Introducción a la Inferencia Estadística

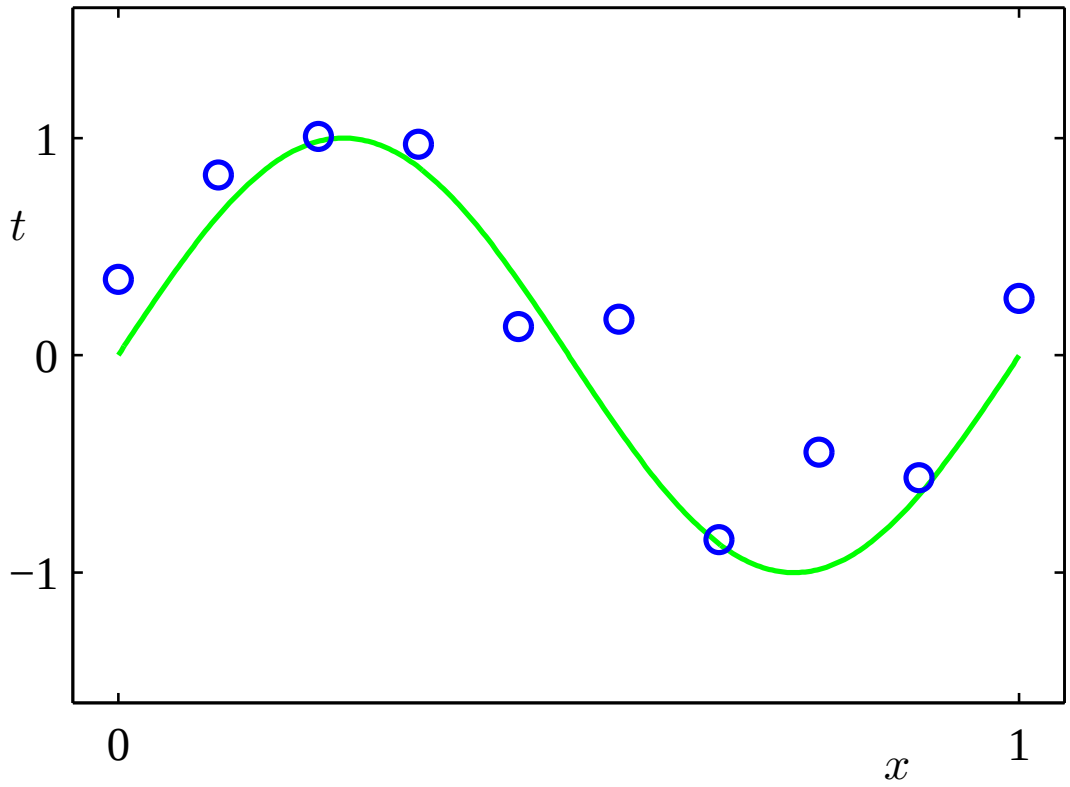
Mauricio A. Álvarez

Modelos probabilísticos profundos
AIR Institute

Contenido

Contenido

Ejemplo Regresión



- **Regresión**: supongamos una función conocida $\text{sen}(2\pi x)$ con ruido aleatorio incluido en la variable objetivo **t**.
- Datos disponibles: $\mathbf{x} \equiv \{x_1, \ldots, x_N\}^\top$, $\mathbf{t} \equiv \{t_1, \ldots, t_N\}^\top$.

Modelo de regresión lineal (I)

- Supongamos t dado como

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon,$$

donde $\epsilon \sim \mathcal{N}(0, \beta^{-1})$.

- Para la función $y(\mathbf{x}, \mathbf{w})$, el modelo lineal asume que

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{M-1} w_i \phi_i(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}),$$

donde $\phi_i(\mathbf{x})$ son funciones base, M es el número de parámetros del modelo, y w_0 es el desplazamiento.

- Igualmente, $\mathbf{w} = [w_0 \cdots w_{M-1}]^\top$, $\boldsymbol{\phi}(\mathbf{x}) = [\phi_0(\mathbf{x}) \cdots \phi_{M-1}(\mathbf{x})]^\top$.

Modelo de regresión lineal (II)

- Usando funciones base polinomiales, $\phi_j(x) = x^j$, se tiene

$$y(x, \mathbf{w}) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j = \mathbf{w}^\top \phi(x),$$

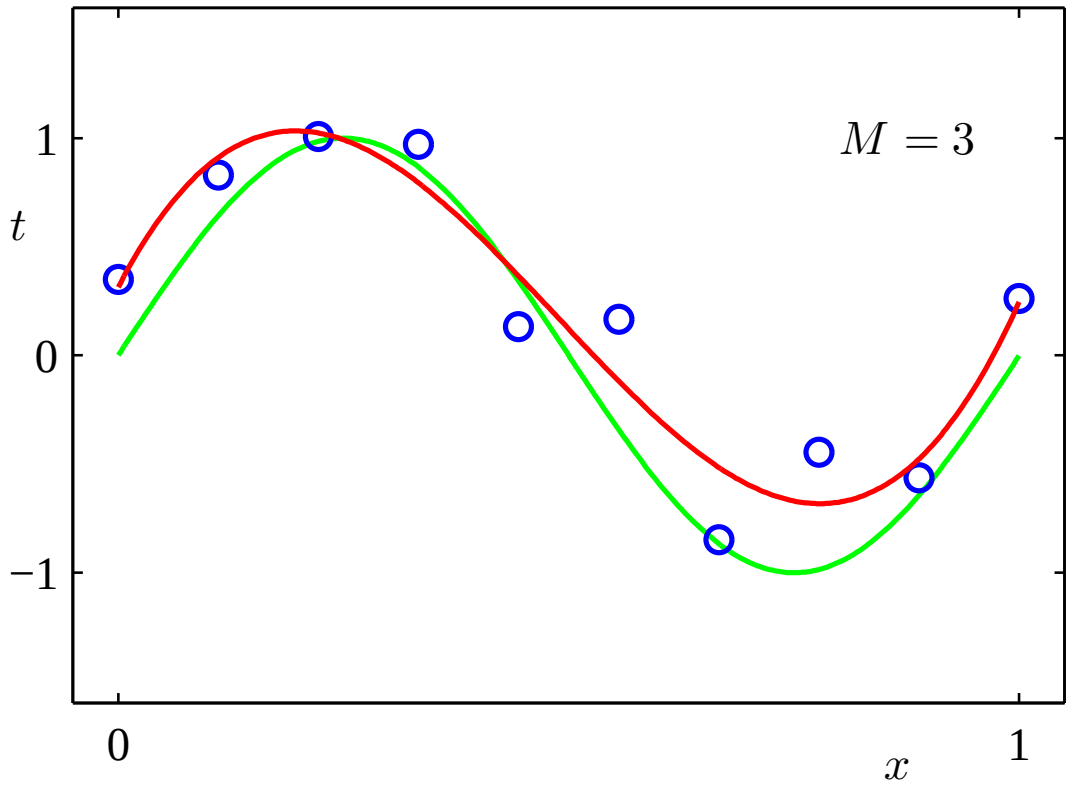
donde $\phi(x) = [1 \; x \; x^2 \cdots x^M]^\top$.

- El problema de estimación consiste en encontrar el valor de los parámetros $\mathbf{w}$, β y M que describen mejor un conjunto de datos $\mathbf{x}$, y $\mathbf{t}$.

- Usualmente, al todo el conjunto de parámetros se le denota por una sola letra, por ejemplo, $\boldsymbol{\theta}$.

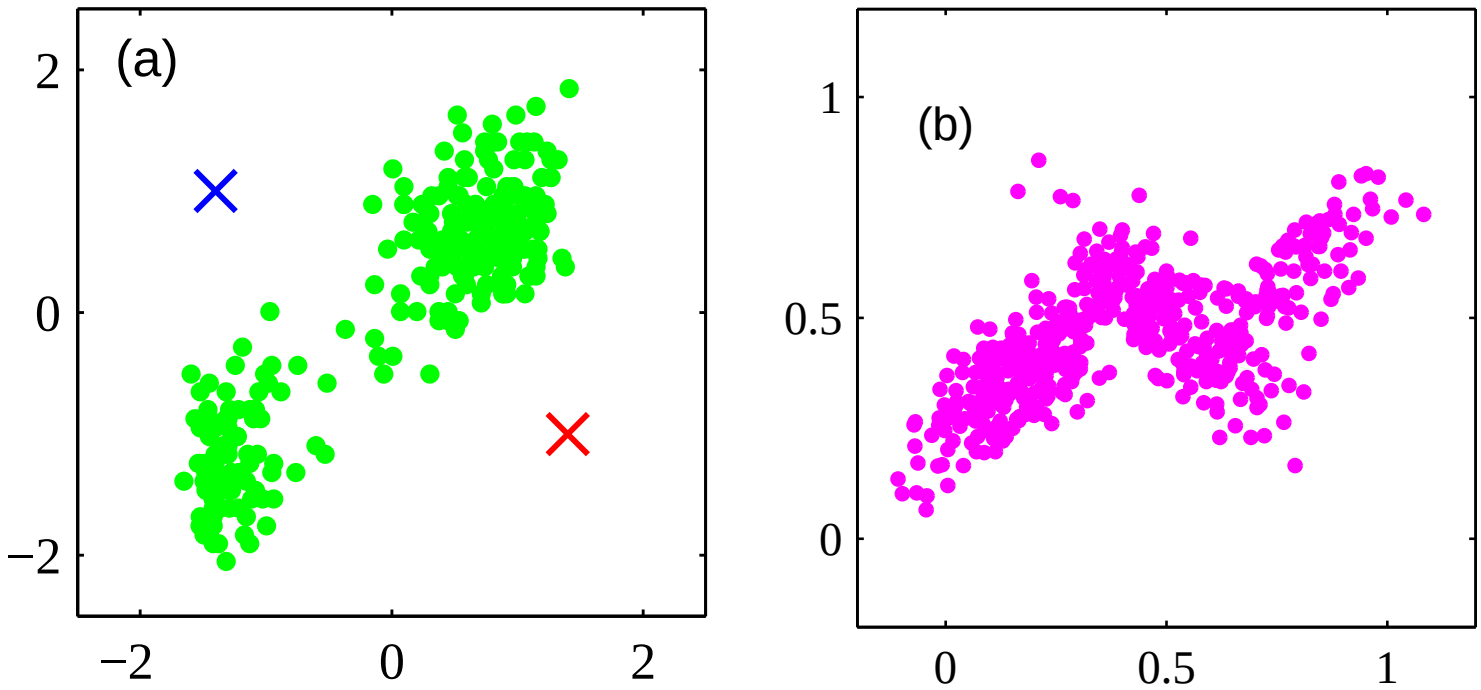
- Para el caso anterior $\boldsymbol{\theta} = \{\mathbf{w}, \beta, M\}$.

Después de resolver el problema de estimación



Contenido

Ejemplo agrupamiento



- **Agrupamiento**: encontrar de forma automática grupos similares presentes en un conjunto de datos.
- Datos disponibles: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, cada uno de dos variables $\mathbf{x}_i = [x_{i,1} \;\; x_{i,2}]^\top$.

Agrupamiento probabilístico

- Una forma de aproximar funciones de probabilidad multimodales es a través de una mezcla de funciones de probabilidad.

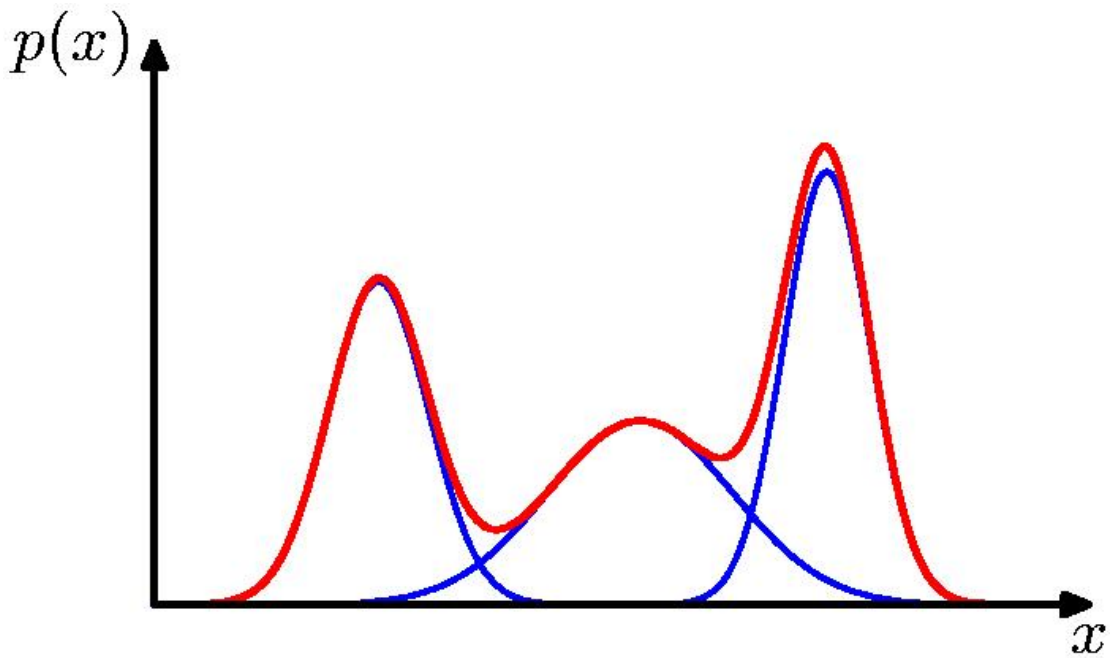
- De las mezclas de funciones de probabilidad, la mezcla de Gaussianas es una de las más conocidas,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

donde K es el número de componentes de la mezcla, y los parámetros π_k son probabilidades que satisfacen

$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1.$$

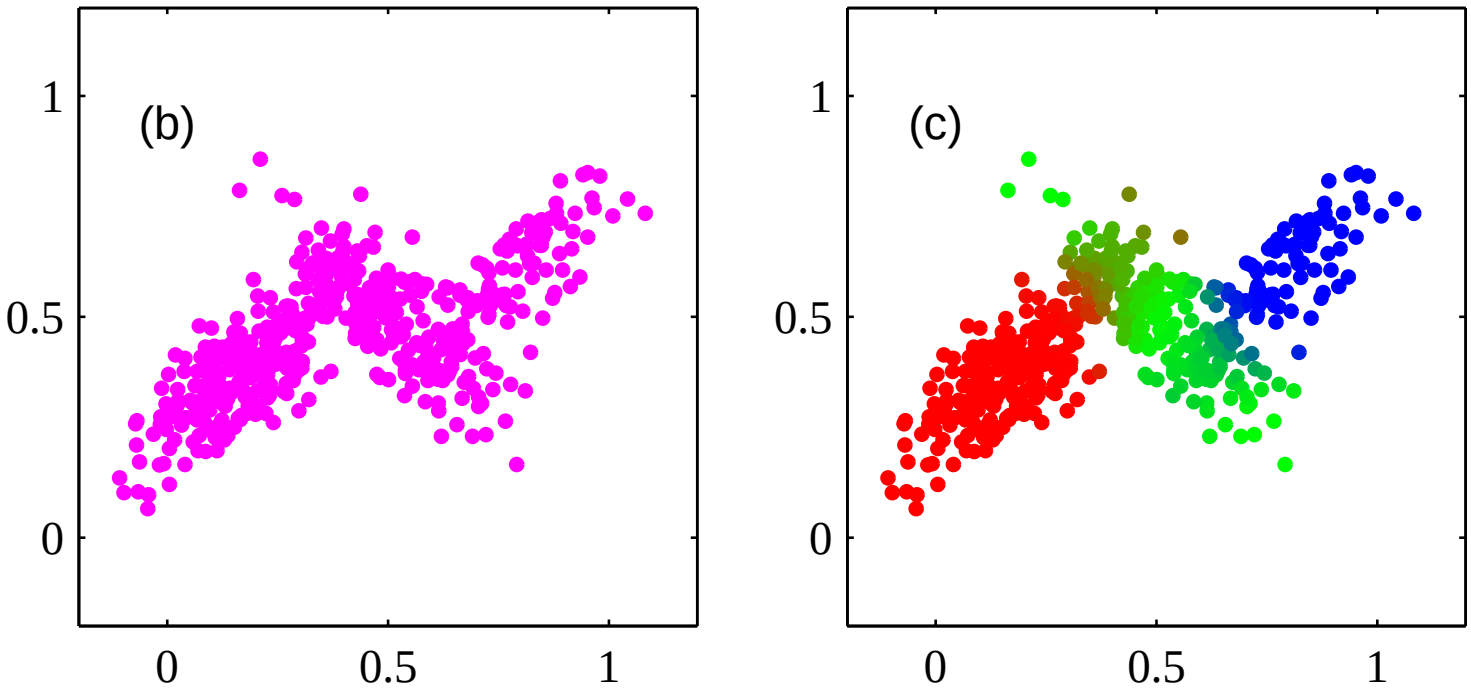
Ejemplo en una dimensión



Agrupamiento probabilístico

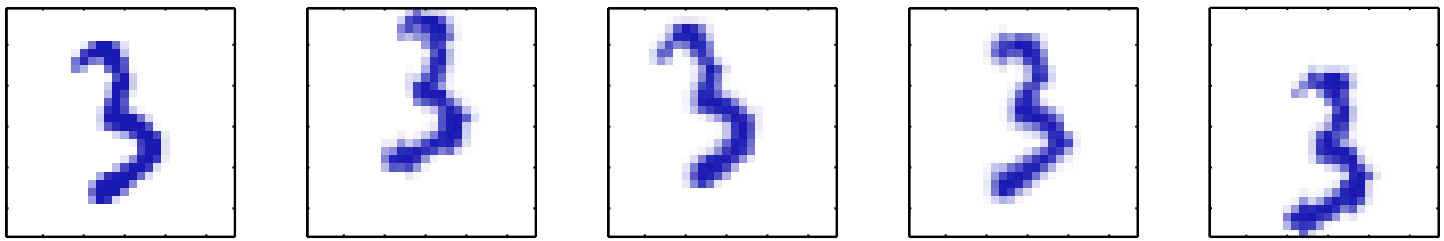
- El problema de estimación consiste en encontrar el valor de los parámetros $\theta = \{K, \{\pi_k\}_{k=1}^K, \{\boldsymbol{\mu}_k\}_{k=1}^K, \{\boldsymbol{\Sigma}_k\}_{k=1}^K\}$ a partir de $\mathbf{X}$.

Después de resolver el problema de estimación



Contenido

Ejemplo reducción de dimensionalidad (I)



- Cada imagen vive en un espacio de $100 \times 100 = 10,000$ dimensiones.
- Cada imagen se puede representar como un vector $\mathbf{x} \in \mathbb{R}^{10,000}$.
- Cada imagen se generó a partir de traslaciones verticales, traslaciones horizontales y rotaciones (3 grados de libertad).

Ejemplo reducción de dimensionalidad (II)

- **Reducción de dimensión**: encontrar de forma automática una representación de baja dimensionalidad $\mathbf{z} \in \mathbb{R}^M$, con $M \ll 10,000$.

- Datos disponibles: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, cada uno de $10,000$ variables $\mathbf{x}_i = [x_{i,1} \ \cdots x_{i,10,000}]^\top$.

Modelo de reducción de dimensión

- El modelo lineal más conocido es el Análisis de Componentes Principales (PCA por su nombre en inglés).

- En su versión probabilística (PPCA) cada datos observado, $\mathbf{x} \in \mathbb{R}^D$ se representa mediante

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

donde $\mathbf{W}^{D \times M}$, $\boldsymbol{\mu}$ es la media de los datos y

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}),$$
$$p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \sigma^2\mathbf{I}).$$

- El problema de estimación consiste en encontrar el valor de los parámetros $\boldsymbol{\theta} = \{M, \mathbf{W}, \boldsymbol{\mu}, \sigma^2\}$.

Contenido

Tipos de estimación: clásica

- La estimación se suele dividir en *estimación clásica, puntual o frecuentista* y *estimación Bayesiana*.

- Estimación clásica
  - perspectiva sin condicional: los métodos de inferencia deben dar buenas respuestas cuando se repite su uso.
  - natural cuando se necesita escribir software que sea útil para mucha gente que usa diferentes bases de datos.
  - los parámetros de interés se asumen como *deterministas*, pero desconocidos.

Tipos de estimación: Bayesiana

- En estimación Bayesiana,
  - perspectiva condicional: la inferencia se debe hacer condicionado a los datos actuales.
  - natural en proyectos de largo plazo que involucran un experto en determinado dominio.
  - se asume que los parámetros son la *realización de una variable aleatoria (ó vector aleatorio)* con alguna función de distribución, y se usa el teorema de Bayes para actualizar el conocimiento que tenemos sobre θ, a partir de una muestra X, para finalmente obtener G.

Más detalles ...

- Capítulo 1 del libro “Statistical decision theory and Bayesian analysis” by James Berger.

- El videolecture de Michael I Jordan “Bayesian or frequentist: which one are you?”
  `http://videolectures.net/mlss09uk_jordan_bfway/`

Contenido

Contenido

Introducción

- En esta sección se habla de las características de un “buen” estimador en el sentido clásico.
- En este contexto, se estudian los estimadores que en *promedio* conducen al valor verdadero del parámetro.
- Dentro de este grupo, el objetivo es encontrar los estimadores que exhiban la *menor variabilidad*.
- Si el estimador tiene estas características, producirá valores cercanos al parámetro verdadero la mayoría de las veces.

Notación

- En adelante el vector de parámetros estimados (obtenidos de la muestra) se denota como $\widehat{\boldsymbol{\theta}}$ y los parámetros verdaderos (obtenidos de la población) se denotan como $\boldsymbol{\theta}$.

- El estimador es una función $g(\cdot)$ tal que

$$\widehat{\boldsymbol{\theta}} = g(\mathcal{D}),$$

donde $\mathcal{D}$ hace referencia a la muestra disponible.

Estimadores insesgados

- Para que un estimador sea *insesgado*, en promedio, el estimador debe conducir al valor verdadero del parámetro desconocido.

- Debido a que el valor del parámetro θ en forma general, se encuentra en el intervalo $a < \theta < b$, la condición de insesgado afirma que no importa cuál sea el valor real de θ, el estimador siempre lo obtendrá en promedio.

- Matemáticamente, un estimador es insesgado (imparcial o centrado), si

$$E(\widehat{\theta}) = \theta, \quad a < \theta < b,$$

donde (a, b) denota el rango de valores posibles de θ.

Ejemplo: constante con ruido Gaussiano

- Considérese las observaciones obtenidas como

$$x_n = A + w_n, \quad n = 1, \ldots, N$$

donde A es el parámetro que necesita estimarse y $w_n, \forall\, n$ es ruido blanco Gaussiano con varianza σ^2.

- Sea $\widehat{A}$ un estimador de A, obtenido como

$$\widehat{A} = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

- Tomando el valor esperado a ambos lados se tiene

$$E[\widehat{A}] = E\left[\frac{1}{N} \sum_{n=1}^{N} x_n\right] = \frac{1}{N} \sum_{n=1}^{N} E[x_n] = \frac{1}{N} \sum_{n=1}^{N} A = A.$$

Estimadores de varianza mínima (I)

- Para encontrar un estimador óptimo es necesario definir algún criterio de optimalidad.
- Supongamos que el criterio de optimalidad es el *error cuadrático medio*

$$\text{mse}(\widehat{\theta}) = E\left[(\widehat{\theta} - \theta)^2\right].$$

- Este criterio mide la desviación media cuadrática promedio entre el estimador y el valor real.
- La adopción de este criterio conduce a estimadores que no son realizables, es decir, estimadores que no se pueden expresar únicamente como función de los datos.

Estimadores de varianza mínima (II)

- El criterio MSE se puede reescribir como

$$\begin{aligned}\text{mse}(\widehat{\theta}) &= E\left\{\left[\left(\widehat{\theta} - E(\widehat{\theta})\right) + \left(E(\widehat{\theta}) - \theta\right)\right]^2\right\} \\ &= \text{var}(\widehat{\theta}) + \left[E(\widehat{\theta}) - \theta\right]^2 \\ &= \text{var}(\widehat{\theta}) + b^2(\theta),\end{aligned}$$

  Lo cual demuestra que el MSE está compuesto de errores debidos a la varianza del estimador así como a su sesgo.

- Supongamos el ejemplo anterior y el siguiente estimador

$$\overline{A} = a\frac{1}{N}\sum_{n=1}^{N} x_n,$$

  para alguna constante a.

Estimadores de varianza mínima (III)

- El objetivo es encontrar el valor de a que minimiza el MSE.
- Como $E[\overline{A}] = aA$ y $\text{var}(\overline{A}) = a^2\sigma^2/N$, se tiene

$$\begin{aligned} \text{mse}(\overline{A}) &= \text{var}(\overline{A}) + b^2(\overline{A}), \\ &= \frac{a^2\sigma^2}{N} + (a-1)^2A^2. \end{aligned}$$

- Diferenciando con respecto a a e igualando a cero se tiene

$$a_{\text{opt}} = \frac{A^2}{A^2 + \sigma^2/N}.$$

- El valor óptimo de a depende del parámetro desconocido A, luego el estimador no es realizable.

Estimadores de varianza mínima (IV)

- Lo anterior se debe a que el sesgo es una función de A.
- Desde un punto de vista práctico el estimador que minimiza el MSE no puede adoptarse.
- Un enfoque alternativo consiste en restringir el sesgo a que sea igual a cero y encontrar el estimador que minimiza la varianza.
- Tal estimador se conoce como *estimador insesgado de mínima varianza* (MVU).

Encontrando el estimador MVU

- No siempre es posible encontrar el estimador MVU para todos los valores de un parámetro.

- Algunos métodos para encontrarlo (si existiese) son los siguientes
  - Se determina el límite inferior de Cramer-Rao (CRLB) y se verifica si algún estimador lo satisface.
  - Se aplica el teorema Rao-Blackwell-Lehmann-Scheffe.
  - Se restringen los posibles estimadores, no sólo a que sean insesgados, si no también lineales. Luego el MVU se busca entre esta clase restringida de estimadores.

Extensión a un vector de parámetros

- Se define

$$E(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} E(\widehat{\theta}_1) \\ E(\widehat{\theta}_2) \\ \vdots \\ E(\widehat{\theta}_p) \end{bmatrix}.$$

- Se define el estimador insesgado como $E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ para cada $\boldsymbol{\theta}$.

- Un estimador MVU tiene la propiedad adicional que $\mathrm{var}(\widehat{\theta}_i)$, para $i = 1, \ldots, p$, es mínima entre todos los posibles estimadores insesgados.

Contenido

Máxima verosimilitud

- El MLE para un parámetro escalar se define como el valor de θ que maximiza la función de verosimilitud $p(\mathbf{x}; \theta)$ para $\mathbf{x}$ fijo.

- La estimación por *máxima verosimilitud* (maximum likelihood estimation - MLE) es uno de los métodos de estimación clásica más empleados en la práctica.

- Es aproximadamente igual al estimador MVU debido a su eficiencia aproximada.

Contenido

Otros tipos de estimación clásica

- Método de mínimos cuadrados.
- Método de los momentos.

Contenido

Introducción

- En la estimación Bayesiana, el parámetro de interés θ se considera una variable aleatoria con fdp $p(\theta)$.

- Se desea encontrar la fdp a posteriori del parámetro θ una vez se tienen los datos $\mathcal{D}$.

- Lo anterior se puede lograr empleando el *teorema de Bayes*

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}.$$

donde

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta.$$

Formas de estimar la fdp a posteriori

- Para muchos modelos probabilísticos, suele ser difícil estimar $p(\theta|\mathcal{D})$.

- Algunos métodos que se suelen emplear para encontrar la fdp a posteriori incluyen
  1. Maximum a posteriori (MAP).
  2. Aproximación por Laplace.
  3. Bayes variacional, KL($q \parallel p$).
  4. Expectation - Propagation (EP), KL($p \parallel q$).
  5. Monte Carlo.
  6. Markov Chain Monte Carlo.