

Mínimos cuadrados: tres interpretaciones

Antonio Sala

Notas de clase sobre control de sistemas multivariables

Dept. Ing. Sistemas y Automática (DISA)
Universitat Politècnica de València (UPV)

*Presentación vídeo disponible en:

<http://personales.upv.es/asala/YT/V/ls3i.html>

Presentación

Motivación:

Las fórmulas de mínimos cuadrados pueden ser obtenidas como resultado de una optimización, de unas operaciones con matrices pseudoinversas, de inferencia estadística.

Objetivos:

Comprender las relaciones entre las tres interpretaciones del significado del óptimo mínimo-cuadrático.

Contenidos:

Interpretación determinista (optimización). Interpretación geométrica (proyección ortogonal). Interpretación estadística (mínima correlación, mínima varianza de error). Conclusiones.

I Mínimos cuadrados: interpretación determinista

Mínimos cuadrados con N datos (trasladados a media cero):

$$<\text{minimizar } J(\theta) = \frac{1}{2} \sum_{i=1}^N (y_k - \theta_{1 \times m} x_{k \ m \times 1})^2 >$$

Objetivo: obtener un modelo que permita predecir y a partir de x , en la forma $y_{predic} = \theta \cdot x$, considerando x **medible**. Denotemos error como $e = (y - y_{predic})$. El gradiente de $J = \frac{1}{2} \sum_{k=1}^n e_k^2$ es:

$$\left. \frac{dJ}{d\theta} \right|_{1 \times m} = \sum_{k=1}^N (y_k - \theta x_k)_{1 \times 1} x_k^T = \sum_{k=1}^N e_k x_k^T =$$

$$\left(\sum_{k=1}^N y_k x_k^T \right)_{1 \times m} - \theta_{1 \times m} \cdot \left(\sum_{k=1}^N x_k x_k^T \right)_{m \times m}$$

Solución: $\frac{dJ}{d\theta} = 0, \Rightarrow$

$$\theta_{opt, 1 \times m} = \left(\sum_{k=1}^N y_k x_k^T \right)_{1 \times m} \cdot \left(\sum_{k=1}^N x_k x_k^T \right)_{m \times m}^{-1}$$

II Mínimos cuadrados: interpretación geométrica

Si se construye matriz de datos $X_{m \times N}$, vector de salidas $Y_{1 \times N}$ tenemos:

$$\left(\sum_{k=1}^N x_k x_k^T \right)_{m \times m} = XX^T, \quad \left(\sum_{k=1}^N y_k x_k^T \right)_{1 \times m} = YX^T$$

$$J_{1 \times 1} = (Y - \theta X)(Y - \theta X)^T$$

$$\theta_{opt, 1 \times m} = YX^T(XX^T)^{-1} = YX^\dagger$$

► El vector (fila) de error $E = Y - \theta_{opt}X$ es ortogonal a la aproximación $\hat{Y} = \theta_{opt}X$.

En efecto:

$$E \hat{Y}^T = (Y - \theta_{opt}X) \cdot X^T \overbrace{\theta_{opt}^T}^{\theta_{opt}^T} = (Y - YX^T(XX^T)^{-1}X) \cdot X^T \overbrace{(XX^T)^{-1}XY^T}^{\theta_{opt}^T} =$$

$$YX^T(XX^T)^{-1}XY^T - YX^T(XX^T)^{-1}XX^T(XX^T)^{-1}XY^T = 0$$

II Mínimos cuadrados: interpretación geométrica

► La aproximación $\hat{Y} = \theta_{opt}X = Y \overbrace{X^T(XX^T)^{-1}X}^{\text{matriz de proyección}}$ es la “**proyección ortogonal**” de $Y_{1 \times N}$ sobre el subespacio m -dimensional generado por las filas de X .

$Y = \overbrace{(Y - \theta_{opt}X)}^{\nu} + \overbrace{\theta_{opt}X}^{\hat{Y}}$, expresada como suma de dos vectores ortogonales ($\nu \perp \hat{Y}$), un vector $\hat{Y}$ combinación lineal de filas de X (componente **explicado**), y otro ν perpendicular a éste (perpendicular a filas de X , componente **no explicado**).

Es una “proyección” porque la “proyección de la proyección” ya no cambia nada:

$$\hat{Y} \cdot X^T(XX^T)^{-1}X = YX^T(XX^T)^{-1}X \cdot X^T(XX^T)^{-1}X = \hat{Y}$$

Informalmente, el componente “no explicado” del componente “explicado” es cero.

III Mínimos cuadrados: interpretación estadística

Matrices de varianzas-covarianzas relacionadas (suponiendo que a todo le hemos restado su media muestral):

$$\Sigma_{xx} = E[xx^T] = \left(\frac{1}{N-1} \sum_{k=1}^N x_k x_k^T \right)_{m \times m} = \frac{1}{N-1} XX^T$$

$$\Sigma_{yx} = E[yx^T] = \left(\frac{1}{N-1} \sum_{k=1}^N y_k x_k^T \right)_{1 \times m} = \frac{1}{N-1} YX^T$$

$$\Sigma_{ex} = E[ex^T] = \left(\frac{1}{N-1} \sum_{k=1}^N e_k x_k^T \right)_{1 \times m}$$

III Mínimos cuadrados: interpretación estadística

$$\theta_{opt, 1 \times m} = YX^T(XX^T)^{-1} = YX^T \frac{1}{N-1} \left(\frac{1}{N-1} \right)^{-1} (XX^T)^{-1} = \Sigma_{yx} \Sigma_{xx}^{-1}$$

Modelo óptimo mínimos cuadrados $\frac{dJ}{d\theta} = 0 \Leftrightarrow$

$\Sigma_{ex} = 0$, **covarianza** (muestral) entre “error” y “datos entrada” es **cero**.

Predicción óptima en sentido “estadístico” (error no correlado con regresores, expresión en función de matrices de varianza-covarianza) coincide con óptimo “determinista” de mínimos cuadrados (si datos tienen media 0)... y, también volviendo a estadística, con **minimización** de la **varianza muestral** del error $\min_{\theta} E(ee^T) = \min_{\theta} \frac{2}{N-1} J(\theta)$.

Conclusiones

- El problema de **mínimos cuadrados** tiene interpretación **determinista** como **optimización** de un índice de coste (derivadas parciales igual a cero).

Ajuste de parámetros

- También, interpretación **geométrica** como: (a) error ortogonal a aproximación, (b) aproximación es la **proyección ortogonal** de medidas sobre *subespacio de regresores*.

Identificación subespacio

- Por último, interpretación **estadística** como: (a) error **no** tiene **correlación** (muestral) con regresores, (b) se minimiza la **varianza** del error.

Identificación con ruido, filtro de Kalman