IECD-325: Diferenciación matricial

Felipe Osorio

felipe.osorio@uv.cl

Notación:

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

Ejemplo:

Podemos escribir los siguientes casos particulares:

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

Diferenciación matricial

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

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

de este modo, introducimos la notación

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

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

$$\mathsf{D}\,\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \mathsf{D}\,f_1(\boldsymbol{x}) \\ \vdots \\ \mathsf{D}\,f_m(\boldsymbol{x}) \end{pmatrix} = \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}^\top},$$

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

Considere la fórmula de Taylor de primer orden,

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

donde,

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

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

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

De este modo, se define

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

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

Definición 1:

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

$$\boldsymbol{f}(\boldsymbol{c}+\boldsymbol{u}) = \boldsymbol{f}(\boldsymbol{c}) + \boldsymbol{A}(\boldsymbol{c})\boldsymbol{u} + \boldsymbol{r}_c(\boldsymbol{u}),$$

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

$$\lim_{u\to 0} \frac{\boldsymbol{r}_c(\boldsymbol{u})}{||\boldsymbol{u}||} = \boldsymbol{0},$$

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

$$\mathsf{d}\, \boldsymbol{f}(\boldsymbol{c};\boldsymbol{u}) = \boldsymbol{A}(\boldsymbol{c})\boldsymbol{u},$$

se denomina primer diferencial de $\boldsymbol{f}$ en $\boldsymbol{c}$ con incremento $\boldsymbol{u}$.

Resultado 1 (Magnus y Neudecker, 1985)[1]:

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

$$\mathsf{d}\, \boldsymbol{f}(\boldsymbol{c}; \boldsymbol{u}) = (\mathsf{D}\, \boldsymbol{f}(\boldsymbol{c}))\boldsymbol{u}.$$

La matriz $\mathsf{D}\, \boldsymbol{f}(\boldsymbol{c}) \in \mathbb{R}^{m \times n}$ se denomina matriz Jacobiana. Tenemos también que

$$\nabla \boldsymbol{f}(\boldsymbol{c}) = (\mathsf{D}\, \boldsymbol{f}(\boldsymbol{c}))^\top$$

es la matriz gradiente de $\boldsymbol{f}$.

[1] Journal of Mathematical Psychology **29**, 474-492.
También es conocido como el "Primer teorema de identificación"

Definición 2:

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

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

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

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

Definición 3:

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

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

Resultado 2:

Sean $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ y $\boldsymbol{D}$ matrices de órdenes apropiados y λ escalar. Entonces

(a) $\boldsymbol{A} \otimes \boldsymbol{B} \otimes \boldsymbol{C} = (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C})$,

(b) $(\boldsymbol{A} + \boldsymbol{B}) \otimes (\boldsymbol{C} + \boldsymbol{D}) = \boldsymbol{A} \otimes \boldsymbol{C} + \boldsymbol{B} \otimes \boldsymbol{C} + \boldsymbol{A} \otimes \boldsymbol{D} + \boldsymbol{B} \otimes \boldsymbol{D}$,

(c) $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{A}\boldsymbol{C} \otimes \boldsymbol{B}\boldsymbol{D}$,

(d) $\lambda \otimes \boldsymbol{A} = \lambda \boldsymbol{A} = \boldsymbol{A} \otimes \lambda$,

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

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

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

Resultado 3:

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

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

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

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

Observación:

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

$$\boldsymbol{a}\boldsymbol{b}^\top = \boldsymbol{a} \otimes \boldsymbol{b}^\top = \boldsymbol{b}^\top \otimes \boldsymbol{a}.$$

Por otro lado, tenemos que

$$\operatorname{vec}(\boldsymbol{a}\boldsymbol{b}^\top) = \operatorname{vec}(\boldsymbol{a} \otimes \boldsymbol{b}^\top) = \operatorname{vec}(\boldsymbol{b}^\top \otimes \boldsymbol{a}) = \boldsymbol{b} \otimes \boldsymbol{a}.$$

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

Resultado 4:

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

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

(b) Si $\boldsymbol{A}, \boldsymbol{B}$ y $\boldsymbol{C}$ son de órdenes adecuados, entonces

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

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

Resultado 5:

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

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

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

$$\operatorname{vec} \boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{f}(\operatorname{vec} \boldsymbol{X})$$

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

$$\operatorname{vec} \mathsf{d}\, \boldsymbol{F}(\boldsymbol{C}; \boldsymbol{U}) = \mathsf{d}\, \boldsymbol{f}(\operatorname{vec} \boldsymbol{C}; \operatorname{vec} \boldsymbol{U})$$

en cuyo caso $\boldsymbol{F}$ tiene matriz Jacobiana

$$\mathsf{D}\, \boldsymbol{F}(\boldsymbol{C}) = \mathsf{D}\, \boldsymbol{f}(\operatorname{vec} \boldsymbol{C})$$

Resultado 6:

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

$$\operatorname{vec} \mathsf{d}\, \boldsymbol{F}(\boldsymbol{C}; \boldsymbol{U}) = (\mathsf{D}\, \boldsymbol{F}(\boldsymbol{C})) \operatorname{vec} \boldsymbol{U}.$$

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

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

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

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

$$\mathsf{d}^2\,\phi = \mathsf{d}(\mathsf{d}\,\phi).$$

Resultado 6:

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

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

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

Observación:

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

- Sea $\boldsymbol{f}(\boldsymbol{x})$ función vectorial $m \times 1$ con argumento $\boldsymbol{x}$, vector n-dimensional, entonces
$$\mathsf{D}\,\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^{m \times n} \qquad \text{sin embargo,} \qquad \mathsf{d}\,\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^{m}$$
- Para funciones matriciales, $\mathsf{d}\,\boldsymbol{F}(\boldsymbol{X})$ tiene la misma dimensión que $\boldsymbol{F}$ sin importar la dimensión de $\boldsymbol{X}$.

Reglas fundamentales:

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

$$\begin{aligned}
&\mathsf{d}\,\alpha = 0, && \mathsf{d}(\alpha u) = \alpha\,\mathsf{d}\,u,\\
&\mathsf{d}(u+v) = \mathsf{d}\,u + \mathsf{d}\,v, && \mathsf{d}(uv) = (\mathsf{d}\,u)v + u(\mathsf{d}\,v)\\
&\mathsf{d}(u/v) = \frac{(\mathsf{d}\,u)v - u(\mathsf{d}\,v)}{v^2}, (v \neq 0), && \mathsf{d}\,u^{\alpha} = \alpha u^{\alpha-1}\,\mathsf{d}\,u,\\
&\mathsf{d}\,e^{u} = e^{u}\,\mathsf{d}\,u, && \mathsf{d}\log u = u^{-1}\,\mathsf{d}\,u, (u > 0)\\
&\mathsf{d}\,\alpha^{u} = \alpha^{u}\log\alpha\,\mathsf{d}\,u, (\alpha > 0).
\end{aligned}$$

Aquí por ejemplo,

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

Reglas fundamentales:

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

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

Otros diferenciales de uso frecuente en Estadística son:

$$
\mathsf{d}\,|\boldsymbol{F}| = |\boldsymbol{F}| \operatorname{tr} \boldsymbol{F}^{-1}\,\mathsf{d}\,\boldsymbol{F}, \qquad \mathsf{d}\log|\boldsymbol{F}| = \operatorname{tr} \boldsymbol{F}^{-1}\,\mathsf{d}\,\boldsymbol{F},
$$

$$
\mathsf{d}\,\boldsymbol{F}^{-1} = -\boldsymbol{F}^{-1}(\mathsf{d}\,\boldsymbol{F})\boldsymbol{F}^{-1}.
$$

Diferenciación matricial

Ejemplo (Mínimos cuadrados):

Considere el problema de optimización

$$\min_{x} \phi(\boldsymbol{x}), \qquad \phi(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2,$$

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

Diferenciando ϕ con relación a $\boldsymbol{x}$, tenemos

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

Notando que

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

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

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

Evidentemente,

$$((\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^\top\boldsymbol{A}\,\mathsf{d}\,\boldsymbol{x})^\top = (\mathsf{d}\,\boldsymbol{x})^\top\boldsymbol{A}^\top)(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$$

Lo anterior permite notar que

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

y usando el primer Teorema de identificación,

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

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

$$\boldsymbol{A}^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = \mathbf{0} \qquad \Longrightarrow \qquad \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{b}.$$

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

$$\widehat{\boldsymbol{x}} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{b} = \boldsymbol{A}^+\boldsymbol{b}. \tag{1}$$

Calculando el segundo diferencial,

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

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

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

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

Observación:

$\widehat{\boldsymbol{x}}$ dado en Ecuación (1) es conocido como solución mínimos cuadrados.