

# Probabilidad y Estadística para Inteligencia Artificial

Dr. Ing. Pablo Briff

Laboratorio de Sistemas Embebidos - FIUBA

*pbriff@fi.uba.ar*

25 de Julio de 2020



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Tabla de Contenidos I

- 1 Estimadores Puntuales
  - Estimador
- 2 Estimación de Cuadrados Mínimos
  - Estimación de Cuadrados Mínimos
  - Cuadrados Mínimos Lineal
- 3 Máxima Verosimilitud
  - Máxima Verosimilitud
- 4 Estimación de Densidad de Probabilidad
  - Estimación de Densidad usando Histograma
  - Estimación de Densidad de Kernel
- 5 Ejercicios Práctico-Teóricos
  - Ejercicio 1
  - Ejercicio 2
  - Ejercicio 3
  - Ejercicio 4
- 6 Bibliografía



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

- Sea  $X$  una v.a. de la cual queremos estimar un parámetro  $\mu$



- Sea  $X$  una v.a. de la cual queremos estimar un parámetro  $\mu$
- Un estimador es una v.a. de la forma  $\hat{X} = g(X)$



- Sea  $X$  una v.a. de la cual queremos estimar un parámetro  $\mu$
- Un estimador es una v.a. de la forma  $\hat{X} = g(X)$
- La función  $g(X)$  define los distintos estimadores de  $X$



- Sea  $X$  una v.a. de la cual queremos estimar un parámetro  $\mu$
- Un estimador es una v.a. de la forma  $\hat{X} = g(X)$
- La función  $g(X)$  define los distintos estimadores de  $X$
- Notamos que al ser  $\hat{X}$  una función de la v.a  $X$  adquiere propiedades estadísticas de  $X$



- Sea  $X$  una v.a. de la cual queremos estimar un parámetro  $\mu$
- Un estimador es una v.a. de la forma  $\hat{X} = g(X)$
- La función  $g(X)$  define los distintos estimadores de  $X$
- Notamos que al ser  $\hat{X}$  una función de la v.a  $X$  adquiere propiedades estadísticas de  $X$
- Ejemplos de estimadores son:



- Sea  $X$  una v.a. de la cual queremos estimar un parámetro  $\mu$
- Un estimador es una v.a. de la forma  $\hat{X} = g(X)$
- La función  $g(X)$  define los distintos estimadores de  $X$
- Notamos que al ser  $\hat{X}$  una función de la v.a  $X$  adquiere propiedades estadísticas de  $X$
- Ejemplos de estimadores son:
- La media muestral  $\bar{X} = \frac{1}{N} \sum_{i=0}^{N-1} x_i$





- Sea  $X$  una v.a. de la cual queremos estimar un parámetro  $\mu$
- Un estimador es una v.a. de la forma  $\hat{X} = g(X)$
- La función  $g(X)$  define los distintos estimadores de  $X$
- Notamos que al ser  $\hat{X}$  una función de la v.a  $X$  adquiere propiedades estadísticas de  $X$
- Ejemplos de estimadores son:
- La media muestral  $\bar{X} = \frac{1}{N} \sum_{i=0}^{N-1} x_i$
- Calcular la esperanza y varianza del estimador  $\bar{X}$



- Definimos las siguientes propiedades de un estimador  $\bar{X}$  del parámetro real  $\mu$ :



# Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador  $\bar{X}$  del parámetro real  $\mu$ :
- Esperanza:  $E[\bar{X}] = E[g(X)]$



# Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador  $\bar{X}$  del parámetro real  $\mu$ :
- Esperanza:  $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias):  $b = E[\bar{X}] - \mu$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador  $\bar{X}$  del parámetro real  $\mu$ :
- Esperanza:  $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias):  $b = E[\bar{X}] - \mu$
- Varianza:  $\text{var}[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador  $\bar{X}$  del parámetro real  $\mu$ :
- Esperanza:  $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias):  $b = E[\bar{X}] - \mu$
- Varianza:  $\text{var}[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$
- Error cuadrático medio (o mean squared error, MSE):  
 $mse = \text{var}[\bar{X}] + b^2$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador  $\bar{X}$  del parámetro real  $\mu$ :
- Esperanza:  $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias):  $b = E[\bar{X}] - \mu$
- Varianza:  $\text{var}[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$
- Error cuadrático medio (o mean squared error, MSE):  
 $mse = \text{var}[\bar{X}] + b^2$
- Estimadores insesgados:  $b = 0$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador  $\bar{X}$  del parámetro real  $\mu$ :
- Esperanza:  $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias):  $b = E[\bar{X}] - \mu$
- Varianza:  $\text{var}[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$
- Error cuadrático medio (o mean squared error, MSE):  
 $mse = \text{var}[\bar{X}] + b^2$
- Estimadores insesgados:  $b = 0$
- Qué es mejor? Poco sesgo o poca varianza?





- Es común querer estimar el valor de una v.a.  $X$  dada una *medición*  $Y$



# Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a.  $X$  dada una *medición*  $Y$
- Por ej.  $Y$  puede ser una versión de  $X$  contaminada por ruido



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a.  $X$  dada una *medición*  $Y$
- Por ej.  $Y$  puede ser una versión de  $X$  contaminada por ruido
- Sea  $\hat{X}$  un estimador de  $X$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a.  $X$  dada una *medición*  $Y$
- Por ej.  $Y$  puede ser una versión de  $X$  contaminada por ruido
- Sea  $\hat{X}$  un estimador de  $X$
- Nos interesa encontrar un estimador tal que el error cuadrático medio sea mínimo (least squares estimation, LSE):



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a.  $X$  dada una *medición*  $Y$
- Por ej.  $Y$  puede ser una versión de  $X$  contaminada por ruido
- Sea  $\hat{X}$  un estimador de  $X$
- Nos interesa encontrar un estimador tal que el error cuadrático medio sea mínimo (least squares estimation, LSE):
- $\min E[(X - \hat{X})^2]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a.  $X$  dada una *medición*  $Y$
- Por ej.  $Y$  puede ser una versión de  $X$  contaminada por ruido
- Sea  $\hat{X}$  un estimador de  $X$
- Nos interesa encontrar un estimador tal que el error cuadrático medio sea mínimo (least squares estimation, LSE):
- $\min E[(X - \hat{X})^2]$
- Demostramos a continuación que el mejor estimador de LSE es  $\hat{X} = E[X]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Llamamos  $\mu = E[X]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Llamamos  $\mu = E[X]$
- Vemos que  $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Estimación de Cuadrados Mínimos

- Llamamos  $\mu = E[X]$
- Vemos que  $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Llamamos  $\mu = E[X]$
- Vemos que  $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$
- Por definición  $E[X - \mu] = 0$ , entonces



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Llamamos  $\mu = E[X]$
- Vemos que  $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[(X - \mu)(\mu - \hat{X})] + E[(\mu - \hat{X})^2]$
- Por definición  $E[X - \mu] = 0$ , entonces
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + E[(\mu - \hat{X})^2]$



# Estimación de Cuadrados Mínimos

- Llamamos  $\mu = E[X]$
- Vemos que  $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$
- Por definición  $E[X - \mu] = 0$ , entonces
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + (\mu - \hat{X})^2$
- Notando que  $E[(X - \mu)^2] = \text{var}[X]$  no depende de  $\hat{X}$ , entonces para tener mínimo  $E[(X - \hat{X})^2]$  debemos minimizar el último término:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos

- Llamamos  $\mu = E[X]$
- Vemos que  $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$
- Por definición  $E[X - \mu] = 0$ , entonces
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + (\mu - \hat{X})^2$
- Notando que  $E[(X - \mu)^2] = \text{var}[X]$  no depende de  $\hat{X}$ , entonces para tener mínimo  $E[(X - \hat{X})^2]$  debemos minimizar el último término:
- $\therefore \hat{X} = \mu \quad \square$



# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$  minimiza  $E[(X - \hat{X})^2|Y = y]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$  minimiza  $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$  es la estimación de cuadrados mínimos de  $X$  dada la observación  $y$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$  minimiza  $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$  es la estimación de cuadrados mínimos de  $X$  dada la observación  $y$
- Para cualquier estimador  $g(Y)$  función de la observación se cumple:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$  minimiza  $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$  es la estimación de cuadrados mínimos de  $X$  dada la observación  $y$
- Para cualquier estimador  $g(Y)$  función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$  minimiza  $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$  es la estimación de cuadrados mínimos de  $X$  dada la observación  $y$
- Para cualquier estimador  $g(Y)$  función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$
- Usando la ley de esperanzas iteradas llegamos a:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$  minimiza  $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$  es la estimación de cuadrados mínimos de  $X$  dada la observación  $y$
- Para cualquier estimador  $g(Y)$  función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$
- Usando la ley de esperanzas iteradas llegamos a:
- $E[(X - E[X|Y])^2] \leq E[(X - g(Y))^2]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$  minimiza  $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$  es la estimación de cuadrados mínimos de  $X$  dada la observación  $y$
- Para cualquier estimador  $g(Y)$  función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$
- Usando la ley de esperanzas iteradas llegamos a:
- $E[(X - E[X|Y])^2] \leq E[(X - g(Y))^2]$
- El resultado se puede extender a  $n$  v.a. condicionales:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos  $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$  minimiza  $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$  es la estimación de cuadrados mínimos de  $X$  dada la observación  $y$
- Para cualquier estimador  $g(Y)$  función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$
- Usando la ley de esperanzas iteradas llegamos a:
- $E[(X - E[X|Y])^2] \leq E[(X - g(Y))^2]$
- El resultado se puede extender a  $n$  v.a. condicionales:
- $E[(X - E[X|Y_1, Y_2, \dots, Y_n])^2] \leq E[(X - g(Y_1, Y_2, \dots, Y_n))^2]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como  $\tilde{X} = X - \hat{X}$ , donde  $\hat{X} = E[X|Y]$  es el estimador óptimo



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como  $\tilde{X} = X - \hat{X}$ , donde  $\hat{X} = E[X|Y]$  es el estimador óptimo
- Notamos que  $E[\tilde{X}] = 0$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como  $\tilde{X} = X - \hat{X}$ , donde  $\hat{X} = E[X|Y]$  es el estimador óptimo
- Notamos que  $E[\tilde{X}] = 0$
- También se cumple que  $E[\tilde{X}|Y = y] = 0$  para todo  $y$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como  $\tilde{X} = X - \hat{X}$ , donde  $\hat{X} = E[X|Y]$  es el estimador óptimo
- Notamos que  $E[\tilde{X}] = 0$
- También se cumple que  $E[\tilde{X}|Y = y] = 0$  para todo  $y$
- Además, el error de estimación es descorrelacionado con la estimación  $\hat{X}$ , es decir  $E[\tilde{X}\hat{X}] = 0$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como  $\tilde{X} = X - \hat{X}$ , donde  $\hat{X} = E[X|Y]$  es el estimador óptimo
- Notamos que  $E[\tilde{X}] = 0$
- También se cumple que  $E[\tilde{X}|Y = y] = 0$  para todo  $y$
- Además, el error de estimación es descorrelacionado con la estimación  $\hat{X}$ , es decir  $E[\tilde{X}\hat{X}] = 0$
- Se cumple la siguiente ley de varianzas:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como  $\tilde{X} = X - \hat{X}$ , donde  $\hat{X} = E[X|Y]$  es el estimador óptimo
- Notamos que  $E[\tilde{X}] = 0$
- También se cumple que  $E[\tilde{X}|Y = y] = 0$  para todo  $y$
- Además, el error de estimación es descorrelacionado con la estimación  $\hat{X}$ , es decir  $E[\tilde{X}\hat{X}] = 0$
- Se cumple la siguiente ley de varianzas:
- $\text{var}[X] = \text{var}[\hat{X}] + \text{var}[\tilde{X}]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

- Calcular  $E[X|Y]$  para todos los  $g(Y)$  es en general complicado



# Cuadrados Mínimos Lineal

- Calcular  $E[X|Y]$  para todos los  $g(Y)$  es en general complicado
- Por simplicidad nos limitamos a estimadores lineales

$$g(Y) = a_1 Y_1 + \dots + a_n Y_n + b$$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Cuadrados Mínimos Lineal

- Calcular  $E[X|Y]$  para todos los  $g(Y)$  es en general complicado
- Por simplicidad nos limitamos a estimadores lineales
$$g(Y) = a_1 Y_1 + \dots + a_n Y_n + b$$
- Para  $n = 1$  tenemos  $g(Y) = aY + b$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Cuadrados Mínimos Lineal

- Calcular  $E[X|Y]$  para todos los  $g(Y)$  es en general complicado
- Por simplicidad nos limitamos a estimadores lineales
$$g(Y) = a_1 Y_1 + \dots + a_n Y_n + b$$
- Para  $n = 1$  tenemos  $g(Y) = aY + b$
- Minimizar:  $E[(X - aY - b)^2]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Cuadrados Mínimos Lineal

- Calcular  $E[X|Y]$  para todos los  $g(Y)$  es en general complicado
- Por simplicidad nos limitamos a estimadores lineales
$$g(Y) = a_1 Y_1 + \dots + a_n Y_n + b$$
- Para  $n = 1$  tenemos  $g(Y) = aY + b$
- Minimizar:  $E[(X - aY - b)^2]$
- Si fijamos  $a$ , es como tener que estimar una v.a.  $X - aY$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Cuadrados Mínimos Lineal

- Calcular  $E[X|Y]$  para todos los  $g(Y)$  es en general complicado
- Por simplicidad nos limitamos a estimadores lineales
$$g(Y) = a_1 Y_1 + \dots + a_n Y_n + b$$
- Para  $n = 1$  tenemos  $g(Y) = aY + b$
- Minimizar:  $E[(X - aY - b)^2]$
- Si fijamos  $a$ , es como tener que estimar una v.a.  $X - aY$
- Entonces  $b = E[X - aY] = E[X] - aE[Y]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Cuadrados Mínimos Lineal

- Calcular  $E[X|Y]$  para todos los  $g(Y)$  es en general complicado
- Por simplicidad nos limitamos a estimadores lineales
$$g(Y) = a_1 Y_1 + \dots + a_n Y_n + b$$
- Para  $n = 1$  tenemos  $g(Y) = aY + b$
- Minimizar:  $E[(X - aY - b)^2]$
- Si fijamos  $a$ , es como tener que estimar una v.a.  $X - aY$
- Entonces  $b = E[X - aY] = E[X] - aE[Y]$
- Reemplazando todo en función de  $a$  queda



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Cuadrados Mínimos Lineal

- Calcular  $E[X|Y]$  para todos los  $g(Y)$  es en general complicado
- Por simplicidad nos limitamos a estimadores lineales
$$g(Y) = a_1 Y_1 + \dots + a_n Y_n + b$$
- Para  $n = 1$  tenemos  $g(Y) = aY + b$
- Minimizar:  $E[(X - aY - b)^2]$
- Si fijamos  $a$ , es como tener que estimar una v.a.  $X - aY$
- Entonces  $b = E[X - aY] = E[X] - aE[Y]$
- Reemplazando todo en función de  $a$  queda
- $E[((X - E[X]) - a(Y - E[Y]))^2] = \sigma_X^2 + a^2\sigma_Y^2 - 2a \operatorname{cov}(X, Y)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

- Minimizando en función de  $a$  queda



# Cuadrados Mínimos Lineal

- Minimizando en función de  $a$  queda
- $a = \rho \frac{\sigma_X}{\sigma_Y}$  con  $\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

- Minimizando en función de  $a$  queda
- $a = \rho \frac{\sigma_X}{\sigma_Y}$  con  $\rho = \frac{\text{COV}(X,Y)}{\sigma_X \sigma_Y}$
- El estimador de cuadrados mínimos lineal de  $X$  basado en  $Y$  es:





# Cuadrados Mínimos Lineal

- Minimizando en función de  $a$  queda
- $a = \rho \frac{\sigma_X}{\sigma_Y}$  con  $\rho = \frac{\text{COV}(X,Y)}{\sigma_X \sigma_Y}$
- El estimador de cuadrados mínimos lineal de  $X$  basado en  $Y$  es:
- $\hat{X} = E[X] + \frac{\text{COV}(X,Y)}{\sigma_Y^2}(Y - E[Y])$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Cuadrados Mínimos Lineal

- Minimizando en función de  $a$  queda
- $a = \rho \frac{\sigma_X}{\sigma_Y}$  con  $\rho = \frac{\text{COV}(X,Y)}{\sigma_X \sigma_Y}$
- El estimador de cuadrados mínimos lineal de  $X$  basado en  $Y$  es:
- $\hat{X} = E[X] + \frac{\text{COV}(X,Y)}{\sigma_Y^2}(Y - E[Y])$
- Es decir necesitamos conocimiento previo de las medias, varianzas y covarianzas de las v.a.



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

- Extendemos el concepto de LSE a una regresión lineal



# Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean  $\mathbf{x} = [x_1, \dots, x_n]^T$ ,  $\mathbf{y} = [y_1, \dots, y_n]^T$  observaciones de dos variables  $X, Y$  en instantes de tiempos distintos  $n$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean  $\mathbf{x} = [x_1, \dots, x_n]^T$ ,  $\mathbf{y} = [y_1, \dots, y_n]^T$  observaciones de dos variables  $X, Y$  en instantes de tiempos distintos  $n$
- Queremos encontrar una relación lineal  $Y = aX + b$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean  $\mathbf{x} = [x_1, \dots, x_n]^T$ ,  $\mathbf{y} = [y_1, \dots, y_n]^T$  observaciones de dos variables  $X, Y$  en instantes de tiempos distintos  $n$
- Queremos encontrar una relación lineal  $Y = aX + b$
- Encontrar  $\beta = [a, b]^T$  tal que  $\|\mathbf{y} - \mathbf{x}\|^2$  sea mínimo



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean  $\mathbf{x} = [x_1, \dots, x_n]^T$ ,  $\mathbf{y} = [y_1, \dots, y_n]^T$  observaciones de dos variables  $X, Y$  en instantes de tiempos distintos  $n$
- Queremos encontrar una relación lineal  $Y = aX + b$
- Encontrar  $\beta = [a, b]^T$  tal que  $\|\mathbf{y} - \mathbf{x}\|^2$  sea mínimo

- Definimos la matriz:  $\mathbf{A} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$



FACULTAD  
DE INGENIERIA  
Universidad de Buenos Aires

# Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean  $\mathbf{x} = [x_1, \dots, x_n]^T$ ,  $\mathbf{y} = [y_1, \dots, y_n]^T$  observaciones de dos variables  $X, Y$  en instantes de tiempos distintos  $n$
- Queremos encontrar una relación lineal  $Y = aX + b$
- Encontrar  $\beta = [a, b]^T$  tal que  $\|\mathbf{y} - \mathbf{x}\|^2$  sea mínimo
- Definimos la matriz:  $\mathbf{A} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$
- Si  $\mathbf{A}$  tiene rango completo, entonces



FACULTAD  
DE INGENIERIA  
Universidad de Buenos Aires



# Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean  $\mathbf{x} = [x_1, \dots, x_n]^T$ ,  $\mathbf{y} = [y_1, \dots, y_n]^T$  observaciones de dos variables  $X, Y$  en instantes de tiempos distintos  $n$
- Queremos encontrar una relación lineal  $Y = aX + b$
- Encontrar  $\beta = [a, b]^T$  tal que  $\|\mathbf{y} - \mathbf{x}\|^2$  sea mínimo

- Definimos la matriz:  $\mathbf{A} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$
- Si  $\mathbf{A}$  tiene rango completo, entonces

- $\beta = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}$



FACULTAD  
DE INGENIERIA  
Universidad de Buenos Aires

- Sean  $\mathcal{X} = \{x_1, \dots, x_N\}$  muestras obtenidas a partir de una realización de las v.a. i.i.d  $X = \{X_1, \dots, X_N\}$



# Máxima Verosimilitud

- Sean  $\mathcal{X} = \{x_1, \dots, x_N\}$  muestras obtenidas a partir de una realización de las v.a. i.i.d  $X = \{X_1, \dots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s)  $\theta$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Máxima Verosimilitud

- Sean  $\mathcal{X} = \{x_1, \dots, x_N\}$  muestras obtenidas a partir de una realización de las v.a. i.i.d  $X = \{X_1, \dots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s) desconocido(s)  $\theta$
- Es decir,  $x_n \sim p(\mathcal{X}|\theta)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Máxima Verosimilitud

- Sean  $\mathcal{X} = \{x_1, \dots, x_N\}$  muestras obtenidas a partir de una realización de las v.a. i.i.d  $X = \{X_1, \dots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s)  $\theta$
- Es decir,  $x_n \sim p(\mathcal{X}|\theta)$
- La idea en Máxima Verosimilitud (MV) es encontrar  $\theta$  tal que la probabilidad de obtener  $\mathcal{X}$  a partir de  $p(\mathcal{X}|\theta)$  sea máxima



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Máxima Verosimilitud

- Sean  $\mathcal{X} = \{x_1, \dots, x_N\}$  muestras obtenidas a partir de una realización de las v.a. i.i.d  $X = \{X_1, \dots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s)  $\theta$
- Es decir,  $x_n \sim p(\mathcal{X}|\theta)$
- La idea en Máxima Verosimilitud (MV) es encontrar  $\theta$  tal que la probabilidad de obtener  $\mathcal{X}$  a partir de  $p(\mathcal{X}|\theta)$  sea máxima
- La verosimilitud del producto, dada la independencia de las v.a. es el producto de las verosimilitudes



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Máxima Verosimilitud

- Sean  $\mathcal{X} = \{x_1, \dots, x_N\}$  muestras obtenidas a partir de una realización de las v.a. i.i.d  $X = \{X_1, \dots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s)  $\theta$
- Es decir,  $x_n \sim p(\mathcal{X}|\theta)$
- La idea en Máxima Verosimilitud (MV) es encontrar  $\theta$  tal que la probabilidad de obtener  $\mathcal{X}$  a partir de  $p(\mathcal{X}|\theta)$  sea máxima
- La verosimilitud del producto, dada la independencia de las v.a. es el producto de las verosimilitudes
- $l(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta) = \prod_{i=1}^N p_{X_i|\Theta}(x_i|\theta)$



- Tomamos el logaritmo de  $l(\theta|\mathcal{X})$  para transformar el producto en sumas





- Tomamos el logaritmo de  $l(\theta|\mathcal{X})$  para transformar el producto en sumas
- Log es una función monótona y no cambia la propiedades de optimalidad



- Tomamos el logaritmo de  $l(\theta|\mathcal{X})$  para transformar el producto en sumas
- Log es una función monótona y no cambia la propiedades de optimalidad
- Entonces la verosimilitud logarítmica (log-likelihood, LL) es:



- Tomamos el logaritmo de  $l(\theta|\mathcal{X})$  para transformar el producto en sumas
- Log es una función monótona y no cambia la propiedades de optimalidad
- Entonces la verosimilitud logarítmica (log-likelihood, LL) es:
- $\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_{i=1}^N \log p_{X_i|\Theta}(x_i|\theta)$



- Tomamos el ejemplo de la distribución Gaussiana



# Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es  $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es  $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para  $\mathcal{X} = \{x_i\}$  con pdf como la anterior, la LL:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es  $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para  $\mathcal{X} = \{x_i\}$  con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$





# Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es  $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para  $\mathcal{X} = \{x_i\}$  con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$
- Para encontrar el estimador de máxima verosimilitud debemos igualar el gradiente de  $\mathcal{L}$  a cero



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es  $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para  $\mathcal{X} = \{x_i\}$  con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$
- Para encontrar el estimador de máxima verosimilitud debemos igualar el gradiente de  $\mathcal{L}$  a cero
- $\nabla \mathcal{L} = [\frac{\partial \mathcal{L}}{\partial \mu} \frac{\partial \mathcal{L}}{\partial \sigma}] = [0 \ 0]$  lo cual da:



# Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es  $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para  $\mathcal{X} = \{x_i\}$  con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$
- Para encontrar el estimador de máxima verosimilitud debemos igualar el gradiente de  $\mathcal{L}$  a cero
- $\nabla \mathcal{L} = [\frac{\partial \mathcal{L}}{\partial \mu} \frac{\partial \mathcal{L}}{\partial \sigma}] = [0 \ 0]$  lo cual da:
- $\mu = \frac{\sum_i x_i}{N}$ , estimador de máxima verosimilitud de la media (insesgado)



# Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es  $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para  $\mathcal{X} = \{x_i\}$  con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$
- Para encontrar el estimador de máxima verosimilitud debemos igualar el gradiente de  $\mathcal{L}$  a cero
- $\nabla \mathcal{L} = [\frac{\partial \mathcal{L}}{\partial \mu} \frac{\partial \mathcal{L}}{\partial \sigma}] = [0 \ 0]$  lo cual da:
- $\mu = \frac{\sum_i x_i}{N}$ , estimador de máxima verosimilitud de la media (insesgado)
- $s^2 = \frac{\sum_i (x_i - \mu)^2}{N}$ , estimador de máxima verosimilitud de la varianza (sesgado)



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

- El estimador de varianza anterior se suele denominar  $s_N^2$



- El estimador de varianza anterior se suele denominar  $s_N^2$
- Es el estimador óptimo de máxima verosimilitud para la distribución Gaussiana



- El estimador de varianza anterior se suele denominar  $s_N^2$
- Es el estimador óptimo de máxima verosimilitud para la distribución Gaussiana
- Es asintóticamente insesgado, es decir el sesgo tiende a cero cuando  $N \rightarrow \infty$



- El estimador de varianza anterior se suele denominar  $s_N^2$
- Es el estimador óptimo de máxima verosimilitud para la distribución Gaussiana
- Es asintóticamente insesgado, es decir el sesgo tiende a cero cuando  $N \rightarrow \infty$
- En el infinito (N grande ), el estimador  $s_N^2$  y el estimador insesgado  $s_{N-1}^2 = \frac{\sum_i (x_i - \mu)^2}{N-1}$  coinciden





# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos  $n$  observaciones  $x_i \in [0, 1]$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos  $n$  observaciones  $x_i \in [0, 1]$
- Un histograma particiona el conjunto  $[0, 1]$  en  $M$  porciones (bins)



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos  $n$  observaciones  $x_i \in [0, 1]$
- Un histograma particiona el conjunto  $[0, 1]$  en  $M$  porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos  $n$  observaciones  $x_i \in [0, 1]$
- Un histograma particiona el conjunto  $[0, 1]$  en  $M$  porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es  $B_1 = [0, \frac{1}{M}]$ ,  $B_2 = [\frac{1}{M}, \frac{2}{M})$ ,  $B_M = [\frac{M-1}{M}, 1]$



# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos  $n$  observaciones  $x_i \in [0, 1]$
- Un histograma particiona el conjunto  $[0, 1]$  en  $M$  porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es  $B_1 = [0, \frac{1}{M}]$ ,  $B_2 = [\frac{1}{M}, \frac{2}{M})$ ,  $B_M = [\frac{M-1}{M}, 1]$
- Para cada  $x \in B_l$ , la estimación de densidad es:



# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos  $n$  observaciones  $x_i \in [0, 1]$
- Un histograma particiona el conjunto  $[0, 1]$  en  $M$  porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es  $B_1 = [0, \frac{1}{M}]$ ,  $B_2 = [\frac{1}{M}, \frac{2}{M})$ ,  $B_M = [\frac{M-1}{M}, 1]$
- Para cada  $x \in B_l$ , la estimación de densidad es:
- $\hat{p}_n(x) = \frac{\text{no. observaciones en } B_l}{n} \times \frac{1}{\text{longitud bin}}$





# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos  $n$  observaciones  $x_i \in [0, 1]$
- Un histograma particiona el conjunto  $[0, 1]$  en  $M$  porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es  $B_1 = [0, \frac{1}{M}]$ ,  $B_2 = [\frac{1}{M}, \frac{2}{M})$ ,  $B_M = [\frac{M-1}{M}, 1]$
- Para cada  $x \in B_l$ , la estimación de densidad es:
- $\hat{p}_n(x) = \frac{\text{no. observaciones en } B_l}{n} \times \frac{1}{\text{longitud bin}}$
- $\hat{p}_n(x) = \frac{M}{n} \sum_{i=1}^n I(X_i \in B_l)$



# Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad  $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos  $n$  observaciones  $x_i \in [0, 1]$
- Un histograma particiona el conjunto  $[0, 1]$  en  $M$  porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es  $B_1 = [0, \frac{1}{M}]$ ,  $B_2 = [\frac{1}{M}, \frac{2}{M})$ ,  $B_M = [\frac{M-1}{M}, 1]$
- Para cada  $x \in B_l$ , la estimación de densidad es:
- $\hat{p}_n(x) = \frac{\text{no. observaciones en } B_l}{n} \times \frac{1}{\text{longitud bin}}$
- $\hat{p}_n(x) = \frac{M}{n} \sum_{i=1}^n I(X_i \in B_l)$
- $I(X_i \in B_l)$  es la función indicador, definida por  $I(X_i \in B_l) = 1$  si  $X_i \in B_l$  y  $I(X_i \in B_l) = 0$  si no



# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$



# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$



# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins  $M$



# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins  $M$
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins  $M$
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad
- Encontramos ahora la varianza del estimador



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins  $M$
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad
- Encontramos ahora la varianza del estimador
- $\text{var}(\hat{p}_n(x)) = M \frac{p(x^*)}{n} + \frac{p^2(x^*)}{n}$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins  $M$
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad
- Encontramos ahora la varianza del estimador
- $\text{var}(\hat{p}_n(x)) = M \frac{p(x^*)}{n} + \frac{p^2(x^*)}{n}$
- La varianza aumenta con el número de bins y decrece con el número de observaciones



# Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*), x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins  $M$
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad
- Encontramos ahora la varianza del estimador
- $\text{var}(\hat{p}_n(x)) = M \frac{p(x^*)}{n} + \frac{p^2(x^*)}{n}$
- La varianza aumenta con el número de bins y decrece con el número de observaciones
- Existe un  $M$  óptimo que minimiza el  $MSE = \text{var} + b^2$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra  $x_i$  y lo transforma en una "montaña" (bump)



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra  $x_i$  y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra  $x_i$  y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad
- Sea el estimador de densidad de kernel:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra  $x_i$  y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad
- Sea el estimador de densidad de kernel:
- $\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra  $x_i$  y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad
- Sea el estimador de densidad de kernel:
- $\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)$
- $K(\cdot)$  es la función de kernel, gralmente. una función suave y simétrica como la Gaussiana



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra  $x_i$  y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad
- Sea el estimador de densidad de kernel:
- $\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)$
- $K(\cdot)$  es la función de kernel, gralmente. una función suave y simétrica como la Gaussiana
- $h > 0$  es el ancho de banda de filtrado



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)



**FACULTAD  
DE INGENIERÍA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel  $K(\cdot)$  debemos considerar que:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel  $K(\cdot)$  debemos considerar que:
- Sea simétrica



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel  $K(\cdot)$  debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$ , es decir que sea pdf



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel  $K(\cdot)$  debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$ , es decir que sea pdf
- $\lim_{t \rightarrow -\infty} K(x) = \lim_{t \rightarrow +\infty} K(x) = 0$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel  $K(\cdot)$  debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$ , es decir que sea pdf
- $\lim_{t \rightarrow -\infty} K(x) = \lim_{t \rightarrow +\infty} K(x) = 0$
- Funciones kernel comunes:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel  $K(\cdot)$  debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$ , es decir que sea pdf
- $\lim_{t \rightarrow -\infty} K(x) = \lim_{t \rightarrow +\infty} K(x) = 0$
- Funciones kernel comunes:
- Gaussiana:  $K(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-x^2}{2}$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel  $K(\cdot)$  debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$ , es decir que sea pdf
- $\lim_{t \rightarrow -\infty} K(x) = \lim_{t \rightarrow +\infty} K(x) = 0$
- Funciones kernel comunes:
- Gaussiana:  $K(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-x^2}{2}$
- Uniforme:  $K(x) = I(-1 \leq x \leq 1)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Cuando el ancho de banda  $h$  es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando  $h$  es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel  $K(\cdot)$  debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$ , es decir que sea pdf
- $\lim_{t \rightarrow -\infty} K(x) = \lim_{t \rightarrow +\infty} K(x) = 0$
- Funciones kernel comunes:
- Gaussiana:  $K(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-x^2}{2}$
- Uniforme:  $K(x) = I(-1 \leq x \leq 1)$
- Epanechnikov (mínimo MSE):  
 $K(x) = \frac{3}{4} \max\{1 - x^2, 0\}$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

- Como dijimos antes, queremos estimar la pdf  $p_X(x)$



# Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf  $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado,  $x_0$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf  $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado,  $x_0$
- El sesgo está dado por:



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf  $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado,  $x_0$
- El sesgo está dado por:
- $b(\hat{p}_n(x_0)) = E[\hat{p}_n(x_0)] - p(x_0) = E\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) - p(x_0)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf  $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado,  $x_0$
- El sesgo está dado por:
$$b(\hat{p}_n(x_0)) = E[\hat{p}_n(x_0)] - p(x_0) = E\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) - p(x_0)$$
- Obviando la demostración tenemos:
$$b(\hat{p}_n(x_0)) = \frac{1}{2} h^2 p''(x_0) \int y^2 H(y) dy + \mathcal{O}(h^3)$$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf  $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado,  $x_0$
- El sesgo está dado por:
  - $b(\hat{p}_n(x_0)) = E[\hat{p}_n(x_0)] - p(x_0) = E\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) - p(x_0)$
  - Obviando la demostración tenemos:
$$b(\hat{p}_n(x_0)) = \frac{1}{2} h^2 p''(x_0) \int y^2 H(y) dy + \mathcal{O}(h^3)$$
- Esto significa que cuando  $h \rightarrow 0$  el bias decrece como  $\mathcal{O}(h^2)$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf  $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado,  $x_0$
- El sesgo está dado por:
  - $b(\hat{p}_n(x_0)) = E[\hat{p}_n(x_0)] - p(x_0) = E\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) - p(x_0)$
  - Obviando la demostración tenemos:
$$b(\hat{p}_n(x_0)) = \frac{1}{2} h^2 p''(x_0) \int y^2 H(y) dy + \mathcal{O}(h^3)$$
- Esto significa que cuando  $h \rightarrow 0$  el bias decrece como  $\mathcal{O}(h^2)$
- El término  $p''(x_0)$  indica que la curvatura de la pdf (desconocida) incrementa el sesgo porque es estimador de kernel suaviza las curvas



- Analizamos ahora la varianza de la estimación de kernel



# Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel

- $$\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y) dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y) dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$
- Es decir la varianza se reduce a una tasa  $\mathcal{O}\left(\frac{1}{nh}\right)$  cuando  $n \rightarrow \infty, h \rightarrow 0$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y) dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$
- Es decir la varianza se reduce a una tasa  $\mathcal{O}\left(\frac{1}{nh}\right)$  cuando  $n \rightarrow \infty, h \rightarrow 0$
- Por último, estudiamos el MSE del estimador:



# Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y) dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$
- Es decir la varianza se reduce a una tasa  $\mathcal{O}\left(\frac{1}{nh}\right)$  cuando  $n \rightarrow \infty, h \rightarrow 0$
- Por último, estudiamos el MSE del estimador:
- $\text{MSE}(\hat{p}_n(x_0)) = b^2 + \text{var} = \mathcal{O}(h^4) + \mathcal{O}\left(\frac{1}{nh}\right)$





# Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y) dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$
- Es decir la varianza se reduce a una tasa  $\mathcal{O}\left(\frac{1}{nh}\right)$  cuando  $n \rightarrow \infty, h \rightarrow 0$
- Por último, estudiamos el MSE del estimador:
- $\text{MSE}(\hat{p}_n(x_0)) = b^2 + \text{var} = \mathcal{O}(h^4) + \mathcal{O}\left(\frac{1}{nh}\right)$
- Podemos elegir  $h$  tal que se minimice el MSE asintóticamente, esto se logra con  $h \propto n^{-1/5}$ , lo cual logra un  $\text{MSE} = \mathcal{O}(n^{-4/5})$



# Estimación de Densidad de Kernel

- La estimación de densidad de kernel es más rápida que el estimador óptimo del método con histograma ( $\mathcal{O}(n^{-2/3})$ ) pero más lenta que máxima verosimilitud ( $\mathcal{O}(n^{-1})$ )



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- La estimación de densidad de kernel es más rápida que el estimador óptimo del método con histograma ( $\mathcal{O}(n^{-2/3})$ ) pero más lenta que máxima verosimilitud ( $\mathcal{O}(n^{-1})$ )
- Sin embargo en estimación de kernel no asumimos ninguna distribución como en máxima verosimilitud



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- La estimación de densidad de kernel es más rápida que el estimador óptimo del método con histograma ( $\mathcal{O}(n^{-2/3})$ ) pero más lenta que máxima verosimilitud ( $\mathcal{O}(n^{-1})$ )
- Sin embargo en estimación de kernel no asumimos ninguna distribución como en máxima verosimilitud
- Solamente asumimos que la distribución original es suave y diferenciable



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Estimación de Densidad de Kernel

- La estimación de densidad de kernel es más rápida que el estimador óptimo del método con histograma ( $\mathcal{O}(n^{-2/3})$ ) pero más lenta que máxima verosimilitud ( $\mathcal{O}(n^{-1})$ )
- Sin embargo en estimación de kernel no asumimos ninguna distribución como en máxima verosimilitud
- Solamente asumimos que la distribución original es suave y diferenciable
- Este es el precio a pagar por una incremento en la flexibilidad de la estimación (i.e., menos hipótesis)



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Ejercicio 1

- Ⓐ Sea  $X$  una v.a. con media  $\mu$  y varianza  $v$
- Ⓑ Sean  $Y_1, \dots, Y_n$  mediciones de la forma  $Y_i = X + W_i$
- Ⓒ Sean  $W_i$  v.a. con media cero y varianza  $v_i$
- Ⓓ Asumamos que  $X, W_1, \dots, W_n$  son independientes
- Ⓔ Demostrar que el estimador lineal de cuadrados mínimos de  $X$  en base a las mediciones  $Y_1, \dots, Y_n$  es:
- Ⓕ 
$$\hat{X} = \frac{(\mu/v) + \sum_{i=1}^n (Y_i/v_i)}{(1/v) + \sum_{i=1}^n (1/v_i)}$$
- Ⓖ Simular para  $n = 10, n = 1000, v = 0.1, v = 100$
- Ⓗ Qué conclusión se puede sacar para valores de  $n$  y/o  $v$  grandes?



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Ejercicio 2

- a) Se tira una moneda 100 veces, y salen 55 cecas.
- b) Encontrar el estimador de máxima verosimilitud de la probabilidad de ceca  $p$
- c) Simular el experimento y encontrar por computadora el valor de la estimación de máxima verosimilitud de  $p$



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

# Ejercicio 3

- a) Una v.a. continua  $X$  responde a un proceso Gaussiano de media cero y  $\sigma^2 = 1$
- b) Simular varias realizaciones de  $X$  y estimar la pdf usando el método del histograma
- c) Variar la cantidad de bins y sacar conclusiones acerca de la calidad de estimación



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires



# Ejercicio 4

- a) Repetir el ejercicio 3 usando un estimador de kernel con una función Gaussiana
- b) Comparar resultados entre ambos ejercicios



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

 E. Alpaydin, *Introduction to machine learning*.

MIT press, 2020.


 “MIT - Lecture Notes Course 6.041-6.431, Fall 2000.”

[https://vfu.bg/en/e-Learning/Math--Bertsekas\\_Tsitsiklis\\_Introduction\\_to\\_probability.pdf](https://vfu.bg/en/e-Learning/Math--Bertsekas_Tsitsiklis_Introduction_to_probability.pdf).

Accessed: 2020-05-15.

 S. M. Kay, *Fundamentals of statistical signal processing*.

Prentice Hall PTR, 1993.

 “Universidad de Washington, Lecture 6: Density Estimation: Histogram and Kernel Density Estimator.” [http://faculty.washington.edu/yenchic/18W\\_425/Lec6\\_hist\\_KDE.pdf](http://faculty.washington.edu/yenchic/18W_425/Lec6_hist_KDE.pdf).

Accessed: 2020-05-15.



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires