

MAT-266: Diferenciación matricial

Felipe Osorio

`fosorios.mat.utfsm.cl`

Departamento de Matemática, UTFSM



Notación:

Denotaremos por ϕ , \mathbf{f} y \mathbf{F} **funciones** escalar, vectorial y matricial, respectivamente mientras que ζ , \mathbf{x} y \mathbf{X} **argumentos** escalar, vectorial y matricial, respectivamente.

Ejemplo:

Podemos escribir los siguientes casos particulares:

$$\begin{array}{lll} \phi(\zeta) = \zeta^2, & \phi(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}, & \phi(\mathbf{X}) = \text{tr}(\mathbf{X}^\top \mathbf{X}), \\ \mathbf{f}(\zeta) = (\zeta, \zeta^2)^\top, & \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}, & \mathbf{f}(\mathbf{X}) = \mathbf{X}\mathbf{a}, \\ \mathbf{F}(\zeta) = \zeta^2 \mathbf{I}_n, & \mathbf{F}(\mathbf{x}) = \mathbf{x}\mathbf{x}^\top, & \mathbf{F}(\mathbf{X}) = \mathbf{X}^\top. \end{array}$$



Considere $\phi : S \rightarrow \mathbb{R}$ con $S \subset \mathbb{R}^n$, se define la derivada de ϕ con relación a $\mathbf{x} \in S$ como

$$\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial \phi}{\partial x_1}, \dots, \frac{\partial \phi}{\partial x_n} \right)^\top = \left(\frac{\partial \phi}{\partial x_i} \right) \in \mathbb{R}^n$$

de este modo, introducimos la notación

$$D\phi(\mathbf{x}) = \frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}^\top} \in \mathbb{R}^{1 \times n}.$$

Ahora, si $\mathbf{f} : S \rightarrow \mathbb{R}^m$, $S \subset \mathbb{R}^n$. Entonces la matriz $m \times n$,

$$D\mathbf{f}(\mathbf{x}) = \begin{pmatrix} Df_1(\mathbf{x}) \\ \vdots \\ Df_m(\mathbf{x}) \end{pmatrix} = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^\top},$$

es la **derivada** o **matriz Jacobiana** de \mathbf{f} . La transpuesta de la matriz Jacobiana $D\mathbf{f}(\mathbf{x})$ se denomina **gradiente** de $\mathbf{f}(\mathbf{x})$.



Considere la fórmula de Taylor de primer orden,

$$\phi(c + u) = \phi(c) + u\phi'(c) + r_c(u),$$

donde,

$$\lim_{u \rightarrow 0} \frac{r_c(u)}{u} = 0.$$

es de orden más pequeño que u conforme $u \rightarrow 0$. Note también que

$$\lim_{u \rightarrow 0} \frac{\phi(c + u) - \phi(c)}{u} = \phi'(c).$$

De este modo, se define

$$d\phi(c; u) = u\phi'(c),$$

como el (primer) diferencial de ϕ en c con incremento u . Esto motiva la siguiente definición.



Definición 1:

Sea $f : S \rightarrow \mathbb{R}^m$, $S \subset \mathbb{R}^n$, si existe una matriz $A \in \mathbb{R}^{m \times n}$, tal que

$$f(c + u) = f(c) + A(c)u + r_c(u),$$

para todo $u \in \mathbb{R}^n$ con $\|u\| < \delta$, y

$$\lim_{u \rightarrow 0} \frac{r_c(u)}{\|u\|} = 0,$$

entonces la función f se dice diferenciable en c . El vector $m \times 1$

$$df(c; u) = A(c)u,$$

se denomina **primer diferencial** de f en c con incremento u .



Resultado 1 (Magnus y Neudecker, 1985)¹:

Sea $f : S \rightarrow \mathbb{R}^m$, $S \subset \mathbb{R}^n$ función diferenciable, $c \in S$ y u un vector n -dimensional. Entonces

$$df(c; u) = (Df(c))u.$$

La matriz $Df(c) \in \mathbb{R}^{m \times n}$ se denomina **matriz Jacobiana**. Tenemos también que

$$\nabla f(c) = (Df(c))^T$$

es la **matriz gradiente** de f .

¹Journal of Mathematical Psychology 29, 474-492.

También es conocido como el “**Primer teorema de identificación**”



Definición 2:

Sea $\mathbf{A} \in \mathbb{R}^{n \times q}$ particionada como

$$\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_q),$$

donde $\mathbf{a}_k \in \mathbb{R}^n$ es la k -ésima columna de \mathbf{A} . Entonces

$$\text{vec}(\mathbf{A}) = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_q \end{pmatrix}.$$

Definición 3:

Sea $\mathbf{A} \in \mathbb{R}^{m \times n}$ y $\mathbf{B} \in \mathbb{R}^{p \times q}$, entonces el producto Kronecker entre \mathbf{A} y \mathbf{B} denotado por $\mathbf{A} \otimes \mathbf{B}$ es la matriz $mp \times nq$ definida como

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{pmatrix}$$



Resultado 2:

Sean A, B, C y D matrices de órdenes apropiados y λ escalar. Entonces

$$(a) \quad A \otimes B \otimes C = (A \otimes B) \otimes C = A \otimes (B \otimes C),$$

$$(b) \quad (A + B) \otimes (C + D) = A \otimes C + B \otimes C + A \otimes D + B \otimes D,$$

$$(c) \quad (A \otimes B)(C \otimes D) = AC \otimes BD,$$

$$(d) \quad \lambda \otimes A = \lambda A = A \otimes \lambda,$$

$$(e) \quad (A \otimes B)^{\top} = A^{\top} \otimes B^{\top},$$

$$(f) \quad (A \otimes B)^{-1} = A^{-1} \otimes B^{-1},$$

$$(g) \quad (A \otimes B)^{-} = A^{-} \otimes B^{-}.$$



Resultado 3:

Sean $A \in \mathbb{R}^{n \times n}$ y $B \in \mathbb{R}^{p \times p}$. Entonces

(a) $\text{tr}(A \otimes B) = \text{tr}(A) \text{tr}(B)$,

(b) $|A \otimes B| = |A|^p |B|^n$,

(c) $\text{rg}(A \otimes B) = \text{rg}(A) \text{rg}(B)$.

Observación:

Si $a \in \mathbb{R}^n$ y $b \in \mathbb{R}^p$, entonces

$$ab^T = a \otimes b^T = b^T \otimes a.$$

Por otro lado, tenemos que

$$\text{vec}(ab^T) = \text{vec}(a \otimes b^T) = \text{vec}(b^T \otimes a) = b \otimes a.$$

Esto sugiere una conexión entre el operador de vectorización, el producto Kronecker y la traza.



Resultado 4:

(a) Si A y B son ámbas matrices de orden $m \times n$, entonces

$$\text{tr } A^\top B = \text{vec}^\top A \text{vec } B,$$

(b) Si A, B y C son de órdenes adecuados, entonces

$$\text{vec } ABC = (C^\top \otimes A) \text{vec } B,$$

donde $\text{vec}^\top A = (\text{vec } A)^\top$.

Resultado 5:

Sean A, B, C y D matrices, tal que, el producto $ABCD$ está definido y es cuadrado, entonces

$$\text{tr } ABCD = \text{vec}^\top D^\top (C^\top \otimes A) \text{vec } B = \text{vec}^\top D (A \otimes C^\top) \text{vec } B^\top.$$



Sea $\mathbf{F} : S \rightarrow \mathbb{R}^{m \times p}$, $S \subset \mathbb{R}^{n \times q}$ una función matricial, podemos notar que

$$\text{vec } \mathbf{F}(\mathbf{X}) = \mathbf{f}(\text{vec } \mathbf{X})$$

esto permite obtener el diferencial de una función matricial considerando la relación

$$\text{vec } d\mathbf{F}(\mathbf{C}; \mathbf{U}) = d\mathbf{f}(\text{vec } \mathbf{C}; \text{vec } \mathbf{U})$$

en cuyo caso \mathbf{F} tiene matriz Jacobiana

$$D\mathbf{F}(\mathbf{C}) = D\mathbf{f}(\text{vec } \mathbf{C})$$

Resultado 6:

Sea $\mathbf{F} : S \rightarrow \mathbb{R}^{m \times p}$, $S \subset \mathbb{R}^{n \times q}$ función diferenciable, $\mathbf{C} \in S$ y \mathbf{U} matriz $n \times q$.
Entonces

$$\text{vec } d\mathbf{F}(\mathbf{C}; \mathbf{U}) = (D\mathbf{F}(\mathbf{C})) \text{vec } \mathbf{U}.$$

con $(D\mathbf{F}(\mathbf{C}))^\top$ la matriz gradiente de \mathbf{F} .



Considere $\phi : S \rightarrow \mathbb{R}$ con $S \subset \mathbb{R}^n$, entonces se define la **matriz Hessiana** como la matriz de segundas derivadas, dada por

$$H \phi(\mathbf{x}) = \frac{\partial^2 \phi(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \frac{\partial}{\partial \mathbf{x}^\top} \left(\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}^\top} \right)^\top = D(D \phi(\mathbf{x}))^\top.$$

Evidentemente, el segundo diferencial de una función escalar está dado por

$$d^2 \phi = d(d \phi).$$

Resultado 6:

Sea $\phi : S \rightarrow \mathbb{R}$, $S \subset \mathbb{R}^n$ dos veces diferenciable, $\mathbf{c} \in S$ y \mathbf{u} vector n -dimensional. Entonces

$$d^2 \phi(\mathbf{c}; \mathbf{u}) = \mathbf{u}^\top (H \phi(\mathbf{c})) \mathbf{u}.$$

donde $H \phi(\mathbf{c}) \in \mathbb{R}^{n \times n}$ es la **matriz Hessiana** de ϕ .



Observación:

Algunas ventajas (prácticas) importantes del cálculo de diferenciales son:

- ▶ Sea $\mathbf{f}(\mathbf{x})$ función vectorial $m \times 1$ con argumento \mathbf{x} , vector n -dimensional, entonces

$$D\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{m \times n} \quad \text{sin embargo,} \quad d\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$$

- ▶ Para funciones matriciales, $d\mathbf{F}(\mathbf{X})$ tiene la **misma** dimensión que \mathbf{F} **sin importar** la dimensión de \mathbf{X} .



Reglas fundamentales:

Considere u y v funciones escalares y α una constante, entonces:

$$d\alpha = 0,$$

$$d(\alpha u) = \alpha du,$$

$$d(u + v) = du + dv,$$

$$d(uv) = (du)v + u(dv)$$

$$d(u/v) = \frac{(du)v - u(dv)}{v^2}, (v \neq 0),$$

$$du^\alpha = \alpha u^{\alpha-1} du,$$

$$de^u = e^u du,$$

$$d \log u = u^{-1} du, (u > 0)$$

$$d\alpha^u = \alpha^u \log \alpha du, (\alpha > 0).$$

Aquí por ejemplo,

$$\phi(x) = u(x) + v(x).$$



Reglas fundamentales:

Análogamente para \mathbf{U} , \mathbf{V} funciones matriciales, α un escalar (constante) y $\mathbf{A} \in \mathbb{R}^{m \times n}$ constante, tenemos

$$\begin{aligned}d\mathbf{A} &= \mathbf{0}, & d(\alpha\mathbf{U}) &= \alpha d\mathbf{U}, \\d(\mathbf{U} + \mathbf{V}) &= d\mathbf{U} + d\mathbf{V}, & d(\mathbf{U}\mathbf{V}) &= (d\mathbf{U})\mathbf{V} + \mathbf{U}d\mathbf{V}, \\d(\mathbf{U} \otimes \mathbf{V}) &= d\mathbf{U} \otimes d\mathbf{V}, & d(\mathbf{U} \odot \mathbf{V}) &= d\mathbf{U} \odot d\mathbf{V}, \\d\mathbf{U}^\top &= (d\mathbf{U})^\top, & d\operatorname{vec} \mathbf{U} &= \operatorname{vec} d\mathbf{U}, \\d\operatorname{tr} \mathbf{U} &= \operatorname{tr} d\mathbf{U}.\end{aligned}$$

Otros diferenciales de uso frecuente en Estadística son:

$$\begin{aligned}d|\mathbf{F}| &= |\mathbf{F}| \operatorname{tr} \mathbf{F}^{-1} d\mathbf{F}, & d\log |\mathbf{F}| &= \operatorname{tr} \mathbf{F}^{-1} d\mathbf{F}, \\d\mathbf{F}^{-1} &= -\mathbf{F}^{-1}(d\mathbf{F})\mathbf{F}^{-1}.\end{aligned}$$



Ejemplo (Mínimos cuadrados):

Considere el problema de optimización

$$\min_x \phi(\mathbf{x}), \quad \phi(\mathbf{x}) = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2,$$

donde $\mathbf{b} \in \mathbb{R}^n$ y $\mathbf{A} \in \mathbb{R}^{n \times p}$ con $\text{rg}(\mathbf{A}) = p$.

Diferenciando ϕ con relación a \mathbf{x} , tenemos

$$d\phi(\mathbf{x}) = d(\mathbf{b} - \mathbf{A}\mathbf{x})^\top (\mathbf{b} - \mathbf{A}\mathbf{x}) + (\mathbf{b} - \mathbf{A}\mathbf{x})^\top d(\mathbf{b} - \mathbf{A}\mathbf{x}).$$

Notando que

$$d(\mathbf{b} - \mathbf{A}\mathbf{x}) = -\mathbf{A} d\mathbf{x},$$

(y análogamente $d(\mathbf{b} - \mathbf{A}\mathbf{x})^\top = -(d\mathbf{x})^\top \mathbf{A}^\top$), obtenemos

$$d\phi(\mathbf{x}) = -(d\mathbf{x})^\top \mathbf{A}^\top (\mathbf{b} - \mathbf{A}\mathbf{x}) - (\mathbf{b} - \mathbf{A}\mathbf{x})^\top \mathbf{A} d\mathbf{x}.$$

Evidentemente,

$$((\mathbf{b} - \mathbf{A}\mathbf{x})^\top \mathbf{A} d\mathbf{x})^\top = (d\mathbf{x})^\top \mathbf{A}^\top (\mathbf{b} - \mathbf{A}\mathbf{x})$$



Lo anterior permite notar que

$$d\phi(\mathbf{x}) = -2(\mathbf{b} - \mathbf{Ax})^\top \mathbf{A} d\mathbf{x},$$

y usando el primer Teorema de identificación,

$$\frac{\partial\phi(\mathbf{x})}{\partial\mathbf{x}} = -2\mathbf{A}^\top(\mathbf{b} - \mathbf{Ax})$$

Desde la condición de primer orden $\partial\phi(\mathbf{x})/\partial\mathbf{x} = \mathbf{0}$, obtenemos

$$\mathbf{A}^\top(\mathbf{b} - \mathbf{Ax}) = \mathbf{0} \quad \implies \quad \mathbf{A}^\top \mathbf{Ax} = \mathbf{Ab}.$$

Cómo \mathbf{A} tiene rango columna completo, el sistema de ecuaciones tiene solución única, dada por

$$\hat{\mathbf{x}} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b} = \mathbf{A}^+ \mathbf{b}. \quad (1)$$



Calculando el segundo diferencial,

$$d^2 \phi(\mathbf{x}) = -2 d(\mathbf{b} - \mathbf{A}\mathbf{x})^\top \mathbf{A} d\mathbf{x} = 2(d\mathbf{x})^\top \mathbf{A}^\top \mathbf{A} d\mathbf{x},$$

por el segundo Teorema de identificación, tenemos que la matriz Hessiana adopta la forma:

$$\frac{\partial^2 \phi(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} = 2\mathbf{A}^\top \mathbf{A},$$

que es definida positiva (para cualquier \mathbf{x}) y por tanto $\hat{\mathbf{x}}$ es mínimo global para ϕ .

Observación:

$\hat{\mathbf{x}}$ dado en Ecuación (1) es conocido como ***solución mínimos cuadrados***.

