Apéndice B

Diferenciación matricial

En esta sección haremos uso de la siguiente notación. ϕ, $\boldsymbol{f}$ y $\boldsymbol{F}$ representan funciones escalar, vectorial y matricial, respectivamente mientras que ζ, $\boldsymbol{x}$ y $\boldsymbol{X}$ argumentos escalar, vectorial y matricial, respectivamente.

A partir de esta convención es directo que podemos escribir los siguientes casos particulares:

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

Existen varias definiciones para la derivada de una función matricial $\boldsymbol{F}(\boldsymbol{X})$ con relación a su argumento (matricial) $\boldsymbol{X}$. En este apéndice nos enfocamos en el cálculo diferencial propuesto por Magnus y Neudecker (1985).

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 $\boldsymbol{f}(\boldsymbol{x})$.

B.1. Aproximación de primer orden

Considere la fórmula de Taylor de primer orden,

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

donde el *resto*

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

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 B.1 (Diferencial de una función vectorial). 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 $\boldsymbol{u}$.

Magnus y Neudecker (1985) mostraron la existencia y unicidad del diferencial $\mathsf{d}\,\boldsymbol{f}(\boldsymbol{c};\boldsymbol{u})$ de una función $\boldsymbol{f} : S \to \mathbb{R}^m$, $S \subset \mathbb{R}^n$ ($\boldsymbol{c} \in S$), dado por

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

también mostraron la regla de la cadena e invarianza de Cauchy para el diferencial y enunciaron su primer teorema de identificación.

TEOREMA B.2 (Primer teorema de identificación). *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 $\boldsymbol{f}$.

Sea $\boldsymbol{f} : S \to \mathbb{R}^m$, $S \subset \mathbb{R}^n$ y $f_i : S \to \mathbb{R}$ el i-ésimo componente de $\boldsymbol{f}$ $(i = 1,\dots,m)$. Sea $\boldsymbol{e}_j$ un vector n-dimensional cuyo j-ésimo elemento es uno y los restantes son cero, y considere

$$\lim_{t\to 0} \frac{f_i(\boldsymbol{c}+t\boldsymbol{e}_j) - f_i(\boldsymbol{c})}{t}$$

si el límite existe, se denomina la j-ésima *derivada parcial* de f_i en $\boldsymbol{c}$ y es denotada por $\mathsf{D}_j f_i(\boldsymbol{c})$. Note que el elemento ij de $\mathsf{D}\boldsymbol{f}(\boldsymbol{c})$ es $\mathsf{D}_j f_i(\boldsymbol{c})$.

B.2. Funciones matriciales

Considere algunos ejemplos de funciones matriciales

$$\boldsymbol{F}(\zeta) = \begin{pmatrix} \cos(\zeta) & \sin(\zeta) \\ -\sin(\zeta) & \cos(\zeta) \end{pmatrix}, \quad \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^\top, \quad \boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{X}^\top, \quad \boldsymbol{X} \in \mathbb{R}^{n\times q}.$$

Antes de considerar el diferencial de una función matricial $\boldsymbol{F} : S \to \mathbb{R}^{m\times p}$, $S \subset \mathbb{R}^{n\times q}$ introducimos dos conceptos preliminares: la vectorización de una matriz y el producto Kronecker.

Definición B.3 (Operador de vectorización). 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 B.4 (Producto Kronecker). 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 B.5. *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 B.6. *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) $\operatorname{rg}(\boldsymbol{A}\otimes\boldsymbol{B}) = \operatorname{rg}(\boldsymbol{A})\operatorname{rg}(\boldsymbol{B})$.

Observe que, 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

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

Estos resultados sugieren una conexión entre el operador de vectorización, el producto Kronecker y la traza. Considere el siguiente resultado

Resultado B.7.

(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*

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

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

Finalmente, tenemos el siguiente resultado

RESULTADO B.8. *Sean $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ y $\boldsymbol{D}$ matrices, tal que, el producto $\boldsymbol{ABCD}$ está definido y es cuadrado, entonces*

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

En los ejemplos introducidos al inicio de esta sección, tenemos

$$\begin{aligned} \operatorname{vec} \boldsymbol{F}(\zeta) &= (\cos(\zeta), -\sin(\zeta), \sin(\zeta), \cos(\zeta))^\top, \\ \operatorname{vec} \boldsymbol{F}(\boldsymbol{x}) &= \operatorname{vec}(\boldsymbol{x}\boldsymbol{x}^\top) = \boldsymbol{x} \otimes \boldsymbol{x}, \\ \operatorname{vec} \boldsymbol{F}(\boldsymbol{X}) &= \operatorname{vec} \boldsymbol{X}^\top = (\boldsymbol{X} \otimes \boldsymbol{I}_q) \operatorname{vec} \boldsymbol{I}_q. \end{aligned}$$

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

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

Las consideraciones anteriores motivan el primer teorema de indentificación para funciones matriciales Magnus y Neudecker (1985)

TEOREMA B.9 (Primer teorema de identificación para funciones matriciales). *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 $\boldsymbol{F}$.

B.3. Matriz Hessiana

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

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

Es posible definir el diferencial de funciones vectoriales y matriciales de manera análoga a la delineada anteriormente. Sin embargo, en este apéndice nos enfocaremos solamente en el cálculo de diferenciales de funciones escalares. El segundo diferencial de una función escalar está dado por

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

Magnus y Neudecker (1985) enunciaron el siguiente teorema de identificación para matrices Hessianas de funciones escalares

TEOREMA B.10 (Segundo teorema de identificación). *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.

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 $\boldsymbol{X}$.

B.4. Reglas fundamentales

A continuación se presentan algunas reglas fundamentales para el cálculo de diferenciales
Considere u y v funciones escalares y α una constante, entonces:
$$\begin{aligned}
&\mathsf{d}\,\alpha = 0, \qquad \mathsf{d}(\alpha u) = \alpha\,\mathsf{d}\,u, \quad \mathsf{d}(u+v) = \mathsf{d}\,u + \mathsf{d}\,v,\\
&\mathsf{d}(uv) = (\mathsf{d}\,u)v + u(\mathsf{d}\,v) \qquad \mathsf{d}(u/v) = \frac{(\mathsf{d}\,u)v - u(\mathsf{d}\,v)}{v^2}, (v \neq 0),\\
&\mathsf{d}\,u^{\alpha} = \alpha u^{\alpha-1}\,\mathsf{d}\,u, \qquad \mathsf{d}\,e^{u} = e^{u}\,\mathsf{d}\,u,\\
&\mathsf{d}\log u = u^{-1}\,\mathsf{d}\,u, (u > 0) \qquad \mathsf{d}\,\alpha^{u} = \alpha^{u}\log\alpha\,\mathsf{d}\,u, (\alpha > 0),
\end{aligned}$$
aquí por ejemplo,
$$\phi(x) = u(x) + v(x).$$
Análogamente para $\boldsymbol{U}$, $\boldsymbol{V}$ funciones matriciales, α un escalar (constante) y $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ constante, tenemos
$$\begin{gathered}
\mathsf{d}\,\boldsymbol{A} = \boldsymbol{0}, \qquad \mathsf{d}(\alpha\boldsymbol{U}) = \alpha\,\mathsf{d}\,\boldsymbol{U},\\
\mathsf{d}(\boldsymbol{U}+\boldsymbol{V}) = \mathsf{d}\,\boldsymbol{U} + \mathsf{d}\,\boldsymbol{V}, \quad \mathsf{d}(\boldsymbol{U}\boldsymbol{V}) = (\mathsf{d}\,\boldsymbol{U})\boldsymbol{V} + \boldsymbol{U}\,\mathsf{d}\,\boldsymbol{V},\\
\mathsf{d}(\boldsymbol{U}\otimes\boldsymbol{V}) = \mathsf{d}\,\boldsymbol{U}\otimes\mathsf{d}\,\boldsymbol{V}, \qquad \mathsf{d}(\boldsymbol{U}\odot\boldsymbol{V}) = \mathsf{d}\,\boldsymbol{U}\odot\mathsf{d}\,\boldsymbol{V},\\
\mathsf{d}\,\boldsymbol{U}^{\top} = (\mathsf{d}\,\boldsymbol{U})^{\top}, \quad \mathsf{d}\operatorname{vec}\boldsymbol{U} = \operatorname{vec}\mathsf{d}\,\boldsymbol{U}, \quad \mathsf{d}\operatorname{tr}\boldsymbol{U} = \operatorname{tr}\mathsf{d}\,\boldsymbol{U}.
\end{gathered}$$
Otros diferenciales de uso frecuente en Estadística son:
$$\begin{gathered}
\mathsf{d}\,|\boldsymbol{F}| = |\boldsymbol{F}|\operatorname{tr}\boldsymbol{F}^{-1}\,\mathsf{d}\,\boldsymbol{F}, \qquad \mathsf{d}\log|\boldsymbol{F}| = \operatorname{tr}\boldsymbol{F}^{-1}\,\mathsf{d}\,\boldsymbol{F},\\
\mathsf{d}\,\boldsymbol{F}^{-1} = -\boldsymbol{F}^{-1}(\mathsf{d}\,\boldsymbol{F})\boldsymbol{F}^{-1}.
\end{gathered}$$