(\hat{A}) = \frac{2}{3}$$



Con  $N \geq 20$ ,

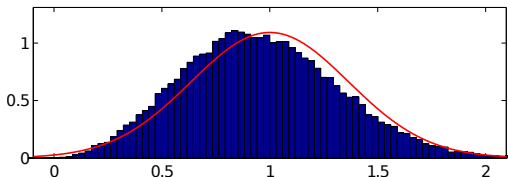
$$\left| \widehat{\mathbb{E}}(\hat{A}) - 1 \right| \leq 0.01.$$

# Ejemplo I

## Ejemplo: Nivel de DC en WGN (varianza y media relacionadas)

- Para comparar la PDF verdadera con la PDF asintótica, se repiten  $M = 20000$  experimentos con  $N = 5$  y  $N = 200$  y se grafican los histogramas de los estimadores.

Histograma con  $N = 5$  en  $M = 20000$  experimentos.



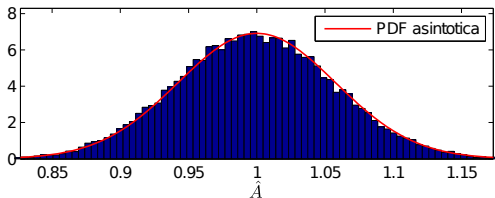
PDF asintótica:

$$\hat{A} \approx \mathcal{N}\left(1, \frac{2/3}{N}\right)$$

Con  $N = 5$ ,

- Se observa sesgo en el estimador
- PDF verdadera no tiene forma Gaussiana.

Histograma con  $N = 200$  en  $M = 20000$  experimentos.



Con  $N = 200$ , se cumplen bien las propiedades asintóticas.

# Ejemplo III

## Ejemplo: MLE de la fase de una senoide

Se quiere estimar la fase  $\phi$  de una senoide contaminada con WGN,

$$x[n] = A \cos(2\pi f_0 n + \phi) + w[n], \text{ con } n = 0, 1, \dots, N-1,$$

donde  $w[n] \sim \mathcal{N}(0, \sigma^2)$ , para todo  $n$ ; se asume  $A$ ,  $f_0$  y  $\sigma^2$  conocidos.

- En este caso no es posible encontrar un estimador MVU mediante la CRLB o estadísticos suficientes.
- La CRLB para el problema es (Kay 1993, ejemplo 3.4),

$$\text{var}(\hat{\phi}) \geq \frac{2\sigma^2}{NA^2}.$$

- Para encontrar el MLE hay que maximizar  $p(\mathbf{x}; \phi)$ , siendo

$$p(\mathbf{x}; \phi) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A \cos(2\pi f_0 n + \phi))^2 \right],$$

que es equivalente a minimizar

$$J(\phi) = \sum_{n=0}^{N-1} (x[n] - A \cos(2\pi f_0 n + \phi))^2.$$

## Ejemplo: MLE de la fase de una senoide

- Diferenciando respecto a  $\phi$  se tiene que

$$\frac{\partial J(\phi)}{\partial \phi} = 2A \sum_{n=0}^{N-1} (x[n] - A \cos(2\pi f_0 n + \phi)) \sin(2\pi f_0 n + \phi).$$

- Al igualar a 0, se llega a que el MLE  $\hat{\phi}$  cumple

$$\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n + \hat{\phi}) = A \sum_{n=0}^{N-1} \cos(2\pi f_0 n + \hat{\phi}) \sin(2\pi f_0 n + \hat{\phi}).$$

- El lado derecho de la igualdad es aproximadamente 0 si  $f_0$  no es cercana a 0 o a  $1/2$  (**ejercicio**). Por lo que el MLE cumple aproximadamente,

$$\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n + \hat{\phi}) = 0.$$

# Ejemplo III

## Ejemplo: MLE de la fase de una senoide

- Utilizando la relación trigonométrica  $\sin(a + b) = \sin a \cos b + \cos a \sin b$ :

$$\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n) \cos(\hat{\phi}) = - \sum_{n=0}^{N-1} x[n] \cos(2\pi f_0 n) \sin(\hat{\phi})$$

- Despejando  $\hat{\phi}$  se obtiene que el MLE es aproximadamente

$$\hat{\phi} = -\arctan \frac{\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n)}{\sum_{n=0}^{N-1} x[n] \cos(2\pi f_0 n)}.$$

- En ese ejemplo, la varianza asintótica del estimador MLE es,

$$\text{var}(\hat{\phi}) = \frac{1}{N \frac{A^2}{2\sigma^2}} = \frac{1}{N\eta} \quad \text{donde } \eta = \frac{\frac{A^2}{2}}{\sigma^2} \text{ es la SNR.}$$

- La PDF asintótica es

$$\hat{\phi} \stackrel{a}{\sim} \mathcal{N}(\phi, 1/(N\eta)).$$

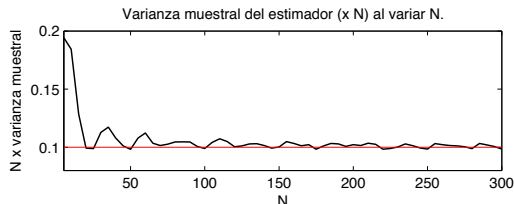
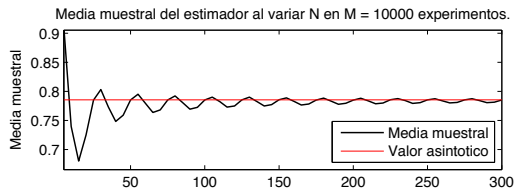


# Ejemplo III

## Ejemplo: MLE de la fase de una senoide

- Para determinar la cantidad de datos para que se cumplan las propiedades asintóticas se realiza una simulación en computadora.

$$A = 1, \quad f_0 = 0.08, \quad \phi = \pi/4, \quad \sigma^2 = 0.05 \quad (\text{con lo cual SNR}=10)$$



**Valores  
asintóticos:**

$$\hat{\phi} = \frac{\pi}{4}, \quad N \text{var}(\hat{A}) = \frac{1}{10}$$

$N > 100$  (approx)  
para alcanzar valores  
asintóticos

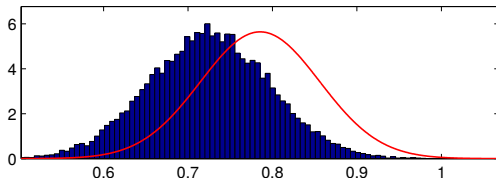
Hay valores de N  
preferenciales para  
el sesgo

# Ejemplo III

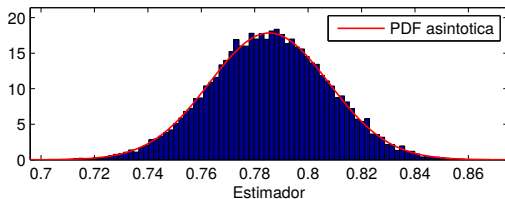
## Ejemplo: MLE de la fase de una senoide

- Para comparar la PDF verdadera con la PDF asintótica, se repite el experimento  $M = 20000$  veces con  $N = 20$  y con  $N = 200$ .

Histograma con  $N = 20$  en  $M = 20000$  experimentos.



Histograma con  $N = 200$  en  $M = 20000$  experimentos.



- La PDF asintótica es

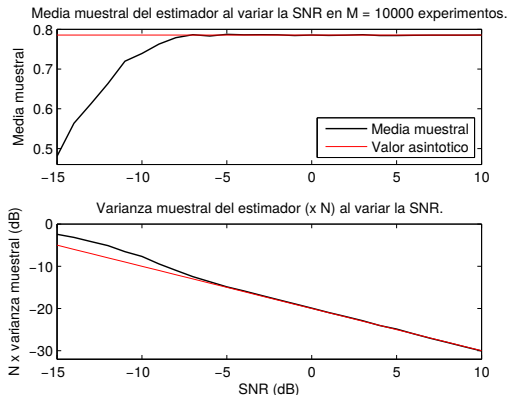
$$\hat{A} \sim \mathcal{N}\left(\frac{\pi}{4}, \frac{1}{10N}\right)$$

- $N = 20$  el sesgo es significativo.
- $N = 200$  se cumplen bien las propiedades asintóticas
- MLE desempeño pobre si conjunto de datos es chico.

# Ejemplo III

## Ejemplo: MLE de la fase de una senoide

- Se quiere analizar el desempeño del estimador al variar la SNR.
- Para eso se repite el experimento fijando  $N$  en 100 y se calcula la media y la varianza cambiando la SNR.



- Si SNR es pequeña, varianza supera CRLB considerablemente.
- La cota se alcanza con SNR altas.
- Condición para estar en regimen asintótico depende de  $N$  pero además de la SNR.

## Ejemplo: MLE de la fase de una senoide

### Observaciones/Resumen

- La PDF asintótica del estimador MLE es válida solo si el conjunto de datos es suficientemente grande.
- En problemas de estimación de parámetros de señales en ruido las condiciones asintóticas también dependen de la SNR
- Para establecer la cantidad de datos necesarios para estar en régimen asintótico, se pueden realizar simulaciones numéricas
- En este ejemplo, el estimador MLE encontrado analíticamente es aproximado. Para encontrar el MLE exacto, se puede recurrir a métodos de optimización para encontrar el cero de la función deseada.

# MLE de parámetros transformados

- En ocasiones, es necesario estimar una **función del parámetro  $\theta$** .
- Por ejemplo, en el problema de estimación del nivel de DC  $A$  en WGN, podría interesar calcular la potencia  $A^2$  de la señal.
- El MLE de una función del parámetro  $\theta$  se obtiene fácilmente a partir del MLE de  $\theta$ .

**Ejemplo: Nivel de DC transformado en WGN** Se consideran los datos

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

donde  $\sigma^2$  es conocido y se quiere estimar el MLE de  $\alpha = \exp(A)$ .

- La PDF de los datos es

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

# MLE de parámetros transformados

## Ejemplo: Nivel de DC transformado en WGN

- Como  $\alpha$  es una transformación biyectiva de  $A$ , es posible re-parametrizar la PDF en función de  $\alpha$ ,

$$p_T(\mathbf{x}; \alpha) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \log \alpha)^2 \right], \quad \alpha > 0. \quad (5)$$

El subíndice  $T$  refleja que la PDF es parametrizada respecto al parámetro transformado.

- Para encontrar el MLE de  $\alpha$ , hay que maximizar (5) en  $\alpha$ , llegando a

$$\sum_{n=0}^{N-1} (x[n] - \log \hat{\alpha}) \frac{1}{\hat{\alpha}} = 0 \quad \text{de donde} \quad \hat{\alpha} = \exp(\bar{x}).$$

- $\hat{A} = \bar{x}$  es el MLE de  $A$ , entonces  $\hat{\alpha} = \exp(\hat{A})$ .
- **Propiedad de invarianza:** El MLE del parámetro transformado es la transformación del MLE del parámetro original.

# MLE de parámetros transformados

## Ejemplo: Nivel de DC transformado en WGN

- Consideremos ahora la transformación  $\alpha = A^2$  para el conjunto de datos dado en el ejemplo anterior.
- Al intentar reparametrizar la PDF de  $A$  respecto a  $\alpha$  se observa que

$$A = \pm\sqrt{\alpha},$$

ya que en este caso la transformación no es uno a uno.

- Para caracterizar todas las posibles PDFs se requiere dos conjuntos de PDFs,

$$p_{T_1}(\mathbf{x}; \alpha) = p(\mathbf{x}; \sqrt{\alpha}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \sqrt{\alpha})^2 \right], \quad \alpha \geq 0$$

$$p_{T_2}(\mathbf{x}; \alpha) = p(\mathbf{x}; -\sqrt{\alpha}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] + \sqrt{\alpha})^2 \right], \quad \alpha > 0$$

- El MLE de  $\alpha$  se obtiene como

$$\hat{\alpha} = \arg \max_{\alpha} \{p_{T_1}(\mathbf{x}; \alpha), p_{T_2}(\mathbf{x}; \alpha)\}.$$

# MLE de parámetros transformados

## Ejemplo: Nivel de DC transformado en WGN

- De manera equivalente, el MLE puede encontrarse como el valor de  $\alpha$  que maximiza la **función de verosimilitud modificada**, construida como

$$\bar{p}_T(\mathbf{x}; \alpha) = \max_{\alpha} \{p_{T_1}(\mathbf{x}; \alpha), p_{T_2}(\mathbf{x}; \alpha)\}, \quad \text{para cada } \alpha \geq 0.$$

- En este ejemplo el MLE de  $\hat{\alpha}$  es

$$\begin{aligned}\hat{\alpha} &= \arg \max_{\alpha \geq 0} \{p_{T_1}(\mathbf{x}; \alpha), p_{T_2}(\mathbf{x}; \alpha)\} \\&= \arg \max_{\alpha \geq 0} \{p(\mathbf{x}; \sqrt{\alpha}), p(\mathbf{x}; -\sqrt{\alpha})\} \\&= \left[ \arg \max_{\sqrt{\alpha} \geq 0} \{p(\mathbf{x}; \sqrt{\alpha}), p(\mathbf{x}; -\sqrt{\alpha})\} \right]^2 \\&= \left[ \arg \max_{-\infty < A < \infty} p(\mathbf{x}; A) \right]^2 \\&= \hat{A}^2 \\&= \bar{x}^2.\end{aligned}$$

- La propiedad de invarianza se cumple aunque la transformación no sea biyectiva.



**Teorema: Propiedad de invarianza del MLE** El MLE del parámetro  $\alpha = g(\theta)$ , donde la PDF  $p(\mathbf{x}, \theta)$  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.

# Extensión a vector de parámetros

**Ejemplo:** Nivel de DC en WGN (estimar  $A$  y  $\sigma^2$ ) Se consideran las observaciones del nivel de continua en WGN,

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

- En este caso, el vector de parámetros es  $\theta = [A, \sigma^2]^T$ .
- La función de verosimilitud logarítmica queda

$$\log p(\mathbf{x}; \theta) = -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2,$$

- y las derivadas son (ejercicio)

$$\frac{\partial \log p(\mathbf{x}; \theta)}{\partial A} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)$$

$$\frac{\partial \log p(\mathbf{x}; \theta)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=0}^{N-1} (x[n] - A)^2.$$

# Extensión a vector de parámetros

## Ejemplo: Nivel de DC en WGN (estimar $A$ y $\sigma^2$ )

- Resolviendo para  $A$  en la primer ecuación se tiene que

$$\frac{1}{\hat{\sigma}^2} \sum_{n=0}^{N-1} (x[n] - \hat{A}) = 0 \quad \text{con lo cual} \quad \hat{A} = \bar{x}.$$

- Mientras que utilizando la segunda ecuación y sustituyendo el valor obtenido de  $\hat{A} = \bar{x}$ , se obtiene

$$-\frac{N}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{n=0}^{N-1} (x[n] - \hat{A})^2 = 0 \quad \text{con lo cual} \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2.$$

- El MLE es por lo tanto

$$\hat{\theta} = \begin{bmatrix} \bar{x} \\ \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 \end{bmatrix}.$$

## 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),

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

- **Kay, S. M.** (1993)  
*Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Capítulo 7.