

El Método de Mínimos Cuadrados

Sergio A. Cruces Álvarez

UNIVERSIDAD DE SEVILLA

1

En esta segunda prueba se pretenden esbozar algunos de los puntos más importantes del método de Mínimos Cuadrados junto con ejemplos de su aplicación práctica.

Este tema se circunscribe en uno de los capítulos iniciales de un curso de doctorado sobre Teoría de la Estimación Paramétrica y su duración real equivaldría a unas 5 horas lectivas.

Se presuponen en el alumno unos conocimientos básicos de álgebra, cálculo vectorial y teoría de probabilidad.

Puntos a Tratar

1. Introducción.
2. Mínimos cuadrados determinista.
3. Mínimos cuadrados estocástico.
4. Conclusiones.
5. Referencias.

2

Los puntos que vamos a tratar en la exposición son:

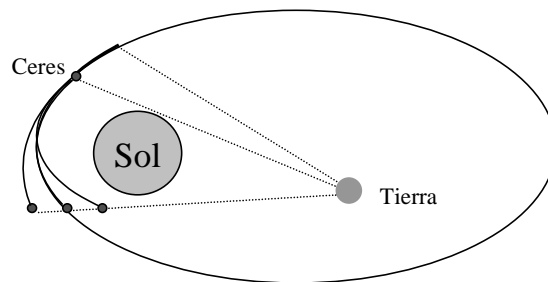
1. La introducción al método de Mínimos Cuadrados trata de motivar al alumno presentando la particular historia que propició su descubrimiento por parte de Gauss.
2. La versión determinista del método que responde a un problema de aproximación. Una gran parte de la exposición se dedica a este punto pues la simplicidad en la formulación del método determinista de mínimos cuadrados lo hace ideal para presentar de una forma coherente y concisa muchos conceptos y sus interrelaciones.
3. El método de Mínimos Cuadrados Estocástico responde a un problema de estimación. Como este método posee una complejidad ligeramente superior al anterior. Aquí no se presentan conceptos nuevos sólo examinan unos pocos aunque, eso sí, con una mayor profundidad.
4. Se repasan las principales conclusiones
5. Se comentan algunas referencias útiles para estudiar y profundizar en el tema.

Introducción



Carl Friedrich Gauss
[1777- 1855]

El método de Mínimos Cuadrados tiene una larga historia que se remonta a principios del siglo XIX.



3

El método de mínimos cuadrados tiene una larga historia que se remonta a los principios del siglo XIX. En Junio de 1801, Zach, un astrónomo que Gauss había conocido dos años antes, publicaba las posiciones orbitales del cuerpo celeste Ceres, un nuevo “pequeño planeta” descubierto por el astrónomo italiano G. Piazzi en ese mismo año. Desafortunadamente, Piazzi sólo había podido observar 9 grados de su órbita antes de que este cuerpo desapareciese tras de el sol. Zach publicó varias predicciones de su posición incluyendo una de Gauss que difería notablemente de las demás. Cuando Ceres fue redescubierto por Zach en Diciembre de 1801 estaba casi exactamente en donde Gauss había predicho.

Aunque todavía no había revelado su método, Gauss había descubierto el método de mínimos cuadrados. En un trabajo brillante logró calcular la órbita de Ceres a partir de un número reducido de observaciones, de hecho, el método de Gauss requiere sólo un mínimo de 3 observaciones y todavía es, en esencia, el utilizado en la actualidad para calcular las órbitas.

Puntos a Tratar

1. Introducción.
- ⇒ 2. Mínimos cuadrados determinista.
 - 2.1. Interpretación geométrica.
 - 2.2. Aumentando el orden de la aproximación.
 - 2.3. Aumentando el número de observaciones.
 - 2.4. Mínimos cuadrados con restricciones.
 - 2.5. Dependencia no lineal.
3. Mínimos cuadrados estocástico.
 - 3.1. Teorema de Gauss-Markov.
 - 3.2. Interpretación Geométrica.
 - 3.3. Teorema de Aitken.
4. Conclusiones.
5. Referencias.

4

En este punto se describe la versión determinista del método de Mínimos Cuadrados. El orden de presentación es el siguiente.

En primer lugar, se presentan “los personajes de la película”, es decir, el modelo de generación de los datos junto con las señales y variables implicadas.

A continuación, se describe el criterio propuesto, su solución para el caso de dependencia lineal con los parámetros y , una interpretación geométrica de ésta que sirve para establecer el fundamental principio de ortogonalidad.

Las soluciones recursivas del método se describen en base al concepto de innovación (nueva información) para los casos de aumentar el número de parámetros o el número de datos.

Se obtiene una solución compacta para una situación en la que se consideren restricciones lineales y , finalmente, se trata el caso de que exista una dependencia no lineal de la señal con los parámetros que se desean estimar.

Mínimos Cuadrados Determinista

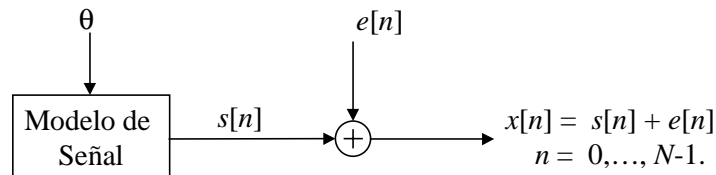


Fig.1: Modelo de las observaciones

		<u>Notación Vectorial.</u>
θ	Vector con los p parámetros de interés.	$\theta = [\theta_1, \dots, \theta_p]^T$
$s[n]$	Señal determinista dependiente de θ .	$\mathbf{s}(\theta) = [s[0], \dots, s[N-1]]^T$
$e[n]$	Ruido y desajustes del modelo.	$\mathbf{e}(\theta) = [e[0], \dots, e[N-1]]^T$
$x[n]$	Observaciones.	$\mathbf{x} = [x[0], \dots, x[N-1]]^T$

5

En el método de Mínimos Cuadrados deseamos minimizar la discrepancia entre los datos observados $x[n]$ y la señal original $s[n]$. Esta señal se genera a través de un modelo que depende un conjunto de parámetros de interés agrupados en el vector θ . Aunque $s[n]$ es completamente determinista la presencia de inexactitudes en el modelo o ruido en los sensores hace que las observemos una versión perturbada de ésta que denotamos por $x[n]$.

A lo largo del tema preferiremos utilizar la notación vectorial por su mayor simplicidad y claridad a la hora de permitir visualizar los resultados.

Mínimos Cuadrados Determinista

✧ Objetivo:

Buscar la *mejor aproximación* $\hat{\theta}$
del sistema sobredeterminado $\mathbf{x} \cong \mathbf{s}(\theta)$



Minimizar $J(\theta) = \|\mathbf{e}(\theta)\|^2$

Soluc. MCD $\hat{\theta} = \arg(\min_{\theta} J(\theta))$

***Método de aproximación:** no utiliza hipótesis probabilísticas sobre los datos, sólo su modelo de generación (Fig.1).

✧ Dependencia lineal

$$\mathbf{s}(\theta) = \mathbf{H}\theta$$

$\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_p]$ matriz de observación ($N \times p$)

6

El objetivo del método de Mínimos Cuadrados Determinista (MCD) es el de elegir el parámetro θ que mejor aproxima la señal original $\mathbf{s}[n]$ a los datos observados $\mathbf{x}[n]$.

El criterio de proximidad que se aplica en este caso es el que resulta de considerar una función de coste o discrepancia $J(\theta)$ que se forma con la norma al cuadrado del error. Así, el estimador de mínimos cuadrados es aquel que minimiza esta función de coste.

Es importante destacar que MCD es un método de aproximación pues no utiliza hipótesis probabilísticas sobre los datos, sólo su modelo de generación.

El caso más sencillo se da cuando asumimos que la dependencia de la señal con los parámetros es lineal. La matriz por la que se multiplican los parámetros para generar la señal recibe el nombre de matriz de observación y sus columnas jugarán un importante papel en este método.

Dependencia Lineal

✧ Solución (si $\text{rango}(\mathbf{H}^T \mathbf{H}) = p$)

$$\left. \begin{aligned} \frac{\partial J(\theta)}{\partial \theta^T} &= -2(\mathbf{x} - \mathbf{H}\theta)^T \mathbf{H} = \mathbf{0} \\ \frac{\partial}{\partial \theta^T} \left(\frac{\partial J(\theta)}{\partial \theta^T} \right)^T &= 2\mathbf{H}^T \mathbf{H} > 0 \text{ (def+)} \end{aligned} \right\} J(\theta) \text{ estrict. convexa.}$$

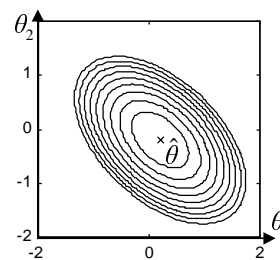
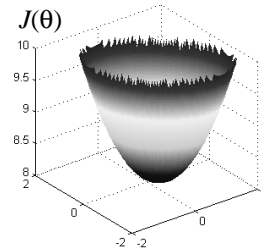
Único mínimo global

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

¿Señal estimada a partir de la mejor aproximación?

$$\Rightarrow \hat{\mathbf{s}} = \mathbf{H}\hat{\theta} = \mathbf{H}\mathbf{H}^+ \mathbf{x}$$

donde $\mathbf{H}^+ = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$ Pseudoinversa de \mathbf{H}



7

Encontrar los parámetros que dan la solución al problema, cuando las columnas de la matriz de observación son linealmente independientes, es muy simple. Basta con calcular la derivada y la matriz Hessiana para observar que ésta última es siempre estrictamente definida positiva, lo cual, garantiza que la función de coste $J(\theta)$ es estrictamente convexa y, por tanto, existe un único mínimo. Este se obtiene simplemente buscando el único punto crítico de la función.

Un razonamiento cualitativo conduce al mismo resultado. Basta observar que $J(\theta)$ es una función positiva con una dependencia cuadrática sobre los parámetros, por tanto, su grafo traza un hiper-paraboloide p -dimensional que tiene un único mínimo y cuyas curvas de nivel son, en el caso más general, hiper-elipsoides.

Una vez obtenidos los parámetros que dan la mejor aproximación, ¿cómo podemos obtener nuestra mejor estima para la señal?... Sólo hace falta sustituir éstos en el modelo de señal para encontrar la respuesta.

Dependencia Lineal

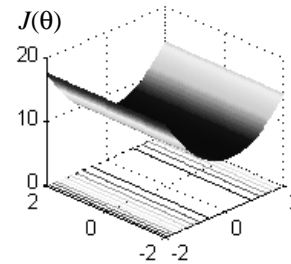
✧ Solución (si $\text{rango}(\mathbf{H}^T \mathbf{H}) < p$) $\Rightarrow \frac{\partial}{\partial \theta^T} \left(\frac{\partial J(\theta)}{\partial \theta^T} \right)^T = 2\mathbf{H}^T \mathbf{H} \geq 0$ (semidef+)

$J(\theta)$ convexa (no estrictamente)
Subespacio de mínimos globales.

Inf. Soluciones $\forall \mathbf{q} \in \mathfrak{R}^n$

$$\hat{\theta} = \mathbf{H}^+ \mathbf{x} + (\mathbf{I} - \mathbf{H}^+ \mathbf{H}) \mathbf{q}$$

Soluc. particular Soluc. homogénea



θ no es identificable, aunque la señal estimada a partir de la mejor aproximación sigue siendo única

$$\hat{\mathbf{s}} = \mathbf{H} \hat{\theta} = \mathbf{H} \mathbf{H}^+ \mathbf{x}$$

8

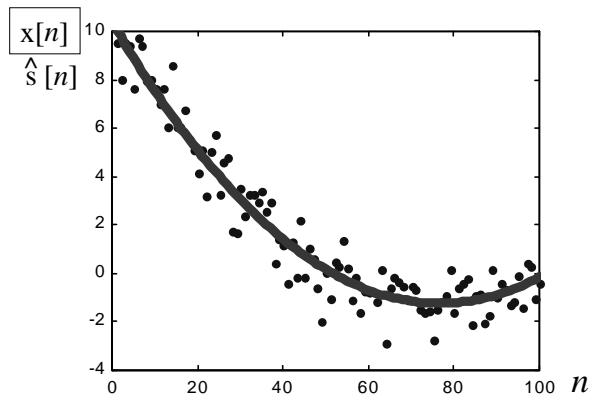
Cuando las columnas de la matriz de observación son linealmente dependientes la solución no es tan sencilla.

El Hessiano es semidefinido positivo, lo cual implica que $J(\theta)$ sigue siendo una función convexa aunque no de forma estricta. Todos los puntos críticos de la función son, por tanto, mínimos. La expresión analítica de estos mínimos se determina igualando la derivada de la función a cero y resolviendo el sistema indeterminado resultante.

En este caso los parámetros originales no son identificables puesto que la solución de MCD no es única. Lo que sí es identificable es la estima de la señal que proporciona el método, pues ésta resulta ser común para todos los parámetros solución.

Ejemplo

Señal cuadrática en ruido: $x[n] = \theta_1 + \theta_2 n + \theta_3 n^2 + e[n]$



$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 2^2 \\ \vdots & \vdots & \vdots \\ 1 & N & N^2 \end{pmatrix}$$

$$\theta_* = \begin{pmatrix} 10 \\ -0.3 \\ 0.002 \end{pmatrix}$$

$$\hat{\theta} = \begin{pmatrix} 10.076 \\ -0.299 \\ 0.0020 \end{pmatrix}$$

$$\hat{s} = \mathbf{H}\hat{\theta}$$

9

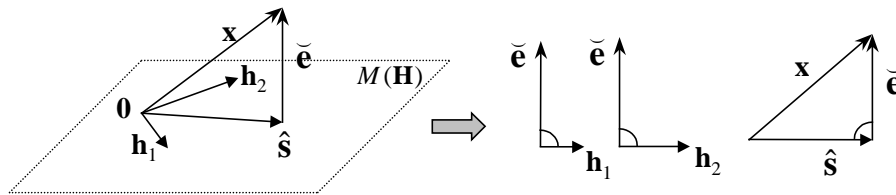
En este ejemplo se obtienen los tres parámetros que mejor ajustan una señal cuadrática a las observaciones y se visualiza gráficamente la señal que proporciona el mejor ajuste según el método de mínimos cuadrados.

Interpretación Geométrica

◇ Principio de Ortogonalidad:

$$\langle \tilde{\mathbf{e}}, \mathbf{h}_i \rangle = 0 \quad i=1, \dots, p.$$

El error mínimo $\tilde{\mathbf{e}}$ es ortogonal al subespacio $M(\mathbf{H}) = \{\mathbf{y} : \mathbf{y} = \mathbf{H}\mathbf{z} \quad \forall \mathbf{z} \in \mathfrak{R}^p\}$ generado por las columnas de \mathbf{H} .



$\hat{\mathbf{s}}$ es la proyección ortogonal de \mathbf{x} sobre $M(\mathbf{H})$

10

La interpretación geométrica de los resultados nos lleva a un principio fundamental: el “principio de ortogonalidad” o “teorema de la proyección”, que es válido para cualquier espacio vectorial de Hilbert y que enunciamos a continuación:

“Sea $M(\mathbf{H})$ el subespacio generado por las columnas de \mathbf{H} , la norma del error entre un vector \mathbf{x} arbitrario y cualquier punto de este subespacio es mínima, si y sólo si, el error es ortogonal al subespacio. Además, el punto del subespacio que minimiza la norma del error (la mejor aproximación de la señal) es único.”

El principio de ortogonalidad está íntimamente ligado a la geometría de los espacios de Hilbert. Una de las consecuencias de este principio es que el teorema de Pitágoras se debe cumplir en los espacios de Hilbert.

La utilidad del principio de ortogonalidad tiene dos vertientes: por una parte nos permite obtener la solución del problema de una forma muy elegante sin recurrir a calcular gradientes y derivadas, por otra, nos proporciona una interpretación geométrica de los resultados que facilita nuestra comprensión del problema y capacidad de intuición.

*Interpretación Geométrica

Demostración del principio de ortogonalidad:

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta^T} \Big|_{\theta=\hat{\theta}} = -2\tilde{\mathbf{e}}^T \mathbf{H} = \mathbf{0}^T &\Rightarrow [\mathbf{h}_1, \dots, \mathbf{h}_N]^T \tilde{\mathbf{e}} = [0, \dots, 0] \\ &\Rightarrow \langle \tilde{\mathbf{e}}, \mathbf{h}_i \rangle = 0 \quad i=1, \dots, p.\end{aligned}$$

Demostración del corolario al principio de ortogonalidad:

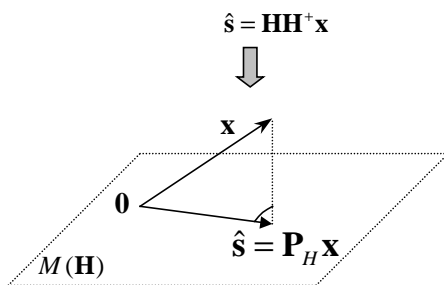
$$\hat{\mathbf{s}} = \mathbf{H}\hat{\theta} \in M(\mathbf{H}) \quad \Rightarrow \langle \tilde{\mathbf{e}}, \hat{\mathbf{s}} \rangle = 0.$$

11

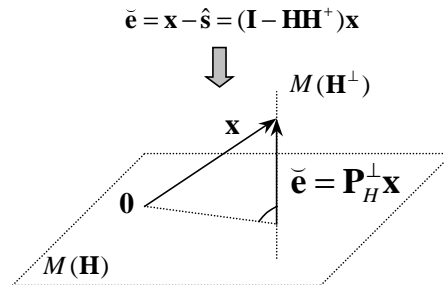
Una vez definido el espacio vectorial de Hilbert y el producto interno entre dos vectores, la demostración del principio de ortogonalidad se obtiene directamente de la ecuación que nos facilita los puntos críticos.

Matrices de Proyección \perp

◇ ¿Cómo es el operador que proyecta \mathbf{x} sobre $M(\mathbf{H})$? ... ¿y sobre $M(\mathbf{H}^\perp)$?



Se define $\mathbf{P}_H = \mathbf{H}\mathbf{H}^+$ como la matriz que proyecta ortogonalmente un vector sobre $M(\mathbf{H})$.



Se define $\mathbf{P}_H^\perp = \mathbf{I} - \mathbf{P}_H$ como la matriz que proyecta sobre el complemento ortogonal de $M(\mathbf{H})$.

12

Llegado este punto nos podemos preguntar ¿cómo es el operador que proyecta ortogonalmente las observaciones \mathbf{x} sobre el subespacio generado por las columnas de \mathbf{H} ?

Como está claro que la señal estimada es el resultado de esta proyección ortogonal, la respuesta a esta pregunta se obtiene directamente identificando este operador en la expresión que nos da la señal estimada como mejor aproximación. Así, comprobamos que el operador buscado es un operador lineal denominado matriz de proyección ortogonal.

Otra posible cuestión podría ser ¿cómo encontrar el operador que proyecta las observaciones sobre el complemento ortogonal del subespacio generado por las columnas de \mathbf{H} ?

Por la estructura del problema, las componente de las observaciones que no se puede expresar sobre $M(\mathbf{H})$ es el error mínimo. Observando la expresión del error en términos de \mathbf{x} resulta fácil identificar el operador de proyección sobre el complemento ortogonal al subespacio.

Matrices de Proyección \perp

Propiedades de las matrices de proyección ortogonal:

1. Son simétricas $\mathbf{P}_H = \mathbf{P}_H^T$
2. Son idempotentes $\mathbf{P}_H \mathbf{P}_H = \mathbf{P}_H$
3. Sus autovalores $\lambda_i = \begin{cases} 0 & \text{si el autovector asociado } \mathbf{q}_i \notin M(\mathbf{H}) \\ 1 & \text{si el autovector asociado } \mathbf{q}_i \in M(\mathbf{H}) \end{cases}$
4. Son únicas para cada subespacio $M(\mathbf{H}) \equiv M(\mathbf{H}') \Leftrightarrow \mathbf{P}_H = \mathbf{P}_{H'}$
5. Sobre subespacios ortogonales *la proyección se desacopla*

Si $\mathbf{H} = (\mathbf{H}_1 | \mathbf{H}_2)$ donde $\mathbf{H}_1 \perp \mathbf{H}_2 \Rightarrow \mathbf{P}_H = \mathbf{P}_{H_1} + \mathbf{P}_{H_2}$

$$\Rightarrow \hat{\mathbf{s}} = \mathbf{P}_{H_1} \mathbf{x} + \mathbf{P}_{H_2} \mathbf{x}$$

13

Las propiedades de las matrices de proyección ortogonal nos permiten encontrar soluciones simples a problemas, en principio, complicados.

Estas propiedades se obtienen directamente de la descomposición en valores singulares de \mathbf{H} y de la definición de matriz de proyección ortogonal.

De entre todas ellas las propiedades 4 y 5 son las más interesantes:

4. Las matrices de proyección ortogonal son únicas para cada subespacio, es decir, que columnas de \mathbf{H} que generen el mismo subespacio tienen la misma matriz de proyección.

5. Si somos capaces de expresar el espacio generado por las columnas de \mathbf{H} sobre dos subespacios ortogonales el problema de la proyección se desacopla de forma independiente sobre cada uno de ellos, de forma que, la proyección conjunta resulta ser la suma de las proyecciones marginales.

Aumentando el Orden

Sea $\hat{\theta}_{k-1} = (\mathbf{H}_{k-1}^T \mathbf{H}_{k-1})^{-1} \mathbf{H}_{k-1}^T \mathbf{x}$ el estimador que nos da la mejor aproximación con $k-1$ parámetros. Cuando añadimos un nuevo parámetro tendremos

$$\hat{\theta}_k = (\mathbf{H}_k^T \mathbf{H}_k)^{-1} \mathbf{H}_k^T \mathbf{x} \quad \text{donde} \quad \mathbf{H}_k = [\mathbf{H}_{k-1}, \mathbf{h}_k]$$

- ◇ ¿Es necesario recalcular la solución cada vez?
- ◇ Existe un método más eficiente?

Existe una *solución recursiva* que aprovecha $\hat{\theta}_{k-1}$ para calcular $\hat{\theta}_k$.

$$\hat{\theta}_k = \begin{bmatrix} \hat{\theta}_{k-1} - \mathbf{H}_{k-1}^+ \mathbf{h}_k \beta \\ \beta \end{bmatrix} \quad \text{con} \quad \beta = \frac{\mathbf{h}_k^T \mathbf{P}_{k-1}^\perp \mathbf{x}}{\mathbf{h}_k^T \mathbf{P}_{k-1}^\perp \mathbf{h}_k}$$

14

En muchas situaciones el modelo de las observaciones nos es desconocido. En esos casos uno puede plantearse el añadir más parámetros para mejorar la estima. Para ello, bastaría volver a resolver el problema tras ampliar la matriz de observación con una nueva columna por cada parámetro añadido a la estima.

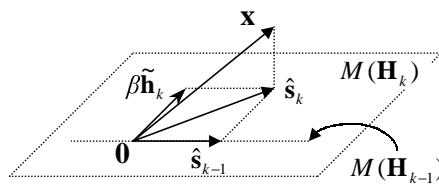
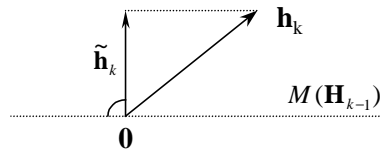
Pero..., ¿existe un método mejor que recalcular la solución cada vez?

La respuesta a esta pregunta es sí. Es posible encontrar una solución recursiva que aprovecha la estima hecha considerando $k-1$ parámetros para obtener la solución que tiene en cuenta un parámetro más. Esta solución recursiva conlleva, a medida que se incrementa el número de parámetros, una reducción en la carga computacional total, a la vez que evita, mediante la utilización del lema de inversión, la inversión explícita de matrices (operación que resulta muy costosa en cuanto a carga computacional).

Aumentando el Orden

Deducción geométrica:

La *innovación* es la componente de \mathbf{h}_k ortogonal a $M(\mathbf{H}_{k-1})$.



¿Por qué es preferible trabajar con la innovación?

$$\left. \begin{aligned} \hat{\mathbf{s}}_k &= \hat{\mathbf{s}}_{k-1} + \beta \tilde{\mathbf{h}}_k \\ \hat{\mathbf{s}}_k &= \mathbf{H}_k \hat{\boldsymbol{\theta}}_k \end{aligned} \right\} \begin{array}{l} \text{De ambas ec. se despeja la} \\ \text{soluc. recursiva para } \hat{\boldsymbol{\theta}}_k \end{array}$$

15

Mediante manipulaciones algebraicas de la solución de orden k es posible obtener directamente la solución recursiva. No obstante, su deducción geométrica resulta mucho más ilustrativa e intuitiva.

La clave de todo está en descomponer la nueva columna de la matriz de observación en dos componentes. Una que contiene información redundante (que ya existía en la aproximación basada en $k-1$ parámetros) y otra componente basada en la nueva información que aparece al añadir un parámetro más al modelo. Esta última componente recibe el nombre de innovación y es, por definición, ortogonal al subespacio generado por el resto de columnas.

Dado que $M([\mathbf{H}_{k-1}, \text{innovación}]) \equiv M(\mathbf{H}_k)$ y \mathbf{H}_{k-1} es ortogonal a la innovación, la proyección total se desacopla sobre estos dos subespacios de forma que la solución es simplemente la suma de las proyecciones de los datos sobre cada uno de ellos. De esta forma, re-escribiendo las matrices de proyección en términos de las variables conocidas se llega a la solución recursiva buscada.

*Aumentando el Orden

Deducción detallada:

$$\tilde{\mathbf{h}}_k = \mathbf{P}_{H_{k-1}}^\perp \mathbf{h}_k$$

$$\hat{\mathbf{h}}_k = \mathbf{P}_{H_{k-1}} \mathbf{h}_k = \mathbf{H}_{k-1} \alpha$$

como $M([\mathbf{H}_{k-1}, \mathbf{h}_k]) \equiv M([\mathbf{H}_{k-1}, \tilde{\mathbf{h}}_k])$

$$\hat{\mathbf{s}}_k = \mathbf{P}_{[H_{k-1}, \tilde{h}_k]} \mathbf{x} = \mathbf{P}_{H_{k-1}} \mathbf{x} + \mathbf{P}_{\tilde{h}_k} \mathbf{x} \rightarrow \frac{\tilde{\mathbf{h}}_k}{\|\tilde{\mathbf{h}}_k\|} \left\langle \mathbf{x}, \frac{\tilde{\mathbf{h}}_k}{\|\tilde{\mathbf{h}}_k\|} \right\rangle = \tilde{\mathbf{h}}_k \beta$$

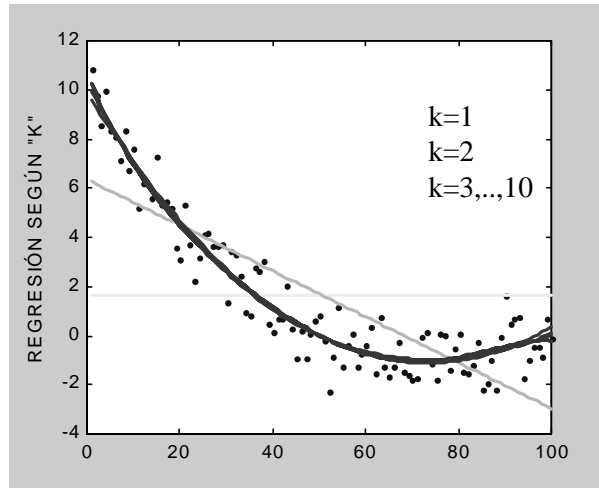
$$\hat{\mathbf{s}}_k = \mathbf{H}_{k-1} \hat{\theta}_{k-1} + (\mathbf{h}_k - \mathbf{H}_{k-1} \alpha) \beta$$

$$= \underbrace{(\mathbf{H}_{k-1}, \mathbf{h}_k)}_{\mathbf{H}_k} \cdot \underbrace{\begin{pmatrix} \hat{\theta}_{k-1} - \alpha \beta \\ \beta \end{pmatrix}}_{\hat{\theta}_k}$$

16

Esta transparencia facilita un análisis más pormenorizado de los pasos que llevan a la solución.

Ejemplo



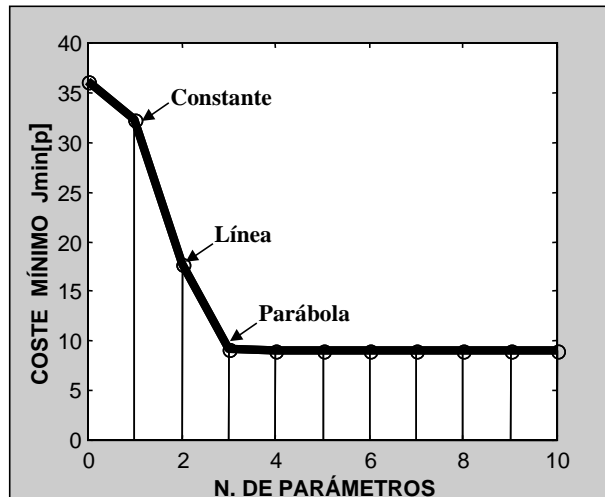
Aumentamos poco a poco el número de parámetros hasta 10.

17

Como ejemplo de este método hemos obtenido de forma recursiva la solución al problema de regresión polinómica de una señal inmersa en ruido. Para ello, utilizamos 100 observaciones y consideramos desde uno hasta un total de diez coeficientes (parámetros), pasando por todos los casos intermedios.

Las observaciones se denotan por puntos azules en la figura, mientras que la componente de señal estimada se representa en trazo continuo con colores que varían según el número de parámetros considerados.

Ejemplo (continuación)



El coste mínimo indica que basta con utilizar 3 parámetros.

18

El coste mínimo nos facilita una idea del número mínimo de parámetros necesarios para representar las observaciones.

En nuestro caso, comprobamos que basta con utilizar 3 parámetros (nuestra señal es aproximadamente cuadrática). Aumentando más el número de parámetros ya no disminuye significativamente el error de representación, más aun, las pequeñas disminuciones en coste se pueden deber a que utilizemos los nuevos grados de libertad que nos proporcionan los parámetros intentar ajustar o modelar el ruido. Esto último no siempre es deseable pues, en la práctica, el modelo más simple que describe de forma adecuada los datos suele resultar preferible.

Aumentando Observaciones

Sea $\hat{\theta}[n-1]$ la estimación a partir de los datos $x[0], \dots, x[n-1]$. Si añadimos un dato más ($x[n]$), el nuevo estimador será

$$\hat{\theta}[n] = (\mathbf{H}^T[n]\mathbf{H}[n])^{-1}\mathbf{H}^T[n]\mathbf{x}_n \quad \text{con} \quad \mathbf{x}_n = \begin{pmatrix} \mathbf{x}_{n-1} \\ x[n] \end{pmatrix} \quad \mathbf{H}[n] = \begin{pmatrix} \mathbf{H}[n-1] \\ \mathbf{h}^T[n] \end{pmatrix}$$

◇ ¿Es necesario recalcular la solución cada vez?
¿Existe un método más eficiente?

Existe una solución secuencial que ofrece una carga computacional de $O(np^2)$ operaciones frente a las $O(n^3)$ de la solución directa.

$$\hat{\theta}[n] = \hat{\theta}[n-1] + K[n] \cdot (x[n] - \mathbf{h}^T[n]\hat{\theta}[n-1])$$

$$\text{donde} \quad K[n] = \frac{(\mathbf{H}^T[n-1]\mathbf{H}[n-1])^{-1}\mathbf{h}[n]}{1 + \mathbf{h}^T[n](\mathbf{H}^T[n-1]\mathbf{H}[n-1])^{-1}\mathbf{h}[n]}$$

19

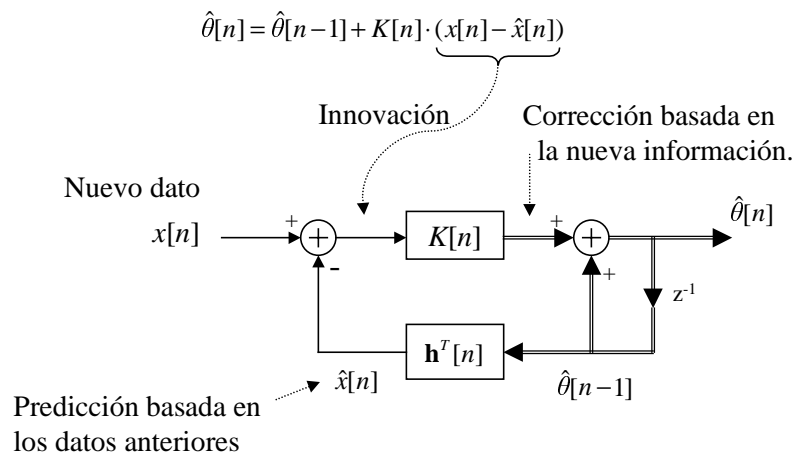
En otras situaciones, el número de parámetros de nuestro modelo es constante pero los datos de los que disponemos se incrementan de forma dinámica con el tiempo, cosa que ocurre, por ejemplo, con el índice de la Bolsa. Como cada vez que nos llega un dato más, resulta deseable aprovechar su información para mejorar nuestra estima actual, se nos plantea de nuevo la pregunta: ¿es necesario re-calcularse toda la solución cada vez que esto ocurre?

Al igual que en el caso precedente, la respuesta es afirmativa. Existe un método secuencial que permite calcular la nueva solución en función de la que se obtuvo sin conocer ese dato adicional. De esta forma no sólo se obtienen ventajas computacionales importantes sino que también se obtiene una visión clara de cuál es la información con la que contribuye cada nuevo dato al modelo de las observaciones.

A esta solución secuencial se llega mediante manipulaciones algebraicas entre las que se utiliza el lema de inversión.

Aumentando Observaciones

Interpretación de la solución:



20

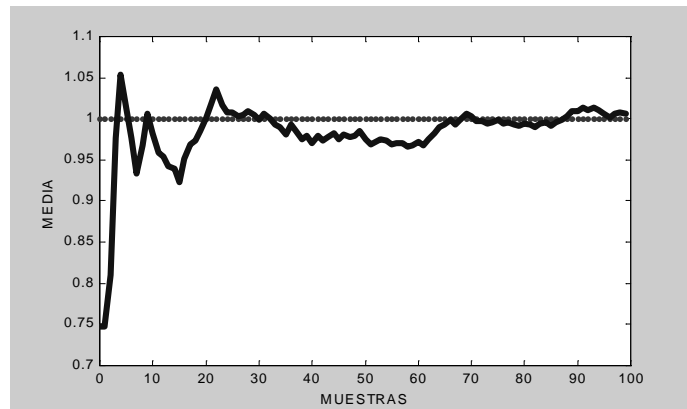
La interpretación de la solución secuencial resulta muy ilustrativa.

A cada nuevo dato se le sustrae la predicción de su valor en base a los datos precedentes, de esta forma se construye la innovación que representa la nueva información aportada por el dato actual. Ésta se multiplica por un factor $K[n]$ que hace las funciones de ganancia y que refleja nuestra confianza en esta información frente a la aportada por los datos precedentes. El resultado de este producto es finalmente la corrección de la estima previa para obtener el estima actual.

Ejemplo

Media de una señal en ruido: $x[n] = \theta + e[n]$

$$\hat{\theta}[n] = \frac{1}{n+1} \sum_{k=0}^n x[k] \Rightarrow \hat{\theta}[n] = \hat{\theta}[n-1] + \frac{1}{n+1} \cdot (x[n] - \hat{\theta}[n-1])$$



21

Como ejemplo de aplicación de la estimación secuencial consideramos un caso muy sencillo: el de estimación de la media de una señal en ruido. El estimador óptimo, que se obtiene por resolución directa de las ecuaciones normales, resulta coincidir para cada conjunto de datos con su media muestral. Por otra parte, su estimación secuencial responde a una ecuación equivalente en la que se desacopla la información pasada y la aportada por el dato actual.

Los resultados se pueden observar en la figura. La media teórica de esta señal es la unidad. Aunque la estima aparece inicialmente muy ruidosa por el escaso número de datos iniciales, la media estimada se acerca progresivamente a la media teórica a medida que el número de datos aumenta.

Ponderación del Error

✧ A veces nos interesa *ponderar cada error por separado*:

$$J(\theta) = \sum_{i=0}^n \mathbf{W}_i \cdot |e[i]|^2 = (\mathbf{x} - \mathbf{H}\theta)^T \mathbf{W}(\mathbf{x} - \mathbf{H}\theta) \quad \text{donde } \mathbf{W} \text{ diagonal def +}$$

Solución: *cambio de variable*

$$\left. \begin{array}{l} \underline{\mathbf{x}} = \mathbf{W}^{1/2} \mathbf{x} \\ \underline{\mathbf{H}} = \mathbf{W}^{1/2} \mathbf{H} \end{array} \right\} J(\theta) = \|\underline{\mathbf{x}} - \underline{\mathbf{H}}\theta\|^2$$

$$\hat{\theta} = \underline{\mathbf{H}}^+ \underline{\mathbf{x}} = (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W} \mathbf{x}$$

22

Resulta fácil imaginar situaciones en las que nuestra confianza en todas las observaciones no sea la misma. En estos casos parece natural ponderar por separado los distintos errores para enfatizar aquellos datos más precisos o fiables.

La solución matemática de este problema es similar a la del problema de mínimos cuadrados determinista ya estudiado, como se comprueba realizando el cambio de variable adecuado.

Restricciones Lineales

✧ Otras veces, la aproximación debe cumplir ciertas *restricciones*:

$$\min_{\theta} J(\theta) \text{ sujeto a } \mathbf{A}\theta = \mathbf{b}$$

Solución: *multiplicadores de Lagrange*

$$L(\theta) = (\mathbf{x} - \mathbf{H}\theta)^T (\mathbf{x} - \mathbf{H}\theta) + \lambda^T (\mathbf{A}\theta - \mathbf{b})$$

$$\begin{cases} \frac{\partial L(\theta)}{\partial \theta} = 0 \\ \mathbf{A}\theta - \mathbf{b} = 0 \end{cases}$$

$$\hat{\theta}_c = \hat{\theta} - (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T (\mathbf{A} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T)^{-1} \cdot (\mathbf{A} \hat{\theta} - \mathbf{b})$$

23

Otro planteamiento distinto consiste en imponer a la aproximación óptima restricciones lineales que, en muchos casos, representan cierta información a priori de la que se dispone sobre el verdadero parámetro de interés, mejorando así la estima.

La solución a este problema con restricciones pasa primero por derivar la función Lagrangiana e igualarla a cero y, en segundo lugar, utilizar las restricciones para hallar los multiplicadores adecuados y así encontrar la solución.

Dependencia No Lineal

✧ ¿Qué podemos hacer cuando $s(\theta)$ depende no linealmente con θ ?

Posibilidades:

A. Casos simples:

1. Transformación de los parámetros.
2. Separabilidad de los parámetros.

B. Casos difíciles:

3. Otros métodos de minimización: NR, GN, etc.

24

Hasta ahora se ha considerado una dependencia lineal de la señal sobre los parámetros, lo cual, nos permitió encontrar soluciones cerradas para el estimador. En general, cuando la dependencia de la señal sobre los parámetros es no lineal, la determinación de soluciones cerradas es muy difícil o imposible. Por ello, se suele recurrir a métodos de búsqueda en rejilla si la dimensionalidad de los parámetros es pequeña, o a métodos iterativos de minimización como el método de Newton-Raphson o el método de Gauss-Newton.

Una de las pocas excepciones a esta regla la constituyen los casos en los que la transformación de parámetros o la separación de parámetros es viable.

Transformación de Parámetros

$$\text{Si } \exists g : \begin{cases} \theta \xleftarrow{\text{Unívoca}} \alpha = g(\theta) \\ s[n] \text{ lineal en } \alpha \end{cases}$$



$$\hat{\alpha} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$$

$$\hat{\theta} = g^{-1}(\hat{\alpha})$$

25

El método de transformación de parámetros consiste en buscar una transformación unívoca de éstos que haga lineal el modelo de la señal en el nuevo espacio.

Si es posible encontrar dicha transformación, se puede resolver el problema de Mínimos Cuadrados directamente sobre el dominio transformado y luego realizar la transformación inversa para obtener el parámetro estimado.

El problema de este método radica en que, en general, la determinación de dicha transformación unívoca suele ser muy difícil.

Ejemplo

Estimación de la amplitud y fase de un coseno en ruido.

$$\begin{aligned}
 x[n] &= A \cos(2\pi f_0 n + \varphi) + e[n] \\
 &= \alpha_1 \cos 2\pi f_0 n + \alpha_2 \sin 2\pi f_0 n + e[n] \quad \Longleftrightarrow \quad \left\{ \begin{array}{l} \alpha = \begin{pmatrix} A \cos \varphi \\ -A \sin \varphi \end{pmatrix} \\ \mathbf{H} = \begin{pmatrix} 1 & 0 \\ \cos 2\pi f_0 & \sin 2\pi f_0 \\ \vdots & \vdots \\ \cos 2\pi f_0 (N-1) & \sin 2\pi f_0 (N-1) \end{pmatrix} \end{array} \right.
 \end{aligned}$$

$$1. \quad \hat{\alpha} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$$

$$2. \quad \begin{pmatrix} \hat{A} \\ \hat{\varphi} \end{pmatrix} = \begin{pmatrix} \sqrt{\hat{\alpha}_1^2 + \hat{\alpha}_2^2} \\ \arctan\left(\frac{-\hat{\alpha}_2}{\hat{\alpha}_1}\right) \end{pmatrix}$$

26

En este ejemplo vamos a ver un caso de dependencia no lineal. Se trata de estimar la amplitud y fase de un coseno, del que se conoce su frecuencia y que se encuentra inmerso en ruido, aplicando el método de transformación de parámetros.

Si descomponemos el coseno con fase en términos de senos y cosenos simples observamos que la transformación a coordenadas cartesianas de la amplitud y de la fase (considerando sólo el argumento principal de ésta) hace lineal la dependencia de la señal sobre este nuevo espacio, al tiempo que la transformación es unívoca.

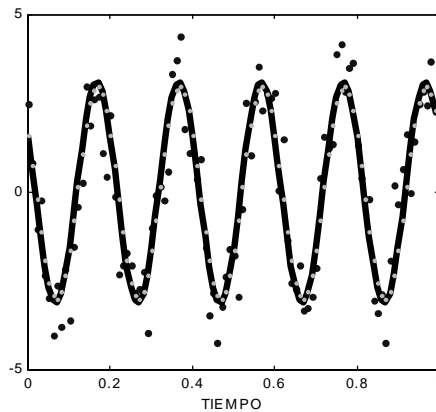
De esta forma, el problema se resuelve fácilmente sobre el dominio transformado y se obtienen los parámetros estimados sin más que invertir la transformación.

Ejemplo (continuación)

$$N = 100$$
$$f_0 = 5\text{Hz.}$$

$$A = 3\text{ V.}$$
$$\varphi = 1\text{ rad.}$$

$$\begin{pmatrix} \hat{A} \\ \hat{\varphi} \end{pmatrix} = \begin{pmatrix} 3.09 \\ 1.04 \end{pmatrix}$$



Señal [negro], observaciones [azul], estima [verde].

27

En esta figura se muestran con los puntos azules 100 observaciones de un coseno inmerso en ruido. El coseno tiene una frecuencia conocida de 5 Hz. y una amplitud y fase desconocidas que nos proponemos estimar. Su componente de señal aparece dibujada en trazo negro continuo sobre la figura.

La estima del coseno se realiza utilizando el estimador obtenido mediante el método de transformación de parámetros y se representa con puntos claros (verdes) en el gráfico.

Sobre la figura se puede comprobar que, a pesar del pequeño tamaño de la población de muestras, la estima resulta tener una gran precisión coincidiendo prácticamente la señal estimada con la señal existente.

Separabilidad de Parámetros

Sea $\theta = (\alpha^T, \beta^T)^T$ el problema se dice *separable* si

$$\mathbf{s} = \mathbf{H}(\alpha) \cdot \beta \quad \begin{cases} \text{no lineal en } \alpha \\ \text{lineal en } \beta \end{cases}$$

$$\Rightarrow \hat{\beta} = (\mathbf{H}^T(\alpha)\mathbf{H}(\alpha))^{-1}\mathbf{H}^T(\alpha)\mathbf{x}$$

y el problema se simplifica a la búsqueda del $\hat{\alpha}$ que minimiza

$$J(\alpha, \hat{\beta}) = \mathbf{x}^T \left(\mathbf{I} - \mathbf{H}(\alpha) (\mathbf{H}^T(\alpha)\mathbf{H}(\alpha))^{-1} \mathbf{H}^T(\alpha) \right) \mathbf{x}$$

28

Algunas veces los parámetros no son directamente transformables aunque sí existe un subconjunto de ellos sobre los que la señal depende linealmente. En estos casos el problema se dice separable.

El problema se simplifica ya que su dimensión se reduce. La expresión óptima de los parámetros lineales se puede encontrar aplicando Mínimos Cuadrados en función del resto. De esta forma, el problema se limita a una búsqueda del mínimo sobre los parámetros que caracterizan la dependencia no lineal de la señal.

Otros métodos

Iteración de Newton-Raphson: busca los ceros de $\left(\frac{\partial J}{\partial \theta^T}\right)^T$

$$\theta^{(k+1)} = \theta^{(k)} - \left(\frac{\partial}{\partial \theta^T} \left(\frac{\partial J}{\partial \theta^T}\right)^T\right)^{-1} \left(\frac{\partial J}{\partial \theta^T}\right)^T \Big|_{\theta=\theta^{(k)}}$$

Iteración de Gauss-Newton: lineariza la dependencia de $\mathbf{s}(\theta)$ sobre θ .

$$s[n; \theta] \approx s[n; \theta^{(k)}] + \underbrace{\frac{\partial s[n; \theta^{(k)}]}{\partial \theta^T}}_{\mathbf{H}(\theta^{(k)})} (\theta - \theta^{(k)})$$

$$\theta^{(k+1)} = \theta^{(k)} + \left(\mathbf{H}^T(\theta^{(k)})\mathbf{H}(\theta^{(k)})\right)^{-1} \mathbf{H}^T(\theta^{(k)}) \cdot (\mathbf{x} - \mathbf{s}(\theta^{(k)}))$$

Problemas de convergencia

29

Cuando los métodos de transformación de parámetros o de separabilidad no son aplicables debemos recurrir a algoritmos de minimización no lineal, que suelen consistir en métodos de búsqueda iterativa de la solución. Dos de los algoritmos de este tipo más famosos son los métodos de Newton-Raphson y de Gauss-Newton.

El método de Newton-Raphson es el conocido método de búsqueda de los ceros de la función derivada. Es un método invariante (que no depende de la base sobre la que estén expresados los parámetros) y posee una convergencia cuadrática cuando ésta se da. Uno de sus principales inconvenientes es que precisa el cálculo explícito del Hessiano de la función.

El método de Gauss-Newton es propiamente un algoritmo de minimización que lineariza la dependencia de la señal sobre los parámetros entorno a su valor nominal en cada iteración. De esta forma, cada una de las funciones de coste resultantes de esta secuencia es cuadrática y se minimiza en una sola iteración.

Desafortunadamente, ambos métodos, pueden evidenciar problemas de convergencia cuando la inicialización se encuentra muy lejos de la solución.

Puntos a Tratar

1. Introducción.
2. Mínimos cuadrados determinista.
 - 2.1. Interpretación geométrica.
 - 2.2. Aumentando el orden de la aproximación.
 - 2.3. Aumentando el número de observaciones.
 - 2.4. Mínimos cuadrados con restricciones.
 - 2.5. Dependencia no lineal.
- ⇒ 3. Mínimos cuadrados estocástico.
 - 3.1. Teorema de Gauss-Markov.
 - 3.2. Interpretación Geométrica.
 - 3.3. Teorema de Aitken.
4. Conclusiones.
5. Referencias.

30

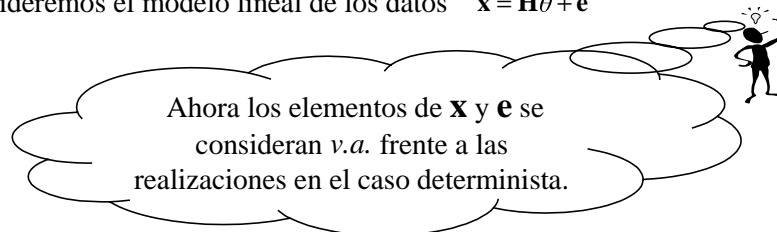
Otra variante del método de Mínimos Cuadrados se considera un método estocástico.

Esta variante presenta similitudes en su formulación y, a la vez, claras diferencias de interpretación con respecto método determinista.

En este apartado plantaremos una de las variantes del problema de Mínimos Cuadrados Estocástico en la que los parámetros del modelo se consideran un vector determinista. Enunciaremos el teorema de Gauss-Markov y presentaremos una interpretación geométrica de éste. Para terminar exponiendo el caso más general del teorema precedente que se conoce bajo el nombre de teorema de Aitken.

Mínimos Cuadrados Estocástico

Consideremos el modelo lineal de los datos $\mathbf{x} = \mathbf{H}\theta + \mathbf{e}$



◇ *Método de estimación:* hipótesis prob. sobre la distribución del error.

caracterización parcial basada en la media y covarianza. \Rightarrow Ventaja frente a ML y técnicas Bayesianas.

31

La radical diferencia entre el método determinista y el método estocástico consiste en que ahora se consideran hipótesis probabilísticas sobre la función de densidad de probabilidad del error y de las observaciones. Esto hace que el método de Mínimos Cuadrados Estocástico se considere un método propio de estimación frente a la versión determinista que se entendía como un método de aproximación.

Ahora los elementos del error y de las observaciones son variables aleatorias distribuidas de acuerdo con una cierta función densidad de probabilidad.

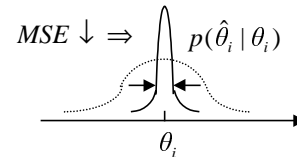
El método de Mínimos Cuadrados Estocástico es una de las varias aproximaciones existentes en teoría de la estimación, siendo las principales alternativas a éste: Máxima Verosimilitud y las técnicas Bayesianas. Estas otras técnicas normalmente requieren una descripción estadística completa de las variables del problema en términos de sus funciones de distribución de probabilidad, mientras que el método de Mínimos Cuadrados Estocástico sólo requiere una caracterización parcial basada en medias y matrices de covarianza.

Mínimos Cuadrados Estocástico

✧ Criterio natural de optimalidad:

$$\forall i \quad MSE_{\hat{\theta}_i} = |\text{sesgo}_{\hat{\theta}_i}|^2 + \text{var}_{\hat{\theta}_i}$$

$$E|\hat{\theta}_i - \theta_i|^2 = |E[\hat{\theta}_i] - \theta_i|^2 + E|\hat{\theta}_i - E[\hat{\theta}_i]|^2$$



✧ Estimador lineal $\hat{\theta}_i = \mathbf{a}_i^T \mathbf{x}$ e insesgado $E[\hat{\theta}_i] = \theta_i \Leftrightarrow \mathbf{a}_i^T \mathbf{h}_j = \delta_{ij}$

$$MSE_{\hat{\theta}_i} = \text{var}_{\hat{\theta}_i} = \mathbf{a}_i^T \mathbf{C}_e \mathbf{a}_i$$

✧ **Objetivo:**

$$\text{Min}_{\hat{\theta}_i} \text{var}_{\hat{\theta}_i} \quad \text{sueto a} \quad \hat{\theta}_i \text{ lineal e insesgado} \quad \forall i = 1, \dots, p$$

$$\text{Min}_{\hat{\theta}_i} \mathbf{a}_i^T \mathbf{C}_e \mathbf{a}_i \quad \text{sueto a} \quad \mathbf{a}_i^T \mathbf{h}_j = \delta_{ij} \quad \forall i, j = 1, \dots, p$$

32

Un criterio natural de optimalidad en estadística consisten en minimizar el error cuadrático medio (MSE) entre el el estimador real y su estima (que es una variable aleatoria). La justificación de este criterio radica en que trata de agrupar la función densidad de probabilidad del estimador sobre el parámetro verdadero. El término MSE se descompone en otros dos con interpretaciones muy significativas: por una parte está el sesgo del estimador al cuadrado mientras que, por la otra, está la varianza del estimador.

En lugar de pretender minimizar el criterio MSE de forma genérica, el método de mínimos cuadrados estocástico limita esa búsqueda a una clase especial de estimadores, la de los estimadores lineales e insesgados.

Así, el problema de mínimos cuadrados estocástico se traduce en minimizar la varianza de cada componente del estimador sujeto a restricciones que garantizan que éste sea lineal e insesgado. La formulación matemática de este problema resulta ser totalmente determinista y su solución se explicita en el siguiente teorema.

Teorema de Gauss-Markov

Si los datos siguen el modelo lineal y el error tiene media cero y matriz de covarianza $\mathbf{C}_e = \sigma_e^2 \mathbf{I}$ el óptimo estimador Lineal e Inssegado de Mínima Varianza de los parámetros es

$$\hat{\theta} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \quad \text{y} \quad \mathbf{C}_{\hat{\theta}} = \sigma_e^2 (\mathbf{H}^T \mathbf{H})^{-1}$$

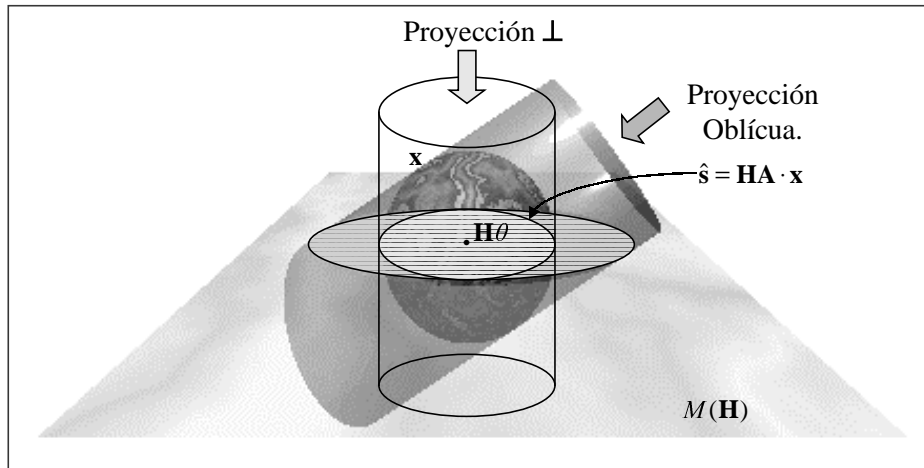
Similar a la soluc. de MCD,
aunque la interpretación es distinta.

33

El teorema de Gauss-Markov contempla la hipótesis adicional de que el error sea de media cero y que la matriz de covarianzas del error tenga la estructura de la matriz identidad escalada por la varianza del error. Bajo estas condiciones, el teorema nos proporciona la expresión del estimador óptimo lineal e inssegado de mínima varianza, así como también, la matriz de covarianza de este estimador.

El sorprendente el hecho de que la solución a este problema sea la misma que la del método determinista puede conducir en muchos casos a una interpretación errónea de este estimador.

Interpretación Geométrica



La proyección $\perp \Rightarrow \min C_s \Rightarrow \min C_{\theta_i} \forall i$

34

La interpretación geométrica del teorema de Gauss-Markov se presenta en la siguiente figura. El plano denota el subespacio generado por las columnas de la matriz de observación. En el centro de éste hay un punto que denota la componente de señal. Sobre la señal se superpone la v.a. ruido que (al ser de varianza constante para cada componente) se representa como una esfera centrada en la componente de señal y con el radio de la desviación típica. La v.a. observaciones es pues la resultante de sumar la señal más la v.a. ruido.

La proyección de la v.a. observación sobre el plano conduce, en general, a una superficie elíptica centrada en el valor verdadero de la señal y que representa la v.a. que estima la componente de señal. La parte común a todas las posibles proyecciones está dada por el área que proporciona la proyección ortogonal. Esta es, pues, la proyección que garantiza una menor varianza en cada una de las componentes del estimador lineal insesgado.

Teorema de Aitken

Si los datos siguen el modelo lineal y el error tiene media cero y matriz de covarianza \mathbf{C}_e el óptimo estimador Lineal e Insesgado de Mínima Varianza de los parámetros es

$$\hat{\theta} = (\mathbf{H}^T \mathbf{C}_e^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}_e^{-1} \mathbf{x} \quad \text{y} \quad \mathbf{C}_{\hat{\theta}} = (\mathbf{H}^T \mathbf{C}_e^{-1} \mathbf{H})^{-1}$$

Similar a MCD con
ponderación del error.

Demonstración: *cambio de variable*

$$\left. \begin{array}{l} \underline{\mathbf{x}} = \mathbf{C}_e^{-1/2} \mathbf{x} \\ \underline{\mathbf{H}} = \mathbf{C}_e^{-1/2} \mathbf{H} \end{array} \right\} \mathbf{C}_e = \mathbf{I} \Rightarrow \hat{\theta} = \underline{\mathbf{H}}^+ \underline{\mathbf{x}} = (\mathbf{H}^T \mathbf{C}_e^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}_e^{-1} \mathbf{x}$$

35

La versión general del estimador de mínimos cuadrados estocástico la proporciona el teorema de Aitken. En este caso se permite al error tener una matriz de covarianzas arbitraria salvo por el hecho de que esta debe ser de rango completo y definida positiva.

La solución general del problema es similar a la del método determinista con ponderación del error cuando se utiliza como matriz de ponderación la inversa de la matriz de covarianzas del error.

La demostración de este teorema se logra a través de un sencillo cambio de variable que haga que la nueva variable aleatoria error cumpla las condiciones del teorema de Gauss-Markov.

Puntos a Tratar

1. Introducción.
2. Mínimos cuadrados determinista.
 - 2.1. Interpretación geométrica.
 - 2.2. Aumentando el orden de la aproximación.
 - 2.3. Aumentando el número de observaciones.
 - 2.4. Mínimos cuadrados con restricciones.
 - 2.5. Dependencia no lineal.
3. Mínimos cuadrados estocástico.
 - 3.1. Teorema de Gauss-Markov.
 - 3.2. Interpretación Geométrica.
 - 3.3. Teorema de Aitken.
- ⇒ 4. Conclusiones.
5. Referencias.

36

En este punto nos disponemos a presentar las conclusiones más importantes y algunas referencias útiles sobre el tema.

Conclusiones Importantes

- ☑ MC determinista \Leftrightarrow método de aproximación.
 - + Interpretación geométrica y principio de ortogonalidad.
 - + Innovación \Leftrightarrow soluc. recursivas.
 - Dependencia no lineal \Leftrightarrow técnicas de minimización.

- ☑ MC estocástico \Leftrightarrow método de estimación.
 - + Caracterización parcial.
 - + Equivalencia matemática entre las soluc MCD y el MCE.
 - ¡Interpretación diferente de MCD!

- ☑ Ambos métodos:
 - + Simplicidad en las hipótesis \Leftrightarrow estimador muy utilizado.

37

El método de Mínimos Cuadrados Determinista es propiamente un método de aproximación. Su interpretación geométrica nos conduce a un importante principio: el principio de ortogonalidad cuya validez es extensible a cualquier espacio de Hilbert.

Otro concepto clave es el de innovación que nos conduce a una interpretación sencilla de las soluciones recursivas.

El caso más complicado es el de dependencia no lineal en la que la solución al problema pocas veces adopta la forma de una expresión cerrada.

El método de Mínimos Cuadrados Estocástico es propiamente un método de estimación. Frente a otros métodos estadísticos éste sólo requiere una caracterización parcial del error.

Existe una equivalencia matemática entre las soluciones al problema determinista y estocástico aunque la interpretación de ambos difiere.

En ambos métodos la simplicidad de sus hipótesis y formulación hacen que estos sean muy utilizados en la práctica para la estimación de los parámetros de interés.

Conclusiones Futuras

- + Otra versión estocástica del método de MC considera a θ como una v.a. \Leftrightarrow Estimación Lineal Bayesiana
- MC no posee consideraciones de optimalidad (salvo si $e \sim$ Gauss.).

38

Algunas de las conclusiones que todavía no hemos podido obtener pero que examinaremos más adelante en el curso, y que aparecen directamente relacionadas con este tema, son las siguientes:

Existe otra versión estocástica del método de Mínimos Cuadrados que considera al vector de parámetros como un vector de variables aleatorias. Esta aproximación recibe el nombre de estimación Bayesiana y se suele utilizar para añadir información conocida a priori sobre el vector de parámetros para, así, intentar mejorar la estima.

El método de Mínimos Cuadrados Estocástico es un método de estimación que, en general, no posee propiedades de optimalidad estadística, salvo cuando el error se distribuye de acuerdo con una función de probabilidad Gaussiana, puesto que, en este caso, el estimador óptimo es una función lineal de las observaciones.

Referencias

- [Kay,93] Kay, Steven M., "Fundamentals of Statistical Signal Processing: Estimation Theory", Prentice Hall, 1993. ISBN: 0-13-345711-7. Capítulos 4,6 y 8.
- [Luenberger,69] Luenberger, David G., "Optimization by Vector Space Methods", John Wiley & Sons, 1969. SBN: 47155359x. Capítulos 3 y 4.
- [Kailath,00] Kaylath, T., Sayed, A.H., Hassibi, B., "Linear Estimation", Prentice Hall, 2000. ISBN: 0-13-022464-2. Capítulos 1 y 2.
- [Haykin,96] Haykin, Simon, "Adaptive Filter Theory", Tercera edición, Prentice Hall, ISBN: 0-13-397985-7. Capítulos 11, 12 y 13.
- [Mendel, 95] Mendel, Jerry M., "Lessons in Estimation Theory for Signal Processing, Communications and Control", Prentice Hall, 1995. ISBN: 0-13-120981-7. Capítulos 1 al 8.
- [Stuard,91] Stuard, A., Ord, J.K. "Kellidall's Advanced Theory of Statistics", Quinta edición, Volumen 2: "Classical Inference and Relationship", editorial Edward Arnold, ISBN: 0-340-52923-7. Capítulos 17, 19 y 28.

39

De entre la bibliografía recomendada sobre el tema de Mínimos Cuadrados destacamos algunos libros en especial:

Las referencias preferidas sobre el tema son [Kay,93] y [Luenberger,69]. La primera por su claridad de exposición y abundante provisión de ejemplos, la segunda referencia por su enfoque geométrico del problema de Mínimos Cuadrados sobre espacios vectoriales de Hilbert.

Otros libros interesantes son [Haykin,96] y [Mendel,95] que están pensados como libros de texto para los últimos cursos de la carrera y, en especial, para cursos de doctorado.

Por último, dos libros para profundizar en el tema son [Kailath,00] (que acaba de aparecer recientemente) y [Stuard,91]. Ambos tratan de una forma mucho más exhaustiva y detallada el problema de estimación lineal en sus distintas variantes.