ESTIMACIÓN

Estas transparencias contienen material adaptado del curso de PATTERN RECOGNITION AND MACHINE LEARNING de Heikki Huttunen y del libro Duda.

APRENDIZAJE AUTOMÁTICO, ESTIMACIÓN Y DETECCIÓN

- Introducción conceptos de teoría de estimación y detección:
    - Estimador de máxima verosimilitud
    - Estimación Bayesiana
    - Detección Neyman Pearson
    - Evaluación de desempeño, métricas.
    - Teoría Detección Bayesiana

ESTIMACIÓN

- Objetivo: Estimar valores de un conjunto de parámetros a partir de los datos. Ejemplos: registros de consumo, audio, valores pixeles de una imágenes, comunicaciones, control
- *Estimación de Parámetros:* Dado un conjunto de datos con N puntos

  los cuales dependen de un parámetro $\theta \in R$,

$$\mathbf{x} = \{x[0], x[1], \ldots, x[N-1]\}$$

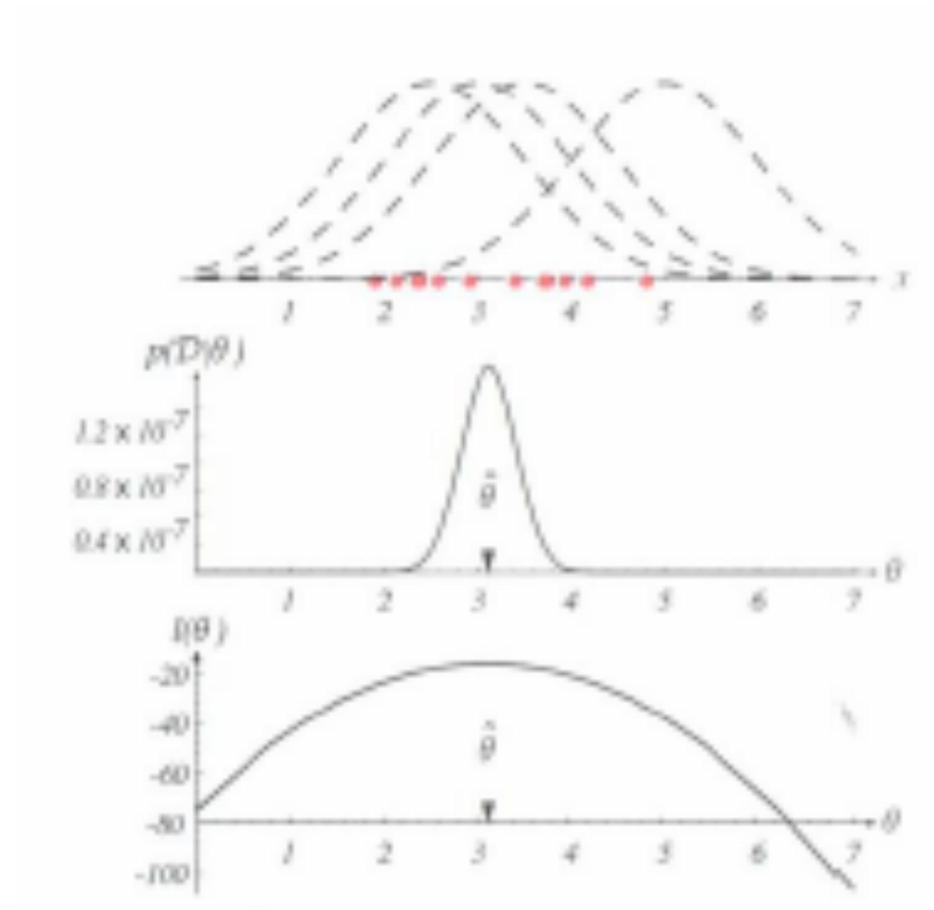
- Queremos diseñar un estimador g(.) para θ.

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

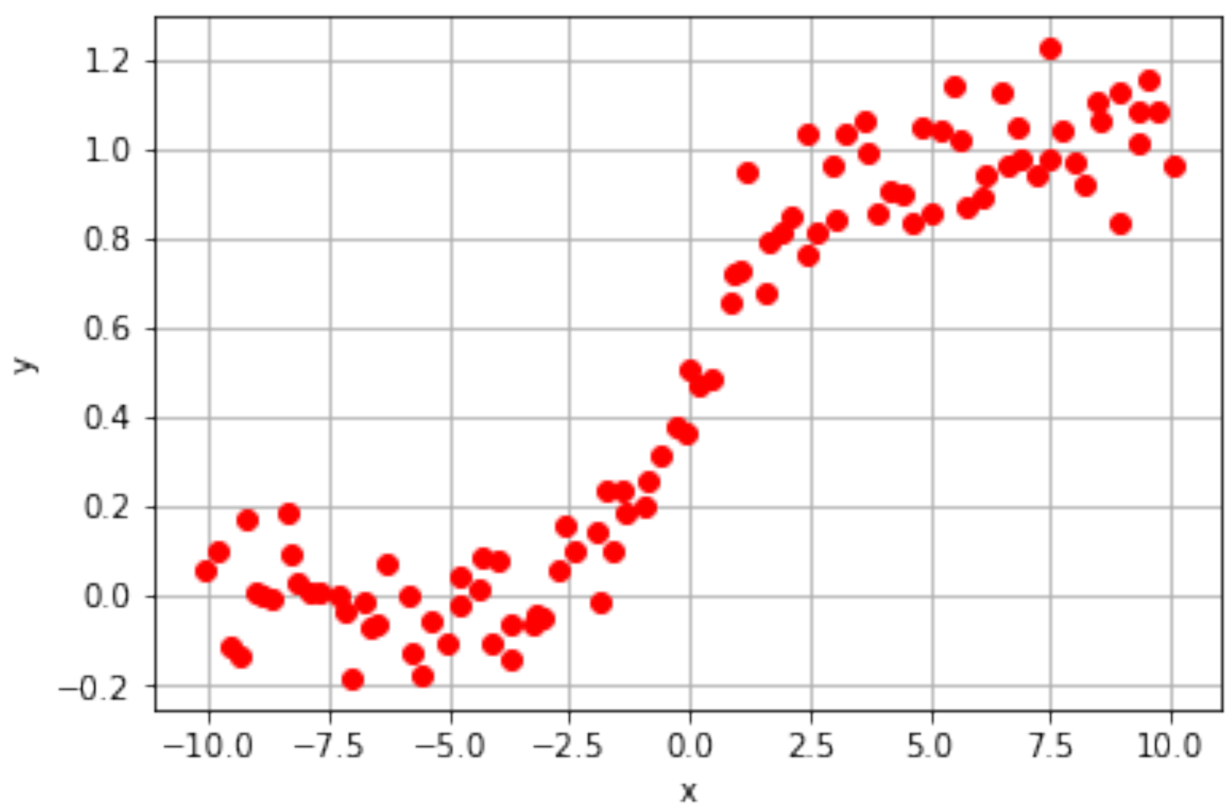
ESTIMACIÓN

¿Que modelo usamos para los datos?

¿Como determinamos los parametros?

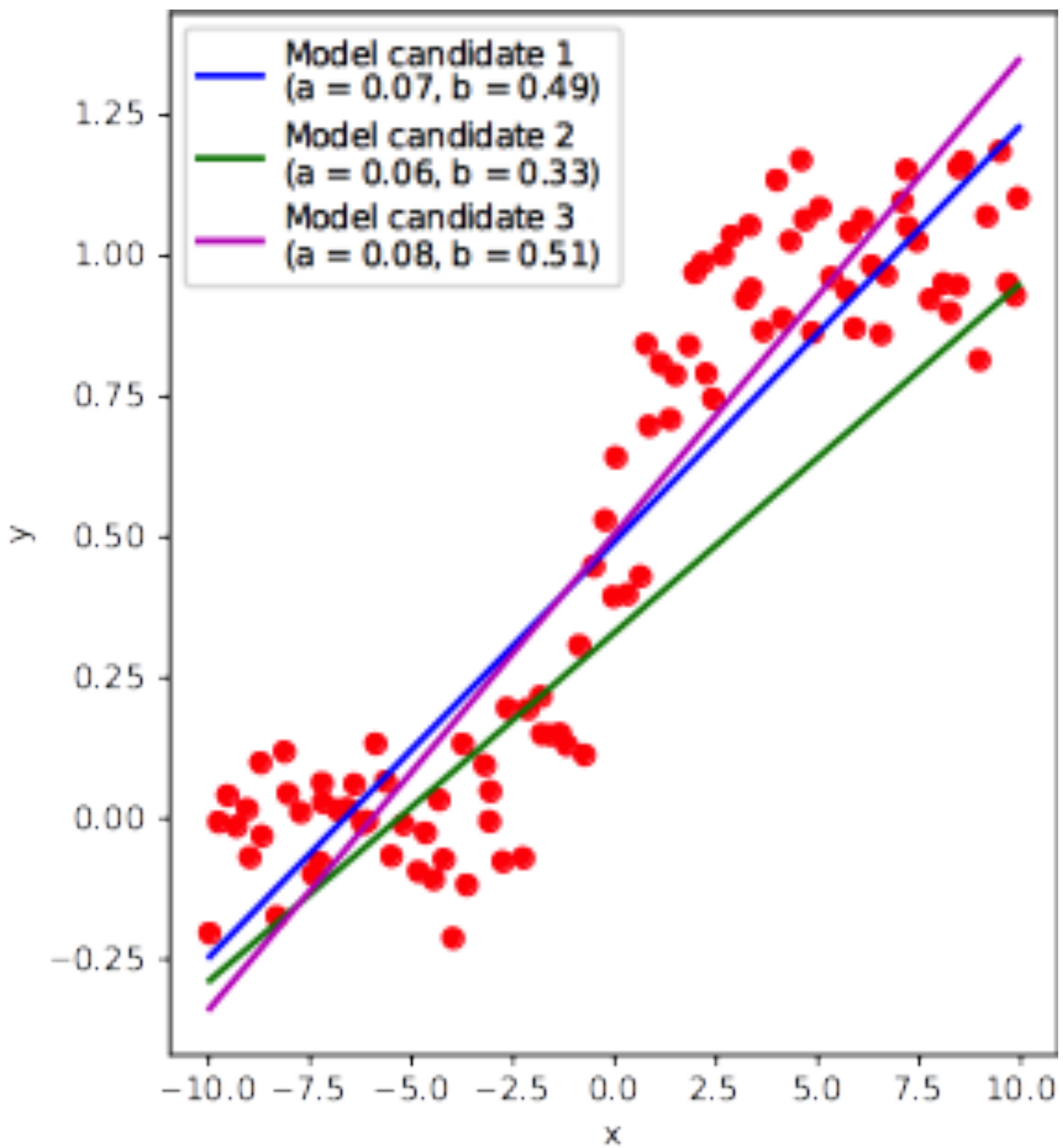


ESTIMACIÓN DE UNA RECTA



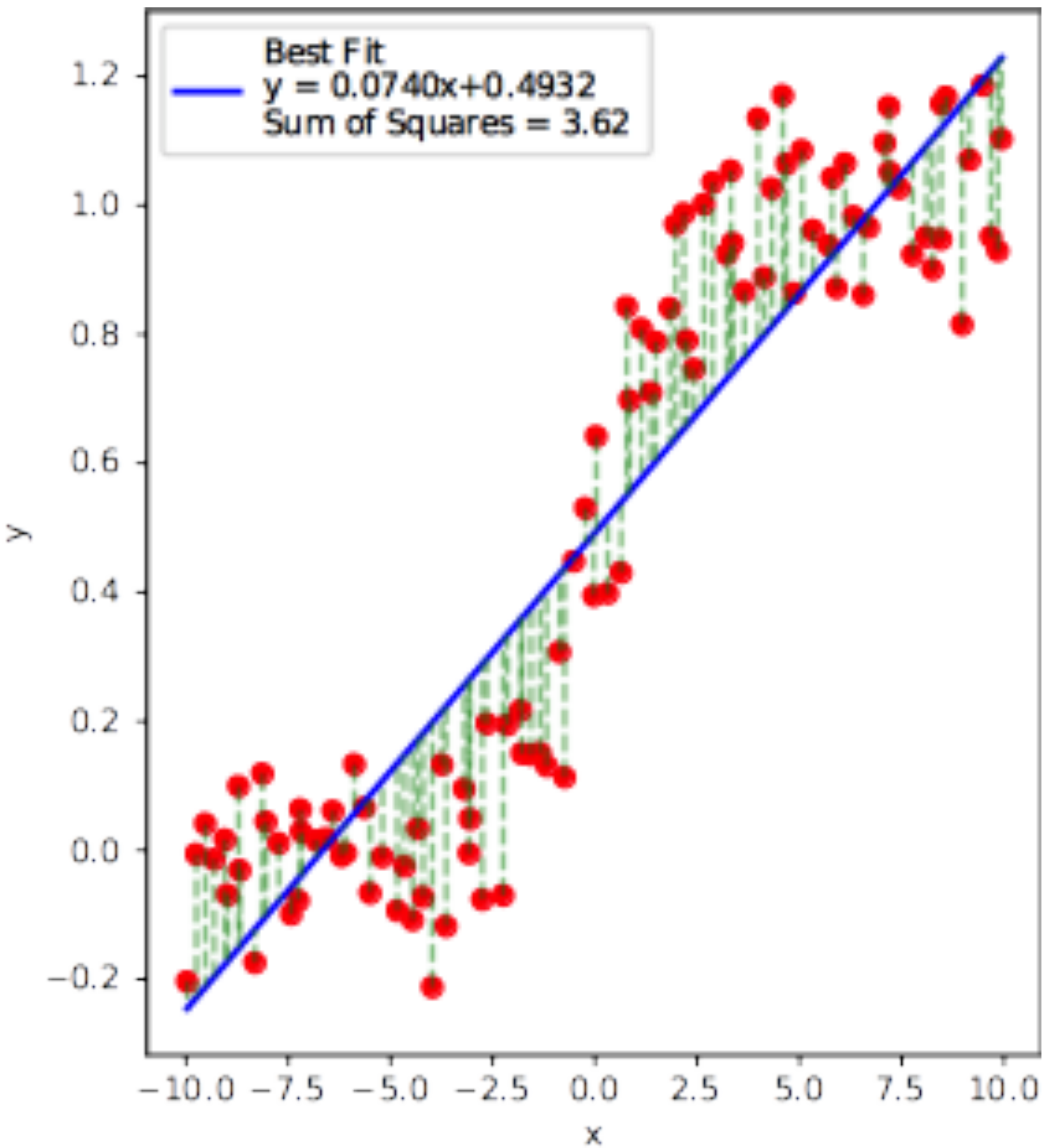
- Supongamos que tenemos la serie temporal de la figura y queremos determinar la relación entre las dos coordenadas.
- Asumiremos una relación lineal: $y[n] = ax[n] + b + w[n]$, con $a \in \mathrm{R}$ y $b \in \mathrm{R}$
- $w[n] \sim \mathrm{N}(0,\sigma 2)$
  N $(0, \sigma 2)$ es una distribución normal con media 0 y varianza $\sigma 2$

ESTIMACIÓN DE UNA RECTA



- Cada valor de a y b representa una recta. Cuál es la mejor recta?
- Depende del criterio que usemos: ej: mínimo error cuadrático.

ESTIMACIÓN DE PARÁMETROS



- $\hat{a} = 0.07401$ and $\hat{b} = 0.49319$
- Minimizan la suma de la distancia al cuadrado entre los puntos rojos y la recta azul.

ESTIMADOR DE MÍNIMOS CUADRADOS

- En esta aproximación se busca minimizar la diferencia al cuadrado entre los datos y el modelo de la señal. No se impone un modelo para el ruido.

$\hat{\theta}_{LS}$ valor de θ que minimiza el error cuadrático:

$$J(\theta) = \sum_{n=0}^{N-1} (y[n] - s[n;\theta])^2.$$

En el caso de la recta , $\boldsymbol{\theta} = [a, b]^T$

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} (y[n] - s[n;\boldsymbol{\theta}])^2 = \sum_{n=0}^{N-1} (y[n] - (ax[n] + b))^2.$$

ESTIMADOR DE MÍNIMOS CUADRADOS

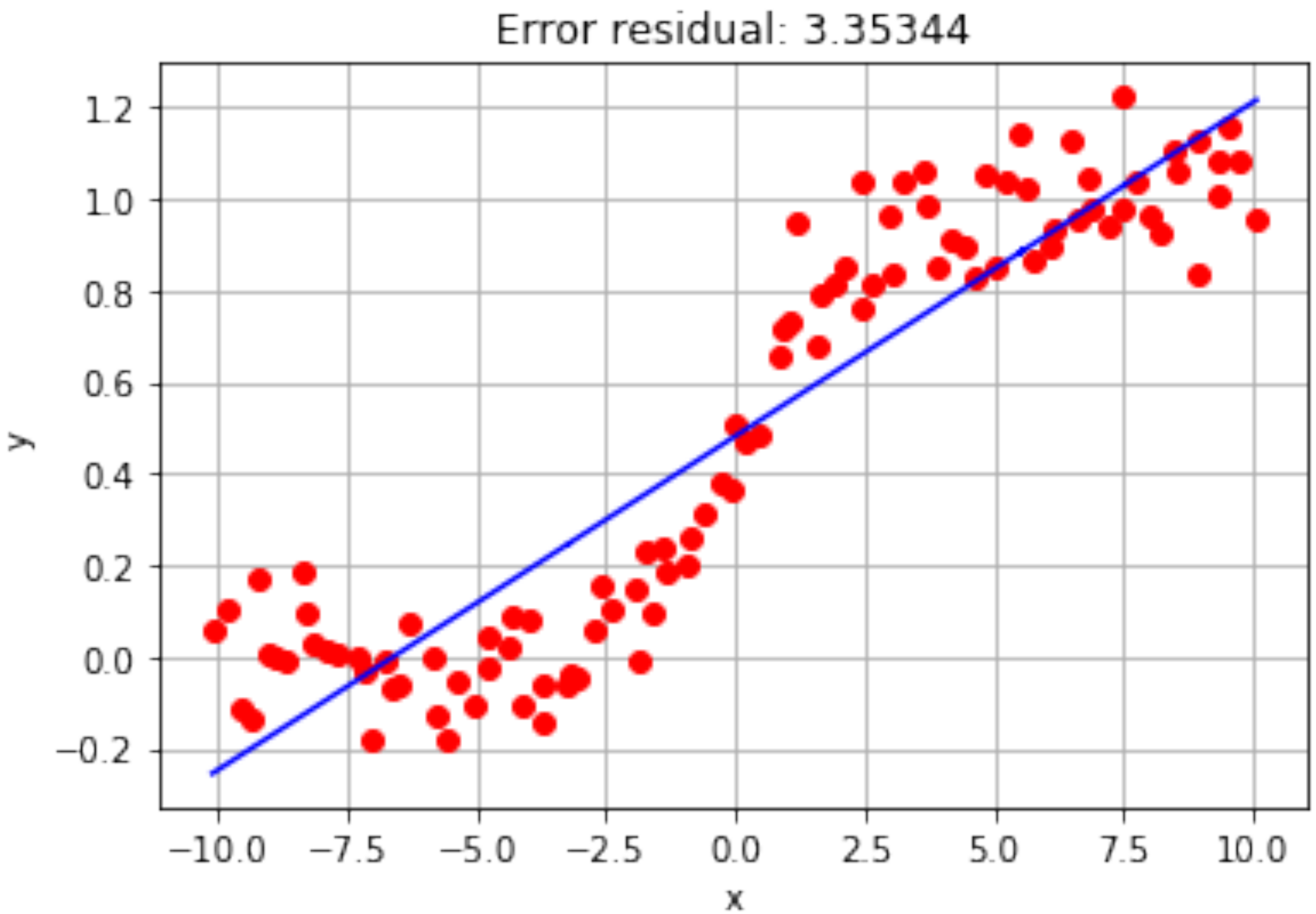
En forma matricial

$$\underbrace{\begin{pmatrix} y[0] \\ y[1] \\ \vdots \\ y[N-1] \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} x[0] & 1 \\ x[1] & 1 \\ x[2] & 1 \\ \vdots & \vdots \\ x[N-1] & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\boldsymbol{\theta}} + \mathbf{w}.$$

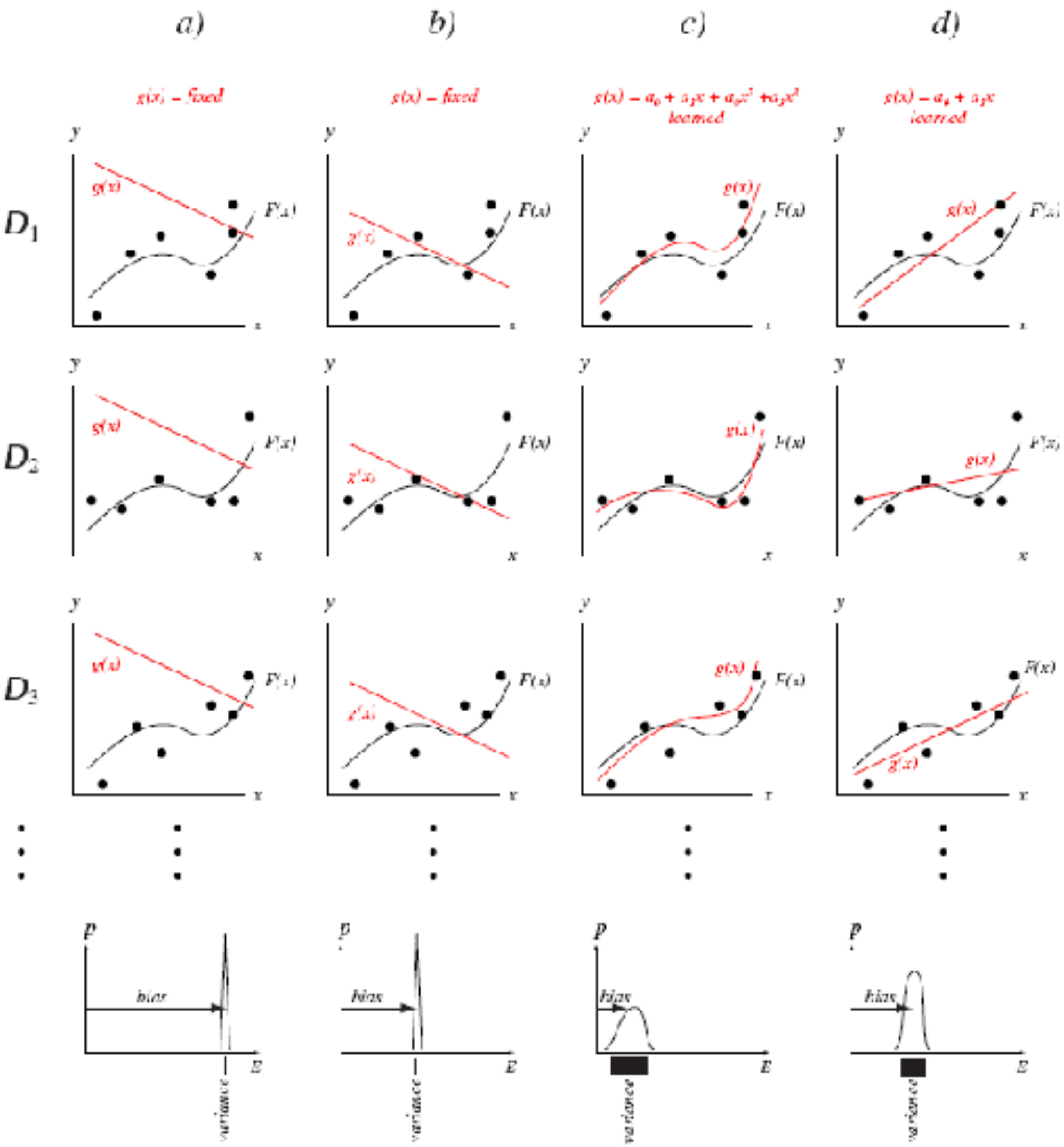
Estimador que minimiza el error: $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{w}$

$$\hat{\boldsymbol{\theta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

ESTIMADOR DE MINIMO CUADRADOS



PROPIEDADES DE UN ESTIMADOR



Compromiso

- Sesgo: $b(\hat{\theta}) = E(\hat{\theta}) - \theta$
  - *Insesgado* $b(\hat{\theta}) = 0$
- Varianza de un estimador :

$var(\hat{\theta}) = E(E(\hat{\theta}) - \hat{\theta})^2$

- Error cuadrático medio:

$MSE(\hat{\theta}) = var(\hat{\theta}) + b^2(\hat{\theta})$

EJ: EVALUACIÓN DE ESTIMADORES

Consideremos muestras de una señal continua en presencia de ruido, a la cuál le queremos estimar su valor: $x[n] = A + w[n],$

en presencia de ruido gaussiano de media nula y varianza desconocida $w[n] \sim \mathcal{N}(0, \sigma^2)$

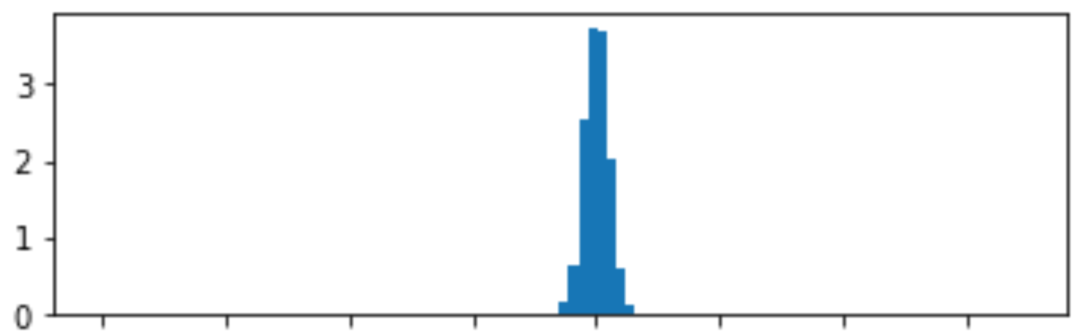
Propondremos dos estimadores posibles para A

¿Cual es mejor? $$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n].$$ $$\check{A} = x[0].$$

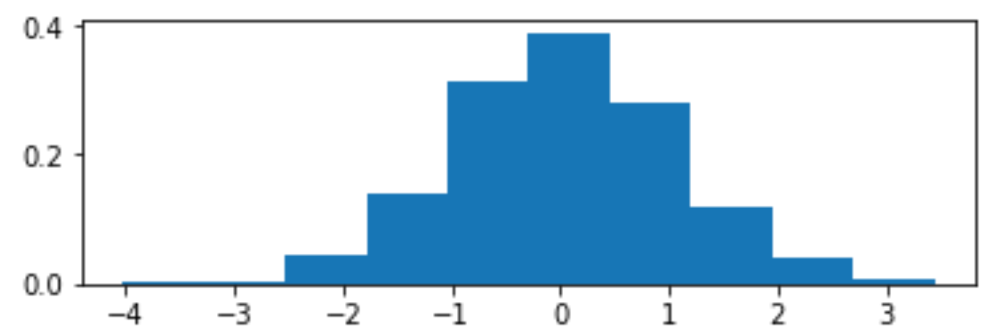
EJ: EVALUACIÓN DE ESTIMADORES

Analizaremos la varianza de cada estimador.

En forma empírica, evaluando 1000 realizaciones:



Estimador media muestran



Estimador primera muestra

EJ: EVALUACIÓN DE ESTIMADORES

En forma analítica:

1. Estimador de media muestral:

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

2. Estimador con la primer muestra: $\mathrm{var}(\check{A}) = \mathrm{var}(x[0]) = \sigma^2.$

El primer estimador tiene una varianza N veces más chica que el primero. ¿Que pasa con el sesgo y el MSE de ambos estimadores?

Verificar que ambos son insesgados y por lo tanto el muestral tiene menor MSE

DISEÑO DE UN ESTIMADOR

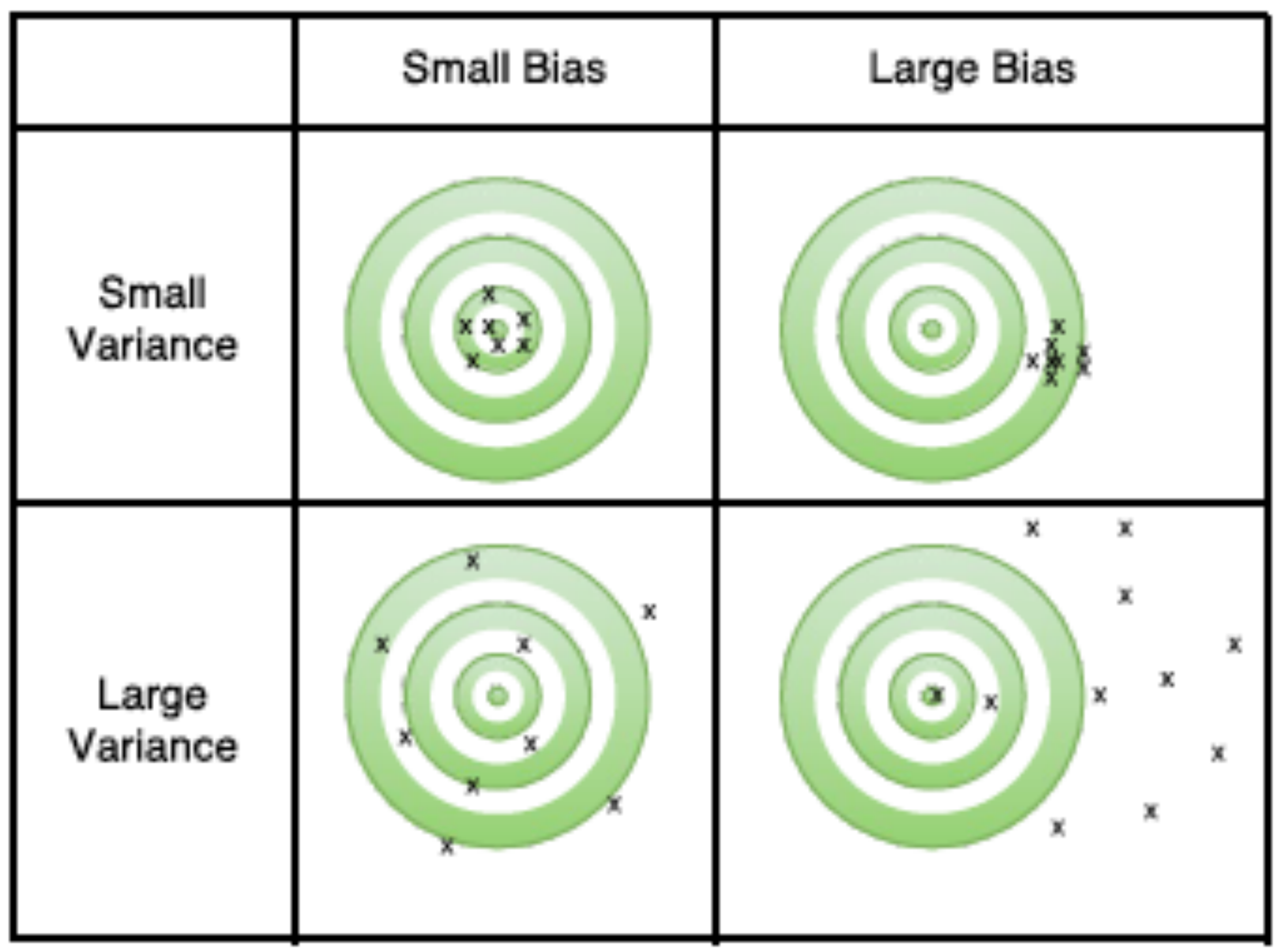
- Dependiendo de la aplicación puede ser ventajoso usar distintos estimadores:

1. **MVU** (Minimun Variance Unbiased Estimator): en forma analítica se determina dentro de todos los estimadores insesgados el de mínima varianza.
2. **MLE** (Maximun Likelihood Estimator): en forma analítica se determina el estimador que maximiza la verosimilitud con respecto a los datos medidos.
3. Otros : **MSE** (Minimun Square Error Estimator)

Nos concentraremos en MLE por ser de utilidad en aprendizaje estadístico.

ESTIMADOR INSESGADO DE MÍNIMA VARIANZA (MVU)

- MVU estimador óptimo pero difícil de encontrar o no existir.



ESTIMADOR MLE

- Método propuesto por Fisher in 1922, publicó los principios básicos en 1912 como estudiante de tercero en el grado.
- MLE es el estimador más usado debido a su aplicabilidad a problemas complejos.
- Idea básica: Encontrar los parámetros que con mayor probabilidad generaron los datos.
- Es un estimador que puede o no ser óptimo, en el sentido de mínima varianza y tampoco garantiza que sea insesgado.

LA FUNCIÓN VEROSIMILITUD

- Consideraremos el ejemplo de estimar la continua en una señal con ruido gaussiano de media nula, con pdf de ruido:

$$p(w[n]) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(w[n])^2\right]$$

- Por simplicidad consideraremos como estimador la primera muestra.

- Como $x[n] = A + w[n]$

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

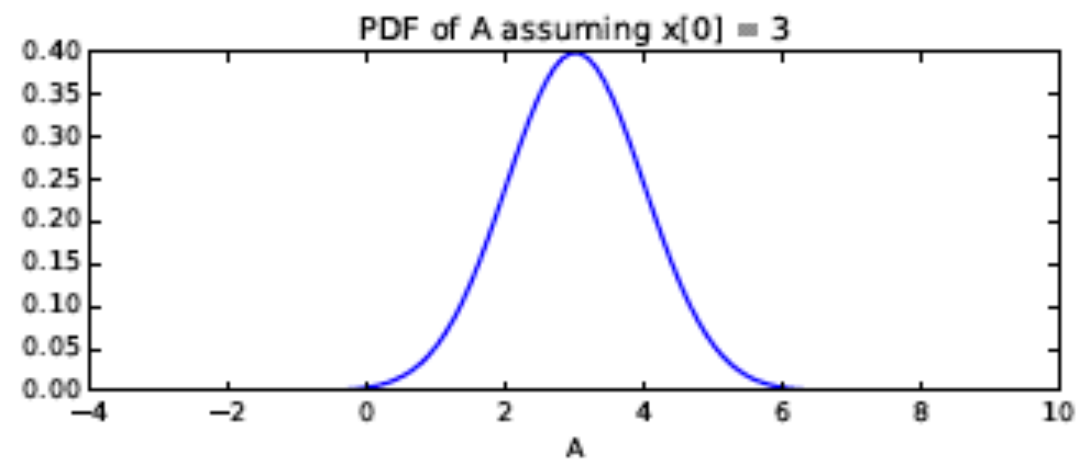
LA FUNCIÓN VEROSIMILITUD

- Si consideramos el caso de estimador x[0],

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

Si x[0]=3, algunos valores de A son más probables que otros y se puede determinar la pdf de A en forma fácil.

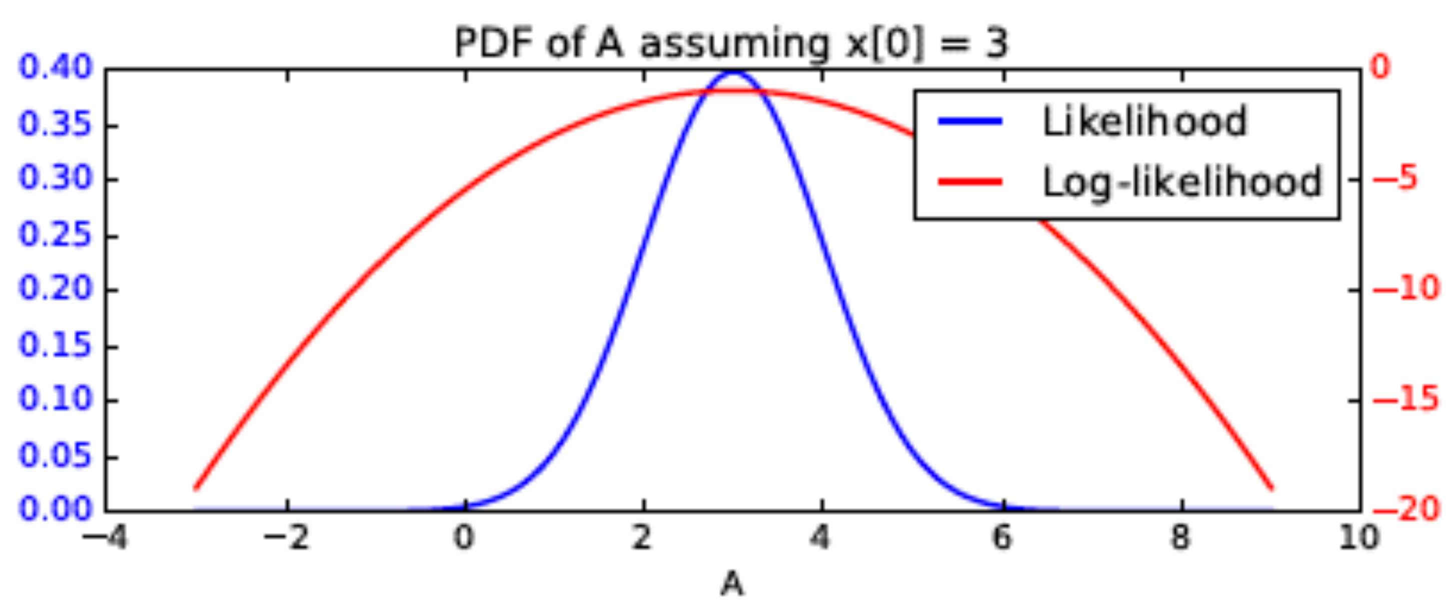
$$\text{pdf of } A = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(3-A)^2\right]$$



Esta función es la *función de verosimilitud* de A y su máximo es la *estimación de máxima verosimilitud.*

LA FUNCIÓN DE VEROSIMILITUD

- Se obtiene al tomar la pdf de los datos como función del parámetro desconocido supuestos los datos dados(fijos).
- Es usual que tenga una forma exponencial (pdf gaussiana) por lo que se toma el logaritmo y se obtiene una función log-verosimilitud.



MLE A PARTIR DE UN EJEMPLO

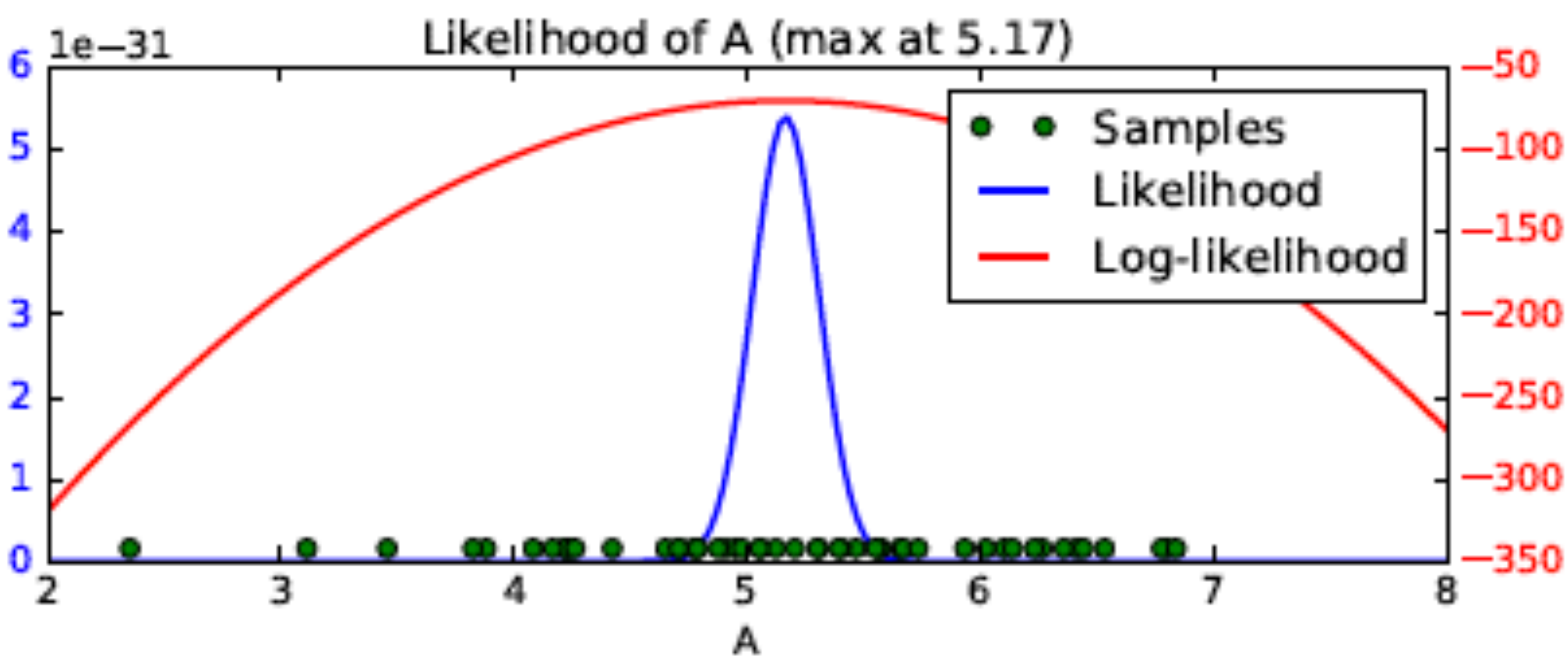
- Vamos a buscar el estimador MLE para el caso de la señal continua contaminada con ruido. Buscaremos el mejor estimador que explica las muestras observadas
- Asumiremos que todas las muestras x[i] son independientes (provienen de muestras de ruido índependientes), entonces:

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

- Para enfatizar que es una función de A, dados los datos a veces se escribe como p(A;x) o L(A;x)

EJEMPLO MLE - DETERMINACIÓN EN FORMA EMPÍRICA

- En el ejemplo se grafica p(A;x) y L(A;x) cuando se consideran 50 muestras que de la señal continua de valor verdadero de A=5, contaminadas con ruido.



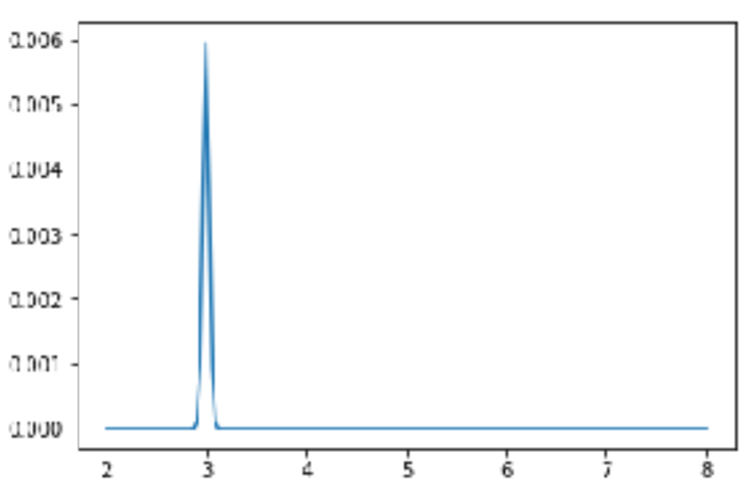
MLE EMPÍRICO EN PYTHON

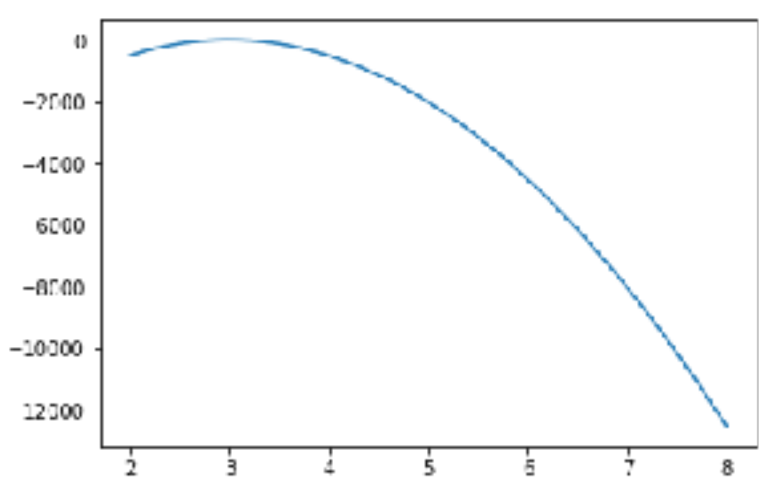
```
# The samples are in array called x0

x = np.linspace(2, 8, 200)
likelihood = []
log_likelihood = []

for A in x:
    likelihood.append(gaussian(x0, A, 1).prod())
    log_likelihood.append(gaussian_log(x0, A, 1).sum())

print ("Max likelihood is at %.2f" % (x[np.argmax(log_likelihood)]))
```





EJEMPLO MLE, DETERMINACIÓN ANALÍTICA

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

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

Para encontrar el máximo respecto a A, derivo:

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

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

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

Media muestral

GENERALIZACIÓN: ESTIMADOR DE MÁXIMA VEROSIMILITUD

- Cuando queremos diseñar un clasificador disponemos de muestras, si además conocemos el modelo para los datos podemos aplicar teoría de estimación y el problema se reduce a estimar los parámetros de las densidades.
- Dos enfoques:
  - Frecuentista: Estimador de Máxima Verosimilitud: Parámetros cantidades determinísticas. Ventajas: Buenas propiedades de convergencia al aumentar muestras de entrenamiento, simple.
  - Bayesiano: Parámetros variables aleatorias con una distribución a priori.

MLE

C conjuntos de datos $D_1 \dots D_c$ clasificados $(D_j \leftrightarrow w$

D_j : realización de un proceso aleatorio iid.

$p(\mathbf{x} / w_j)$ tiene forma paramétrica conocida

Ej : $p(\mathbf{x} / w_j) = N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

Notación para explicitar dependencia : $p(\mathbf{x} / w_j, \boldsymbol{\theta}_j)$

con $\boldsymbol{\theta}_j$ vector de parámetros desconocidos.

$\Longrightarrow$ c problemas de estimación desacoplados

MLE

D_i conjunto de muestras, de clase $w_i \rightarrow$ estimar $p(\mathbf{x}/w_i, \boldsymbol{\theta}_i)$

Notación simplificada : $D \leftarrow D_i$, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_i$ $D = \{\mathbf{x}_1 \ldots \mathbf{x}_n\}$ $\mathbf{x}_i \in R^d$

$$iid \rightarrow \boxed{p(D/\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k / \boldsymbol{\theta})}$$

$p(D/\boldsymbol{\theta})$: verosimilitud de $\boldsymbol{\theta}$ respecto a D

$$\boxed{\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(D/\boldsymbol{\theta})}$$

valor de $\boldsymbol{\theta}$ que más concuerda con las observaciones.

MLE

Como la función logaritmica es creciente estricta :

$$\boxed{\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(D/\boldsymbol{\theta})} \qquad l(\boldsymbol{\theta}) : \text{log – verosimilitud}$$

Condición necesaria para el estimador ML : $\boxed{\nabla_{\boldsymbol{\theta}} l\big|_{\hat{\boldsymbol{\theta}}} = 0}$

- Verificar que es un máximo (Hessiana definida negativa)
- Testear todos los máximos locales para encontrar máximo global

MLE – EJ: GAUSSIANA Μ Y Σ DESCONOCIDAS

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(D / \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \sum_{k=1}^{n} \left\{ \log\left[\left(2\pi\right)^{d} \det(\boldsymbol{\Sigma}^{-1}) \right] - (\mathbf{x}_k - \boldsymbol{\mu})^{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \right\}$$

Vamos a buscar los ceros del gradiente con respecto a μ y Σ^{-1}

$$\nabla_{\boldsymbol{\mu}} l = \sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}) \rightarrow \boxed{\hat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k}$$

Media muestral

MLE – EJ: GAUSSIANA Μ Y Σ DESCONOCIDAS

$$\nabla_{\Sigma^{-1}} l = \frac{1}{2}\sum_{k=1}^{n} \nabla_{\Sigma^{-1}} (\log(2\pi)^d \det\mathbf{\Sigma}) - \nabla_{\Sigma^{-1}} traza\left[\mathbf{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T\right]$$

$$\nabla_{\Sigma^{-1}} l = \frac{1}{2}\sum_{k=1}^{n} \mathbf{\Sigma}^T - (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T = 0$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T$$

covarianza muestral $\quad \boldsymbol{\mu} \leftarrow \hat{\boldsymbol{\mu}}$ $\quad \boxed{\hat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T}$

MLE – EJ: GAUSSIANA Μ Y Σ DESCONOCIDAS

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \qquad \mathbf{x}_k \text{ iid} \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$E(\hat{\boldsymbol{\mu}}) = \frac{1}{n}\sum_{k=1}^{n}E(\mathbf{x}_k) = \boldsymbol{\mu} \quad \rightarrow \boxed{\hat{\boldsymbol{\mu}} \quad \text{insesgado}}$$

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_n)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_n)^T$$

$$E(\hat{\boldsymbol{\Sigma}}_n) = \frac{1}{n}\sum_{k}\left[E(\mathbf{x}_k\mathbf{x}_k^T) + \frac{1}{n^2}\sum_{ij}E(\mathbf{x}_i\mathbf{x}_j^T) - \frac{1}{n}\sum_{i}E(\mathbf{x}_k\mathbf{x}_i^T) - \frac{1}{n}\sum_{i}E(\mathbf{x}_k\mathbf{x}_i^T)\right]$$

MLE – EJ: GAUSSIANA Μ Y Σ DESCONOCIDAS

$$E\left[(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_j-\boldsymbol{\mu})^T\right]=E\left(\mathbf{x}_i\mathbf{x}_j{}^T\right)-\boldsymbol{\mu}\boldsymbol{\mu}^T$$

$$E(\hat{\boldsymbol{\Sigma}}_n)=\frac{1}{n}\sum_k\left[\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T+\frac{1}{n^2}\left(n\boldsymbol{\Sigma}+n^2\boldsymbol{\mu}\boldsymbol{\mu}^T\right)-\frac{2}{n}\left(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T n\right)\right]$$

$$E(\hat{\boldsymbol{\Sigma}}_n)=\left(1-\frac{1}{n}\right)\boldsymbol{\Sigma}=\left(\frac{n-1}{n}\right)\boldsymbol{\Sigma}\neq\boldsymbol{\Sigma}$$

$$\boxed{\hat{\boldsymbol{\Sigma}}_n:\quad\left\{\begin{matrix}\text{sesgado}\\ \text{asintoticamente insesgado}\end{matrix}\right\}}$$

PROPIEDADES DE UN ESTIMADOR

- Independientemente de si asumimos θ determinista o aleatorio, su estimación es una variable aleatoria función de las observaciones.
- Para caracterizar un estimador se calcula su sesgo, su varianza y su error cuadrático medio.

$$\boxed{\begin{aligned} &MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] \\ &\operatorname{var}\left(\hat{\theta}\right) = E\left(\hat{\theta} - E\left(\hat{\theta}\right)\right)^2 \quad b\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta \end{aligned}}$$

COMPROMISO SESGO VARIANZA DE UN ESTIMADOR

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

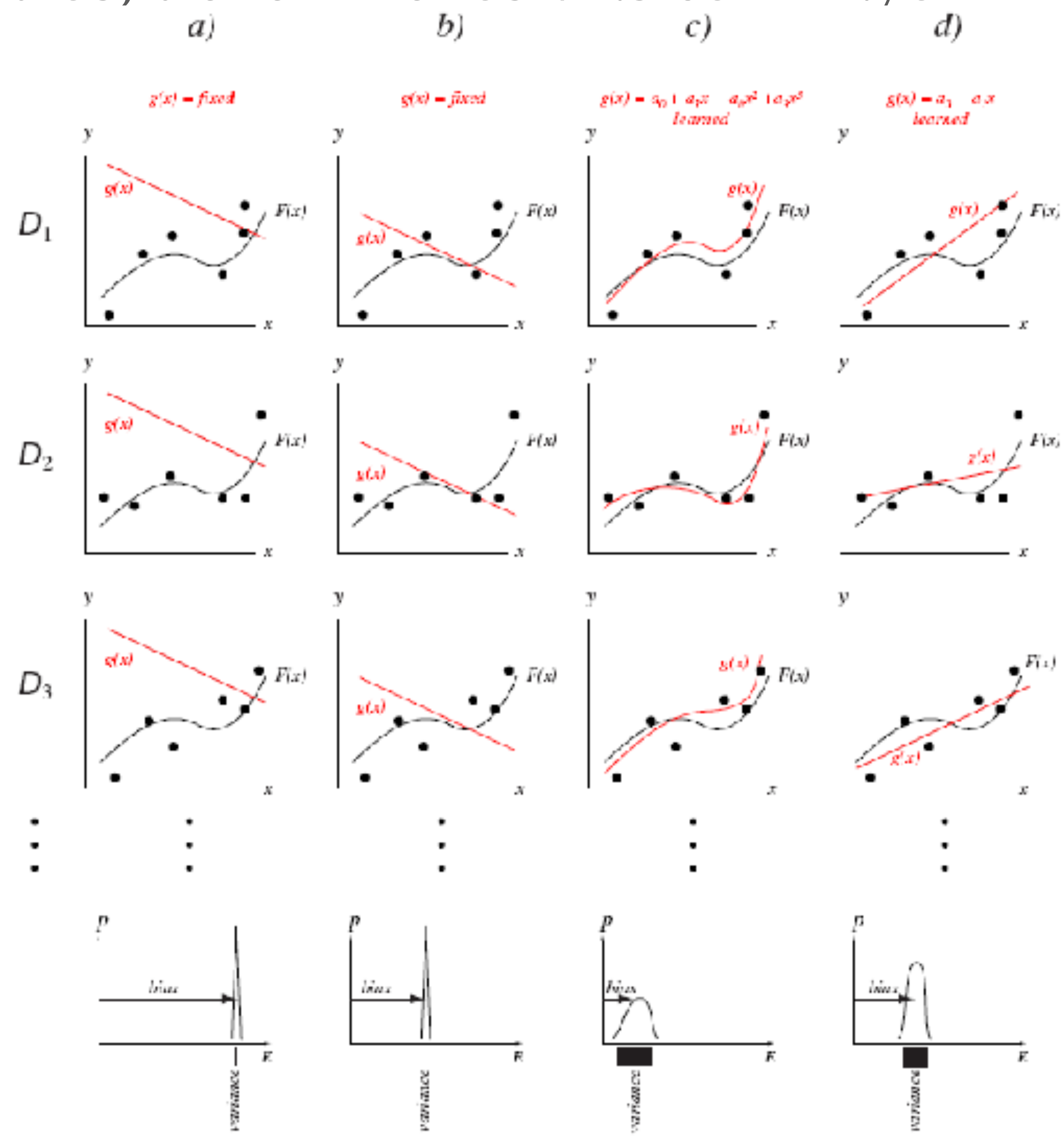
$$= \mathrm{var}\left(\hat{\theta}\right) + 2\left(E\left(\hat{\theta}\right) - \theta\right)E\left[\hat{\theta} - E\left(\hat{\theta}\right)\right] + \left[E\left(\hat{\theta}\right) - \theta\right]^2$$

$$\boxed{MSE(\hat{\theta}) = \mathrm{var}\left(\hat{\theta}\right) + b^2\left(\hat{\theta}\right)}$$

$$MSE(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \mathrm{var}\left(\hat{\theta}_i\right) + b^2\left(\hat{\theta}_i\right)$$

PROPIEDADES DE UN ESTIMADOR

- Procedimientos con mayor flexibilidad para adaptarse a los datos, mayor número de parámetros, tienen menos vías con mayor varianza.



COMPROMISO SESGO - VARIANZA

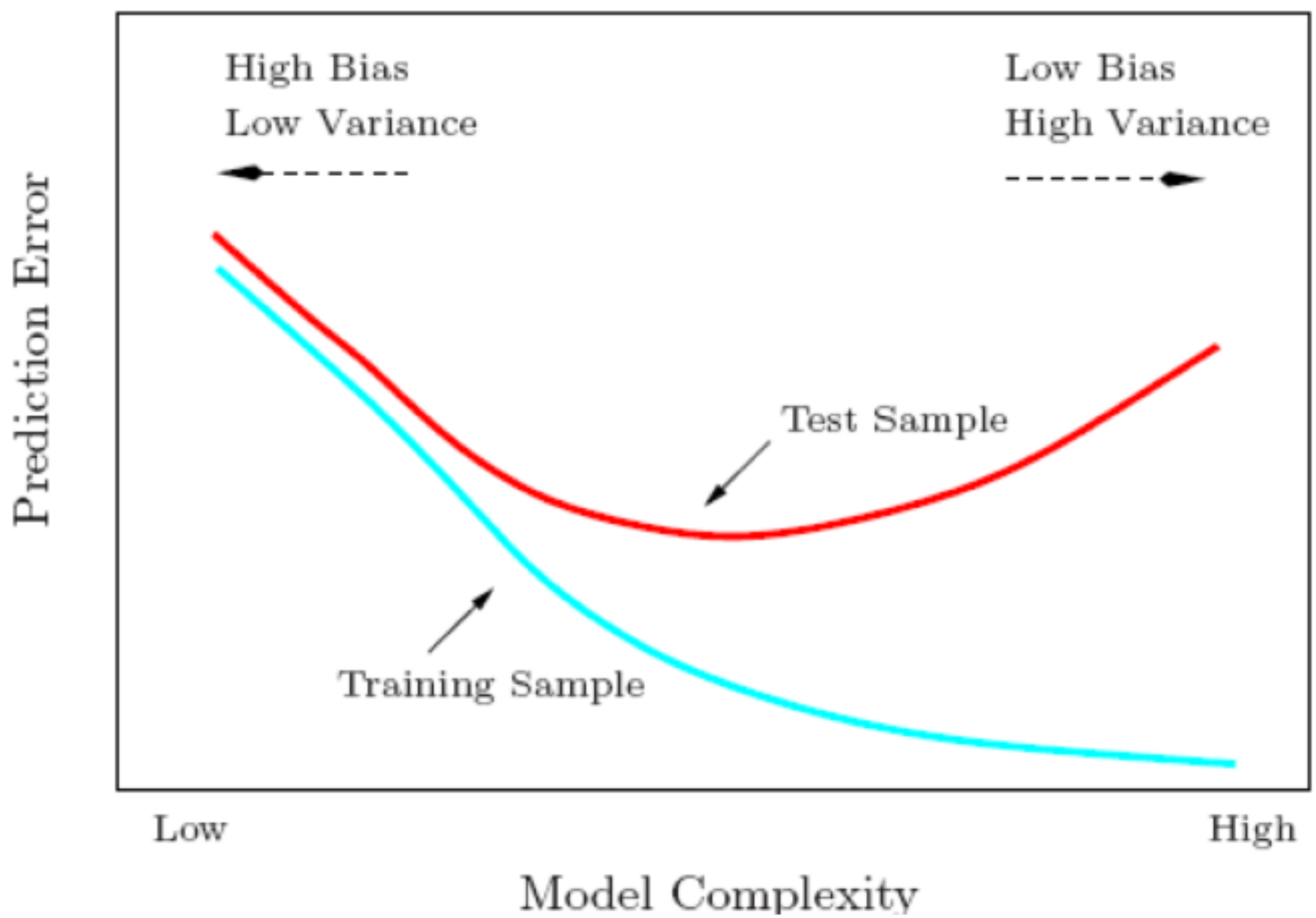


Fig. 2.11 Hastie

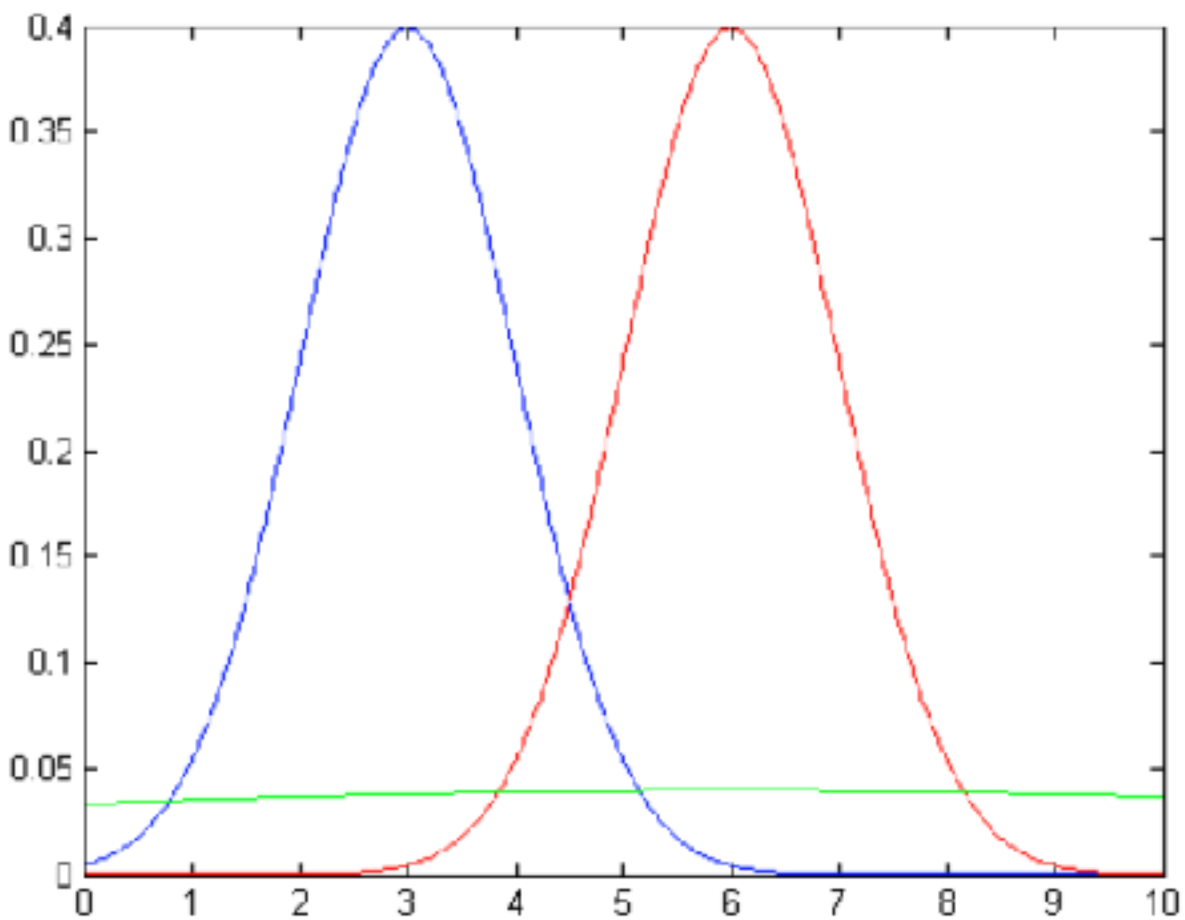
PROPIEDADES DE UN ESTIMADOR MLE

- Es deseable obtener estimadores insesgados
- Compromiso sesgo varianza: En algunos casos introducir un pequeño sesgo reduce mucho la varianza y por lo tanto el error cuadrático medio.
- La cota de Cramer-Rao nos da para un problema de estimación determinado la varianza mínima de un estimador insesgado,
- Un estimador es eficiente si es de varianza mínima.
- MLE es asintóticamente eficiente y asintóticamente insesgado.

PROPIEDADES ASISTÓTICAS DE MLE

- Cuando el número de muestras tiende a infinito MLE sigue una ley gaussiana con media nula y con varianza mínima.
- En problemas de reconocimiento de patrones con conjuntos de entrenamiento con muchos datos un estimador asintóticamente insesgado y eficiente es aceptable.
- Si el modelo es adecuado un estimador MLE da buenos resultados.
- ¿Que pasa si el modelo no lo es?

PROPIEDADES DE MLE



$$p(x/w_1) = N(3,1) \quad \text{azul}$$

$$p(x/w_2) = N(6,10) \quad \text{verde - distribución real de los datos}$$

$$p(x/w_2) = N(\hat{\mu},1) \quad \text{roja - modelo asumido } \hat{u}$$

umbral propuesto : 4,5

umbral óptimo ≈ 5

ESTIMACIÓN BAYESIANA

- Estimación de densidades utilizando toda la información disponible: Prioris y Datos.
- Hipótesis:

1. $p(x/\theta)$: conocida pero no se conoce el vector de parámetros en forma exacta.
2. Conocimiento a priori de θ en $p(\theta)$.
3. Resto del conocimiento a cerca de θ está contenido en el conjunto D de muestras tomadas en forma iid de acuerdo a $p(x)$ desconocida.

ESTIMACIÓN BAYESIANA

$$D = \bigcup_{1}^{c} D_i \qquad D_i \bigcap D_j = \phi$$

$D_i \leftrightarrow w_i$: muestras de entrenamiento clase i

D : conjunto de muestras de entrenamiento

$\mathbf{x}$: una muestra sin clasificar

$$P(w_i / \mathbf{x}, D) = \frac{p(\mathbf{x} / w_i, D) P(w_i / D)}{\sum_{j=1}^{c} p(\mathbf{x} / w_j, D) P(w_j / D)}$$

supondremos : $P(w_i / D) = P(w_i)$ prioris conocidas

$\forall i \neq j$ las muestras D_i no tienen influencia sobre $p(\mathbf{x} / w_j, D)$:

esto es $\quad p(\mathbf{x} / w_j, D) = p(\mathbf{x} / w_j, D_j) \;\; \forall j$

$$\rightarrow P(w_i / \mathbf{x}, D) = \frac{p(\mathbf{x} / w_i, D_i) P(w_i)}{\sum_{j=1}^{c} p(\mathbf{x} / w_j, D_j)\, P(w_j)}$$

Podemos tratar cada clase de forma independiente

para aliviar notación : $\quad D_i = D, \quad w_i = w$

$$P(w / \mathbf{x}, D) = \frac{p(\mathbf{x} / w, D) P(w)}{p(\mathbf{x} / D)}$$

DISTRIBUCIÓN DE LOS PARÁMETROS

- Supondremos que la densidad $p(x)$ es paramétrica de forma conocida y parámetros θ desconocidos ($p(x/\theta)$ completamente conocida)
- La observación de muestras aporta nueva información y da lugar a la probabilidad a posteriori $p(\theta/D)$ que esperamos que sea más en pico en torno al verdadero valor de θ que el prior $p(\theta)$ conocida.

DISTRIBUCIÓN DE PARÁMETROS

Objetivo: Encontrar $p(\mathbf{x}/D)$ que es lo más cerca que puedo estar de $p(\mathbf{x})$.

$$p(\mathbf{x}/D) = \int p(\mathbf{x}, \boldsymbol{\theta}/D)d\boldsymbol{\theta}$$

$$p(\mathbf{x}, \boldsymbol{\theta}/D) = p(\mathbf{x}/\boldsymbol{\theta}, D)\,p(\boldsymbol{\theta}/D)$$

$$\Rightarrow \boxed{p(\mathbf{x}/D) = \int p(\mathbf{x}/\boldsymbol{\theta})\,p(\boldsymbol{\theta}/D)d\boldsymbol{\theta}}$$

$$\text{usando Bayes:} \quad p(\boldsymbol{\theta}/D) = \frac{p(D/\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(D/\boldsymbol{\theta})\,p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

EJEMPLO: GAUSSIANA M DESCONOCIDA

$$p(x/\theta) = N(\mu, \sigma^2) \quad \sigma^2 : \text{conocido},$$

$$\text{prior sobre } \mu \quad p(\mu) = N(\mu_0, \sigma_0^{\ 2}),$$

$$\mu_0 : \text{lo que creemos} \quad \sigma_0^{\ 2} : \text{incertidumbre}$$

$$p(\mu / D) = \frac{p(D/\mu)p(\mu)}{\int p(D/\mu)p(\mu)d\mu}$$

$$D = \{x_1, \ldots\ldots x_n\} \quad \text{independientes}$$

$$p(\mu / D) = \alpha \prod_{k=1}^{n} p(x_k / \mu) p(\mu) = \alpha' \exp\left(-\frac{1}{2} \left(\sum_{k=1}^{n} \frac{(x_k - \mu)^2}{\sigma^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^{\ 2}} \right) \right)$$

$$= \alpha'' \exp\left(-\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^{\ 2}} \right) \mu^2 - 2 \left(\frac{1}{\sigma^2} \sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^{\ 2}} \right) \mu \right] \right)$$

$$\boxed{p(\mu / D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right]}$$

Identificando coeficientes :

$$\begin{cases} \dfrac{1}{\sigma_n^{\;2}} = \dfrac{\mathrm{n}}{\sigma^2} + \dfrac{1}{\sigma_0^{\;2}} \\[2ex] \dfrac{\mu_n}{\sigma_n^{\;2}} = \dfrac{n\hat{\mu}_n}{\sigma^2} + \dfrac{\mu_0}{\sigma_0^{\;2}} \end{cases}$$

$$\begin{cases} \mu_n = \dfrac{n\sigma_0^{\;2}}{n\sigma_0^{\;2} + \sigma^2}\hat{\mu}_n + \dfrac{\sigma^2}{n\sigma_0^{\;2} + \sigma^2}\mu_0 \\[2ex] \sigma_n^2 = \dfrac{\sigma^2\sigma_0^{\;2}}{n\sigma_0^{\;2} + \sigma^2} \end{cases}$$

EJEMPLO: GAUSSIANA M DESCONOCIDA

i) $\underset{n \to \infty}{\sigma_n \to 0}$: al aumentar la cantidad de muestras disminuye incertidumbre

ii) $\underset{n \to \infty}{\mu_n \to \hat{\mu}_n}$: la influencia del prior disminuye

iii) si $\sigma_0 >> \sigma \to \mu_n \approx \hat{\mu}_n$: confiamos más en los datos que en los priors

si $\sigma_0 = 0 \to \mu_n = \mu_0$: tenemos confianza ∞ $\mu_n = \mu_0$.

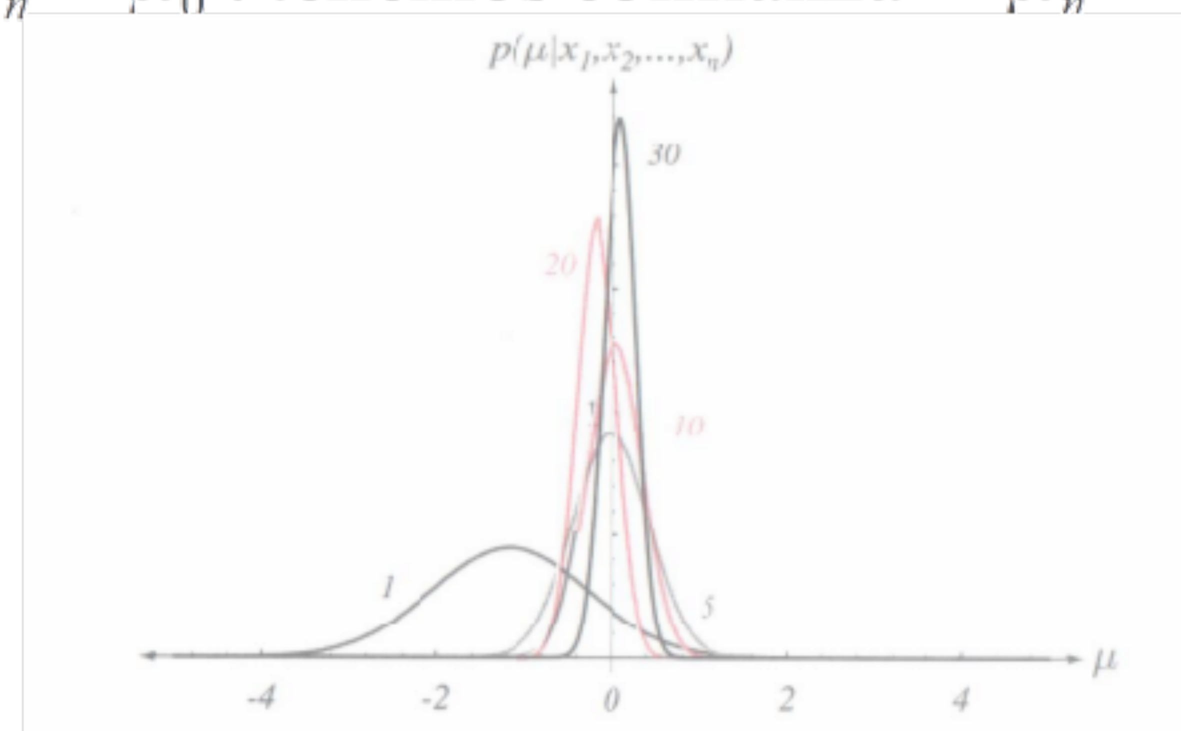


Fig. 3.2 Duda

Obtenida densidad a posteriori, podemos calcular $p(x/D)$

$$\boxed{p(x/D) = \int p(x/\mu)p(\mu/D)d\mu}$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_n)^2}{\sigma_n^{\ 2}} \right] \right) d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_n} f(\sigma,\sigma_n,x) \exp\left(-\frac{1}{2} \frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^{\ 2}} \right)$$

donde $\quad f(\sigma,\sigma_n,x) = \int_R \exp\left(-\frac{1}{2} \frac{\sigma^2+\sigma_n^{\ 2}}{\sigma^2\sigma_n^{\ 2}} \left(\mu - \frac{\sigma_n^{\ 2}x+\sigma^2\mu_n}{\sigma^2+\sigma_n^{\ 2}} \right)^2 \right) d\mu$

$$\boxed{f(\sigma,\sigma_n,x) = \int_R \exp\left(-\frac{1}{2}\left(\frac{\mu-\alpha}{\beta} \right)^2 \right) d\mu = \sqrt{2\pi}\beta \quad \text{con} \quad \beta = \frac{\sigma\sigma_n}{\sqrt{\sigma^2+\sigma_n^{\ 2}}}}$$

$$\Rightarrow \quad \boxed{p(x/D) = N(\mu_n, \sigma^2 + \sigma_n{}^2)}$$

σ_n : incertidumbre en la estimación de μ_n

σ : incertidumbre nuestra medida

Esto es válido para cada clase,

determinamos $p(x/w_j, D_j)$ $j = 1..c$

Clasificación : $\quad P(w_j / x, D) = kp(x / w_j, D_j)P(w_j)$

$$\boxed{\text{Decido} \quad x \in w_{j^*} \quad \text{con} \quad j^* = \arg\max_j P(w_j / x, D)}$$

ESTIMACIÓN BAYESIANA

- A diferencia de MLE que para la estimación de $p(x/D)$ tiene en cuenta una estimación puntual de los parámetros en la estimación Bayesiana integra la densidad a posteriori $p(\theta/D)$.
- Para el caso gaussiano multivariado el resultado es análogo considerando vectores medias y matrices covarianza.
- ¿Comó hacemos los cálculos con densidades cualesquiera?

APRENDIZAJE BAYESIANO RECURSIVO INCREMENTAL

$$D = \{\mathbf{x}_1 \ldots \mathbf{x}_n\} \quad \text{iid} \approx p(\mathbf{x}/\boldsymbol{\theta}) \rightarrow p(D/\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_k/\boldsymbol{\theta})$$

$$p(\boldsymbol{\theta}/D) = \frac{p(D/\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(D/\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

$$\text{Notemos:} \quad D^i = \{\mathbf{x}_1 \ldots \mathbf{x}_i\} \quad i = 1, \ldots n$$

$$p(D^n/\boldsymbol{\theta}) = p(D^{n-1}/\boldsymbol{\theta})\,p(\mathbf{x}_n/\boldsymbol{\theta})$$

$$p(\boldsymbol{\theta}/D^0) = p(\boldsymbol{\theta}) \quad \text{prior}$$

$$\Rightarrow p(\boldsymbol{\theta}/D^n) = \frac{p(\mathbf{x}_n/\boldsymbol{\theta})\,p(\boldsymbol{\theta}/D^{n-1})}{\int p(\mathbf{x}_n/\boldsymbol{\theta})\,p(\boldsymbol{\theta}/D^{n-1})\,d\boldsymbol{\theta}}$$

VÍNCULO CON MLE

Si $p(D/\boldsymbol{\theta})$ tiene un pico pronunciado en $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$

y $p(\hat{\boldsymbol{\theta}}) \neq 0$ con $p(\boldsymbol{\theta})$ suave en un entorno de $\hat{\boldsymbol{\theta}}$,

como $p(\boldsymbol{\theta} / D) = p(D/\theta)\, p(\boldsymbol{\theta})$,

$p(\boldsymbol{\theta} / D)$ también tiene un pico pronunciado en $\hat{\boldsymbol{\theta}}$

$$\boxed{p(\mathbf{x} / D) = \int p(\mathbf{x} / \boldsymbol{\theta}) p(\boldsymbol{\theta} / D) d\boldsymbol{\theta} \approx p(\mathbf{x} / \hat{\boldsymbol{\theta}}) \quad \text{verosimilitud}}$$

ESTIMADOR MÁXIMO A POSTERIORI (MAP)

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} / D) = \arg\max_{\theta} \left\{ \ln p(D / \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \right\}$$

$$\text{si } \ \mathrm{p}(\boldsymbol{\theta}) = cte \ \Rightarrow \boxed{\hat{\boldsymbol{\theta}}_{MAP} = \hat{\boldsymbol{\theta}}_{MLE}}$$

El estimador MAP no está bien visto por los Bayesianos ya que reduce una densidad a un valor determinista

COMPARACIÓN MLE Y ESTIMACIÓN BAYESIANA

- Para prioris razonables ambas soluciones son equivalentes cuando n $\rightarrow\infty$.
- ¿Qué pasa con conjunto de datos limitados?

1. Complejidad:
    1. MLE: Cálculo diferencial , métodos gradiente.
    2. Bayesiano: Integración multidimensional.
2. Interpretabilidad:
    1. MLE: más fácil de intrepretar.
    2. Bayesiano: promedio ponderado de los modelos, refleja incertidumbre.
3. Confianza en la información a priori.
    1. MLE: asume la forma paramétrica original
    2. Bayesiano: no asume la forma paramétrica original. Ej gaussiana varianza conocida.