GEOMETRÍA E INFORMACIÓN

OPTATIVO

Mariela Adelina Portesi
Pedro Walter Lamberti
Steeve Zozor

Versión completa del 28 de octubre de 2019

Facultad de Ciencias Exactas

UNIVERSIDAD NACIONAL DE LA PLATA

Editorial de la Universidad de La Plata

Esto es una dedicatoria
del libro.

Agradecimientos

Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla.

Esto es un epígrafe con texto simulado.
Esto es un epgrafe con texto simulado.
AUTOR DEL EPÍGRAFE, TÍTULO DE LA OBRA

PRÓLOGO

Este libro surge de la experiencia de los autores en el dictado del curso semestral “Métodos de geometría diferencial en teoría de la información”, que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. **. . . acabar**

Los autores

ADVERTENCIA

Este libro surge de la experiencia de los autores en el dictado del curso semestral “Métodos de geometría diferencial en teoría de la información”, que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. **. . . acabar**

Mariela A. Portesi

Grenoble, Junio de 2016

Índice

Referencias

NOTACIONES

Operaciones, conjuntos & números

$\mid$	"tal que".
$\vee$	"o", *i. e.,* $A \ \vee \ B$ signifca tener A o B no necesariamente exclusivamente.
$\wedge$	"y", *i. e.,* $A \ \wedge \ B$ signifca tener a la vez A y B.
$\exists$	"existe".
$\exists!$	"existe un único".
$\forall$	"para todo los/las".
$\{\cdots\}$	Conjunto
$\in$	"es un elemento de" o "partenece a" $x \in A$ significa que el elemento x partenece al conjunto A.
$\notin$	"no es elemento de" o "no partenece a".
$\subset$	"es incuido en" $B \subset A$ significa que el conjunto B es adentro del conjunto A: $x \in B \Rightarrow x \in A$.
$\setminus$	Conjunto privado de: $A \setminus B = \{x \in A \mid x \notin B\}$.
$\cup$	Union: $A \cup B = \{x \mid (x \in A) \ \vee \ (x \in B)\}$.
$\bigcup_{i=1}^{n}$	Union de conjuntos: $\bigcup_{i=1}^{n} A_i = A_1 \cup \cdots \cup A_n$.
$\cap$	Intersección: $A \cap B = \{x \mid (x \in A) \ \wedge \ (x \in B)\}$.
$\bigcap_{i=1}^{n}$	Intersección de conjuntos: $\bigcap_{i=1}^{n} A_i = A_1 \cap \cdots \cap A_n$.
$\times$	Producto cartesiano $x \in A \times B \Leftrightarrow x = (a, b)$ con $a \in A$ y $b \in B$.
$\bigtimes_{i=1}^{n}$	Productos cartesianos: $\bigtimes_{i=1}^{n} A_i = A_1 \times \cdots \times A_n$.
$\emptyset$	Conjunto vacío, a veces denotado también $\{\}$.
$\mathbb{N}$	Enteros naturales.
$\mathbb{Z}$	Enteros relativos.
$\mathbb{Q}$	Números racionales.

$\mathbb{R}$	Números reales.
$\mathbb{C}$	Números complejos.
$\mathbb{K}^*$	Conjunto $\mathbb{K}$ privado de 0, $\mathbb{K}^* = \{x \in \mathbb{K} \mid x \neq 0\}$ ($\mathbb{K} = \mathbb{N}, \mathbb{Z}, \mathbb{R}$ o $\mathbb{C}$).
$\mathbb{K}_+$	Elementos de $\mathbb{K}$ no negativos, $\mathbb{K}_+ = \{x \in \mathbb{K} \mid x \geq 0\}$ ($\mathbb{K} = \mathbb{N}, \mathbb{Z}$ o $\mathbb{R}$).
$\mathbb{K}^d = \underbrace{\mathbb{K}\times\cdots\times\mathbb{K}}_{d \text{ veces}}$	Se usará la misma notación para los vectores, $M_{d,1}(\mathbb{K}) \equiv \mathbb{K}^d$ (ver más abajo): $x \in \mathbb{K}^d \Leftrightarrow x = (x_1, \ldots, x_d) \equiv \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$ con $x_i \in \mathbb{K}$.
$[a\,;\,b]$	Intervalo real cerrado en a y b, $[a\,;\,b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$
$(a\,;\,b)$	Intervalo real abierto en a y b, $(a\,;\,b) = \{x \in \mathbb{R} \mid a < x < b\}$
$[a\,;\,b)$	Intervalo real cerrado en a y abierto en b, $[a\,;\,b) = \{x \in \mathbb{R} \mid a \leq x < b\}$
$(a\,;\,b]$	Intervalo real abierto en a y cerrado en b, $(a\,;\,b] = \{x \in \mathbb{R} \mid a < x \leq b\}$
$\imath$	Imaginario puro, $\imath = \sqrt{-1}$
$\Re e\,\{\cdot\}$	Parte real de un complejo (componente a componente sobre $\mathbb{C}^d$)
$\Im m\,\{\cdot\}$	Parte imaginaria de un complejo (componente a componente sobre $\mathbb{C}^d$)
$\cdot^*$	Conjugación compleja (componente a componente sobre $\mathbb{C}^d$)
$\lvert\cdot\rvert$	Valor absoluto o modulus tratando de un real o complejo Cardenal (posiblemente infinito) tratando de un conjunto discreto Volumen (posiblemente infinito) tratando de un conjunto continuo
$\lVert\cdot\rVert_p$	Norma p, $\lVert x\rVert_p = \left(\sum_{i=1}^{d} \lvert x\rvert^p\right)^{\frac{1}{p}}$.
$\lVert\cdot\rVert$ o $\lvert\cdot\rvert$	Norma 2.
$\mathbb{S}_d$	Esfera unitaria de $\mathbb{R}^d$, $\mathbb{S}_d = \left\{x \in \mathbb{R}^d \mid \lVert x\rVert = 1\right\}$

$\mathbb{S}_d(c,r)$	Esfera de $\mathbb{R}^d$, centrada en $c \in \mathbb{R}^d$ y de rayo r $\mathbb{S}_d(c,r) = \left\{x \in \mathbb{R}^d \;\middle|\; \|x-c\| = r\right\}$, $\mathbb{S}_d \equiv \mathbb{S}_d(0,1)$
$\mathbb{SC}_d$	Esfera unitaria de $\mathbb{C}^d$, $\mathbb{SC}_d = \left\{z \in \mathbb{C}^d \;\middle|\; \|z\| = 1\right\}$
$\mathbb{B}_d$	Bola abierta unitaria de $\mathbb{R}^d$, $\mathbb{B}_d = \left\{x \in \mathbb{R}^d \;\middle|\; \|x\| < 1\right\}$
$\mathbb{B}_d(c,r)$	Bola abierta de $\mathbb{R}^d$, centrada en $c \in \mathbb{R}^d$ y de rayo r $\mathbb{B}_d(c,r) = \left\{x \in \mathbb{R}^d \;\middle|\; \|x-c\| < r\right\}$, $\mathbb{B}_d \equiv \mathbb{B}_d(0,1)$
Δ_{k-1}	$(k-1)$-simplex estandar [1] de $\mathbb{R}^k_+$, o simplex de probabilidad, i.e., $\Delta_{k-1} = \left\{x \in \mathbb{R}^k_+ \;\middle|\; \sum_{i=1}^{k} x_i = 1\right\}$
$\mathrm{P}_{k,d}$	Partición (con las permutaciones) de $k \in \mathbb{N}$ de tamaño $d \in \mathbb{N}^*$, $\mathrm{P}_{k,d} = \left\{n = \begin{bmatrix} n_1 & \cdots & n_d \end{bmatrix}^t \in \mathbb{N}^d \;\middle|\; \sum_{i=1}^{d} n_i = k\right\}$
$\mathfrak{S}_n$	Conjunto de las permutaciones de $\{1, \ldots, n\}$
$C^k(I)$	Conjunto de las functiones $f : \mathbb{R} \mapsto \mathbb{R}$ k veces diferenciable y de derivadas continuas
$\mathfrak{P}_d$	Conjuntos (convexo) de las funciones $f : \mathbb{R}^d \mapsto \mathbb{C}$ continuas, a simetría hermítica $f(-x) = f^*(x)$, definida no negativas, $\forall\, n \in \mathbb{N}^*,\ a_i \in \mathbb{C}, x_i \in \mathbb{R}^d,\ i = 1, \ldots, n,\ \sum_{i,j=1}^{n} a_i a_j^* f(x_j - x_i) \geq 0$, con $f(0) = 1$.
$\mathfrak{P}$	Conjuntos (convexo) $\bigcap_{d=1}^{+\infty} \mathfrak{P}_d$ de todas las funciones continuas, a simetría hermítica, definida positivas, de valor unidad al origen.
$\mathfrak{EP}_d$	Conjuntos (convexo) de las funciones $f : \mathbb{R}_+ \mapsto [-1\,;\,1]$ continuas con $f(0) = 1$ tal que $F : x \mapsto f\left(\|x\|^2\right) \in \mathfrak{P}_d$.
$\mathfrak{EP}$	Conjuntos (convexo) $\bigcap_{d=1}^{+\infty} \mathfrak{EP}_d$ de las funciones $f : \mathbb{R}_+ \mapsto [-1\,;\,1]$ continuas con $f(0) = 1$ tal que $F : x \mapsto f\left(\|x\|^2\right) \in \mathfrak{P}$.
$\mathfrak{CM}_n$	Conjuntos (convexo) de las funciones $f \in C^{n-1}(\mathbb{R}_+)$ tales que $\forall\, k = 0, \ldots, n-1,\ (-1)^k f^{(k)} \geq 0$ y $(-1)^{n-1} f^{(n-1)}$ es decreciente. Dicho de otra manera, cuando $n \geq 2$, $(-1)^k f^{(k)}$ es convexa, $k = 0, \ldots, n-2$.
$\mathfrak{CM}$	Conjuntos (convexo) $\bigcap_{n=0}^{+\infty} \mathfrak{CM}_n$ de las funciones $f \in C^{\infty}(\mathbb{R}_+)$ tales que $\forall\, k \in \mathbb{N},\ (-1)^k f^{(k)} \geq 0$.

[1] **Politopio, convex hull $\{\mathbb{1}_i\}_{i=1}^k$ (ver notaciones matriciales pagina ??).**

Vectores, matrices, tensores

$\mathcal{M}_d(\mathbb{K}) \equiv \mathbb{K}^d$	Espacio de vector $M = \begin{bmatrix} M_1 \\ \vdots \\ M_d \end{bmatrix}$ con $M_i \in \mathbb{K}$ y $\mathbb{K} = \mathbb{R}$ o $\mathbb{C}$.
$\mathcal{M}_{d,d'}(\mathbb{K})$	Espacio de matrices M, de tamaño $d \times d'$, de componentes $M_{i,j} \in \mathbb{K}$; $\mathcal{M}_{d,1} \equiv \mathbb{K}^d$ así que se usará esta notación cuando [2] $d, d' > 1$.
$\mathcal{M}_{d_1,\ldots,d_K}(\mathbb{K})$	Espacio de tensores [3] M, de orden K, de tamaño $d_1 \times \cdots \times d_K$, de componentes $M_{i_1,\ldots,i_K} \in \mathbb{K}$; De nuevo, con esta notación se supone que $d_i > 1, i = 1, \ldots, K$.
$.^*$	Conjugación, componente por componente, M^* es de componentes $M^*_{i_1,\ldots,i_K}$
$.^t$	Transpuesta, $\forall M \in \mathcal{M}_d(\mathbb{K})$, $M^t = \begin{bmatrix} M_1 & \cdots & M_d \end{bmatrix}$ y $\forall M \in \mathcal{M}_{d,d'}(\mathbb{K})$, M^t es de componente (i,j)-ésima $M_{i,j}$
$.^\dagger$	Transconjugada, $M^\dagger = (M^*)^t$
$\Re e\{\cdot\}$	Parte real, $\Re e\{M\}$ es de componente $\Re e\{M_{i_1,\ldots,i_K}\}$
$\Im m\{\cdot\}$	Parte imaginaria, $\Im m\{M\}$ es de componente $\Im m\{M_{i_1,\ldots,i_K}\}$
$\otimes$	Producto tensorial o producto externo [4]: para $A \in \mathcal{M}_{d_1,\ldots,d_K}(\mathbb{K})$, $B \in \mathcal{M}_{d'_1,\ldots,d'_K}(\mathbb{K})$ el producto externo $A \bigtimes B$ es el tensor de orden $K + L$ de componente $(i_1, \ldots, i_K, j_1, \ldots, j_L)$-ésima $A_{i_1,\ldots,i_K} B_{j_1,\ldots,j_L}$. Nota: para $a \in \mathcal{M}_d(Kset)$, $b \in \mathcal{M}_{d'}(\mathbb{K})$, $a \otimes b = ab^t$.
$.^{\otimes k}$	k veces el producto externo, $A^{\otimes k} = \underbrace{A \otimes \cdots \otimes A}_{k \text{ veces}}$.
$S_d(\mathbb{K})$	Conjunto de matrices de $\mathbb{K}$ simetricas, $S_d(\mathbb{K}) = \{M \in \mathcal{M}_{d,d}(\mathbb{K}) \mid M^t = M\}$.
$H_d(\mathbb{C})$	Conjunto de matrices de $\mathbb{C}$ hermiticas (a simetría hermitica), $H_d(\mathbb{C}) = \{M \in \mathcal{M}d, d(\mathbb{C}) \mid M^\dagger = M\}$. Notar que se tiene $H_d(\mathbb{R}) \equiv S_d(\mathbb{R})$.

[2] Si una dimensión es 1, se saca de las dimensiones (caso degenerado).

[3] De hecho, el termino "tensor" tiene una significación física. El termino "tabla" sería más adecuado. Pero en este libro, usaremos tensor, del hecho que se usa frecuentemente, por abuso de denominación.

$P_d(\mathbb{K})$	Conjunto de matrices hermiticas semidefinida positivas: $P_d(\mathbb{K}) = \left\{M \in H_d(\mathbb{K}) \;\middle|\; \forall\, x \in \mathbb{K}^d,\; x^\dagger M x \geq 0\right\}.$
$P_d^+(\mathbb{K})$	Conjunto de matrices hermiticas definida positivas: $P_d^+(\mathbb{K}) = \left\{M \in H_d(\mathbb{K}) \;\middle|\; \forall\, x \neq 0 \in \mathbb{K}^d,\; x^\dagger M x > 0\right\}.$
$\mathcal{P}_{d,k}(\mathbb{K})$	Conjunto [5] de k-uplet de matrices de $P_d(\mathbb{K})$ sumando a la identidad [6], $\mathcal{P}_{d,k}(\mathbb{K}) = \left\{(M_1, \ldots, M_k) \in P_d(\mathbb{K})^k \;\middle|\; \sum_{i=1}^{k} M_i = I\right\}$
$\mathfrak{S}_d(\mathbb{K})$	Conjunto de las matrices de permutaciones de $\mathcal{M}_{d,d}(\mathbb{K})$ $\mathfrak{S}_d(\mathbb{K}) = \left\{\Pi = \left[\mathbb{1}_{\sigma(1)} \quad \cdots \quad \mathbb{1}_{\sigma(d)}\right]^t = \sum_{i=1}^{d} \mathbb{1}_i \mathbb{1}_{\sigma(i)}^t, \quad \sigma \in \mathfrak{S}_d\right\}$ Tiene exactamente un 1 en cada linea y en cada columna, y 0 en los otros componentes. Notar: $\forall\, \Pi \in \mathfrak{S}_d(\mathbb{K}), \quad \Pi^{-1} = \Pi^t.$
$\mathcal{P}_{d,k}^+(\mathbb{K})$	Conjunto [5] de k-uplet de matrices de $P_d^+(\mathbb{K})$ sumando a la identidad [6], $\mathcal{P}_{d,k}^+(\mathbb{K}) = \left\{(M_1, \ldots, M_k) \in P_d^+(\mathbb{K})^k \;\middle|\; \sum_{i=1}^{k} M_i = I\right\}$
$\geq$	$A \geq B$ significa que $(A - B) \in P_d(\mathbb{K})$
$>$	$A > B$ significa que $(A - B) \in P_d^+(\mathbb{K})$
$\mathbb{1}$	Vector de componentes iguales a 1, $\mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.
$\mathbb{1}_i$	Vector de componentes j-ésima iguales a $\mathbb{1}_{\{i\}}(j)$.
diag	Matriz diagonal con los componentes del vector argumento en su diagonal, para $v \in \mathbb{K}^d$, $\mathrm{diag}(v) = \sum_{i=1}^{d} \left(\mathbb{1}_i \mathbb{1}_i^t\right) v_i = \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & v_d \end{bmatrix}.$

[4] Se encuentra también el símbolo $\otimes$ para el producto de Kronecker: para matrices por ejemplo $A \otimes B$ será la matriz de componentes bloc $A_{i,j}B$. De hecho, es equivalente a "desarollar" el tensor de forma matricial, o equivalente de la vectorización de una matriz (ver ej. (Magnus & Neudecker, 1979)).

[5] **Este conjunto es un convexo de $P_d(\mathbb{K})^k$. Pero todavia no me queda claro que estructura tiene. Si no restringimos a matrices diagonales, es un simplex (?, ?, p. 329). Si no, no le se.**

[6] Se dice que el k-uplet satisface a la *resolución de la identidad*, o a la *relación de completud*.

0	Se notará el vector o una matriz de componentes ceros como el caso escalar, $0 \equiv \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$.
$I_d \quad , \quad I$	Matriz cuadrada $d \times d$, identidad, $I \equiv I_d = \mathrm{diag}(\mathbb{1})$ (se saca el índice cuando no hay ambiguedad)
$G^{(i,j)}$	Matriz de agregación de $\mathcal{M}_{k-1,k}(\mathbb{K})$. Para $i < j$, multiplicado a un vector $x \in \mathbb{K}^k$ se saca la j-ésima componente de x y se reemplaza la i-ésima por $x_i + x_j$ y similarmente por simetría para $i > j$: $\forall\, i < j, \quad G^{(i,j)} = \begin{bmatrix} I_{j-1} & \mathbb{1}_i & 0 \\ 0 & 0 & I_{k-j} \end{bmatrix}, \quad G^{(j,i)} = G^{(i,j)}$.
Tr	Traza de una matriz (cuadrada) de $\mathcal{M}_{d,d}(\mathbb{K})$, $\mathrm{Tr}\, M = \sum_{i=1}^{d} M_{i,i}$.
$\det$	Determinante [7] de una matriz (cuadrada) de $\mathcal{M}_{d,d}(\mathbb{K})$, $\det M = \sum_{\sigma \in \mathfrak{S}_d} \varepsilon(\sigma) \prod_{i=1}^{d} M_{\sigma(i),i}$ con $\varepsilon(\sigma) = \prod_{1 \le i \le j \le d} \mathrm{sign}\,(\sigma(j) - \sigma(i))$ signatura de la permutación σ.
$\lvert\cdot\rvert$	Valor absoluto del determinante de una matriz (cuadrada) de $\mathcal{M}_{d,d}(\mathbb{K})$.
$\cdot^{-1}$	Matriz inversa (cuando existe), $MM^{-1} = M^{-1}M = I$
$\cdot^{\frac{1}{2}}$	Para $M \in \mathcal{P}_d^+(\mathbb{K})$, $M^{\frac{1}{2}}$ es la única matriz de $\mathcal{P}_d^+(\mathbb{K})$ tal que $M^{\frac{1}{2}} M^{\frac{1}{2}} = M$ (Horn & Johnson, 2013; Magnus & Neudecker, 1999)
$\cdot^{-\frac{1}{2}}$	Para $M \in \mathcal{P}_d^+(\mathbb{K})$, $M^{-\frac{1}{2}} = \left(M^{-1}\right)^{\frac{1}{2}} = \left(M^{\frac{1}{2}}\right)^{-1}$ (Horn & Johnson, 2013; Magnus & Neudecker, 1999)
$\cdot^{-t}$	$M^{-t} = (M^t)^{-1} = \left(M^{-1}\right)^t$
$\cdot^{-*}$	$M^{-*} = (M^*)^{-1} = \left(M^{-1}\right)^*$
$\cdot^{-\dagger}$	$M^{-\dagger} = \left(M^\dagger\right)^{-1} = \left(M^{-1}\right)^\dagger$
$\lVert\cdot\rVert_p$	Norma p de Hölder, $\lVert M \rVert_p = \left(\sum_{i,j} \lvert M_{i,j} \rvert^p\right)^{\frac{1}{p}}$

[7] $\varepsilon(\sigma)$ es la signatura de la permutación σ, $\varepsilon(\sigma) = \prod_{1 \le i \le j \le d} \mathrm{sign}\,(\sigma(j) - \sigma(i))$ con sign función "signo".

$\|\cdot\|_F$	Norma de Frobenius, $\|M\|_F = \sqrt{\mathrm{Tr}\,(MM^\dagger)} = \|M\|_2$
$\|\cdot\|_{q,p}$	Norma de Hölder inducida, $\|M\|_{q,p} = \sup_{x \in \mathcal{M}_{d',1}(\mathbb{K})^*} \dfrac{\|Mx\|_q}{\|x\|_p}$.

En todo el libro, acordar la notacion con $\mathcal{M}$ para los conjuntos de matrices, normas...

Funciones

$\lfloor\cdot\rfloor$	Parte entera inferior, $\lfloor x \rfloor = \max\{n \in \mathbb{Z} \mid n \leq x\}$, *i. e.,* $x - 1 < \lfloor x \rfloor \leq x$.
$\lceil\cdot\rceil$	Parte entera superior, $\lceil x \rceil = \min\{n \in \mathbb{Z} \mid n \geq x\}$ *i. e.,* $x \leq \lceil x \rceil < x + 1$.
$\sum$	Suma de elementos, $\sum_{i=n}^{m} x_i = \begin{cases} x_n + x_{n+1} + \cdots + x_m & \text{si} \quad m \geq m \\ 0 & \text{si} \quad m < n \quad \text{(por convención)} \end{cases}$.
$\prod$	Producto de elementos, $\prod_{i=n}^{m} x_i = \begin{cases} x_n . x_{n+1} \ldots x_m & \text{si} \quad m \geq m \\ 1 & \text{si} \quad m < n \quad \text{(por convención)} \end{cases}$.
sign	Función signo, $\mathrm{sign}(x) = \begin{cases} -1 & \text{si} \quad x < 0 \\ 0 & \text{si} \quad x = 0 \\ 1 & \text{si} \quad x > 0 \end{cases}$.
log	Logaritmo natural (de base e).
$\log_b$	Logaritmo de base $b > 0$, $\log_b x = \dfrac{\log x}{\log b}$.
$\mathbb{1}_A$	Function indicadora del conjunto A: $\mathbb{1}_A(x) = \begin{cases} 1 & \text{si} \quad x \in A \\ 0 & \text{si} \quad x \notin A \end{cases}$
$(\,\cdot\,)_+$	$(x)_+ = \max(x, 0) = \mathbb{1}_{\mathbb{R}_+}(x)$
Γ , !	Función Gamma o factorial (Abramowitz & Stegun, 1970; Andrews, Askey & Roy, 1999; Gradshteyn & Ryzhik, 2015), $\Gamma(a) = \int_{\mathbb{R}_+} x^{a-1}\, e^{-x}\, dx, \quad a \in \mathbb{R}_+^* \qquad \forall\, n \in \mathbb{N}, \quad \Gamma(n+1) = n! = \prod_{i=1}^{n} i.$

Γ_d	Función gamma multivariada (Anderson, 2003; Gupta & Nagar, 1999) $\Gamma_d(x) = \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^{d} \Gamma\left(x - \frac{j-1}{2}\right), \quad x > \frac{d-1}{2}.$
$(x)^{\overline{n}}$	Factorial creciente o símbolo de Pochhammer usual (a veces denotado $x^{(n)}$) (Graham, Knuth & Patashnik, 1994), $(x)^{\overline{n}} = \prod_{i=0}^{n-1}(x+i)$ con la convención $(x)^{\overline{0}} = 1.$
$(x)^{\underline{n}}$	Factorial decreciente (a veces denotado $(x)_n$) (Graham et al., 1994), $(x)^{\underline{n}} = \prod_{i=0}^{n-1}(x-i)$ con la convención $(x)^{\underline{0}} = 1.$
$\binom{n}{k}$	Coefficiente binomial $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}.$
$B(\cdot,\cdot)$	Función beta $B(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$
$B(\cdot)$	Función de Dirichlet [8] (Andrews et al., 1999, Teo. 1.8.6) o (Gupta & Nagar, 1999) $B(a) = \frac{\prod_{i=1}^{k}\Gamma(a_i)}{\Gamma\left(\sum_{i=1}^{k} a_i\right)}, \quad a \in \mathbb{R}_+^{*\,k}.$
$B_d(\cdot)$	Función de Dirichlet multivariada (Gupta & Nagar, 1999) $B_d(a) = \frac{\prod_{i=1}^{k}\Gamma_d(a_i)}{\Gamma_d\left(\sum_{i=1}^{k} a_i\right)}, \quad a \in \left(\frac{d-1}{2}\,;+\infty\right)^k.$
${}_1F_1$	Función confluent hipergeométrica (Abramowitz & Stegun, 1970; Andrews et al., 1999; Gradshteyn & Ryzhik, 2015) ${}_1F_1(a;b;z) = \sum_{m\in\mathbb{N}} \frac{(a)^{\overline{m}}\, z^m}{(b)^{\overline{m}}\, m!}.$
${}_2F_1$	Función hipergeométrica (Abramowitz & Stegun, 1970; Andrews et al., 1999; Gradshteyn & Ryzhik, 2015) ${}_2F_1(a_1,a_1;b;z) = \sum_{m\in\mathbb{N}} \frac{(a_1)^{\overline{m}}\,(a_2)^{\overline{m}}\, z^m}{(b)^{\overline{m}}\, m!}.$
$\Phi_2^{(k)}$	Función confluent hipergeométrica k-variada o forma confluente de series de Lauricella (Srivastava & Karlsson, 1985, § 1.4, ec. (8)) o (Humbert, 1922; Appell, 1925; ?, ?, ?; Erdélyi, 1940). $\Phi_2^{(k)}(a;b;z) = \sum_{m\in\mathbb{N}^k} \frac{\prod_{i=1}^{k}(a_i)^{\overline{m_i}}\, z_i^{m_i}}{(b)^{\overline{m_1+\cdots+m_k}}\,\prod_{i=1}^{k} m_i!}$

[8] Eso es nada más que une beta generalizada a más de dos variables. Por eso, usamos la misma notación.

J_α	Función Bessel de primera especie y de orden α (Abramowitz & Stegun, 1970; Gradshteyn & Ryzhik, 2015; Watson, 1922; Gray & Mathew, 1895).
K_α	Función Bessel modificada de segunda especie y de orden α (Abramowitz & Stegun, 1970; Gradshteyn & Ryzhik, 2015; Watson, 1922; Gray & Mathew, 1895) o Función de MacDonald (MacDonald, 1898).
$\frac{\partial}{\partial x}$	Derivada parcial con respeto a la variable x.
$\nabla_x \quad , \quad .'$	Gradiante con respect a la variable x, para $f : \mathbb{R}^d \mapsto \mathbb{R}, \quad \nabla_x f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}$ Se denotará f' cuando $d = 1$.
J_f	Jacobiana, para $f : \mathbb{R}^d \mapsto \mathbb{R}^{d'}, \quad \mathrm{J}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_{d'}}{\partial x_1} & \cdots & \frac{\partial f_{d'}}{\partial x_d} \end{bmatrix}$, fíjense que si $d' = 1$, $\mathrm{J}_f^t \equiv \nabla_x f$.
$\mathcal{H}_x \quad , \quad .''$	Hessiana, para $f : \mathbb{R}^d \mapsto \mathbb{R}, \quad \mathcal{H}_x f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$. Se denotará f'' cuando $d = 1$.
$\mathcal{H}_{x,y}$	Hessiana "compuesta", para $f : \mathbb{R}^d \times \mathbb{R}^{d'} \mapsto \mathbb{R}, \quad \mathcal{H}_{x,y} f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial y_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial y_{d'}} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial y_1} & \cdots & \frac{\partial^2 f}{\partial x_d \partial y_{d'}} \end{bmatrix}$.
$f^{(k)}$	Derivada de orden $k \geq$ para funciones $f : \mathbb{R} \mapsto \mathbb{R}$.
$\circ$	Composición de funciones, $f \circ g(x) = f\left(g(x)\right)$.
f^{-1}	Inversión de f, $f^{-1}(y) = \{x \mid f(x) = y\}$ o inversa de un conjunto por f, $f^{-1}(B) = \{x \mid f(x) \in B\}$.

$\max_{\mathcal{X}}$	Máximo de una función real sobre el dominio $\mathcal{X}$: $\exists\, x_0 \in \mathcal{X} \mid \forall\, x \in \mathcal{X},\ f(x) \leq f(x_0) = \max_{\mathcal{X}} f.$ No siempre existe un máximo.
$\sup_{\mathcal{X}}$	Supremo de una función sobre el dominio $\mathcal{X}$, $\sup_{\mathcal{X}} f$ es el más pequeño real F_s tal que $\forall\, x \in \mathcal{X}, f(x) \leq F_s$. Siempre existe, pero no siempre se alcanza el supremo; cuando se alcanza $\sup_{\mathcal{X}} f = \max_{\mathcal{X}} f$.
$\min_{\mathcal{X}}$	Mínimo de una función sobre el dominio $\mathcal{X}$, $\exists\, x_0 \in \mathcal{X} \mid \forall\, x \in \mathcal{X},\ f(x) \geq f(x_0) = \min_{\mathcal{X}} f.$ No siempre existe un mínimo.
$\inf_{\mathcal{X}}$	Infimo de una función sobre el dominio $\mathcal{X}$, $\inf_{\mathcal{X}} f$ es el más grande real F_i tal que $\forall\, x \in \mathcal{X}, f(x) \geq F_i$. Siempre existe, pero no siempre se alcanza el infimo; cuando se alcanza $\inf_{\mathcal{X}} f = \min_{\mathcal{X}} f$.
argmáx	Operador argumento máximo, $\operatorname{argmáx}_{\mathcal{X}} f$ es el(los) $x \in \mathcal{X}$ que maximiza(n) f, $\operatorname{argmáx}_{\mathcal{X}} f = \{x_0 \in \mathcal{X} \mid f(x_0) = \max_{\mathcal{X}} f\}$.
argmín	Operador argumento mínimo, $\operatorname*{argmín}_{\mathcal{X}} f$ es el(los) $x \in \mathcal{X}$ que minimiza(n) f, $\operatorname{argmín}_{\mathcal{X}} f = \{x_0 \in \mathcal{X} \mid f(x_0) = \min_{\mathcal{X}} f\}$.

CAMBIAR EL POCHHAMMER EN LO QUE SIGUE con la macro

Erdélyi, A., Beitrag zur Theorie der konfluenten hypergeometrischen Funktionen von mehreren Veränderlichen, Wiener Sitzungsberichte 146 (1937), 431-467, ec. 7.2

P. Appell and J. Kampé de Fériet, Fonctions hypergéométriques et hypersphériques, Gauthier Villars, Paris, 1926, pp. 124-125, 116

H. Exton, Multiple hypergeometric functions and applications, Ellis Horwood, Chichester, U.K., 1976 P42

Probabilidad

Ω	Espacio fundamental.
μ	Medida.
E	Esperanza o promedio estadistico.
Cov	Covarianza.

pCov	Pseudo-covarianza.
Curt	Curtosis.
ExCurt	Curtosis por excesso.
Asim	Asimetría.
$\overset{\mathrm{d}}{=}$	Igualdad en distribucion: $X \overset{\mathrm{d}}{=} Y$ significa que X e Y tienen la misma distribución de probabilidad.
$\overset{\mathrm{d}}{\to}$	Límite en distribucion: $X_t \xrightarrow[t \to T]{\mathrm{d}} Y$ significa que la distribución de probabilidad de X_t tiende a la de Y cuando $t \to T$.

CAPÍTULO 1
Elementos de teoría de probabilidades

While writing my book I had an argument with Feller.
He asserted that everyone said "random variable"
and I asserted that everyone said "chance variable."
We obviously had to use the same name in our books,
so we decided the issue by a stochastic procedure.
That is, we tossed for it and he won.
J. L. DOOB, STATISTICAL SCIENCE (1953)

Anadir por lo menos un poco

- **Las nociones de convergencia (aparece en casos limites de leyes, en el TCL). Ver (Ash & Doléans-Dade, 1999, Cap. 2.8, 6), (Billingsley, 2012, Cap. 5), (Athreya & Lahiri, 2006, Sec. 9, p. 287), (Brockwell & Davis, 1987, Cap. 6), (Jacob & Protters, 2003, Caps. 17, 18).**
- **Cotas de Chernoff con la MGF o PGF**
- **Hablar del problema de Hamburger?**
- **teorema de Polya?**
- **Dibujar en el plano complejo Φ_X? Unas curvas son lindas.**
- **Ejemplos: von Mises y vonMises-Fisher? Cantor (singular...)?**
- **Hablar de simulación (inversion OK; mezcla? rejección? multivariada via la condicional?) VEr Kotz vol 2 por ejemplo**
- **Hablar de cuantiles, de fractiles, de modo (mode)**

1.1 Introducción

A pesar de que las nociones de azar (que proviene del árabe *zahr* que significa dado, flor) o de aleatoriedad (del latín *alea* que es suerte, dado) son muy antiguas (Serrano Marugán, 2000), el matemático italiano y jugador de dados y cartas Gerolamo Cardano es "probablemente" uno de los primeros en tratar matemáticamente el concepto de probabilidad en el siglo XVI, en su libro sobre los juegos de azar escrito en 1564 pero publicado en 1663 (Cardano, 1663) (ver (Bellhouse, 2005) o (Hald, 1990, Cap. 4)). La denominación de probabilidad, ella, viene de Aristote y designaba una percepción de una idea. Tomó su sentido más actual solamente durante la edad media en Europa, por mala traducción de la escritura de Aristote. Después de la primeras semillas, debido a Cardano, hay que mencionar los franceses Pierre de Fermat y Blaise Pascal en el medio del siglo XVII (Pascal, 1679) o (Hald, 1990, Cap. 5), y el neerlandese C. Huygens (Huygens, 1657) o (Hald, 1990, Cap. 6)), que fueron claramente unos de los primeros a desarrollar la teoría de las probabilidades. Más tarde, pasos importante fureon debidos al suizo Jacob Bernoulli (miembro de una dinastía de matemáticos) (Bernoulli, 1713, en latín) o ((E. D. Sylla, Translator), 1713; Hald, 1990; David & Edwards, 2001; Hald, 2006) y al franco-inglés Abraham de Moivre (de Moivre, 1756; Hald, 1990; David & Edwards, 2001; Hald, 2006). Hasta la época de Bernoulli, el enfoque era puramente discreto, es decir que el conjunto de estados posibles era discreto de tamaño finito (6 caras de un dado, 32 tarjetas, 2 caras de una moneda,...). La meta de la mayoría de los estudios eran dedicados a los juegos (dados, cartas), problema de seguro/riesgo, o estudios sociales en problaciones.

El francés Pierre Simon Laplace (Laplace, 1812, 1814; de Laplace, 1820; ?, ?) fue quizás uno de los primeros en proveer un aporte importante al desarrollo de la teoría de las probabilidades en los siglos XVIII-XIX, a través del punto de vista "frecuentista" y combinatorial (ver también (Hald, 1990, Caps. 13, 15 & 22) o (David & Edwards, 2001; Hald, 2006)). En la misma época, hay que mencionar C. F. Gauss, matemático muy prolífico, quien trabajó, entre muchas cosas, en la predicciíon de la trayectoría del planetisimo Céres (Gauss, 1809, 1810) o (Hald, 2006, Cap. 7). Proponiendo un error cuadratico, apareció implicitamente la ley Normal, o Gausiana, que tiene su nombre, a pesar de que la desarollo más Laplace (aun que, sobre el mismo problema, propuso el un error tipo L^1, vínculado a la ley doble-exponencial o de Laplace) (Laplace, 1809a, 1809b, 1812, 1814; de Laplace, 1820). A veces la ley de Gauss, quizas la más importante en la teoría de las probabilidades, llamada también gausiana o normal, es conocida como ley de Laplace-Gauss.

Un paso muy importante, especialmente tratanto de aleatoriadidad continua (ej. medida de una velocidad, que puede tomar cualquier valor real si tomamos en cuenta la dirección), fue debido entre otros a Kolmogorov en 1933 que se apoyó sobre trabajos de Richard von Mises (von Mises, 1932) y también sobre la teoría de la medida y de la integración, debidas entre otros a Émile Borel y Henri Lebesgue (Borel, 1898, 1909; Lebesgue, 1904, 1918; Halmos, 1950), para formalizar analíticamente la teoria de las probabilidades (Kolmogorov, 1956; Barone & Novikoff, 1978; Jacob & Protters, 2003).

Este punto de vista permite tratar formalmente el caso de variables discretas, continuas, o mezcla de ambas, que sean escalares o multivariadas, en un marco único y muy poderoso, sin perder las intuición que lleva el punto de vista frecuencista.

1.2 Probabilidades

El concepto de *probabilidad* es importante en situaciones donde el resultado de un dado proceso o medición es incierto, cuando la salida de una experiencia no es totalmente previsible. La probabilidad de un evento es una medida que se asocia con cuán probable es el evento o resultado.

Una definición de probabilidad se puede dar en base a la enumeración exhaustiva de los resultados posibles de un experimento o proceso, suponiendo que el conjunto de posibilidades es completo en el sentido de que una de ellas debe ocurrir o debe ser verdad. Si el proceso tiene K resultados distinguibles, mutuamente excluyentes e igualmente probables (esto es, no se prefiere una posibilidad frente a otras), y si k de esos K resultados tienen un dado atributo, la probabilidad asociada a dicho atributo en un dado proceso es $\frac{k}{K}$. Por ejemplo, sorteando un número entre los naturales del 1 al 10, la probabilidad de "obtener un número par" es $\frac{5}{10} = \frac{1}{2}$.

Otra definición de probabilidad se basa en la frecuencia relativa de ocurrencia de un evento. Si en una cantidad K muy grande de procesos independientes cierto atributo aparece k veces, se identifica a la probabilidad asociada a un proceso o ensayo con la frecuencia relativa de ocurrencia $\frac{k}{K}$ del atributo (van Brakel, 1976; Hald, 1990; David & Edwards, 2001; Shafer & Vovk, 2006, & Ref.).

Los axiomas de Kolmogorov proveen requisitos suficientes para determinar completamente las propiedades de la medida de probabilidad $P(A)$ que se puede asociar a un evento A entre un conjunto de resultados o eventos de un proceso.

Llamemos Ω al *espacio muestral* o *espacio fundamental*, que es el espacio de *muestras* (*outcomes*, en inglés) $\omega \in \Omega$. Se asocia $\mathcal{A}$ a una colección de sub-conjuntos de Ω, donde los elementos de $\mathcal{A}$ son llamados *eventos*. Por ejemplo, para un dado de 6 caras, Ω es el conjunto de caras que se pueden etiquetar con los números naturales del 1 al 6 (o también con las letras *a, b, c, d, e, f*, u otro etiquetado), y $\mathcal{A}$ tiene los eventos A "es un número natural par" y B "es un número natural impar". En el caso de analizar el tiempo de vida de un aparato, $\Omega \equiv \mathbb{R}_+$. El conjunto de resultados posibles se supone conocido, aún cuando se desconozca de antemano el resultado de una prueba.

Entre los eventos se pueden considerar operaciones y definiciones análogas a las de la teoría de conjuntos (ver entre otros (Spiegel, 1976; Brémaud, 1988; Mandel & Wolf, 1995; Sierpiński, 1975, 1976; Borel, 1898, 1909)):

- Combinación o unión de eventos: $A \cup B$ implica que se da A, ó B, o ambos (por ejemplo, para un dado de 6 caras, si A son los eventos "cara par" y B los eventos "cara menor o igual a 3", resulta $A \cup B = \{1\,;\,2\,;\,3\,;\,4\,;\,6\}$). Según la literatura, se denota a veces $A + B$ o $A \vee B$.

- Intersección de eventos: $A \cap B$ implica que se dan ambos A y B (en el ejemplo precedente, $A \cap B = \{2\}$). Se denota a veces (A, B) o $A \wedge B$.
- Complemento de un evento: $\bar{A}$ indica que no se da A. Se denota a veces $-A$ o $A^{\complement}$ (en el ejemplo precedente, $\bar{A} = \{1\,;\,3\,;\,5\}$).
- Eventos *disjuntos* o *mutuamente excluyentes* o *incompatibles*: son aquellos que no se superponen, se anota $A \cap B = \emptyset$ donde $\emptyset = \bar{\Omega}$ denota el *evento nulo* (evento que no puede ocurrir, es el complemento de Ω). Por ejemplo los eventos "cara par" y "cara impar" son incompatibles.
- Denotaremos también $A \setminus B$ cuando el evento A se realiza pero no B. Se lo denota también $A - B$, que es también $A \cap \bar{B}$ (en el ejemplo precedente, $A \setminus B = \{4\}$).

Esto es ilustrado en la Fig. 1-1 empleando lo que se conoce como diagramas de Venn [9]. La unión y la intersección de eventos satisfacen las mismas reglas que en la teoría de conjuntos, es decir cada una es conmutativa $A \cup B = B \cup A$, $A \cap B = B \cap A$, asociativa $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$, distributiva con respecto a la otra $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (ver por ejemplo (Jeffrey, 1948, 1973; Halmos, 1950; Feller, 1971; Brémaud, 1988; Mandel & Wolf, 1995; Ibarrola, Pardo & Quesada, 1997; Lehmann & Casella, 1998; Athreya & Lahiri, 2006; Cohn, 2013; Hogg, McKean & Craig, 2013)).

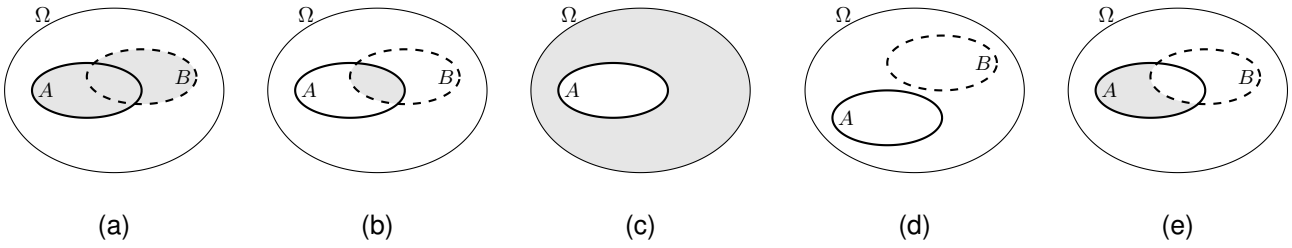


Figura 1-1: Ilustración de las operaciones entre eventos: (a) unión $A \cup B$, (b) intersección $A \cap B$, (c) complemento $\bar{A}$, (d) eventos excluyentes $A \cap B = \emptyset$, y (e) $A \setminus B$. A es representado en línea llena, B en línea discontinua; en (a)-(c) y (e), el resultado de la operación es la zona sombreada.

Formalmente, se define de manera abstracta un espacio medible $(\Omega, \mathcal{A})$ de la manera siguiente ((Halmos, 1950; Feller, 1968, 1971; Brémaud, 1988; Ibarrola et al., 1997; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013); ver también (Barone & Novikoff, 1978; Borel, 1898; Sierpiński, 1918, 1975, 1976, & Ref.) para notas históricas):

[9] Este tipo de diagramas fue popularizado por el inglés John Venn en 1880, pero en su trabajo (Venn, 1880) da la paternidad al matemático suizo Leonhard Euler, uno de los primeros en usar tal representación en el siglo XVIII en sus famosas "Cartas a una princesa alemana, acerca de diversas cuestiones de física y filosofía" (ver (Euler, 1768, L 102-105, pp. 95-126)), o antes a Christian Weise y Johan Christian Langius (Langius, 1712); apareció aún en trabajos de Gottfried Wilhelm Leibniz en el siglo anterior.

Definición 1-1 (Espacio medible)**.** $(\Omega, \mathcal{A})$, *formado por un espacio muestral* Ω *y una colección* $\mathcal{A}$ *de conjuntos de* Ω, *es llamado espacio medible si satisface los requisitos*

1. $\emptyset \in \mathcal{A}$,
2. *si* $A \in \mathcal{A}$, *entonces* $\bar{A} \in \mathcal{A}$,
3. *la unión numerable de conjuntos de* $\mathcal{A}$ *queda en* $\mathcal{A}$ ($\mathcal{A}$ *es cerrado por la unión numerable).*

Con estas propiedades, $\mathcal{A}$ *es llamada una* σ*-álgebra. Los elementos de* $\mathcal{A}$ *son dichos medibles.*

Es sencillo mostrar que Ω también está en $\mathcal{A}$, y que $\mathcal{A}$ es cerrado por la intersección numerable. Un ejemplo de σ-álgebra sobre $\Omega = \{1\,;\,2\,;\,3\,;\,4\,;\,5\,;\,6\}$ puede ser $\mathcal{A} = \big\{\emptyset\,;\,\Omega\,;\,\{1\,;\,2\,;\,3\}\,;\,\{4\,;\,5\,;\,6\}\big\}$.

A partir de $(\Omega, \mathcal{A})$, se asocia una noción de probabilidad P a un dado evento. Esta queda determinada por los siguientes requisitos llamados *Axiomas de Kolmogorov* (ver por ejemplo (Spiegel, 1976; Kolmogorov, 1956; Shafer & Vovk, 2006; von Plato, 2005)):

1. $P(A) \geq 0 \;\; \forall\, A \in \mathcal{A}$.
2. Si $A_1, \ldots, A_i, \ldots$ son eventos mutuamente excluyentes de $\mathcal{A}$, entonces $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$.
3. $P(\Omega) = 1$.

Más formalmente, se define un *espacio de probabilidad* o *espacio probabilístico* de la manera siguiente (Halmos, 1950; Feller, 1968, 1971; Brémaud, 1988; Ibarrola et al., 1997; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Jacob & Protters, 2003; Cohn, 2013):

Definición 1-2 (Espacio de medida y espacio probabilístico)**.** *Sea* $(\Omega, \mathcal{A})$ *un espacio medible. Una función* $\mu : \mathcal{A} \mapsto \mathbb{R}_+$ *tal que*

1. $\mu(\emptyset) = 0$, *y*
2. *para cualquier conjunto numerable* $\{A_i\}_{i \in I}$ (I *numerable) de elementos mutuamente excluyentes de* $\mathcal{A}$ *se tiene* $\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$,

es llamada función medida o medida σ*-aditiva, y el espacio* $(\Omega, \mathcal{A}, \mu)$ *es llamado espacio de medida.*

- *Cuando* μ *es tal que existe un conjunto numerable* $\{A_i\}_{i \in I}$ (I *numerable) de elementos de* $\mathcal{A}$ *tal que* $\Omega = \cup_{i \in I} A_i$ *y* $\forall\, i \in I, \quad \mu(A_i) < +\infty$ *finito, la medida se dice* σ*-finita y el espacio de medida se dice* σ*-finito.*
- *Cuando* μ *está acotada por arriba,* $\mu(\Omega) < +\infty$, *la medida se dice finita y el espacio de medida también se dice finito.*
- *Además, si* $\mu(\Omega) = 1$, *la medida es dicha medida de probabilidad. En general, se la denota* P. *En este caso, el espacio* $(\Omega, \mathcal{A}, P)$ *es llamado espacio probabilístico.*

(ver también (Kolmogorov & Fomin, 1961, Cap. 5 & 6)). Es importante notar que una combinación lineal positiva de medidas es una medida, pero el producto de dos medidas no es una medida más.

A partir de los axiomas de Kolmogorov se pueden probar varios corolarios y propiedades:

- la probabilidad de un evento seguro o cierto es 1;
- la probabilidad de un evento que no puede ocurrir es 0: por ejemplo, $P(\emptyset) = 0$;
- el rango de las probabilidades está acotado: $0 \leq P(A) \leq 1 \ \ \forall\, A \in \mathcal{A}$;
- condición de normalización: si $\Omega = \bigcup_{i=1}^{n} A_i$, con A_i mutuamente excluyentes, entonces $\sum_{i=1}^{n} P(A_i) = 1$; el conjunto $\{A_i\}_{i=1}^{n}$ se dice *conjunto completo de eventos posibles excluyentes entre sí* y es ilustrado en la Figura 1-2;
- si A es subconjunto de B, lo que escribiremos $A \subset B$, es decir que si B se realiza, A se realiza también (pero no necesariamente al revés), entonces $P(A) \leq P(B)$; es ilustrado en la Figura 1-2.

(a) (b)

Figura 1-2: Ilustración de: (a) conjunto completo de eventos posibles excluyentes entre sí; (b) inclusión entre eventos, donde A está en gris oscuro mientras que B está en gris (claro y oscuro).

La probabilidad $P(A \cap B)$ del evento $A \cap B$ se llama también *probabilidad conjunta* de A y B. Se demuestra que

- $P(A \cap B)$ está acotada: $0 \leq P(A \cap B) \leq \min\{P(A), P(B)\}$ (viene de $A \cap B \subset A$ y $A \cap B \subset B$);
- si A y B son mutuamente excluyentes, entonces $P(A \cap B) = 0$ (viene de $A \cap B = \emptyset$);
- si $\{B_j\}_{j=1}^{m}$ es un conjunto completo de eventos posibles excluyentes entre sí, entonces $\sum_{j=1}^{m} P(A \cap B_j) = P(A)$ (viene de $\{A \cap B_j\}_j$ mutuamente excluyentes y $\bigcup_j (A \cap B_j) = A \cap \left(\bigcup_j B_j\right) = A \cap \Omega = A$).

En el caso de eventos no necesariamente mutuamente excluyentes, se prueba que la *ley de composición* o *fórmula de inclusión-exclusión* es

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B),$$

y que para n eventos resulta

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P\left(A_i\right).$$

La igualdad vale en el caso especial de eventos mutuamente excluyentes (recuperando el segundo axioma de Kolmogorov).

Se prueba también que si $\{A_i\}_{i=1}^{+\infty}$ es una secuencia *creciente* de eventos, *i. e.,* $\forall\, i \geq 1, \quad A_i \subset A_{i+1}$, entonces

$$P\left(\bigcup_{i=1}^{+\infty} A_i\right) = \lim_{i \to +\infty} P(A_i).$$

Por otro lado, si $\{A_i\}_{i=1}^{+\infty}$ es una secuencia *decreciente* de eventos, *i. e.,* $\forall\, i \geq 1, \quad A_{i+1} \subset A_i$, entonces

$$P\left(\bigcap_{i=1}^{+\infty} A_i\right) = \lim_{i \to +\infty} P(A_i).$$

Podemos preguntarnos cuál es la probabilidad de un evento A, si sabemos que se da cierto evento B. Por ejemplo, para un dado de 6 caras equilibrado, cuál es la probabilidad de tener un número par sabiendo que tenemos un número menor o igual a 3. La respuesta se encuentra en la noción de *probabilidad condicional* (Hausdorff, 1901; Jeffrey, 1948, 1973; Brémaud, 1988; Mandel & Wolf, 1995; Ash & Doléans-Dade, 1999; Jacob & Protters, 2003; Shafer & Vovk, 2006):

Definición 1-3 (Probabilidad condicional)**.** *La probabilidad condicional de A dado B, denotado $P(A|B)$, se define como la razón entre la probabilidad del evento conjunto y la probabilidad de que se dé B (cuando éste es un evento de probabilidad no nula):*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

En el ejemplo precedente, la probabilidad condicional va a ser $P(A|B) = \frac{P(A \cup B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

Claramente del hecho de que P es una medida de probabilidad se tiene

$$P(A|B) \geq 0.$$

Luego, de $A \cap B \subseteq B$ resulta $P(A \cap B) \leq P(B)$; es decir,

$$P(A|B) \leq 1.$$

Además, $P(\Omega \cap B) = P(B)$ dando

$$P(\Omega|B) = 1.$$

Para cualquier conjunto $\{A_i\}$ de eventos mutuamente excluyentes, los $(A_i \cap B)$ son también mutuamente excluyentes, así que $P\left(\left(\bigcup_i A_i\right) \bigcap B\right) = P\left(\bigcup_i \left(A_i \bigcap B\right)\right) = \sum_i P\left(A_i \bigcap B\right)$ dando

$$P\left(\bigcup_i A_i \middle| B\right) = \sum_i P\left(A_i | B\right).$$

Dicho de otra manera, $P(A|B)$ es una medida de probabilidad [10]. Diversas situaciones de probabilidades condicionales son ilustradas en la Fig. 1-3.

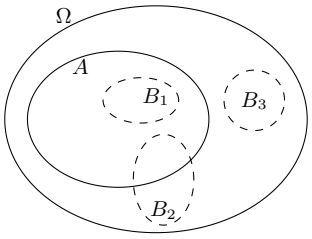


Figura 1-3: Ilustración de la probabilidad condicional con A interior de la curva en línea llena y unos B_i interiores de las curvas en líneas discontinuas. Se tiene $\omega \in B_1 \Rightarrow \omega \in A$ así que $P(A|B_1) = 1$; por otro lado, $\omega \in B_3 \Rightarrow \omega \notin A$ así que $P(A|B_3) = 0$. Entre estas situaciones extremas, si $P(\bar{A} \cap B_2) \neq 0$ y $P(A \cap B_2) \neq 0$ tenemos $0 < P(A|B_2) < 1$ (se puede tomar el ejemplo de probabilidad de un evento iguale a su superficia sobre la de Ω para ver estas propiedades en este caso particular).

Algunas propiedades interesantes son las siguientes:

- $P(A \cap B|B) = P(A|B)$ $\big($viene de $P(A \cap B \cap B) = P(A \cap B)\big)$;
- si A y B son mutuamente excluyentes, obviamente $P(A|B) = 0$;
- si $B \subseteq C$, entonces $P(A|B \cap C) = P(A|B)$ $\big($viene de $P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(A \cap B)}{P(B)}$, pues $B \cap C = B\big)$;
- condición de normalización: si $\{A_i\}_{i=1}^n$ es un conjunto completo de resultados posibles mutuamente excluyentes, entonces $\sum_{i=1}^n P(A_i|B) = 1$;
- relación entre probabilidades condicionales inversas: $P(B|A) = \dfrac{P(B)}{P(A)} P(A|B)$, de donde $P(A|B)$ y $P(B|A)$ coinciden sólo cuando A y B tienen la misma probabilidad;

Las dos propiedades importantes son la fórmula de probabilidad total y la fórmula de Bayes (ver (Brémaud, 1988; Jacob & Protters, 2003; Bayes, 1763; Barnard, 1958) [11]). Les vamos a ver en forma de lemma:

Teorema 1-1 (Fórmula de probabilidad total)**.** *Sea $J \subseteq \mathbb{N}$ y $\{B_j\}_{j \in J}$ un conjunto completo de eventos*

[10] Se puede definir un espacio de probabilidad $(\Omega_B, \mathcal{A}_B, P_B)$ donde $P_B(A) \equiv P(A|B)$. La notación $P_B(A)$ suele utilizarse en la literatura pero no la usaremos en esta obra para no confundirla con la medida de probabilidad de una variable aleatoria que definiremos en la sección siguiente.

[11] La obra del matemático y religioso inglés Thomas Bayes fue de hecho recopilada y publicada después de su muerte por Richard Price.

no nulos mutuamente excluyentes, i. e., tal que $B_i \cap B_j = \emptyset$ *si* $i \neq j$ *y* $\bigcup_{j \in J} B_j = \Omega$. *Entonces*

$$P(A) = \sum_{j \in J} P(A|B_j)P(B_j)$$

Demostración. Escribimos $A = A \cap \left(\bigcup_j B_j\right) = \bigcup_j (A \cap B_j)$. Los $A \cap B_j$ son mutuamente excluyentes, y $P(A \cap B_j) = P(A|B_j)P(B_j)$ lo que cierra la prueba. □

Lema 1-1 (Fórmula de Bayes)**.** *Sea* $J \subseteq \mathbb{N}$ *y* $\{B_j\}_{j \in J}$ *un conjunto completo de eventos no nulos mutuamente excluyentes. Entonces*

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j \in J} P(A|B_j)P(B_j)};$$

Demostración. Esta fórmula resuelta de la definición de la probabilidad condicional y de la fórmula de probabilidad total. □

Veamos ahora la noción de independencia entre dos eventos. Por ejemplo, si se tiran dos dados sobre sendas mesas, no hay ninguna razón para que la muestra de uno "influya" la del otro. Dicho de otra manera, dos eventos son independientes si el conocimiento de uno no lleva ninguna "información" sobre el otro (Brémaud, 1988; Mandel & Wolf, 1995; Ash & Doléans-Dade, 1999; Hausdorff, 1901; Jacob & Protters, 2003; Borel, 1909):

Definición 1-4 (Independencia estadística)**.** *Dos eventos* A *y* B *se dicen estadísticamente independientes si la probabilidad condicional de* A *dado* B *es igual a la probabilidad incondicional de* A*:*

$$P(A|B) = P(A).$$

Es equivalente al hecho de que la probabilidad conjunta se factoriza:

$$P(A \cap B) = P(A)P(B).$$

Por inducción, la condición necesaria y suficiente para que n eventos $A_1, \ldots, A_n$ sean mutuamente estadísticamente independientes es que la probabilidad conjunta se factorice como

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Se deduce que los eventos mutuamente excluyentes no son estadísticamente independientes.

Es importante notar que la independencia mutua no es equivalente a la independencia por pares de eventos, como lo ilustra el ejemplo siguiente.

Ejemplo 1-1 (Independencia mutua vs por pares)**.** *Tiramos 2 dados independientemente y consideramos los eventos:* $A_i, i = 1, 2$ *"el dado* i *es par" y* A_3 *"la suma de ambos dados es impar". Es claro que* A_1 *y* A_2 *son independientes y además para* $i = 1$ *o 2,* $P(A_i \cap A_3) = \frac{1}{4} = P(A_i)P(A_3)$*, mientras que* $P(A_1 \cap A_2 \cap A_3) = 0 \neq \frac{1}{8}$*: los eventos son independientes por pares, pero no son mutuamente independientes (Hogg et al., 2013).*

Definición 1-5 (Independencia condicional)**.** *Dos eventos A y B se dicen estadísticamente independientes condicionalmente a un tercer evento C, si la probabilidad conjunta de A y B condicionalmente a C es igual al producto de la probabilidad de A condicionalmente a C por la de B condicionalmente a C:*

$$P(A \cap B|C) = P(A|C)P(B|C).$$

Si $P(B|C) \neq 0$, es equivalente a $P(A|B \cap C) = P(A|C)$.

Es importante notar que dos eventos pueden ser independientes, pero no serlo condicionalmente a un tercero, como lo ilustra el ejemplo siguiente.

Ejemplo 1-2 (Independencia incondicional pero no condicional)**.** *Teniendo dos monedas bien equilibradas y tirándolas de manera independiente, consideramos los eventos A "la primera faz es una cruz", B "la segunda faz es una cara", C "las faces son idénticas". Claramente $P(A \cap B) = \frac{1}{4} = P(A)P(B)$, mientras que $P(A \cap B|C) = 0 \neq P(A|C)P(B|C) = \frac{1}{16}$.*

Al revés, dos eventos pueden ser condicionalmente independientes a un tercero, pero ser dependientes.

Ejemplo 1-3 (Independencia condicional pero no incondicional)**.** *Sea Alice tirando una moneda bien equilibrada y denotamos A el evento "era una cruz". Claramente $P(A) = \frac{1}{2}$. Suponemos que Alice transmite el resultado a Bob a través de un intermediario Charlie con una probabilidad ε de mentir a Charlie, y llamamos C el evento "Alice dijo a Charlie que era una cruz". Tenemos que $P(C) = P(C|A)P(A) + P(C|\overline{A})P(\overline{A}) = (1-\varepsilon)\frac{1}{2} + \varepsilon\frac{1}{2} = \frac{1}{2}$. Suponemos ahora que Charlie transmite a Bob lo que le dijo Alice, con una probabilidad ϑ de mentir (independientemente de Alice) y llamamos B el evento "Charlie dijo a Bob que era una cruz". Es de nuevo sencillo ver que $P(B) = \frac{1}{2}$. Ahora, $P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = 2\,P(A \cap B \cap C)$. El evento $A \cap B \cap C$ es era una cruz y Alice no mintó y Charlie tampoco, es decir, por la independencia: $P(A \cap B|C) = (1-\varepsilon)(1-\vartheta)$. Inmediatamente $P(B|C) = 1-\vartheta$ y $P(A|C) = 2\,P(A \cap C)$ siendo $A \cap C$ el evento "era una cruz y Alice no mintió", i.e. $P(A|C) = 1-\varepsilon$. En conclusión, $P(A \cap B|C) = P(A|C)P(B|C)$: A y B son independientes condicionalmente a C. Ahora, $P(A \cap B) = P(A \cap B \cap C) + P(A \cap B \cap \overline{C}) = \frac{1}{2}(1-\varepsilon)(1-\vartheta) + \frac{1}{2}\varepsilon\vartheta \neq \frac{1}{4} = P(A)P(B)$ en general: A y B no resultan independientes. Este ejemplo es una instancia de lo que se llama un proceso de Markov, que vamos a ver un poco más en el cápitulo 2.*

1.3 Variables aleatorias y distribuciones de probabilidad

En un experimento o un dado proceso, los posibles resultados son típicamente números reales, siendo cada número un evento. Luego los resultados son mutuamente excluyentes. Se considera a esos números como valores de una *variable aleatoria* X a valores reales, que puede ser discreta, continua o mixta.

Formalmente, la noción de variable aleatoria se apoya sobre la noción de función medible. Por esta formalización, vamos a necesitar definir la integración de manera general, más allá del enfoque de Riemann ("a la Lebesgue"), así como la noción de derivada de una medida con respecto a otra para definir densidades de probabilidad, en analogía a la densidad de masa en mecánica por ejemplo (Lebesgue, 1904, 1918; Kolmogorov & Fomin, 1961; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013).

1.3.1 Consideraciones preliminaries: Teorías de la medida y de la integración.

La primera noción que subyace a la definición formal de variable aleatoria es la de función medible:

Definición 1-6 (Función medible)**.** *Sean* $(\Omega, \mathcal{A})$ *y* $(\Upsilon, \mathcal{B})$ *dos espacios medibles. Una función* $f : \Omega \mapsto \Upsilon$ *se dice* $(\mathcal{A}, \mathcal{B})$*-medible si*

$$\forall\, B \in \mathcal{B}, \quad A \equiv f^{-1}(B) = \left\{\omega \in A \;\middle|\; f(\omega) \in B\right\} \in \mathcal{A}.$$

Dicho de otra manera, la pre-imágen de un elemento dado de $\mathcal{B}$ *(elemento medible) pertenece a* $\mathcal{A}$ *(elemento medible). Por abuso de escritura, se dice más simplemente que* $f : (\Omega, \mathcal{A}) \mapsto (\Upsilon, \mathcal{B})$ *es medible.*

Además, a partir de un espacio de medida y una función f medible, se puede definir una medida imágen sobre el espacio de llegada (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013):

Teorema 1-2 (Teorema de la medida imágen)**.** *Sean* $(\Omega, \mathcal{A}, \mu)$ *un espacio de medida,* $(\Upsilon, \mathcal{B})$ *un espacio medible y* $f : (\Omega, \mathcal{A}) \mapsto (\Upsilon, \mathcal{B})$ *una función medible. Sea* μ_f *tal que*

$$\forall\, B \in \mathcal{B}, \quad \mu_f(B) = \mu\left(f^{-1}(B)\right).$$

Entonces, μ_f *es una medida sobre el espacio medible* $(\Upsilon, \mathcal{B})$*, i. e.,* $(\Upsilon, \mathcal{B}, \mu_f)$ *define un espacio de medida. Además,* $\mu(\Omega) = \mu_f(\Upsilon)$ *(posiblemente infinitas). Se dice que* μ_f *es la medida imágen de* μ *por* f*.*

Demostración. Por definición, claramente $\mu_f \geq 0$. Además, obviosamente $f^{-1}(\emptyset) = \emptyset$ dando $\mu_f(\emptyset) = \mu(\emptyset) = 0$. Luego, para un conjunto numerable $\{B_j\}$ de elementos de $\mathcal{B}$ disjuntos entre sí, las pre-imágenes de los B_j también son disjuntos entre sí (para $k \neq j$ no se puede tener $\omega \in f^{-1}(B_j) \cap f^{-1}(B_k)$ sino ω tendría dos imágenes distintas por f). Entonces $f^{-1}\left(\bigcup_j B_j\right) = \bigcup_j f^{-1}(B_j)$. Esto implica que $\mu_f\left(\bigcup_j B_j\right) = \mu\left(f^{-1}\left(\bigcup_j B_j\right)\right) = \mu\left(\bigcup_j f^{-1}(B_j)\right) = \sum_j \mu\left(f^{-1}(B_j)\right) = \sum_j \mu_f(B_j)$. Finalmente, necesariamente $f^{-1}(\Upsilon) = \Omega$ (obviamente $f(\Omega) \subseteq \Upsilon$) lo que cierra la prueba [12] □

[12] De hecho, se puede probar sencillamente que la pre-imágen de una unión numerable (disjuntos o no) es la unión de las pre-

A continuación, necesitaremos tratar de funciones medibles teniendo una propiedad (P) salvo sobre un conjunto de medida μ igual a cero. Más generalmente viene acá la noción de propiedad *casi siempre*:

Definición 1-7 (Propiedad (e igualdad) μ-casi siempre)**.** *Una funcion medible f se dice tener una propiedad (P) dada μ-casi siempre, si y solamente si la tiene excepto sobre un conjunto de medida nula,*

$$\mu\left(\{\omega \mid f(\omega) \text{ no satisface (P) }\}\right) = 0.$$

Por ejemplo, dos funciones medibles f_1 y f_2 $(\Omega, \mathcal{A}, \mu) \to (\Upsilon, \mathcal{B})$ son iguales μ-casi siempre,

$$f_1 = f_2 \quad (\mu\text{-c.s.})$$

si y solamente si son iguales excepto sobre un conjunto de medida nula,

$$\mu\left(\{\omega \mid f_1(\omega) \neq f_2(\omega)\}\right) = 0.$$

Un espacio que juega un rol particular es $\mathbb{R}^d$, al cual se puede asociar una σ-álgebra particular conocida como *σ-álgebra de Borel* (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a, 2007b; Cohn, 2013):

Definición 1-8 ($\mathbb{R}^d$ y Borelianos)**.** *Para cualquier $d \geq 1$ entero, llamamos Borelianos $\mathcal{B}(\mathbb{R}^d)$ de $\mathbb{R}^d$ a la σ-álgebra más pequeña generada por los productos cartesianos $\bigtimes_{i=1}^{d}(-\infty\,;\, b_i]$ (similarmente, por los abiertos de $\mathbb{R}^d$, o también para los productos cartesianos de intervalos $\bigtimes_{i=1}^{d}(a_i\,;\, b_i]$), i. e., uniones numerables, intersecciones numerables, complementos de estos intervalos. $\mathcal{B}(\mathbb{R}^d)$ es también llamado σ-álgebra de Borel de $\mathbb{R}^d$.*

Se necesita ahora definir la noción de integración de una función medible con respecto a una medida:

Definición 1-9 (Medida e integración)**.** *Para una medida cualquiera, sobre un espacio de medida $(\Omega, \mathcal{A}, \mu)$, se define la integración a partir de*

$$\forall\, A \in \mathcal{A}, \quad \int_A d\mu(\omega) = \int_\Omega \mathbb{1}_A(\omega)\, d\mu(\omega) = \mu(A),$$

donde $\mathbb{1}_A$ es la función indicadora del conjunto A (ver notaciones). $d\mu(\omega)$ se escribe a veces también $\mu(d\omega)$, medida de un "infinitésimo". Claramente, por propiedades de una medida, para A_i, A_j disjuntos

---

imágenes; lo mismo occure para la intersección y además la pre-imágen del complemento es el complemento de la pre-imágen. Esto se conoce como *leyes de de Morgan* (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Hogg et al., 2013) (ver también (Kolmogorov & Fomin, 1957, Cap. 1) y (Kolmogorov & Fomin, 1961, Caps. 5 & 6)).

$\mathbb{1}_{A_i} + \mathbb{1}_{A_j} = \mathbb{1}_{A_i \cup A_j}$, *dando* $\int_\Omega \left(\mathbb{1}_{A_i} + \mathbb{1}_{A_j}\right) d\mu(\omega) = \mu(A_i \cup A_j) = \mu(A_i) + \mu(A_j) = \int_\Omega \mathbb{1}_{A_i} d\mu(\omega) + \int_\Omega \mathbb{1}_{A_j} d\mu(\omega)$ *y entonces, sin perdida de generalidad para un conjunto* $\{A_j\}$ *numerable y* $\{a_j\}$ *reales no negativos, la integral de la función escalonada* $\sum_j a_j \mathbb{1}_{A_j}$ *es dada por*

$$\int_\Omega \left(\sum_j a_j \mathbb{1}_{A_j}(\omega) \right) d\mu(\omega) = \sum_j a_j \int_\Omega \mathbb{1}_{A_j}(\omega)\, d\mu(\omega).$$

Para los A_i *disjuntos es la consecuencia directa de la propiedad precedente, y si* A_i, A_j *no son disjuntos. De hecho, sufice considerar* $A_i \setminus A_j,\ A_j \setminus A_i, A_i \cap A_j$ *con* $A \setminus B = \{\omega \mid \omega \in A \ \wedge \ \omega \notin B\}$ *y respectivamente los coefficientes* $a_i,\ a_j,\ a_i + a_j$ *para volver al caso de conjuntos disjuntos.*

Antes de definir la integración de una función real, medible, cualquiera, el último paso que falta es el siguiente:

Teorema 1-3 (Función medible como límite)**.** *Sea* $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, no negativa y medible. Existe una sucesión* $\{g_n\}_{n\in\mathbb{N}}$ *creciente de funciones escalonadas que converge simplemente (punto a punto) hacia* g*.*

Demostración. La sucesión $g_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\, \mathbb{1}_{g^{-1}\left(\left[\frac{k}{2^n}\,;\,\frac{k+1}{2^n}\right)\right)} + n\, \mathbb{1}_{g^{-1}([n\,;+\infty))}$ es escalonada, creciente y converge hacia g (notar que esta sucesión comparte la idea que subyace a la integración de Riemann).

□

De este resultado, se puede generalizar la noción de integración de una función real:

Definición 1-10 (Integración de una función real)**.** *Sea* $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, no negativa y medible, y* $\{g_n\}_{n\in\mathbb{N}}$ *una sucesión creciente de funciones escalonadas que converge simplemente hacia* g*. Por definición,*

$$\int_\Omega g(\omega)\, d\mu(\omega) = \lim_{n\to\infty} \int_\Omega g_n(\omega)\, d\mu(\omega).$$

Notar de que el límite puede ser infinita.

Sea ahora $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *medible cualquiera. Se verifica sencillamente que también* $|g|$ *(valor absoluto) es medible y, por definición,* g *se dice* μ*-integrable si la integral de* $|g|$ *es finita,*

$$g \text{ es } \mu\text{-integrable} \quad \Leftrightarrow \quad \int_\Omega |g(\omega)|\, d\mu(\omega) < +\infty.$$

Además, se escribe $g = g_+ + g_-$ *con* $g_+ = \max(g, 0)$ *y* $g_- = \min(g, 0)$*. Es sencillo ver de que si* g *es medible,* g_+ *y* g_- *son medibles. Si* g *es* μ*-integrable, necesariamente* g_+ *y* g_- *son* μ*-integrables, y, por definición*

$$\int_\Omega g(\omega)\, d\mu(\omega) = \int_\Omega g_+(\omega)\, d\mu(\omega) - \int_\Omega (-g_-(\omega))\ d\mu(\omega).$$

A continuación, damos unos teoremas que serán muy útiles más adelante, sin detallar las pruebas. Por esto, el lector se puede referir a (Ash & Doléans-Dade, 1999; Lieb & Loss, 2001; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013).

Teorema 1-4 (Teorema de convergencia monótona)**.** *Sea* $\{f_n\}_{n\in\mathbb{N}}$ *una sucesión creciente de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$*, positivas, convergiendo simplemente hacia una función* f *medible. Entonces*

$$\lim_{n\to+\infty} \int_\Omega f_n(\omega) d\mu(\omega) = \int_\Omega f(\omega) d\mu(\omega).$$

De hecho se prueba este teorema a partir de la definición de integración. Este teorema da una condición simple permitiendo intercambiar integración y límite.

Corolario 1-1. *Sea* $\{f_n\}_{n\in\mathbb{N}}$ *una sucesión de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$*, positivas, tal que la serie* $\sum_n f_n$ *converge simplemente hacia una función* f*,* μ*-integrable. Entonces*

$$\int_\Omega \sum_{n\in\mathbb{N}} f_n(\omega) d\mu(\omega) = \int_\Omega f(\omega) d\mu(\omega).$$

Es una consecuencia del teorema de convergencia monótona, considerando la sucesión creciente $\{\sum_{k=0}^n f_k\}_{n\in\mathbb{N}}$.

Teorema 1-5 (Teorema de convergencia dominada)**.** *Sea* $\{f_n\}_{n\in\mathbb{N}}$ *una sucesión creciente de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$ *convergiendo simplemente hacia una función* f*, medible. Suponemos que existe una función* μ*-integrable* g *que domina la sucesión, i. e.,* $\forall\, \omega \in \Omega, \quad |f_n(\omega)| \leq g(\omega)$*. Entonces*

$$\lim_{n\to+\infty} \int_\Omega f_n(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\, d\mu(\omega) \leq \int_\Omega g(\omega)\, d\mu(\omega).$$

Este teorema da una condición suficiente muy útil y muy usada para asegurarse de que se puede intercambiar límite e integración.

El último teorema que vamos a necesitar permite intercambiar dos integraciones. Antes, necesitamos definir la noción de espacio medible producto y medida producto.

Definición 1-11 (Espacio medible producto, medida producto)**.** *Sean dos espacios de medida* $(\Omega_1, \mathcal{A}_1, \mu_1)$ *y* $(\Omega_2, \mathcal{A}_2, \mu_2)$*. Llamamos espacio medible producto* $(\Omega, \mathcal{A})$ *al espacio del producto cartesiano* $\Omega = \Omega_1 \times \Omega_2$ *con la* σ*-álgebra* $\mathcal{A}$ *generada por los productos cartesianos* $A_1 \times A_2$ *donde* $A_i \in \mathcal{A}_i, i = 1, 2$*. Además, llamamos medida producto* μ *definida sobre* $\mathcal{A}$ *a la medida tal que* $\forall\, (A_1, A_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \quad \mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$*.*

Teorema 1-6 (Teorema de Fubini)**.** *Sea* $(\Omega, \mathcal{A}, \mu)$ *espacio de medida producto de* $(\Omega_1, \mathcal{A}_1, \mu_1)$ *y* $(\Omega_2, \mathcal{A}_2, \mu_2)$ *donde* μ *es la medida producto. Sea* f *una función integrable sobre* $(\Omega, \mathcal{A}, \mu)$ *entonces*

- $\omega_1 \mapsto f(\omega_1, \omega_2)$ *es* μ_1*-integrable (*μ_2*-c.s.)* *y* $\omega_2 \mapsto f(\omega_1, \omega_2)$ *es* μ_2*-integrable (*μ_1*-c.s.),*
- $\omega_1 \mapsto \displaystyle\int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)$ *es* μ_1*-integrable* *y* $\omega_2 \mapsto \displaystyle\int_{\Omega_1} f(\omega_1, \omega_2)\, d\mu_1(\omega_1)$ *es* μ_2*-integrable.*

Además,

$$\int_{\Omega_1\times\Omega_2} f(\omega)\, d\mu(\omega) = \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega)\, d\mu_2(\omega_2) \right) d\mu_1(\omega_1) = \int_{\Omega_2} \left(\int_{\Omega_1} f(\omega)\, d\mu_1(\omega_1) \right) d\mu_2(\omega_2)$$

Teorema 1-7 (Integral a parámetro: continuidad y diferenciabilidad)**.** *Sea I un compacto de $\mathbb{R}^d$ y $\{f(\cdot,t)\}_{t\in I}$ una familia de funciones medibles sobre $(\Omega,\mathcal{A},\mu)$, tal que $t\mapsto f(\omega,t)$ sea continua sobre I (μ-c.s.). Si existe una función μ-integrable g tal que*

$$\forall\, t\in I,\quad \forall\,\omega\in\Omega,\quad |f(\omega,t)|\le g(\omega),$$

entonces $\omega\to f(\omega,t)$ es μ-integrable y la función $t\mapsto\displaystyle\int_\Omega f(\omega,t)\,d\mu(\omega)$ es continua sobre I. Además, si $f(\omega,\cdot)$ es diferenciable sobre I y si existe una función μ-integrable h tal que

$$\forall\, t\in I,\quad \forall\,\omega\in\Omega,\quad \|\nabla_t f(\omega,t)\|\le h(\omega),$$

donde ∇_t indica el gradiente, i. e., el vector de componentes $\frac{\partial}{\partial t_1},\dots,\frac{\partial}{\partial t_d}$, entonces la función $t\mapsto\displaystyle\int_\Omega f(\omega,t)\,d\mu(\omega)$ es diferenciable sobre I, y

$$\nabla_t\int_\Omega f(\omega,t)\,d\mu(\omega)=\int_\Omega \nabla_t f(\omega,t)\,d\mu(\omega).$$

Básicamente, este teorema es consecuencia del teorema de convergencia dominada.

Seguimos esta sección con la noción de derivada de una medida con respecto a otra, dando una definición muy general de densidad:

Definición 1-12 (Densidad de una medida)**.** *Sean μ y ν dos medidas cualesquiera sobre un espacio medible $(\Omega,\mathcal{A})$. Si existe una función real no negativa $p:\Omega\mapsto\mathbb{R}_+$ medible tal que*

$$\forall\, A\in\mathcal{A},\quad \nu(A)=\int_A p(\omega)\,d\mu(\omega),$$

p es llamada densidad de ν con respecto a μ, denotada

$$p=\frac{d\nu}{d\mu},$$

también llamada derivada de Radon-Nikodým.

Notar que dos funciones pueden cumplir esta definición, por ejemplo si son iguales μ-casi siempre. De hecho, si dos funciones $p_1=p_2$ (μ-c.s.), y C es el conjunto donde no son iguales, siendo de medida nula, de $\displaystyle\int_A p_1(\omega)d\mu(\omega)=\int_{A\setminus C} p_1(\omega)d\mu(\omega)=\int_{A\setminus C} p_2(\omega)d\mu(\omega)=\int_A p_2(\omega)d\mu(\omega)$ se ve que dos funciones iguales casi siempre pueden ser densidad de una medida con respecto a una otra.

Es sencillo ver que si $\mu(A)=0$, necesariamente $\nu(A)=0$. De eso viene la noción de absoluta continuidad:

Definición 1-13 (Absoluta continuidad)**.** *Sean μ y ν dos medidas sobre un espacio medible $(\Omega,\mathcal{A})$. Se dice que ν es absolutamente continua con respecto a μ, denotado*

$$\nu\ll\mu\qquad \textit{si}\ \ \forall\, A\in\mathcal{A},\quad \mu(A)=0\ \Rightarrow\ \nu(A)=0.$$

De hecho, se muestra la recíproca de la definición Def. 1-12 a través de lo que se conoce como teorema de Radon-Nikodým (Nikodym, 1930; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013):

Teorema 1-8 (Radon-Nikodým)**.** *Sean dos medidas μ y ν, entonces*

$$\nu \ll \mu \quad \Longleftrightarrow \quad \nu \text{ admite una densidad con respecto a } \mu.$$

Además, esta densidad $\frac{d\nu}{d\mu}$ es única en el sentido de que si dos funciones cumplen la definición, son iguales μ-casi siempre.

En todo lo que sigue, hablaremos de "la" densidad de una medida, salvo si se necesita explícitamente tener en cuenta esta sutileza.

A continuación, dos lemas van a ser muy útiles especialmente en el Capítulo 2, tratando con dos (o más) medidas y densidades.

Lema 1-2. *Sean ν y μ dos medidas sobre $(\Omega, \mathcal{A})$ tales que $\nu \ll \mu$. Entonces, para cualquier función medible f,*

$$\int_\Omega f(\omega)\, \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\, d\nu(\omega)$$

Demostración. Tomando $f = \mathbb{1}_A$, de la definición Def. 1-12 se obtiene

$$\int_\Omega \mathbb{1}_A(\omega)\, \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \int_A \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \nu(A) = \int_A d\nu(\omega)$$

Se cierra la prueba usande el teorema 1-3 y la definición 1-10, tratando f con su parte positiva y negativa separadamente. □

Lema 1-3. *Sean ν, μ y λ tres medidas sobre $(\Omega, \mathcal{A})$ y suponemos $\nu \ll \lambda$ y $\lambda \ll \mu$. Entonces*

- $\nu \ll \mu$ *(transitividad);*
- *equivalentemente, el soporte (ensemble de puntos que no anula la función) de $\dfrac{d\nu}{d\mu}$ está incluído (μ-casi siempre) en el soporte de $\dfrac{d\lambda}{d\mu}$;*
- $\dfrac{d\nu}{d\lambda}\dfrac{d\lambda}{d\mu} = \dfrac{d\nu}{d\mu}$ *(μ-c.s.).*

Demostración. El primer resultado viene de la definición de la absoluta continuidad $\mu(A) = 0 \quad \Rightarrow \quad \lambda(A) = 0 \quad \Rightarrow \quad \nu(A) = 0$. El segunndo resultado se obtiene escribiendo la medida en su forma integral. Adémas, por definición de la densidad, $\forall A \in \mathcal{A}, \quad \nu(A) = \displaystyle\int_A \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega)$. Luego, aplicando el lema anterior a $f = \mathbb{1}_A \frac{d\nu}{d\mu}$ se obtiene que, también, $\nu(A) = \displaystyle\int_A \frac{d\nu}{d\lambda}(\omega)\, d\lambda(\omega) = \int_A \frac{d\nu}{d\lambda}(\omega)\, \frac{d\lambda}{d\mu}(\omega)\, d\mu(\omega)$, lo que cierra la prueba. □

Unas medidas que juegan un rol particular son las medidas de Lebesgue o medidas discretas.

Definición 1-14 (Medida de Lebesgue)**.** *La medida de Lebesgue* μ_L *sobre* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *se define tal que para cualquier producto cartesiano de intervalos,*

$$\mu_L\left(\bigtimes_{i=1}^{d}(a_i\,;\,b_i)\right) = \prod_{i=1}^{d}(b_i - a_i).$$

Acá, notamos dos hechos interesantes:

- μ_L es σ-finita. Viene de $\mathbb{R}^d = \cup_{i=1}^{d}\cup_{j_i\in\mathbb{Z}}\bigtimes_{i=1}^{d}(j_i\,;\,j_i+1]$ conjuntamente a $\mu_L(\bigtimes_{i=1}^{d}(j_i\,;\,j_i+1]) = 1 < +\infty$.
- $\forall\, A \in \mathcal{A}, \quad \mu_L(A) = |A|$ volumen de A.
- Para una función g suficientemente "suave", la integración con respecto a la medida de Lebesgue coincide naturalmente con la integración de Riemann.

La medida de Lebesgue es así natural para la integración. Luego, en lo que sigue, al mencionar igualdad μ_L-casi siempre, diremos simplemente "casi siempre" $(c.s.)$, entendiendo que es con respecto a la medida de Lebesgue. De la misma manera, hablando de densidad, sin precisiones, se entenderá que es con respecto a μ_L.

Al "contrario" de la medida de Lebesgue, medidas discretas son también particulares. La más "elemental" es conocida como *medida de Dirac*, dando lugar a medidas discretas:

Definición 1-15 (Medida de Dirac y medida discreta)**.** *La medida de Dirac al punto* x_0*, denotada* δ_{x_0}*, es tal que*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad \delta_{x_0}(B) = \mathbb{1}_B(x_0)$$

Dado un conjunto $\mathcal{X} = \{x_i\}_i$ *discreto (finito o infinito numerable), llamaremos medida discreta a la medida definida por*

$$\mu_{\mathcal{X}} = \sum_i \delta_{x_i}$$

(en general, son definidas como combinaciones lineales positivas, siendo éste un caso particular).

Notar que,

- $\mu_{\mathcal{X}}$ es σ-finita (se muestra con el mismo enfoque que para la medida de Lebesgue).
- $\forall\, A \in \mathcal{A}, \quad \mu_{\mathcal{X}}(A) = |\mathcal{X} \cap A|$ cardenal de $\mathcal{X} \cap A$.
- Para una función g medible,

$$\int_{\mathbb{R}^d} g(x)\, d\delta_{x_k}(x) = g(x_k) \qquad \text{y} \qquad \int_{\mathbb{R}^d} g(x)\, d\mu_{\mathcal{X}}(x) = \sum_{x\in\mathcal{X}} g(x),$$

  luego la integración se vuelve una suma. Se prueba saliendo de g de la forma $g = \mathbb{1}_C$ y del Teorema 1-3 conjuntamente con las definiciones Def. 1-10 y Def. 1-15.

Con esta serie de definiciones, tenemos todo lo necesario para introducir la definición de variables/vectores aleatorios reales y sus caracterizaciones.

1.3.2 Variables aleatorias y vectores aleatorios. Distribución de probabilidad.

Empezamos con la noción de variable aleatoria real, que corresponde como el resultado de un experimento o de un evento dado (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988):

Definición 1-16 (Variable aleatoria real)**.** *Una variable aleatoria real es una función medible*

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$$

donde la medida P_X sobre $\mathcal{B}(\mathbb{R})$ es la medida imagen de P. P_X es frecuentemente llamada distribución de probabilidad o ley de la variable aleatoria X. En lo que sigue, escribiremos para cualquier $B \in \mathcal{B}(\mathbb{R})$ los eventos

$$(X \in B) \equiv X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\},$$

así que, por definición,

$$P_X(B) = P(X \in B).$$

Para ilustrar esta definición, tomando el ejemplo de un dado, Ω es discreto y representa las caras, mientras que los números (se asocia a cada cara un número real) serán la imagen de Ω por X (ej. $X(\omega_j) = j, \quad j = 1, \ldots, 6$).

Notar que, por las propiedades de una medida sobre una σ-álgebra, para caracterizar completamente la distribución P_X es suficiente conocerla sobre los intervalos de la forma $(-\infty\,;\, b]$. Esto da lugar a la definición de función de repartición (a veces llamada función distribución por abuso de denominación) (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988; Hogg et al., 2013):

Definición 1-17 (Función de repartición)**.** *La función de repartición F_X de una variable aleatoria X se define como*

$$F_X(x) = P_X((-\infty\,;\, x]) = P(X \leq x).$$

A veces, por abuso de lenguaje, se denomina a F_X ley de la variable aleatoria. Se encuentra también en la literatura la terminología función densidad cumulativa (cdf, por "cumulative density function" en inglés).

Naturalmente, de las propiedades de una medida de probabilidad se tiene:

- $0 \leq F_X(x) \leq 1$;

- $\lim_{x\to-\infty} F_X(x) = 0$ y $\lim_{x\to+\infty} F_X(x) = 1$ (viene de $P_X(\emptyset) = 0$ y $P_X(\mathbb{R}) = 1$);
- F_X es creciente $\big($viene de que $x_1 \leq x_2 \Leftrightarrow (-\infty\,;\, x_1] \subseteq (-\infty\,;\, x_2]\big)$;
- F_X no es necesariamente continua (lo vamos a ver más adelante); pero en cada punto x, F_X es continua por la derecha (ver inciso anterior).

Cuando se trabaja con $d \geq 2$ variables aleatorias es conveniente definir un *vector aleatorio* de dimensión d, y apelar para su estudio a nociones del álgebra lineal y a notación matricial. Se tiene el vector aleatorio d-dimensional $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$, caracterizado por d-uplas de variables aleatorias reales. Como en el caso univariado, se define este vector de la siguiente manera (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988):

Definición 1-18 (Vector aleatorio real)**.** *Un vector aleatorio real es una función medible*

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X).$$

donde $\mathcal{B}(\mathbb{R}^d)$ son los borelianos de $\mathbb{R}^d$, σ-álgebra generada por los productos cartesianos $(-\infty\,;\, b_1] \times \cdots \times (-\infty\,;\, b_d]$ y donde la medida P_X sobre $\mathcal{B}(\mathbb{R}^d)$ es la medida imagen de P llamada distribución de probabilidad de la variable aleatoria (o vector aleatorio) X. Como en el caso escalar, para cualquier $B \in \mathcal{B}(\mathbb{R}^d)$

$$(X \in B) \equiv X^{-1}(B) = \{\omega \in \Omega \;\big|\; X(\omega) \in B\} \qquad \textit{y} \qquad P_X(B) = P(X \in B).$$

Nota: a veces tenemos que considerar el caso de matrices aleatorias, o funciones medibles a valores matriciales. Dado que se puede poner en biyección una matriz con un vector (por ejemplo poniendo cada columna "debajo" de su columna antecedente). Volveremos más adelante en matriz aleatorias.

De las propiedades de una medida sobre una σ-álgebra, para caracterizar completamente la distribución P_X, de nuevo es suficiente conocerla sobre los elementos de la forma $\bigtimes_{i=1}^{d}(-\infty\,;\, b_i]$, *i. e.,* la función de repartición multivariada (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988; Hogg et al., 2013):

Definición 1-19 (Función de repartición multivariada)**.** *La función de repartición F_X de un vector aleatorio X es definida en $x = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}^d$ por*

$$F_X(x) = P_X\left(\bigtimes_{i=1}^{d}(-\infty\,;\, x_i]\right) = P\left(\bigcap_{i=1}^{d}(X_i \leq x_i)\right).$$

Por abuso de notación, escribiremos en lo que sigue

$$F_X(x) = P(X \leq x),$$

dando por entendido que $(X \leq x)$ es el evento $\bigcap_{i=1}^{d}(X_i \leq x_i)$.

De nuevo, de las propiedades de una medida de probabilidad surge:

- $0 \leq F_X(x) \leq 1$;
- $\lim\limits_{\forall i, x_i \to -\infty} F_X(x) = 0$ y $\lim\limits_{\forall i, x_i \to +\infty} F_X(x) = 1$;
- F_X es creciente con respecto a cada variable x_i.

Para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1; \ldots; d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ es obviamente un vector aleatorio k-dimensional. Es entonces sencillo ver que

$$F_{X_{I_k}}(x_{I_k}) = \lim_{\forall i \notin I_k, x_i \to +\infty} F_X(x)$$

$\Big($viene de que $\bigcap_{j=1}^k (X_{i_j} \leq x_{i_j}) = \Big(\bigcap_{j=1}^k (X_{i_j} \leq x_{i_j})\Big) \bigcap \Big(\bigcap_{i \notin I_k} (X_i \in \mathbb{R})\Big)\Big)$. Esta función se llama *función de repartición marginal* de F_X.

Cerramos estas generalidades con el caso de variables independientes:

Definición 1-20 (Independencia)**.** *Sean d variables aleatorias X_i y $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$. Las X_i son mutuamente independientes si y solamente si, para cualquier ensemble de conjuntos $B_i \in \mathcal{B}(\mathbb{R})$, los eventos $(X_i \in \mathcal{B}_i)$ son mutuamente independientes, i. e.,*

$$P_X\left(\bigtimes_{i=1}^d B_i\right) = \prod_{i=1}^d P_{X_i}(B_i).$$

Es equivalente a

$$F_X(x) = \prod_{i=1}^d F_{X_i}(x_i).$$

La ley del vector aleatorio se factoriza en este caso. Necesariamente, $\mathcal{X} = X(\Omega)$ es de la forma $\mathcal{X} = \bigtimes_{i=1}^d \mathcal{X}_i$ con $\mathcal{X}_i = X_i(\Omega)$, producto cartesiano.

Es importante notar que no es equivalente a tener la independencia por pares, como lo ilustramos al final de la sección precedente.

Más allá de este enfoque general, dos casos particulares de variables aleatorias son de interés: las variables discretas y las continuas. En el primer caso $X(\Omega)$ es discreto, finito o no. En las subsecciones siguientes estudiamos las particularidades de cada caso.

Para fijar notaciones, en todo lo que sigue escribiremos

$$\mathcal{X} = X(\Omega)$$

conjunto de llegada de X, o conjunto de valores que puede tomar la variable aleatoria. A veces, por simplicidad, se considera a $\mathcal{X}$ como el espacio muestral y se olvida que X sea una función medible entre espacios de probabilidades, *i. e.,* se trabaja en $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X)$ como en el espacio pre-imagen. Además, a veces, y por abuso, llamaremos frecuentemente a $\mathcal{X}$ *dominio de definición* de la medida de probabilidad P_X, siendo $P_X(\mathbb{R}^d \setminus \mathcal{X}) = 0$.

1.3.3 Variable aleatoria discreta

Definición 1-21 (Variable aleatoria discreta)**.** *Una variable aleatoria se dice discreta cuando $\mathcal{X} = X(\Omega)$ es discreto, finito o infinito numerable. En otras palabras, los posibles valores de una variable aleatoria discreta X consisten en un conjunto contable (finito o infinito numerable) de números reales, $\mathcal{X} = \{x_j\}$ y se puede escribir a X como una variable escalonada (ver ej. (Athreya & Lahiri, 2006; Hogg et al., 2013)),*

$$X = \sum_j x_j \mathbb{1}_{A_j} \quad \textit{con} \quad A_j = X^{-1}(\{x_j\}).$$

Notar que Ω no es necesariamente discreto. Por ejemplo, si ω es la posición de un punto sobre una linéa, y se tiene $X(\omega) = 0$ si ω está a la izquierda de un umbral y $X(\omega) = 1$ si ω está a su derecha, luego $\mathcal{X} = \{0\,;\,1\}$ mientras que Ω no es discreto.

En el caso de una variable aleatoria discreta X, las probabilidades $P_X(\{x_j\}) = P(X = x_j),\ x_j \in \mathcal{X}$ caracterizan completamente esta variable aleatoria (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Hogg et al., 2013):

Definición 1-22 (Función de masa de probabilidad)**.** *Se define la función de masa de probabilidad de X, variable aleatoria discreta tomando sus valores sobre $\mathcal{X}$ por*

$$p_X(x) \equiv P(X = x) = P_X(\{x\}) \quad x \in \mathcal{X}.$$

Por abuso de denominación, llamaremos en este libro a p_X distribución de probabilidad. Además, usaremos también la notación

$$p_X = \begin{bmatrix} \cdots & p_X(x_i) & \cdots \end{bmatrix}^t$$

que llamaremos vector de probabilidad, de tamaño $|\mathcal{X}|$, posiblemente infinito.

En la figura 1-4-(a) se muestra una representación gráfica de una distribución de probabilidad discreta.

Notar que siendo P_X una medida de probabilidad, $p_X \geq 0$ y está obviamente normalizada en el sentido de que

$$\sum_{x \in \mathcal{X}} p_X(x) = 1.$$

Dicho de otra manera, en el caso finito $|\mathcal{X}| = \alpha < +\infty$, el vector de probabilidad p_X partenece al simplex estandar o simplex de probabilidad $p_X \in \Delta_{\alpha-1}$ (ver notaciones).

Volviendo a la medida de probabilidad P_X se nota que

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \sum_{x \in \mathcal{X} \cap B} p_X(x) = \int_B dP_X(x),$$

lo que da para la función de repartición

$$F_X(x) = \sum_{x_j \leq x} p_X(x_j).$$

De esta forma, se justifica la denominación *cumulativa* para F_X. También, se puede ver de inmediato que F_X es una función discontinua, con saltos finitos (en x_j, salto de altura $p_X(x_j)$). Esto es ilustrado figura 1-4-(b).

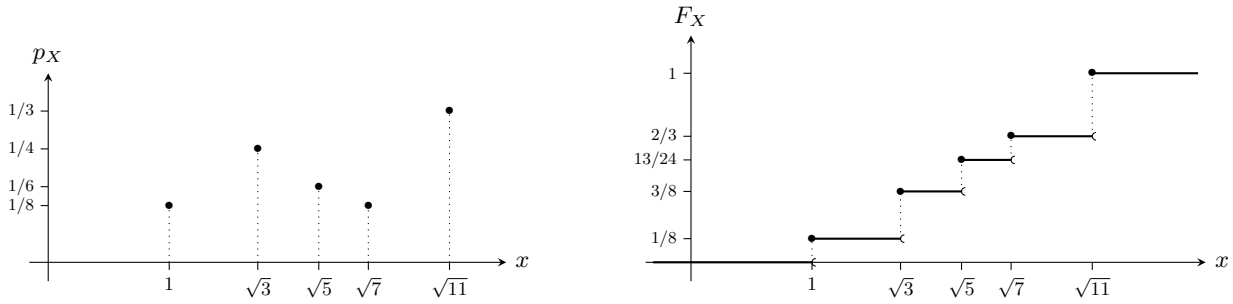


Figura 1-4: Ilustración de una distribución de probabilidad discreta (a), y la función de repartición asociada (b), con $\mathcal{X} = \{\, 1\,;\, \sqrt{3}\,;\, \sqrt{5}\,;\, \sqrt{7}\,;\, \sqrt{11}\,\}$ y $p_X = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} & \frac{1}{6} & \frac{1}{8} & \frac{1}{3} \end{bmatrix}^t$.

Un caso especial se tiene cuando un valor x_k es cierto o seguro, y no ocurre ninguno de los otros valores x_j $(j \neq k)$. La forma de la distribución es: $p_X(x) = 1$ si $x = x_k$ y 0 si no; el vector de probabilidad se escribirá $p_X = \mathbb{1}_k$ (ver notaciones; el vector posiblemente de dimensión infinita). Se denota también con el *símbolo de Kronecker* $\delta_{j,k} = 1$ si $j = k$ y 0 si no. En este libro, evitaremos usar este símbolo para no confundirlo con la medida de Dirac. Sin embargo, resuelta que P_X es precisamente la medida de Dirac en x_k (ver Def. 1-15).

Otra situación particular es la de *equiprobabilidad* o *distribución uniforme* cuando $|\mathcal{X}| = \alpha < +\infty$, $\alpha \in \mathbb{N}^*$. La forma de la distribución es: $p_X(x_j) = \frac{1}{\alpha} \quad \forall\, j = 1, \ldots, \alpha$, *i. e.,* $p_X = \begin{bmatrix} \frac{1}{\alpha} & \cdots & \frac{1}{\alpha} \end{bmatrix}^t$ o, en términos de medida, $P_X = \frac{1}{\alpha} \sum_{j=1}^{\alpha} \delta_{x_j}$. La función de repartición resulta una función escalonada, con saltos de altura $\frac{1}{\alpha}$ en cada x_j, $\quad j = 1\,,\, \ldots\,,\, \alpha$.

De manera general, la medida de probabilidad de una variable discreta se escribe como combinación convexa de medidas de Dirac,

$$P_X = \sum_j p_j \delta_{x_j}, \quad \text{con} \quad p_j = P(X = x_j) \geq 0, \quad \sum_j p_j = 1,$$

i. e., como una medida discreta.

Para comparar dos distribuciones discretas es útil reordenar cada vector de probabilidad permutando sus elementos hasta listarlos de forma decreciente. El vector reordenado a partir de p se anota $p^{\downarrow}$, de modo que $p_1^{\downarrow} \geq p_2^{\downarrow} \geq \ldots \geq p_\alpha^{\downarrow}$. En el ejemplo del caso con certeza se tiene $p^{\downarrow} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^t$,

mientras que la distribución uniforme no varía. La comparación de dos vectores de probabilidad se puede apoyar sobre la noción de mayorización:

Definición 1-23 (Mayorización)**.** *Un vector de probabilidad (distribución)* p *mayorizado por un vector de probabilidad (distribución)* q*, denotado* $p \prec q$*, se define como:*

$$p \prec q \qquad \textit{sii} \qquad \sum_{i=1}^{k} p_i^{\downarrow} \leq \sum_{i=1}^{k} q_i^{\downarrow} \quad \forall\, k = 1\,, \ldots \alpha - 1, \qquad \textit{y} \qquad \sum_{i=1}^{\alpha} p_i^{\downarrow} = \sum_{i=1}^{\alpha} q_i^{\downarrow}$$

(siendo las últimas sumas iguales a 1). Si los alfabetos de definición de p *y* q *son de tamaños diferentes,* α *es el tamaño más grande y la distribución sobre el alfabeto más corto es completada por estados de probabilidad 0 (sería equivalente a añadir estados ficticios de probabilidad nula).*

Por ejemplo, $\begin{bmatrix} 0{,}40 & 0{,}30 & 0{,}20 & 0{,}10 \end{bmatrix}^t \prec \begin{bmatrix} 0{,}50 & 0{,}30 & 0{,}15 & 0{,}05 \end{bmatrix}^t$ (ver Fig. 1-5-(a)).

Es importante resaltar que la mayorización provee un *orden parcial* (no total) entre distribuciones, existiendo pares de distribuciones tales que ninguna mayoriza a la otra. Por ejemplo, $\begin{bmatrix} 0{,}50 & 0{,}35 & 0{,}10 & 0{,}05 \end{bmatrix}^t$ y $\begin{bmatrix} 0{,}60 & 0{,}20 & 0{,}17 & ,03 \end{bmatrix}^t$ no se comparan por mayorización (ver Fig. 1-5-(b)).

Es interesante notar que la siguiente propiedad es válida para toda distribución p de tamaño α (Marshall, Olkin & Arnold, 2011, p. 9, (6)-(8)):

$$\begin{bmatrix} \frac{1}{\alpha} & \frac{1}{\alpha} & \cdots & \frac{1}{\alpha} \end{bmatrix}^t \prec p \prec \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^t .$$

En este sentido, los casos particulares de equiprobabilidad y de certeza, se dice que son distribuciones extremas. Notamos que uno implica ignorancia máxima en el resultado de la variable mientras que el otro corresponde a conocimiento completo. Veremos en el capítulo 2 el rol importante de este orden sobre las medidas informacionales.

La relación de mayorización es ilustrada en la figura 1-5, donde se representan las sumas parciales en función de k, llamadas *curvas de Lorentz* [13] (Marshall et al., 2011; Lorenz, 1905). Gráficamente, $p \prec q$ es equivalente a tener la curva de Lorentz asociada a p por debajo de la asociada a q.

1.3.4 Variable aleatoria continua

En varios contextos, una variable aleatoria puede tomar valores en un conjunto no numerable, por ejemplo cualesquiera de los números en cierto intervalo de la recta real. Ya no es una variable discreta. En las variables que no son discretas, el caso particular de interés es el de variables continuas (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Hogg et al., 2013):

[13] Se prueba sencillamente que estas curvas son crecientes y cóncavas.

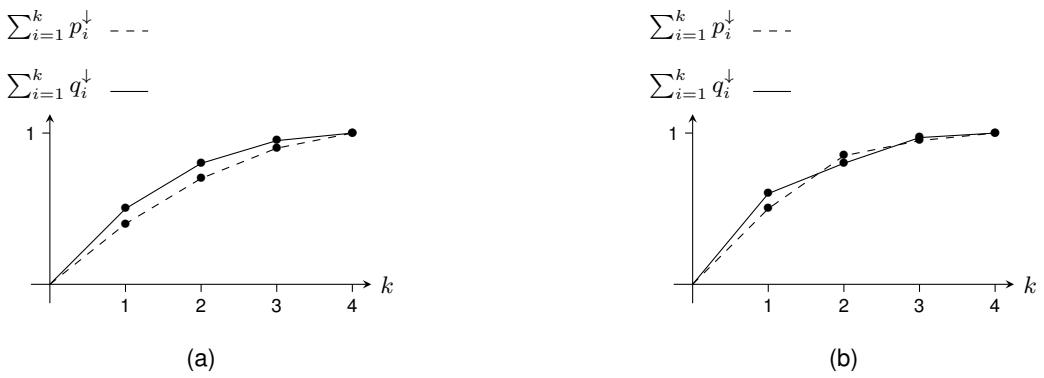


Figura 1-5: Orden parcial por mayorización: sumas parciales para $k = 1\,,\,2\,,\,3\,,\,4$ (a) para los vectores de probabilidades $p = \begin{bmatrix}0{,}40 & 0{,}30 & 0{,}20 & 0{,}10\end{bmatrix}^t$ (línea punteada) y $q = \begin{bmatrix}0{,}50 & 0{,}30 & 0{,}15 & 0{,}05\end{bmatrix}^t$ (línea llena) y (b) para los vectores de probabilidad $p = \begin{bmatrix}0{,}50 & 0{,}35 & 0{,}10 & 0{,}05\end{bmatrix}^t$ (línea punteada) y $q = \begin{bmatrix}0{,}60 & 0{,}20 & 0{,}17 & 0{,}03\end{bmatrix}^t$ (línea llena). En el caso (a), $p \prec q$ mientras que en el caso (b), $p \not\prec q$ y $q \not\prec p$ (no están ordenadas por mayorización).

Definición 1-24 (Variable aleatoria continua)**.** *Una variable aleatoria X se dice continua si su función de repartición F_X es continua sobre $\mathbb{R}$.*

Cuando se puede, es conveniente asociar una *función densidad de probabilidad* (comúnmente anotada por su sigla en inglés: pdf, por *probability density function*). La definición de tal densidad se apoya en la definición 1-12 aplicada a la medida de probabilidad P_X:

Definición 1-25 (Variable aleatoria que admite una densidad de probabilidad)**.** *Sea X una variable aleatoria continua y P_X su medida de probabilidad. Por definición, se dice que X admite una densidad de probabilidad con respecto a una medida μ sobre $\mathbb{R}$ si $P_X \ll \mu$ (teorema de Radon-Nikodým 1-8).*
En general, nos enfocamos en la medida (llamada de referencia) $\mu = \mu_L$ de Lebesgue. Denotando $d\mu_L(x) \equiv dx$, la definición se reduce a: Si existe una función no negativa p_X medible sobre $\mathbb{R}$ tal que

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \int_B p_X(x)\, dx,$$

entonces se dice que X admite una densidad y p_X es llamada densidad de probabilidad de X (dando por entendido "con respecto a la medida de Lebesgue"). Notando que $P_X(B) = P_X(B \cap \mathcal{X})$, el soporte de p_X es necesariamente $\mathcal{X} = X(\Omega)$ (i. e., $p_X(\overline{\mathcal{X}}) = \{0\}$ y $p_X(\mathcal{X}) > 0$), y

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \int_{B\cap\mathcal{X}} p_X(x)\, dx.$$

Para la función de repartición F_X tenemos entonces

$$F_X(x) = \int_{-\infty}^{x} p_X(u)\, du$$

(esta expresión es valida para cualquier medida μ, densidad con respecto a esta medida de referencia, e integración sobre $(-\infty\,;\,x]$ con el "diferencial" $d\mu(x)$). Dicho de otra manera, si F_X es (continua y)

derivable sobre $\mathbb{R}$, al menos por partes, X admite una densidad de probabilidad (con respecto a la medida de Lebesgue) y [14]

$$p_X(x) = \frac{dF_X(x)}{dx}.$$

Por abuso de terminología, en lo que sigue llamaremos a p_X también distribución de probabilidad, a pesar de que no tiene el mismo sentido que la masa de probabilidad del caso discreto.

La escritura integral de F_X justifica de nuevo la denominación *cumulativa* para F_X. Además, se puede ver por ejemplo que en este caso $P(a < X \leq b) = \int_a^b p_X(x)\,dx = F_X(b) - F_X(a)$ y que claramente

$$\forall x \in \mathbb{R}, \quad P_X(\{x\}) = P(X = x) = 0,$$

esto es, $\{x\}$ es de medida P_X nula. Similarmente, cualquier conjuntos numerable de $\mathbb{R}$ es de medida P_X nula.

Notar que aún cuando $0 \leq F_X \leq 1$, p_X puede ser mayor que uno. Por ejemplo, para $F_X(x) = 2\,x\,\mathbb{1}_{[0\,;\,\frac{1}{2})}(x) + \mathbb{1}_{[\frac{1}{2}\,;\,+\infty)}(x)$, que define una función de repartición, $p_X(x) = 2\mathbb{1}_{[0\,;\,\frac{1}{2})}(x)$. No es contradictorio en el sentido de que p_X no es una probabilidad, sino que $p_X(x)\,dx$ puede ser visto como la probabilidad de hallar a la variable con valores en el "intervalo infinitesimal entre x y $x+dx$". Finalmente, la condición de normalización se escribe

$$\int_{\mathcal{X}} p_X(x)\,dx = \int_{\mathbb{R}} p_X(x)\,dx = 1.$$

En la figura 1-6-(a) se muestra una representación gráfica de una función densidad de probabilidad para una variable continua que admite una densidad, y en la figura 1-6-(b) la función de repartición correspondiente.

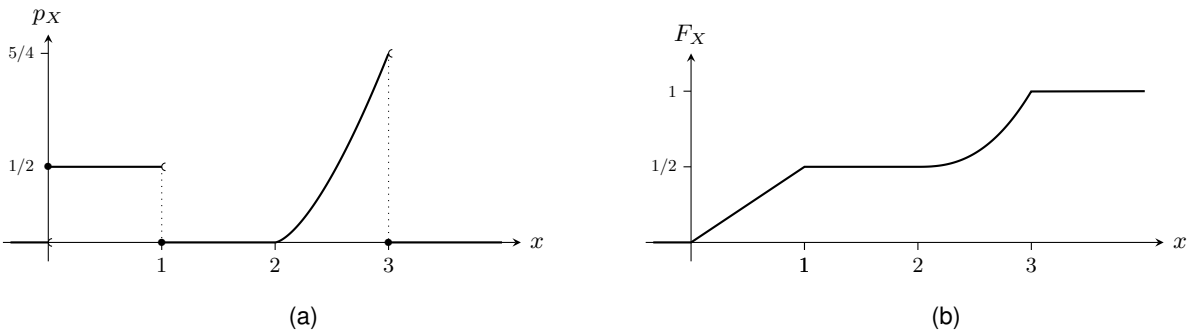


Figura 1-6: Ilustración de: (a) una distribución de probabilidad continua, y (b) la función de repartición asociada, con $\mathcal{X} = [0\,;\,1) \cup [2\,;\,3)$ y $p_X(x) = \frac{1}{2}\mathbb{1}_{[0\,;\,1)}(x) + \frac{5(x-2)^{\frac{3}{2}}}{4}\mathbb{1}_{[2\,;\,3)}(x)$, *i. e.*, $F_X(x) = \frac{x}{2}\,\mathbb{1}_{[0\,;\,1)}(x) + \frac{1}{2}\mathbb{1}_{[1\,;\,2)}(x) + \frac{1+(x-2)^{\frac{5}{2}}}{2}\mathbb{1}_{[2\,;\,3)}(x) + \mathbb{1}_{[3\,;\,+\infty)}(x)$.

[14] Recordar que, rigurosamente, la igualdad debe ser entendida "casi siempre".

Notar que una variable aleatoria puede no ser ni continua, ni discreta, como se ilustra en el ejemplo siguiente:

Ejemplo 1-4 (Ejemplo de variable mixta)**.** *Sean* U *y* V *variables continuas, independientes, de densidad de probabilidad* $p_U = p_V = \mathbb{1}_{[0\,;\,1)}$ *(*U *y* V *son uniformes sobre* $[0\,;\,1)$*) y sea* $X = V\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$*, es decir* $X(\omega) = V(\omega)$ *si* $U(\omega) < \frac{1}{2}$ *y* 1 *si no. Entonces de la fórmula de probabilidades totales,* $F_X(x) = P(X \leq x) = P\left((X \leq x)\,\middle|\,\left(U < \frac{1}{2}\right)\right) P\left(U < \frac{1}{2}\right) + P\left((X \leq x)\,\middle|\,\left(U \geq \frac{1}{2}\right)\right) P\left(U \geq \frac{1}{2}\right)$ *i. e.,* $F_X(x) = \frac{1}{2}P\left((V \leq x)\,\middle|\,\left(U < \frac{1}{2}\right)\right) + \frac{1}{2}P\left((1 \leq x)\,\middle|\,\left(U \geq \frac{1}{2}\right)\right)$*. Ahora, de la independencia de* U *y* V*, tenemos* $F_X(x) = \frac{1}{2}F_V(x) + \frac{1}{2}\mathbb{1}_{[1\,;\,+\infty)}(x)$ *es decir*

$$F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;\,+\infty)}(x).$$

Esta función de repartición es representada en la figura 1-7: no es discreta ni continua. Entonces, a pesar de que $\mathcal{X} = [0\,;\,1]$ *sea un intervalo real,* X *no es continua (y tampoco puede ser discreta).*

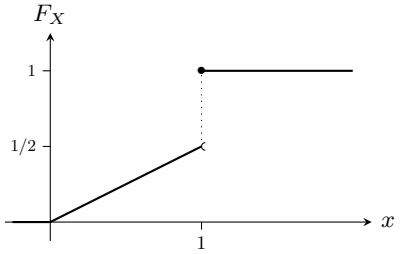


Figura 1-7: Función de repartición $F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;\,+\infty)}(x)$ asociada a $X = V\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ con U y V variables independientes, continuas, uniformes sobre $\mathcal{X} = [0\,;\,1)$. F_X no es tipo escalón, así que X no es discreta. A pesar de que $\mathcal{X} = [0\,;\,1]$ es un intervalo, de la presencia del salto en $x = 1$ tampoco X es continua.

Volvemos a las variables discretas X sobre $\mathcal{X} = \{x_j\}_j$, de medida de probabilidad de la forma $P_X = \sum_j p_j\delta_{x_j}$. Considerando la medida discreta $\mu_{\mathcal{X}} = \sum_j \delta_{x_j}$, es claro que $P_X \ll \mu_{\mathcal{X}}$. Entonces, formalmente, P_X admite una densidad con respecto a la medida discreta $\mu_{\mathcal{X}}$ y esta densidad, definida sobre $\mathcal{X}$ es $p_X(x) = P(X = x)$. A pesar de que sea una tautología, esto justifica que usemos la escritura p_X (minúscula) tanto en el caso discreto como en el caso continuo, y que hablemos (por abuso de terminología) de distribución de probabilidad en ambos casos.

Recordemos que cualquier medida de probabilidad (caso continuo o no) se escribe también con una integral $P_X(B) = \int_B dP_X(x)$ y que en el caso discreto cierto $X = x_k$, la medida de probabilidad P_X es la medida de Dirac. A veces, por abuso de escritura $dP_X(x)$ es denotado $\delta_{x_k}(x)\,dx$ o $\delta(x - x_k)\,dx$ donde ahora δ es llamada *distribución (delta) de Dirac*. Se puede ver este Dirac como una densidad de probabilidad $p_X(x)$ con respecto a la medida de Lebesgue pero no es una función "ordinaria" dado que P_X no es diferenciable con respecto a la medida de Lebesgue. Se la llama *función generalizada*

o *distribución de Schwartz* [15]. En particular, $F_X(x) = \mathbb{1}_{\mathbb{R}_+}(x - x_k)$ ($\mathbb{1}_{\mathbb{R}_+}$ es conocido también como *función de Heaviside* [16]) y en el sentido de las distribuciones, $\frac{dF_X}{dx} = \delta_{x_k}$. Además, se usan en general las propiedades, para cualquier function f y real x_0,

$$f(x)\delta(x - x_0) = f(x_0)\delta(x - x_0) \qquad \text{y} \qquad \int_{\mathbb{R}} f(x)\delta(x - x_0)\, dx = f(x_0),$$

pero hay que entender la integración a través de la medida Dirac (insistimos en el hecho de que esta notación es un abuso de escritura, ej. (Gel'fand & Shilov, 1964)). Usando las distribuciones de Dirac, se puede unificar el tratamiento de las variables aleatorias discretas con las continuas en términos de densidad (con respecto a la medida de Lebesgue): si una variable aleatoria discreta toma los valores x_j con probabilidades $p_j = P(X = x_j)$ respectivamente, entonces formalmente se puede describir mediante una variable aleatoria continua X con "función densidad de probabilidad" $p_X(x) = \sum_j p_j\, \delta(x - x_j)$. Insistimos en el hecho de que rigurosamente debemos trabajar con medidas, como lo hemos formalizado al principio de este capítulo.

Terminamos mencionando el resultado siguiente, probado en (Athreya & Lahiri, 2006, Ec. (2.5) p. 47 & teorema 4.1.1) por ejemplo [17]:

Teorema 1-9 (Descomposición de una medida de probabilidad)**.** *Cualquier función de repartición $F_X(x)$ se descompone como combinación convexa de una función de repartición F_{d} discreta y una función de repartición F_{c} continua:*

$$\exists\, a \in [0\,;\,1] \quad \textit{tal que} \quad F_X(x) = aF_{\mathrm{d}}(x) + (1 - a)F_{\mathrm{c}}(x)$$

En términos de medida, o como corolario de (Athreya & Lahiri, 2006, teorema 4.1.1), cualquier medida de probabilidad P_X se descompone como la combinación convexa de una medida discreta P_{d} y una continua P_{c},

$$\exists\, a \in [0\,;\,1],\ \widetilde{\mathcal{X}} \textit{ discreto} \quad \textit{tal que} \quad P_X = aP_{\mathrm{d}} + (1 - a)P_{\mathrm{c}} \qquad \textit{con} \qquad P_{\mathrm{d}} \ll \mu_{\widetilde{\mathcal{X}}} \quad \textit{y} \quad P_{\mathrm{c}} \ll \mu_L$$

Entonces, $P_X \ll \mu_{\widetilde{\mathcal{X}}} + \mu_L$, i. e., admite una densidad con respecto a la medida σ-finita $\mu_{\widetilde{\mathcal{X}}} + \mu_L$.

---

[15] La teoría de distribuciones valió a Laurent Schwarz la medalla Fields en 1950. Entre otros trabajos, se probó que el Dirac, visto como distribución de Schwartz o función generalizada, tiene una "representación integral" $\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\imath tx} dt$ o más rigurosamente transformada de Fourier de $x \mapsto 1$ en el sentido de las funciones generalizadas o distribuciones. Esto muestra claramente su carácter no ordinario (la integral siendo divergente en el sentido usual). Esto va más allá de la meta del capítulo y el lector se podrá referir a (Schwartz, 1966; Gel'fand & Shilov, 1964, 1968) por ejemplo.

[16] Viene del nombre del físico ingles Oliver Heaviside quien estudio las ecuaciones electromagnéticas de Maxwell.

[17] Básicamente, se muestra que $\widetilde{X} = \{x \in \mathcal{X} \mid \ p(x) = F_X(x) - \liminf_{u \to x} F(u) > 0\}$ es numerable. A continuación, $F(x) - \sum_{\widetilde{x} \in \widetilde{\mathcal{X}}} p(\widetilde{x})\mathbb{1}_{(-\infty\,;\,\widetilde{x}]}(x)$ es continua, y se recupera la descomposición con $a = \sum_{\widetilde{x} \in \widetilde{\mathcal{X}}} p(\widetilde{x})$.

Dicho de otra manera, cualquier variable aleatoria es mixta, como en el ejemplo 1-4. Del teorema, es discreta cuando $a = 1$, y continua cuando $a = 0$.

1.3.5 Vector aleatorio discreto

Un ejemplo de vector aleatorio discreto puede ser dado por un conjunto de dados (que podrían ser dependientes si están ligados por un hilo por ejemplo).

Definición 1-26 (Vector aleatorio discreto)**.** *Un vector aleatorio d-dimensional se escribe $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ y $\mathcal{X} = X(\Omega) \subset \bigtimes_{i=1}^{d} \mathcal{X}_i$ donde $\mathcal{X}_i = X_i(\Omega)$. Se dice que X es discreto cuando $\mathcal{X} \subseteq \mathbb{N}^d$, es discreto, finito o infinito numerable.*

Obviamente, la medida de probabilidad en los $x = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}^t \in \bigtimes_{i=1}^{d} \mathcal{X}_i$ caracteriza completamente este vector aleatorio:

Definición 1-27 (Función de masa de probabilidad conjunta)**.** *Por definición, la función de masa de probabilidad de X, vector aleatorio discreto que toma sus valores sobre $\mathcal{X} \subset \bigtimes_i \mathcal{X}_i$, está dada por*

$$p_X(x) \equiv P(X = x) = P\left(\bigcap_{i=1}^{d} (X_i = x_i)\right) \quad \forall\, x_i \in \mathcal{X}_i,\ 1 \leq i \leq d.$$

Se la llama también función de masa de probabilidad conjunta de los X_i o, por abuso de denominación, llamaremos a p_X distribución de probabilidad (conjunta). Notar que $\mathcal{X}$ no es necesariamente igual al producto cartesiano de los $\mathcal{X}_i$.
En el caso multivariado, la notación vectorial es más delicada de usar: p_X sería un "tensor" (o tabla) d-dimensional (una matriz para $d = 2$, una "tabla" 3 dimensonal para $d = 3$,... ; ver notaciones). Pero es posible usar una notación vectorial, recordando que $\mathbb{N}^d$ puede ser puesto en biyección con $\mathbb{N}$, y dada una biyección usarla para etiquetar los componentes de p_X puestos en vector. En el caso finito $\mathcal{X}_i = \{x_{j_i}\}_{j=1}^{\alpha_i}$ con $\alpha_i = |\mathcal{X}_i| < +\infty$, se puede organizar los componentes tales que $p_X(x_{j_1}, \ldots, x_{j_d})$ sea la j-ésima componente del vector p_X con $j = 1 + \sum_{i=1}^{d} (j_i - 1) \prod_{k=i+1}^{d} \alpha_k$.

De nuevo, se puede interpretar p_X como densidad con respecto a la medida discreta $\mu_{\mathcal{X}}$, Def. 1-15.

Similarmente al caso escalar $d = 1$, la función de repartición de un vector aleatorio discreto d-dimensional es tipo escalón d-dimensional, *i. e.,* compuesto de partes de hiperplanos d-dimensionales, F_X constante sobre $[x_{(j-1)_1}\,;\, x_{j_1}) \times \cdots \times [x_{(j-1)_d}\,;\, x_{j_d})$. Además, las componentes son mutuamente independientes si y solamente si la función de répartición se factoriza, o equivalentemente la función de masa se factoriza, *i. e.,*

$$X_i \text{ mutuamente independientes} \quad \Leftrightarrow \quad p_X(x) = \prod_{i=1}^{d} p_{X_i}(x_i).$$

En notación tensorial, $p_X = p_{X_1} \otimes \cdots \otimes p_{X_d}$, producto externo (ver notaciones).

Al final, de la fórmula de calculo de función de repartición marginal vista en la sección 1.3.2, para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1\,;\,\ldots\,;\,d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ tiene como probabilidad marginal o distribución marginal

$$p_{X_{I_k}}(x_{I_k}) = \sum_{\forall i \notin I_k, x_i \in \mathcal{X}_i} p_X(x).$$

1.3.6 Vector aleatorio continuo

Como para el caso de una variable, se puede considerar cualquier medida de referencia μ sobre $\mathbb{R}^d$ para definir una noción de densidad (d-variada), pero en general nos enfocamos en la medida de Lebesgue.

Definición 1-28 (Vector aleatorio continuo y densidad de probabilidad multivariada)**.** *Un vector aleatorio $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ se dice continuo si su función de repartición F_X es continua sobre $\mathbb{R}^d$. Como en el caso escalar, por definición 1-12 (y la recíproca evocada después de la definición), se dice que X admite una densidad de probabilidad p_X con respecto a una medida μ sobre $\mathbb{R}^d$ si $P_X \ll \mu$. De nuevo, nos enfocamos en la medida (llamada de referencia) $\mu = \mu_L$ de Lebesgue: si existe una función no negativa y medible $p_X : \mathbb{R}^d \mapsto \mathbb{R}$ tal que*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_X(B) = \int_B p_X(x)\, dx = \int_{B \cap \mathcal{X}} p_X(x)\, dx$$

con $\mathcal{X} = X(\Omega)$ soporte de p_X y $d\mu_L(x) \equiv dx = dx_1 \cdots dx_d$, entonces se dice que X admite una densidad y p_X es llamada densidad de probabilidad de X (entendiendo "con respecto a la medida de Lebesgue"), o también densidad de probabilidad conjunta de los X_i. En particular,

$$F_X(x) = \int_{\times_{i=1}^d (-\infty\,;\,x_i]} p_X(u)\, du$$

o, equivalentemente, para F_X (continua y) derivable sobre $\mathbb{R}^d$ (con respecto a la medida de Lebesgue), por lo menos por partes,

$$p_X(x) = \frac{\partial^d F_X(x)}{\partial x_1 \cdots \partial x_d}.$$

Usaremos la terminología (por abuso) de distribución de probabilidad.

Como en el caso escalar, $p_X \geq 0$, pero no es necesario que sea menor que 1, y satisface la condición de normalización

$$\int_{\mathcal{X}} p_X(x)\, dx = \int_{\mathbb{R}^d} p_X(x)\; dx = 1.$$

El teorema 1-9 se cumple también en el caso d-dimensional: cualquier medida de probabilidad P_X (resp. función de repartición F_X) se descompone como combinación convexa de una medida de

probabilidad (resp. función de repartición) discreta y una continua. En otros términos existe un $\widetilde{X}$ discreto tal que $P_X \ll \mu_{\widetilde{\mathcal{X}}} + \mu_L$ (σ-finita).

Mencionamos que las d variables aleatorias $X_1, \ldots, X_d$, componentes de un vector aleatorio X, son independientes si y solamente si la función de repartición se factoriza, lo que da, derivando esta,

$$X_i \text{ mutuamente independientes} \quad \Leftrightarrow \quad p_X(x) = \prod_{i=1}^{d} p_{X_i}(x_i).$$

Seguimos esta sección mencionando que, de la fórmula de cálculo de función de repartición marginal vista en la sección 1.3.2, para un subconjunto $I_k = (i_1\,, \ldots\,, i_k)$ de $1 \leq k \leq d$ elementos de $\{1\,; \ldots\,; d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ tiene la *densidad de probabilidad marginal*

$$p_{X_{I_k}}(x_{I_k}) = \int_{\times_{i \notin I_k} \mathcal{X}_i} p_X(x) \prod_{i \notin I_k} dx_i = \int_{\mathbb{R}^{d-k}} p_X(x) \prod_{i \notin I_k} dx_i.$$

En particular, la función densidad de probabilidad marginal que caracteriza a la variable aleatoria X_i es la ley que se obtiene integrando la densidad de probabilidad conjunta sobre todas las variables excepto la i-ésima.

Como en el caso discreto, se puede querer comparar dos distribuciones de probabilidad, y por eso reordenar o rearreglar una densidad de probabilidad. Como en el caso discreto, denotamos $p_X^{\downarrow}$ a la densidad con rearreglo simétrico. Primero, se necesita definir el rearreglo simétrico de un ensemble y luego de una densidad de probabilidad:

Definición 1-29 (Rearreglo simétrico de un conjunto)**.** *Sea $\mathcal{P} \subset \mathbb{R}^d$ abierto de volumen finito, $|\mathcal{P}| < +\infty$. El rearreglo simétrico $\mathcal{P}^{\downarrow}$ de $\mathcal{P}$ es la bola centrada en 0 de igual volumen que $\mathcal{P}$, i. e.,*

$$\mathcal{P}^{\downarrow} = \left\{ x \in \mathbb{R}^d \;\middle|\; \frac{2\pi^{\frac{d}{2}} \|x\|^d}{\Gamma\left(\frac{d}{2}\right)} < |\mathcal{P}| \right\} = \mathbb{B}_d \left(0\,,\, \frac{1}{\sqrt{\pi}} \left(\frac{|\mathcal{P}| \Gamma\left(\frac{d}{2}\right)}{2} \right)^{\frac{1}{d}} \right),$$

Esto se ilustra en la figura 1-8-a.

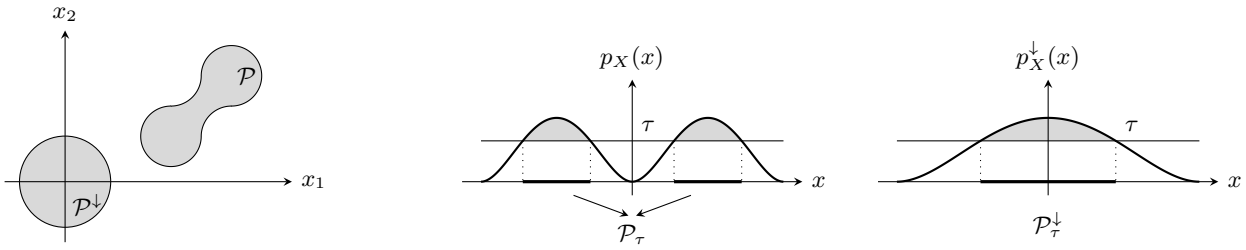
Definición 1-30 (Rearreglo simétrico de una densidad de probabilidad)**.** *Sea p_X una densidad de probabilidad y sean $\mathcal{P}_t = \{y \;|\; p_X(y) > t\} \subset \mathbb{R}^d$ para cualquier $t > 0$, sus conjuntos de niveles. La densidad de probabilidad con rearreglo simétrico $p_X^{\downarrow}$ de p_X se define como* [18]

$$p_X^{\downarrow}(x) = \int_0^{+\infty} \mathbb{1}_{\mathcal{P}_u^{\downarrow}}(x)\, du.$$

De $\forall\, t < \tau \Leftrightarrow \mathcal{P}_\tau \subseteq \mathcal{P}_t \Leftrightarrow \mathcal{P}_\tau^{\downarrow} \subseteq \mathcal{P}_t^{\downarrow}$, es sencillo ver que si $x \in \mathcal{P}_\tau^{\downarrow}$, entonces $x \in \mathcal{P}_t^{\downarrow}$, lo que conduce a $p_X^{\downarrow}(x) > \tau$, y vice-versa. Más allá, sobre $\mathcal{P}_{\tau+d\tau} \setminus \mathcal{P}_\tau$ la función p_X "vale" τ y sobre $\mathcal{P}_{\tau+d\tau}^{\downarrow} \setminus \mathcal{P}_\tau^{\downarrow}$ la

[18] Se prueba que esta función, positiva por definición, suma a 1. Además, por construcción, depende únicamente de $\|x\|$ y decrece con $\|x\|$.

función $p_X^{\downarrow}$ "vale" también τ, lo que da $\int_{\mathcal{P}_\tau^{\downarrow}} p_X^{\downarrow}(x)\,dx = \int_{\mathcal{P}_\tau} p_X(x)\,dx$ (ver (Lieb & Loss, 2001; Wang & Madiman, 2004) para une prueba más rigurosa). La representación de la definición es conocida como representación tarta en capas ("layer cake representation" en ingles). Esto es ilustrado en la figura 1-8-b para el caso escalar $d = 1$. Notar que, por construcción, $p_X^{\downarrow}$ cae el la familia de densidades esféricas



(a) (b)

Figura 1-8: (a): Ilustración del rearreglo simétrico $\mathcal{P}^{\downarrow}$ de un conjunto $\mathcal{P}$, siendo $\mathcal{P}^{\downarrow}$ la bola centrada en 0 de mismo volumen que $\mathcal{P}$, en el caso bi-dimensional $d = 2$. (b) Construcción del rearreglo $p_X^{\downarrow}$ en un contexto escalar (ver ejemplo 1-5 para $d = 1,\ \nu = 5,\ m = 1$): dado un τ, se busca $\mathcal{P}_\tau$ (izquierda) y se deduce $\mathcal{P}_\tau^{\downarrow}$ (derecha); dado un x, se busca el mayor t tal que $x \in \mathcal{P}_t^{\downarrow}$, siendo entonces este t máximo igual a $p_X^{\downarrow}(x)$ (derecha); además, por construcción, las superficies en grise son iguales.

(ver sección 1.10) (Lord, 1954; Fang, Kotz & Ng, 1990; Cambanis, Huang & Simons, 1981; Eaton, 1981).

A partir de esta definición del rearreglo, se puede ahora extender la noción de mayorización del caso discreto al caso continuo de la manera siguiente:

Definición 1-31 (Mayorización en el contexto continuo)**.** *Una densidad de probabilidad p se dice mayorizada por una distribución q sii:*

$$p \prec q \quad \textit{sii} \quad \int_{\mathbb{B}_d(0,r)} p^{\downarrow}(x)\,dx \le \int_{\mathbb{B}_d(0,r)} q^{\downarrow}(x)\,dx \quad \forall\, r > 0, \quad \textit{y} \quad \int_{\mathbb{R}^d} p^{\downarrow}(x)\,dx = \int_{\mathbb{R}^d} q^{\downarrow}(x)\,dx,$$

(las últimas integrales son obviamente iguales a 1).

Nota: la función

$$\mathcal{L}_p(r) = \int_{\mathbb{B}_d(0,r)} p^{\downarrow}(x)\,dx$$

da el equivalente de la curba de Lorentz en el contexto continuo, así que la relación de mayorización se interpreta graficamente de la misma manera que en el caso discreto (excepto que, contrariamente al caso discreto, la curva no es necesariamente concava).

Ejemplo 1-5. *Consideramos la densidad de probabilidad d-dimensional mezcla* [19] *de Student-r (ver*

[19] Una mezcla de ley es definida como combinación convexa $f = \sum_{i=1}^{k} \alpha_i f_i$ de leyes f_i con $\alpha = \begin{bmatrix} \alpha_1 & \cdots & \alpha_k \end{bmatrix}^t \in \Delta_{k-1}$.

sección 1.10.2.10)

$$p_X(x) = \alpha \left(\left(1 - \|x+m\|^2\right)_+^{\frac{\nu-d}{2}} + \left(1 - \|x-m\|^2\right)_+^{\frac{\nu-d}{2}} \right)$$

con

$$\nu > d-2, \qquad m \in \mathbb{R}^d \setminus \mathbb{B}_d \qquad \textit{y} \qquad \alpha = \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)} \quad \textit{coeficiente de normalización}$$

y $(\cdot)_+ = \max(x, 0)$ *(ver notaciones). Esta densidad de probabilidad es dibujada figura 1-8-((b) izquierda) para* $d = 1,\ \nu = 5, m = 1$ *y figura 1-9-(a) para* $d = 2,\ \nu = 6,\ m = \frac{1}{\sqrt{d}}\mathbb{1}$*. El dominio de definición, el máximo, y la matriz de covarianza (ver sección 1.10.2.10) son dados por*

$$\mathcal{X} = \mathbb{B}_d(-m, 1) \cup \mathbb{B}_d(m, 1), \qquad \max_{\mathcal{X}} p_X(x) = \alpha, \qquad \Sigma_X = \frac{1}{\nu+2} I + mm^t$$

Ahora, para $\tau > \alpha$, $\mathcal{P}_\tau = \emptyset$ *y para cualquier* $\tau \in [0\,;\,\alpha]$ *buscando los* x *tal que* $p_X(x) > \tau$ *(notando que las bolas en* $\mathcal{X}$ *son disjuntas) conduce a*

$$\mathcal{P}_\tau = \mathbb{B}_d(-m, \beta_\tau) \cup \mathbb{B}_d(m, \beta_\tau) \qquad \textit{con} \qquad \beta_\tau = \sqrt{1 - \left(\frac{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)\,\tau}{\Gamma\left(\frac{\nu}{2}+1\right)} \right)^{\frac{2}{\nu-d}}}$$

Notando que las bolas que constituan $\mathcal{P}_\tau$ *son disjuntas, queda claro que el volumen de* $\mathcal{P}_\tau$ *es dado por* $|\mathcal{P}_\tau| = 2\ |\mathbb{B}_d(0, \beta_\tau)| = \left|\mathbb{B}_d\left(0, 2^{\frac{1}{d}}\beta_\tau\right)\right|$*, que conduce al dominio rearreglado,*

$$\mathcal{P}_\tau^{\downarrow} = \mathbb{B}_d\left(0, 2^{\frac{1}{d}}\beta_\tau\right)$$

Ahora, se muestra sencillamente que $x \in \mathcal{P}_u^{\downarrow}$ *es equivalente a*

$$u < \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)} \left(1 - \frac{\|x\|^2}{4^{\frac{1}{d}}}\right)_+^{\frac{\nu-d}{2}}$$

De la definición 1-30 obtenemos al final

$$p_X^{\downarrow}(x) = \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)} \left(1 - \frac{\|x\|^2}{4^{\frac{1}{d}}}\right)_+^{\frac{\nu-d}{2}}$$

dibujada figura 1-8-((b) derecha) para $d = 1, \nu = 5$ *y figura 1-9 para* $d = 2, \nu = 6$*. Pasando, se reconoce una ley Student-*r *con un grado de libertad* ν*.*

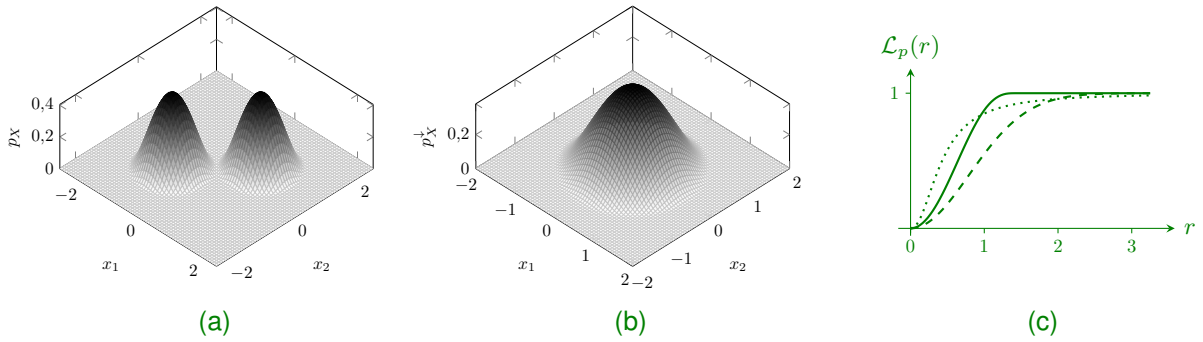
Finalmente, la curva de Lorentz es dada por

$$\begin{aligned} \mathcal{L}_{p_X}(r) &= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)} \int_{\mathbb{B}_d(0,r)} \left(1 - \frac{\|x\|^2}{4^{\frac{1}{d}}}\right)^{\frac{\nu-d}{2}} \mathbb{1}_{\mathbb{B}_d\left(0,2^{\frac{1}{d}}\right)}(x)\,dx \\ &= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)} \frac{2\,\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^r \rho^{d-1} \left(1 - \frac{\rho^2}{4^{\frac{1}{d}}}\right)^{\frac{\nu-d}{2}} \mathbb{1}_{\left[0\,;\,2^{\frac{1}{d}}\right]}(\rho)\,d\rho \end{aligned}$$

pasando en coordenadas hiperesférica (ver (Lord, 1954; Fang et al., 1990; Cambanis et al., 1981) y sección 1.10). Por el cambio de variable $u = \frac{\rho^2}{4^{\frac{1}{d}}}$ *obtenemos*

$$\mathcal{L}_{p_X}(r) = \frac{\displaystyle\int_0^{\frac{r^2}{4^{\frac{1}{d}}}} u^{\frac{d}{2}-1}(1-u)^{\frac{\nu-d}{2}}\,du}{B\left(\frac{d}{2}\,,\,\frac{\nu-d}{2}+1\right)}$$

conocida como función beta incompleta (Gradshteyn & Ryzhik, 2015, Ec. 8.392), tomada en $\frac{r^2}{4^{\frac{1}{d}}}$. La figura 1-9-(c) dibuja $\mathcal{L}_{p_X}(r)$ para $d = 2$, $\nu = 6$, la de la ley gausiana esférica g_X de covarianza $\frac{\mathrm{Tr}\,\Sigma_X}{d} I$ y la de la ley Student-t esférica s_X de covarianza $\frac{\mathrm{Tr}\,\Sigma_X}{d} I$ y con $\nu' = 2{,}25$ grado de libertad (ver sección 1.10). Eso ilustra graficamente la relación de mayorización $g_X \prec p_X$ y que ambas $s_X \not\prec p_X$ y $p_X \not\prec s_X$.



(a) (b) (c)

Figura 1-9: (a) Densidad de probabilidad p_X del ejemplo 1-5 en el contexto bi-dimensional $d = 2$ y para $\nu = 6$, $m = \frac{1}{\sqrt{d}}\mathbb{1}$. (b) Densidad rearreglada $p_X^{\downarrow}$. (c) Equivalente continua de la curva de Lorentz para la densidad p_X (linea llena), la densidad gausiana esférica g_X con covarianza de misma traza que la de p_X (linea con guiones) y la densidad Student-t con covarianza de misma traza que la de p_X y $\nu' = 2{,}25$ grado de libertad (linea punteada): $g_X \prec p_X$ pero $s_X \not\prec p_X$ y $p_X \not\prec s_X$.

1.4 Transformación de variables y vectores aleatorios

En esta sección nos interesamos en los efectos sobre una variable o un vector aleatorio. Por ejemplo, en un juego con dos dados, nos puede interesar la ley de la suma que daría el número de casilla que debemos adelantar en un juego de la oca.

Teorema 1-10 (Transformación medible de un vector aleatorio)**.** *Sea $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ una variable aleatoria, y $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$ una función medible. Entonces, $Y = g(X)$ es una variable aleatoria $(\Omega, \mathcal{A}) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$. Además, la medida imagen P_Y está vinculada a P_X por*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^{d'}), \quad P_Y(B) = P_X(g^{-1}(B)).$$

Demostración. Este resultado es obvio. Siendo g una función medible (recordar Def. 1-6), para todo $B \in \mathcal{B}(\mathbb{R}^{d'})$, por definición $g^{-1}(B) \in \mathcal{B}(\mathbb{R}^d)$. Además, si P_X es la medida (de probabilidad) asociada al espacio de salida de g, el resultado es consecuencia del teorema de la medida imagen 1-2. □

(Ver ej. (Mukhopadhyay, 2000; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Bogachev, 2007b; Cohn, 2013)).

Es sencillo probar que cualquier combinación de funciones medibles queda medible, cualquier producto (adecuado) de funciones medibles queda medible, y que si $\{f_k\}_{k=1}^{d'}$ son $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}))$-medibles, entonces $f = (f_1, \ldots, f_{d'})$ es $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^{d'}))$-medible (Athreya & Lahiri, 2006).

Mencionamos que si $\mathcal{X} = X(\Omega)$ es discreto, entonces $\mathcal{Y} = g(\mathcal{X}) = Y(\Omega)$ será discreto también, y:

Teorema 1-11 (Función de masa por transformación medible)**.** *Sean X, vector aleatorio d-dimensional discreto, $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$ una función medible, e $Y = g(X)$ necesariamente discreto d'-dimensional sobre $\mathcal{Y} = g(\mathcal{X})$. La distribución de Y está vinculada con la de X por la relación*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{x \in g^{-1}(y)} p_X(x).$$

Demostración. El resultado es inmediato. □

En particular, si g es inyectiva (necesariamente biyectiva de $\mathcal{X}$ en $\mathcal{Y}$), el vector de probabilidad queda invariante, $p_Y = p_X$; solamente cambian los estados.

Es importante mencionar que con $\mathcal{Y}$ discreto, $\mathcal{X}$ no es necesariamente discreto (Athreya & Lahiri, 2006). Por ejemplo, $Y = \mathbb{1}_{X>0}$ es tal que $\mathcal{Y} = \{0\,;\,1\}$ a pesar de que $\mathcal{X}$ puede no ser discreto.

Tratar con variables aleatorias continuas resulta más delicado. Vimos en el ejemplo precedente que el carácter continuo puede perderse por transformación. De la misma manera, en un ejemplo de la sección anterior, vimos que $Y = X_1 \mathbb{1}_{X_2>0}$ con X_i independientes uniformes no es continua ni discreta. En el enfoque de variables continuas, una clase importante de funciones en las cuales nos vamos a interesar son las funciones continuas (y diferenciables):

Lema 1-4 (Continuidad y carácter medible)**.** *Sea $g : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ continua. Entonces, g es $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^{d'}))$-medible.*

Demostración. Por continuidad, la pre-imagen de un abierto de $\mathbb{R}^{d'}$ por g es un abierto de $\mathbb{R}^d$ y entonces es en $\mathcal{B}(\mathbb{R}^d)$. La prueba se cierra recordando la definición de $\mathcal{B}(\mathbb{R}^{d'})$, σ-álgebra generada por los abiertos de $\mathbb{R}^{d'}$. □

En lo que sigue, nos interesamos más especialmente en el caso de funciones $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. De hecho, si $d' < d$, es sencillo llegar al caso considerado añadiendo $d - d'$ transformaciones. Por ejemplo, con $d = 2$ si nos interesa $X_1 + X_2$, se puede considerar $\begin{bmatrix} X_1 + X_2 & X_2 - X_1 \end{bmatrix}^t$ y llegar a la variable de interés por cálculo de marginal. Si $d' > d$ la situación es más delicada, $g(Y)$ viviendo sobre una variedad d-dimensional de $\mathbb{R}^{d'}$.

En el caso de vectores aleatorios continuos X que admitend una densidad de probabilidad, una pregunta natural es entonces saber si se conserva la continuidad y la existencia de una densidad, así como su forma. La respuesta se da en el teorema siguiente (Brémaud, 1988; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Cohn, 2013; Hogg et al., 2013):

Teorema 1-12 (Densidad de probabilidad por transformación continua inyectiva diferenciable)**.** *Sea X, vector aleatorio d-dimensional continuo que admite una densidad de probabilidad p_X, y sea $g : \mathbb{R}^d \mapsto$*

$\mathbb{R}^d$ *una función continua inyectiva y diferenciable tal que* $|\mathrm{J}_g| > 0$ *(ver notaciones), Sea* $Y = g(X)$*. Entonces* Y *es continua y admite una densidad de probabilidad* p_Y *de soporte* $\mathcal{Y} = g(\mathcal{X}) = Y(\Omega)$ *tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = p_X(g^{-1}(y)) \left|\mathrm{J}_{g^{-1}}(y)\right|.$$

Demostración. Por definición, admitiendo X una densidad y siendo g medible,

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_Y(B) = P_X(g^{-1}(B)) = \int_{g^{-1}(B)\cap\mathcal{X}} p_X(x)\, dx.$$

Por cambio de variables $x = g^{-1}(y)$ (siendo g inyectiva, el antecedente es único por definición) y notando que $g\left(g^{-1}(B) \cap \mathcal{X}\right) = B \cap \mathcal{Y}$,

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_Y(B) = \int_{B\cap\mathcal{Y}} p_X(g^{-1}(y)) \left|\mathrm{J}_{g^{-1}}(y)\right| dy$$

lo que cierra la prueba [20]. □

El caso escalar puede ser visto como caso particular, dando:

Corolario 1-2. *Sean* X*, variable aleatoria continua que admite una densidad de probabilidad* p_X*,* $g : \mathbb{R} \mapsto \mathbb{R}$ *una función continua, inyectiva y diferenciable, e* $Y = g(X)$*. Entonces* Y *es continua y admite una densidad de probabilidad* p_Y *tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = p_X(g^{-1}(y)) \left|\frac{dg^{-1}(y)}{dy}\right|.$$

De hecho, se pueden ver estos resultados esquemáticamente como una "conservación" de probabilidad, $p_X(x)dx = p_Y(y)dy$, el volumen dy estando relacionado al dx a traves de la matriz Jacobiana (ver nota de pie **??**).

Una forma alternativa de derivar este corolario consiste en salir de la función de repartición, notando que g es necesariamente monótona [21]: si $y \notin \mathcal{Y}$, necesariamente $p_Y = 0$ ($F_Y(y) = 1$ si $y > \sup \mathcal{Y}$ y $F_Y(y) = 0$ si $y < \inf \mathcal{Y}$) y para cualquier $y \in \mathcal{Y}$,

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \begin{cases} P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) & \text{si } g \text{ es creciente} \\ P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) & \text{si } g \text{ es decreciente} \end{cases}.$$

---

[20] La aparición del Jacobiano viene del mismo enfoque que el cambio de variables en la integración de Riemann. De hecho, como lo hemos visto, $\mu_L(B) = |B|$ es el volumen y de la definición misma del determinente, para cualquier matriz cuadrada el volumen se escribe $\mu_L(MB) = |MB| = |M||B| = |M|\mu_L(B)$ donde la misma escritura $|\cdot|$ representa el valor absoluto del determinente de una matriz. Esta notación se justifica precisamente por su significación de volumen, y el resultado es inmediato para $g(x) = Mx$. La forma para una transformación más general se obtiene a partir de un desarollo de Taylor al orden 1 de la transformación, haciendo aparecer el determinente del Jacobiano (Athreya & Lahiri, 2006; Cohn, 2013).

[21] Notar que $P(X \geq x) = 1 - P(X < x) = 1 - P(X \leq x) + P(X = x)$, pero siendo X continua, $P(X = x) = 0$.

El resultado se obtiene calculando las derivadas del primer y último términos respecto de la variable transformada y.

Si g no es inyectiva, g^{-1} es multivaluada o multiforme. En este caso, se puede todavía tratar el problema, particionando $\mathbb{R}^d$ en conjuntos donde g es inyectiva, dando

Teorema 1-13 (Densidad de probabilidad por transformación continua no inyectiva diferenciable)**.** *Sea X, vector aleatorio d-dimensional continuo que admite una densidad de probabilidad p_X, y sea $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ una función continua y diferenciable. Denotamos $\{\mathcal{X}_{[k]}\}_{k=0}^m$ la partición de $\mathcal{X}$ tal que $|\mathrm{J}_g(y)| = 0$ sobre $\mathcal{X}_{[0]}$, y para todo $k \geq 1$ se tiene $g : \mathcal{X}_{[k]} \mapsto \mathcal{Y}$ inyectiva y tal que $|\mathrm{J}_g(y)| > 0$. Suponemos que $\mathcal{X}_{[0]}$ es de medida P_X nula, notamos g_k^{-1} la función inversa de g sobre $g(\mathcal{X}_{[k]})$ (rama k-ésima de la función multivaluada g^{-1}), $\mathrm{J}_{g_k^{-1}}$ su matriz Jacobiana, e $I(y) = \{k \mid y \in g(\mathcal{X}_{[k]})\}$ los índices tales que y tiene un inverso por g_k. Esto es ilustrado en la figura 1-10 para $d = 1$. Entonces Y es continua y admite una densidad de probabilidad p_Y tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{k \in I(y)} p_X(g_k^{-1}(y)) \left|\mathrm{J}_{g_k^{-1}}(y)\right|.$$

En el caso escalar $d = 1$ esto se formula

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{k \in I(y)} p_X(g_k^{-1}(y)) \left|\frac{dg_k^{-1}(y)}{dy}\right|.$$

Esto se ilustra el la figura 1-10.

Demostración. Basta escribir $g^{-1}(B) = \bigcup_{k=0}^m \left(g^{-1}(B) \cap \mathcal{X}_{[k]}\right)$, unión de borelianos disjuntos. Siendo $g^{-1}(B) \cap \mathcal{X}_{[k]} = g_k^{-1}(B)$, y siendo $\mathcal{X}_{[0]}$ de medida nula, se obtiene

$$P_Y(B) = P_X(g^{-1}(B)) = \sum_{k=1}^m \int_{g_k^{-1}(B)} p_X(x)\, dx$$

Se nota ahora que $g_k(g_k^{-1}(B)) = B \cap g(\mathcal{X}_{[k]}) \subset B$ pero no es necesariamente B. Sin embargo, Por cambio de variables $x = g_k^{-1}(y)$ en cada integral, tenemos intonces

$$P_Y(B) = \sum_{k=1}^m \int_{B \cap g(\mathcal{X}_{[k]})} p_X\left(g_k^{-1}(y)\right) \left|\mathrm{J}_{g_k^{-1}}(y)\right| \, dy = \int_B \sum_{k=1}^m \mathbb{1}_{g(\mathcal{X}_{[k]})}(y)\, p_X\left(g_k^{-1}(y)\right) \left|\mathrm{J}_{g_k^{-1}}(y)\right| \, dy$$

Por definición de $I(y)$, $y \in B \cap g(\mathcal{X}_{[k]}) \;\Leftrightarrow\; k \in I(y)$, lo conduce finalmente a

$$P_Y(B) = \int_{B \cap g(\mathcal{X}_{[k]})} \sum_{k \in I(y)} p_X\left(g_k^{-1}(y)\right) \left|\mathrm{J}_{g_k^{-1}}(y)\right| \, dy$$

□

Ejemplo 1-6 (Ejemplo de transformación no biyectiva)**.** *Sea X definida sobre $\mathcal{X} = \mathbb{R}$ y la transformación de variables $Y = X^2$. Se tiene $y = g(x) = x^2$, continua diferenciable de derivada nula sobre $\mathcal{X}_{[0]} = \{0\}$, de medida nula, cuyas inversas son $g_1^{-1}(y) = -\sqrt{y}$ sobre $\mathcal{X}_{[1]} = \mathbb{R}_-^*$ y $g_2^{-1}(y) = +\sqrt{y}$ sobre $\mathcal{X}_{[2]} = \mathbb{R}_+^*$; luego $p_Y(y) = \frac{p_X(\sqrt{y}) + p_X(-\sqrt{y})}{2\sqrt{y}}$, sobre $\mathcal{Y} = \mathbb{R}_+^*$.*

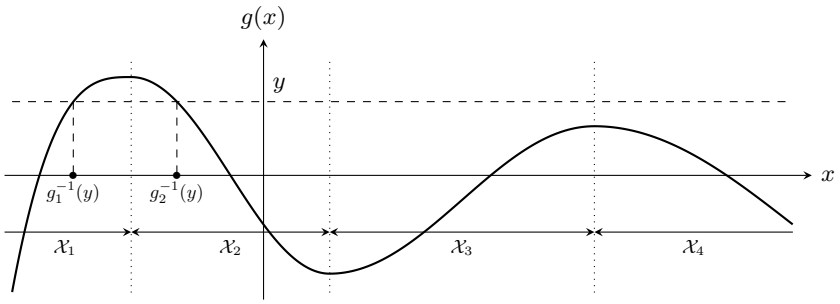


Figura 1-10: Ilustración de una transformación g no inyectiva, tal que $\mathcal{X}_{[0]} = \{x \big| g'(x) = 0\}$, representado por los valores de x en las lineas punteadas. Es de medida de Lebesgue nula. Se indican los dominios $\mathcal{X}_{[k]}$. La línea discontinua da un nivel y y los puntos en el eje x representan $g_k^{-1}(y),\ k \in I(y)$; en el ejemplo, $I(y) = \{1\,;\,2\}$ y, suponiendo que $\mathcal{X} = \mathbb{R}$, $F_Y(y) = F_X(g_1^{-1}(y)) + 1 - F_X(g_2^{-1}(y))$.

De nuevo, en el caso escalar, se puede salir de la función de repartición

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \sum_{k=1}^{m} P\left(X \in \mathcal{X}_{[k]} \cap g_k^{-1}(-\infty\,;\,y]\right)$$

(siendo $\mathcal{X}_{[0]}$ de medida nula, sobre este dominio la probabilidad es cero). Sea $\mathcal{Y}_{[k]} = g_k(\mathcal{X}_{[k]})$. Ahora, si $y \notin I(y)$,

$$P\left(X \in \mathcal{X}_{[k]} \cap g_k^{-1}(-\infty\,;\,y]\right) = \begin{cases} P(X \in \mathcal{X}_{[k]}) & \text{si} \quad y > \sup \mathcal{Y}_{[k]} \\ 0 & \text{si} \quad y < \text{ínf}\, \mathcal{Y}_{[k]} \end{cases}$$

dando una derivada nula. Si $y \in I(y)$,

$$P\left(X \in \mathcal{X}_{[k]} \cap g_k^{-1}(-\infty\,;\,y]\right) = \begin{cases} F_X(g_k^{-1}(y)) - F_X(\text{ínf}\, \mathcal{Y}_{[k]}) & \text{si} \quad g_k \quad \text{es creciente} \\ F_X(\sup \mathcal{Y}_{[k]}) - F_X(g_k^{-1}(y)) & \text{si} \quad g_k \quad \text{es decreciente} \end{cases}.$$

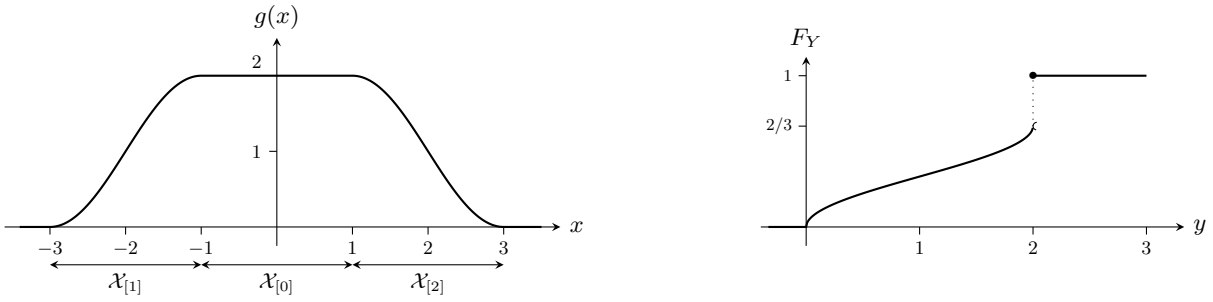
El resultado sigue diferenciando estas expresiones. Se ilustra esto también en la figura 1-10.

Una tercera alternativa, a pesar de que sea delicado, es apoyarse en la teoría de distribuciones y expresar como $p_Y(y) = \int_{\mathcal{X}} p_X(x)\, \delta(y - g(x))\, dx$, donde se usa la expansión de la función delta en términos de sus ceros: $\delta(y - g(x)\,) = \sum_{k \in I(y)} \frac{1}{|g_k'(g_k^{-1}(y))|} \delta(x - g_k^{-1}(y))$ (Mandel & Wolf, 1995).

Es importante notar que la condición $\mathcal{X}_{[0]}$ de medida nula es importante. Al contrario, Y no resulta continua como se puede ver en el ejemplo siguiente.

Ejemplo 1-7 (Transformación con $P_X\left(\mathcal{X}_{[0]}\right) \neq 0$)**.** *Sea X uniforme sobre $\mathcal{X} = (3\,;\,3)$ e $Y = g(X)$ con $g(x) = \left(1 + \cos\left((|x| - 1)\frac{\pi}{2}\right)\right) \mathbb{1}_{(1\,;\,3)}(|x|) + 2\mathbb{1}_{[0\,;\,1]}(|x|)$. Esta función se representa en la figura 1-11-(a). Claramente, g es continua y diferenciable sobre $\mathcal{X}$, pero con $\mathcal{X}_{[0]} = [-1\,;\,1]$ que no es de medida nula. Saliendo de $F_Y(y) = P(g(X) \leq y)$ se calcula sencillamente $F_Y(y) = \frac{2}{3}\left(1 - \frac{1}{\pi}\arccos(y-1)\right)\mathbb{1}_{[0\,;\,2)} + \mathbb{1}_{[2\,;\,+\infty)}(y)$, ilustrada en la figura 1-11-(b). Claramente F_Y es discontinua en $y = 2$: Y no es continua.*

Un ejemplo de cambio de transformación puede sevir a calcular la densidad de probabilidad de una suma:



(a) (b)

Figura 1-11: (a): Gráfica de $g(x) = \left(1 + \cos\left((|x| - 1)\frac{\pi}{2}\right)\right)\mathbb{1}_{(1\,;\,3)}(|x|) + 2\mathbb{1}_{[0\,;\,1]}(|x|)$. Suponiendo que $\mathcal{X} = (-3\,;\,3)$, claramente $\mathcal{X}_{[0]} = [-1\,;\,1]$ no es de medida nula. (b): Para X uniforme sobre $\mathcal{X}$, la variable $Y = g(X)$ resulta con función de repartición F_Y no continua.

Ejemplo 1-8 (Distribución de la suma de vectores aleatorios)**.** *Sean X e Y dos vectores aleatorios d-dimensionales conjuntamente continuos, de densidad de probabilidad conjunta $p_{X,Y}$, y sea el vector*

$$V = X + Y.$$

Queremos calcular la densidad de probabilidad de V. Para esto, se puede considerar la transformación biyectiva

$$g : (x, y) \mapsto (u, v) = (x, x + y).$$

Entonces

$$g^{-1}(u, v) = (u, v - u)$$

y la matriz Jacobiana es

$$J_{g^{-1}} = \begin{bmatrix} I & -I \\ 0 & I \end{bmatrix}$$

donde la identidad I y la matriz nula 0 son en $\mathcal{M}_{d,d}(\mathbb{R})$ conjuntos de matrices $d \times d$ (ver notaciones). Claramente $\left|J_{g^{-1}}\right| = 1$ así que

$$p_{U,V}(u, v) = p_{X,Y}(u, v - u)$$

como lo podiamos intuir. Además, por marginalización, inmediatamente

$$p_V(v) = \int_{\mathbb{R}^d} p_{X,Y}(u, v - u)\, du.$$

Si X e Y son independientes, $p_{U,V}(u, v) = p_X(u)p_Y(v - u)$ y la fórmula integral se escribe

$$p_V(v) = \int_{\mathbb{R}^d} p_X(u)p_Y(v - u)\, du = \int_{\mathbb{R}^d} p_Y(u)p_X(v - u)\, du$$

(por cambio de variable en la secunda expresión). Esta fórmula es conocida como producto de convolución entre las funciones[22] *p_X y p_Y y como lo podemos ver, es conmutativo.*

[22]Un producto de convolution entre funciones se define entre cualquieras funciones, que sean densidades de probabilidad o

1.5 Leyes condicionales

Al considerar un par de vectores aleatorios X e Y, una pregunta natural puede ser cómo caracterizar el vector Y si "observamos $X = x$". En otras palabras, la pregunta es describir la ley de Y "sabiendo (o observando) que $X = x$". En lo que sigue, para fijar notación, consideramos $(X, Y) : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}, \mathcal{B}(\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}))$ tal que X es d_X-dimensional e Y es d_Y-dimensional (incluyendo los casos escalares). Escrimiremos de nuevo $\mathcal{X} = X(\Omega)$ e $\mathcal{Y} = Y(\Omega)$.

Caso X discreta: Un caso sencillo a estudiar es cuando $\mathcal{X}$ es discreto. En este caso, para cualquier $x \in \mathcal{X}$, tenemos $P_X(x) = P(X = x) \neq 0$ y de la definición de la probabilidad condicional Def. 1-3, $\forall\, B \in \mathcal{B}(\mathbb{R}^{d_Y}), \quad P(Y \in B | X = x) = \dfrac{P\big((Y \in B) \cap (X = x)\big)}{P(X = x)}$ define una medida de probabilidad que llamamos medida de probabilidad condicional. Siendo una medida de probabilidad, nos referimos a la subsección anterior para definir una función de repartición tomando $B = \prod_{i=1}^{d} (-\infty\,;\, y_i]$, caracterizando completamente la medida de probabilidad:

Definición 1-32 (Medida de probabilidad y función de repartición condicional (X discreto))**.** *Por cualquier $x \in \mathcal{X}$, la medida condicional de Y, condicionalmente a $(X = x)$, se define por*

$$\forall\, B \in \mathcal{B}\left(\mathbb{R}^{d_X}\right), \qquad P_{Y|X=x}(B) = P(Y \in B | X = x),$$

y la función de repartición condicional se define por,

$$\forall\, x \in \mathcal{X},\ y \in \mathcal{Y}, \qquad F_{Y|X=x}(y) = P(Y \leq y |\, X = x) = \frac{P\left((Y \leq y) \cap (X = x)\right)}{P\left(X = x\right)}.$$

Ahora, cuando Y también es discreta, se puede definir la función de masa discreta de probabilidad condicional, y si Y es continua y admite una densidad de probabilidad, se puede definir una densidad de probabilidad condicional:

Definición 1-33 (Función de masa o densidad de probabilidad condicional (X discreta))**.** *Por definición, cuando $\mathcal{Y}$ es discreta, la función de masa de probabilidad condicional de Y condicionalmente a $X = x$ es,*

$$\forall\, x \in \mathcal{X},\ y \in \mathcal{Y}, \quad p_{Y|X=x}(y) = P\left(Y = y |\, X = x\right) = \frac{P\big((Y = y) \cap (X = x)\big)}{P(X = x)}.$$

Si Y es continua, es sencillo ver que $P_{Y|X=x} \ll P_Y$, i.e., para $B \in \mathcal{B}(\mathbb{R}^{d_Y})$, $P_Y(B) = 0 \Rightarrow P_{Y|X=x}(B) = 0$. Si Y admite una densidad con respecto a la medida de Lebesgue, $P_Y \ll \mu_L$ medida de Lebesgue, es

---

no. Una condición suficiente para que existe tal producto es que las funciones que se convoluan sean L^1 (Golberg, 1961; Stein & Weiss, 1971; Pinsky, 2009) (es el caso de densidad de probabilidad).

claro que también $P_{Y|X=x} \ll \mu_L$, y por teorema de Radon-Nikodým 1-8, $P_{Y|X=x}$ admite una densidad de probabilidad (con respecto a la medida de Lebesgue) que denotaremos $p_{Y|X=x}$,

$$\forall\, B \in \mathcal{B}(\mathbb{R}^{d_Y}), \quad P_{Y|X=x}(B) = \int_B p_{Y|X=x}(y)\, dy.$$

A partir de la función de repartición, obtenemos

$$p_{Y|X=x}(y) = \frac{\partial^{d_Y} F_{Y|X=x}(y)}{\partial y_1 \dots \partial y_{d_Y}}.$$

Caso general: Cuando X es continua, el problema es más sutil porque $P(X = x) = 0$. Entonces, no se puede usar la definición de la probabilidad condicional, siendo el evento $(X = x)$ de probabilidad nula. Sin embargo, se pueden seguir los pasos de Rényi (Rényi, 2007, Cap. 5), de Feller (Feller, 1971, Cap. 10) o Ash & Doléans-Dade (Ash & Doléans-Dade, 1999, Sec. 5.3) por ejemplo para resolver el problema, llegando en el contexto continuo a un resultado intuitivo como en el caso discreto.

Sea $B \in \mathcal{B}(\mathbb{R}^{d_Y})$ tal que $P(Y \in B) \neq 0$ y definimos $\nu_B(A) = P\big((X \in A) \cap (Y \in B)\big)$ sobre $\big(\mathbb{R}^{d_X}, \mathcal{B}(\mathbb{R}^{d_X})\big)$. Es sencillo ver que siendo dado B, ν_B define una medida. Además, $\nu_B \ll P_X$, i.e., para $A \in \mathcal{B}(\mathbb{R}^{d_X})$, $P_X(A) = P(X \in A) = 0 \;\Rightarrow\; 0 = P\Big((X \in A) \cap (Y \in B)\Big) = \nu_B(A)$. Por teorema de Radon-Nikodým 1-8, ν_B admite una densidad $g_B = \frac{d\nu_B}{dP_X}$ con respecto a P_X,

$$\forall\, A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right), \quad P\big((X \in A) \cap (Y \in B)\big) = \int_A g_B(x)\, dP_X(x).$$

Claramente $g_B \geq 0$, y de $P(X \in A) = P\big((X \in A) \cap (Y \in B)\big) + P\big((X \in A) \cap (Y \in \overline{B})\big)$ para cualquier $A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right)$, se escribe $\forall\, A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right), \quad 0 \leq P\big((X \in A) \cap (Y \in \overline{B})\big) = \int_A dP_X(x) - \int_A g_B(x)\, dP_X(x)$, lo que permite concluire que $0 \leq g_B \leq 1$. Con el mismo razonamiento, se puede ver que $g_B(\mathbb{R}^{d_Y}) = 1$ y que g_B es σ-aditiva. En realidad, $g_B \leq 1$ P_X-casi siempre, pero olvidando esta subtilesa, g_B define una medida de probabilidad, que llamaremos *medida de probabilidad condicional*. Por continuación se define una función de repartición condicional de la misma manera que se definó la función de repartición. En resumen:

Definición 1-34 (Medida de probabilidad y función de repartición condicional)**.** *La medida de probabilidad condicional de $P_{Y|X=x}$ es definida tal que*

$$\forall\, (A, B) \in \mathcal{B}\left(\mathbb{R}^{d_X}\right) \times \mathcal{B}\left(\mathbb{R}^{d_Y}\right), \quad P\big((X \in A) \cap (Y \in B)\big) = \int_A P_{Y|X=x}(B)\, dP_X(x).$$

Tomando $B = \bigtimes_i (-\infty\,;\, y_i]$ se obtiene la función de repartición condicional a partir de

$$\forall\, A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right),\, y \in \mathcal{Y}, \quad P\big((X \in A) \cap (Y \leq y)\big) = \int_A F_{Y|X=x}(y)\, dP_X(x).$$

Adémas, si X admite una densidad de probabilidad p_X, $dP_X = p_X dx$ y tomando $A = \bigtimes_i (-\infty\,;\, x_i]$ se obtiene

$$F_{X,Y}(x, y) = \int_{\bigtimes_i (-\infty\,;\, x_i]} F_{Y|X=x}(y)\, p_X(x)\, dx$$

o, por diferenciación, para cualquier $y \in \mathcal{Y}$ *y* $x \in \mathcal{X}$ *(i. e., tal que* $p_X(x) \neq 0$*),*

$$F_{Y|X=x}(y) = \frac{\dfrac{\partial^{d_X} F_{X,Y}(x,y)}{\partial x_1 \dots \partial x_{d_X}}}{p_X(x)}.$$

Claramente, $\left(\mathbb{R}^{d_Y}, \mathcal{B}(\mathbb{R}^{d_Y}), P_{Y|X=x}\right)$ define un espacio de probabilidad y, a veces, por abuso de escritura, denotaremos $Y|X=x$ la variable aleatoria $(\Omega, \mathcal{A}, P) \to \left(\mathbb{R}^{d_Y}, \mathcal{B}(\mathbb{R}^{d_Y}), P_{Y|X=x}\right)$.

En el contexto de variable aleatorias independientes, intuitivamente conocer a X no va a llevar "información" sobre Y, lo que se formaliza de la manera siguiente:

Lema 1-5 (Probabilidad condicional e independencia)**.** *Sean* X *e* Y *vectores aleatorios independientes, entonces*

$$\forall\, x \in \mathcal{X},\ B \in \mathcal{B}(\mathbb{R}^{d_Y}), \quad P_{Y|X=x}(B) = P_Y(B)$$

A continuación, obviamente,

$$F_{Y|X=x}(y) = F_Y(y)$$

Demostración. Inmediatamente, de la independencia, tenemos

$$\begin{aligned} P\big((X \in A) \cap (Y \in B)\big) &= P(X \in A)\, P(Y \in B) \\ &= P_Y(B) \int_A dP_X(x) \\ &= \int_A P_Y(B)\, dP_X(x) \end{aligned}$$

lo que cierra la prueba. □

Ahora, tomando $A = \mathcal{X}$ en la primera fórmula que define la medida de probabilidad condicional se recupera el equivalente continuo de la fórmula de probabilidad total:

Teorema 1-14 (Fórmula de probabilidad total (caso general))**.** *Sean* X *e* Y *vectores aleatorias y* $\mathcal{X} = X(\Omega), \quad \mathcal{Y} = Y(\Omega)$*. Entonces*

$$\forall\, B \in \mathcal{B}\left(\mathbb{R}^{d_Y}\right), \qquad P(Y \in B) = \int_{\mathcal{X}} P_{Y|X=x}(B)\, dP_X(x)$$

lo que da en termino de función de repartición condicional

$$F_Y(y) = \int_{\mathcal{X}} F_{Y|X=x}(y)\, dP_X(x)$$

Si P_X *admite una densidad, se escribe todo notando que* $dP_X(x) = p_X(x) d\mu_L(x) \equiv p_X(x) dx$*.*

Por último, si (X, Y) admite una densidad, en sencillo ver que $P_{Y|X=x} \ll \mu_L$, y entonces $P_{Y|X=x}$ admite una densidad que llamaremos *densidad de probabilidad condicional*. Sean $A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right)$ y $B \in \mathcal{B}\left(\mathbb{R}^{d_Y}\right)$,

$$\begin{aligned} P\Big((X \in A) \cap (Y \in B)\Big) &= \int_{A \times B} p_{X,Y}(x,y)\, dx\, dy \\ &= \int_B \left(\int_A \frac{p_{X,Y}(x,y)}{p_X(x)}\, dy \right) p_X(x)\, dx \end{aligned}$$

Entonces, si $p_X(x) \neq 0$, tenemos

$$P_{Y|X=x}(A) = \int_A \frac{p_{X,Y}(x,y)}{p_X(x)}\, dy.$$

Teorema 1-15 (Densidad de probabilidad condicional)**.** *Si (X,Y) admite una densidad de probabilidad, la medida de probabilidad condicional $P_{Y|X=x}$ admite una densidad, llamada densidad de probabilidad condicional definida por*

$$\forall\, x \in \mathcal{X}, \quad p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

definida sobre $\mathcal{Y}$. Claramente, a partir de la función de repartición condicional resulta que

$$p_{Y|X=x} = \frac{\partial^{d_Y} F_{Y|X=x}}{\partial y_1 \dots \partial y_{d_Y}}.$$

De hecho, esta construcción rigurosa coincide con la intuición que podemos tener en este caso continuo. Por ejemplo, podemos pensar a $F_{Y|X=x}(y)$ como caso límite de $P\left(Y \leq y \,\middle|\, x \leq X \leq x+\delta x\right) = \frac{P\big((Y \leq y) \cap (x \leq X \leq x+\delta x)\big)}{P\left(x \leq X \leq x+\delta x\right)} = \frac{F_{X,Y}(x+\delta x, y) - F_{X,Y}(x,y)}{F_X(x+\delta x) - F_X(x)}$ cuando δx tiende a 0. En el caso escalar, se calcula por ejemplo haciendo un desarrollo de Taylor del numerador y del denominador a orden 1, o usando la regla de l'Hôpital [23] para re-obtener la función de repartición condicional de la definición **??**. En el caso multivariado, hace falta hacer los desarollos hasta el orden d_X para concluir.

Notar que:

- si X e Y son independientes,
$$p_{Y|X=x} = p_Y;$$
- por la expresión $p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$, por integración con respecto a y obtenemos la condición de normalización
$$\int_{\mathcal{Y}} p_{Y|X=x}(y)\, dy = 1.$$

También, se escribe la fórmula de probabilidad total a través de las densidades por la expresión $p_{X,Y}(x,y) = p_{Y|X=x}(y)\, p_X(x)$ y luego por integración con respecto a x:

Teorema 1-16 (Fórmula de probabilidad total (caso con densidades))**.** *Si (X,Y) admite una densidad de probabilidad conjunta $p_{X,Y}$, entonces Y tiene una densidad de probabilida que se recupera a través de la fórmula*

$$p_Y(y) = \int_{\mathcal{X}} p_{Y|X=x}(y)\, p_X(x)\, dx.$$

De la expresión la densidad condicional, $p_{X,Y}(x,y) = p_{Y|X=x}(y)\, p_X(x) = p_{X|Y=y}(x)\, p_Y(y)$, y de la fórmula de probabilidad total se recupera sencillamente el equivalente continuo de la formula de Bayés:

---

[23] De hecho, esta regla es debido al suizo J. Bernoulli que tuvo un acuerdo financiero con el Guillaume François Antoine, marqués de l'Hôpital, permitiendolo de publicar unos resultados de Bernoulli bajo su nombre.

Teorema 1-17 (Fórmula de Bayes (caso continuo))**.**

$$\forall\, y \in \mathcal{Y},\ x \in \mathcal{X}, \qquad p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)\, p_Y(y)}{\displaystyle\int_{\mathcal{Y}} p_{X|Y=y}(x)\, p_Y(y)\, dy}.$$

Volvemos ahora al ejemplo 1-8:

Ejemplo 1-9 (Distribución condicional de la suma de vectores aleatorios)**.** *Sea $V = X + Y$, con X e Y vectores d-dimensionales. Introduciendo $U = X$ obtuvimos $p_{U,V}(u, v) = p_{X,Y}(u, v - u)$ dando también $p_V(v) = \displaystyle\int_{\mathbb{R}^d} p_{X,Y}(u, v - u)\, du$. Entonces, recordando que $U = X$, se obtiene*

$$p_{V|X=x}(v) = \frac{p_{X,Y}(x, v - x)}{p_X(x)} = \frac{p_{X,Y}(x, v - x)}{\displaystyle\int_{\mathbb{R}^d} p_{X,Y}(x, v - x)\, dv},$$

dando en el caso X e Y independientes

$$p_{V|X=x}(v) = p_Y(v - x).$$

Esto corresponde a la intución de que, con $V = X + Y$, fijando $X = x$ el vector aleatorio V es nada más que Y desplazado en x. Pero hay que tomar muchas precauciones con este razonamiento, valido únicamente cuando X e Y son independientes. En caso contrario, fijando X no coincide con un desplazamiento por la dependencia (esquematícamente, fijando X no sólo mueve Y sino que "cambia" su estadística).

1.6 Esperanza, momentos, identidades y desigualdades

Como lo hemos introducido, la noción formal de probabilidad nació en el contexto de juego (cartas, dados), bajo entre otros el impulso del matemático italiano y jugador de dados y cartas Gerolamo Cardano en el siglo XVI, y aún más bajo el trabajo muy profundo de la dinastía Bernoulli, y en particular Jacob Bernoulli (Bernoulli, 1713, en latín) o ((E. D. Sylla, Translator), 1713). En particular, Bernoulli se interesó no solamente al resultado de un sorteo, impredictible, pero en lo que pasa cuando se hace un gran numéro de sorteos. Así salió la idea de "resultado promedio". De hecho, la noción de promedio es probablemente debido a B. Pascal: se encuentran, por lo menos semillas de esta noción en una carta que mandó a Fermat en 1654 (David & Edwards, 2001), o un poco más tarde debido a C. Huygens (Huygens, 1657; David & Edwards, 2001; Hald, 1990) (?, ?, ?, ?, ?, ?). Bernoulli hizo el paso decisivo, y en su trabajo, ha ido más allá: probó la ley de gran número que vamos a ver más adelante. Sin ir tan lejos, en el caso de una variable X aleatoria discreta (ej. un dado, que puede tomar valores en $\{1\,;\,\ldots\,;\,6\}$), haciendo varios sorteos, el promedio de estes sorteos va a ser $\sum_i \widetilde{p}_i x_i$ con x_i los valores que puede tomar la variable y $\widetilde{p}_i$ la proporción de x_i que se obtuvo en el sorteo. De hecho, intuitivamente (es la visión frecuencista), $\widetilde{p}_i$ va a tender a $p_i = P(X = x_i)$ cuando el número de

sorteo tiende al infinito. La definición del valor promedio, o media, de una variable aleatoria cualquiera se formaliza mas rigorosamente en en marco de la teoria de la medida, pero coincide con la intuición.

1.6.1 Media de un vector aleatorio

Una variable aleatoria X tiene asociado un *promedio* o *media* (también llamado *valor esperado o de expectación o esperanza matemática*) que se obtiene pesando cada valor de X con la medida de probabilidad asociada a ese valor (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006),

Definición 1-35 (Media o valor/vector medio)**.** *Formalmente, la media de una variable aleatoria X integrable es definida por*

$$\mathrm{E}[X] = \int_{\Omega} X(\omega)\, dP(\omega).$$

Por el teorema de la medida imagen 1-2, esta media se escribe también a partir de la medida de probabilidad P_X como

$$\mathrm{E}[X] = \int_{\mathbb{R}} x\, dP_X(x).$$

En el caso vectorial d-dimensional, hay que entender la media, o vector medio, como un vector de componentes i-ésima la media $\mathrm{E}[X_i]$ de la componente i-ésima X_i de X, dando

$$\mathrm{E}[X] = \int_{\mathbb{R}^d} x\, dP_X(x).$$

A veces, se encuentra también la notación $\langle x \rangle$ o $\langle x \rangle_{p_X}$ para el valor medio, especialmente en la literatura de física.

La secunda formulación del valor medio se proba sencillamente, empezando por $X = \mathbb{1}_A$ para unos $A \in \mathcal{B}(\mathbb{R})$. Entonces $P_X = (1 - P(A))\delta_0 + P(A)\delta_1$. Luego $\int_{\Omega} \mathbb{1}_A(\omega) dP(\omega) = P(A) = (1 - P(A)) \times 0 + P(A) \times 1 = \int_{\mathbb{R}} x dP_X(x)$. Se cierra la prueba con el teorema 1-3 dando cualquier función medible como límite de funciones escalonadas, y por la definición 1-10 de la integral de cualquier función medible.

Luego, de la distribución marginal $P_{X_i}(B) = \int_{\mathbb{R}^{i-1} \times B \times \mathbb{R}^{d-i}} dP_X(x)$, se obtiene $\mathrm{E}[X_i] = \int_{\mathbb{R}^d} x_i\, dP_X(x)$, dando la última formulación en el caso vectorial.

Una variable aleatoria X se dice integrable cuando $E[|X|] < \infty$. De la misma manera, un vector aleatorio admite una media si y solamente si cada componente es integrable. Veremos más adelante que existen variables aleatorias que no admiten una media.

Más allá de la formulación matemática, $\mathrm{E}[X]$ representa la posición alrededor de la cual se "distribuye las probabilidades de occurencia". Es el equivalente probabilístico de centro de gravedad o barycentro en mecánica.

En el caso de variables aleatorias discretas, de soporte $\mathcal{X}$ discreto finito o numerable, inmediatamente

$$\mathrm{E}[X] = \sum_{x \in \mathcal{X}} x\, P(X = x) = \sum_{x \in \mathcal{X}} x\, p_X(x).$$

Fijense de que $\mathrm{E}[X]$ no partenece necesariamente a $\mathcal{X}$:

Ejemplo 1-10. *Sea* X *uniforme sobre* $\mathcal{X} = \{1\,;\,3\,;\,7\}$*, i. e.,* $\forall\, i \in \mathcal{X},\quad P(X = i) = \frac{1}{3}$*. Se calcula* $\mathrm{E}[X] = 1 \times \frac{1}{3} + 3 \times \frac{1}{3} + 7 \times \frac{1}{3} = \frac{11}{3} \notin \mathcal{X}$*. Tampoco es el promedio de los valores extremos.*

Cuando $|\mathcal{X}| = +\infty,\ X$ no es necesariamente integrable:

Ejemplo 1-11. *Sea* $\mathcal{X} = \mathbb{N}^*$ *con* $P(X = x) = \frac{6}{\pi^2\, x^2}$*. Claramente,* $\sum_x \frac{6}{\pi^2\, x}$ *diverge, así que* X *no tiene una media.*

En el caso de vectores aleatorios continuos, obtenemos la expresión siguiente de la media (o vector medio):

$$\mathrm{E}[X] = \int_{\mathbb{R}^d} x\, p_X(x)\, dx.$$

Las mismas observaciones que hicimos en el caso discreto se encuentra en el caso continuo:

Ejemplo 1-12. *Sea* X *de densidad de probabilidad* $p_X(x) = \frac{1}{2}\mathbb{1}_{[0\,;\,1)}(x) + \frac{3\sqrt{x-2}}{4}\mathbb{1}_{[2\,;\,3)}(x)$ *como ilustrado figura Fig. 1-6. Se calcula* $\mathrm{E}[X] = \frac{31}{20} \notin \mathcal{X} = [0\,;\,1] \cup [2\,;\,3]$.

Ejemplo 1-13. *Un ejemplo de vector aleatorio no teniendo media es dado en el caso de una distribución de Cauchy-Lorentz (ver más adelante)* $p_X(x) = \dfrac{\alpha}{(1 + x^t x)^{\frac{d+1}{2}}}$ *donde* α *es un factor de normalización.*

En el caso general, para calcular la media, hay que pasar por la distribución P_X, como en el ejemplo 1-4:

Ejemplo 1-14 (Continuación del ejemplo 1-4)**.** *Sea* $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ *con* U *y* V *variables aleatorias independientes de distribución uniformas sobre* $[0\,;\,1)$*, i. e.,* $p_U(x) = \mathbb{1}_{[0\,;\,1)}(x)$*. De* $(X \in B) \Leftrightarrow \big((U < \frac{1}{2}) \cap (V \in B)\big) \cup \big((U \geq \frac{1}{2}) \cap (1 \in B)\big)$*, del hecho de que los eventos de la unión son incompatibles y de la independencia de* U *y* V *(o saliendo de la función de repartición), se obtiene* $P_X(B) = \frac{1}{2}P_V(B) + \frac{1}{2}\delta_1(B)$*. A continuación,* $\mathrm{E}[X] = \dfrac{1}{2}\displaystyle\int_{\mathbb{R}} dP_V(x) + \frac{1}{2}\int_{\mathbb{R}} d\delta_1(x) = \frac{1}{2}\int_{\mathbb{R}} p_V(x)\, dx + \frac{1}{2}\times 1 = \frac{1}{2}\int_0^1 dx + \frac{1}{2} = \frac{3}{4}$.

Una nota intersante es de que, en el caso escalar, para $X \geq 0$ admitiendo una media, se obtiene

$$\mathrm{E}[X] = \int_{\mathbb{R}_+} P(X > t)\, dt = \int_{\mathbb{R}_+} (1 - F_X(t))\ dt.$$

Se proba saliendo de $x = \displaystyle\int_0^x dt = \int_{\mathbb{R}_+} \mathbb{1}_{(t\,;\,+\infty)}(x)\, dt$ dando $\mathrm{E}[X] = \displaystyle\int_{\mathbb{R}} \left(\int_{\mathbb{R}_+} \mathbb{1}_{(t\,;\,+\infty)}(x)\, dt\right) dP_X(x) = \int_{\mathbb{R}_+} \left(\int_{\mathbb{R}} \mathbb{1}_{(t\,;\,+\infty)}(x)\, dP_X(x)\right) dt$ por el teorema de Fubini Th. 1-6. Se cierra la prueba observando que la integral interior es nada más que $P(X > t)$. En el caso discreto con $\mathcal{X} = \mathbb{N}$, viene inmediatamente $\displaystyle\sum_{t\in\mathbb{N}} P(X > t)$ que podemos probar directamente saliendo

de $P(X = t) = P(X > t) - P(X > t - 1)$. En el caso de variable admitiendo una densidad, se lo obtiene también haciendo una integración por partes [24]

Esta fórmula se aplica al ejemplo 1-4 que tratamos:

Ejemplo 1-15 (Continuación del ejemplo 1-4)**.** *Sea* $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ *con* U *y* V *variables aleatorias independientes de distribución uniformas sobre* $[0\,;\,1)$*. Tratando de este ejemplo, obtuvimos* $F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;\,+\infty)}(x)$*. A continuación, reobtenemos* $\mathrm{E}[X] = \displaystyle\int_0^1 \left(1 - \frac{x}{2}\right) dx = \frac{3}{4}$.

.

Terminamos esta sección con la propiedad de linealidad de la esperanza matemática E, como consecuencia de la linealidad de la integración y definición de la distribución marginal: para cualquier conjunto de vectores aleatorios $\{X_i\}$ integrables, cualquieras matrices $\{C_i\}$ dadas (deterministicas) de dimensiones compatibles con las de X (incluyendo el caso escalar), y b vecto dado (cierto),

$$\mathrm{E}\left[\sum_i C_i X_i + b\right] = \sum_i C_i\,\mathrm{E}\left[X_i\right] + b$$

(la integrabilidad de la suma se proba a partir de la desigualdad triangular).

1.6.2 Momentos de un vector aleatorio

Si X es una variable aleatoria, para cualquier función medible f, $f(X)$ también lo es. Se puede entonces definir su valor medio, si existe. A pesar de necesitar evaluar la distribución de probabilidad de $Y = f(X)$, el valor medio se calcula a partir del de X:

Teorema 1-18 (Teorema de transferencia)**.** *Sea* X *un vector aleatorio* d*-dimensional y* $f : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ *una función medible tal que* $f(X)$ *sea integrable. Entonces*

$$\mathrm{E}\left[f(X)\right] = \int_\Omega f(X(\omega))\,dP(\omega) = \int_{\mathbb{R}^{d'}} f(x)\,dP_X(x).$$

En particular, en el caso $\mathcal{X} = X(\Omega)$ *discreto se obtiene*

$$\mathrm{E}\left[f(X)\right] = \sum_{x\in\mathcal{X}} f(x)\,P(X = x)$$

y para X *continuo admitiendo una densidad de probabilidad*

$$\mathrm{E}\left[f(X)\right] = \int_{\mathbb{R}^d} f(x)\,p_X(x)\,dx.$$

---

[24] En el caso discreto, hay que tener precauciones separando la series de una diferencia de terminos. En el caso X continuo admitiendo una densidad, hay que estudiar bien el coomportamiento de $t \mapsto (1 - F_X(t))$ al infinito.

Demostración. Sea $B \in \mathcal{B}(\mathbb{R}^d)$ y consideramos $f(x) = \mathbb{1}_B(x)$. Entonces, $\mathcal{Y} = \{0\,;\,1\}$ y inmediatamente

$$P_Y = P_X(B)\,\delta_1 + (1 - P_X(B))\,\delta_0.$$

Entonces

$$\mathrm{E}[f(X)] = \int_{\mathbb{R}} P_X(B)\,d\delta_1 + \int_{\mathbb{R}} (1 - P_X(B))\,d\delta_0 = P_X(B) = \int_{\mathbb{R}^d} \mathbb{1}_B(x)\,dP_X(x).$$

En el caso $d' = 1$, para $f \geq 0$, se cierra entonces la prueba usando el teorema 1-3, escribiendo f como límite creciente de una sucesión de funciones escalonadas, y la definición 1-10 de la integración real. El caso $d' > 1$ es nada mas que $d' = 1$, componente a componente. □

De manera general, estas medias son llamadas *momentos* de la variable aleatoria X. Los momentos relevantes usuales son los siguientes:

- para el "monomio" $f(x) = x^{\otimes k}$ producto externo [25] de x K veces siendo $k \in \mathbb{N}^*$, se obtiene el tensor conteniendo todo los k-*ésimo momentos (ordinarios)* de X:

  $$m_k[X] \equiv \mathrm{E}\left[X^{\otimes k}\right] = \int_{\mathbb{R}^d} x^{\otimes k}\,dP_X(x)$$

  que tiene unidades de $\prod_{j=1}^{k} X_{i_j}$ (de X_i^k si los componentes de X tienen la misma "unidad"). Se escriben también los momentos de orden k bajo la escritura (con el mismo m, por abuso)

  $$m_{i_1,\ldots,i_k}[X] = \mathrm{E}\left[\prod_{j=1}^{k} X_{i_j}\right] \quad \text{con} \quad (i_1,\ldots,i_k) \in \{1,\ldots,d\}^k$$

  que son los componentes de $m_k[X]$. Se puede incluir el caso $k = 0$ con la convención $x^{\otimes 0} = 1$, que corresponde a la condición de normalización: $m_0[X] = \int_{\mathbb{R}} dP_X(x) = 1$. La media es el primer momento: $m_1[X] = \mathrm{E}[X] = m_X$. Típicamente, los primeros momentos son más relevantes que los de órdenes mayores, para la caracterización de una distribución. Para $k = 2$, en el caso escalar, el momento de orden 2 es el análogo del momento de inercia de la mecánica.
  Por ejemplo, para la distribución uniforme $p_X(x) = \frac{1}{b-a}$ en el intervalo $[a\,;\,b]$, resulta $m_k[X] = \frac{b^{k+1}-a^{k+1}}{(k+1)(b-a)}$. En particular, $m_1[X] = \frac{a+b}{2}$, valor medio del intervalo.
  Fijense de que $X^{\otimes k}$ no es siempre integrable, por ejemplo, en el caso con densidad, si $p_X(x)$ tiene soporte (semi)infinito, necesariamente la función p_X debe tender a 0 cuando $\|x\| \to \infty$. Si $p_X(x)$ es *de largo alcance*, en el sentido de que no cae a 0 suficientemente rápido con x para x grandes, algunos momentos pueden no existir. Por ejemplo, la distribución de probabilidad de Cauchy–Lorentz (o función de Breit–Wigner), dada por $p_X(x) =$

[25] Recuerdense de que $x \otimes x$ es una matriz teniendo como componentes $x_i x_j$; entonces $x^{\otimes k}$ en un tensor k-dimensional teniendo como componentes $\left[x^{\otimes k}\right]_{i_1,\ldots,i_k} = \prod_{j=1}^{k} x_{i_j}$.

$\frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{\frac{d+1}{2}}\,|R|^{\frac{1}{2}}}\left(1+(x-x_0)^t R^{-1}(x-x_0)\right)^{-\frac{d+1}{2}}$ sobre $\mathbb{R}^d$, con $R \in P_d^+(\mathbb{R})$, $x_0 \in \mathbb{R}^d$, no tiene momentos finitos de orden $k \geq 1$.

- Frecuentemente (especialmente en el caso de variables discretas X sobre $\mathcal{X} = \mathbb{N}$), resulta útil introducir los *momentos factorials* de X mediante

$$f_{l_1,\ldots,l_d}[X] = \mathrm{E}\left[\prod_{i=1}^{d} (X_i)^{\underline{l_i}}\right]$$

con $(x)^{\underline{l}}$ factorial decreciente (ver notaciones). Se puede ver que cuando $\mathcal{X} = \{0\,;\,\ldots\,;\,n\}$, $n \in \mathbb{N}$, $f_{l_1,\ldots,l_d}[X] = 0$ cuando hay por lo menos un $l_i > n$. Cuidense que la notación para estos momentos es un poco diferente de la con los momentos porque resuelte más delicado ppner estos en un tensor. Eso viene de que, por ejemplo, $x_i x_i \neq (x)^{\underline{l}}$; no se puede "separar" los variables en la factorial como se hace en la potencia apareciendo en los momentos.

- Los *momentos centrales* se definen alrededor de la media $\mathrm{E}[X]$, *i. e.*,, como el tensor de los k-ésimo momentos de la *desviación* $X - \mathrm{E}[X]$:

$$\zeta_k[X] \equiv \mathrm{E}\left[(X - \mathrm{E}[X])^{\otimes k}\right].$$

Se escribe también

$$\zeta_{i_1,\ldots,i_k}[X] = \mathrm{E}\left[\prod_{j=1}^{k} \left(X_{i_j} - \mathrm{E}[X_{i_j}]\right)\right] \quad \text{con} \quad (i_1,\ldots,i_k) \in \{1,\ldots,d\}^k$$

Se deduce que si la distribución de probabilidad satisface a una simetría central con respecto a la media, *i. e.*, $X - m_X \overset{\mathrm{d}}{=} -(X - m_X)$ donde $\overset{\mathrm{d}}{=}$ significa que los vectores aleatorios tiene la misma distribución de probabilidad, entonces todos los momentos centrales impares son nulos. Los momentos (centrales) brindan medidas que caracterizan la distribución. Cuando existen:

1. El primer momento, o media:

$$m_X = \mathrm{E}[X].$$

2. El segundo momento central se conoce como *matriz de covarianza*. En el caso escalar, hablamos de *varianza*, o *dispersión* o también *desviación cuadrática media*.

$$\Sigma_X \equiv \mathrm{Cov}[X] \equiv \zeta_2[X] = \mathrm{E}\left[(X - m_X)\,(X - m_X)^t\right].$$

Se conoce también la matriz inversa Σ_X^{-1} bajo la denominaciíon *matriz de precisión*. En el caso escalar, la varianza se escribe en general

$$\mathrm{Var}[X] \equiv \sigma_X^2 = \mathrm{E}\left[(X - m_X)^2\right]$$

y es una medida del cuadrado del ancho efectivo de una densidad de probabilidad (o vector de probabilidad). Para dos vectores aleatorios X e Y respectivamente d y d'-dimensonal (con d' no necesariamente igual a d), hablamos de *covarianza entre X e Y*, que se escribe

$$\Sigma_{X,Y} \equiv \mathrm{Cov}[X,Y] = \mathrm{E}\left[(X - m_X)\,(Y - m_Y)^t\right].$$

Esta matriz partenece a $\mathcal{M}_{d,d'}(\mathbb{R})$ y

$$\Sigma_{X,Y} = \Sigma_{Y,X}^t.$$

.

Se puede notar que, para dos componentes $i \neq j$ de X, Σ_X tiene como (i,j)-ésima componente la covarianza $\mathrm{Cov}[X_i, X_j] = \mathrm{E}\left[(X_i - m_{X_i})\left(X_j - m_{X_j}\right)\right]$ entre las variables X_i y X_j y tiene las varianzas de los X_i en su diagonal. Además, es sencillo ver que $\Sigma_X \in P_d(\mathbb{R})$, *i. e.,* $\mathrm{Cov}[X]$ es simétrica, por construcción, y $\mathrm{Cov}[X] \geq 0$ ($\forall\, \mu \in \mathbb{R}, \quad \mu^t \Sigma_X \mu \geq 0$; en el caso escalar la varianza es no negativa), con igualdad sólo cuando $P_X = \delta_{x_0}$ para un x_0 dado, esto es, cuando no hay incerteza sobre el resultado. De la desigualdad de Cauchy-Bunyakovsky-Schwarz (ver corolario **??**, pagina **??**) se proba sencillamente que

$$|\mathrm{Cov}[X_i, X_j]|^2 \leq \sigma_{X_i}^2 \sigma_{X_j}^2,$$

así que se define también el *coeficiente de correlación* que es adimensional y toma valores entre -1 (variables completamente anticorrelacionadas) y 1 (variables completamente correlacionadas) como:

$$\rho_{i,j} = \rho_{j,i} \equiv \frac{\mathrm{Cov}[X_i, X_j]}{\sigma_{X_i}\sigma_{X_j}}.$$

Como ejemplo, dadas X_1 y $X_2 = aX_1 + b$ que fluctúan en fase ($a > 0$) o al revés ($a < 0$), se tiene la relación entre desviaciones $X_2 - \mathrm{E}[X_2] = a\big(X_1 - \mathrm{E}[X_1]\big)$, conduciendo a $\rho_{1,2} = \frac{a}{|a|} = \pm 1$.
También, se puede ver que

$$\mathrm{Var}\left[\|X - \mathrm{E}[X]\|\right] = \mathrm{Tr}\,\Sigma_X.$$

La covarianza está bien definida si $\|X\|$ es una variable aleatoria de cuadrado integrable, esto es, cuando $E\left[\|X\|^2\right] < \infty$. Se proba sencillamente (desallorando el "producto" y usando la linealidad de la esperanza) que

$$\mathrm{Cov}[X, Y] = \mathrm{E}\left[XY^t\right] - m_X m_Y^t$$

conocido como *fórmula de König-Huygens* o también *translación de Steiner*. En el caso escalar y $X = Y$, es el equivalente al secundo teorema de König de la mecánica [26] (Koenigio, 1751) (ver también (Landau & Lifshitz, 1976, § 8)). Es también el equivalente del teorema de Huygens o de Steiner de la mecánica relacionando el momento de inertia de un solido con respeto al origen en función del momento de inertia con respeto al centro de

[26] Expresa la energía cinética de un sistema de párticulas de masa m, velocidades v_i, como la suma de la energía cinética con respecto al centro de masa $\frac{1}{2}m\sum_i (v_i - v_{\mathrm{cm}})^2$ (equivalente mecánico de la varianza) y de la energía cinética de centro de masa $\frac{1}{2}mv_{\mathrm{cm}}^2$ (equivalente mecánico del cuadrado de la media).

masa (Teodorescu, 2007, § 1.2.8) o (Haas, 1929, p. 104-112). Además, inmediatamente,

$$\forall\, A \in \mathcal{M}_{n,d}(\mathbb{R}),\ B \in \mathcal{M}_{n',d'}(\mathbb{R}),\ a \in \mathbb{R}^n,\ b \in \mathbb{R}^{n'}, \quad \mathrm{Cov}[AX + a, BY + b] = A\,\mathrm{Cov}[X, Y]B^t.$$

En el caso escalar, $d = 1$, lo que es conocido también como el *ancho* de una distribución está dado por la *desviación estándar*

$$\sigma_X = \sqrt{\mathrm{Var}[X]}$$

tiene las mismas unidades de X, y se usa para normalizar los momentos centrales de orden superior. El *ancho relativo* es otra medida que caracteriza la distribución, dado por $\frac{\sigma_X}{m_X} = \sqrt{\frac{\mathrm{E}[X^2]}{m_X^2} - 1}$ cuando $m_X \neq 0$.
Dado un vector aleatorio X, teniendo en cuenta que los dos primeros momentos dan las características más importantes de la distribución de probabilidad, puede resultar conveniente hacer una transformación de variable aleatoria a la llamada *variable estándar*: $Y \equiv \Sigma_X^{-\frac{1}{2}}(X - m_X)$, donde $\Sigma^{-\frac{1}{2}}$ es la única matriz simétrica definida positiva tal que su cuadrado es igual a Σ^{-1} (Horn & Johnson, 2013; Magnus & Neudecker, 1999) que entonces tiene media igual a 0 y una matriz de covarianza igual al identidad I (en el caso escalar, desviación estándar igual a 1).

3. En el caso escalar, el tercer momento central permite definir el *coeficiente de asimetría* o más sencillamente *asimetría* (o skewness en ingles) (Spiegel, 1976; Pearson, 1905):

$$\mathrm{Asim}[X] \equiv \gamma_X \equiv \mathrm{E}\left[\left(\frac{X - m_X}{\sigma_X}\right)^3\right] = \frac{\zeta_3[X]}{\sigma_X^3},$$

momento de orden 3 de la variable estandar, que resulta adimensional y puede tener signo positivo o negativo, anulándose para distribuciones que son simétricas respecto del valor medio. Cuando es positivo, significa que hay más peso a la derecha de la media, y vice-versa. En el contexto multivariado, la extensión natural es el tensor de orden 3 de los momentos del vector centrado y estandardizado $\Sigma_X^{-\frac{1}{2}}(X - m_X)$ (Móri, Rohatgi & Székely, 1994):

$$\mathrm{Asim}[X] \equiv \gamma_X \equiv \mathrm{E}\left[\left(\Sigma_X^{-\frac{1}{2}}(X - m_X)\right)^{\otimes 3}\right].$$

Sin embargo, se puede querer un resumen de la asimetría a través un número escalar o un vector. Se encuentran en la literatura varias proposiciones en esta dirección (Mardia, 1970; Malkovich & Afifi, 1973; Isogai, 1982; Srivastava, 1984; Móri et al., 1994; Balakrishnan, Brito & Quiroz, 2007; Kollo, 2008; Balakrishnan & Scarpa, 2012). La mayoría de estas extensiones se relacionan con γ_X (ej. suma de los terminos al cuadrado en (Mardia, 1970), vector tal que la i-ésima componente es la suma en j de las componentes (j, j, i)-ésima de γ_X en (?, ?) o el vector tal que su i-ésima componente es la suma en j, k de las componentes (j, k, i)-ésima de γ_X en (Kollo, 2008)). Todas coinciden con γ_X o su cuadrado en el caso escalar y cada medida es invariente por transformación $X \mapsto AX + b$ con $A \in \mathcal{M}_{d,d}(\mathbb{R})$ invertible y $b \in \mathbb{R}^d$.

4. En el caso escalar, el cuarto momento central da lugar a la *curtosis* (Pearson, 1905; Westfall, 2014):

$$\mathrm{Curt}[X] \equiv \kappa_X \equiv \mathrm{E}\left[\left(\frac{X - m_X}{\sigma_X}\right)^4\right] = \frac{\zeta_4[X]}{\sigma_X^4},$$

momento de orden 4 de la variable estandar, que posibilita diferenciar entre distribuciones altas y angostas. Veremos más adelante de que para la densidad Gausiana $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-m)^2}{2\,\sigma^2}\right)$, $m_X = m$, $\sigma_X = \sigma$, $\gamma_X = 0$, $\kappa_X = 3$. Se dice de que p_X es alta y angosta, o sub-gausiana, o *con colas livianas* o también platicúrtica cuando $\kappa_X < 3$, y se dice bajas y anchas o sobre-gausiana, o *con colas pesadas* o también leptocúrtica cuando $\kappa_X > 3$ (para $\kappa_X = 3$ la distribución es a veces dicha mesocúrtica). A veces, se define entonces la *curtosis por exceso*

$$\mathrm{ExCurt}[X] \equiv \bar{\kappa}_X = \mathrm{Curt}[X] - 3 = \mathrm{E}\left[\left(\frac{X - m_X}{\sigma_X}\right)^4\right] - 3 = \frac{\zeta_4[X]}{\sigma_X^4} - 3.$$

Más que el pico de distribución, la curtosis (por excesso) describe las colas de una distribución (pesadas para el curtosis por exceso o livianas en el caso contrario) (Westfall, 2014). Como para la asimetría, la extensión natural multivariada de la curtosis es el tensor de orden 4

$$\mathrm{Curt}[X] \equiv \kappa_X \equiv \mathrm{E}\left[\left(\Sigma_X^{-\frac{1}{2}}(X - m_X)\right)^{\otimes 4}\right]$$

Se muestra de nuevo que en el contexto gausiano multivariado $p_X(x) = (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-m)^t\Sigma^{-1}(x-m)\right)$ (ver más adelante), $m_X = m$, $\Sigma_X = \Sigma$, $\gamma_X = 0$, $\kappa_X = \sum_{i,j=1}^{d}\Big(\left(\mathbb{1}_i\mathbb{1}_i^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_j^t\right) + \left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_i\mathbb{1}_j^t\right) + \left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_i^t\right)\Big)$ así que se puede definir una curtosis multivariada por exceso (sub-entendido, con respeto a la gausiana) como:

$$\mathrm{ExCurt}[X] \equiv \bar{\kappa}_X = \mathrm{Curt}[X] - \sum_{i,j=1}^{d}\Big(\left(\mathbb{1}_i\mathbb{1}_i^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_j^t\right) + \left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_i\mathbb{1}_j^t\right) + \left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_i^t\right)\Big).$$

Se encuentren en la literatura varias medidas alternativas dando un resumen de la curtosis a través un número escalar o una matriz (Mardia, 1970; Malkovich & Afifi, 1973; Srivastava, 1984; Móri et al., 1994; Kollo, 2008; Seber, 2004). La mayoría de estas extensiones se relacionan con κ_X (ej. suma de las componentes (i,i,j,j) de κ_X en (Mardia, 1970), matriz de (i,j)-ésima componente igual a la suma en k de las componentes (i,k,k,j)-ésima de κ_X en (?, ?) o la matriz de (i,j)-ésima componente igual a la suma en k,l de las componentes (k,l,i,j)-ésima de γ_X en (Kollo, 2008)). Todas coinciden con κ_X (resp. $\bar{\kappa}_X$) en el caso escalar y cada medida es invariente por transformación $X \mapsto AX + b$ con $A \in \mathcal{M}_{d,d}(\mathbb{R})$ invertible y $b \in \mathbb{R}^d$.

Fijense de que, en el contexto escalar $d = 1$, se vinculan los momentos centrales y los momentos ordinarios directamente de las definiciones:

$$\zeta_k[X] = \sum_{l=0}^{k}\binom{k}{l}(-m_X)^{k-l}\,m_l[X] \qquad \text{y} \qquad m_k[X] = \sum_{l=0}^{k}\binom{k}{l}m_X^{k-l}\zeta_l[X]$$

para cualquier $k \in \mathbb{N}$, siendo $m_0[X] = \zeta_0[X] = 1$. Por ejemplo, $\zeta_2[X] = m_2[X] - m_1[X]^2$ que es nada más que la relación de König-Huyggens, mientras que $\zeta_3[X] = m_3[X] - 3m_1[X]m_2[X] + 2m_1[X]^3$. En el contexto multivariado, las relaciónes momentos–momentos centrales toman las expresiónes

$$\zeta_{i_1,\dots,i_k}[X] = \sum_{I\subset\{1,\dots,k\}} (-1)^{k-|I|} m_J[X] \prod_{l\notin I} m_{X_{i_l}} \qquad \text{y} \qquad m_{i_1,\dots,i_k}[X] = \sum_{I\subset\{1,\dots,k\}} \zeta_I[X] \prod_{l\notin I} m_{X_{i_l}}$$

donde, por abuso de scritura $m_J \equiv m_{j_1,\dots,j_{|J|}}$ (y similarmente para ζ).

1.6.3 Independencia, identidades y desigualdades

Una primera relación interesante concierna el caso de variables independientes y como se comporta la covarianza de estas:

Lema 1-6 (Independencia y covarianza)**.** *Sean X e Y dos vectores aleatorios integrables. Si son independientes, entonces*

$$\mathrm{E}[XY^t] = \mathrm{E}[X]\,\mathrm{E}[Y]^t \qquad \textit{i. e.,} \qquad \mathrm{Cov}[X,Y] = 0.$$

En particular, para X con componentes independientes, $\mathrm{Cov}[X]$ es una matriz diagonal.

Demostración. Sean $X = \sum_j x_j \mathbb{1}_{A_j}$ e $Y = \sum_k y_k \mathbb{1}_{B_k}$ dos variables escalonadas. Entonces, $A_j = (X = x_j)$ y $B_k = (Y = y_k)$. Luego

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_{j,k} x_j y_k \,\mathrm{E}\left[\mathbb{1}_{A_j}\mathbb{1}_{B_k}\right] \\
&= \sum_{j,k} x_j y_k \,\mathrm{E}\left[\mathbb{1}_{A_j\cap B_k}\right] \\
&= \sum_{j,k} x_j y_k P(A_j \cap B_k) \\
&= \sum_{j,k} x_j y_k P(X = x_j)P(Y = y_k) \quad \text{(de la independencia)}
\end{aligned}$$

dando el resultado para variables escalonadas. Se cierra la prueba para variables positivas como límite de crecientes de funciones escalonadas, y variables reales tratando las partes positivas y negativas aparte. El caso vectorial se deduce trabajando con pares de componentes. □

Notar que la reciproca es falsa en general:

Ejemplo 1-16 (Uniforme sobre el disco unitario)**.** *Sea $X = (X_1, X_2)$ uniforme sobre el disco unitario o bola unitaria 2-dimensional $\mathbb{B}_2$ (ver notaciones), i.e., $p_X(x) = \frac{1}{\pi}\mathbb{1}_{\mathbb{S}^2}(x)$. Claramente, los X_i no pueden ser independientes del hecho que $\mathcal{X}_i = [-1\,;\,1]$ y $\mathcal{X} \neq \mathcal{X}_1 \times \mathcal{X}_2$ (es estrictamente incluido en el*

producto cartesiano). Por simetría central de p_X*, es sencillo ver que* $\mathrm{E}[X_1 X_2] = 0$ *y similarmente* $\mathrm{E}[X_i] = 0$*: a pesar de que los* X_i *no sean independientes,* $\mathrm{Cov}[X_1, X_2] = 0$.

La consecuencia de la independencia sobre la covarianza facilita frecuentemente los calculos de media. Volviendo al ejemplo 1-4:

Ejemplo 1-17 (Continuación del ejemplo 1-4)**.** *Tratando de la media de* $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ *con* U *y* V *variables independientes de distribución uniformas sobre* $(0\,;\,1)$*, se calcula gracia a la linealidad y a la independencia,* $\mathrm{E}[X] = \mathrm{E}[V]\,\mathrm{E}\left[\mathbb{1}_{U<\frac{1}{2}}\right] + \mathrm{E}\left[\mathbb{1}_{U\geq\frac{1}{2}}\right] = \frac{1}{2}\times\frac{1}{2} + \frac{1}{2} = \frac{3}{4}$ *como lo hemos obtenido usando* P_X *en Ej. 1-14 o la positividad en Ej. 1-15.*

Una otra consecuencia de esta proposición trata de un conjunto de vectores aleatorios $\{X_i\}$ y un conjunto de matrices de dimensiones adecuadas,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i \Sigma_{X_i} A_i^t + \sum_{j\neq i} A_i\,\mathrm{Cov}[X_i, X_j] A_j^t.$$

En particular, en el caso escalar,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i^2\,\mathrm{Var}[X_i] + \sum_{j\neq i} A_i A_j\,\mathrm{Cov}[X_i, X_j].$$

Si los X_i son independientes, entonces las covarianzas conjuntas son nulas así que, respectivamente,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i \Sigma_{X_i} A_i^t \qquad \text{y} \qquad \mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i^2 \sigma_{X_i}^2.$$

Si el teorema da una implicación de la independencia, de hecho existe una reciproca que toma la forma siguiente:

Teorema 1-19 (Independencia y momentos)**.** *Sean* X *e* Y *dos vectores aleatorios. Son independientes si y sólo si* $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ *para todos pares de funciones* f *y* g*, medibles y acotadas de dimensiones adecuadas.*

Demostración. Se puede referirse a (Feller, 1971; Jacob & Protters, 2003) para una prueba rigurosa. En el caso escalar, el principio consiste a ver f y g como límites de funciones escalonadas. Para $f(x) = \sum_i a_i \mathbb{1}_{A_i}(x)$ y $g(y) = \sum_j b_j \mathbb{1}_{B_j}(y)$ se obtiene $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ si y sólo si $\sum_{i,j} a_i b_j \left(P((X\in A_i)\cap(Y\in B_j)) - P(X\in A_i)P(Y\in B_j)\right) = 0$. Básicamente, eso debe valer para cualquieras A_i, B_j y a_i, b_j, así que el término entre parentesis debe ser cero, lo que es nada más que la definición de la independencia de X e Y. El caso vectorial se entiende por pares de componentes. □

Relaciones tambíen muy útiles son conocidas como *Desigualdades de Chebyshev* (Bienaymé, 1853a; Tchébichev, 1867; Markov, 1884; Olkin & Pratt, 1958; Ferentinos, 1982; Navarro, 2013; Stellato, Van Parys & Goulart, 2017). Estas desigualdades dan una cota superior a la probabilidad de que una cantidad que fluctúa aleatoriamente exceda cierto valor umbral, aún sin conocer detalladamente la forma de la distribución de probabilidad.

Teorema 1-20 (Desigualdades de Chebyshev)**.** *Sea un vector aleatorio d-dimensional X y una función $g : \mathbb{R}^d \mapsto \mathbb{R}_+$ medible tal que $g(X)$ sea integrable. Entonces,*

$$\forall\, a > 0, \quad P(g(X) \geq a) \;\leq\; \frac{\mathrm{E}[g(X)]}{a}.$$

Demostración. Sea $\mathcal{D}_a = \{x \in \mathcal{X} \;\big|\; g(x) \geq a\} \subset \mathcal{X}$. Entonces, g siendo non negativa,

$$\mathrm{E}[g(X)] = \int_{\mathcal{X}} g(x)\, dP_X(x) \geq \int_{\mathcal{D}_a} g(x)\, dP_X(x) \geq \int_{\mathcal{D}_a} a\, dP_X(x) = aP(X \in \mathcal{D}_a).$$

Se cierra la prueba notando de que $(X \in \mathcal{D}_a) = (g(X) \geq a)$. □

Existen varias formas similares, que son de hecho casos particulares de estas desigualdades.

Corolario 1-3 (Bienaymé–Chebyshev)**.** *Sea X un vector aleatorio d-dimensional admitiendo una esperanza m_X y una covarianza Σ_X. Entonces,*

$$\forall\, \varepsilon > 0, \qquad P\left(\left\|\Sigma_X^{-\frac{1}{2}}(X - m_X)\right\| > \varepsilon\right) \;\leq\; \frac{d}{\varepsilon^2}.$$

Viene del teorema incial aplicado a $\Sigma_X^{-\frac{1}{2}}(X - m_X)$, $g(x) = \|x\|^2$ y $a = \varepsilon^2$.

Corolario 1-4 (Markov)**.** *Sea X un vector aleatorio y $\varphi \geq 0$ una función no decreciente tal que $\varphi(\|X\|)$ sea integrable. Entonces,*

$$\forall\, \varepsilon \geq 0, \quad \textit{tal que} \quad \varphi(\varepsilon) \neq 0, \qquad P\left(\|X\| > \varepsilon\right) \;\leq\; \frac{\mathrm{E}\left[\varphi\left(\|X\|\right)\right]}{\varphi(\varepsilon)}.$$

La versión inicial des esta desigualdad trataba de funciones $\varphi(u) = u^r$, $r > 0$. Viene del teorema incial aplicado a $g(x) = \varphi\left(\|x\|\right)$ y $a = \varphi(\varepsilon)$, notando de que $\left(\varphi\left(\|X\|\right) \geq \varphi(\varepsilon)\right) = \left(\|X\| \geq \varepsilon\right)$ por la no decrecencia de φ. El caso anterior (una vez la variable centrada) es nada más que un caso especial.

Estas relaciones afirman que cuanto más chica es la varianza, más se concentra la variable en torno a su media. Ambas cotas son en general débiles, como se lo puede ver en el ejemplo siguiente

Ejemplo 1-18. *La desigualdad de Bienaymé–Chebyshev indica que la probabilidad de encontrar una fluctuación superior a $\varepsilon = 3\sigma_X$, tres desviaciones estándar alrededor de la media, está por debajo de $1/9$; el cálculo para una distribución típica como la Gaussiana, $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{(x-m_X)^2}{2}\right)$ ajusta dicha probabilidad por debajo de $0{,}003$.*

Una desigualdad muy importante que usaremos frecuentemente en el cápitulo siguiente, trata de funciones convexa, y del efecto sobre la media de una vector aleatorio.

Definición 1-36 (Función convexa)**.** *Por definición, una función $\phi : \mathcal{X} \subset \mathbb{R}^d \mapsto \mathbb{R}$ con $\mathcal{X}$ un convexo, es convexa si para cualquier $\pi_1 \in [0\,;\,1], \pi_2 = 1 - \pi_1$ y $x_1, x_2 \in \mathbb{R}^d$,*

$$\phi(\pi_1 x_1 + \pi_2 x_2) \leq \pi_1 \phi(x_1) + \pi_2 \phi(x_2).$$

ϕ es dicha estrictamente convexa si la desigualdad es estricta, salvo si $x_2 = x_1$.
Se puede ver de que si ϕ es dos veces diferenciable, su matriz Hessiana es simétrica no negativa,

$\mathcal{H}\phi \in P_d(\mathbb{R})$.

Se muestra por recurrencia que para cualquier conjunto $\{x_i\}_i$ *numerable de elementos de* $\mathcal{X}$ *y reales positivos* $\{\pi_i\}_i$ *tales que* $\sum_i \pi_i = 1$*,*

$$\phi\left(\sum_i \pi_i x_i\right) \leq \sum_i \pi_i \phi(x_i).$$

Dicho con palabras, la funcción del barycentro (combinación convexa) de los x_i *es debajo del barycentro de los* $\phi(x_i)$*. Eso es ilustrado el la figura Fig. 1-12.*

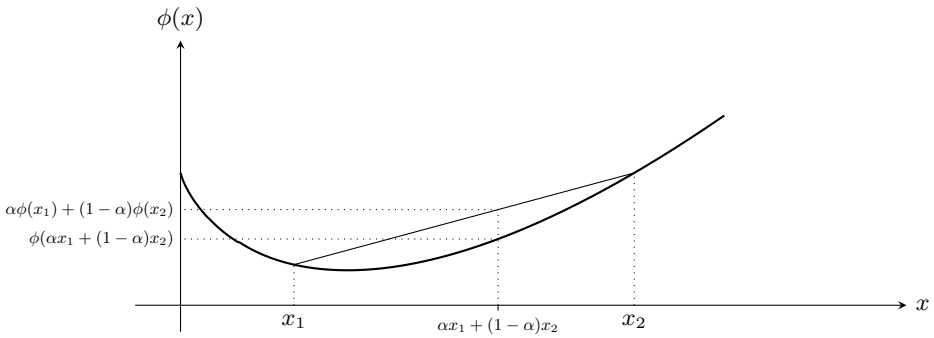


Figura 1-12: Ejemplo de función ϕ convexa: la cuerda, conteniendo los barycentros de $\{\phi(x_1)\,;\,\phi(x_2)\}$, es siempre arriba de la curba, *i. e.,* función de los barycentros de $\{x_1\,;\,x_2\}$.

Intuitivamente, la media teniendo un sabor de barycentro, se intuye de que la media de $\phi(X)$ va a ser arriba de la función de la media de X. Es precisamente el teorema de Jensen [27] (Jensen, 1906; Feller, 1971; Brémaud, 1988; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013):

Teorema 1-21 (Desigualdad de Jensen)**.** *Sea* X *integrable y definida sobre* $\mathcal{X} \subset \mathbb{R}^d$*, convexo y* $\phi : \mathcal{X} \mapsto \mathbb{R}$ *función convexa. Entonces*

$$\mathrm{E}[\phi(X)] \geq \phi\left(\mathrm{E}[X]\right).$$

Si ϕ *es estrictamente convexa, la igualdad se alcanza si y solamente si* X *es determinista casi siempre.*

Demostración. Sea $X = \sum_i x_i \mathbb{1}_{A_i}$ variable escalonada. Entonces $\phi(X) = \sum_i \phi(x_i)\mathbb{1}_{A_i}$, dando

$$\mathrm{E}[\phi(X)] = \sum_i P(A_i)\phi(x_i) \geq \phi\left(\sum_i P(A_i)x_i\right) = \phi\left(\mathrm{E}[X]\right)$$

con igualdad (cuando la convexidad es estricta) si y solamente si todos los x_i son iguales. Se cierra la prueba tomando $X \geq 0$ como límite de sucesión de funciones escalonadas (teorema 1-3), y cualquier X tratando de la parte positiva y negativa. El caso vectorial se trata componente a componente para X

[27] En (Jensen, 1906) se trata del en el caso discreto y integral; en (Hölder, 1889; Hadamard, 1893) se encuentran las primeras semillas de esta desigualdad, y entre otros (Jessen, 1931a, 1931b; Perlman, 1974; Rudin, 1991) para versiones más generales.

en termino de límite. Tomando el límite, la condición x_i todos iguales vuelve "casi todos" los x_i deben ser iguales, *i. e.,* X debe ser constante casi siempre. □

Terminamos esta sección con una desigualdad también muy útil, y conocida en los espacios de Hilbert, conocida como *desigualdad de Hölder* (Hölder, 1889; Shohat, 1929):

Teorema 1-22 (Desigualdad de Hölder)**.** *Sean* X *e* Y *dos vectores aleatorios* d*-dimensionales y* $r > 1$ *real y* $r^* > 1$ *tal que* $\frac{1}{r} + \frac{1}{r^*}$*, llamado conjugado de Hölder de* r*. Entonces,*

$$\left|\mathrm{E}\left[X^t Y\right]\right| \leq \mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}} \mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}}.$$

Se obtiene la igualda si y solamente si existe un λ *tal que* $X = \lambda Y$ *casi siempre.*

Demostración. Obviamente, $|\mathrm{E}\left[X^t Y\right]| \leq \mathrm{E}\left[|X^t Y|\right]$. Luego, de la convexidad de la función $-\log$ se obtiene la desigualdad $\log(|ab|) = \frac{1}{r}\log|a|^r + \frac{1}{r^*}\log|b|^{r^*} \leq \log\left(\frac{|a|^r}{r} + \frac{|b|^{r^*}}{r^*}\right)$ con igualdad si y solamente si a es proporcional a b. Ahora, para dos vectores a y b, tenemos $|a^t b| \leq \sum_i |a_i b_i|$; se aplica la desigualdad con el logaritmo a cada $|a_i b_i|$ para obtener la desigualdad de Young $|a^t b| \leq \frac{\|a\|_r^r}{r} + \frac{\|b\|_{r^*}^{r^*}}{r^*}$ con igualdad si y solamente si los vectores son proporcional. A continuación, denotando

$$\widetilde{X} = \frac{X}{\mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}}} \qquad \text{y} \qquad \widetilde{Y} = \frac{Y}{\mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}}}$$

tenemos

$$\mathrm{E}\left[\left|X^t Y\right|\right] = \mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}} \mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}} \mathrm{E}\left[\left|\widetilde{X}^t \widetilde{Y}\right|\right].$$

De la desigualdad de Young, se obtiene entonces

$$\mathrm{E}\left[\left|\widetilde{X}^t \widetilde{Y}\right|\right] \leq \frac{\mathrm{E}\left[\left\|\widetilde{X}\right\|_r^r\right]}{r} + \frac{\mathrm{E}\left[\left\|\widetilde{Y}\right\|_{r^*}^{r^*}\right]}{r^*} = \frac{1}{r} + \frac{1}{r^*} = 1,$$

lo que cierra la prueba. □

Un corolario es conocido como desigualdad de Cauchy-Bunyakovsky-Schwarz [28] para $p = \frac{1}{2}$:

Corolario 1-5 (Desiguladad de Cauchy-Bunyakovsky-Schwarz)**.** *Sean* X *e* Y *dos vectores aleatorios* d*-dimensionales. Entonces*

$$\left|\mathrm{E}\left[X^t Y\right]\right|^2 \leq \mathrm{E}\left[\|X\|^2\right] \mathrm{E}\left[\|Y\|^2\right].$$

Se obtiene la igualda si y solamente si existe un λ *tal que* $X = \lambda Y$ *casi siempre.*

[28] Esta desigualdad, fue probada por Cauchy para sumas en 1821 (Cauchy, 1821), para integrales por Bunyakovsky en 1859 (Bouniakowsky, 1859) y más elegamente por Schwarz en 1888 (Schwarz, 1888) en un enfoque más general. Ver también (Steele, 2004).

Nota: se puede probar esta desigualdad considerando el polinomio $\mathrm{E}\left[\|\lambda X + Y\|^2\right] \geq 0$, del secundo orden en λ. Siendo no negativa para cualquier λ el discriminente debe ser no positivo, conduciendo a la desigualdad.

Se notará que de este corolario se puede víncular la asimetría y la curtosis via la desigualdad $\gamma_X^2 \leq \kappa_X$.

De hecho, se puede ver $\mathrm{E}[X^t Y]$ como un producto escalar entre variables aleatorias. La sola subtileza es que $\mathrm{E}\left[\|X\|^2\right] = 0$ conduce a $X = 0$ casi siempre, *i. e.,* se puede tener $X \neq 0$ pero con medida de probabilidad igual a cero (ej. puntos ω "aislados" en el contexto continuo).

Un otro corolario de la desiguldad de Hölder concierna el comportamiento de $s \mapsto \mathrm{E}\left[\|X\|_s^s\right]^{\frac{1}{s}}$ dado X:

Corolario 1-6 (Crecencia de $s \mapsto \mathrm{E}\left[\|X\|_s^s\right]^{\frac{1}{s}}$)**.** *Sea X vectore aleatorio d-dimensionales. Entonces*

$$s \mapsto \mathrm{E}\left[\|X\|_s^s\right]^{\frac{1}{s}} \quad \textit{es creciente}$$

Demostración. Aplicando la desigualdad de Hölder a $\|X\|_s^s$ y 1 se obtiene

$$\forall\, r > 1, \quad \mathrm{E}\left[\|X\|_s^s\right] \leq \mathrm{E}\left[\|X\|_{rs}^{rs}\right]^{\frac{1}{r}}$$

Se cierra la prueba elevando la desigualdad a la potencia $\frac{1}{s}$ y notando que $t = rs > s$. □

Varias otras desigualdades se encuentran en la literatura (ver por ejemplo en (Shohat, 1929, y notas debidas a Pearson) para unas de las más antiguas), así que no se puede ser exaustivo. Presentamos en esta sección las principales.

1.7 Esperanza condicional

Vimos en la sección 1.5 que una pregunta natural era, dados dos vectores aleatorios X e Y, de caracterizar el vector Y si "observamos X". Más adelante, nos podemos interesar a la media de Y cuando observamos X. Una manera intuitiva es de definir tal media cuando "sabemos" que $X = x$ a partir de la ley condicional $P_{Y|X=x}$ (Feller, 1968, 1971; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Spiegel, 1976; Kolmogorov, 1956; Jacob & Protters, 2003; Billingsley, 2012):

Definición 1-37 (Esperanza condicional)**.** *Sean X e Y dos vectores aleatorios respectivamente d_X y d_Y-dimensionales, y sea la función*

$$f(x) \equiv \mathrm{E}[Y|X = x] = \int_{\mathbb{R}^{d_Y}} y\, dP_{Y|X=x}(y)$$

Se define la esperanza de Y condicionalmente a X como siendo la variable aleatoria

$$\mathrm{E}[Y \mid X] = f(X)$$

Claramente, $\mathrm{E}[Y|X = x]$ siendo una esperanza vínculado al espacio de probabilidad $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_{Y|X=x}\big)$, heride de las propiedades de la media. Entre otros, es lineal, lo que da

$$\mathrm{E}\left[\sum_i a_i Y_i \,\middle|\, X\right] = \sum_i a_i \,\mathrm{E}[Y_i \,|\, X]$$

satisface la desiguadad de Jensen, para ϕ convexa

$$\mathrm{E}\left[\phi(Y) \,|\, X\right] \geq \phi\big(\,\mathrm{E}[Y \,|\, X]\,\big)$$

entre otros.

Como en el caso de medida de probabilidad, cuando dos variables son independientes, condicionar no cambia la esperanza:

Lema 1-7 (Esperanza condicional e independencia)**.** *Cuando X e Y son independientes, la esperanza condicional coincide con la de Y,*

$$X \text{ e } Y \text{ independientes} \quad \Rightarrow \quad \mathrm{E}[Y|X] = \mathrm{E}[Y]$$

Demostración. El resultado viene inmediatamente de $P_{Y|X=x} = P_Y$ (ver lema 1-5). □

La media condicional se revela muy útil y poderoso para evaluar esperanzas de variables aleatorias por ejemplo gracia a la formula de la esperanza total, equivalente de las formulas de probabilidad total lema 1-1, lema 1-14 y lema 1-16.

Teorema 1-23 (Media total)**.** *La media (total) del vector aleatorio Y concide con la media de la esperanza condicional, i. e.,*

$$\mathrm{E}[Y] = \mathrm{E}\left[\,\mathrm{E}\,[Y|X]\,\right]$$

Más generalmente, para cualquier función medible f,

$$\mathrm{E}[f(Y)] = \mathrm{E}\left[\,\mathrm{E}\,[f(Y)|X]\,\right]$$

Demostración. De la fórmula de probabilidad total tenemos para cualquier $B \in \mathcal{B}(\mathbb{R}^{d_Y})$,

$$\begin{aligned}
\int_B dP_Y(y) &= P(Y \in B) \\
&= \int_{\mathbb{R}^{d_X}} P_{Y|X=x}(B)\, dP_X(x) \\
&= \int_{\mathbb{R}^{d_X}} \left(\int_B dP_{Y|X=x}(y)\right) dP_X(x)
\end{aligned}$$

es decir

$$\int_{\mathbb{R}^{d_Y}} \mathbb{1}_B(y)\, dP_Y(y) = \int_{\mathbb{R}^{d_X}} \left(\int_{\mathbb{R}^{d_Y}} \mathbb{1}_B(y)\, dP_{Y|X=x}(y)\right) dP_X(x)$$

i. e.,

$$\mathrm{E}[\mathbb{1}_B(Y)] = \mathrm{E}\left[\,\mathrm{E}[\mathbb{1}_B(Y)|X]\right]$$

Ahora, se usa la linealidad para cualquier función escalonada f, y luego por el teorema de convergencia monótona 1-4, para cualquier función medible,

$$\mathrm{E}[f(Y)] = \mathrm{E}\left[\,\mathrm{E}[f(Y)|X]\right].$$

□

Un otro resultado importante, permitiendo frecuentemente simplificar la evaluacíon de momentos a partir de esperanza condicional es el siguiente:

Teorema 1-24. *Para cualquier funciónes medibles f, g, tenemos*

$$\mathrm{E}\left[\,f(X)\,g(Y)\mid X\right] = f(X)\;\mathrm{E}\left[g(Y)|X\right]$$

Más generalmente, para h medible

$$\mathrm{E}\left[\,h(X,Y)\mid X=x\right] = \mathrm{E}\left[h(x,Y)|X=x\right]$$

lo que se simplifica si además X e Y son independientes:

$$X \text{ e } Y \text{ independientes} \quad\Rightarrow\quad \mathrm{E}\left[\,h(X,Y)\mid X=x\right] = \mathrm{E}\left[h(x,Y)\right]$$

Demostración. De la definición de la medida de probabilidad condicional tenemos para cualquier $A \in \mathcal{B}(\mathbb{R}^{d_X}),\quad B \in \mathcal{B}(\mathbb{R}^{d_Y}),\quad C \in \mathcal{B}(\mathbb{R}^{d_X})$,

$$P\big((X\in A)\cap(Y\in B)\cap(X\in C)\big) = \int_C P_{X,Y|X=x}(A\times B)\,dP_X(x)$$

pero, también

$$P\big((X\in A)\cap(Y\in B)\cap(X\in C)\big) = \int_{A\cap C} P_{Y|X=x}(B)\,dP_X(x)$$

Entonces,

$$P_{X,Y|X=x}(A,B) = \mathbb{1}_A(x)\,P_{Y|X=x}(B)$$

A continuación,

$$\int_{\mathbb{R}^{d_X}\times\mathbb{R}^{d_Y}} \mathbb{1}_A(u)\,\mathbb{1}_B(v)\,dP_{X,Y|X=x}(u,v) = \mathbb{1}_A(x)\int_{\mathbb{R}^{d_Y}} \mathbb{1}_B(v)\,dP_{Y|X=x}(v)$$

Entonces, por linealidad, aplicando este resultado a funciones escalonadas, y luego por el teorema de convergencia monótona, para cualquieras f, g medibles,

$$\int_{\mathbb{R}^{d_X}\times\mathbb{R}^{d_Y}} f(u)\,g(v)\,dP_{X,Y|X=x}(u,v) = f(x)\int_{\mathbb{R}^{d_Y}} g(v)\,dP_{Y|X=x}(v)$$

es decir, por definición de la esperanza condicional,

$$\mathrm{E}[f(X)\,g(Y)|X=x] = f(x)\;\mathrm{E}[g(Y)|X=x]$$

lo que cierra la prueba de la primera identidad. Las otras se prueban con los mismos pasos. □

Un resultado que sirve a veces como definición, en el contexto de variable de cuadrado integrable, se víncula con la idea de aproximar una variable por una funcción de una otra:

Teorema 1-25. *Sea Y de cuadrado integrable, la esperanza condicional $E[Y|X]$ es la única variable $Z = f(X)$, función de X de cuadrado integrable, minimizando el error promedio cuadratico $\mathrm{E}[\|Y - Z\|^2]$. Dicho de otra manera, con el criterio de error cuadratico promedio mínimo, $E[Y|X]$ es la "mejor" función de X (en el sentido de la distancia inducida por el producto escalar) aproximando Y.*

Demostración. Usando la fórmula de esperanza total, y el teorema 1-24, se escribe

$$\begin{aligned} \mathrm{E}\left[\|Y - f(X)\|^2\right] &= \mathrm{E}\left[\mathrm{E}\left[\|Y - f(X)\|^2 \mid X\right]\right] \\ &= \mathrm{E}\left[f(X)^2 - 2f(X)\,\mathrm{E}[Y|X] + \mathrm{E}\left[Y^2|X\right]\right] \end{aligned}$$

Ahora, buscando $\lambda \equiv f(x)$ minimizando $\|\lambda\|^2 - 2\lambda^t\,\mathrm{E}[Y|X = x] + \mathrm{E}\left[\|Y\|^2|X = x\right]$ para cualquier $x \in \mathcal{X}$, se minimizará el promedio en X. Inmediatamente, notando que buscamos el mínimo de un paraboliode de concavivad por arriba, anulando el gradiente en λ so obtiene $\lambda \equiv f(x) = \mathrm{E}[Y|X = x]$, el único mínimo, lo que cierra la prueba. □

Este resultado es muy conocido en el mundo de la estimación donde se quiere aproximar una variable minimizando el error cuadratico promedio (Kay, 1993; Robert, 2007).

Corolario 1-7. *Sea Y de cuadrado integrable, La esperanza condicional $E[Y|X]$ es la única variable $Z = f(X)$ de cuadrado integrable tal que para cualquier función medible g tal que $g(X)$ es de cuadrado integrable, $\mathrm{E}[g(X)^t Y] = \mathrm{E}[g(X)^t Z]$.*

Demostración. Como lo hemos visto en la sección anterior, $(U, V) \mapsto \mathrm{E}[U^t V]$ define un producto escalar. Según el teorema de proyección ortogonal (ver figura 1-13 y (Jacob & Protters, 2003; Athreya & Lahiri, 2006; ?, ?)), el único $f(X)$ que minimiza $\mathrm{E}\left[\|Y - f(X)\|^2\right]$, es la proyección ortogonal de Y sobre el espacio $\sigma^2(X) = \left\{g(X) \mid g \text{ es medible con } \wedge\ g(X) \text{ es de cuadrado integrable}\right\}$. En otros terminos, la desviación $Y - \mathrm{E}[Y|X]$ entre Y y $\mathrm{E}[Y|X]$ es ortogonal a cualquier $g(X) \in \sigma^2(X)$, y reciprocamente si $Y - f(X)$ es ortogonal a cualquier $g(X)$, $f(X)$ es necesariamente la proyección ortogonal de Y sobre $\sigma^2(X)$, lo que cierra la prueba. □

A veces, este resultado sirve también como definición de la esperanza condicional.

1.8 Funciones generadoras

Como lo hemos visto, un vector aleatorio es completamente definida por su medida de probabilidad P, o equivalemente por la medida imagen P_X, o a través de la función de repartición F_X. Sin embargo, bajo el impulso de Laplace en el siglo XVII (entre otros), se introdujo caracterizaciones alternativas a través de transformaciones de la medida de probabilidad, conocidas como *funciones generadoras*

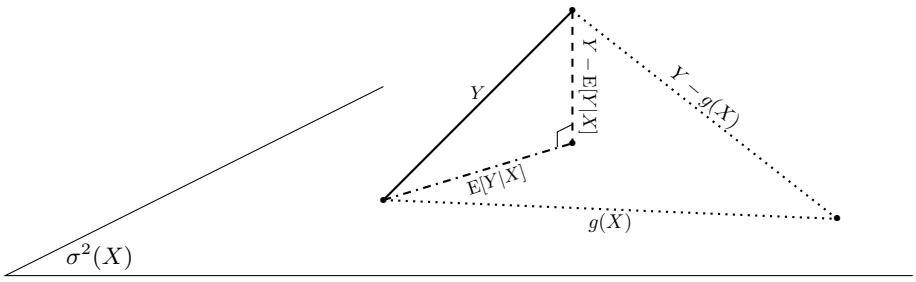


Figura 1-13: Ilustración del teorema de proyección ortogonal. El espacio $\sigma^2(X) = \{g(X) \mid g \text{ es medible con } \wedge \; g(X) \text{ es de cuadrado integrable}\}$ es representado por el plano y el vector representa Y. La linea punteada representa la desviación $Y - g(X)$ entre Y y $g(X)$ dado $g(X) \in \sigma^2(X)$, siendo su cuadrado promedio mínimo cuando $g(X) = \mathrm{E}[Y|X]$ corresponde a la proyección ortogonal de Y. La linea con guiones representa la desviación de cuadrado promedio mínimo y la linea mixta, la proyección ortogonal $\mathrm{E}[Y|X]$.

o *funciones generatrices* [29] (Laplace, 1812, 1814; de Laplace, 1820). Existen varias funciones, cuyas tienes propiedades particulares que vamos a ver en las subsecciones siguientes. Entre otros, estas funciones dadas como valores de expectación de funciones de la variable aleatoria (discreta o continua), con un parámetro real o complejo, permiten hallar fácilmente los distintos momentos de una distribución de probabilidad.

1.8.1 Función generadora de probabilidad

De manera general, siguiendo el enfoque de A. de Moivre (ver nota de pie 29) dada una sucesión $\{a_n\}_{n\in\mathbb{N}}$, se define la función generadora dicha *ordinaria* de la sucesión como $G(\{a_n\}_{n\in\mathbb{N}}, z) = \sum_{n\in\mathbb{N}} a_n z^n$. A veces, esta serie es conocida como transformada en z de la sucesión $\{a_n\}_{n\in\mathbb{N}}$. Tratando de variables aleatorias discretas sobre $\mathbb{N}$, con $p_n = P_X(n) = P(X = n)$, se puede definir así la función generadora asociada a la sucesión p_n y se puede ver que no es nada más que el momento $\mathrm{E}\left[z^X\right]$. De manera general, la función generadora de probabilidad se define de la manera siguiente (Feller, 1968; Johnson, Kotz & Balakrishnan, 1997; Mukhopadhyay, 2000; Athreya & Lahiri, 2006):

Definición 1-38 (función generadora de probabilidad o de momentos factoriales)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ *vector aleatorio* d*-dimensional definido sobre* $\mathcal{X} \subset \mathbb{R}^d$*. La función definida por*

$$G_X(z) = \mathrm{E}\left[\prod_{i=1}^{d} z_i^{X_i}\right] \quad \textit{con} \quad z = \begin{bmatrix} z_1 & \cdots & z_d \end{bmatrix}^t \in \mathbb{C}^d$$

[29] De hecho, de manera general, se introdujeron tales funciones en un marco más general, asociado a sucesiones de números, bajo el impulso de A. de Moivre (de Moivre, 1730); ver también (Stirling, 1730; Euler, 1741, 1750; de Moivre, 1756) o (Knuth, 1997, Sec. 1.2.9).

es conocida como función generadora de probabilidad *o* función generadora de momentos factoriales *de* X.

Esta función est definida sobre un producto cartesiano de anillos [30] en el plano complejos, $r_i \leq |z_i| \leq R_i$ con $r_i \leq 1$ y $R_i \geq 1$.

La denominación *generadora de probabilidad* (pgf para *probability generating function* en ingles) se entiende sencillamente del hecho siguiente:

Lema 1-8. *Cuando* $\mathcal{X} = \mathbb{N}^d$ *para cualquier* $x = \begin{bmatrix} x_1 & \dots & x_d \end{bmatrix}^t \in \mathrm{P}_{k,d} = \left\{ x = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}^t \in \mathbb{N}^d \;\middle|\; \sum_{i=1}^d x_i = k \right\}$ *(ver notaciones)*

$$\frac{1}{\prod_{i=1}^d x_i!} \left. \frac{\partial^k G_X}{\partial z_1^{x_1} \dots \partial z_d^{x_d}} \right|_{z=0} = P_X(x) = P(X = x)$$

Demostración. Se puede escribir la función G_X bajo su forma de generadora ordinaria $G_X(z) = \sum_{x \in \mathbb{N}^d} \left(\prod_{i=1}^d z_i^{x_i} \right) P(X = x)$ con $x = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}^t$. A continuación, se nota que la serie converge uniformamente por lo menos en la bola $\mathbb{B}_d \equiv \mathbb{B}_d(1)$, probando que G_X es diferenciable en $\mathbb{B}_d$, así que esta series es nada mas que el desarollo de Taylor de G_X al torno de $z = 0$ (o, equivalentemente, se puede diferenciar la suma y tomar la derivada en $z = 0$), lo que cierra la prueba. □

De este resultado, se puede notar que, en el caso discreto, hay una relación uno-a-uno entre la medida de probabilidad P_X y la función generadora de probabilidad G_X. En el caso general, veremos en la subsección **??** que para z_j de la forma $z_j = e^{\imath u_j}$ con $u_j \in \mathbb{R}$ la transformación se inversa, de manera que se puede recuperar la medida de probabilidad P_X a partir de G_X. Dicho de otra manera, como la medida P_X, la función G_X caracteriza completamente el vector aleatorio X.

---

[30] Para $i = 1, \dots, d$, sean los reales r_i, R_i tales que $0 \leq r_i \leq R_i$. Consideramos $j = \begin{bmatrix} j_1 & \cdots & j_d \end{bmatrix}^t \in \{-1\,;\,+1\}^d$ y a continuación $\mathcal{D}_j = \bigtimes_{i=1}^d (j_i \mathbb{R}_+)$ $(-\mathbb{R}_+ = \mathbb{R}_-)$, los 2^d hipercuadrantes de $\mathbb{R}^d$. Notamos $\rho_{j_i} = r_i \mathbb{1}_{\{-1\}}(j_i) + R_i \mathbb{1}_{\{+1\}}(j_i)$. Supongamos que en cada hipercuadrante $\int_{\mathcal{D}_j} \prod_{i=1}^d \rho_{j_i}^{x_i} \, dP_X(x) < +\infty$. Es sencillo ver que para $x_i \in (j_i \mathbb{R}_+)$, $r_i \leq |z_i| \leq R_i \;\Rightarrow\; |z_i|^{x_i} \leq \rho_{j_i}^{x_i}$. Entonces, por teorema de convergencia dominada las integrales $\int_{\mathcal{D}_j} \prod_{i=1}^d z_i^{x_i} \, dP_X(x)$ convergen, y por consecuencia, $\int_{\mathbb{R}^d} \prod_{i=1}^d z_i^{x_i} \, dP_X(x)$ converge. Ahora, en el producto cartesiano de los circulos unitarios $|z_i| = 1$ tenemos $\left| \prod_{i=1}^d z_i^{x_i} \right| = 1$, de integral sobre $\mathcal{D}_j$ convergente (vale $P_X(\mathcal{D}_j) \leq 1$). De nuevo, por teorema de convergencia dominada, $\int_{\mathcal{D}_j} \prod_{i=1}^d z_i^{x_i} \, dP_X(x)$ converge sobre el producto cartesiano de circulos unitarios, lo que asegura la existancia de $R_i \geq 1$ y $r_i \leq 1$ (un anillo puede ser resctricto al circulo).

Aparece que la función generadora G_X se vincula también con los momentos factoriales, justificando su secunda denominación, *generadora de momentos factoriales* (fmgf para *factorial moments generating function* en ingles):

Lema 1-9. *Para cualquier* $k = \begin{bmatrix} k_1 & \cdots & k_d \end{bmatrix}^t \in \mathrm{P}_{K,d}$, *derivando* G_X *se proba que, cuando existen* [31]

$$\left.\frac{\partial^K G_X}{\partial z_1^{k_1} \cdots \partial z_d^{k_d}}\right|_{z=1} = \mathrm{E}\left[\prod_{i=1}^{d} (X_i)^{\underline{k_i}}\right]$$

momento factorial [32] *de* X.

De este resultado, se ve por ejemplo que, cuando existen, se recuperan los momentos de X a través de las derivadas de G_X:

- $G_X(1) = 1$, condición de normalización.
- $\nabla_z G_X(1) = \mathrm{E}[X]$.
- $\mathcal{H}_z G_X(1) + \mathrm{diag}\left(\nabla_z G_X(1)\right) = \mathrm{E}\left[XX^t\right]$ donde $\mathcal{H}_z$ es la matrice Hessiana y $\mathrm{diag}(a)$ matriz diagonal de componentes (i,i)-ésima igual a a_i (ver notaciones). Entonces la matriz de covarianza es dada por $\mathrm{Cov}[X] = \mathcal{H}_z G_X(1) + \mathrm{diag}\left(\nabla_z G_X(1)\right) - \nabla_z G_X(1) \nabla_z^t G_X(1)$.

La función G_X tiene unas propiedades permitiendo por ejemplo de manejar sencillamente distribuciones de probabilidades de combinaciones lineales de vectores aleatorios independientes, como lo vamos a ver a través del teorema siguiente.

Teorema 1-26. *Sean* X *e* Y *dos vectores aleatorios* d*-dimensionales independientes,* $a = \begin{bmatrix} a_1 & \cdots & a_d \end{bmatrix}^t \in \mathbb{R}^d$ *y* $b = \begin{bmatrix} b_1 & \cdots & b_d \end{bmatrix}^t \in \mathbb{R}^d$. *Entonces para cualquier* $z = \begin{bmatrix} z_1 & \cdots & z_d \end{bmatrix} \in \mathbb{C}^d$ *(donde existen las funciones):*

$$G_{\mathrm{diag}(a)X+b}(z) = \left(\prod_{i=1}^{d} z_i^{b_i}\right) G_X\left(z_1^{a_1}, \ldots, z_d^{a_d}\right),$$

$$G_{X+Y}(z) = G_X(z)\, G_Y(z)$$

y para $z \in \mathbb{C}$

$$G_X\left(z^{a_1}, \ldots, z^{a_d}\right) = G_{a^t X}(z)$$

Demostración. El primer resultado es inmediato, escribiendo $z_i^{a_i X_i + b_i} = z_i^{b_i} \left(z_i^{a_i}\right)^{X_i}$. El secundo viene de $z_i^{X_i+Y_i} = z_i^{X_i} z^{Y_i}$ conjuntamente con el teorema 1-19 con $f(X) = \prod_{i=1}^{d} z_i^{X_i}$ y $g(Y) = \prod_{i=1}^{d} z_i^{Y_i}$. El tercer resultado es consecuencia de $\prod_{i=1}^{d} \left(z^{a_i}\right)^{X_i} = z^{\sum_{i=1}^{d} a_i X_i}$. □

---

[31] En el caso extremo, un anillo del dominio de convergencia de la serie dando G_X puede ser restricto al circulo unitario, así que no hay garantia que las derivadas en $z = 1$ existen.

[32] Recuerdense que $(x)^{\underline{n}} = \prod_{i=0}^{n-1}(x-i), \quad n > 0$ símbolo de Pochhammer, con la convención $(x)_0 = 1$; ver nota de pie **??**.

Estos resultados permiten manejar sencillamente la medida de probabilidad de combinaciones lineales de vectores aleatorios independientes y de marginales a través esta función generadora.

De la tercera identidad, se puede hacer un paso más tratando de sumas aleatorias de vectores aleatorios:

Teorema 1-27. *Sea* $\{X_n\}_{n\in\mathbb{N}}$ *una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de probabilidad)* P_X *(resp.* G_X*) y* N *una variable aletoria definida sobre* $\mathbb{N}$*, independiente de los* X_n*. Sea el vector aleatorio* $S_N = \sum_{n=0}^{N} X_n$*. Entonces*

$$G_{S_N}(z) = G_N\big(G_X(z)\big),$$

Demostración. Usando la formula de esperanza total del teorema. 1-23, se escribe

$$\begin{aligned} G_{S_N}(z) &= \mathrm{E}\left[z^{\sum_{n=0}^{N} X_n}\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[z^{\sum_{n=0}^{N} X_n}\,\middle|\, N\right]\right] \\ &= \mathrm{E}\left[G_X(z)^N\right] \end{aligned}$$

□

1.8.2 Función generadora de momentos

Como lo hemos visto, la función generadora de probabilidad permite recucperar los momentos de un vector aleatorio a través de combinaciones de sus derivadas. Con una pequaña modificación, se puede definir una función permitiendo recuperar más directamente los momentos, de manera siguiente (Feller, 1968; Johnson et al., 1997; Mukhopadhyay, 2000; Athreya & Lahiri, 2006):

Definición 1-39 (función generadora de momentos)**.** *La* función generadora de momentos *(mgf para* moment generating function *en ingles) de un vector aleatorio* d*-dimensional se define como*

$$M_X(u) = \mathrm{E}\left[e^{u^t X}\right]$$

para $u \in \mathbb{C}^d$*.*

De esta definición se nota inmediatamente que

$$M_X(u) = G_X\left(e^u\right) \qquad \text{donde} \qquad e^u = \begin{bmatrix} e^{u_1} & \cdots & e^{u_d} \end{bmatrix}^t$$

Entonces, como G_X, la generadora de los momentos caracteriza completamente el vector aleatorio X. Además, de este vínculo entre G_X y M_X, y del dominio de definición de G_X, queda claro

que M_X es definida sobre un producto cartesiano de franjas del plano complejo, $v_i \leq \Re e\{u_i\} \leq V_i$ donde $v_i \leq 0 \leq V_i$ son llamados *indices de convergencia*. En el caso de variables escalares denotando $s = -u$, esta función se interpreta como la transformada bilateral de Laplace-Stieltjes de la medida P_X. Si además P_X admite una densidad p_X, esta función se interpreta entonces como la transformada bilateral de Laplace usual de p_X. Se refiera por ejemplo a (Widder, 1946) para más detalles sobre esta transformación (y la nota de pie 90 de la sección 1.10.4 más adelante para el caso unilateral usual).

Se muestra que esta función es continua en su dominio de definición, en particular en un entorno de $u = 0$ donde queda positiva. Eso viene de la continuidad de $x \mapsto e^{u^t x}$ y de la convergencia uniforme de la integral en el dominio de definición (ver secciones anteriores y el teorema de convergencia dominada). Además, si admite un desarollo de Taylor en este punto, la generadora de los momentos permite recuperar directamente los momentos a través de derivadas, sin hacer combinaciones:

Lema 1-10. *Para cualquier* $k > 0, \quad (i_1, \ldots, i_k) \in \{1, \ldots, d\}^k$ *derivando* M_X *se proba que, cuando existen*

$$\left.\frac{\partial^k M_X}{\partial u_{i_1} \cdots \partial u_{i_k}}\right|_{u=0} = \mathrm{E}\left[\prod_{j=1}^{k} X_{i_j}\right] = m_{i_1,\ldots,i_k}[X]$$

momento de orden k *de* X.

En particular, se recuperan

- $M_X(0) = 1$, condición de normalización.
- $\nabla_u M_X(0) = \mathrm{E}[X]$ promedio,
- $\mathcal{H}_u M_X(0) = \mathrm{E}\left[XX^t\right]$, *i. e.,* $\mathrm{Cov}[X] = \mathcal{H}_u M_X(0) - \nabla_u M_X(0) \nabla_u^t M_X(0)$ matriz de covarianza.

Como la función G_X, la generadora de los momentos tiene unas propiedades similares a las de los teoremas 1-26 y 1-27:

Teorema 1-28. *Sean* X *e* Y *dos vectores aleatorios* d*-dimensionales independientes,* A *una matriz de* $\mathcal{M}_{d',d}(\mathbb{R})$ *y* $b = \begin{bmatrix} b_1 & \cdots & b_{d'} \end{bmatrix}^t \in \mathbb{R}^{d'}$*. Entonces para cualquier* $u = \begin{bmatrix} u_1 & \cdots & u_{d'} \end{bmatrix}^t \in \mathbb{C}^{d'}$ *(donde la función existe):*

$$M_{AX+b}(u) = e^{u^t b} M_X\left(A^t u\right),$$

y para cualquier $u = \begin{bmatrix} u_1 & \cdots & u_d \end{bmatrix}^t \in \mathbb{C}^d$ *(donde la función existe):*

$$M_{X+Y}(u) = M_X(u)\, M_Y(u)$$

Además, para $\{X_n\}_{n\in\mathbb{N}}$ *una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de momentos)* P_X *(resp.* M_X*)* N *una variable aleatoria definida sobre* $\mathbb{N}$*, independiente de los* X_n*, y* $S_N = \sum_{n=0}^{N} X_n$*,*

$$M_{S_N}(u) = G_N\big(M_X(u)\big),$$

Demostración. Las pruebas siguen punto a punto los mismos pasos que las de los teoremas 1-26 y 1-27. □

1.8.3 Función característica

Si la función generadora de momentos permite recuperar los momentos de un vector aleatorio, no es definida sobre todo $\mathbb{C}^d$. Sin embargo, cuando $\Re e\{u_i\} = 0$, esta función es siempre definida. Entonces, una función generadora muy útil que se usa frecuentemente es la de momentos para este tipo de argumentos, lo que es conocida como función característica y que es al final definida sobre $\mathbb{R}^d$ de manera siguiente (Lukacs, 1961; Golberg, 1961; Feller, 1968; Stein & Weiss, 1971; Johnson et al., 1997; Mukhopadhyay, 2000; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Sasvári, 2013):

Definición 1-40 (función característica)**.** *La* función característica *(cf para* characteristic function *en ingles) de un vector aleatorio d-dimensional se define como*

$$\Phi_X(\omega) = \mathrm{E}\left[e^{\imath\omega^t X}\right]$$

para $\omega \in \mathbb{R}^d$.

De esta definición se nota inmediatamente que

$$\Phi_X(\omega) = M_X(\imath\omega) = G_X\left(e^{\imath\omega}\right) \qquad \text{donde} \qquad e^{\imath\omega} = \begin{bmatrix} e^{\imath u_1} & \cdots & e^{\imath u_d} \end{bmatrix}^t$$

De hecho, se puede definir esta función para un argumento complejo, pero es equivalente a volver a la definición de la generadora de momentos.

En su forma general, la función característica se escribe

$$\Phi_X(\omega) = \int_{\mathbb{R}^d} e^{\imath\omega^t x}\, dP_X(x)$$

y es relacionada a la transformada de Fourier-Stieltjes de la medida P_X (Pinsky, 2009, Chap. 5). Cuando P_X admite una densidad p_X, la función es una transformada de Fourier usual de la densidad p_X, introducida bajo el impulso de Fourier en 1822 para estudiar la difusión del calor (Fourier, 1822).

Insistamos sobre el hecho que la importancia de esta función reside en que siempre existe y está bien definida, dado que $\int_{\mathbb{R}^d} \left|e^{\imath\omega^t x}\right| \, dP_X(x) = \int_{\mathbb{R}^d} dP_X(x) = 1$.

Como para las generadoras ya introducidas, la función característica permite recuperar directamente los momentos a través de derivadas:

Lema 1-11. *Para cualquier $k > 0, \quad (i_1, \ldots, i_k) \in \{1, \ldots, d\}^k$, derivando Φ_X se proba que, cuando existen*

$$(-\imath)^k \left.\frac{\partial^k \Phi_X}{\partial\omega_{i_1}\cdots\partial\omega_{i_k}}\right|_{\omega=0} = \mathrm{E}\left[\prod_{j=1}^{k} X_{i_j}\right] = m_{i_1,\ldots,i_k}[X]$$

momento de orden k de X.

En particular, se recuperan

- $\Phi_X(0) = 1$, condición de normalización.
- $-\imath\nabla_\omega M_X(0) = \mathrm{E}[X]$ promedio,
- $-\mathcal{H}_\omega M_X(0) = \mathrm{E}\left[XX^t\right]$, *i. e.,* $\mathrm{Cov}[X] = -\mathcal{H}_\omega M_X(0) + \nabla_\omega M_X(0)\nabla_\omega^t M_X(0)$ matriz de covarianza.

Notar que Φ_X no es siempre diferenciable en $\omega = 0$; Por ejemplo, en el caso de la distribución de Cauchy–Lorentz univariada [33] $p_X(x) = \frac{\gamma}{\pi(\gamma^2+(x-x_0)^2)}$ con $\gamma > 0$, resulta $\Phi_X(\omega) = e^{-\imath x_0\omega - \gamma|\omega|}$. Esta función está definida para todo ω, pero no es derivable en $\omega = 0$, lo que coincide con el hecho de que no están definidos los momentos mayor o igual a uno para esta densidad de probabilidad.

Resumimos algunas otras propiedades importantes de la función característica:

Teorema 1-29 (Propiedades principales de la función característica)**.**

1. *Φ_X es una función medible y continua en $\mathbb{R}^d$ (Pinsky, 2009, Prop. 5.2.1). Eso es una consecuencia del teorema de convergencia dominada (ver teorema 1-5).*
2. *$\Phi_X(0) = 1$: Eso es inmediato escribiendo la integral, siendo P_X una medida de probabilidad.*
3. *$|\Phi_X(\omega)| \leq 1 = \Phi_X(0)$: $|\Phi_X(\omega)|$ es máxima en $\omega = 0$. Eso viene directamente de $\left|e^{\imath\omega^t x}\right| = 1$. Eso significa también que*
$$\Phi_X(\mathbb{R}^r) \subset \{z \in \mathbb{C} \mid |z| \leq 1\}$$
4. *$\Phi_X(-\omega) = \Phi_X^*(\omega)$: Φ_X tiene una símetria hermitica.*
5. *$\Phi_X \in \mathfrak{P}_d$, es definida no negativa, i. e., para un conjunto arbitrario de $n \geq 1$ números complejos $a_1, \ldots, a_n$ y n vectores $w_1, \ldots, w_n$ de $\mathbb{R}^d$, se cumple*
$$\sum_{k,l=1}^{n} a_k^* a_l \Phi_X(w_l - w_k) \geq 0$$
Dicho de otra manera, la matriz de componente (k,l)-ésima $\Phi_X(w_l - w_k)$ es a hermitica (símetria hermítica dada por la propiedad anterior, y no negativa definida), i. e., matriz de $P_n(\mathbb{C})$ (ver notaciones). Esta positividad viene de $\sum_{k,l=1}^{n} a_k^* a_l e^{\imath(w_l - w_k)^t x} = \left|\sum_l a_l e^{\imath w_l^t x}\right|^2 \geq 0$.

Si una función $f : \mathbb{R}^d \mapsto \mathbb{C}$ es a simetría hermitica y definida positivas, obviamente $f(0) \in \mathbb{R}_+$ ($n = 1, a_1 = 1, x_1 = 0$). Además, tomando $a_1 = 1$, $a_2 = \pm 1$ y $w_1 = 0$, para cualquier $w_2 = w \in \mathbb{R}^d$ se obtiene $|f(w)| \leq |\Re e\{f(w)\}| \leq f(0)$: $|f|$ es máxima en 0. En lo que sigue, denotaremos el convexo (ver notaciones)

$$\mathfrak{P}_d = \left\{\Phi : \mathbb{R}^d \mapsto \mathbb{C} \text{ continuas, a simetría hermitica y definida positivas con } \Phi(0) = 1\right\}$$

[33] Lo mismo occure en la extensión multivariada (Samorodnitsky & Taqqu, 1994).

(la clase de las función continuas, a simetría hermítica y definidas positivas es definida bajo un factor real positivo). Se notará tambien, que, por proyeción del argumento de la función,

$$\mathfrak{P}_1 \supset \mathfrak{P}_2 \cdots \supset \mathfrak{P}$$

De hecho, existe una recíproca al teorema 1-29, debido a S. Bochner [34] (Bochner, 1932, 1959; Golberg, 1961; Pinsky, 2009; Sasvári, 2013)

Teorema 1-30 (Bochner). *Una función $\Phi : \mathbb{R}^d \mapsto \mathbb{C}$ es continua, definida no negativa con $\Phi(0) = 1$, i. e., $\Phi \in \mathfrak{P}_d$, si y solamente existe una medida μ de probabilidad sobre $\mathcal{B}(\mathbb{R}^d)$ tal que*

$$\forall\, \omega \in \mathbb{R}^d, \quad \Phi(\omega) = \int_{\mathbb{R}^d} e^{\imath\, \omega^t x} d\mu(x)$$

Dicho de otra manera, cualquier función continua, definida positiva con $\Phi(0) = 1$ es la función característica de un vector aleatorio.

En el teorema, vimos que la transformada de Fourier-Stieltjes de una medida de probabilidad P_X es medible, continua, definida no negativa, con $\Phi(0) = 1$. La recíproca es más dificil a probar y necesita lemas adicionales. Se puede encontrar una linda prueba en (Sasvári, 2013, Sec. 1.7) adonde dejamor el lector.

Como lo hemos notado en las subsecciones anteriores, la función característica define completamente el vector aleatorio. En particular, hay una relación uno-uno (casi siempre) entre Φ_X y la medida P_X. En particular, existe una fórmula de inversión permitiendo volver a la medida P_X a partir de Φ_X (Ash & Doléans-Dade, 1999; Sasvári, 2013):

Teorema 1-31 (Fórmula de inversión). *Sea X vector aleatorio d-dimensional de función característica Φ_X. Sea $A = \bigtimes_{i=1}^{d} (a_i\,;\, b_i) \in \mathcal{B}(\mathbb{R}^d)$ y $\partial A = \bigtimes_{i=1}^{d} [a_i\,;\, b_i] \setminus A$ su borde. Entonces [35],*

$$\lim_{w_1 \to +\infty, \ldots, w_d \to \infty} \frac{1}{(2\pi)^d} \int_{\times_{j=1}^{d} [-w_j\,;\, w_j]} \Phi_X(\omega) \prod_{j=1}^{d} \frac{e^{-\imath\, a_j \omega_j} - e^{-\imath\, b_j \omega_j}}{\imath\, \omega_j}\, d\omega$$

$$=$$

$$\int_{\mathbb{R}^d} \prod_{j=1}^{d} \left(\mathbb{1}_{(a_j\,;\, b_j)}(x_j) + \frac{1}{2} \mathbb{1}_{\{a_j\,;\, b_j\}}(x_j) \right) dP_X(x)$$

En particular, cuando P_X vale 0 sobre el borde de A, es decir $P_X(\partial A) = 0$, se obtiene

---

[34] De hecho, lo probó Bochner en el caso escalar $d = 1$, pero se extiende al caso multivariado.

[35] Se prolonga la función $\frac{e^{-\imath\, a_j \omega_j} - e^{-\imath\, b_j \omega_j}}{\imath\, \omega_j}$ en $\omega_j = 0$ por su límite $\lim_{\omega_j \to 0} \frac{e^{-\imath\, a_j \omega_j} - e^{-\imath\, b_j \omega_j}}{\imath\, \omega_j} = b_j - a_j$.

$$\lim_{w_1\to+\infty,\ldots,w_d\to\infty} \frac{1}{(2\pi)^d}\int_{\times_{j=1}^d[-w_j\,;\,w_j]} \Phi_X(\omega)\prod_{j=1}^d \frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,d\omega \;=\; P_X(A).$$

Nota: el límite $\lim_{T\to+\infty}\int_{-T}^{T}$ se nota a veces $\mathrm{vp}\int_{\mathbb{R}}$, integral *en valor principal*.

Demostración. Por definición de la función característica, tenemos

$$\int_{\times_{j=1}^d[-w_j\,;\,w_j]} \Phi_X(\omega)\prod_{j=1}^d \frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,d\omega = \int_{\times_{j=1}^d[-w_j\,;\,w_j]}\int_{\mathbb{R}^d} e^{\imath\omega^t x}\,dP_X(x)\prod_{j=1}^d \frac{e^{-\imath a_j\omega_j}-e^{-\imath b_j\omega_j}}{\imath\omega_j}\,d\omega$$

Ahora, notando que $\left|\frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,e^{\imath\,\omega^t x}\right| \le b_j-a_j$ es uniformamente acotado, se puede evocar el teorema de Fubini 1-6 para intercambiar las integrales, así que, con $e^{\imath\,\omega^t x}=\prod_{j=1}^d e^{\imath\,\omega_j x_j}$, tenemos

$$\int_{\times_{j=1}^d[-w_j\,;\,w_j]} \Phi_X(\omega)\prod_{j=1}^d \frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,d\omega = \int_{\mathbb{R}^d}\left(\prod_{j=1}^d\int_{-w_j}^{w_j} \frac{e^{-\imath\,\omega_j(a_j-x_j)}-e^{-\imath\,\omega_j(b_j-x_j)}}{\imath\,\omega_j}\,d\omega_j\right) dP_X(x)$$

Se nota que

$$\begin{aligned}\int_{-w_j}^{w_j} \frac{e^{-\imath\,\omega_j(a_j-x_j)}-e^{-\imath\,\omega_j(b_j-x_j)}}{\imath\,\omega_j}\,d\omega_j &= -\int_{-w_j}^{w_j} \frac{e^{+\imath\,\omega_j(a_j-x_j)}-e^{+\imath\,\omega_j(b_j-x_j)}}{\imath\,\omega_j}\,d\omega_j\\ &= \int_{-w_j}^{w_j} \frac{\sin(\omega_j(b_j-x_j))-\sin(\omega_j(a_j-x_j))}{\omega_j}\,d\omega_j\end{aligned}$$

por cambio de variables $\omega_j\to-\omega_j$ en la primera linea, tomando entonce la media suma de los terminos derecho/izquierdo dando la secunda linea. Seguimos notando que

$$\int_{-w}^{w}\frac{\sin(\omega(c-x))}{\omega}\,d\omega = \operatorname{sign}(c-x)\int_{-w|c-x|}^{w|c-x|}\frac{\sin(\omega)}{\omega}\,d\omega$$

es decir, de $\lim_{T\to+\infty}\int_{-T}^{T}\frac{\sin\omega}{\omega}d\omega=\pi$ (Gradshteyn & Ryzhik, 2015, Ec. 3.721), se obtiene

$$\lim_{w\to+\infty}\int_{-w}^{w}\frac{\sin(\omega(c-x))}{\omega}\,d\omega = \pi\operatorname{sign}(c-x)$$

Se acaba la prueba de $\left|\frac{\sin(\omega_j(b_j-x_j)-\sin(\omega_j(a_j-x_j))}{\omega_j}\right|<2$ conjuntamente al teorema de convergencia dominada 1-5 permitiendo permutar integral y límite, y de $\operatorname{sign}(b_j-x_i)-\operatorname{sign}(a_j-x_i)=2\,\mathbb{1}_{(a_j\,;\,b_j)}(x_j)+\mathbb{1}_{\{a_j\,;\,b_j\}}(x_j)$. □

Dos teoremas de inversión en los casos particular continuo y discreto permiten respectivamente volver a la densidad de probabilidad o a la masa de probabilidad.

Teorema 1-32 (Inversión, caso continuo)**.** *Si Φ_X es integrable, entonces P_X admite una densidad tal que*

$$p_X(x)=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\Phi(\omega)\,e^{-\imath\,\omega^t x}\,d\omega$$

Demostración. Varias pruebas existen (ej. (Sasvári, 2013, p. 21)). Una bastante directa puede ser de salir de la fórmula general del teorema 1-31, de fijar a y poner $b \equiv x \in \mathbb{R}^d$. Se toma la derivada $\frac{\partial^d}{\partial x_1 \cdots \partial x_d}$ de la integral $\frac{1}{(2\pi)^d} \int_{\times_{j=1}^d [-w_j\,;\,w_j]} \Phi_X(\omega) \prod_{j=1}^d \frac{e^{-\imath\, a_j \omega_j} - e^{-\imath\, b_j \omega_j}}{\imath\, \omega_j}\, d\omega$ y se evoca el teorema de convergencia dominada para intercambiar derivación e integración, y luego tomar el límite, para obtener el resultado. □

Teorema 1-33 (Inversión, caso discreto)**.** *Para cualquier* $x \in \mathbb{R}^d$*,*

$$\lim_{w_1 \to \infty, \ldots, w_d \to \infty} \frac{1}{2^d w_1 \ldots w_d} \int_{\times_{j=1}^d [-w_j\,;\,w_j]} \Phi(\omega)\, e^{-\imath\, \omega^t x}\, d\omega = P_X(x)$$

Demostración. Por definición de la función característica, y aplicando el teorema de Fubini como en el caso general (mismo enfoque),

$$\begin{aligned}
\frac{1}{2^d w_1 \ldots w_d} \int_{\times_{j=1}^d [-w_j\,;\,w_j]} \Phi_X(\omega) e^{-\imath\omega^t x}\, d\omega &= \frac{1}{2^d w_1 \ldots w_d} \int_{\times_{j=1}^d [-w_j\,;\,w_j]} \int_{\mathbb{R}^d} e^{\imath\omega^t y}\, dP_X(y)\, e^{-\imath\omega^t x} d\omega \\
&= \int_{\mathbb{R}^d} \left(\prod_{j=1}^d \frac{1}{2w_j} \int_{-w_j}^{w_j} e^{\imath\,(y_j - x_j) w_j} dw_j \right) dP_X(y) \\
&= \int_{\mathbb{R}^d} \prod_{j=1}^d \frac{\sin(w_j(y_j - x_j))}{(y_j - x_j) w_j} dP_X(y)
\end{aligned}$$

con el límite $\lim_{y_j \to x_j} \frac{\sin(w_j(y_j - x_j))}{w_j(y_j - x_j)} = 1$. Además, con el mismo enfoque que en el cas general acotando el integrande, por el teorema de convergencia dominada, se puede intercambiar límite e integral, así que $\lim_{x_j \to \infty} \frac{\sin(w_j(y_j - x_j))}{w_j(y_j - x_j)} = \mathbb{1}_{x_j}(y_j)$, lo cierra la prueba □

Como las funciones G_X y M_X, la función característica tiene entre otros propiedades similares a las de los teoremas 1-28 y **??**:

Teorema 1-34. *Sean* X *e* Y *dos vectores aleatorios* d*-dimensionales independientes,* A *una matriz de* $\mathcal{M}_{d',d}(\mathbb{R})$ *y* $b = \begin{bmatrix} b_1 & \cdots & b_{d'} \end{bmatrix}^t \in \mathbb{R}^{d'}$*. Entonces para cualquier* $\omega = \begin{bmatrix} \omega_1 & \cdots & \omega_{d'} \end{bmatrix}^t \in \mathbb{R}^{d'}$*:*

$$\Phi_{AX+b}(\omega) = e^{\imath\omega^t b}\, \Phi_X\left(A^t \omega\right),$$

y para cualquier $\omega = \begin{bmatrix} \omega_1 & \cdots & \omega_d \end{bmatrix}^t \in \mathbb{R}^d$*:*

$$\Phi_{X+Y}(\omega) = \Phi_X(\omega)\, \Phi_Y(\omega)$$

Además, para $\{X_n\}_{n \in \mathbb{N}}$ *una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de momentos)* P_X *(resp.* M_X*) e* N *una variable aleatoria definida sobre* $\mathbb{N}$*, independiente de los* X_n*, y* $S_N = \sum_{n=0}^N X_n$*,*

$$\Phi_{S_N}(\omega) = G_N\big(\Phi_X(\omega)\big),$$

Gracia a la función caractrística, se muestra también que para un vector aleatorio d-dimensional X, conocer la distribución de cualquier $a^t X$ para $a \in \mathbb{S}_d$ permite conocer la distribución de X (Muirhead, 1982; Bilodeau & Brenner, 1999; Sasvári, 2013):

Teorema 1-35 (Cramér-Wold)**.** *Sea X vector aleatorio d-dimensional. La distribución p_X de X es completamente determinada por el conjunto de distribuciones $\{p_{a^t X},\, \forall\, a \in \mathbb{S}_d\}$ de las variables $a^t X$.*

Demostración. Por definición de la función característica, $\forall\, \omega \in \mathbb{R}$

$$\begin{aligned} \Phi_{a^t X}(\omega) &= \mathrm{E}\left[e^{\imath\, \omega a^t X}\right] \\ &= \Phi_X(\omega a) \end{aligned}$$

Se concluye notando que $\left\{a\omega \;\middle|\; a \in \mathbb{S}_d,\, \omega \in \mathbb{R}\right\} = \mathbb{R}^d$. □

Un otro resultado interesante se víncula a la noción de mezcla de escala y toma la forma siguiente:

Teorema 1-36. *Sea X vector aleatorio de función característica Φ_X y M variable aleatoria independiente de X y de medida de probabilidad P_M. Entonces, la función característica de la mezcla de escala MX es dada por*

$$\Phi_{MX}(\omega) = \int_{\mathbb{R}} \Phi_X(m\omega)\, dP_M(m)$$

Demostración. Por definición de la función característica, la fórmula de esperanza total 1-23 y el teorema 1-24, se tiene

$$\begin{aligned} \Phi_{MX}(\omega) &= \mathrm{E}\left[e^{\imath\, \omega^t MX}\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[e^{\imath\, \omega^t MX} \,\middle|\, M\right]\right] \\ &= \int_{\mathbb{R}} \mathrm{E}\left[e^{\imath\, \omega^t mX} \,\middle|\, M = m\right] dP_M(m) \\ &= \int_{\mathbb{R}} \mathrm{E}\left[e^{\imath\, \omega^t mX}\right] dP_M(m) \\ &= \int_{\mathbb{R}} \Phi_X(m\omega)\, dP_M(m) \end{aligned}$$

□

De la función característica (o también de la generadoda de momentos) se puede probar resultados sobre la escritura estocástica de vectores aleaorios que se va a revelar frecuentement muy útil:

Teorema 1-37. *Sean dos vectores aleatorios d-dimensional X e Y de misma distribución de probabilidad. Entonces, para cualquier función $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ medible, $f(X)$ y $f(Y)$ tienen la misma distribución de probabilidad (d' puede ser menor, igual o mayor que d), i. e.,*

$$X \overset{\mathrm{d}}{=} Y \quad \Longrightarrow \quad f(X) \overset{\mathrm{d}}{=} f(Y)$$

Demostración. Este resultado se encuentra entre otros en (Fang et al., 1990, aserción 2 p. 13) o (Zolotarev, 1986) y se proba sencillamente por la función característica:

$$\Phi_{f(X)}(\omega) = \mathrm{E}\left[e^{\imath\,\omega^t f(X)}\right] = \int_{\mathbb{R}^d} e^{\imath\,\omega^t f(x)}\, dP_X(x) = \int_{\mathbb{R}^d} e^{\imath\,\omega^t f(x)}\, dP_Y(x) = \Phi_{f(Y)}(\omega)$$

Se concluye del caractér uno-uno entre la función característica y la medida de probabilidad. □

1.8.4 Función generadora de cumulantes y secunda función característica

A veces aparece más comodo trabajar con momentos especiales llamados *cumulantes*. Esos fueron introducidos por T. N. Thiele en los años 1889 (Thiele, 1889, 1903; ?, ?, ?, ?) y aparecen bajo esta denominación en un papel de R. Fisher & J. Wishart cuatro decadas después (?, ?). Estos cumulantes aparecen a través del logaritmo de la función generadora de momentos. Como lo vamos a ver en la seción 1.10.3 tratando de la familia exponencial, aparece que el logarotmo de la generadora de momentos tiene una significación física, permitiendo de calcular energias libres o internas en física estadistica a través de lo que se conoce como función de partición (ver (?, ?; Gibbs, 1902; Fisher, 1930) y más adelante sección 1.10.3). Se refiera también a (?, ?, ?, ?, ?) para lo que sigue.

Definición 1-41 (Generadora de los cumulantes)**.** *Sea X vector aleatorio y $M_X(u) = \mathrm{E}\left[e^{u^t X}\right]$ su generadora de momentos. Entonces, se define la función* generadora de los cumulantes *como*

$$C_X(u) = \log M_X(u) = \log\left(\mathrm{E}\left[e^{u^t X}\right]\right)$$

Del hecho que M_X es compleja, hace falta entender el logaritmo como el de un número complejo (?, ?; Carrier, Krook & Pearson, 2005). De la continuidad de M_X, con $M_X(0) = 1$, la generadora de los momentos es real positiva por lo menos en un enterno de $u = 0$, así que la generadora de los cumulantes va a ser definida por lo menos en un enterno de $u = 0$.

A partir de esta definición, se define los dichos *cumulantes* de la misma manera que para los momentos:

Definición 1-42 (Cumulantes)**.** *Sea X vector aleatorio d-dimensional y $C_X(u) = \log M_X(u)$ la generadora de los momentos. Esta función es definida por lo menos en un entorno de $u = 0$, $C_X(0) = 0$ y si admite en desarollo de Taylor al torno de $u = 0$, se escribe*

$$C_X(u) = \sum_{k=1}^{+\infty} \sum_{(i_1,\ldots,i_k)\in\{1,\ldots,d\}^k} \kappa_{i_1,\ldots,i_k}[X]\, \frac{u_{i_1}\ldots u_{i_k}}{k!}$$

donde el tensor $\kappa_k[X]$ de order k y de componentes $\kappa_{i_1,\ldots,i_k}[X]$ llamados cumulentes*,*

$$\kappa_{i_1,\ldots,i_k}[X] = \left.\frac{\partial^k C_X}{\partial u_{i_1}\ldots\partial u_{i_k}}\right|_{u=0}$$

es llamado cumulente de orden k de X.

Se podra notar que

- $\kappa_1[X] = \nabla C_X(0) = m_X$ media de X;
- $\kappa_2[X] = \mathcal{H}C_X(0) = \Sigma_X = \zeta_2[X]$ covarianza de X;
- $\kappa_3[X] = \zeta_3[X]$ momento centrado de orden 3;
- Pero, $\forall\, k > 3, \quad \kappa_k[X] \neq \zeta_k[X]$.

De hecho, se puede relacionar momentos y cumulantes con el enfoque de derivada de funciones compuestas (Stanley, 1999, Teo. 5.1.4) o (Hardy, 2006). Eso conduce a las relaciones siguientes (formula de Leonov y Shiryayev (Leonov & Shiryaev, 1959; Shiryayev, 1984; ?, ?; Brillinger, 2001)):

Lema 1-12 (Relación momentos-cumulantes). *Sea X vector aleatorio. Cuando existen, los momentos y cumulantes de son relacionados por las formulas*

$$\begin{cases} m_{i_1,\ldots,i_k}[X] &= \displaystyle\sum_{\pi\in\Pi_k}\prod_{B\in\pi}\kappa_{i_B}[X] \\ \kappa_{i_1,\ldots,i_k}[X] &= \displaystyle\sum_{\pi\in\Pi_k}(-1)^{|\pi|-1}\Gamma(|\pi|)\prod_{B\in\pi} m_{i_B}[X] \end{cases}$$

donde Π_k es el conjunto de las particiones [36] *de $\{1,\ldots,k\}$ y m_{j_B} denota el momento de indices $\{j_l \mid l \in B\}$.*

Además, momentos centrales y cumulantes se vinculan bajo la forma

$$\begin{cases} \zeta_{i_1,\ldots,i_k}[X] &= \displaystyle\sum_{\pi\in\Pi_k}\prod_{B\in\pi}\left(\kappa_{i_B}[X] - m_{X_{i_B}}\mathbb{1}_{\{1\}}(|B|)\right) \\ \kappa_{i_1,\ldots,i_k}[X] &= m_{X_{i_1}}\mathbb{1}_{\{1\}}(k) + \displaystyle\sum_{\pi\in\Pi_k}(-1)^{|\pi|-1}\Gamma(|\pi|)\prod_{B\in\pi}\zeta_{i_B}[X] \end{cases}$$

Demostración. Recordamosnos que para $k \in \mathbb{N}^*, \quad (i_1,\ldots,i_k) \in \{1,\ldots,d\}^k$,

$$m_{i_1,\ldots,i_k}[X] = \left.\frac{\partial^k M_X}{\partial u_{i_1}\cdots\partial u_{i_k}}\right|_{u=0} \quad \text{y} \quad \kappa_{i_1,\ldots,i_k}[X] = \left.\frac{\partial^k C_X}{\partial u_{i_1}\cdots\partial u_{i_k}}\right|_{u=0}$$

Salimos ahora de la formula de Hardy (Hardy, 2006, Prop. 1), generalizando la formula de Faà di Bruno (Faà di Bruno, 1855; Faà de Bruno, 1857) [37]: Para $h(x) = f(g(u)), \quad \forall\, n \in \mathbb{N}^*, \quad \forall\,(i_1,\ldots,i_n) \in \{1,\ldots,d\}^n$,

$$\frac{\partial^n h}{\partial u_{i_1}\cdots\partial u_{i_n}} = \sum_{\pi\in\Pi_n} f^{(|\pi|)}(g(u))\prod_{B\in\pi}\frac{\partial^{|B|}g}{\prod_{j\in B}\partial u_{i_j}}$$

---

[36] Por ejemplo, $\Pi_3 = \Big\{\{\{1,2,3\}\},\ \{\{1,2\},\{3\}\},\ \{\{1,3\},\{2\}\},\ \{\{2,3\},\{1\}\}$; cuando $\pi = \{\{1,2\},\{3\}\}$, tenemos el producot para $B = \{1,2\}$ con los indices i_1, i_2 y $B = \{3\}$ con el indice i_3.

[37] Esta formula da la derivada k-ésima de una función compuesta escalar mediante los polinimios incompletos de Bell (Bell, 1928). La formula es asociada al matemático italiano Francesco Faà di Bruno que publicó su fórmula en los años 1855, 1857. Pero fue probada desde 1800 por el matemático francés Louis François Antoine Arbogast (?, ?).

donde $f^{(l)}$ es la l-ésima derivada de f. Cuando se lo aplica a $g = C_X$ y $f(u) = \exp(u)$, dando $f^{(n)}(u) = \exp(u)$, se obtiene inmediatamente la relación dando los momentos en función de los cumulantes. Al revés, cuando se lo aplica a $g = M_X$ y $f(u) = \log u$, dando $f^{(n)}(u) = \frac{(-1)^{n-1}\Gamma(n)}{u^n}$, se obtiene inmediatamente la relación dando los cumulantes en función de los momentos.

Para lo de los momentos centrales, se nota simplemente que para $k = 1$ el cumulante y el momento coinciden, y que para $k > 1$ lo cumulantes de X y de $X - m_X$ coinciden. □

De hecho, se podriá definir así de manera *ad-hoc* los cumulantes, que las generadoras admitan o no desarollos de Taylor en $u = 0$.

Para cerrar esta sección, se puede evocar el hecho que la generadora de los momentos no esta siempre bien definida sobre todo $\mathbb{C}^d$, mientras que la función característica es siempre bien definida. A pesar de que no sea necesario (el comportamiento en $u = 0$ es importante), se usa frecuentemente el logaritmo de la función característica para definir los cumulantes.

Definición 1-43 (Secunda función característica)**.** *Sea X vector aleatorio y $\Phi_X(\omega) = \mathrm{E}\left[e^{\imath\omega^t X}\right]$ su función característica. Etonces, se define la* secunda función característica *como*

$$\Psi_X(\omega) = \log \Phi_X(\omega) = \log\left(\mathrm{E}\left[e^{\imath\omega^t X}\right]\right)$$

De nuevo del hecho que Φ_X es compleja, hace falta entender el logaritmo como el de un número complejo (?, ?; Carrier et al., 2005), y de la continuidad de Φ_X, con $\Phi_X(0) = 1$, la función característica es real positiva por lo menos en un enterno de $u = 0$, así que Ψ va a ser definida por lo menos en un entorno de $u = 0$. Si admite un desarollo de Taylor, se muestra inmediatamente que los cumulantes satisfacen la relación siguiente:

Lema 1-13 (Cumulantes a partir de la secunda función característica)**.** *Sea X vector aleatorio d-dimensional y $\Psi_X(u) = \log \Phi_X(u)$ su secunda función característica. Si admite en desarollo de Taylor, los cumulantes de orden k satisfacen*

$$\kappa_{i_1,\ldots,i_k}[X] = (-\imath)^k \left.\frac{\partial^k \Psi_X}{\partial\omega_{i_1}\ldots\partial\omega_{i_k}}\right|_{\omega=0}, \qquad (i_1,\ldots,i_k) \in \{1,\ldots,d\}^k$$

Varias propiedades de las funciones C_X y Ψ_X se deducen de las de M_X y de Φ_X. Nos enfocamos en la propiedad relacionana a combinaciones lineales de vectores independientes, con consecuencias sobre los cumulantes:

Teorema 1-38. *Sean X e Y dos vectores aleatorios d-dimensionales independientes, A una matriz de $\mathcal{M}_{d',d}(\mathbb{R})$ y $b \in \mathbb{R}^{d'}$. Entonces para cualquier $u \in \mathbb{C}^{d'}$ (donde las funciones existen):*

$$C_{AX+b}(u) = u^t b + C_X\left(A^t u\right) \qquad y \qquad \Psi_{AX+b}(\omega) = \imath\,\omega^t b + \Psi_X\left(A^t\omega\right)$$

y para cualquier $u \in \mathbb{C}^d$ o $\omega \in \mathbb{R}^d$ (donde la funciones existen):

$$C_{X+Y}(u) = C_X(u) + C_Y(u) \qquad y \qquad \Psi_{X+Y}(\omega) = \Psi_X(\omega) + \Psi_Y(\omega)$$

Las consecuencias de este sobre los cumulantes es que [38]*,*

$$\forall\, a \in \mathbb{R}, \quad \kappa_K[aX] = a^K \kappa_K[X] \qquad \textit{y} \qquad \kappa_K[X+Y] = \kappa_K[X] + \kappa_K[Y]$$

Demostración. Las relaciones tratanto de C_X y Ψ_X son consecuencias directas de los teoremas 1-28 y **??**, tomando el logaritmo de M_X o de Φ_X respectivamente. Las relaciones sobre los cumulantes es entonces consecuencias de las de C_X (o de Ψ_X) y de la definición de los cumulantes a través derivadas de C_X (o de Ψ_X). □

Se notará que, remarcablemente y contrariamente a los momentos, los cumulantes de cualquier orden son funciones lineales de variables aleatorias independientes (para los momentos, vale soló hasta el orden 3).

1.9 Vectores aleatorios complejos y matrices aleatorias en algunas palabras.

En varios campos, como en mecánica cuántica, procesamiento de señales, estadística, estamos frente a datos complejas o puestas en matrices. Aún que se puede poner en biyección $\mathbb{C}$ y $\mathbb{R}^2$ o, similarmente, $\mathcal{M}_{d,d'}(\mathbb{R})$ y $\mathbb{R}^{dd'}$, puede parecer más natural trabajar conservando la estructura de los datos. Además, permite frecuentemente usar escrituras más compactas que pasar por vectores reales.

En el contexto complejo por ejemplo, en el marco de las comunicaciones se modulan las señales en general en fase y cuadratura, es decir con ambos un seno y un coseno. Formalmente, se puede considerar modulaciones exponencial complejas, dando lugar a vectores aleatorias. Es un ejemplo entre otros que motivan el estudio de vectores aleatorios a valores complejas (Lapidoth, 2017; Schreier & Scharf, 2003; Eriksson & Koivunen, 2006; Eriksson, Ollila & Koivunen, 2009; ?, ?; Park, 2018).

Con respeto al caso de matrices, los primeros estudios de tales matrices se encuentrand entre los años 1928 y 1955 (Wishart, 1928; von Neumann & Goldstine, 1947; Wigner, 1955). Desde esta época, fueron estudiados intensivamente que sea formalmente o en el contexto de grande dimensión (Marčenko & Pastur, 1967; Boutet de Monvel, Pastur & Shcherbina, 1995). Sin embargo, trabajar con matrices, *i. e.,* conservando la estructura, implica ciertas matices en las escrituras de las distribuciones de probabilidades, momentos o funciones generadoras. Los vamos a ver brevemente, dejando el lector a libros o articulos especializados como (Akemann, Baik & Di Francesco, 2015; Gupta & Nagar, 1999; Anderson, A., Guionnet & Zeitouni, 2010; Livan, Novaes & Vivo, 2018; Edelman & Rao, 2005; Mehta, 2004; Carmeli, 1983; Dawid, 1981; Mezzadri & Snaith, 2008; Tulino & Verdu, 2004) para tener más detalles.

[38] La primera relación se generaliza sencillamente con A matricial, pero para dar expresiones compactas, se necesita introducir más profundamente calculos tensiorales, lo que va más allá del enfoque de este libro.

1.9.1 Vectores aleatorios complejos

Formalmente, un vector aleatorio complejo se define de la misma manear que en el caso real:

Definición 1-44 (Vector aleatorio complejo)**.** *Un vector aleatorio complejo es una función medible*

$$Z : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{C}^d, \mathcal{B}(\mathbb{C}^d), P_Z).$$

donde $\mathcal{B}(\mathbb{C}^d)$ son los borelianos de $\mathbb{C}^d$, σ-álgebra generada por los productos cartesianos $\big((-\infty\,;\, b_1] + \imath(-\infty\,;\, c_1]\big) \times \cdots \times \big((-\infty\,;\, b_d] + \imath(-\infty\,;\, c_d]\big)$ y donde la medida P_Z sobre $\mathcal{B}(\mathbb{C}^d)$ es la medida imágen de P. Como en el caso real,

$$(Z \in B) \equiv Z^{-1}(B) = \{\omega \in \Omega \;\big|\; Z(\omega) \in B\} \qquad \textit{y} \qquad P_Z(B) = P(Z \in B).$$

Sin embargo, se puede poner en biyección $\mathbb{C}^d$ y $\mathbb{R}^{2d}$ de tal manera de que se puede definir naturalmente un vector complejo aleatorio a partir de un vector aleatorio real de la manera alternativa equivalente (Lapidoth, 2017, Cap. 17):

Definición 1-45 (Vector aleatorio complejo – definición equivalente)**.** *Un vector aleatorio complejo se define como*

$$Z = X + \imath Y$$

donde $\widetilde{Z} \equiv \begin{bmatrix} X \\ Y \end{bmatrix}$ es un vector aleatorio de $\mathbb{R}^{2d}$. La medida de probabilidad imágen es entonces

$$P_Z \equiv P_{\widetilde{Z}} = P_{X,Y}$$

Resuelte de esta definición equivalente los hechos siguientes:

- La función de repartición de Z se escribe como la función de repartición conjunta de X e Y,

  $$F_Z \equiv F_{\widetilde{Z}} = F_{X,Y}$$

  Notando de que es una función de x e y, F_Z hace aparecer explicitamente ambos z y z^* complejos conjugados.

- Si la medida $P_{\widetilde{Z}}$ admite una derivada de Radon-Nykodým con respeto a la medida de Lebesgue sobre $\mathbb{R}^{2d}$, se define la densidad de probabilidad de Z como $f_Z \equiv f_{\widetilde{Z}} = f_{X,Y}$. A partir de la función de repartición, se escribe entonces o a través de la derivada $(2d)$-ésima de $F_{X,Y}$ con respeto a las compontentes x_i e y_i. Equivalentemente, por calculo de Wirtinger (Remmert, 1991) [39]

  $$f_Z = \imath \left(\frac{\partial^{2d} F_Z}{\partial z_1^2 \cdots \partial z_d^2} - \frac{\partial^{2d} F_Z}{\partial {z_1^*}^2 \cdots \partial {z_d^*}^2} \right)$$

[39] Hay que entender $z_i = x_i + \imath\, y_i$ y $z_i^* = x_i - \imath\, y_i$ como si fueran dos variables y, a continuación, ver las derivadas como

- Los momentos de orden k siendos definido a partir de las componentes de X y de Y, se definen también bajo la forma

$$m_{i_1,\ldots,i_l\,;\,i'_1,\ldots,i'_n}[Z] = \mathrm{E}\left[\prod_{j=1}^{l} Z_{i_j} \prod_{j=1}^{n} Z^*_{i'_j}\right] \quad \text{con} \quad l+n=k, \quad (i_1,\ldots i_l, i'_1 \ldots i'_n) \in \{1,\ldots,d\}^k$$

y similarmente para los momentos centrales $\zeta_{i_1,\ldots,i_l\,;\,i'_1,\ldots,i'_n}[Z]$ y los cumulantes a partir del logaritmo de la función característica por ejemplo (er más adelante). En particular,

  - La media de $Z = X + \imath Y$ es definida por

$$m_Z = \mathrm{E}[Z] = \mathrm{E}[X] + \imath\,\mathrm{E}[Y]$$

    La media de Z^* no lleva información más de orden 1 ($m_{Z^*} = \mathrm{E}[Z^*] = m^*_Z$).

  - La matriz de covarianza es definida por

$$\Sigma_Z \equiv \mathrm{Cov}[Z] \equiv \mathrm{E}\left[(Z-m_Z)(Z-m_Z)^\dagger\right]$$

    con $Z^\dagger = (Z^*)^t$ el transconjugado (transpuesta conjugada, ver notaciones). Fijense de que, volviendo al vector $\widetilde{Z}^t = \begin{bmatrix} X^t & Y^t \end{bmatrix}$ tenemos por un lado

$$\Sigma_{\widetilde{Z}} = \begin{bmatrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma^t_{X,Y} & \Sigma_Y \end{bmatrix}$$

    conteniendo todas las convarianzas, y por el otro lado,

$$\Sigma_Z = (\Sigma_X + \Sigma_Y) - \imath\left(\Sigma_{X,Y} - \Sigma^t_{X,Y}\right)$$

    Se puede ver que la covarianza de Z no contiene todos los terminos de orden 2. Por eso, se define también la *pseudo-covarianza*, sin terminos conjugados,

$$\check{\Sigma}_Z \equiv \mathrm{pCov}[Z] \equiv \mathrm{E}\left[(Z-m_Z)(Z-m_Z)^t\right]$$

    Ahora, se puede ver que

$$\check{\Sigma}_Z = (\Sigma_X - \Sigma_Y) + \imath\left(\Sigma_{X,Y} + \Sigma^t_{X,Y}\right)$$

    Entonces, se recupera inmediatamente Σ_X, Σ_Y y $\Sigma_{X,Y}$ a partir de Σ_Z y $\check{\Sigma}_Z$; Claramente, los momentos centrales de orden 2 son dados por ambas Σ_Z y $\check{\Sigma}_Z$.

Los momentos así definidos heriden naturalmente de las propiedades de las del caso real.

$\frac{\partial}{\partial z_i} = \frac{1}{2}\left(\frac{\partial}{\partial x_i} - \imath\,\frac{\partial}{\partial y_i}\right)$ y $\frac{\partial}{\partial z^*_i} = \frac{1}{2}\left(\frac{\partial}{\partial x_i} + \imath\,\frac{\partial}{\partial y_i}\right)$ o, al revés $\frac{\partial}{\partial x_i} = \frac{\partial}{\partial z_i} + \frac{\partial}{\partial z^*_i}$ y $\frac{\partial}{\partial y_i} = \imath\,\frac{\partial}{\partial z_i} - \imath\,\frac{\partial}{\partial z^*_i}$. El calculo diferencial así hecho es conocido como calculo de W. Wirtinger (Remmert, 1991; Wirtinger, 1927), bajo el impulso de W. Wirtinger en 1927.. De hecho, fue introducido antes , por lo menos desde H. Poincaré en los años 1899 (Poincaré, 1899).

- Se puede ver que $\Sigma_Z \in P_d(\mathbb{C})$ (hermitica semi-definida positiva), es decir que $\Sigma_Z = \Sigma_Z^\dagger$ y $\forall\, \mu \in \mathbb{C}, \quad \mu^\dagger \Sigma_Z \mu \geq 0$. Al revés, $\check{\Sigma}_Z \notin P_d(\mathbb{C})$; esta matriz es solamente simétrica $\check{\Sigma}_Z = \check{\Sigma}_Z^t \in S_d(\mathbb{C})$.

- Las generadoras son respectivamente equivalentes a las de $\widetilde{Z}$, o usando a la vez Z y Z^*. Por ejemplo, para la función característica, se la puede definir de argumento complejo como

$$\Phi_Z(\omega) = \mathrm{E}\left[e^{\imath \Re e\{\omega^\dagger Z\}}\right] \quad \text{con} \quad \omega \in \mathbb{C}^d$$

  (ver por ejemplo (Lapidoth, 2017, Cap. 17)). De nuevo se define la secunda función característica tomando el logaritmo $\Psi_X(\omega) = \log \Phi_X(\omega)$. Las funciones generadoras así definidas heriden naturalmente de las propiedades de las del caso real. Entre otros, interpretando la función característica como función de ambos ω y ω^*, y derivando como si serían variables "independientes" (ver nota de pie 39):

  - $\mathrm{E}[Z] = -2\,\imath\, \nabla_{\omega^*} \Phi_Z|_{\omega=0}$,
  - $\mathrm{Cov}[Z] = -4\, \mathcal{H}_{\omega^*,\omega} \Phi_Z|_{\omega=0} + 4\, \nabla_{\omega^*} \Phi_Z \nabla_\omega^t \Phi_Z|_{\omega=0}$,
  - $\mathrm{pCov}[Z] = -4\, \mathcal{H}_{\omega^*} \Phi_Z|_{\omega=0} + 4\, \nabla_{\omega^*} \Phi_Z \nabla_{\omega^*}^t \Phi_Z|_{\omega=0}$

  donde $\mathcal{H}_{\omega^*,\omega}$ significa que se diferencia en ω^* y luego en ω (o vice-versa).

En el marco de vectores complejos, aparece una subclase particular invariante por rotación: los vectores circulares.

Definición 1-46 (Vector aleatorio complejo circular)**.** *Un vector aleatorio complejo Z es dicho* circular *en torno*[40] *a un vector $\mu \in \mathbb{C}^d$ si para cualquier $\theta \in [0\,;\, 2\pi)$,*

$$e^{\imath\theta}\,(Z - \mu) \overset{\mathrm{d}}{=} Z - \mu$$

donde $\overset{\mathrm{d}}{=}$ significa "igualdad en distribución" (ver notaciones).

Los vectores circulares tienen propiedades particulares importantes:

- Si Z es circular en torno a un vector μ y admite una media, entonces

$$m_Z = \mathrm{E}[Z] = \mu$$

  Eso viene del hecho que $e^{\imath\theta}\,\mathrm{E}\,[Z - \mu] = \mathrm{E}\left[e^{\imath\theta}(Z - \mu)\right] = \mathrm{E}\,[Z - \mu]$. Entonces, para cualquier $\theta \in [0\,;\, 2\pi)$, $\left(1 - e^{\imath\theta}\right)\mathrm{E}\,[Z - \mu] = 0$, lo que proba que $\mathrm{E}\,[Z - \mu] = 0$.

- Si Z es circular en torno a un vector μ y admite momentos de orden 2, entonces la pseudo-covarianza es cero,

$$\check{\Sigma}_Z = \mathrm{pCov}[Z] = 0$$

[40] En la literatura, la noción de circular es dada para $\mu = 0$ (Lapidoth, 2017, Def. 24.3.2), pero se extiende sin costo adicional al caso de la definición dada en este libro.

Recordandose que $m_Z = \mu$, eso viene del hecho que $e^{2\imath\theta}\,\mathrm{E}\,[(Z-m_Z)(Z-m_Z)^t] = \mathrm{E}\left[\left(e^{\imath\theta}(Z-m_Z)\right)\left(e^{\imath\theta}(Z-m_Z)\right)^t\right] = \mathrm{E}\,[(Z-m_Z)(Z-m_Z)^t]$. Entonces, para cualquier $\theta \in [0\,;\,2\pi)$, $\left(1-e^{2\,\imath\theta}\right)\mathrm{E}\,[(Z-m_Z)(Z-m_Z)^t] = 0$, lo que cierra la prueba. La consecuencia es que en el contexto circular,

$$\Sigma_X = \Sigma_Y \quad \text{y} \quad \Sigma_{X,Y}^t = -\Sigma_{X,Y}$$

Fijense de que si la pseudo-covarianza de un vector aleatorio complejo es cero, eso no implica de que el vector es circular. Por ejemplo, sea Z tomando valores sobre $\mathcal{Z} = \{1+\imath\,;\,1-\imath\,;\,-1+\imath\,;\,-1-\imath\}$ con probabilidades $p = \left[\frac{1}{3} \quad \frac{1}{4} \quad \frac{1}{5} \quad \frac{13}{60}\right]^t$. No puede ser circular porque, por ejemplo $e^{\imath\frac{\pi}{4}}Z$ toma sus valores en $\{\sqrt{2}\,;\,-\sqrt{2}\,;\,\imath\sqrt{2}\,;\,-\imath\sqrt{2}\} \neq \mathcal{Z}$ o, por ejemplo, $e^{\imath\frac{\pi}{2}}Z$ toma sus valores en $\mathcal{Z}$ pero con el vector de probabilidad $p' = \left[\frac{1}{5} \quad \frac{1}{3} \quad \frac{13}{60} \quad \frac{1}{4}\right]^t \neq p$.

Cuando la pseudo-covarianza es cero se dice a veces que el vector es circular al orden 2. Más precisamente, en la literatura, se usa la definición siguiente (Lapidoth, 2017, Def. 17.4.1):

Definición 1-47 (Vector aleatorio complejo propio)**.** *Un vector aleatorio complejo Z es dicho* propio *(proper en ingles) si admite momentos hasta el orden 2 y ambos,*

$$\mathrm{E}[Z] = 0, \qquad \mathrm{pCov}[Z] = 0$$

Se podría ampliar esta definición hablando de vector propio en torno a un vector μ, conservando solamente la nulidad de la pseudo-covarianza. Como lo vimos tratando de vectores circulares, tenemos la implicación siguiente:

Teorema 1-39 (Circularidad)**.** *Sea Z vector alatorio complejo. Entonces,*

$$Z \text{ circular en torno a } m \qquad \Longrightarrow \qquad Z \text{ propio en torno de } m$$

Los vectores propios tienen propiedades particulares, entre otros las siguientes.

Teorema 1-40 (Conservación del carácter propio por transformación lineal)**.** *Sea Z vector aleatorio complejo propio de $\mathbb{C}^d$, entonces, para cualquier matriz $A \in \mathcal{M}_{d',d}(\mathbb{C})$, el vector aleatorio AZ es propio.*

Demostración. La prueba es obiva, notando que $\mathrm{E}[AZ] = A\,\mathrm{E}[Z] = 0$ y que $\mathrm{pCov}[AZ] = A\,\mathrm{pCov}[Z]A^t = 0$. □

Teorema 1-41 (Carácter propio y proyección)**.** *Un vector aleatorio Z complejo de $\mathbb{C}^d$ es propio si y solamente si para cualquier $c \in \mathbb{C}^d$, la variable c^tZ es propia.*

Demostración. Claramente, de $\mathrm{E}\,[c^tZ] = c^t\,\mathrm{E}[Z]$ y $\mathrm{pCov}\,[c^tZ] = c^t\,\mathrm{pCov}[Z]c$, tenemos que si Z es propio, $\mathrm{E}[Z] = 0 \Rightarrow \mathrm{E}\,[c^tZ] = 0$ y $\mathrm{pCov}[Z] = 0 \Rightarrow \mathrm{pCov}\,[c^tZ] = 0$.
Reciprocamente, si para cualquier c la variable c^tZ es propia, se puede elegir d vectores c_i^t puestas en una matriz C invertible. Entonces $\mathrm{E}[CZ] = 0$ por hypotesis, lo que da $C\,\mathrm{E}[Z] = 0 \Rightarrow \mathrm{E}[Z] = 0$ de la invertibilidad. De la misma manera, tenemos por hypothesis $\mathrm{pCov}[CZ] = 0$ lo que significa que $C\,\mathrm{pCov}[Z]C^t = 0 \Rightarrow \mathrm{pCov}[Z] = 0$ de la invertibilidad de C. □

1.9.2 Matrices aleatorias reales

De la misma manera que se puede querer trabajar con vectores complejos, a veces los datos son naturalmente puestas en matrices aleatorias. Por ejemplo, se puede querer estimar una matriz de covarianza a partir de una secuencia de vectores aleatorios $X_i,\ i = 1, \ldots, n$ de media cero y de misma ley. Un estimador natural es de reemplazar el promedio estadistico por un promedio empirico usando las observaciones/los vectores aleatorios $\widehat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^t$ (en practica, se reemplaza los X_i por observaciones/sampleos x_i y se evalua entonces un sampleo del estimador). Claramente $\widehat{\Sigma}_X$ es aleatoria por construcción, y tiene naturalmente la estructura de una matriz.

En lo que sigue, nos enfocaremos en las matrices reales. Para el caso complejo, se prodrá referirse a la subsección anterior. Notando que se puede poner en biyección $\mathcal{M}_{d,d'}(\mathbb{R})$ y $\mathbb{R}^{dd'}$, une manera de tratar de matrices aleatorias X puede ser de trabajar con su vectorización $\mathrm{vec}\,(X) = \begin{bmatrix} X_1^t & \cdots & X_{d'}^t \end{bmatrix}^t$ donde X_i es la i-ésima columna de X, las vectorizaciones de las operaciones matriciales (Magnus & Neudecker, 1979, Cap. 2) (ver también (Neudecker & Wansbeek, 1983; Harville, 2008)), y referirse a la sección tratando de vectores aleatorios. Pero resulte a veces más directo conservar la estructura matricial y trabajar con esa.

Formalmente, una matriz aleatorio real se define de la misma manear que en el caso de vectores reales, de la manera siguiente:

Definición 1-48 (Matriz aleatoria real)**.** *Una matriz aleatoria real es una función medible*

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathcal{M}_{d,d'}(\mathbb{R}), \mathcal{B}\left(\mathcal{M}_{d,d'}(\mathbb{R})\right), P_X)\,.$$

donde $\mathcal{B}\left(\mathcal{M}_{d,d'}(\mathbb{R})\right)$ son los borelianos de $\mathcal{M}_{d,d'}(\mathbb{R})$, σ-álgebra generada por los productos cartesianos $\bigtimes_{i=1,j=1}^{i=d,j=d'} (-\infty\,;\, b_{i,j}]$ y donde la medida P_X sobre $\mathcal{B}\left(\mathcal{M}_{d,d'}(\mathbb{R})\right)$ es la medida imágen de P. Nuevamente,

$$(X \in B) \equiv X^{-1}(B) = \{\omega \in \Omega \;\big|\; X(\omega) \in B\} \qquad \textit{y} \qquad P_X(B) = P(X \in B).$$

En lo que sigue, nos vamos a concentrar sobre dos situaciones particulares: (i) el caso general de matrices de $\mathcal{M}_{d',d}$ y (ii) el conjunto de matrices simétricas $S_d(\mathbb{R})$ (o un subconjunto) que tiene la particularidad de tener componentes iguales. El el último caso, a veces resuelte más comodo tener en cuenta esta simetria, es de decir que la matriz tiene a los más $\frac{d(d+1)}{2}$ componentes linealmente independientes.

Caso general De manera general, trabajamos en el contexto de matrices no necesiaramente cuadradas, y con componentes potencialemnte linealmente independientes (sin simetria particular). En este marco general, X viviendo sobre $\mathcal{X} \subseteq \mathcal{M}_{d,d'}(\mathbb{R})$:

- La función de repartición F_X es la distribución conjunta de las componentes $X_{i,j}$. Si admite una densidad, se define como $p_X = \dfrac{\partial^{dd'} F_X}{\prod_{i=1}^{d}\prod_{j=1}^{d'} \partial x_{i,j}}$.

- Los momentos y cumulantes se definen como en el caso de vectores; Por ejemplo, todos los momentos de orden k son dados por

$$m_k[X] = \mathrm{E}\left[X^{\otimes k}\right]$$

donde $\cdot^{\otimes k}$ es k veces el producto externo (ver notaciones), y similarmente para los momentos centrales $\zeta_K - k[X]$ y a través de la secunda función característica para los cumulantes $\kappa_K[X]$ (ver más adelante). En particular,

  - La media es definida por

$$m_X = \mathrm{E}[X]$$

  - La covarianza es definida por

$$\Sigma_X \equiv \mathrm{Cov}[X] = \mathrm{E}\left[(X - m_X) \otimes (X - m_X)\right] = \mathrm{E}\left[X \otimes X\right] - m_X \otimes m_X$$

Σ_X es un tensor de orden 4 de componentes $(\Sigma_X)_{i,j,k,l} = \mathrm{Cov}[X_{i,j}X_{k,l}]$..
Más generalmente, tratando de covarianza etre dos matrices aleatorias X e Y, se define el tensor covarianza conjunta como

$$\Sigma_{X,Y} \equiv \mathrm{Cov}\left[X, Y\right] : \mathrm{E}\left[(X - m_X) \otimes (Y - m_Y)\right] = \mathrm{E}\left[X \otimes Y\right] - m_X \otimes m_Y$$

- Se puede escribir también las funciones generadoras con una forma matricial; por ejemplo, tratando de la función caractéristica, va a ser una función de dd' variables que se puede poner en una matriz de $\mathcal{M}_{d,d'}(\mathbb{R})$ de tal manera que

$$\Phi_X(\omega) = \mathrm{E}\left[e^{\imath\, \mathrm{Tr}(\omega^t X)}\right] \quad \text{con} \quad \omega \in \mathcal{M}_{d,d'}(\mathbb{R})$$

De nuevo se define la secunda función característica tomando el logaritmo $\Psi_X(\omega) = \log \Phi_X(\omega)$.

- Ahora es sencillo ver de que, si existent, se puede recuperar los momentos por diferenciación como en el caso de vectores,

$$-\imath \left.\frac{\partial \Phi_X}{\partial \omega_{i,j}}\right|_{\omega=0} = \mathrm{E}[X_{i,j}]$$

o

$$-\left.\frac{\partial^2 \Phi_X}{\partial \omega_{i,j}\partial \omega_{k,l}}\right|_{\omega=0} = \mathrm{E}\left[X_{i,j}X_{k,l}\right]$$

Se prodría referirse a (Magnus & Neudecker, 1999, Cap. 8) para usar las reglas de derivación matricial para hacer los calculos en la mayoría de los casos que se encuentran en la literatura (vamos a ver ejemplos en la sección 1.10).

- Al final, se notara que si las columnas de X son independientes, se factorisa función característica a partir de las de cada columna: Sean $X = \begin{bmatrix} X_{(1)} & \cdots & X_{(d')} \end{bmatrix}$ con los $X_{(i)}$ d-dimensionales, y $\omega = \begin{bmatrix} \omega_{(1)} & \cdots & \omega_{(d')} \end{bmatrix}$ con $\omega_{(i)} \in \mathbb{R}^d$. Entonces

$$X_{(i)}, \text{ independientes} \qquad \Rightarrow \qquad \Phi_X(\omega) = \prod_{i=1}^{d'} \Phi_{X_{(i)}}(\omega_{(i)})$$

  Es inmediato de $e^{\imath \operatorname{Tr}(\omega^t X)} = e^{\imath \sum_{i=1}^{d'} \omega_{(i)}^t X_{(i)}} = \prod_{i=1}^{d'} e^{\imath \omega_{(i)}^t X_{(i)}}$ y de la independencia (ver teorema 1-19 o teorema 1-34).

Caso simétrico En el contexto simétrico, *i. e.,* el espacio de llegada es $\mathcal{X} \subseteq S_d(\mathbb{R})$ (ej. el cono $P_d^+(\mathbb{R})$), aparece que por lo más la matriz tiene $\frac{d(d+1)}{2}$ componentes linealmente independientes. A veces resuelte más comodo definir la función característica en este caso con $\omega \in S_d(\mathbb{R})$ para respectar las simetrias del problema y no tener ninguna degenerencia de esa misma (ver por ejemplo (Peddada & Richards, 1991; Anderson, 2003)). A veces, es aún dificil o imposible calcular en todo $\mathcal{M}_{d,d}(\mathbb{R})$. Eso tiene consecuencias:

- Si existen, se puede recuperar los momentos por diferenciación como en el caso de vectores, pero hay que tener en cuenta el hecho que si $i \neq j$, $\omega_{i,j} = \omega_{j,i}$ aparece dos veces en ω. Entonces, por ejemplo, se puede ver inmediatamente que

$$-\imath \left. \frac{\partial \Phi_X}{\partial \omega_{i,j}} \right|_{\omega=0} = \left(2 - \mathbb{1}_{\{i\}}(j)\right) \mathrm{E}[X_{i,j}]$$

  o que

$$- \left. \frac{\partial^2 \Phi_X}{\partial \omega_{i,j} \partial \omega_{k,l}} \right|_{\omega=0} = \left(2 - \mathbb{1}_{\{i\}}(j)\right) \left(2 - \mathbb{1}_{\{l\}}(k)\right) \mathrm{E}\left[X_{i,j} X_{k,l}\right]$$

1.10 Algunos ejemplos de distribuciones de probabilidad

En esta sección, vamos a ver unos ejemplos de distribuciones que se encuentran frecuentemente en problema prácticos de varias areas cientificas (estadística, física, ingenería,...). Daremos las caracteristicas de cada ley presentada, así que sus propiedades remarcables. El numéro de leyes de probababilidad es tan importante que es dificil, para no decir imposible, ser exahustivo. Además, existen muchas relaciones entre leyes. En esta sección, vamos a ver algunas leyes que aparecen frecuentemente, y nos enfocaremos solamente sobre algunos vínculos entre leyes (los principales). Para tener más detalles, se puede referirse a los libros especializados en este marco, como por ejemplo (Spiegel, 1976; Johnson, Kotz & Kemp, 1992; Johnson et al., 1997; Johnson, Kotz & Balakrishnan, 1995a, 1995b; Kotz, Balakrishnan & Johnson, 2000; Gupta & Nagar, 1999; Fang et al., 1990; Samorodnitsky & Taqqu, 1994).

1.10.1 Distribuciones de variable discreta

1.10.1.1. Variable real con certeza

El caso $X = a \in \mathbb{R}^d$ deterministico ($\forall\, \omega,\ X(\omega) = a$) puede ser ver visto como un caso degenerado de vector aleatorio. Visto así, sus características principales vistas en las secciones anteriores son resimidas en la tabla siguiente:

Dominio de definición	$\mathcal{X} = \{a\}, \quad a \in \mathbb{R}^d$
Distribución de probabilidad	$p_X(x) = \mathbb{1}_{\{a\}}(x)$
Promedio	$m_X = a$
Covarianza [41]	$\Sigma_X = 0$
Generadora de probabilidad	$G_X(z) = \prod_{i=1}^{d} z_i^{a_i}$ para $z_i \in \mathbb{C}$ si $a_i \geq 0$ y $\mathbb{C}^*$ si no
Generadora de momentos	$M_X(u) = e^{a^t u}$ para $u \in \mathbb{C}^d$
Función característica	$\Phi_X(\omega) = e^{\imath\, a^t \omega}$

La función de masa y función de repartición son representadas en la figura Fig. 1-14 en el caso escalar.

(a) (b)

Figura 1-14: Ilustración de una distribución cierta (a), y la función de repartición asociada (b).

[41] Siendo cero la covarianza, no se define ni la asimetría, ni la curtosis. Sin embargo, de una manera se puede decir que la ley no es asimétrica, y con cola livianas (no hay colas).

Notar que todo se extiende al caso complejo sin costo adicional.

Poner acá la ley de los gran números? Más notas historicas.

1.10.1.2. Ley uniforme sobre un "intervalo" de $\mathbb{Z}$

Se denota $X \sim \mathcal{U}\{a\,;\,b\}$ con $(a,b) \in \mathbb{Z}^2,\ b \geq a$. Las características de X son las siguientes:

Parámetros	$(a,b) \in \mathbb{Z}^2,\ b \geq a$
Dominio de definición	$\mathcal{X} = \{a\,;\,a+1\,;\,\ldots\,;\,b\}$
Distribución de probabilidad	$p_X(x) = \frac{1}{b-a+1}$
Promedio	$m_X = \dfrac{a+b}{2}$
Varianza	$\sigma_X^2 = \dfrac{(b-a)(b-a+2)}{12}$
Asimetría	$\gamma_X = 0$ para $b \neq a$ (ver más adelante)
Curtosis por exceso	$\bar{\kappa}_X = -\dfrac{6}{5}\dfrac{(b-a)(b-a+2)+2}{(b-a)(b-a+2)}$ para $b \neq a$ (ver más adelante)
Generadora de probabilidad	$G_X(z) = \dfrac{z^a - z^{b+1}}{1-z}$ para [42] $z \in \mathbb{C}$ si $a \geq 0$ y $\mathbb{C}^*$ sino
Generadora de momentos	$M_X(u) = \dfrac{e^{au} - e^{(b+1)u}}{1-e^u}$ para [43] $u \in \mathbb{C}$
Función característica	$\Phi_X(\omega) = \dfrac{e^{\imath a\omega} - e^{\imath(b+1)\omega}}{1-e^{\imath\omega}}$

La distribución de masa de probabilidad y función de repartición de una variable uniforme $\mathcal{U}\{a\,;\,b\}$ son representadas en la figura Fig. 1-15.

Cuando $b = a$, la variable tiende a una variable cierta $X = a$. En este caso, siendo la varianza cero, no se puede definir ni asimetría (pero no hay asimetría), ni curtosis (pero la ley no tiene colas, *i. e.,* colas livianas), como lo hemos visto en la sección anterior.

La distribución uniforme aparece por ejemplo en el tiro de un dado equilibriado con $a = 1,\ b = 6$.

Se notará que esta ley da un ejemplo más en lo cual la media no es necesariamente en $\mathcal{X}$. Tipicamente, en el ejemplo del dado, la media es $m_X = \frac{7}{2} \notin \{1\,;\,\ldots\,;\,6\}$.

---

[42] En el caso límite $z \to 1$, $\lim_{z\to 1} \frac{z^a - z^{b+1}}{1-z} = b+1-a$

[43] En el caso límite $u \to 0$, $\lim_{u\to 0} \frac{e^{au} - e^{(b+1)u}}{1-e^u} = b+1-a$, y similarmente para la función característica.

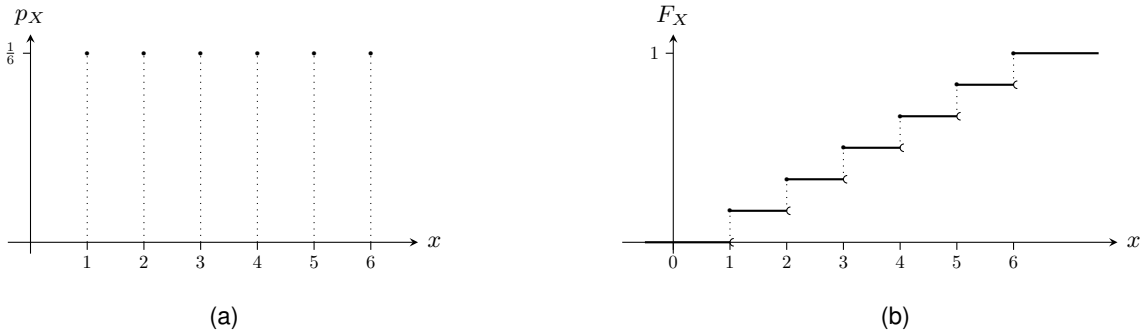


Figura 1-15: Ilustración de una densidad de probabilidad uniforme (a), y la función de repartición asociada (b). $a = 1,\ b = 6$ (ej. dado equilibriado).

1.10.1.3. Ley de Bernoulli

Esta ley aparece cuando se hace una experiencia con dos estados posibles, tipo un tiro de moneda. Apareció en trabajos muy antiguos, entre otros el de J. Bernoulli tratando de la ley de gran números (Bernoulli, 1713; Hald, 1990; David & Edwards, 2001).

Se denota $X \sim \mathcal{B}(p)$ con $p \in [0\,;\,1]$ y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \{0\,;\,1\}$
Parámetro	$p \in [0\,;\,1]$
Distribución de probabilidad	$p_X(1) = 1 - p_X(0) = p$
Promedio	$m_X = p$
Varianza	$\sigma_X^2 = p\,(1-p)$
Asimetría	$\gamma_X = \dfrac{1-2\,p}{\sqrt{p\,(1-p)}}$ para $p \notin \{0\,;\,1\}$ (ver más adelante)
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{1-6\,p+6\,p^2}{p\,(1-p)}$ para $p \notin \{0\,;\,1\}$ (ver más adelante)
Generadora de probabilidad	$G_X(z) = 1-p+pz$ sobre $\mathbb{C}$
Generadora de momentos	$M_X(u) = 1-p+p\,e^u$ sobre $\mathbb{C}$
Función característica	$\Phi_X(\omega) = 1-p+p\,e^{\imath\omega}$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-16.

Notar que cuando $p = 0$ (resp. $p = 1$) la variable es cierta $X = 0$ (resp. $X = 1$). En estos casos, nuevamente, siendo la varianza cero, no se puede definir ni asimetría (pero no hay asimetría), ni curtosis

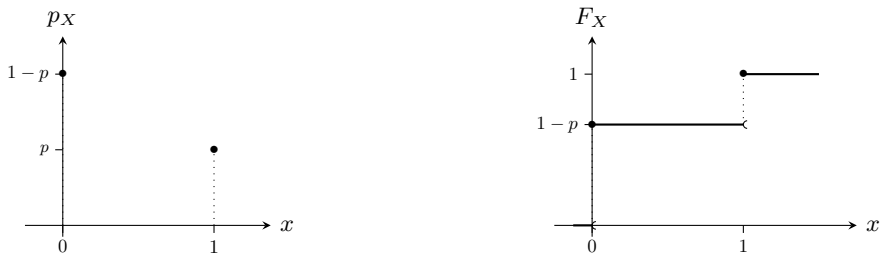


Figura 1-16: Ilustración de una distribución de probabilidad de Bernoulli (a), y la función de repartición asociada (b), con $p = \frac{1}{3}$.

(pero la ley no tiene colas, *i. e.,* colas livianas), como ya lo hemos visto anterioramente.

Se notará también que la ley de Bernoulli tiene una propiedad de reflexividad trivial:

Lema 1-14 (Reflexividad)**.** *Sea* $X \sim \mathcal{B}(p)$. *Entonces*

$$1 - X \sim \mathcal{B}(1-p)$$

Demostración. El resultado es inmediato de $P(1 - X = 1) = P(X = 0) = 1 - p$. □

1.10.1.4. Ley binomial

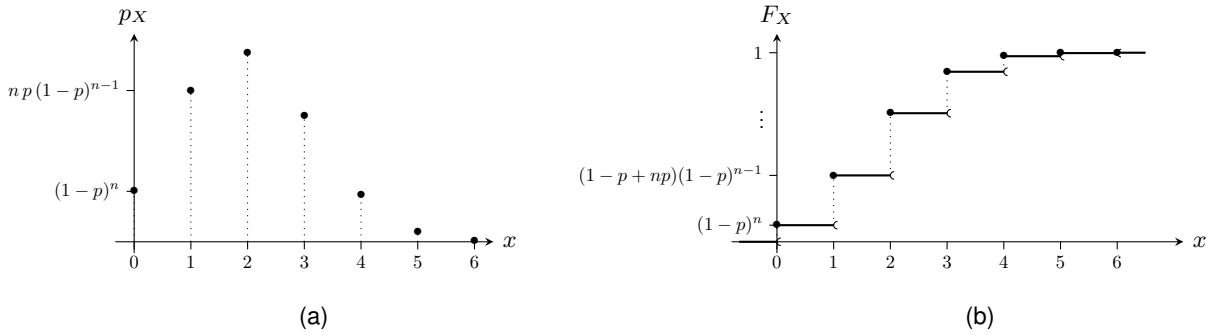
Esta ley apareció en trabajos muy antiguos, de nuevo y naturalmente, entre otros, en de J. Bernoulli en 1713 (Bernoulli, 1713; Hald, 1990; David & Edwards, 2001). Se la puede ver como una extension de la ley de Bernoulli a $n \geq 1$ tiros de una moneda, contando por ejemplo cuantas veces aparecen una cara.

Se denota $X \sim \mathcal{B}(n, p)$ con $n \in \mathbb{N}^*, \quad p \in [0\,;\,1]$ y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \{0\,;\,\ldots\,;\,n\}$
Parámetros	$n \in \mathbb{N}^*, \quad p \in [0\,;\,1]$
Distribución de probabilidad	$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$
Promedio	$m_X = np$
Varianza	$\sigma_X^2 = np(1-p)$
Asimetría	$\gamma_X = \dfrac{1-2p}{\sqrt{np\,(1-p)}}$ para $p \notin \{0\,;\,1\}$ (ver más adelante)
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{1-6\,p+6\,p^2}{n\,p\,(1-p)}$ para $p \notin \{0\,;\,1\}$ (ver más adelante)

Generadora de probabilidad	$G_X(z) = (1-p+pz)^n$ sobre $\mathbb{C}$
Generadora de momentos	$M_X(u) = (1-p+p\,e^u)^n$ sobre $\mathbb{C}$
Función característica	$\Phi_X(\omega) = (1-p+p\,e^{\imath\omega})^n$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-17.



(a) (b)

Figura 1-17: Ilustración de una distribución de probabilidad binomial (a), y la función de repartición asociada (b), con $n = 6, \quad p = \frac{1}{3}$.

Otros ilustraciones para otros p?

Cuando $n = 1$, se recupera la lei de Bernoulli $\mathcal{B}(p) \equiv \mathcal{B}(1,p)$. Adémas, se muestra sencillamente usando la generadora de probabilidad que

Lema 1-15. *Sean* $X_i \sim \mathcal{B}(p), \quad i = 1,\dots,n$ *independientes, entonces*

$$\sum_{i=1}^{n} X_i \sim \mathcal{B}(n,p)$$

De este resultado, se puede notar que, por ejemplo, le distribución binomial aparece en el conteo de eventos independientes de misma probabilidad entre n.

También, la ley binomial tiene una propiedad de reflexividad, consecuencia directa de la de Bernoulli:

Lema 1-16 (Reflexividad)**.** *Sea* $X \sim \mathcal{B}(n,p)$. *Entonces*

$$n - X \sim \mathcal{B}(n, 1-p)$$

Demostración. El resultado es inmediato de la propiedad de reflexividad de la ley de Bernoulli, conjuntamente al lema 1-15. Alternativamente, se nota que $P(n - X = x) = P(X = n - x) = \binom{n}{n-x} p^{n-x}(1-p)^x = \binom{n}{x}(1-p)^x p^{n-x}$ notando que $\binom{n}{n-x} = \binom{n}{x}$. □

Si tomamos el ejemplo de una moneda que se tira n veces de maneras independientes, con probabilidad p que aparezca una cara, X representa el número de caras tiradas. Entonces, $n - X$ es el número de secas: en $n - X$ se intercambian los roles de la cara y de la seca.

Nota que cuando $p = 0$ (resp. $p = 1$) la variable es cierta $X = 0$ (resp. $X = n$). En estos casos, nuevamente, siendo la varianza cero, no se puede definir ni asimetría (pero no hay asimetría), ni curtosis (pero la ley no tiene colas, *i. e.,* colas livianas), como ya lo hemos visto anterioramente.

1.10.1.5. Ley binomial negativa

Formas particulares de esta ley aparecieron en los años 1679 en trabajos de Blaise Pascal (Pascal, 1679; Hald, 1990; David & Edwards, 2001) [44] y un poco más tarde de P. R. de Montmort (de Montmort, 1713, p. 233-248). Esta ley aparece cuando se repite una experencia binaria éxito/fracasco con probabilidad p de éxito, de manera independiente hasta el r-ésimo fracasco (r fijo), contendo el número de excitos obtenidos cuando se para la experiencia.

Se denota $X \sim \mathcal{B}_-(r,p)$ con $r \in \mathbb{N}^*, \quad p \in [0\,;1)$ y sus características son las siguientes:

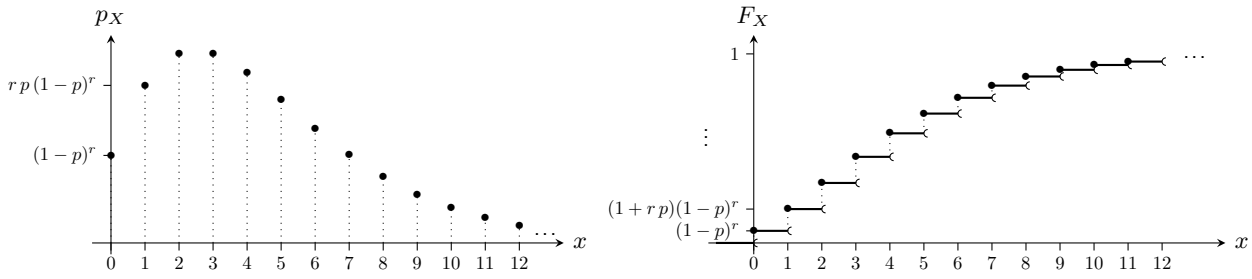
Dominio de definición	$\mathcal{X} = \mathbb{N}$
Parámetros	$r \in \mathbb{N}^*, \quad p \in [0\,;1)$
Distribución de probabilidad	$p_X(x) = \binom{x+r-1}{x} p^x (1-p)^r$
Promedio	$m_X = \dfrac{rp}{1-p}$
Varianza	$\sigma_X^2 = \dfrac{rp}{(1-p)^2}$
Asimetría	$\gamma_X = \dfrac{1+p}{\sqrt{rp}}$ para $p \neq 0$ (ver más adelante)
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{1+4p+p^2}{rp}$ para $p \neq 0$ (ver más adelante)
Generadora de probabilidad	$G_X(z) = \left(\dfrac{1-p}{1-pz}\right)^r$ para $\lvert z\rvert < p^{-1}$
Generadora de momentos	$M_X(u) = \left(\dfrac{1-p}{1-p\,e^u}\right)^r$ para $\Re e\,\{u\} < -\ln p$
Función característica	$\Phi_X(\omega) = \left(\dfrac{1-p}{1-p\,e^{i\omega}}\right)^r$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-18.

Otras ilustraciones para otros r, p?

Recordar que esta ley aparece cuando se repite una experencia binaria $X_i \in \{0\,;1\}, i = 1, \ldots$ con $P(X_i = 1) = p$ de manera independiente (X_i independientes) hasta que r variables valen 0, con

[44] De hecho aparece en una carta de 1654 que mandó B. Pascal a P. de Fermat, publicada mucho tiempo después.



(a) (b)

Figura 1-18: Ilustración de una distribución de probabilidad binomial negativa (a), y la función de repartición asociada (b), con $r = 3, \quad p = \frac{3}{5}$.

r fijo. El número de excito X sigue una ley $\mathcal{B}_-(r,p)$ (el calculo es directo). Dicho de otra manera, $X = \sum_{i=1}^{N} X_i$ con N variable aleatoria tal que $X_N = 0$ y $r = \sum_{i=1}^{N}(1 - X_i)$: condicionalmente a N, la variable X es binomial de parámetro p, *i. e.,* $P(X = x|N = n) = \binom{n}{x} p^x (1-p)^{n-x}$. Se puede ver que $P(N = n) = \binom{n}{r-1}(1-p)^r p^{n-r}$ y la ley de la binomial negativa se recupera a través del teorema de probabilidad total 1-1 o también, a través del teorema 1-27.

Esta distribución se generaliza para $r \in \mathbb{R}_+^*$ pero se pierde la interpretación que vímos en el párafo anterior.

Nota: cuando $p = 0$ la variable es cierta $X = r$. De nuevo, siendo la varianza cero en este caso, no se puede definir ni asimetría (pero no hay asimetría), ni curtosis (pero la ley no tiene colas, *i. e.,* colas livianas).

1.10.1.6. Ley multinomial

Esta ley es una generalización de la ley binomial y aparece por ejemplo cuando se repite una experiencia a k estados n veces de manera independiente y nos interesamos a la probabilidad que el primer evento aparece n_1 veces, el secundo n_2 veces, ...(ej. para $k = 6$, contamos los números de 1, de 2, ...cuando tiramos n veces este dado). Apareció también esta ley por la primera vez en el trabajo de J. Bernoulli (Bernoulli, 1713; Hald, 1990; David & Edwards, 2001) (ver también el ensayo de Montmort de 1708 con otras notaciones (de Montmort, 1713)).

Se denota $X \sim \mathcal{M}(n,p)$ con $n \in \mathbb{N}^*$ y $p = \begin{bmatrix} p_1 & \cdots & p_k \end{bmatrix}^t \in \Delta_{k-1}$ the $(k-1)$-simplex estandar (ver figura 1-30-(a) más adelante, y notaciones). Entonces, a pesar de que se escribe X de manera k-dimensional, el vector partenece a un espacio claramente $d = k - 1$ dimensional y en el caso $k = 2$ se recupera la ley binomial. El dominio de definición es claramente $\mathrm{P}_n k$ (ver notaciones). Las características de $X \sim \mathcal{M}(n,p)$ son las siguientes:

Dominio de definición	$\mathcal{X} = \mathrm{P}_n k$
Parámetros	$n \in \mathbb{N}^*, \quad p \in \Delta_{k-1}$
Distribución de probabilidad	$p_X(x) = \dfrac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i}$
Promedio	$m_X = np$
Covarianza	$\Sigma_X = n\left(\mathrm{diag}(p) - pp^t\right)$
Generadora de probabilidad	$G_X(z) = \left(p^t z\right)^n$ para $z \in \mathbb{C}^k$
Generadora de momentos	$M_X(u) = \left(p^t e^u\right)^n$, $e^u = \begin{bmatrix} e^{u_1} & \cdots & e^{u_k} \end{bmatrix}^t$ para $u \in \mathbb{C}^k$
Función característica	$\Phi_X(\omega) = \left(p^t e^{\imath\omega}\right)^n$

De hecho, se puede considerar que el vector aleatorio es $(k-1)$-dimensional $\widetilde{X} = \begin{bmatrix} \widetilde{X}_1 & \cdots & \widetilde{X}_{k-1} \end{bmatrix}^t$ definido sobre el dominio $\widetilde{\mathcal{X}} = \left\{x \in \{0\,;\,\ldots\,;\,n\}^{k-1},\ \sum_{i=1}^{k-1} x_i \leq n\right\}$. El parámetro de $\widetilde{X}$ es entonces $\widetilde{p} = \begin{bmatrix} p_1 & \cdots & p_{k-1} \end{bmatrix}^t \in \left\{q \in [0\,;\,1]^{k-1} \;\middle|\; \sum_{i=1}^{k-1} q_i \leq 1\right\}$. La masa de probabilidad de $\widetilde{X}$ se escribe obviamente $p_{\widetilde{X}}(x) = \frac{n!}{\prod_{i=1}^{k-1} x_i!(n-\sum_{i=1}^{k-1} x_i)!} \prod_{i=1}^{k-1} p_i^{x_i} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{n-\sum_{i=1}^{k-1} x_i}$. Se notará al final que $G_{\widetilde{X}}(\widetilde{z}) = G_X\left(\begin{bmatrix} \widetilde{z} & 1 \end{bmatrix}^t\right)$ y al revés $G_X(z) = z_k^n\, G_{\widetilde{X}}\left(\begin{bmatrix} \frac{z_1}{z_k} & \cdots & \frac{z_{k-1}}{z_k} \end{bmatrix}^t\right)$. Similarmente, $M_{\widetilde{X}}(\widetilde{u}) = M_X\left(\begin{bmatrix} \widetilde{u} & 0 \end{bmatrix}^t\right)$ y $M_X(u) = e^{n\,u_k} M_{\widetilde{X}}\left(\begin{bmatrix} u_1 - u_k & \cdots & u_{k-1} - u_k \end{bmatrix}^t\right)$ (y similarmente para Φ_X y $\Phi_{\widetilde{X}}$).

Se puede ver también que $\Sigma_X \mathbb{1} = 0$ así que $\Sigma_X \notin P_k^+(\mathbb{R})$. Eso es la consecuencia directa del hecho de que X k-dimensional, vive sobre Δ_{k-1}, $(k-1)$-dimensional. Aparentemente, siendo Σ_X no invertible, no se puede definir ni asimetría, ni curtosis. Sin embargo, habría que considerar $\widetilde{X}$, de promedio $n\widetilde{p}$ y de covarianza el bloque $(k-1)\times(k-1)$ de Σ_X, que es ahora invertible. $\gamma_{\widetilde{X}}$ y $\kappa_{\widetilde{X}}$ son bien definidos. Las expresiones, demasiado pesadas, no son dadas acá.

Deos ejemplos de masa de probabilidad de esta ley son representadas en la figura Fig. 1-19.

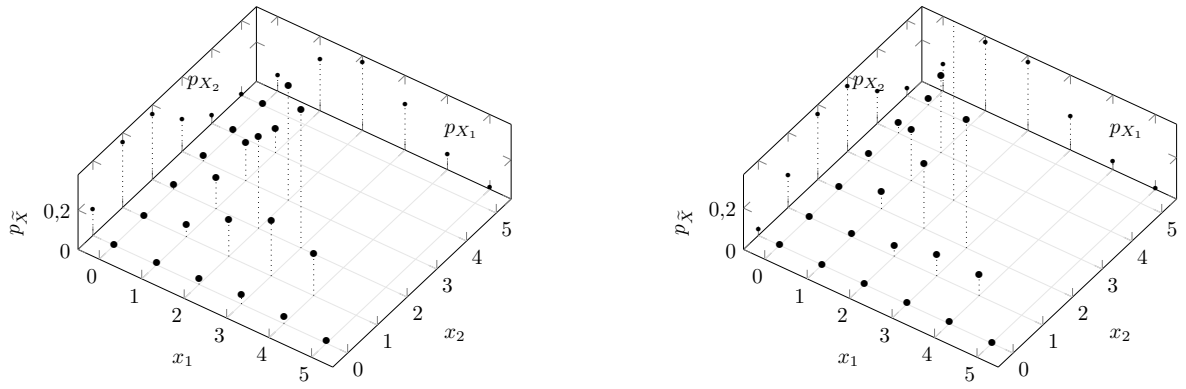
Notar: cuando $p = \mathbb{1}_i$, la variable es cierta $X = n\mathbb{1}_i$.

Otros ilustraciones para otros n, p?

Vectores de distribución multinomial tienen una propiedade notable con respecto a una permutación de variable, parecidas a la de la binomial:

Lema 1-17 (Efecto de una permutación)**.** *Sea* $X \sim \mathcal{M}(n,p)$, $p \in \Delta_{k-1}$ *y* $\Pi \in \mathfrak{S}_k(\mathbb{R})$ *matriz de permutación (ver notaciones). Entonces*

$$\Pi X \sim \mathcal{M}(n, \Pi p)$$



(a) (b)

Figura 1-19: Ilustración de una distribución de probabilidad multinomial para $k = 3$ del vector $(k-1)$-dimensional $\widetilde{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^t$ $(X_3 = 1 - X_1 - X_2)$ con las marginales p_{X_1}, p_{X_2}. Es dibujada solamente la distribución sobre $\widetilde{\mathcal{X}}$, siendo esta nula afuera de $\widetilde{\mathcal{X}}$. Los parámetros son $n = 5$ y $p = \begin{bmatrix} \frac{2}{5} & \frac{1}{3} & \frac{4}{15} \end{bmatrix}^t$ (a), $p = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{bmatrix}^t$ (b).

Demostración. El resultado es inmediato saliendo de la función característica y aplicando el teorema 1-34 (recordar que $\Pi^{-1} = \Pi^t$). Más directamente, notando la permutation σ tal que $\Pi = \sum_{i=1}^k \mathbb{1}_i \mathbb{1}_{\sigma(i)}^t$, se puede ver que $P(\Pi X = x) = P(X = \Pi^{-1}x) = \dfrac{n!}{\prod_{i=1}^k x_{\sigma^{-1}(i)}!} \prod_{i=1}^k p_i^{x_{\sigma^{-1}(i)}} = \dfrac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_{\sigma(i)}^{x_i}$ por cambio de indices. □

Además, la ley multinomial exhibe una stabilidad reemplazando dos componentes por su suma:

Lema 1-18 (Stabilidad por agregación)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_k \end{bmatrix}^t \sim \mathcal{M}(n, p)$, $p \in \Delta_{k-1}$ *y* $G^{(i,j)}$ *matriz de agrupación de las* (i, j)*-ésima componentes (ver notaciones). Entonces,*

$$G^{(i,j)}X \sim \mathcal{M}\left(n, G^{(i,j)}p\right)$$

Este resultado es intuitivo del hecho que vuelve a agrupar los estados i e j en un estado, que tiene entonces la probabilidad $p_i + p_j$ de aparecer.

Demostración. Suponemos $i < j$ (el otro caso se recupera por simetría). A partir de la función característica y el teorema 1-34 se tiene,

$$\begin{aligned} \forall\, \omega \in \mathbb{R}^{k-1}, \quad \Phi_{G^{(i,j)}X}(\omega) &= \Phi_X\left(G^{(i,j)\,t}\omega\right) \\ &= \left(\sum_{l=1}^k p_l\, e^{\imath \left(G^{(i,j)\,t}\omega\right)_l}\right)^n \end{aligned}$$

Ahora, se nota que ${G^{(i,j)}}^t\omega = \begin{bmatrix}\omega_1 & \cdots & \omega_{j-1} & \omega_i & \omega_{j+1} & \cdots & \omega_{k-1}\end{bmatrix}^t$, entonces

$$\begin{aligned}\forall\, \omega \in \mathbb{R}^{k-1}, \quad \Phi_{G^{(i,j)}X}(\omega) &= \left(\sum_{l=1, l\neq j}^{k} p_l\, e^{\imath\, \omega_l} + p_j\, e^{\imath\, \omega_i}\right)^n \\ &= \left(\sum_{l=1, l\neq i, l\neq j}^{k} p_l\, e^{\imath\, \omega_l} + (p_i + p_j)\, e^{\imath\, \omega_i}\right)^n\end{aligned}$$

lo que cierra la prueba. Se puede tener un enfoque más directo, con los mismos pasos que en la prueba del lema 1-23 tratando de la ley hipergeometrica multivaluada. □

De este lema, aplicado de manera recursiva, se obtiene los corolarios siguientes:

Corolario 1-8. *Sea* $X \sim \mathcal{M}(n, p)$*, entonces* $X_i \sim \mathcal{B}(n, p_i)$.

Al final, por una análisis combinatorial, se muestra sencillamente un resultado similar al de la binomial como suma de Bernoulli independientes:

Lema 1-19. *Sean* $U_i, \quad i = 1, \ldots, n$ *discretas sobre* $\mathcal{U} = \{1\,;\,\ldots\,;\,k\}$ *de masa de probabilidad* $p_{U_i} = p \in \Delta_{k-1}$*, independientes, y* $X_i = \mathbb{1}_{U_i}$ *vectores aleatorios* k*-dimensionales (son, por construcción, independientes). Entonces*

$$\sum_{i=1}^{n} X_i \sim \mathcal{M}(n, p)$$

Nota: esta ley se generaliza de la misma manera que para la binomial negativa, dando una ley multinomial negativa o, de manera equivalente, generalizando la binomial negativa a más de dos clases se obtiene la ley multinomial negativa. **Anadirlo en una seccion?**

1.10.1.7. Ley hipergeométrica

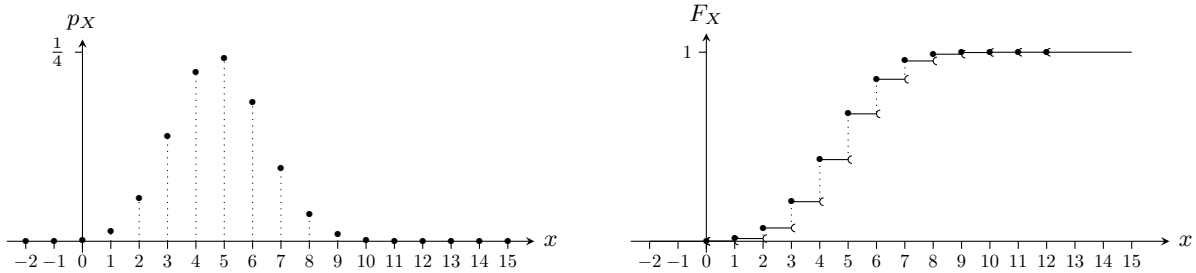
Esta ley aparece por ejemplo cuando se hace una experiencia con una población de tamaño n (ej. n bolas en una urna), que pueden partenecer a dos clases, con k número de elementos de la primera clase (a veces dicho estados de éxito; ej. k bolas negras), $n - k$ número de elementos de la secunda clase, y se hace m tiros sin reemplazamiento. X es el número de tiros parteneciendo en la primera clase (número de éxitos). Esta ley apareció en trabajos de de Moivre en 1710 (de Moivre, 1710; Hald, 1990; David & Edwards, 2001).

Se denota $X \sim \mathcal{H}(n, k, m)$ con $n \in \mathbb{N}^*, \quad k \in \{0\,;\,\ldots\,;\,n\}, \quad m \in \{0\,;\,\ldots\,;\,m\}$ y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \{\text{máx}(0, k + m - n)\,;\,\ldots\,;\,\text{mín}(k, m)\}$

Parámetros	$n \in \mathbb{N}^*$ (población) $k \in \{0\,;\,\ldots\,;\,n\}$ (número de estados exitosos) $m \in \{0\,;\,\ldots\,;\,n\}$ (número de tiros)
Distribución de probabilidad	$p_X(x) = \frac{\binom{k}{x}\binom{n-k}{m-x}}{\binom{n}{m}}$
Promedio	$m_X = \frac{m}{n}\,k$
Varianza [45]	$\sigma_X^2 = \begin{cases} \frac{m\,(n-m)}{n^2(n-1)}\,k\,(n-k) & \text{si} \quad n > 1 \\ 0 & \text{si} \quad n = 1 \end{cases}$
Generadora de probabilidad	$G_X(z) = \frac{\binom{n-k}{m}}{\binom{n}{m}}\ {}_2F_1(-m,-k;n-m-k+1;z)$ sobre $\mathbb{C}$
Generadora de momentos	$M_X(u) = \frac{\binom{n-k}{m}}{\binom{n}{k}}\ {}_2F_1\left(-m,-k;n-m-k+1;e^u\right)$ sobre $\mathbb{C}$
Función característica	$\Phi_X(\omega) = \frac{\binom{n-k}{m}}{\binom{n}{m}}\ {}_2F_1\left(-m,-k;n-m-k+1;e^{\imath\,\omega}\right)$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-20.



(a) (b)

Figura 1-20: Ilustración de una distribución de probabilidad hipergeométrica (a), y la función de repartición asociada (b), con $n = 100, \quad k = 12, \quad m = 40$.

Otros ilustraciones para otros n, k, m?

Poner la asimetría (ya lo tengo)? El Curtosis (lo tengo que simplificar)? muy pesadas... Momento factorial $f_q = \frac{(m)^{\underline{q}}(k)^{\underline{q}}}{(n)^{\underline{q}}}$ permitiendo calcular todo.

Notar: la variable resuelta cierta en los casos siguientes

- $m = 0 \;\Rightarrow\; X = 0$: no se sortean elementos, así que siempre se sortea 0 elementos de la primera clase;

[45] En el caso degenerado $n = 1$, o $m = 0$, o $m = 1 = n$; en ambos casos, la variable es cierta (ver fin de la subsección).

- $m = n \Rightarrow X = k$: si se sortean todos los elementos de la población, se sortean todos los k de la primera clase;
- $k = 0 \Rightarrow X = 0$: si la primera clase no tiene elementos, no se puede tirar elementos de esta clase;
- $k = n \Rightarrow X = m$: al revés si la secunda clase no tiene elementos, todos los sorteados partenecen a la primera clase.
- Para $n = 1$, necesariamente aparece unas de las cuatros situaciones anteriores (de hecho, dos a la vez).

En cada de estos casos, la varianza vale cero así que no se puede definir ni γ_X, ni κ_X. A fuera de estos casos, las expresiones de estas cantidades son dejados al lector como exercicio.

La ley tiene propiedades de reflexividad del mismo tipo que para la ley binomial:

Lema 1-20 (Reflexividad)**.** *Sea* $X \sim \mathcal{H}(n, k, m)$. *Entonces*

$$m - X \sim \mathcal{H}(n, n-k, m) \qquad \textit{y} \qquad k - X \sim \mathcal{H}(n, k, n-m)$$

Demostración. El primer resultado es inmediato de $P(m - X = x) = P(X = m - x) = \frac{\binom{k}{m-x}\binom{n-k}{x}}{\binom{n}{m}}$. El secundo de $P(k - X = x) = P(X = k - x) = \frac{\binom{k}{k-x}\binom{n-k}{m-k+x}}{\binom{n}{m}} = \frac{\binom{k}{x}\binom{n-k}{n-m-x}}{\binom{n}{n-m}}$ notando que $\binom{a}{b} = \binom{a}{a-b}$. □

Se puede ver que si en una urna con bolas negras y blancas, con k bolas negras, y X es el número de bolas negras sorteadas, $m - X$ representa las bolas blancas sorteadas. Es decir que en $m - X$ se intercambia los roles de las bolas negras y blancas. De la misma manera, $k - X$ representa las bolas negras que quedan en la urna, entre las $n - m$ que quedan, es decir que en $k - X$ se intercambia los roles de las bolas sorteadas y las que quedan en la urna.

1.10.1.8. Ley hipergeométrica negativa

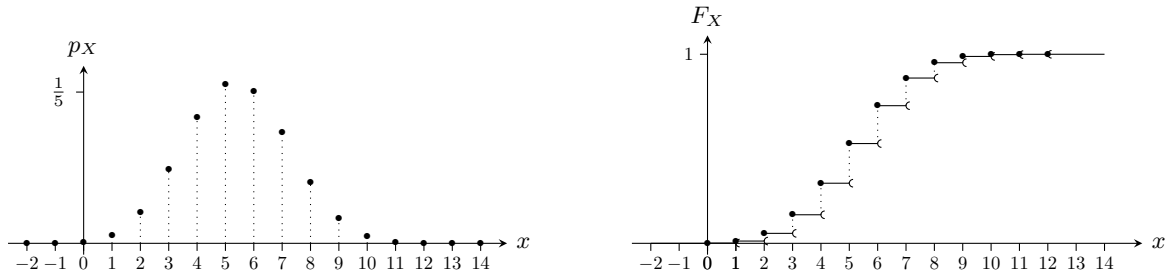
Parece que se encuentran las primeras huellas de esta ley en trabajos del marquesano de Condorcet en 1785 (Marquis de Condorcet, 1785). Esta ley aparece por ejemplo cuando se hace una experiencia del mismo tipo que para la hipergeométrica, con una población de tamaño n (ej. n bolas en una urna), que pueden partenecer a dos clases, con k numéro de elementos de la primera clase, dichos estados de éxito (ej. k bolas negras), $n - k$ numéro de elementos de la secunda clase, dichos estados de fracasco. Pero en lugar de hacer m tiros fijos, se hace tiros hasta que r elementos de la secunda clase (fracascos) sean tiradas. X es el número de tiros parteneciendo en la primera clase (número de éxitos). Es decir que cuando $X = x$, tenemos k elementos de la primera clase en los "primeros" $x + r - 1$ tiros, el últimos parteneciendo a la seconda clase.

Se denota $X \sim \mathcal{H}_-(n, k, r)$ con $n \in \mathbb{N}^*$, $k \in \{0; \ldots; n\}$, $m \in \{0; \ldots; n-k\}$ y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \{0\,;\,\ldots\,;\,k\}$
Parámetros	$n \in \mathbb{N}^*$ (población) $k \in \{0\,;\,\ldots\,;\,n\}$ (número de estados exitosos) $r \in \{0\,;\,\ldots\,;\,n-k\}$ (número de fracascos para parar)
Distribución de probabilidad [46]	$p_X(x) = \begin{cases} \frac{\binom{x+r-1}{x}\binom{n-r-x}{k-x}}{\binom{n}{k}} & \text{si} \quad r > 0 \\ \mathbb{1}_{\{0\}}(x) & \text{si} \quad r = 0 \end{cases}$
Promedio	$m_X = \dfrac{rk}{n-k+1}$
Varianza	$\sigma_X^2 = \dfrac{rk(n+1)(n-k-r+1)}{(n-k+1)^2(n-k+2)}$
Generadora de probabilidad	$G_X(z) = \frac{\binom{n-r}{k}}{\binom{n}{k}}\ {}_2F_1(r,-k;r-n;z)$ sobre $\mathbb{C}$
Generadora de momentos	$M_X(u) = \frac{\binom{n-r}{k}}{\binom{n}{k}}\ {}_2F_1\left(r,-k;r-n;e^u\right)$ sobre $\mathbb{C}$
Función característica	$\Phi_X(\omega) = \frac{\binom{n-r}{k}}{\binom{n}{k}}\ {}_2F_1\left(r,-k;r-n;e^{\imath\,\omega}\right)$

Poner asimetría y curtosis? Expresiones muy pesadas... Momento factorial $f_q = \frac{(r)^{\overline{q}}(k)^{\underline{q}}}{(n-k+1)^{\overline{q}}}$ permitiendo calcular todo.

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-21.



(a) (b)

Figura 1-21: Ilustración de una distribución de probabilidad hipergeométrica negativa (a), y la función de repartición asociada (b), con $n = 100, \quad k = 12, \quad r = 40$.

[46] Para los $x + r - 1$ primeros tiros, de la primera clase hay $\binom{k}{x}$ combinaciones posibles, y $\binom{n-k}{r-1}$ de la seconda clase, sobre los $\binom{n}{x+r-1}$ combinaciones posibles en total. Para el último tiro, quedan $n - k - (r-1)$ posibilidades de la seconda clase sobre los $n - x - (r-1)$ elementos que quedan.

Otros ilustraciones para otros n, k, r?

Notar: cuando $k = 0$, la variable es cierta $X = r$ (se sortean solamente elementos de la seconda clase, así que para siempre cuando se han tirados r elementos); cuando $r = 0$, también la variable es cierta $X = 0$ (no se sortan bolas, así que no hay de la primera clase).

1.10.1.9. Ley hipergeometrica multivariada

Esta ley aparece por ejemplo cuando se generaliza la ley hipergeométrica con $c > 2$ clases con k_i numéro de elementos de la clase i, $\sum_i k_i = n$. Se estudia esta ley, entre otros, por la primera vez, en el ensayo de Montmort en 1708 (de Montmort, 1713), o más tarde, en 1740, en trabajos de Simpson (Simpson, 1740; Hald, 1990; David & Edwards, 2001).

Se denota $X \sim \mathcal{HM}(n, k, m)$ con $n \in \mathbb{N}$, $k = \begin{bmatrix} k_1 & \cdots & k_c \end{bmatrix}^t \in \mathrm{P}_n c$ (ver notaciones) $m \in \{0\,;\,\ldots\,;\,n\}$.

Entonces, como en el caso de la ley multinomial, a pesar de que se escribe X de manera c-dimensional, el vector partenece a un espacio claramente $(c-1)$-dimensional. Notar que en el caso $c = 2$ se recupera la ley hipergeométrica.

Sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \left\{ x \in \bigtimes_{i=1}^{c} \{0\,;\,\ldots\,;\,k_i\} \;\middle|\; \sum_{i=1}^{c} x_i = m \right\}$
Parámetros	$n \in \mathbb{N}^*$ (población) $c \in \mathbb{N}^*$ (número de clases) $k \in \mathrm{P}_{n,c}$ (números de elementos de cada clase) $m \in \{0\,;\,\ldots\,;\,n\}$ (número de tiros)
Distribución de probabilidad	$p_X(x) = \dfrac{\prod_{i=1}^{c} \binom{k_i}{x_i}}{\binom{n}{m}}$
Promedio	$m_X = \dfrac{m}{n}\, k$
Covarianza	$\Sigma_X = \begin{cases} \frac{m(n-m)}{n^2(n-1)} \left(n \operatorname{diag}(k) - k k^t\right) & \text{si} \quad n > 1 \\ 0 & \text{si} \quad n = 1 \end{cases}$

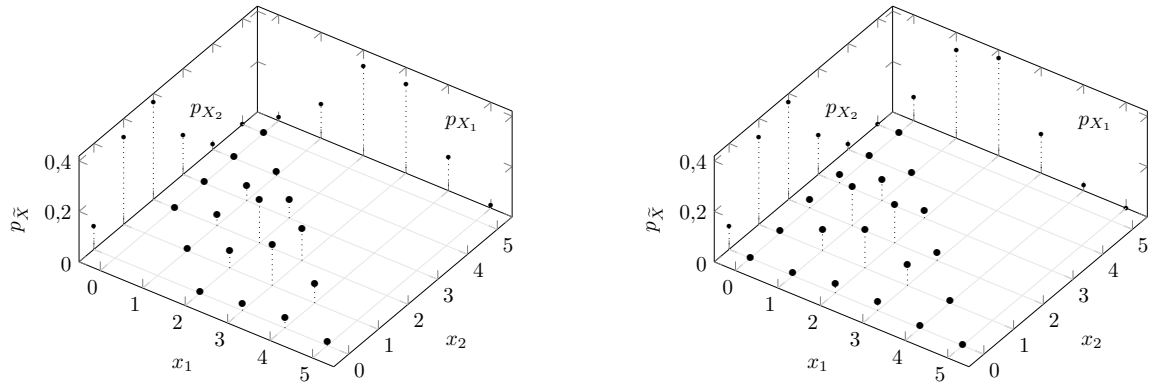
De hecho, se puede considerar que el vector aleatorio es $(c-1)$-dimensional $\widetilde{X} = \begin{bmatrix} \widetilde{X}_1 & \cdots & \widetilde{X}_{c-1} \end{bmatrix}^t$ definido sobre el dominio $\widetilde{\mathcal{X}} = \left\{ \widetilde{x} \in \bigtimes_{i=1}^{c-1} \{0\,;\,\ldots\,;\,k_i\} \;\middle|\; \text{máx}\left(0\,,\, \sum_{i=1}^{c-1} k_i + m - n\right) \leq \sum_{i=1}^{c-1} \widetilde{x}_i \leq m \right\}$. Los parámetros de $\widetilde{X}$ son entonces $n \in \mathbb{N}^*$, $m \in \{0\,;\,\ldots\,;\,n\}$ y $\widetilde{k} = \begin{bmatrix} k_1 & \cdots & k_{c-1} \end{bmatrix}^t \in \left\{ q \in \{0\,;\,\ldots\,;\,n\}^{c-1} \;\middle|\; \sum_{i=1}^{c-1} q_i \leq n \right\}$.

A continuación, la masa de probabilidad de $\widetilde{X}$ es naturalmente $p_{\widetilde{X}}(x) = \dfrac{\prod_{i=1}^{c-1} \binom{k_i}{x_i} \binom{n-\sum_{i=1}^{c-1} k_i}{m-\sum_{i=1}^{c-1} x_i}}{\binom{n}{m}}$.

Similarmente al caso multinomial, se puede ver que $\Sigma_X \mathbb{1} = 0$ así que $\Sigma_X \notin P_k^+(\mathbb{R})$. De nuevo, es la consecuencia directa del hecho de que X c-dimensional, vive sobre una variedad $(c-1)$-dimensional. Aparentemente, siendo Σ_X no invertible, no se puede definir ni asimetría, ni curtosis. Sin embargo, habría para esta ley también que considerar $\widetilde{X}$, de promedio $\frac{m}{n}\widetilde{k}$ y de covarianza el bloque $(c-1)\times(c-1)$ de Σ_X, que es ahora invertible. $\gamma_{\widetilde{X}}$ y $\kappa_{\widetilde{X}}$ son bien definidos. Las expresiones, demasiado pesadas, no son dadas acá (se deja al lector como exercicio).

Ver si se calcula Phi

La masa de probabilidad es representada en la figura Fig. 1-22.



(a) (b)

Figura 1-22: Ilustración de una distribución de probabilidad hipergeométrica multivariada para $c = 3$ del vector $(c-1)$-dimensional $\widetilde{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^t$ $(X_3 = m - X_1 - X_2)$ con las marginales p_{X_1}, p_{X_2}. Es dibujada solamente la distribución sobre $\widetilde{\mathcal{X}}$, siendo esta cero afuera de $\widetilde{\mathcal{X}}$. Los parámetros son $n = 18$, $m = 5$, $k = \begin{bmatrix} 9 & 6 & 3 \end{bmatrix}^t$ (a), $k = \begin{bmatrix} 6 & 6 & 6 \end{bmatrix}^t$ (b).

Notar: cuando $c = 2$ se recupera la ley hipergeométrica; además X resuelta cierta en los casos siguientes (ver subesección anterior para las explicaciones/ilustraciones):

- $m = 0 \Rightarrow X = 0$;
- $m = n \Rightarrow X = k$;
- $k = n\mathbb{1}_i \Rightarrow X = m\mathbb{1}_i$.

Vectores de distribución hipergeométricas multivaluada tienen propiedades notables similares a las de la hipergeométricas y de la multinomial, a saber de tipo reflexividad, con respecto a una permutación de variable y con respecto a una agregación.

Lema 1-21 (Reflexividad)**.** *Sea* $X \sim \mathcal{HM}(n,k,m)$*. Entonces*

$$k - X \sim \mathcal{HM}(n,k,n-m)$$

Demostración. Sea $Y = k - X$. De $P(Y=y) = P(k-X=y) = P(X=k-y) = \frac{\prod_{i=1}^{c}\binom{k_i}{k_i-y_i}}{\binom{n}{m}} = \frac{\prod_{i=1}^{c}\binom{k_i}{y_i}}{\binom{n}{n-m}}$ notando que $\binom{a}{b} = \binom{a}{a-b}$. Se cierra la prueba recordandose que $\sum_{i=1}^{c} k_i = n$ y $\sum_{i=1}^{c} x_i = m$, dando $\sum_{i=1}^{c} k_i = n$ y $\sum_{i=1}^{c} y_i = n-m$. □

Como el en contexto escalar, si en una urna tenemos bolas de c colores diferentes, con un número k_i para el i-ésimo color, X_i es el numéro de este color que se sorteó y $k_i - X_i$ representan las de este color que quedan en la urna, entre las $n - m = \sum_{i=1}^{c}(k_i - X_i)$ que quedan, es decir que en $k - X$ se intercambia los roles de las bolas sorteadas y las que quedan en la urna.

Lema 1-22 (Efecto de una permutación)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_c \end{bmatrix}^t \sim \mathcal{HM}(n,k,m)$ *y* $\Pi \in \mathfrak{S}_c(\mathbb{R})$ *matriz de permutación. Entonces*

$$\Pi X \sim \mathcal{HM}(n, \Pi k, m)$$

Demostración. La prueba sigue paso paso la de la multinomial. Notando la permutation σ tal que $\Pi = \sum_{i=1}^{c} \mathbb{1}_i \mathbb{1}_{\sigma(i)}^t$, se puede ver que $P(\Pi X = x) = P(X = \Pi^{-1}x) = \dfrac{\prod_{i=1}^{c}\binom{k_i}{x_{\sigma^{-1}(i)}}}{\binom{n}{m}} = \dfrac{\prod_{i=1}^{c}\binom{k_{\sigma(i)}}{x_i}}{\binom{n}{m}}$

por cambio de indices. □

Lema 1-23 (Stabilidad por agregación)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_c \end{bmatrix}^t \sim \mathcal{HM}(n,k,m)$ *y* $G^{(i,j)}$ *matriz de agrupación de las* (i,j)*-ésima componentes (ver notaciones). Entonces,*

$$G^{(i,j)}X \sim \mathcal{HM}\left(n, G^{(i,j)}k, m\right)$$

Este resultado es intuitivo del hecho que vuelve a agrupar las clases i e j en una clase, que tiene entonces $k_i + k_j$ elementos.

Demostración. Del lema precediente, notando que existen matrices de permutación [47] $\Pi_k \in \mathfrak{S}_k(\mathbb{R})$ y $\Pi_{k-1} \in \mathfrak{S}_{k-1}(\mathbb{R})$ tal que $G^{(i,j)} = \Pi_{k-1}\, G^{(c-1,c)}\, \Pi_k$, se puede concentrarse en el caso $(i,j) = (c-1,c)$.

[47] Π_k pone las componentes i e j el las posiciones $c-1$ y c, sin cambiar el orden de las precedientes; Π_{k-1} trazlada la última componente en la posición $\min(i,j)$.

Ahora, claramente,

$$\begin{aligned} P(G^{(c-1,c)}X = x) &= P\left(\bigcap_{i=1}^{c-2}(X_i = x_i) \cap (X_{c-1} + X_c = x_{c-1})\right) \\ &= \sum_{t=0}^{x_{c-1}} P\left(\bigcap_{i=1}^{c-2}(X_i = x_i) \cap (X_{c-1} = t) \cap (X_c = t - x_{c-1})\right) \\ &= \frac{\prod_{i=1}^{c-2}\binom{k_i}{x_i}}{\binom{n}{m}} \sum_{t=0}^{x_{c-1}} \binom{k_{c-1}}{t}\binom{k_c}{x_c - t} \end{aligned}$$

Se cierra la prueba de la identidad de Chu-Vandermonde [48] $\sum_{t=0}^{l}\binom{r}{t}\binom{s}{l-t} = \binom{r+s}{l}$ (Knuth, 1997, Ec. (21), p. 59) o (Gradshteyn & Ryzhik, 2015, Ec. 0.156). □

De este lema, aplicado de manera recursiva, se obtiene el corolario siguiente:

Corolario 1-9. *Sea* $X \sim \mathcal{HM}(n, k, m)$*, entonces* $X_i \sim \mathcal{H}(n, k_i, m)$.

Nota: esta ley se generaliza de la misma manera que para la hipergeométrica negativa, dando una ley hipergeométrica negativa multivariada o, de manera equivalente, generalizando la hipergeométrica negativa a más de dos clases se obtiene la ley hipergeométrica negativa. **Anadirlo en una seccion?**

1.10.1.10. Ley geométrica

La ley geométrica es un caso particular de la ley binomial negativa para $r = 1$, como ya lo hemos evocado. Dicho de otra manera, esta distribución aparece en el conteo de une repetición de una experiencia de manera independiente hasta que occure un evento de probabilidad p; por ejemplo el número de tiro de un dado equilibriado hasta que occurre un "6" sigue una ley geométrica de parámetro $p = \frac{1}{6}$.

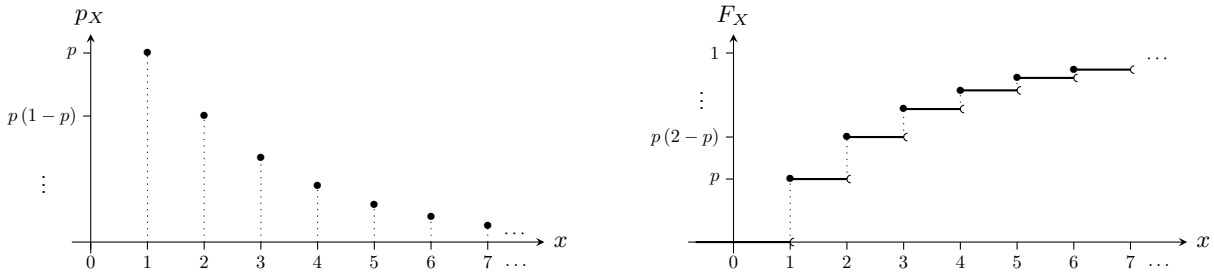
Se denota $X \sim \mathcal{G}(p)$ con $p \in (0\,;\,1]$ y sus características son las siguientes:

[48] Este identidad es debido a A.-T. Vandermonde en 1772, pero esta conocida desde 1303 por el matemático chino Chu Shi-Chieh, explicando la denominación de este identidad (Andersen & Larsen, 1994) o (Askey, 1975, p. 59-60). Se prueba escribiendo $(1+x)^{r+s} = (1+x)^r(1+x)^s$ y desallorando con la fórmula del binomio cada potencia.

Dominio de definición	$\mathcal{X} = \mathbb{N}^*$
Parámetro	$p \in (0\,;\,1]$
Distribución de probabilidad	$p_X(x) = (1-p)^{x-1}p$ (convención $0^0 = 1$)
Promedio	$m_X = \frac{1}{p}$
Varianza	$\sigma_X^2 = \dfrac{1-p}{p^2}$
Asimetría	$\gamma_X = \dfrac{2-p}{\sqrt{1-p}}$ para $p \neq 1$

Curtosis por exceso	$\bar{\kappa}_X = \dfrac{6-6\,p+p^2}{1-p}$ para $p \neq 1$
Generadora de probabilidad	$G_X(z) = \dfrac{pz}{1-(1-p)z}$ para $\lvert z\rvert < \frac{1}{1-p}$
Generadora de momentos	$M_X(u) = \dfrac{p\,e^u}{1-(1-p)\,e^u}$ para $\Re e\,\{u\} < -\ln(1-p)$
Función característica	$\Phi_X(\omega) = \dfrac{p\,e^{\imath\omega}}{1-(1-p)\,e^{\imath\omega}}$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-23. **Otros**



(a) (a)

Figura 1-23: Ilustración de una distribución de probabilidad geométrica (a), y la función de repartición asociada (b), con $p = \frac{1}{3}$.

ilustraciones para otros p?

Como ya lo hemos evocado, esta ley esta vínculada con la binomial negativa, siendo un caso particular:

Lema 1-24 (Vínculo con la ley binomial negativa)**.** *Sea* $X \sim \mathcal{B}_-(1,p)$*, entonces tenemos también*

$$X \sim \mathcal{G}(1-p).$$

Si volvemos a la representation de $X \sim \mathcal{B}_-(1,p)$ como $X = \sum_{i=1}^{N} X_i$ con $X_i \sim \mathcal{B}(p)$ independientes, N tal que $X_N = 0$ y $\sum_{i=1}^{N}(1-X_i) = 1$, aparece que, también, $N \sim \mathcal{G}(1-p)$.

Notar que cuando $p = 1$ la variable es cierta $X = 1$. De nuevo, de la nulidad de la varianza, no se puede definir ni asimetría, ni curtosis (ver observaciones en ejemplos anteriores).

1.10.1.11. Ley de Poisson

Esta ley fue introducida por Poisson en 1837 como caso límite de la ley binomial para n grande, con el producto np fijo (Poisson, 1837, Cap. 3), (Hald, 1990; David & Edwards, 2001). Se interesó Poisson en su estudio al comportamentio probabilístico del conteo de experiencia de Bernoulli bajo la hipotesis de independencia (dando lugar a la ley binomial) en ciencia humana, para una población importante (n grande), pero con un valor promedio dado. De hecho, se conocía esta ley, también como caso límite

de la binomial, por lo menos desde un trabajo de de Moivre unas decadas antes (de Moivre, 1710). Apareció también más tarde en muchos procesos físicos, como el conteo de desintegración atomica por secundo en un material radioactivo, o, (aproximadamente) a través del conteo de partículas que caen en una pequeña superficia, cuanto se tiran partículas uniformamente en una grande superficia en trabajos de W. S. Gosset [49] (Student, 1907).

Se denota $X \sim \mathcal{P}(\lambda)$ con $\lambda \in \mathbb{R}_+^*$ llamada *taza*, y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \mathbb{N}$
Parámetro	$\lambda \in \mathbb{R}_+^*$
Distribución de probabilidad	$p_X(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$
Promedio	$m_X = \lambda$
Varianza	$\sigma_X^2 = \lambda$
Asimetría	$\gamma_X = \dfrac{1}{\sqrt{\lambda}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{1}{\lambda}$
Generadora de probabilidad	$G_X(z) = e^{\lambda(z-1)}$ para $z \in \mathbb{C}$
Generadora de momentos	$M_X(u) = e^{\lambda(e^u-1)}$ para $u \in \mathbb{C}$
Función característica	$\Phi_X(\omega) = e^{\lambda\,(e^{\imath\omega}-1)}$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-24.

Otras ilustraciones para otros λ?

Adémas, se muestra sencillamente usando la generadora de probabilidad que

Lema 1-25 (Stabilidad). *Sean* $X_i \sim \mathcal{P}(\lambda_i), \quad i = 1, \ldots, n$ *independientes, entonces*

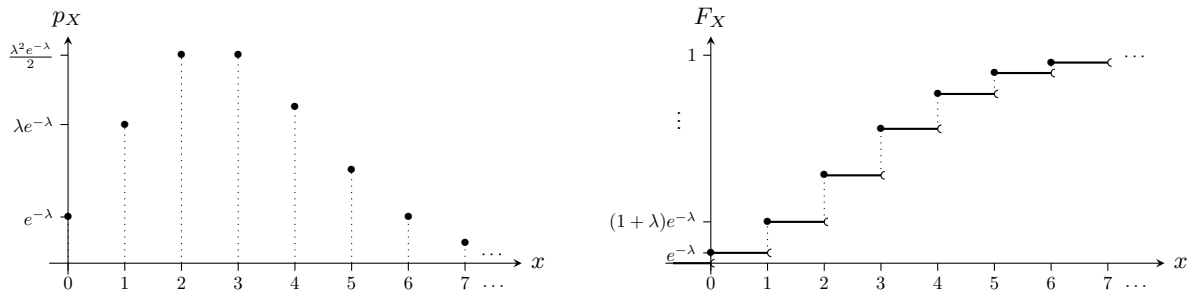
$$\sum_{i=1}^{n} X_i \sim \mathcal{P}\left(\sum_{i=1}^{n} \lambda_i\right)$$

Como lo hemos introducido, la ley de Poisson esta vínculada a la ley binomial, como caso límite:

Lema 1-26 (Vínculo con la ley binomial). *Sean* $X_n \sim \mathcal{B}\left(n\,,\,\frac{\lambda}{n}\right)$ *con* $\lambda > 0$ *fijo, entonces*

$$X_n \xrightarrow[n\to\infty]{\mathrm{d}} X \sim \mathcal{P}(\lambda)$$

[49] Fue connocido bajo en nombre “Student”; ver nota de pie 70.



(a) (b)

Figura 1-24: Ilustración de una distribución de probabilidad de Poisson (a), y la función de repartición asociada (b), con $\lambda = 3$.

donde $\xrightarrow{\mathrm{d}}$ *significa que el límite es en distribución (ver notaciones).*

Demostración. Se sale de la forma de la distribución binomial y de la formula de Stirling [50]: $\log\Gamma(z) = \left(z-\frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + o(1)$ en $z \to +\infty$ (Stirling, 1730; Abramowitz & Stegun, 1970; Gradshteyn & Ryzhik, 2015). □

Aparece que la ley de Poisson esta vínculada también a la ley binomial negativa, también como caso límite:

Lema 1-27 (Vínculo con la binomial negativa)**.** *Sean* $X_r \sim \mathcal{B}_-\left(\frac{\lambda}{r+\lambda}, r\right)$ *con* $\lambda > 0$ *fijo, entonces*

$$X_r \xrightarrow[r\to\infty]{\mathrm{d}} X \sim \mathcal{P}(\lambda)$$

Demostración. Se sale de nuevo la forma de la distribución binomial negativa y de la formula de Stirling para probarlo. □

Más allá del contexto discreto, esta ley esta también vínculada a ley exponencial, por el processo dicho de Poisson. Si eventos pueden aparecer de manera aleatoria en el tiempo tal que, entre dos eventos, el tiempo sigue una ley exponencial de parámetro λ, y que estos tiempos son independientes, entonces dado un intervalo T de tiempo, el número de estos eventos sigue una ley de Poisson de parámetro λT. Lo vamos a ver en el ejemplo de la ley exponencial más adelante.

Al final, notar que cuando $\lambda = 0$ la variable es cierta $X = 0$ (usando la convención $0^0 = 1$).

1.10.1.12. Distribución seria de potencia (power series distributions)?

[50] De hecho, esta formula es probablemente debida previamente a A. De Moivre (de Moivre, 1733, 1756; Pearson, 1924; Le Cam, 1986; Dutka, 1991; Deming, 1933), y fue mejorada por Stirling más tarde. Fue mejorada aún más por el famoso matemático S. Ramanujan recientemente (Andrew & Berndt, 2013, § 4.1).

Estadística de los números de ocupación de niveles energéticos: distribuciones de Maxwell–Boltzmann, de Fermi–Dirac, y de Bose–Einstein (Rényi, 2007, p. 37-38)
Leyes de los grandes números; DeMoivre-Laplace; F? inverse gamma, Rayleigh (Gamma), Rice, chi cuadrado no central?

Everett "The Cambridge Dictionary of Statistics", Cambridge Univ Press, 2006 (3rd Ed.); Hazeinke Michel Ed. (2001) "Probability Distributions", Springer

1.10.2 Distribuciones de variable continua

Antes de ir más adelante, notamos que, tratando de un vector aleatorio X continuo, de densidad de probabilidad $p_X(x)$, para cualquier $a \in \mathbb{R}^*$ el vector $Y_a = aX$ va a ser obviamente continuo, de densidad de probabilidad $p_{Y_a}(y) = \frac{1}{|a|} p_X\left(\frac{y}{a}\right)$. Ahora, cuando $a \to 0$, queda claro que el vector Y_a tiende al vector 0, deterministico. O, de un punto de vista de variable aleatoria, Y_a tiende al vector cierto, discreto. Un punto que puede parecer sopredente es que tal vector no admite una densidad de probabilidad más. De hecho la densidad p_{Y_a} tiende a una función generalizada, o distribución de Schwarz, más precisamente la "función Dirac", como lo hemos visto en el fin de la sección 1.3.4. Como lo hemos enfatizado, en tal caso preferimos seguir trabajando con la medida de probabilidad, que tiende a la medida de Dirac.

1.10.2.1. Distribución uniforme sobre un intervalo

Esta distribución es la más natural que se usa cuando queremos modelar una falta de información sobre una variable, sabiendo que vive en un espacio de volumen finito: sin a priori más, una tendencia natural/intuitiva es de asignar la "misma probabilidad" a cada punto del conjunto. En particular, aparece así naturalmente en la inferencia bayesiana que consiste a modelar como aleatorio un parámetro que se quierre inferir [51] (Robert, 2007) (la ley es dicha ley *a priori*; ver también (Bayes, 1763) o (Laplace, 1812, 1814; de Laplace, 1820); tal a priori es conocido como a priori de Laplace).

Se denota $X \sim \mathcal{U}([a\,;\,b])$. Las características de X son las siguientes:

---

[51] En la inferencia bayesiana, nos interesamos al paremetro (posiblemente multivariado) θ subyancente a una distribución. Por ejemplo, sabemos tener observaciones sorteados de una distribución de Poisson, pero con el parámetro λ desconocido y nos interesamos a $\theta \equiv \lambda$. El enfoque bayesiano consiste a considerar el parámetro Θ aleatorio, tal que la distribución de las observaciones sea vista como distribución condicional $p_{X|\Theta=\theta}(x)$, llamada distribución de sampleo. Dadas las observaciones $X = x$, la meta es de determinar la distribución dicha a posteriori $p_{\Theta|X=x}$ a partir de la cual se puede hacer estimación de θ dadas las observaciones, calcular intervalos de confianza, etc. Se interpreta como distribución explicando el parámetro a partir de las observaciones. Por eso, el metodo se apoya sobre la regla de Bayes $p_{\Theta|X=x}(\theta) \propto p_{X|\Theta=\theta}(x) p_\Theta(\theta)$ así que se necesita elegir una distribución p_Θ dicha a priori. **Ver nota de pie en el cap 2 a modificar**.

Dominio de definición	$\mathcal{X} = [a\,;\,b]$
Parámetros	$(a, b) \in \mathbb{R},\ b > a$
Densidad de probabilidad	$p_X(x) = \frac{1}{b-a}$
Promedio	$m_X = \dfrac{a+b}{2}$
Varianza	$\sigma_X^2 = \dfrac{(b-a)^2}{12}$
Asimetría	$\gamma_X = 0$ para $b \neq a$

Curtosis por exceso	$\bar{\kappa}_X = -\frac{6}{5}$ para $b \neq a$
Generadora de momentos	$M_X(u) = \frac{e^{bu} - e^{au}}{u}$ para [52] $u \in \mathbb{C}$
Función característica	$\Phi_X(\omega) = \frac{e^{\imath a\omega} - e^{\imath b\omega}}{\imath\,\omega}$

Obviamente, se puede escribir $X \overset{\mathrm{d}}{=} a + (b-a)U$ donde $\overset{\mathrm{d}}{=}$ significa que la equalidad es en distribución (las variables tienen la misma distribución de probabilidad), con $U \sim \mathcal{U}([0\,;\,1])$ llamada *uniforme estandar*.

La densidad de probabilidad y función de repartición de la variable estandar son representadas en la figura Fig. 1-25.

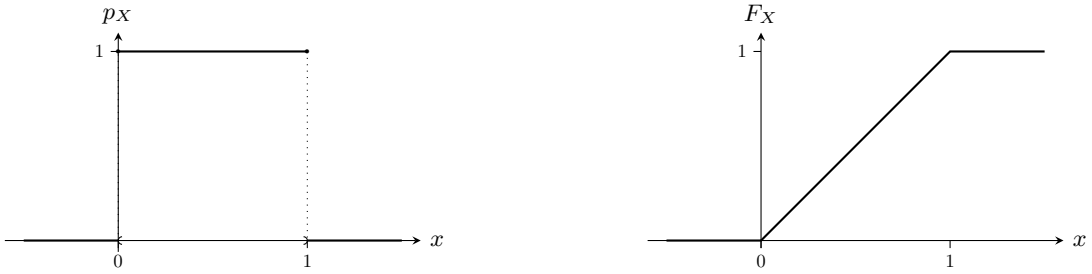


Figura 1-25: Ilustración de una densidad de probabilidad uniforme (a), y la función de repartición asociada (b).

Una nota importante es que cada ley continua es vínculada a la ley uniforme sobre $(0\,;\,1)$ de la manera siguiente:

Lema 1-28 (Inversión)**.** *Sea X, continua sobre $\mathcal{X} \subset \mathbb{R}$, de función de repartición F_X. Entonces*

$$U \equiv F_X(X) \sim \mathcal{U}(0\,;\,1)$$

Reciprocamente, definiendo la función de repartición inversa (o quantile)

$$F_X^{-1}(u) = \text{ínf}\{x \mid F(x) \geq u\}$$

si $V \sim \mathcal{U}(0\,;\,1)$*,*

$$Y = F_X^{-1}(V) \quad \Rightarrow \quad F_Y(y) = F_X(y)$$

Cuando F_X se inversa sencillamente eso da una manera sencilla de tirar sampleos de función de repartición F_X a partir de sampleos tirados según una ley uniforme.

[52] En el caso límite $u \to 0$, $\lim_{u\to 0} \frac{e^{bu}-e^{au}}{u} = b - a$, y similarmente para la función característica

Demostración. Inmediatamente, F_X siendo creciente,

$$\begin{aligned} P(U \leq u) &= P(F_X(X) \leq u) \\ &= P(X \leq F_X^{-1}(u)) \\ &= F_X\left(F_X^{-1}(u)\right) \end{aligned}$$

Similarmente

$$\begin{aligned} P(Y \leq y) &= P(F_X^{-1}(V) \leq y) \\ &= P(V \leq F_X(y)) \\ &= F_X(y) \end{aligned}$$

□

De manera general, para cualquier ensemble $\mathcal{D} \subset \mathbb{R}^d$ de volumen $|\mathcal{D}|$ la variable uniforma sobre $\mathcal{D}$ tiene la densidad con respecto a la medida "natural" sobre $\mathcal{D}$ (Lebesque, discreta,...) constante sobre $\mathcal{D}$,

$$p_X(x) = \frac{1}{|\mathcal{D}|}\mathbb{1}_{\mathcal{D}}(x)$$

La media va a ser el centro de gravedad de $\mathcal{D}$.

Vamos a ver en el capítulo 2 que esta distribución es la distribución definida sobre un conjunto de volumen finito que maximiza la entropía, *i. e.,* que es la "menos informativa". Por ejemplo, si se busca un parámetro modelizado como aleatorio (enfoque bayesiano), definido sobre un conjunto de volumen finito, sin a priori más, una tendencia natural/intuitiva es de asignar la "misma probabilidad" a cada punto del conjunto. Es por eso que aparece así naturalmente en la inferencia bayesiana (Robert, 2007).

Notar que cuando $b \to a$, la variable tiende a una variable cierta $X = a$ (ver principio de esta sección).

1.10.2.2. Distribución normal o gaussiana multivariada real

En el caso escalar, esta ley parece aparecer por unas de las primeras veces en trabajos de de Moivre como approximación de la ley binomial para n grande, usando la formula de Stirling (de Moivre, 1730, 1733, 1756; Pearson, 1924; Pearson, de Moivre & Archibald, 1926; Deming, 1933; Hald, 1984, 1990; Johnson et al., 1995a; David & Edwards, 2001; Hald, 2006). Se puede ver también el trabajo de F. Galton, quien construyó un experimento, la caja dicha de Galton, que ilustra por una parte como se puede obtener la ley binomial como suma de Bernoulli, y la convergencia a la gaussiana (Galton, 1889, Figs. 7-9, p. 63) o (Pearson, 1920, p. 38). Aparte de Moivre, la ley gaussiana fue desarollado mucho por los matemáticos como Gauss en el estudio del movimiento de planetas con perturbaciones (predicción de la trayectoria de Céres) (Gauss, 1809; Pearson, 1924; David & Edwards, 2001; Hald, 2006), basado en trabajos de A. M. Legendre (Legendre, 1805; David & Edwards, 2001; Hald, 2006), o Laplace en

mismos tipos de problemas (Laplace, 1809a, 1809b, 1812, 1814; de Laplace, 1820; Pearson, 1924; David & Edwards, 2001; Hald, 2006). De hecho, apoyandose en trabajos de de Moivre, la formalizó antes y más claramente Laplace, quien revandicó entonces su partenidad (ver por ejemplo (Pearson, 1920)). Por eso, esta ley es también conocida como ley de Laplace-Gauss.

En el contexto multivariado, la extensión natural de la ley binomial siendo la ley multinomial, es sin sorpresa que se introdujo la gaussiana multivaluada como approximación de la multinomial. Este trabajo es debido entre otros a J. L. Lagrange en los años 1770, con correcciones debido unas decadas después a A. de Morgan (de Morgan, 1838). Pero apareció antes en el caso bidimensional, en particular a través del estudio del coeficiente de correlación entre variables aleatorias (ver por ejemplo trabajos de Galton (Galton, 1877a, 1877b; Pearson, 1920)).

A pesar de que parece menos natural en la modelisación de fenomenos aleatorios que leyes uniformes, la ley gaussiana es seguramente unas de las más importante en probabilidad, sino que la más importante y la más expendida en la naturaleza. Eso viene sin duda del teorema del límite central. En dos palabras, cuando se suman un número importante de variables aleatorias (independientes, de misma ley, admitiendo una varianza, o con menos restricciones (Athreya & Lahiri, 2006, Cap. 11)), correctamente normalizado, esta suma tiende a una gaussiana [53]. En la naturaleza, se puede ver el ruido (señales) como suma de un número importante de fuentes de ruido independientes, justificando el modelo gaussiano (Feller, 1971; Le Cam, 1986; Ash & Doléans-Dade, 1999; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Rényi, 2007; Billingsley, 2012). Adémas, como lo vamos a ver en el capítulo 2, esta ley es la de incerteza máxima (maximizando la entropía) teniendo una dada varianza. Aparece naturalmente en termodínamica (gaz perfecto, con un número muy alto de particulas) (Maxwell, 1867; Boltzmann, 1896, 1898; Gibbs, 1902; Jaynes, 1965). En estimación, bajo la hipotesis gaussiana, los estimadores de parámetros minimizando el error cuadrático promedio son generalmente lineal (Kay, 1993; Robert, 2007). Todas estas consideraciones dan a la ley gaussiana un rol central en la teoría de las probabilidades.

Se denota $X \sim \mathcal{N}(m, \Sigma)$ con $m \in \mathbb{R}^d$ y $\Sigma \in P_d^+(\mathbb{R})$ conjunto de las matrices de $\mathcal{M}_{d,d}(\mathbb{R})$ símetricas definidas positivas. Las características de la gaussiana son las siguientes:

[53] De hecho, la approximación de la ley binomial por una gaussiana cuando n es grande es una caso particular del teorema, siendo la binomial una suma de Bernoulli independientes.

Dominio de definición	$\mathcal{X} = \mathbb{R}^d$
Parámetros	$m \in \mathbb{R}^d,\ \Sigma \in P_d^+(\mathbb{R})$
Densidad de probabilidad	$p_X(x) = \dfrac{1}{(2\pi)^{\frac{d}{2}} \lvert\Sigma\rvert^{\frac{1}{2}}} e^{-\frac{1}{2}(x-m)^t \Sigma^{-1}(x-m)}$
Promedio	$m_X = m$
Covarianza	$\Sigma_X = \Sigma$
Asimetría	$\gamma_X = 0$

Curtosis por exceso	$\bar{\kappa}_X = 0$
Generadora de momentos	$M_X(u) = e^{u^t \Sigma u + u^t m}$ para $u \in \mathbb{C}^d$
Función característica	$\Phi_X(\omega) = e^{-\frac{1}{2}\omega^t \Sigma \omega + \imath \omega^t m}$

Nota: trivialmente, se puede escribir $X \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}} N + m$ con $N \sim \mathcal{N}(0, I)$ donde N es dicha *gaussiana estandar* o *centrada-normalizada*. Las características de X son vínculadas a las de N (y vice-versa) por transformación afine (ver secciones anteriores).

La densidad de probabilidad gaussiana y la función de repartición en el caso escalar son representadas en la figura Fig. 1-26-(a) y (b) y una densidad en un contexto bi-dimensional figura Fig. 1-26(c).

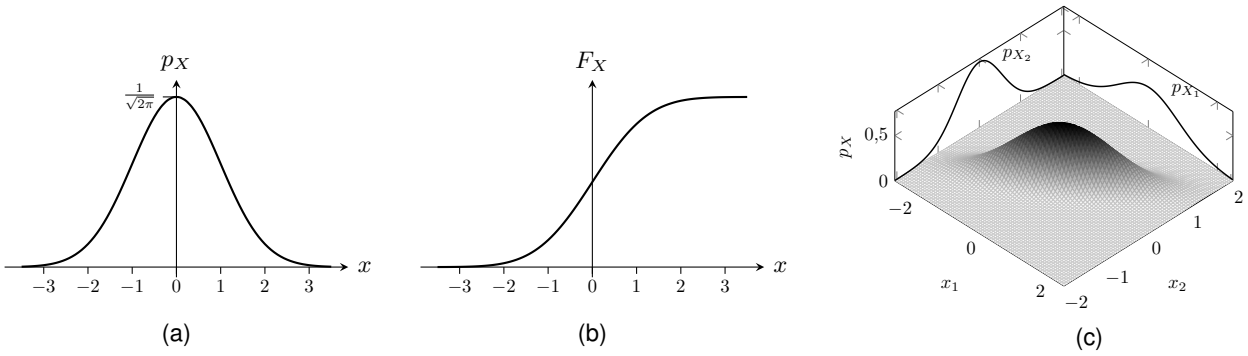


Figura 1-26: Ilustración de una densidad de probabilidad gaussiana escalar estandar (a), y la función de repartición asociada (b), así que una densidad de probabilidad gaussiana bi-dimensional centrada y de matriz de covarianza $\Sigma_X = R(\theta)\Delta^2 R(\theta)^t$ con $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ matriz de rotación y $\Delta = \operatorname{diag}\left(\begin{bmatrix} 1 & a \end{bmatrix}\right)$ matriz de cambio de escala, y sus marginales $X_1 \sim \mathcal{N}\left(0, \cos^2\theta + a^2 \sin^2\theta\right)$ y $X_2 \sim \mathcal{N}\left(0, \sin^2\theta + a^2 \cos^2\theta\right)$ (ver más adelante). En la figura, $a = \frac{1}{4}$ y $\theta = \frac{\pi}{6}$.

La gaussiana tiene un par de propiedades particulares:

Lema 1-29 (Gaussiana y cumulantes)**.** *Sea X vector aleatorio de media m, covarianza Σ y de secunda función característica admtiendo un desarollo de Taylor. Entonces*

$$\kappa_k[X] = 0 \quad \forall\, k \geq 4 \quad \Longleftrightarrow \quad X \sim \mathcal{N}(m, \Sigma)$$

Demostración. La pueba es inmediata del lema 1-13,

$$\kappa_k[X] = 0 \quad \forall\, k \geq 4 \quad \Longleftrightarrow \quad \Psi_X(\omega) = \imath\, \omega^t m - \frac{1}{2}\omega^t \Sigma \omega$$

lo que es nada más que la secunda función característica de la gaussiana, esa determiniendo completamente la ley. □

Teorema 1-42 (Stabilidad)**.** *Sean* $A_i, i = 1, \ldots, n$ *matrices de* $\mathcal{M}_{d',d}(\mathbb{R})$, $d' \leq d$ *de rango lleno,* $b_i \in \mathbb{R}^{d'}$ *y* $X_i \sim \mathcal{N}(m_i, \Sigma_i)$ *independientes, entonces*

$$\sum_{i=1}^{n} (A_i X_i + b_i) \sim \mathcal{N}\left(\sum_{i=1}^{n} (m_i + b_i) \, , \, \sum_{i=1}^{n} A_i \Sigma_i A_i^t \right)$$

En particular, cualquier combinación lineal de los componentes de un vector gaussiano da una gaussiana. Reciprocamente, si cualquier combinación lineal de los componentes de un vector aleatorio sigue una ley gaussiana, entonces el vector es gaussiano.

Demostración. Este resultato se proba usando función característica de la gaussiana, conjuntamente al teorema 1-34. □

Corolario 1-10 (Media empírica)**.** *Sean* $X_i \sim \mathcal{N}(m, \Sigma)$, $i = 1, \ldots, n$ *independientes. Entonces,*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}\left(m \, , \, \frac{1}{n} \Sigma \right)$$

$\overline{X}$ *es llamada* media empírica [54]*, y es un estimador "natural" de la media de un vector aleatorio a partir de copias independientes de misma ley.*

Teorema 1-43 (Independencia)**.** *Sea* $X \sim \mathcal{N}(m, \Delta)$ *con* $\Delta = \operatorname{diag}\left(\begin{bmatrix} \sigma_1^2 & \cdots & \sigma_d^2 \end{bmatrix}^t \right)$ *diagonal. Entonces los componentes* $X_i \sim \mathcal{N}(m_i, \sigma_i^2)$ *son independientes.*

Demostración. Este resultato se proba trivialmente escribiendo la densidad de probabilidad, notando que se factorisa. □

Hemos visto que cuando un vector tiene componentes independientes, la matriz de covarianza es diagonal (lema 1-6), pero que la recíproca es falsa en general. El último teorema muestra que la recíproca vale en el caso gaussiano.

Volvemos ahora al rol central de la gaussiana como modelo probabilistico muy frecuente de fenomeos aleatorios. Este rol particular viene del teorema del límite central que ya introdujimos. A veces, es conocido como teorema de Lindenberg-Feller (por lo menos la forma con condiciones más debiles que en la formulación original). Para unas de las formulaciones originales, se puede referirse al trabajo de Laplace de 1809 o de 1912 (Laplace, 1809a, 1809b, 1812, 1814; de Laplace, 1820). El nombre "central" viene de un documento de G. Pólya de 1920, titulado "Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem" ("Sobre el teorema del límite central del cálculo probabilístico y el problema de los momentos; el teorema es central… (Polya, 1920; Le Cam, 1986)). Se enuncia de manera siguiente (Spiegel, 1976; Brockwell & Davis, 1987; Lehmann & Casella, 1998; Ash & Doléans-Dade, 1999; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Billingsley, 2012):

[54] Es la estimación óptima de la media m a partir de los X_i en el sentido del error cuadratico promedio mínimo, o en el sentido de la verosimlitud máxima (Kay, 1993; Robert, 2007).

Teorema 1-44 (Teorema del límite central)**.** *Sea* $\{X_i\}_{i\in\mathbb{N}^*}$ *una sucesión de vectores aleatorios independientes, de misma ley, y que admiten un promedio* m *y una matriz de covarianza* Σ*. Entonces*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - m) \xrightarrow[n\to+\infty]{\mathrm{d}} Y \sim \mathcal{N}(0,\Sigma)$$

donde $\xrightarrow{\mathrm{d}}$ *significa que el límite es en distribución (ver notaciones).*

Demostración. Hay varias pruebas de este resultado. Quizás la más simple se apoya sobre la función cáracteristica. Sin perdida de generalidad, supongamos que $m = 0$. Sea $Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$. Sea ω fijo. Por independencia y relaciones del teorema 1-34 [55]:

$$\begin{aligned}
\Phi_{Y_n}(\omega) &= \left(\Phi_{X_i}\left(\frac{\omega}{\sqrt{n}}\right)\right)^n \\
&= \left(\Phi_{X_i}(0) + \frac{1}{\sqrt{n}}\,\omega^t\,\nabla_\omega\Phi_{X_i}(0) + \frac{1}{2n}\,\omega^t\,\mathcal{H}_\omega\Phi_{X_i}(0)\,\omega + o\left(n^{-1}\right)\right)^n \\
&= \left(1 - \frac{1}{2n}\,\omega^t\,\Sigma\,\omega + o\left(n^{-1}\right)\right)^n \\
&\xrightarrow[n\to+\infty]{} \exp\left(-\frac{1}{2}\omega^t\Sigma\omega\right)
\end{aligned}$$

porque $\Phi_{X_i}(0) = 1$, X_i siendo de media nula el gradiente de la función característica se cancela en $\omega = 0$, y $\mathcal{H}_\omega\Phi_{X_i}(0) = -\Sigma$. Se reconoce ahora la función característica de la gaussiana, lo que prueba que la función característica de Y_n converge simplemente hacia la función característica de la gaussiana. Se cierra la pruba usando el teorema de convergencia de Lévy, diciendo que la convergencia simple de la función característica implica la convergencia en distribución (Ash & Doléans-Dade, 1999; Billingsley, 2012; Athreya & Lahiri, 2006). □

En particular, la media empírica hechas a partir de vectores independientes de media m, admitiendo una covarianza Σ y de misma ley (no necesariamente gaussiana), tiende a ser gaussiana de media m y covarianza $\frac{1}{n}\Sigma$.

Existen varias variantes de este teorema que enunciamos, sin dar la prueba. Dejamos el lector a libros más especializados como (Ash & Doléans-Dade, 1999; Billingsley, 2012; Athreya & Lahiri, 2006; Lindeberg, 1922).

Teorema 1-45 (Teorema de Lindenberg-Feller)**.** *Sean* $\{X_i\}_{i\in\mathbb{N}^*}$ *vectores aleatorios independientes, no necesariamente de misma distribución de probabilidad, con* X_i *de media* $m_i = \mathrm{E}[X_i]$ *y de matriz de covarianza* $\Sigma_i \in P_d^+(\mathbb{R})$*. Sean* $C_n = \sum_{i=1}^{n}\Sigma_i$*,* c_n^2 *al autovalor más pequeña de* C_n*, y* $Y_n = C_n^{-\frac{1}{2}}\sum_{i=1}^{n}(X_i - \mathrm{E}[X_i])$*.*

$$\textit{Si}\quad \lambda_n > 0 \quad \textit{y}\quad \forall\,\varepsilon > 0,\quad \lim_{n\to+\infty}\sum_{i=1}^{n}\mathrm{E}\left[\left\|\frac{X_i - m_i}{c_n}\right\|^2 \mathbb{1}_{[\varepsilon\,;+\infty)}\left(\left\|\frac{X_i - m_i}{c_n}\right\|\right)\right] = 0$$

[55] $o\left(n^{-1}\right)$ significa que el termino que queda, digamos ε es tal que $n\varepsilon$ tiende a cero cuando n tiende al infinito.

entonces

$$Y_n \xrightarrow[n\to+\infty]{\mathrm{d}} Y \sim \mathcal{N}(0, I)$$

En numerosos libros, este teorema es dado en el caso escalar. Se extiende sencillamente al caso multivariado gracia a lo que es conocido como *teorema de Cramér-Wold*, diciendo que una secuencia de vectores aleatorios $Y_n \xrightarrow{\mathrm{d}} Y$ si y solamente para cualquier $u \in \mathbb{R}^d$ $u^t Y_n \xrightarrow{\mathrm{d}} u^t Y$ (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Billingsley, 2012).

Sin dar la prueba, la condición de Lindenberg dice que si la suma de las "dispersiónes" de los vectores normalizados por los que es basicamente la varianza la más pequeña de los componentes de la suma (una vez diagonalizada) se concentra asintoticamente, la suma renormalizada de lo vectores centrados tiende a la gaussiana (en distribución).

Se puede ver que se satisface la condición de Lindeberg en el caso de variables independientes de misma ley, del hecho que $C_n = n\Sigma$, lo que da $c_n^2 = nc^2$ con c^2 autovalor más pequeña de Σ. A continuación da la condición $\lim_{n\to\infty} \mathrm{E}\left[\|X_i - m_i\|^2 \mathbb{1}_{[\varepsilon\,;+\infty)}\left(\left\|\frac{X_i - m_i}{\sqrt{n}c}\right\|\right)\right] = 0$, satisfecha porque el argumento de la función indicadora tiende a 0 (casi siempre).

Un otro caso "trivial" aparece cuando la secuencia es uniformamente acotada, *i. e.,* $\forall i, \quad \|X_i\| \leq M$. Se puede retomar los argumentos anteriores, remplazando las variables por la cota.

Nota: si se satisface la condición dicha *de Lyapunov*, *i. e.,* si existe $\delta > 0$ tal que

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}\left[\frac{\|X_i - m_i\|^2}{c_n^{2+\delta}}\right] = 0,$$

entonces se satisface la condición de Lindeberg (Ash & Doléans-Dade, 1999). Frecuentemente, es más sencillo verificar la condición más fuerte de Lyapunov para probar la convergencia de una suma de vectores aleatorios a la gaussiana.

Aparece que se puede aún debilitar la condición de independencia sin perder la convergencia a la gaussiana. Para más detalles, ver por ejemplo (Brockwell & Davis, 1987, Sec. 6.4).

1.10.2.3. Distribución normal o gaussiana multivariada complejas

Por definición, un vector aleatorio complejo d-dimensional $Z = X + \imath Y$ es gaussiano significa que el vector $2d$-dimensional $\widetilde{Z} = \begin{bmatrix} X^t & Y^t \end{bmatrix}^t$ es gaussiano. Se puede entonces referirse en el caso de vectores gaussianos, pero como lo hemos presentado en la sección 1.9.1, es frecuentemente más comodo trabajar con Z en lugar de $\widetilde{Z}$.

En el caso general, la gaussiana real siendo completamente descrita por su media y su matriz de covarianza, la gaussiana compleja va a ser completamente definida por la media, la matriz de covarianza y la pseudo-covarianza (ver Sec. 1.9 por las relaciones entre la covarianza de $\widetilde{Z}$ y estas matrices).

Se denota $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$ con $m \in \mathbb{C}^d$, $\Sigma \in P_d^+(\mathbb{C})$ conjunto de las matrices de $\mathcal{M}_{d,d}(\mathbb{C})$ hermíticas definidas positivas, y $\check{\Sigma} \in S_d(\mathbb{C})$ conjunto de las matrices de $\mathcal{M}_{d,d}(\mathbb{C})$ symmetricas (ver

notaciones). Un caso particular aparece cuando Z es propio en torno a m, lo que es equivalente en el caso gaussiano a tener Z circular (ver más adelante) en torno a m, dado cuando $\check{\Sigma} = 0$: en este caso usaremos la misma notación, $Z \sim \mathcal{CN}(m, \Sigma)$. Las características de la gaussiana compleja son las siguientes (Lapidoth, 2017; Picinbono, 1996; Goodman, 1963; van den Bos, 1995; Schreier & Scharf, 2003; Eriksson & Koivunen, 2006):

Dominio de definición	$\mathcal{Z} = \mathbb{C}^d$
Parámetros	$m \in \mathbb{C}^d,\ \Sigma \in P_d^+(\mathbb{C}),\ \check{\Sigma} \in S_d(\mathbb{C})$
Densidad de probabilidad Caso general: Caso circular:	 $p_Z(z) = \dfrac{1}{\pi^d \lvert\Sigma\rvert^{\frac{1}{2}} \lvert P\rvert^{\frac{1}{2}}}\, e^{-(z-m)^\dagger P^{-1}(z-m) + \Re e\{(z-m)^t R^t P^{-1}(z-m)\}}$ con [56] $P = \Sigma - \check{\Sigma}\,\Sigma^{-*}\,\check{\Sigma}^\dagger, \quad R = \check{\Sigma}^\dagger\,\Sigma^{-1}$. $p_Z(z) = \dfrac{1}{\pi^d \lvert\Sigma\rvert}\, e^{-(z-m)^\dagger \Sigma^{-1}(z-m)}$
Promedio	$m_Z = m$
Covarianza	$\Sigma_Z = \Sigma$
Pseudo-covarianza	$\check{\Sigma}_Z = \check{\Sigma}$
Función característica Caso general: Caso circular:	 $\Phi_Z(\omega) = e^{-\frac{1}{4}\omega^\dagger \Sigma \omega - \frac{1}{4}\Re e\{\omega^\dagger \check{\Sigma} \omega^*\} + \imath \Re e\{\omega^\dagger m\}}, \quad \omega \in \mathbb{C}^d$ $\Phi_Z(\omega) = e^{-\frac{1}{4}\omega^\dagger \Sigma \omega + \imath \Re e\{\omega^\dagger m\}}, \quad \omega \in \mathbb{C}^d$

Notar que en el caso escalar propio (circular), la varianza de Z es $\sigma_Z^2 = 2\sigma^2$. El coefficiente 2 viene del hecho que Z contiene dos componentes independientes de varianza σ^2.

Los vectores aleatorios complejos van a compartir las propiedades del caso real, siendo equivalente a un vector $2d$-dimensional gaussiano real.

Primero, los cumulantes de orden superior o igual a 4 valen cero:

Lema 1-30 (Gaussiana compleja y cumulantes)**.** *Sea Z vector aleatorio complejo de media m, de covarianza Σ, de pseudo-covarianze $\check{\Sigma}$ y de secunda función característica admtiendo un desarollo de*

[56] En (Picinbono, 1996) la expresión es ligieramente diferente, pero se recupera usando la simetría $\check{\Sigma}^* = \check{\Sigma}^\dagger$. Recordar que $\cdot^{-*} = (\cdot^*)^{-1}$ (ver notaciones).

Taylor. Entonces para cualquier

$$\kappa_{i_1,\ldots,i_l,i'_1,\ldots,i'_m}[Z] = 0 \quad \forall\, (i_1,\ldots,i_l,i'_1,i'_m \in \{1,\ldots,d\}^{l+m},\, l+m \geq 4 \quad \Longleftrightarrow \quad X \sim \mathcal{CN}(m,\Sigma,\check{\sigma})$$

Secundo, como en el caso real, la gaussiana es estable por combinación lineal de vectores independientes:

Teorema 1-46 (Stabilidad)**.** *Sean* $A_i, i = 1,\ldots,n$ *matrices de* $\mathcal{M}_{d',d}(\mathbb{C})$, $d' \leq d$ *de rango lleno,* $b_i \in \mathbb{C}^{d'}$ *y* $Z_i \sim \mathcal{CN}(m_i,\Sigma_i,\check{\Sigma}_i)$ d*-dimensionales, independientes, entonces*

$$\sum_{i=1}^{n}(A_iZ_i+b_i) \sim \mathcal{CN}\left(\sum_{i=1}^{n}(m_i+b_i)\,,\,\sum_{i=1}^{n}A_i\Sigma_iA_i^\dagger\,,\,\sum_{i=1}^{n}A_i\check{\Sigma}_iA_i^t\right)$$

En particular, cualquier combinación lineal de los componentes de un vector gaussiano complejo da una gaussiana compleja. Reciprocamente, si cualquier combinación lineal de los componentes de un vector aleatorio sigue una ley gaussiana compleja, entonces el vector es gaussiano complejo.

El corolario 1-10 se extiende naturalmente al caso complejo:

Corolario 1-11 (Media empírica)**.** *Sean* $Z_i \sim \mathcal{CN}\left(m,\Sigma,\check{\Sigma}\right)$, $i = 1,\ldots,n$ *independientes. Entonces,*

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n}Z_i \sim \mathcal{CN}\left(m\,,\,\frac{1}{n}\Sigma\,,\,\frac{1}{n}\check{\Sigma}\right)$$

Además, en el caso complejo se tiene una estabilidad combinando Z y Z^*:

Teorema 1-47. *Sean* $A \in \mathcal{M}_{d',d}(\mathbb{C})$, $B \in \mathcal{M}_{d',d}(\mathbb{C})$ *tales que ambas* $A+B$ *y* $A-B$ **sean de rango lleno**, $c \in \mathbb{C}^{d'}$ *y* $Z \sim \mathcal{CN}(m,\Sigma,\check{\Sigma})$ d*-dimensonal, entonces*

$$AZ + BZ^* + c \sim \mathcal{CN}\left(\mu, C, \check{C}\right)$$

con

$$\begin{aligned}\mu &= Am + Bm^* + c\\ C &= A\Sigma A^\dagger + B\Sigma^* B^\dagger + A\check{\Sigma}B^\dagger + B\check{\Sigma}^*A^\dagger\\ \check{C} &= A\check{\Sigma}A^t + B\check{\Sigma}B^t + A\Sigma B^t + B\Sigma^*A^t\end{aligned}$$

Demostración. Tomando la forma real $2d$-dimensional $Z = X + \imath Y$ con X, Y reales, es en biyección con $\widetilde{Z} = \begin{bmatrix}X\\Y\end{bmatrix}$ y entonces Z^* en biyección con $\widetilde{Z^*} = \begin{bmatrix}X\\-Y\end{bmatrix}$. Eso da $AZ + BZ^* + c$ en biyección con $\begin{bmatrix}A+B & 0\\0 & A-B\end{bmatrix}\begin{bmatrix}X\\Y\end{bmatrix} + \begin{bmatrix}\Re e\,\{c\}\\ \Im m\,\{c\}\end{bmatrix}$. Notando que $\begin{bmatrix}A+B & 0\\0 & A-B\end{bmatrix}$ es de rango lleno, por el teorema 1-42 este vector es gaussiano, lo que proba que $AZ + BZ^* + c$ es gaussiano complejo. Las formas de la media, covarianza y pseudo-covarianza siguen de calculos directos de la expresión $AZ + BZ^* + c$. □

Evidentemente, se puede combinar los dos teoremas anteriores.

El teorema del límite central y sus variantes se recuperan del caso real.

Teorema 1-48 (Teorema del límite central (caso complejo))**.** *Sea* $\{Z_i\}_{i\in\mathbb{N}^*}$ *una sucesión de vectores aleatorios independientes, de misma ley, y que admiten un promedio* m*, una matriz de covarianza* Σ *y una matriz de pseudo-covarianza* $\check{\Sigma}$*. Entonces*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{m}(Z_i - m) \xrightarrow[n\to+\infty]{\mathrm{d}} Z \sim \mathcal{CN}\left(0, \Sigma, \check{\Sigma}\right)$$

donde $\xrightarrow{\mathrm{d}}$ *significa que el límite es en distribución (ver notaciones).*

Como en el caso real, aparece que la media empírica hechas a partir de vectores complejos independientes de media m, admitiendo una covarianza Σ una pseudo-covarianza $\check{\Sigma}$, y de misma ley (no necesariamente gaussiana), tiende a ser gaussiana compleja de media m, de covarianza $\frac{1}{n}\Sigma$, y de pseudo-covarianza $\frac{1}{n}\check{\Sigma}$.

No lo presentamos, pero se transpone sencillamente el teorema de Lindenberg-Feller 1-45 al caso complejo.

Notamos también que, en el caso circular, se puede escribir naturalmente $Z \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}}N + m$ con $N \sim \mathcal{CN}(0, I)$ donde N es dicha *Gaussiana estandar* o *centrada-normalizada*. Eso se generaliza en dos direcciones. La primera pone también en juega una gaussiana estandar (Lapidoth, 2017):

Teorema 1-49. *Sea* $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$*. Entonces, existen matrices (no únicas)* $A \in \mathcal{M}_{d,d}(\mathbb{C})$*,* $B \in \mathcal{M}_{d,d}(\mathbb{C})$ *tales que*

$$Z \overset{\mathrm{d}}{=} AW + BW^* + m$$

con $W \sim \mathcal{CN}(0, I)$ *gaussiana estandar.*

Demostración. Inmediatamente

$$Z = \begin{bmatrix} I & \imath I \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \overset{\mathrm{d}}{=} \begin{bmatrix} I & \imath I \end{bmatrix} M \begin{bmatrix} U \\ V \end{bmatrix}$$

con $U \sim \mathcal{N}(0, I)$ y $V \sim \mathcal{N}(0, I)$ independientes, y M tal que $MM^t = \begin{bmatrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_{X,Y}^t & \Sigma_Y \end{bmatrix}$ (ej. raiz cuadrade de esta matriz de $P_{2d}^+(\mathbb{R})$, o descomposición de Cholesky (Horn & Johnson, 2013; Bhatia, 2007)). Ahora, volviendo a la forma compleja tenemos

$$Z \overset{\mathrm{d}}{=} \begin{bmatrix} I & \imath I \end{bmatrix} M \begin{bmatrix} I & I \\ -\imath I & \imath I \end{bmatrix} \begin{bmatrix} \frac{1}{2}(U + \imath V) \\ \frac{1}{2}(U - \imath V) \end{bmatrix}$$

Se cierra la prueba denotando

$$\begin{bmatrix} A & B \end{bmatrix} = \begin{bmatrix} I & \imath I \end{bmatrix} M \begin{bmatrix} I & I \\ -\imath I & \imath I \end{bmatrix}$$

y notando que $W \equiv \frac{1}{2}(U + \imath V) \sim \mathcal{CN}(0, I)$. □

Notar que, usando la descomposición de Cholesky, tenemos M triangular inferior [57], y entonces bloc-triangular inferior $M = \begin{bmatrix} \alpha & 0 \\ \beta & \gamma \end{bmatrix}$. Eso conduce, a $MM^t = \begin{bmatrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_{X,Y}^t & \Sigma_Y \end{bmatrix} = \begin{bmatrix} \alpha\alpha^t & \alpha\beta^t \\ \beta\alpha^t & \beta\beta^t + \gamma\gamma^t \end{bmatrix}$. Eso da por ejemplo $\alpha = \Sigma_X^{\frac{1}{2}}$, $\beta = \Sigma_{X,Y}^t \Sigma_X^{-\frac{1}{2}}$ y $\gamma = \left(\Sigma_Y - \Sigma_{X,Y}^t \Sigma_X^{-1} \Sigma_{X,Y}\right)^{\frac{1}{2}}$. A continuación, $A = \alpha + \gamma + \imath\beta$ y $B = \alpha - \gamma + \imath\beta$. Una solución posible es entonces

$$\begin{cases} A &= \Sigma_X^{\frac{1}{2}} + \left(\Sigma_Y - \Sigma_{X,Y}^t \Sigma_X^{-1} \Sigma_{X,Y}\right)^{\frac{1}{2}} + \imath \Sigma_{X,Y}^t \Sigma_X^{-\frac{1}{2}} \\ B &= \Sigma_X^{\frac{1}{2}} - \left(\Sigma_Y - \Sigma_{X,Y}^t \Sigma_X^{-1} \Sigma_{X,Y}\right)^{\frac{1}{2}} + \imath \Sigma_{X,Y}^t \Sigma_X^{-\frac{1}{2}} \end{cases}$$

Se puede re-escribir estas matrices a partir de Σ y $\check{\Sigma}$ usando las relaciones de la sección 1.9.1.

La secunda extensión pone en juega una sola gaussiana compleja sin su conjugada (Eriksson & Koivunen, 2006; Schreier & Scharf, 2003):

Teorema 1-50. *Sea $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$. Entonces, existe una matriz $C \in \mathcal{M}_{d,d}(\mathbb{C})$ tal que*

$$Z \overset{\mathrm{d}}{=} CW + m$$

con $W \sim \mathcal{CN}(0, I, \Delta)$ con $\Delta \in P_d(\mathbb{R})$ (real) diagonal.

Demostración. Eso viene de teoremas de diagonalización conjunta. Más precisamente, siendo $\Sigma \in P_d^+(\mathbb{C})$ y $\check{\Sigma} \in S_d(\mathbb{C})$, se aplica el teorema (Horn & Johnson, 2013, Teo. 7.6.5) diciendo que existe una matriz no singular (invertible) C tal que $\Sigma = CC^\dagger$ y $\check{\Sigma} = C\Delta C^t$ con Δ real diagonal con elementos positivos ($\Delta \in P_d(\mathbb{R})$ diagonal). Inmediatamente, por el teorema 1-46, tenemos

$$C^{-1}(Z - m) \overset{\mathrm{d}}{=} W \sim \mathcal{CN}(0, I, \Delta)$$

lo que cierra la prueba. □

Al final, vímos en la sección 1.9.1 que si un vector es circular, entonces su pseudo-covarianza es nula, pero la recíproca no vale en general. Aparece que en el contexto gaussiano tenemos la recíproca:

Teorema 1-51 (Circularidad)**.** *Sea $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$. Entonces,*

$$Z \text{ circular en torno a } m \iff Z \text{ propio en torno a } m$$

Demostración. Vímos la directa en la sección 1.9.1, teorema 1-39. Reciprocamente, si Z es propio en torno a m, por definición $\check{\Sigma} = 0$ y el resultado viene de la forma de la función característica por ejemplo: $\Phi_{Z-m}(\omega) = e^{-\frac{1}{4}\omega^\dagger \Sigma \omega} = \Phi_{Z-m}\left(e^{\imath\theta}\omega\right) = \Phi_{e^{\imath\theta}(Z-m)}(\omega)$. □

1.10.2.4. Distribución exponencial

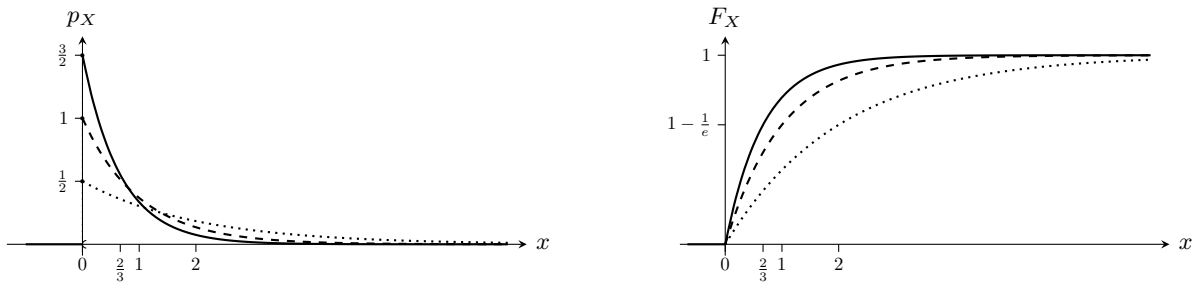
[57] Se puede hacer el mismo razonamiento con la forma triangular superior; se cambia los roles de X e Y en las matrices de covarianza.

A pesar de que sea un caso particular de la distribución Gamma que vamos a ver más adelante, estudiada por Pearson desde el año 1895 (Pearson, 1895), o apareció quizas un poco antes en trabajos de L. Boltzmann o de Whitworth (Balakrishnan & Basu, 1995) (como caso limíte de la ley de Poisson), apareció esta ley de manera "propia" mucho más tarde, entre otros en 1930 en (Kondo, 1930, Ec. (46)).

Se denota $X \sim \mathcal{E}(\lambda)$ con $\lambda \in \mathbb{R}_+^*$ llamada *taza* (inversa de *escala*), y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \mathbb{R}_+$
Parámetro	$\lambda \in \mathbb{R}_+^*$
Densidad de probabilidad	$p_X(x) = \lambda e^{-\lambda x}$
Promedio	$m_X = \dfrac{1}{\lambda}$
Varianza	$\sigma_X^2 = \dfrac{1}{\lambda^2}$
Asimetría	$\gamma_X = 2$
Curtosis por exceso	$\bar{\kappa}_X = 6$
Generadora de momentos	$M_X(u) = \dfrac{\lambda}{\lambda - u}$ para $\Re e\{u\} < \lambda$
Función característica	$\Phi_X(\omega) = \dfrac{\lambda}{\lambda - \imath\omega}$

Unas densidades de probabilidad y funciones de repartición asociadas son representadas en la figura Fig. 1-27 para varios parámetros.



(a) (b)

Figura 1-27: Ilustración de una densidad de probabilidad exponencial (a), y la función de repartición asociada (b), con $\lambda = \frac{3}{2}$ (linea llena), $\lambda = 1$ (linea guionada) y $\lambda = \frac{1}{2}$ (linea punteada).

La ley exponencial es conocida como siendo *sin memoria*, es decir, si buscamos X visto como un

tiempo (ej. tiempo de desintegración de un atómo radioactivo) tal que

$$\forall\, x_0 \geq 0,\, x \geq 0, \quad P(X > x + x_0 | X > x_0) = P(X > x)$$

i. e., la probabilidad que $X > x + x_0$ (extra tiempo después de x_0) condicionalmente a $X > x_0$ es exactamente la de $X > x$ (se olvidó x_0). De hecho, tenemos de la definición de la probabilidad condicional

$$\forall\, x_0 \geq 0,\, x \geq 0, \quad P(X > x + x_0 | X > x_0) = \frac{P\big((X > x + x_0) \cap (X > x_0)\big)}{P(X > x_0)}$$

Ahora, de $(X > x + x_0) \subset (X > x_0)$ se obtiene

$$\forall\, x_0 \geq 0,\, x \geq 0, \quad P(X > x + x_0 | X > x_0) = \frac{1 - F_X(x + x_0)}{1 - F_X(x_0)} = 1 - F_X(x)$$

De la densidad de probabilidad, tenemos para $x > 0$, $F_X(x) = \left(1 - e^{-\lambda x}\right) \mathbb{1}_{\mathbb{R}_+}(x)$, así que $1 - F_X(x) = e^{-\lambda x}$. Poniendo este resultado el la expresión anterior se obtiene finalmente

$$\forall\, x_0 \geq 0,\, x \geq 0, \quad P(X > x + x_0 | X > x_0) = e^{-\lambda x} = 1 - F_X(x) = P(X > x)$$

Como lo hemos evocado tratando de la ley de Poisson, esta es vínculada intimamente a la ley exponencial a través del processo dicho de poisson:

Lema 1-31 (Vínculo con la ley de Poisson)**.** *Sea $T_0 = 0$ y $\forall\, n \in \mathbb{N}^*$ las variables aleatorias positivas T_n tales que $T_{n+1} - T_n \geq 0$ son independientes y de distribución $\mathcal{E}(\lambda)$. Fijamos $T > 0$ y sea X el número de variables T_n que partenecen a $(0\,;\, T)$, i. e., $T_X < T < T_{X+1}$. Entonces*

$$X \sim \mathcal{P}(\lambda T)$$

Dicho de otra manera, si tenemos eventos que aparecen en tiempos aleatorios tales que los incrementos de tiempos entre eventos son independientes y de distribución exponencial de taza λ, el número de eventos en un intervalo de tiempo T dado sigue una ley de Poisson, de taza λT proporcional al intervalo, y proporcional a la taza de la ley exponencial. El parámetro λ representa la taza de evento por unidad de tiempo.

Demostración. Por definición,

$$\begin{aligned} P(X = n) &= P(X \leq n) - P(X \leq n - 1) \\ &= P(T_{n+1} > T) - P(T_n > T) \end{aligned}$$

Es decir

$$P(X = n) = F_{T_n}(T) - F_{T_{n+1}}(T)$$

Ahora, notando que

$$T_n = \sum_{i=0}^{n-1} (T_{i+1} - T_i)$$

de la independencia de los incrementos de tiempo, y de las propiedades de la función característica, tenemos

$$\Phi_{T_n}(\omega) = \frac{\lambda^n}{(1 - \imath\,\omega)^n}$$

De la fórmula de inversion del teorema 1-32 se prueba que [58]

$$p_{T_n}(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} \mathbb{1}_{\mathbb{R}_+}(x)$$

Con integraciones por partes, se obtiene sencillamente

$$F_{T_n}(T) = 1 - \sum_{i=0}^{n-1} \frac{\lambda^i T^i e^{-\lambda T}}{i!}$$

Se concluye poniendo este resultado en la expresión $P(X = n) = F_{T_n}(T) - F_{T_{n+1}}(T)$ que obtuvimos. □

En física, se modela la ley de tiempo de desintegración como siendo exponencial, y se supone que las desintegraciones son independientes. Eso justifica modelo de Poisson para modelizar el número de desintegración durante un tiempo dado.

Una otra característica de esta ley es su stabilidad con respecto al operador no lineal mínimo:

Lema 1-32 (Stabilidad por el mín)**.** *Sean* $X_i \sim \mathcal{E}(\lambda),\ i = 1, \ldots, n$ *independientes. Entonces,*

$$\min_{i=1,\ldots,n} X_i \equiv X \sim \mathcal{E}(n\lambda)$$

Demostración. Inmediatamente, para cualquier $x \geq 0$

$$\begin{aligned} 1 - F_X(x) &= P(X > x) \\ &= P\left(\bigcap_{i=1}^{n} (X_i > x)\right) \\ &= \prod_{i=1}^{n} P(X_i > x) \\ &= e^{-n\lambda x} \end{aligned}$$

La secunda linea viene de la equivalencia entre los eventos $\min_{i=1,\ldots,n} X_i > x$ y $\bigcap_{i=1}^{n} (X_i > x)$ y la tercera de la independencia de los X_i. □

1.10.2.5. Distribución gamma

[58] Una manera es de hacer una integración en el plano complejo y usar los lemas de Jordan y teorema de residuos (Carrier et al., 2005) o (Ablowitz & Fokas, 2003, Cap. 4). Nota: de hecho se reconoce en Φ_{T_n} la función característica de una ley gamma $\mathcal{G}(n, \lambda)$, ley que vamos a ver en la sección 1.10.2.5.

Como lo introdujimos en el ejemplo de la ley exponencial, esta familia de leyes fue estudiada por primera vez al fin del siglo XIV, bajo el impulso de Pearson (Pearson, 1895). De hecho, según Lancaster (Lancaster, 1966) se encuentran trazas de esta ley en trabajos de Laplace como posterior distribución en inferencia Bayesiana (elementos conduciendo a la ley gamma) para la estimación de la dispersión $\frac{1}{\sigma^2}$ de una ley gaussiana. De hecho, la distribución gamma aparece frecuentemente en problemas de inferencia Bayesiana como distribución a priori conjugado [59] del parámetro λ de la ley de Poisson (Robert, 2007). Se encuentren también trazas de esta ley en trabajos de J. Bienaymé como distribución límite del promedio centrado y renormalizado de los componentes de un vector de ley multinomial (Bienaymé, 1838; Lancaster, 1966).

Se denota $X \sim \mathcal{G}(a,b)$ con $a \in \mathbb{R}_+^*$ llamado *parámetro de forma* y $b \in \mathbb{R}_+^*$ llamada *taza* (inversa de *escala*). Las características son:

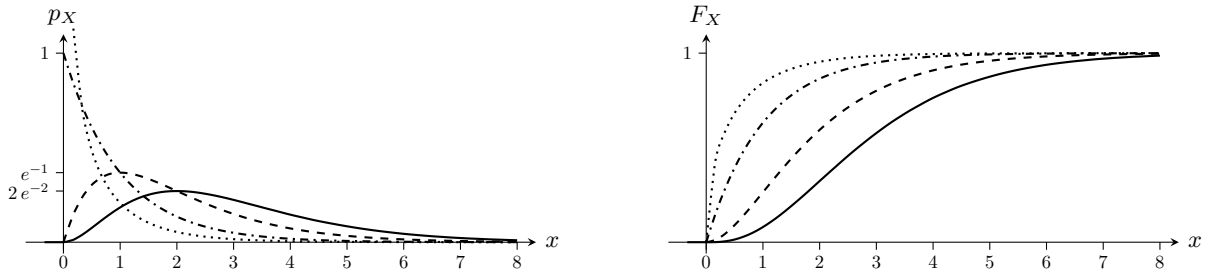
Dominio de definición	$\mathcal{X} = \mathbb{R}_+$
Parámetros	$a \in \mathbb{R}_+^*$ (forma), $b \in \mathbb{R}_+^*$ (taza)
Densidad de probabilidad	$p_X(x) = \dfrac{b^a\, x^{a-1}\, e^{-bx}}{\Gamma(a)}$
Promedio	$m_X = \dfrac{a}{b}$
Varianza	$\sigma_X^2 = \dfrac{a}{b^2}$
Asimetría	$\gamma_X = \dfrac{2}{\sqrt{a}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{6}{a}$
Generadora de momentos	$M_X(u) = \left(1 - \dfrac{u}{b}\right)^{-a}$ para $\Re e\,\{u\} < b$
Función característica	$\Phi_X(\omega) = \left(1 - \dfrac{u\omega}{b}\right)^{-a}$

Nota: trivialmente, se puede escribir $X \overset{\mathrm{d}}{=} \frac{1}{b}G$ con $G \sim \mathcal{G}(a,1)$ donde G es estandardizada

[59] Ver nota de pie 51 por la explicación del enfoque bayesiano que consiste a calcular la distribución a posteriori $p_{\Theta|X=x}(\theta) \propto p_{X|\Theta=\theta}(x)p_\Theta(\theta)$ usando la ley de los datos parametrizado por θ que queremos inferir, modelizado aleatorio. Por eso, como se lo ve en la fórmula de Bayes, se necesita elegir una distribución a priori p_Θ. Vímos que una elección posible es tomarla uniforme. Puede ser problematico por ejemplo cuando θ vive en un espacio de volumen inifinito (a priori impropio), aún si se lo usa frecuentemente (en estimación es equivalente a considerar la verosimulitud). Una otra elección posible es tomar el a priori en una familia parametrizada tal que la distribución a posterior partenece también a esta familia: es lo que se llama *a priori conjugado* para la ley de sampleos $p_{X|\Theta=\theta}$. La idea es que si vienen observaciones, en lugar de re-calcular la ley a posteriori, se puede actualizar solamente los parámetros (llamados hiperparámetros).

o normalizada. De nuevo, las características de X son vínculadas a las de G (y vice-versa) por transformación lineal (ver secciones anteriores).

Unas densidades de probabilidad gamma y las funciones de repartición asociadas son representadas en la figura Fig. 1-28 para varios a y $b = 1$.



(a) (b)

Figura 1-28: Ilustración de una densidad de probabilidad gamma (a), y la función de repartición asociada (b). $b = 1$ y $a = 0{,}5$ (linea punteada), 1 (linea mixta), 2 (linea guionada) y 3 (linea llena).

Cuando $a \in \mathbb{N}^*$ es entero, la ley es a veces conocida como ley de Erlang, del nombre de un ingeniero danés trabajando en (fundador de la) teoría de colas (Cox, 1962; Erlang, 1909, 1925; Brockmeyer, Halstrøm & Jensen, 1948). Si $a = \frac{n}{2}$ con n entero y $\beta = \frac{1}{2}$, se conoce tambié como ley *chi-cuadrado* con n grados de libertad (ver ej. (Johnson et al., 1995a)).

Notar que $X \sim \mathcal{G}(1, b)$ es una variable exponencial de parámetro b, *i. e.,* $X \sim \mathcal{E}(b)$. Cuando $a < 1$, la densidad p_X diverge para $x \to 0$ (divergencia integrable). Además, se muestra también sencillamente con las funciones características que:

Lema 1-33 (Stabilidad)**.** *Sean* $X_i \sim \mathcal{G}(a_i, b)\,,\ i = 1, \ldots, n$ *independientes. Entonces*

$$\sum_{i=1}^{n} X_i \sim \mathcal{G}\left(\sum_{i=1}^{n} a_i\,,\, b\right)$$

En particular, la suma de variables independientes de ley exponencial de mismos parámetro sigue una distribución de Erlang de parámetro de forma n.

Además, se muestra sencillamente por cambio de variables y con la función característica un vínculo con variables gaussianas:

Lema 1-34 (Vínculo con la gaussiana)**.** *Sean* $X_i \sim \mathcal{N}\left(0, \sigma^2\right)\,,\ i = 1, \ldots, n$ *independientes. Entonces*

$$\sum_{i=1}^{n} X_i^2 \sim \mathcal{G}\left(\frac{n}{2}\,,\, \frac{1}{2\,\sigma^2}\right)$$

(ley chi-cuadrado precisamente con n *grados de libertad).*

1.10.2.6. Distribución matriz-variada de Wishart

Este ejemplo es una generalización matriz-variada de la distribución gamma. Se puede ver una matriz como un vector, guardando por ejemplo sus columnas una bajo la precediente. Sin embargo, tal distribución apareciendo naturalmente en un contexto de estimación de matriz de covarianza (ver más adelante), es más natural verla matriz-variada. Tal distrubución fue introducida en $d = 2$ dimensiones por R. Fisher en 1915 (?, ?), y el caso general es debido a J. Wishart (Wishart, 1928; Muirhead, 1982; Bilodeau & Brenner, 1999; Gupta & Nagar, 1999; Anderson, 2003; Seber, 2004). Aparece también naturalmente en problema de inferencia Bayesiana como distribución a priori conjugado de la matriz de precisión Σ^{-1} (inversa de la covarianza) de la ley gaussiana multivariada (Robert, 2007, Ec. (4.4.5)) (ver notas de pie 51 y 59).

Se denota $X \sim \mathcal{W}(V, \nu)$ donde el dominio de definición es $P_d^+(\mathbb{R})$, conjunto matrices simetricas definida positivas, $V \in P_d^+(\mathbb{R})$ parámetro de escala y $\nu > d - 1$ es el grado de libertad. Las características de la distribución son las siguientes:

Dominio de definición [60]	$\mathcal{X} = P_d^+(\mathbb{R}),\ d \in \mathbb{N}^*$
Parámetros	$V \in P_d^+(\mathbb{R})$ (escala) y $\nu > d-1$ (grado de libertad)
Densidad de probabilidad [61]	$p_X(x) = \dfrac{\lvert x\rvert^{\frac{\nu-d-1}{2}}\, e^{-\frac{1}{2}\operatorname{Tr}(V^{-1}x)}}{2^{\frac{d\nu}{2}}\, \lvert V\rvert^{\frac{\nu}{2}}\, \Gamma_d\left(\frac{\nu}{2}\right)}$
Promedio	$m_X = \nu V$
Covarianza	$\Sigma_X = \nu \sum_{i,j=1}^{d} \Big(\big(\mathbb{1}_i \mathbb{1}_j^t V\big) \otimes \big(\mathbb{1}_j \mathbb{1}_i^t V\big) + \big(V \mathbb{1}_i \mathbb{1}_j^t V\big) \otimes \big(\mathbb{1}_i \mathbb{1}_j^t\big) \Big)$
Función característica [62]	$\Phi_X(\omega) = \lvert I - 2\,\imath\,\omega V\rvert^{-\frac{\nu}{2}}, \quad \omega \in S_d(\mathbb{R})$

Fijense que p_X no es la distribución conjunta de los componentes de X: el hecho que X sea una matriz aleatoria de $P_d^+(\mathbb{R})$ impone vínculos sobre sus compnentes; entre otros, $X_{i,j} = X_{j,i}$.

Inmediatamente, si $d = 1$, la distribución de Whishart $\mathcal{W}(V, \nu)$ se reduce a la distribución Gamma

---

[60] De hecho, se puede considerar que la matriz aleatoria es equivalent a tener un vector $\frac{d(d+1)}{2}$-dimensional; por la simetria, claramente X tiene solamente $\frac{d(d+1)}{2}$ componentes diferentes; además, se puede probar que cualquier matriz $A \in P_d^+(\mathbb{R})$ se factoriza bajo la forma $A = LL^t$ con L triangular inferior con elementos positivos sobre su diagonal, llamado factorización de Cholesky (Cholesky, 2005; Gupta & Nagar, 1999; Bhatia, 2007; Harville, 2008; Horn & Johnson, 2013) y reciprocamente. Eso muestra que A se define a partir de $\frac{d(d+1)}{2}$ "grado de libertad".

[61] La densidad de probabilidad corresponde a la densidad conjunta de los $\frac{d(d+1)}{2}$ elementos $X_{i,j},\ 1 \leq i \leq j \leq d$ (Wishart, 1928; Peddada & Richards, 1991; Sultan & Tracy, 1996; Muirhead, 1982; Bilodeau & Brenner, 1999; Gupta & Nagar, 1999; Anderson, 2003; Seber, 2004).

[62] Se prueba que la función generadora de momentos no existe en general.

$\mathcal{G}\left(\frac{\nu}{2}\,,\,\frac{1}{2V}\right)$. De este hecho, se la podría ver como extensión matriz-variada de la distribución gamma. La distribución de Wishart tiene varias otras propiedades como las siguientes.

Lema 1-35 (Stabilidad por transformación lineal)**.** *Sea* $X \sim \mathcal{W}(V,\nu)$ *y* $A \in \mathcal{M}_{d,d'}(\mathbb{R})$ *con* $d' \leq d$*, de rango lleno. Entonces*

$$A^t X A \sim \mathcal{W}\left(A^t V A, \nu\right)$$

En particular, si $d' = 1$*,* $A^t X A \sim G\left(\frac{\nu}{2}\,,\,\frac{1}{2\,A^t V A}\right)$*. Más allá, tomando* $A = \mathbb{1}_j$*, aparece de que las componentes diagonales de* X *son de distribución gamma,* $X_{j,j} \sim \mathcal{G}\left(\frac{\nu}{2}\,,\,\frac{1}{2\,V_{j,j}}\right)$*.*

Demostración. El resultado es inmediato saliendo de la función característica [62] y notando de que

$$\begin{aligned}
\Phi_{A^t X A}(\omega) &= \mathrm{E}\left[e^{\imath\,\mathrm{Tr}\left(\omega^t A^t X A\right)}\right] \\
&= \Phi_X\left(A\omega^t A^t\right) \\
&= \left|I - 2\imath A\omega A^t V\right|^{-\frac{\nu}{2}} \\
&= \left|I - 2\imath\omega A^t V A\right|^{-\frac{\nu}{2}}
\end{aligned}$$

de $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ (Harville, 2008) y de la identidad de Sylvester (Sylvester, 1851; Akritas, Akritas & Malaschonok, 1996) o (Harville, 2008, § 18.1) $|I + AB| = |I + BA|$. . □

De hecho, si los elementos diagonales son de distribución gamma, no es el caso de los elementos no-diagonales (Seber, 2004; Anderson, 2003) o (Gupta & Nagar, 1999, Teo. 3.3.4). De eso resuelte delicado llamar la distribución como gamma matriz-variada.

Lema 1-36 (Stabilidad por suma)**.** *Sea* $X_i \sim \mathcal{W}(V,\nu_i)$*,* $i = 1,\ldots,n$ *independientes. Entonces*

$$\sum_{i=1}^{n} X_i \sim \mathcal{W}\left(V\,,\,\sum_{i=1}^{n}\nu_i\right)$$

Demostración. El resultado es inmediato saliendo de la función caractéristica [62] y notando que como en el context vectorial $\Phi_{\sum_i X_i} = \prod_i \Phi_{X_i}$. □

La distribución de Wishart aparece naturalmente en problemas de estimación de matriz de covarianza en el contexto gaussiano (Muirhead, 1982; Bilodeau & Brenner, 1999; Gupta & Nagar, 1999; Anderson, 2003; Seber, 2004; Kotz & Nadarajan, 2004):

Lema 1-37 (Vínculo con vectores gaussianos)**.** *Sean* $X_i \sim \mathcal{N}(0,V)$*,* $i = 1,\ldots,n > d-1$ *independientes y la matriz* $S = \sum_{i=1}^{n} X_i X_i^t$ *llamada* matriz de dispersión *(scatter matrix en inglés). Entonces,* $S \in P_d^+(\mathbb{R})$ *(c. s.) (*S *es simétrica definida positiva casi siempre, o con probabilidad uno) y*

$$S = \sum_{i=1}^{n} X_i X_i^t \sim \mathcal{W}(V,n).$$

Eso de re-escribe de manera compacta bajo la forma

$$X X^t \sim \mathcal{W}(V,n) \qquad \textit{con} \qquad X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}$$

Demostración. De $\mathrm{Tr}\left(\omega^t S\right) = \sum_{i=1}^{n} \mathrm{Tr}\left(\omega^t X_i X_i^t\right) = \sum_{i=1}^{n} X_i^t \omega X_i^t$ (siendo ω simétrica), y de la independencia, tenemos inmediatamente

$$\Phi_S(\omega) = \prod_{i=1}^{n} \mathrm{E}\left[e^{\imath\, X_i^t \omega X_i^t}\right] = \left(\mathrm{E}\left[e^{\imath\, X_i^t \omega X_i^t}\right]\right)^n$$

Luego,

$$\begin{aligned}
\mathrm{E}\left[e^{\imath\, X_i^t \omega X_i^t}\right] &= \int_{\mathbb{R}^d} (2\pi)^{-\frac{d}{2}} |V|^{-\frac{1}{2}} e^{\imath\, x^t \omega x^t - \frac{1}{2} x^t V^{-1} x}\, dx \\
&= |V|^{-\frac{1}{2}} \left|V^{-1} - 2\imath\,\omega\right|^{-\frac{1}{2}} \int_{\mathbb{R}^d} (2\pi)^{-\frac{d}{2}} \left|\left(V^{-1} - 2\imath\,\omega\right)^{-1}\right|^{-\frac{1}{2}} e^{-\frac{1}{2} x^t \left(V^{-1} - 2\imath\,\omega\right) x}\, dx
\end{aligned}$$

Se puede probar que la integral vale uno (moralmente, integral de una densidad de probabilidad gaussiana) (Muirhead, 1982, Teo. 2.1.11). La prueba se cierra por $\left|V^{-1} - 2\imath\,\omega\right|^{-\frac{1}{2}} = |V|^{\frac{1}{2}} \left|I - 2\imath\,\omega V\right|^{-\frac{1}{2}}$: se reconoce en Φ_S la función característica de la ley de wishart (a condición que $n > d-1$). □

Se puede referirse también a (Gupta & Nagar, 1999; Anderson, 2003; Seber, 2004), o a (Bilodeau & Brenner, 1999, Ej. 7.2) para pruebas alternativas.

Este resultado permite también probar el lema **??** para $\nu = n$ entero escribiendo $X \overset{\mathrm{d}}{=} \sum_{i=1}^{n} X_i X_i^t$ así que $A^t X A \overset{\mathrm{d}}{=} \sum_{i=1}^{n} A^t X_i X_i^t A = \frac{1}{n} \sum_{i=1}^{n} \left(A^t X_i\right)\left(A^t X_i\right)^t$ y notando que los $A^t X_i \sim \mathcal{N}(0, A^t V A)$ son independientes (Seber, 2004). Además, permite re-obtener las expreciones del promedio y de las covarianzas [63]. Notar que cuando los X_i tienen un promemedio, el lema conduce a lo que es conocido como Wishart no central (Anderson, 2003; Seber, 2004).

Este lema se extiende a través de lo que aparece como combinaciones ortonormales (Muirhead, 1982; Gupta & Nagar, 1999; Bilodeau & Brenner, 1999; Anderson, 2003; Seber, 2004; Kotz & Nadarajan, 2004):

Lema 1-38 (Vínculo con vectores gaussianos y matriz idempotenta)**.** *Sean* $X_i \sim \mathcal{N}(0, V),\ i = 1, \ldots, n > d-1$ *independientes y denotamos* $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}$ *matriz de* $\mathcal{M}_{d,n}(\mathbb{R})$*. Sea* $A \in S_n(\mathbb{R})$ *idempotenta, es decir* $A^2 = A$ *de rango* $q > d-1$*. Entonces*

$$XAX^t \sim \mathcal{W}(V, q)$$

Demostración. A siendo idempotenta, existe una matriz $B \in \mathcal{M}_{n,q}(\mathbb{R})$ tal que [64] $A = BB^t$ y

---

[63] Para la covarianza, su usa la formula $\mathrm{E}[Y_1 Y_2 Y_3 Y_4] = \mathrm{E}[Y_1 Y_2]\,\mathrm{E}[Y_3 Y_4] + \mathrm{E}[Y_1 Y_3]\,\mathrm{E}[Y_2 Y_4] + \mathrm{E}[Y_1 Y_4]\,\mathrm{E}[Y_2 Y_3]$ para $Y = \begin{bmatrix} Y_1 & Y_2 & Y_3 & Y_4 \end{bmatrix}^t$ vector gaussiano, formula que se obtiene por ejemplo a partir de la función característica de un vecor gaussiano.

[64] A siendo idempotenta, si x es un autovector con autovalor λ tenemos $\lambda x = Ax = A^2 x = \lambda^2 x$, es decir $\lambda(\lambda - 1) = 0$: las autovalores son 0 o 1. Del rango q, y $A \in S_n(\mathbb{R})$ se diagonaliza bajo la forma $A = \begin{bmatrix} B & \widetilde{B} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} B^t \\ \widetilde{B}^t \end{bmatrix}$ con $B \in \mathcal{M}_{n,q}(\mathbb{R})$ y $\begin{bmatrix} B & \widetilde{B} \end{bmatrix}$ de columnas ortonormales (Harville, 2008, Teo. 21.5.7) o (Horn & Johnson, 2013).

$B^tB = I$ (Harville, 2008). Entonces,

$$XAX^t = YY^t \qquad \text{con} \qquad Y = XB = \begin{bmatrix} Y_1 & \cdots & Y_q \end{bmatrix}$$

Ahora, del teorema 1-42 los Y_i son gaussianos de media nula. Además

$$\begin{aligned} \mathrm{E}\left[Y_i Y_j^t\right] &= \sum_{k,l=1}^{n} B_{ki} B_{lj} \, \mathrm{E}\left[X_k X_l^t\right] \\ &= \left(\sum_{k=1}^{n} B_{ki} B_{kj}\right) V \end{aligned}$$

de la independencia de los X_k que son de covarianza V. De $B^tB = I$ *i. e.,* $\sum_{k=1}^{n} B_{ki}B_{kj} = \mathbb{1}_{\{i\}}(j)$ tenemos que los Y_i son independientes (gaussianos de covarianza nula para $i \neq j$), de covarianza V. La prueba se cierra del lema 1-37. □

En particular, la distribución, de Wishart aparece en la estimación de la matriz de covarianza de un vector gaussiano a partir de copias independientes de vectores gaussianos independientes de mismos parametros (Kotz & Nadarajan, 2004; Bilodeau & Brenner, 1999; Anderson, 2003; Seber, 2004; Gupta & Nagar, 1999):

Corolario 1-12. *Sean $X_i \sim \mathcal{N}(m, \Sigma)$, $i = 1, \ldots, n > d-1$ independientes, y sea la media empírica (ver corolario 1-10)*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sean las matrices aleatorias [65]

$$\overline{\Sigma}_m = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)(X_i - m)^t \qquad \textit{y} \qquad \overline{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \left(X_i - \overline{X}\right)^t$$

Entonces

$$\overline{\Sigma}_m \sim \mathcal{W}\left(\frac{1}{n}\Sigma, n\right) \qquad \textit{y} \qquad \overline{\Sigma} \sim \mathcal{W}\left(\frac{1}{n-1}\Sigma, n-1\right)$$

Demostración. Denotamos $\widetilde{X}_i = X_i - m \sim \mathcal{N}(0, \Sigma)$ independientes, $\widetilde{X} = \begin{bmatrix} \widetilde{X}_1 & \cdots & \widetilde{X}_n \end{bmatrix}$ matriz de $\mathcal{M}_{d,n}(\mathbb{R})$ y $\mathbb{1} \in \mathbb{R}^n$ vector de componentes iguales a 1. Entonces, denotando $\overline{\widetilde{X}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - m) = \frac{1}{n}\sum_{i=1}^{n} X_i - m$ tenemos $X_i - \overline{X} = \widetilde{X}_i - \overline{\widetilde{X}}$ y se muestra sencillamente que

$$\overline{\Sigma}_m = \frac{1}{n} \widetilde{X} \widetilde{X}^t \qquad \text{y} \qquad \overline{\Sigma} = \frac{1}{n-1} \widetilde{X} \left(I - \frac{\mathbb{1}\mathbb{1}^t}{n}\right) \widetilde{X}^t$$

Además, se proba sencillamente que $I - \frac{\mathbb{1}\mathbb{1}^t}{n} \in S_n(\mathbb{R})$ es idempotenta de rango $n-1$. El resultado es consecuencia de los lemas 1-35 (por la normalización) y de 1-37 y 1-38 respectivamente. □

---

[65] La primera matriz aparece por ejemplo cuando quereMemos estimar Σ, conociendo m, y la secinda cuando no se conoce m. En cada caso, la renormalización asegurida que la media del estimador es precisamente Σ, es decir que los estimadores son sin sesgo.

Las distribuciones de Wishart tienen varias propiedades más que se encuentren por ejemplo en los libros especialisados (Muirhead, 1982; Gupta & Nagar, 1999; Anderson, 2003; Seber, 2004; ?, ?). Entre estas, terminaremos esta sección con la factorización de una matriz de distribución de Wishart de la manera siguiente: sea X definido sobre $P_d^+(\mathbb{R})$; entonces, $\forall\, \omega,\quad X(\omega) \in P_d^+(\mathbb{R})$ puede escribirse por factorización de Cholesky, es decir $X(\omega) = T(\omega)T(\omega)^t$ con T triangular inferior con elementos positivos sobre su diagonal (ver nota de pie [60]). En el contexto de Wishart (de parametro (I, ν)), aparece que se puede caracterizar la distribución de los coefficientes. Eso es conocido como *descomposición de Bartlett* (Bartlett, 1934a; Muirhead, 1982; Bilodeau & Brenner, 1999; Gupta & Nagar, 1999; Anderson, 2003; ?, ?) y se formaliza en el caso Wishart general de la manera siguiente:

Teorema 1-52 (Descomposición de Bartlett)**.** *Sea $X \sim \mathcal{W}(V, \nu)$ y $V = LL^t$ factorización de Cholesky de V con L triangular inferior. Entonces, tenemos*

$$X \overset{\mathrm{d}}{=} LUU^tL^t$$

con U de componentes independientes, $U_{ii} > 0$, tales que

$$\begin{cases} U_{ij} = 0 & \textit{si} \quad j > i \ \textit{(}T\ \textit{triangular inferior)} \\ U_{ij} \sim \mathcal{N}(0,1) & \textit{si} \quad j < i \\ U_{ii}^2 \sim \mathcal{G}\left(\frac{\nu-i+1}{2}\,,\, \frac{1}{2}\right) & \end{cases}$$

Demostración. Primero, se nota del lema 1-35 que $X \overset{\mathrm{d}}{=} LYL^t$ con $Y \sim \mathcal{W}(I, \nu)$, de densidad $p_Y(y) = \frac{|y|^{\frac{\nu-d-1}{2}} e^{-\frac{1}{2}\operatorname{Tr}(y)}}{2^{\frac{d\nu}{2}}\Gamma_d\left(\frac{\nu}{2}\right)}$. Ahora, sea la factorización de Cholesky de $Y = UU^t$ (*i. e.,* para cada ω se factoriza $Y(\omega)$ dando $U(\omega)$) y la transformaciíon $g : Y \mapsto U$. Más precisamente, la densidad p_X siendo la de los $\frac{d(d+1)}{2}$ componentes diferentes de X (parte triangular inferior), $g : (X)_{1\le j\le i\le d} \mapsto (U)_{1\le j\le i\le d}$. Según el teorema 2.1.9 de (Muirhead, 1982), el valor absoluto del determinente de la jacobiana de g^{-1} es dado por

$$\left|\mathrm{J}_{g^{-1}}(u)\right| = 2^d \prod_{i=1}^{d} u_{ii}^{d-i+1}$$

(ver también la prueba del teorema 7.2.1 de (Anderson, 2003) o (Bilodeau & Brenner, 1999, Prop. 2.22)).

Se nota ahora que siendo u triangular inferior,

$$\operatorname{Tr}(x) = \sum_{i=1}^{d} x_{ii} = \sum_{1\le j\le i\le d} u_{ij}^2, \qquad |x| = |u|^2 = \prod_{i=1}^{d} u_{ii}^2.$$

Del teorema 1-12 obtenemos para $u_{ii} > 0$ (ver notaciones para la escritura de Γ_d función gamma

multivariada, como producto de funciones Gamma)

$$\begin{aligned} p_U(u) &= \left|\mathrm{J}_{g^{-1}}(u)\right| p_X(uu^t) \\ &= 2^d \prod_{i=1}^{d} u_{ii}^{d-i+1} \frac{\prod_{i=1}^{d} u_{ii}^{\nu-d-1} e^{-\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{i} u_{ij}^2}}{2^{\frac{d\nu}{2}} \pi^{\frac{d(d-1)}{4}} \prod_{i=1}^{d} \Gamma\left(\frac{\nu-i+1}{2}\right)} \\ &= \left(\prod_{i=1}^{d} \frac{2^{1-\frac{\nu-i+1}{2}} u_{ii}^{\nu-i} e^{-\frac{1}{2}t_{ii}^2}}{\Gamma\left(\frac{\nu-i+1}{2}\right)}\right) \left(\prod_{1\le j<i\le d} \frac{e^{-\frac{t_{ij}^2}{2}}}{\sqrt{2\pi}}\right) \end{aligned}$$

Por productos, se concluye que las conmponentes son independientes. Luego, en el secundo producto se reconoce un producto de leyes gaussianas estandares, así que los $U_{ij},\ 1 \le j < i \le d$ son gaussianas estandares. Al final, la densidad de U_{ii} siendo $p_{U_{ii}}(u) = \frac{2^{1-\frac{\nu-i+1}{2}} u^{\nu-i} e^{-\frac{1}{2}u^2}}{\Gamma\left(\frac{\nu-i+1}{2}\right)}$, por transformacíon (ver corolario 1-2), la densidad de U_{ii}^2 es $p_{U_{ii}^2}(v) = \frac{2^{-\frac{\nu-i+1}{2}} v^{\frac{\nu-i+1}{2}-1} e^{-\frac{1}{2}v}}{\Gamma\left(\frac{\nu-i+1}{2}\right)}$: es la ley Gamma $\mathcal{G}\left(\frac{\nu-i+1}{2}, \frac{1}{2}\right)$ (ver subsección 1.10.2.5). □

Nota: la distribución de U_{ii} es a veces llamada *raiz-gamma*.

Esta descomposición permite de sortear sencillamente matrices de distribuciones de Wishart, con grado de libertad no necesariamente entero [66].

1.10.2.7. Distribución beta

Estas distribuciones fueron popularizadas por Pearson en los años 1895 (Pearson, 1895, 1916; David & Edwards, 2001) bajo la denominación Pearson tipo I en su estudio de la teoria de la evolución y la modelización con variables asimétricas. De hecho, apareció mucho tiempo antes, en trabajos de Bayes publicado en un papel postumo por R. Price en 1763 (Bayes, 1763). Aparentemente, la denominación estandar "beta" es debido al estadístico, demógrafo y sociólogo italiano C. Gini en 1911 su estudio del "sex ratio" (desequilibrio entre los nacimientos de muchachos/muchachas) con un enfoque bayesiano (Gini, 1911; Forcina, 2017; David & Edwards, 2001). La distribución beta aparece precisamente, entre otros, en problema de inferencia bayesiana como distribución a priori conjugado del parámetro p de la ley binomial (Robert, 2007) (ver notas de pie 51 y 59).

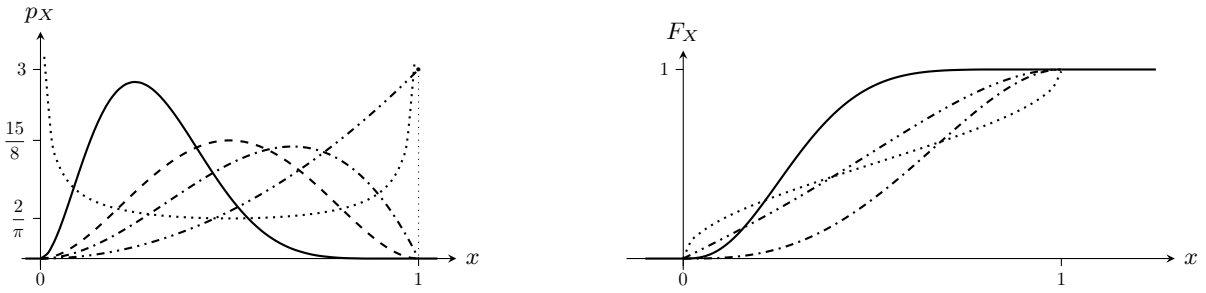
Se denota $X \sim \beta(a,b)$ con $(a,b) \in \mathbb{R}_+^{*2}$ llamados *parámetros de forma*. Las características son:

[66] De hecho, en la literatura, esta descomposición aparece casi siempre con ν entero, pero la prueba no necesite este vínculo.

Dominio de definición	$\mathcal{X} = [0\,;\,1]$
Parámetros	$(a, b) \in \mathbb{R}_+^{*\,2}$ (forma)
Densidad de probabilidad	$p_X(x) = \dfrac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$

Promedio	$m_X = \dfrac{a}{a+b}$
Varianza	$\sigma_X^2 = \dfrac{ab}{(a+b)^2(a+b+1)}$
Asimetría	$\gamma_X = \dfrac{2\,(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{6\,\big((a-b)^2(a+b+1) - ab(a+b+2)\big)}{a\,b\,(a+b+2)\,(a+b+3)}$
Generadora de momentos	$M_X(u) = {}_1F_1\,(a, a+b\,;\, u)$ para $u \in \mathbb{C}$
Función característica	$\Phi_X(\omega) = {}_1F_1\,(a, a+b\,;\, \imath\omega)$

Unas densidades de probabilidad y funciones de repartición asociadas son representadas en la figura Fig. 1-29 para varios a y b.



(a) (b)

Figura 1-29: Ilustración de una densidad de probabilidad beta (a), y la función de repartición asociada (b). $(a, b) = (0{,}5\,,\, 0{,}5)$ (linea punteada), $(3\,,\, 1)$ (linea mixta doble punteada), $(3\,,\, 2)$ (linea mixta), $(3\,,\, 3)$ (linea guionada), $(3\,,\, 7)$ (linea llena).

Notar que se recupera la ley uniforme sobre $[0\,;\, 1]$ para $a = b = 1$. Se conoce la ley de $Y = 2\,B - 1$ con $B \sim \beta\left(\frac{1}{2}, \frac{1}{2}\right)$ como *ley arcseno*.

Variables beta tienen también unas propiedades notables. Primero, por cambio de variables, se demuestra el lema siguiente:

Lema 1-39 (Reflexividad)**.** *Sea* $X \sim \beta(a, b)$*. Entonces*

$$1 - X \sim \beta(b, a)$$

Lema 1-40 (Un vínculo con la ley exponencial)**.** *Sea* $X \sim \beta(a, 1)$*. Entonces*

$$-\log X \sim \mathcal{E}(a)$$

Demostración. El resultado es inmediato de la fórmula de transformación del corolario 1-2. □

Lema 1-41 (Un vínculo con la ley uniforme)**.** *Sea* $X \sim \mathcal{U}([0\,;\,1])$ *y* $a > 0$*. Entonces*

$$U^{\frac{1}{a}} \sim \beta(a, 1)$$

Demostración. El resultado es inmediato de la fórmula de transformación del corolario 1-2. □

Lema 1-42 (Un vínculo con la ley gamma)**.** *Sea* $X \sim \mathcal{G}(a, c)$ *e* $Y \sim \mathcal{G}(b, c)$ *independientes. Entonces*

$$\frac{X}{X+Y} \sim \beta(a, b)$$

(independientemente de c*). Además,* $\frac{X}{X+Y}$ *y* $X + Y$ *son independientes.*

Demostración. La independencia de c es obvia del hecho de que para cualquier $\theta > 0$, $\theta^{-1}X \sim \mathcal{G}(a, \theta c)$ e $\theta^{-1}Y \sim \mathcal{G}(b, \theta c)$, la independencia con respeto a c viniendo de $\frac{\theta^{-1}X}{\theta^{-1}X+\theta^{-1}Y} = \frac{X}{X+Y}$. Entonces, se puede considerar $c = 1$ sin perdida de generalidad. Ahora, sea la transformación

$$\begin{array}{rccl} g\ : & \mathbb{R}_+^2 & \mapsto & [0\,;\,1] \times \mathbb{R}_+ \\ & (x, y) & \to & (u, v) = \left(\frac{x}{x+y}\,,\, x + y\right) \end{array}$$

Entonces, la transformación inversa se escribe

$$g^{-1}(u, v) = (uv\,,\, (1 - u)v)$$

de matriz Jacobiana

$$\mathrm{J}_{g^{-1}} = \begin{bmatrix} v & u \\ -v & 1-u \end{bmatrix}$$

Del teorema de cambio de variables teorema **??**, notando que $\left|\mathrm{J}_{g^{-1}}\right| = v$ y de la independencia de X e Y, se obtiene para el vector aleatorio $W = \begin{bmatrix} U & V \end{bmatrix}^t$ la densidad de probabilidad, definida sobre $[0\,;\,1] \times \mathbb{R}_+$, como

$$\begin{aligned} p_W(u, v) &= p_X(uv)\, p_Y((1-u)v)\, v \\ &= \frac{(uv)^{a-1}\ e^{-uv}}{\Gamma(a)} \times \frac{((1-u)v)^{b-1}\ e^{-(1-u)v}}{\Gamma(b)} \times v \\ &= \frac{u^{a-1}(1-u)^{b-1}}{B(a, b)} \times \frac{v^{a+b-1}e^{-v}}{\Gamma(a+b)} \end{aligned}$$

Inmediatamente, factorizandose, aparece claramente que U y V son independientes. Además, se reconoce en el primer factor la densidad beta de parámetros (a, b). Pasando, se recupera el hecho que $X + Y \sim \mathcal{G}(a + b, 1)$. □

Lema 1-43 (Stabilidad por producto)**.** *Sea* $X \sim \beta(a, b)$ *e* $Y \sim \beta(a + b, c)$ *independientes. Entonces*

$$XY \sim \beta(a, b + c)$$

Demostración. Sean $U \sim \mathcal{G}(a,1)$, $V \sim \mathcal{G}(b,1)$ y $W \sim \mathcal{G}(c,1)$ independientes y sean $X = \frac{U}{U+V}$, $Y = \frac{U+V}{U+V+W}$ y $Z = U+V+W$. Del lema anterior $X \sim \beta(a,b)$ y $Y \sim \beta(a+b,c)$. Sea la transformación

$$\begin{array}{rccl} g: & \mathbb{R}_+^3 & \mapsto & [0\,;\,1]^2 \times \mathbb{R}_+ \\ & (u,v,w) & \to & (x,y,z) = \left(\frac{u}{u+v}\,,\,\frac{u+v}{u+v+w}\,,\,u+v+w\right) \end{array}$$

Entonces, la transformación inversa se escribe

$$g^{-1}(x,y,z) = (xyz\,,\,(1-x)yz\,,\,z(1-y))$$

de matriz Jacobiana

$$\mathrm{J}_{g^{-1}} = \begin{bmatrix} yz & xz & xy \\ -yz & (1-x)z & (1-x)y \\ 0 & -z & 1-y \end{bmatrix}$$

De nuevo, del teorema de cambio de variables teorema **??**, notando que $\left|\mathrm{J}_{g^{-1}}\right| = yz^2$ y de la independencia de U,V,W, se obtiene para el vector aleatorio $T = \begin{bmatrix} X & Y & Z \end{bmatrix}^t$ la densidad de probabilidad probabilidad

$$\begin{aligned} p_T(x,y,z) &= p_u(xyz)\,p_V((1-x)yz)\,p_W(y(1-z))\,yz^2 \\ &= \frac{(xyz)^{a-1}\,e^{-xyz}}{\Gamma(a)} \times \frac{((1-x)yz)^{b-1}\,e^{-(1-x)yz}}{\Gamma(b)} \times \frac{(z(1-y))^{c-1}\,e^{-z(1-y)}}{\Gamma(c)} \times yz^2 \\ &= \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \times \frac{y^{a+b-1}(1-y)^{c-1}}{B(a+b,c)} \times \frac{z^{a+b+c-1}e^{-z}}{\Gamma(a+b+c)} \end{aligned}$$

Eso proba que X, Y y Z son independientes (las densidades se factorizan). Además,

$$XY = \frac{U}{U+V} \times \frac{U+V}{U+V+W} = \frac{U}{U+V+W} \sim \beta(a,b+c)$$

el último resultado como consecuencia de los lemas 1-42 y 1-33. Eso cierra la prueba. □

Lema 1-44 (Ley gamma como caso límite de beta)**.** *Sea* $X_n \sim \beta(a,n)$. *Entonces*

$$nX_n \xrightarrow[n\to+\infty]{\mathrm{d}} X \sim \mathcal{G}(a,1)$$

con $\xrightarrow{\mathrm{d}}$ *límite es en distribución*

Demostración. De la fórmula de transformación tenemos la distribución de nX_n

$$\begin{aligned} p_{nX_n}(x) &= \frac{1}{n}, \frac{\left(\frac{x}{n}\right)^{a-1}\left(1-\frac{x}{n}\right)^{n-1}}{B(a,n)}\,\mathbb{1}_{(0\,;\,1)}\left(\frac{x}{n}\right) \\ &= \frac{x^{a-1}}{\Gamma(a)}\,\frac{\Gamma(n+a)}{n^a\Gamma(n)}\,\left(1-\frac{x}{n}\right)^{n-1}\,\mathbb{1}_{(0\,;\,n)(x)} \end{aligned}$$

El resultado sigue notando que $\mathbb{1}_{(0\,;\,n)} \to \mathbb{1}_{\mathbb{R}_+^*}$, $\left(1-\frac{x}{n}\right)^{n-1} \to e^{-x}$ y de la fórmula de Stirling (ver sección 1.10.1.11). □

La distribución beta se generaliza al caso matriz-variada X definido sobre $\mathcal{X}$ tal que X y $I - X$ partenecen a $P_d^+(\mathbb{R})$; se denota $X \sim \beta_d(a, b)$ donde $(a, b) \in \mathbb{R}_+^{*2}$ y la densidad est dada por $p_X(x) = \frac{|x|^{a-\frac{d+1}{2}}|I-x|^{b-\frac{d+1}{2}}}{B_p([a \quad b]^t)}, \quad (a, b) \in \left(\frac{d-1}{2}; +\infty\right)^2$. Se refiera a (Gupta & Nagar, 1999, Cap. 5) para tener más detalles.

1.10.2.8. Distribución de Dirichlet

Esta distribución tiene su nombre de integrales on a simplex estudiados por M. Lejeune-Dirichlet y J. Liouville en 1839 (Gupta & Richards, 2001; Lejeune-Dirichlet, 1839; Liouville, 1839). Es una extensión multivariada de las variables beta a veces conocida como *beta multivariada* (Olkin & Rubin, 1964). Escribiendo la forma de la distribución solamente con la variables x_i, la integral permitiendo normalizarla es precisamente la estudiada por Lejeune-Dirichlet y Liouville.

Se nota $X \sim \mathcal{D}ir(a)$ con $a \in \mathbb{R}_+^{*k}$ y X vive sobre el $(k-1)$-simplex estandar Δ_{k-1}. a es llamado parámetro de forma. Como en el caso de vectores de distribución multinomial, a pesar de que se escribe X de manera k-dimensional, el vector partenece a una variedad $d = k - 1$ dimensional y en el caso $k = 2$ se recupera la ley beta. A veces se parametriza la ley con un parámetro escalar $\alpha > 0$ y un vector del simplex estandar $\bar{a} \in \Delta_{k-1}$ tal que

$$a = \alpha\bar{a}, \quad \textit{i. e.,} \quad \alpha = \sum_{i=1}^{k} a_i, \quad \bar{a} = \frac{a}{\alpha}$$

α es conocido como parámetro de *concentración* y el vector $\bar{a}$ como *medida de base*.

Las características de un vector de Dirichlet son:

Dominio de definición	$\mathcal{X} = \Delta_{k-1},\ k \in \mathbb{N} \setminus \{0\,;\,1\}$
Parámetros	$a = \alpha\,\bar{a} \in \mathbb{R}_+^{*\,k}$ (forma) con $\alpha \in \mathbb{R}_+^*$ (concentración) y $\bar{a} \in \Delta_{k-1}$ (medida de base)
Densidad de probabilidad [67]	$p_X(x) = \dfrac{\prod_{i=1}^k x_i^{a_i-1}}{B(a)}$
Promedio	$m_X = \bar{a}$
Covarianza [68]	$\Sigma_X = \dfrac{\operatorname{diag}(\bar{a}) - \bar{a}\bar{a}^t}{1+\alpha}$
Generadora de momentos	$M_X(u) = \Phi_2^{(k)}(a, \alpha\,;\,u)$ para $u \in \mathbb{C}$
Función característica	$\Phi_X(\omega) = \Phi_2^{(k)}(a, \alpha\,;\,\imath\omega)$

De nuevo, se puede considerar que el vector aleatorio es $(k-1)$-dimensional $\widetilde{X} = \begin{bmatrix}\widetilde{X}_1 & \cdots & \widetilde{X}_{k-1}\end{bmatrix}^t$ definido sobre el hipertriangulo $\widetilde{\mathcal{X}} = \mathcal{T}_{k-1} = \left\{\widetilde{x} \in [0\,;\,1]^{k-1} \;\middle|\; \sum_{i=1}^{k-1}\widetilde{x}_i \leq 1\right\}$, proyección del simplex sobre el hiperplano $x_k = 0$. Así, $\widetilde{X}$, tiene una densidad con respeto a la medida de Lebesgue usual, y es dada por $p_{\widetilde{X}}(\widetilde{x}) = \frac{\prod_{i=1}^{k-1}\widetilde{x}_i^{a_i-1}\left(1-\sum_{i=1}^{k-1}\widetilde{x}_i\right)^{a_k-1}}{B(a)}$. A final, se notará que $\Phi_{\widetilde{X}}(\widetilde{\omega}) = \Phi_X\left(\begin{bmatrix}\widetilde{\omega} & 0\end{bmatrix}^t\right)$ y $\Phi_X(u) = e^{\omega_k}\Phi_{\widetilde{X}}\left(\begin{bmatrix}\omega_1-\omega_k & \cdots & \omega_{k-1}-\omega_k\end{bmatrix}^t\right)$ (y similarmente para G_X con respeto a $G_{\widetilde{X}}$). Notar también que, la forma de la función generadora de momento viene directamente de la escritura de las series de Taylor de $e^{u_i x_i}$ o de la forma integral de la función confluente hipergeométrica (Phillips, 1988).

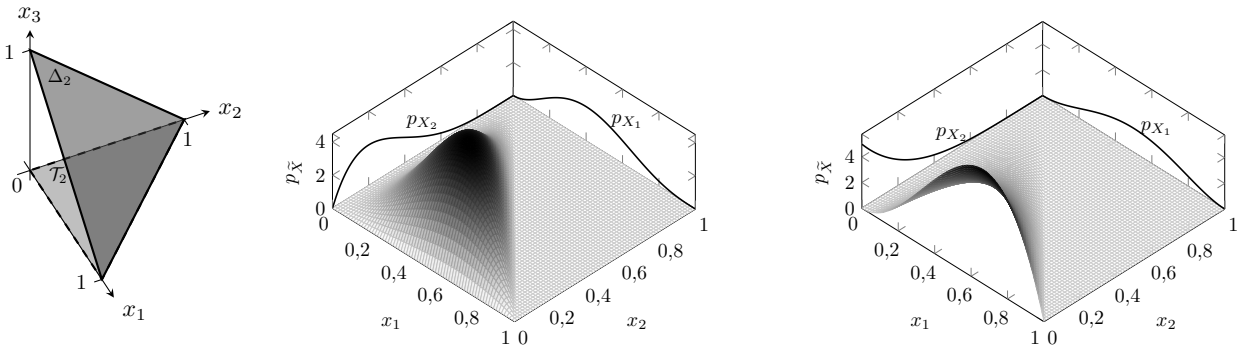
Naturalmente, $\Sigma_X\mathbb{1} = 0$ así que de nuevo $\Sigma_X \notin P_k^+(\mathbb{R})$, como consecuencia directa del hecho que X k-dimensional, vive sobre Δ_{k-1}, $(k-1)$-dimensional. De nuevo, para definir asimetría y curtosis habría que considerar $\widetilde{X}$, de promedio $\begin{bmatrix}\bar{a}_1 & \cdots & \bar{a}_{k-1}\end{bmatrix}^t$ y de covarianza el bloque $(k-1)\times(k-1)$ de Σ_X, que es ahora invertible. $\gamma_{\widetilde{X}}$ y $\kappa_{\widetilde{X}}$ son bien definidos. Las expresiones, demasiado pesadas, no son dadas acá.

La figura Fig. 1-30 representa el dominio de definición del vector (a) y la densidad de probabildad con las marginales (ver más adelante) para $k = 3$ y dos ejemplos de parámetro a.

Vectores de distribución de Dirichlet tienen también unas propiedades notables, parecidas a las de la beta:

[67] La densidad de probabilidad es dada con respeto a la medida de Lebesgue restricta al simplex Δ_{k-1}.

[68] Ver nota de pie **??**.



(a) (b) (c)

Figura 1-30: Ilustración del dominio Δ_{k-1} de definición de la ley de Dirichlet para $k = 3$ (grise oscuro), con el dominio $(k-1)$-dimensional $\mathcal{T}_{k-1}$ del vector $\widetilde{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^t$ $(X_3 = 1 - X_1 - X_2)$ (grise claro) (a), y densidad de probabilidad de $\widetilde{X}$ con las marginales p_{X_1}, p_{X_2}. Los parámetros son $a = \begin{bmatrix} 3 & 2 & 2 \end{bmatrix}^t$ (b) y $a = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix}^t$ (c).

Lema 1-45 (Reflexividad)**.** *Sea* $X \sim \mathcal{D}ir(a)$, $a \in \mathbb{R}_+^{*k}$ *y* $\Pi \in \mathfrak{S}_k(\mathbb{R})$ *matriz de permutación. Entonces*

$$\Pi X \sim \mathcal{D}ir\left(\Pi a\right)$$

Demostración. El resultado es inmediato por cambio de variables $x \to \Pi x$, la Jacobiana siendo Π, de valor absoluto determinente igual a 1 (ver sección 1.4). □

Además, se muestra una stabilidad reemplazando dos componentes por su suma:

Lema 1-46 (Stabilidad por agregación)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_k \end{bmatrix}^t \sim \mathcal{D}ir(a)$, $a = \begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix}^t \in \mathbb{R}_+^{*k}$ *y* $G^{(i,j)}$ *matriz de agrupación de las* (i,j)*-ésima componentes (ver notaciones). Entonces,*

$$G^{(i,j)} X \sim \mathcal{D}ir\left(G^{(i,j)} a\right)$$

Demostración. Se puede probar este resultado a partir de la función característica, usando las propiedades de la función confluent hipergeométrica (Srivastava & Karlsson, 1985; Humbert, 1922; Appell, 1925; ?, ?, ?; Erdélyi, 1940). Pero se puede también tener un enfoque más directo. Del lema precediente, notando que existen matrices de permutación [69] $\Pi_k \in \mathfrak{S}_k(\mathbb{R})$ y $\Pi_{k-1} \in \mathfrak{S}_{k-1}(\mathbb{R})$ tal que $G^{(i,j)} = \Pi_{k-1} G^{(1,2)} \Pi_k$, se puede concentrarse en el caso $(i,j) = (1,2)$. Sea el cambio de variables $g : x = (x_1, \ldots, x_k) \mapsto u = (u_1, \ldots, u_k) = (x_1, x_1 + x_2, x_3, \ldots, x_k)$. Entonces $g^{-1}(u) = (u_1, u_2 - u_1, u_3, \ldots, u_k)$ es de determinente de matriz Jacobiana igual a 1 dando para

[69] Π_k pone las componentes i e j el las posiciones 1 y 2, sin cambiar el orden de las siguientes; Π_{k-1} trazlada la primera componente en la posición $\mín(i,j)$.

$U = g(X)$ la densidad

$$p_U(u) = \frac{u_1^{a_1-1}\,(u_2-u_1)^{a_2-1}\prod_{i=3}^k u_i^{a_i-1}}{B(a)}$$

sobre $g\,(\Delta_{k-1})$. Para $u_2 \in [0\,;\,1]$ tenemos $u_1 \in [0\,;\,u_2]$ así que, por marginalización en u_1 obtenemos la densidad

$$\begin{aligned} p_{G^{(1,2)}X}(u_2,\ldots,u_k) &= \frac{\prod_{i=3}^k u_i^{a_i-1}}{B(a)} \int_0^{u_2} u_1^{a_1-1}\,(u_2-u_1)^{a_2-1}\,du_1 \\ &= \frac{\prod_{i=3}^k u_i^{a_i-1}}{B(a)}\, u_2^{a_1+a_2-1} \int_0^1 v_1^{a_1-1}\,(1-v_1)^{a_2-1}\,dv_1 \end{aligned}$$

con el cambio de variables $u_1 = u_2 v_1$. Se cierra la prueba notando que la integral vale $B(a_1,a_2)$ y que $\frac{B(a_1,a_2)}{B(a)} = \frac{1}{B\left(G^{(1,2)}a\right)}$. □

De este lema, aplicado de manera recursiva, se obtiene en corolario siguiente:

Corolario 1-13. *Sea $X \sim \mathcal{D}ir(a)$, entonces $X_i \sim \beta\,(a_i\,,\,\alpha - a_i)$.*

Naturalmente, la ley de Dirichelt siendo una extensión de la ley beta, existe también un vínculo entre esta ley y variables de distribución gamma:

Lema 1-47 (Vínculo con la ley gamma)**.** *Sea X vector k-dimensional de componentes $X_i \sim \mathcal{G}(a_i,c)$, $i=1,\ldots,k$ independientes y a vector de componente i-ésima a_i. Entonces*

$$\frac{X}{\sum_{i=1}^k X_i} \sim \mathcal{D}ir(a)$$

(independientemente de c). Además, $\frac{X}{\sum_{i=1}^k X_i}$ y $\sum_{i=1}^k X_i$ son independientes.

Demostración. La prueba sigue exactamente los mismos pasos que la del lema 1-42 trabajando con $\widetilde{X}$. □

Naturalmente, la distribución de Dirichlet, extensión de la ley beta, aparece entre otros en problema de inferencia bayesiana como distribución a priori conjugado del parámetro p de la ley multinomial (Robert, 2007), extensión de la ley binomial.

La distribución de Dirichlet se generaliza al caso matriz-variada X definido sobre $\mathcal{P}_{d,k}(\mathbb{R})$, conjuntos de k-uplet de matrices de $P_d^+(\mathbb{R})$ cumpliando la relación de completud (ver notaciones); se denota $X \sim \mathcal{D}ir_d(a)$ donde $a \in \left(\frac{d-1}{2}\,;\,+\infty\right)^k$ la densidad est dada por $p_X(x) = \frac{\prod_{i=1}^k |x_i|^{a_i-\frac{d+1}{2}}}{B_d(a)}$. Se refiera a (Gupta & Nagar, 1999, Cap. 6) para tener más detalles.

1.10.2.9. Distribución Student-t multivariada

En el caso escalar, esta ley fue introducida inicialmente por F. R. Helmert (Helmert, 1875, 1876; Sheynin, 1995) y J. Lüroth (Lüroth, 1876; Pfanzagl, 1996). Pero es más conocida por su introducción

por William Sealy Gosset [70] en 1908, trabajando sobre variables centradas normalizadas por el promedio y varianza empíricos (Student, 1908). Fue estudiada entre otros intensivamente por el famoso matemático R. Fisher (Fisher, 1925a). En la literatura, esta ley es conocida bajo los nombres *Student*, *Student-*t o simplemente t*-distribución* o aún bajo el nombre *Pearson tipo IV* en el caso escalar y *Pearson tipo VII* (para $\frac{\nu+d}{2}$ entero; ver más abajo), debido a la familia de Pearson (Pearson, 1895; Johnson et al., 1995a, 1995a; Kotz et al., 2000; Fang et al., 1990). Esta distribución aparece como a priori conjugado de la media de una gaussiana en inferencia bayesiana (Robert, 2007; Kotz & Nadarajan, 2004).

Se denota con $X \sim \mathcal{T}_\nu(m, \Sigma)$ con $m \in \mathbb{R}^d$, $\Sigma \in P_d^+(\mathbb{R})$ conjunto de las matrices de $\mathcal{M}_{d,d}(\mathbb{R})$ simétricas definidas positivas. m es llamado *parámetro de posición* (no es la media que puede no existir), Σ es llamada *matriz característica* (no es [proporcional a] la covarianza que puede no existir) y $\nu > 0$ llamado *grado de libertad*. Las características de una Student-t son las siguientes:

Dominio de definición	$\mathcal{X} = \mathbb{R}^d$
Parámetro	$\nu \in \mathbb{R}_+^*$ (grado de libertad), $m \in \mathbb{R}^d$ (posición), $\Sigma \in P_d^+(\mathbb{R})$ (matriz característica)
Densidad de probabilidad	$p_X(x) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\pi^{\frac{d}{2}}\nu^{\frac{d}{2}}\Gamma\left(\frac{\nu}{2}\right)\,\lvert\Sigma\rvert^{\frac{1}{2}}}\left(1+\frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu}\right)^{-\frac{\nu+d}{2}}$
Promedio	$m_X = m$ si $\nu > 1$; no existe si no [71].
Covarianza [72]	$\Sigma_X = \frac{\nu}{\nu-2}\,\Sigma$ si $\nu > 2$; no existe si no [71].
Asimetría	$\gamma_X = 0$ si $\nu > 3$; no existe si no [71].
Curtosis por exceso	$\bar{\kappa}_X = \frac{2}{\nu-4}\sum_{i,j=1}^{d}\left(\left(\mathbb{1}_i\mathbb{1}_i^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_j^t\right)+\left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_i\mathbb{1}_j^t\right)+\left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_i^t\right)\right)$ si $\nu > 4$; no existe si no [71].
Función característica [73]	$\Phi_X(\omega) = \frac{\nu^{\frac{\nu}{4}}}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{\nu}{2}\right)}\,e^{\imath\omega^t m}\,\left(\omega^t\Sigma\omega\right)^{\frac{\nu}{4}}\,K_{\frac{\nu}{2}}\left(\sqrt{\nu\,\omega^t\Sigma\omega}\right)$

[70] De hecho, W. S. Gosset fue un estudiante trabajando en la fábrica de cerveza irlandesa Guiness sobre estadísticas relacionadas a la química de la cerveza. Hay varias versiones sobre el hecho que se publicó este trabajo bajo el nombre "Student". Una es que fue para que no se sabe que la fábrica estaba trabajando sobre estas estadísticas para estudiar la calidad de la cerveza (Wendl, 2016).

[71] De manera general, esta ley admite momentos de orden k si y solamente si $\nu > k$.

Nota: nuevamente se puede escribir $X \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}} S + m$ con $S \sim \mathcal{T}_\nu(0, I)$ donde S es dicha *Student-t estandar* y las características de X son vínculadas a las de S (y vice-versa) por transformación lineal (ver secciones anteriores).

Densidades de probabilidad Student-t estandar y funciones de repartición asociadas en el caso escalar son representadas en la figura Fig. 1-31-(a) y (b) para varios ν, y una densidad en un contexto bi-dimensional figura Fig. 1-31(c).

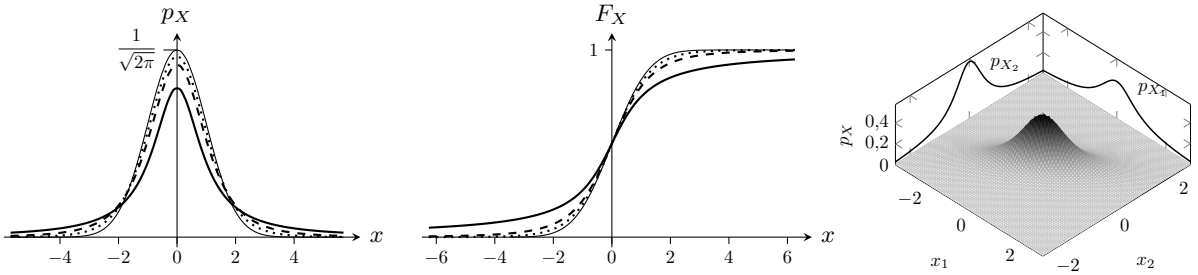


Figura 1-31: Ilustración de una densidad de probabilidad Student-t escalar estandar (a), y la función de repartición asociada (b) con $\nu = 1$ (linea llena), $\nu = 3$ (linea guionada), $\nu = 7$ (linea punteada) y $\nu \to +\infty$ (linea llena fina; ver más adelante) grado de libertad, así que una densidad de probabilidad Student-t bi-dimensional con $\nu = 1$ grado de libertad, centrada, y de matriz característica $\Sigma = R(\theta)\Delta^2 R(\theta)^t$ con $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ matriz de rotación y $\Delta = \operatorname{diag}\left(\begin{bmatrix} 1 & a \end{bmatrix}\right)$ matriz de cambio de escala, y sus marginales $X_1 \sim \mathcal{T}_\nu\left(0, \cos^2\theta + a^2\sin^2\theta\right)$ y $X_2 \sim \mathcal{T}_\nu\left(0, \sin^2\theta + a^2\cos^2\theta\right)$ (ver más adelante). En la figura, $a = \frac{1}{3}$ y $\theta = \frac{\pi}{6}$.

Nota: el caso $\nu = 1$ es conocido como distribución de *Cauchy* o *Cauchy-Lorentz* o *Lorentzian* o *Breit-Wigner* (Cauchy, 1853a, 1853b; ?, ?; Bienaymé, 1853b; Breit & Wigner, 1936; Stigler, 1974; Samorodnitsky & Taqqu, 1994; ?, ?). Es un caso particular también de distribución α-estables (Samorodnitsky & Taqqu, 1994). En particular, una combinación lineal de variables de Cauchy independientes queda de Cauchy. Pero, no viola el teorema del límite central del hecho de que una variable de Cauchy no admite covarianza.

Contrariamente al caso gaussiano, de la forma de la densidad de probabilidad, es claro que si la matriz Σ es diagonal, la densidad no factoriza, así que las componentes del vector no son indepen-

[72] Fijense de que Σ no es la covarianza, pero es proporcional a la covarianza. . . cuando existe. Se podría imaginar renormalizar la ley tal que Σ_X y Σ coinciden, pero no sería posible en el caso $\nu \leq 2$.

[73] Se muestra sencillamente que la función generatriz de momentos puede existir si y solamente si $\Re e\{u\} = 0$. La función generadora de momentos restricta al producto cartesiano de rectas $\Re e\{u\} = 0$ es nada más que la función característica. Además, esta función fue calculdada, especialmente en el caso multivariado, relativamente recientemente (Sutradhar, 1986; Hurst, 1995; Kibria & Joarder, 2006; Song, Park & Kim, 2014).

dientes. Este ejemplo muestra claramente que la reciproca del lema 1-6 es falsa en general.

Sin embargo, las distribuciones Student-t tienen varias propiedades notables.

Lema 1-48 (Stabilidad por transformación lineal)**.** *Sea* $X \sim \mathcal{T}_\nu(m, \Sigma)$, A *matriz de* $\mathcal{M}_{d',d}(\mathbb{R})$ *con* $d' \leq d$*, y de rango lleno y* $b \in \mathbb{R}^{d'}$*. Entonces*

$$AX + b \sim \mathcal{T}_\nu(Am + b, A\Sigma A^t)$$

En particular los componentes de X *son student-*t*,*

$$X_i \sim \mathcal{T}_\nu(m_i, \Sigma_{i,i})$$

Demostración. La prueba es inmediata usando la función característica y sus propiedades por transformación lineal. La condición sobre A es necesaria y suficiente para que $A\Sigma A^t \in P_{d'}^+(\mathbb{R})$. □

Lema 1-49 (Vínculo con las distribuciones gamma y gaussiana (mezcla de escala gaussiana))**.** *Sea* $V \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ *y* $G \sim \mathcal{N}(0, I)$ *con* $\nu > 0$ *y* G d*-dimensional e independiente de* V*. Entonces, para* $\Sigma \in P_d^+(\mathbb{R})$ *y* $m \in \mathbb{R}^d$*,*

$$\frac{\Sigma^{\frac{1}{2}} G}{\sqrt{V}} + m \sim \mathcal{T}_\nu(m, \Sigma)$$

Demostración. Sea $X = \frac{G}{\sqrt{V}}$. De la nota siguiendo la tabla de características es necesario y suficiente probar que $X \sim \mathcal{T}_\nu(0, I)$. Lo más simple es de salir de la formula de probabilidad total del teorema 1-16, notando que de la independencia, condicionalmente a $V = v$ la variable es gaussiana de covarianza $\frac{1}{v} I$,

$$p_{G|V=v}(x) = (2\pi)^{-\frac{d}{2}} v^{\frac{d}{2}} e^{-\frac{x^t x v}{2}}$$

Entonces, multiplicando $p_{G|V=v}$ por p_V y por marginalización, obtenemos

$$\begin{aligned} p_X(x) &= \frac{\nu^{\frac{\nu}{2}}}{2^{\frac{\nu+d}{2}} \pi^{\frac{d}{2}} \Gamma\left(\frac{\nu}{2}\right)} \int_{\mathbb{R}_+} v^{\frac{\nu+d}{2}-1} e^{-\frac{x^t x+\nu}{2} v} \, dv \\ &= \frac{\nu^{\frac{\nu}{2}} (\nu + x^t x)^{-\frac{\nu+d}{2}}}{\pi^{\frac{d}{2}} \Gamma\left(\frac{\nu}{2}\right)} \int_{\mathbb{R}_+} u^{\frac{\nu+d}{2}-1} e^{-u} \, du \\ &= \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{(\pi\nu)^{\frac{d}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^t x}{\nu}\right)^{-\frac{\nu+d}{2}} \end{aligned}$$

La secunda linea viene del cambio de variables $u = \frac{x^t x+\nu}{2} v$ y la tercera reconociendo en la integral la función gamma (ver notaciones). □

Nota: este lema permite también probar el lema **??** escribiendo $AX + b \overset{\mathrm{d}}{=} \sqrt{\frac{\nu}{V}} A\Sigma^{\frac{1}{2}} G + Am + b$.

Lema 1-50 (Límite gaussiana)**.** *Sea* $X_\nu \sim \mathcal{T}_\nu(m, \Sigma)$ *vector Student-*t *parametizado por* ν *su grado de libertad. Entonces*

$$X_\nu \xrightarrow[\nu \to \infty]{\mathrm{d}} = X \sim \mathcal{N}(m, \Sigma)$$

con $\xrightarrow{\mathrm{d}}$ *límite es en distribución.*

Demostración. La prueba es inmediata tomando el logaritmo de la densidad de probabilidad, usando la formula de Stirling $\log\Gamma(z) = \left(z-\frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + o(1)$ en $z \to +\infty$ (Stirling, 1730; Abramowitz & Stegun, 1970; Gradshteyn & Ryzhik, 2015) y $-\frac{d+\nu}{2}\log\left(1+\frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu}\right) = -\frac{d+\nu}{2}\left(\frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu} + o\left(\nu^{-1}\right)\right) = -\frac{(x-m)^t\Sigma^{-1}(x-m)}{2} + o(1)$. □

Las variables Student-t tienen varias representaciones estocásticas, relacionadas a la gaussiana (Fang et al., 1990; Anderson, 2003; Kotz & Nadarajan, 2004; Ando & Kaufman, 1965):

Demostración. □

Nota: este lema permite también probar el lema 1-48 escribiendo $AX + b \overset{\mathrm{d}}{=} \frac{A\Sigma^{\frac{1}{2}}G}{sqrtV} + Am + b$.

Lema 1-51 (Relación con la distribución de Wishart)**.** *Sea* $W \sim \mathcal{W}(\Sigma^{-1}, \nu+d-1)$ $d\times d$ *Wishart con* $\Sigma \in P_d^+(\mathbb{R})$*,* $Y \sim \mathcal{N}(0, \nu I)$ *con* $\nu > 0$ *e* Y *independiente de* W*. Entonces, para* $m \in \mathbb{R}^d$*,*

$$W^{-\frac{1}{2}}Y + m \sim \mathcal{T}_\nu(m, \Sigma)$$

Demostración. Sea $X = W^{-\frac{1}{2}}Y$. De la nota siguiendo la tabla de características es necesario y suficiente probar que $X \sim \mathcal{T}_\nu(0,\Sigma)$. Ahora, de la independencia tenemos

$$p_{X|W=w}(x) = (2\pi\nu)^{-\frac{d}{2}}|w|^{\frac{1}{2}}e^{-\frac{x^twx}{2\nu}}$$

Denotamos por $D = \left\{w_{ij},\, 1\le j\le i\le d \;\middle|\; w\in P_d^+(\mathbb{R})\right\}$ y, por abuso de escritura, $dw = \prod_{1\le j\le i\le d} dw_{ij}$. Entonces, multiplicando $p_{X|W=w}$ por p_W y por marginalización, obtenemos

$$\begin{aligned}
p_X(x) &= \int_D \frac{|w|^{\frac{\nu-1}{2}}e^{-\frac{x^twx}{2\nu}-\frac{1}{2}\operatorname{Tr}(\Sigma w)}}{2^{\frac{d(\nu+d)}{2}}(\pi\nu)^{\frac{d}{2}}\left|\Sigma^{-1}\right|^{\frac{\nu+d-1}{2}}\Gamma_d\left(\frac{\nu+d-1}{2}\right)}dw \\
&= \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{(\pi\nu)^{\frac{d}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left|\Sigma+\frac{xx^t}{\nu}\right|^{-\frac{\nu+d}{2}}|\Sigma|^{\frac{\nu+d-1}{2}}\int_D \frac{|w|^{\frac{\nu+d-d-1}{2}}e^{-\frac{1}{2}\operatorname{Tr}\left(\left[\Sigma+\frac{xx^t}{\nu}\right]w\right)}}{2^{\frac{d(\nu+d)}{2}}\left|\left(\Sigma+\frac{xx^t}{\nu}\right)^{-1}\right|^{\frac{\nu+d}{2}}\Gamma_d\left(\frac{\nu+d}{2}\right)}dw \\
&= \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{(\pi\nu)^{\frac{d}{2}}\Gamma\left(\frac{\nu}{2}\right)|\Sigma|^{\frac{1}{2}}}\left(1+\frac{x^t\Sigma^{-1}x}{\nu}\right)^{-\frac{\nu+d}{2}}\int_D \frac{|w|^{\frac{\nu+d-d-1}{2}}e^{-\frac{1}{2}\operatorname{Tr}\left(\left[I+\frac{xx^t}{\nu}\right]w\right)}}{2^{\frac{d(\nu+d)}{2}}\left|\left(I+\frac{xx^t}{\nu}\right)^{-1}\right|^{\frac{\nu+d}{2}}\Gamma_d\left(\frac{\nu+d}{2}\right)}dw
\end{aligned}$$

Para $a,b\in\mathbb{R}^d$, $M\in\mathcal{M}_{d,d}(\mathbb{R})$, en la secunda linea usamos la identidad $a^tMb = \operatorname{Tr}(ba^tM)$ y $\Gamma_d\left(x-\frac{1}{2}\right) = \frac{\Gamma\left(x-\frac{d}{2}\right)}{\Gamma(x)}\Gamma_d(x)$ (ver notaciones) y en la tercera linea usamos $\left|\Sigma+\frac{xx^t}{\nu}\right| = |\Sigma|\left|I+\frac{\Sigma^{-1}xx^t}{\nu}\right|$ y la identidad de Sylvester (Sylvester, 1851) o (Harville, 2008, § 18.1) $|I+ab^t| = 1+b^ta$. Se concluye que $X\sim\mathcal{T}_\nu(0,\Sigma)$ reconociendo en el factor de la integral como la distribución $\mathcal{T}_\nu(0,\Sigma)$ y en el integrande la distribución de Wishart $\mathcal{W}\left(\left(I+\frac{xx^t}{\nu}\right),\,\nu+d\right)$ que suma entonces a la unidad. □

Como lo hemos introducido, la distribución Student-t aparece naturalmente en el marco de la estimación, especialmente a través de la estimación empírica de la media y covarianza (Muirhead, 1982; Gupta & Nagar, 1999; Bilodeau & Brenner, 1999; Anderson, 2003; Seber, 2004):

Teorema 1-53. *Sean* $X_i \sim \mathcal{N}(m,\Sigma)$, $i=1,\ldots,n>d-1$ *independientes, y sea la media empírica (ver corolario 1-10)*

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

y la covarianza empírica construida a partir de la media empírica (ver corolario 1-12)

$$\overline{\Sigma}=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)^t$$

Entonces:

- $\overline{X}-m \sim \mathcal{N}\left(0\,,\,\frac{1}{n}\,\Sigma\right)$ *y* $\overline{\Sigma} \sim \mathcal{W}(\frac{1}{n-1}\Sigma\,,\,n-1)$ *son independientes;*
- $\sqrt{\frac{n(n-d)}{n-1}}\,\overline{\Sigma}^{\,-\frac{1}{2}}\,\left(\overline{X}-m\right) \sim \mathcal{T}_{n-d}\left(0\,,\,I\right)$

Demostración. Se refiera a los corolarios 1-10 y 1-12 por lo de las distribuciones de $\overline{X}-m$ y de $\overline{\Sigma}$ respectivamente.

A continuación, sean $\widetilde{X}_i=X_i-m$ y $\widetilde{X}=\left[\widetilde{X}_1 \quad \cdots \quad \widetilde{X}_n\right]$. Obviamente

$$\overline{\widetilde{X}}\equiv\overline{X}-m=\frac{1}{n}\,\widetilde{X}\mathbb{1}$$

con $\mathbb{1}\in\mathbb{R}^n$ vector de componentes iguales a 1 y vimos en la prueba del corolario 1-12 que

$$\overline{\Sigma}=\frac{1}{n-1}\widetilde{X}\left(I-\frac{\mathbb{1}\mathbb{1}^t}{n}\right)\widetilde{X}^t$$

$A=I-\frac{\mathbb{1}\mathbb{1}^t}{n}\in P_n(\mathbb{R})$ es idemponenta de rango 1, con $A\mathbb{1}=0$, así que por diagonalización (Horn & Johnson, 2013; ?, ?, ?) tenemos

$$A=P\begin{bmatrix}I_{n-1} & 0\\ 0 & 0\end{bmatrix}P^t \qquad \text{con} \qquad P=\left[B \quad \tfrac{1}{\sqrt{n}}\mathbb{1}\right]$$

$PP^t=PP^t=I$ y

$$B\in\mathcal{M}_{n,n-1}(\mathbb{R}) \quad \text{tal que} \quad B^tB=I \text{ y } \mathbb{1}^tB=0$$

Ahora, poniendo la descomposición diagonal de A en $\overline{\Sigma}$ obtenemos (ver corolario 1-12)

$$\overline{\Sigma}=\frac{1}{n-1}\,YY^t \qquad \text{con} \qquad Y=\widetilde{X}B$$

Luego, de la gaussianidad y independencia de los $\widetilde{X}_i$ tenemos, para $\widetilde{x}=\left[\widetilde{x}_1 \quad \cdots \quad \widetilde{x}_n\right]\in\mathcal{M}_{d,n}(\mathbb{R})$

$$\begin{aligned}
p_{\widetilde{X}}(\widetilde{x}) &= (2\pi)^{-\frac{nd}{2}}|\Sigma|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\widetilde{x}_i^t\Sigma^{-1}\widetilde{x}_i\right)\\
&= (2\pi)^{-\frac{nd}{2}}|\Sigma|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\operatorname{Tr}\left(\Sigma^{-1}\widetilde{x}_i\widetilde{x}_i^t\right)\right)\\
&= (2\pi)^{-\frac{nd}{2}}|\Sigma|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}\widetilde{x}\,\widetilde{x}^t\right)\right)
\end{aligned}$$

Sea la transformación $\begin{bmatrix} Y & \sqrt{n}\widetilde{\overline{X}} \end{bmatrix} = \widetilde{X}P$, *i. e.,* $\widetilde{X} = \begin{bmatrix} Y & \sqrt{n}\widetilde{\overline{X}} \end{bmatrix} P^t$. Se nota que $|P| = 1$ y por transformación (ver teorema 1-12), para $y \in \mathcal{M}_{d,n-1}(\mathbb{R})$ y $x \in \mathbb{R}^d$

$$\begin{aligned} p_{Y,\sqrt{n}\widetilde{\overline{X}}}(y,x) &= (2\pi)^{-\frac{nd}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \operatorname{Tr}\left(\Sigma^{-1} \begin{bmatrix} y & x \end{bmatrix} P^t P \begin{bmatrix} y^t \\ x \end{bmatrix} \right) \right) \\ &= (2\pi)^{-\frac{nd}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \operatorname{Tr}\left(\Sigma^{-1} \left(yy^t + xx^t \right) \right) \right) \\ &= (2\pi)^{-\frac{(n-1)d}{2}} |\Sigma|^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2} \operatorname{Tr}\left(\Sigma^{-1} yy^t \right) \right) \times (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^t \Sigma^{-1} x \right) \end{aligned}$$

Claramente, de la factorización de las distribuciones, $Y = XB$ y $\sqrt{n}\widetilde{\overline{X}}$ son independientes, es decir que $\frac{1}{n-1} YY^t = \overline{\Sigma}$ y $\widetilde{\overline{X}} = \overline{X} - m$ son independientes, lo que cierra la prueba del primer item. Pasando, la forma de $p_{Y,\sqrt{n}\widetilde{\overline{X}}}(y,x)$ confirma que $\overline{X} - m$ es gaussiana centrada de covarianza $\frac{1}{n}\Sigma$, y que los Y_i son independientes gaussianos, dando la distribución de Wishart del lema **??** para la covarianza empírica.

A continuación,

$$\sqrt{\frac{n(n-d)}{n-1}}\, \overline{\Sigma}^{-\frac{1}{2}} \left(\overline{X} - m \right) = \frac{1}{\sqrt{n-1}}\, \Sigma^{-\frac{1}{2}} \left(\Sigma^{-1} \overline{\Sigma} \Sigma^{-1} \right)^{-\frac{1}{2}} \left(\sqrt{n(n-d)}\, \Sigma^{-\frac{1}{2}} \left(\overline{X} - m \right) \right)$$

Del teorema 1-42 y del lema 1-35 tenemos

$$\sqrt{n(n-d)}\, \Sigma^{-\frac{1}{2}} \left(\overline{X} - m \right) \sim \mathcal{N}(0, (n-d)I) \qquad \text{y} \qquad \Sigma^{-1} \overline{\Sigma} \Sigma^{-1} \sim \mathcal{W}\left(\left((n-1)\Sigma \right)^{-1}, n - d + d - 1 \right)$$

Se cierra la prueba usando los lemas 1-51 y **??**. □

Más propiedades de esta distribución se encuentran en libros especializados, por ejemplo (Kotz & Nadarajan, 2004) completamente dedicado a esta distribución.

La distribución Student-t se generaliza al caso complejo Z definido sobre $\mathbb{C}^d$; se denota $Z \sim \mathcal{CT}_\nu(m, \Sigma)$ donde $m \in \mathbb{C}^d$, $\Sigma \in P_d^+(\mathbb{C})$ y la densidad es dada por

$$p_Z(z) = \frac{\Gamma\left(d + \frac{\nu}{2}\right)}{\pi^d \nu^d \Gamma\left(\frac{\nu}{2}\right) |\Sigma|} \left(1 + \frac{(z-m)^\dagger \Sigma^{-1} (z-m)}{\nu} \right)^{-\frac{\nu}{2} - d}$$

(ver por ejemplo (Kotz & Nadarajan, 2004, § 5.12 y ref.) para una versión muy parecida).

También, la distribución Student-t se generaliza al caso matriz-variada X definido sobre $M_{d,d'}(\mathbb{R})$; se denota $X \sim \mathcal{T}_\nu(M, \Sigma, \Omega)$ donde $M \in M_{d,d'}(\mathbb{R})$, $\Sigma \in P_d^+(\mathbb{R})$, $\Omega \in P_{d'}^+(\mathbb{R})$ y la densidad es dada por $p_X(x) = \dfrac{\Gamma_d\left(\frac{\nu+d+d'-1}{2}\right)}{\pi^{\frac{\nu d}{2}} \Gamma_d\left(\frac{\nu+d-1}{2}\right) |\Sigma|^{\frac{d'}{2}} |\Omega|^{\frac{d}{2}}} \left| I + \Sigma^{-1}(x-M)\,\Omega^{-1}\,(x-M)^t \right|^{-\frac{\nu+d+d'-1}{2}}$. Se refiera a (Dickey, 1967), (Gupta & Nagar, 1999, Cap. 4) o (?, ?, §5.11 y ref.) para tener más detalles.

1.10.2.10. Distribución Student-r multivariada

Estas distribuciones aparecieron esencialmente a través el sistema dicho de Pearson en el fin del siglo XIX (Pearson, 1895; Johnson et al., 1995a, 1995a; Kotz et al., 2000; Fang et al., 1990). Más especialmente son conocidos como *Pearson tipo IIIa*α. A veces, se encuentra en física la denominación *Student-*r; aparecen como maximizantes de la entropía de Rényi, o de la de Tsalis de la misma manera que las Student-t usual, pero cuando el indice entrópico es mayor que la unidad (ver capitulo 2 para tener más detalles) (Johnson & Vignat, 2007; Costa, Hero III & Vignat, 2003; Vignat, Hero III & Costa, 2004; Tsallis, 1988, 1999).

Se denota con $X \sim \mathcal{R}_\nu(m,\Sigma)$ con $m \in \mathbb{R}^d$, $\Sigma \in P_d^+(\mathbb{R})$. m es llamado *parámetro de posición*, Σ es llamada *matriz característica* y $\nu > d-2$ llamado *grado de libertad*. Las características de una Student-r son las siguientes:

Dominio de definición	$\begin{aligned}\mathcal{X} &= m + \Sigma^{\frac{1}{2}}\,\mathbb{B}_d\left(0,\sqrt{\nu+2}\right)\\ &= \left\{x\in\mathbb{R}^d \;\middle|\; \Sigma^{-\frac{1}{2}}(x-m)\in\mathbb{B}_d\left(0,\sqrt{\nu+2}\right)\right\}\end{aligned}$
Parámetro	$\nu > d-2$ (grado de libertad), $m\in\mathbb{R}^d$ (posición), $\Sigma\in P_d^+(\mathbb{R})$ (matriz caracerística)
Densidad de probabilidad	$p_X(x) = \dfrac{\Gamma\left(\frac{\nu}{2}+1\right)}{\pi^{\frac{d}{2}}(\nu+2)^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)\,\lvert\Sigma\rvert^{\frac{1}{2}}}\left(1-\dfrac{(x-m)^t\Sigma^{-1}(x-m)}{\nu+2}\right)_+^{\frac{\nu-d}{2}}$
Promedio	$m_X = m$
Covarianza	$\Sigma_X = \Sigma$
Asimetría	$\gamma_X = 0$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{-2}{\nu+4}\sum_{i,j=1}^{d}\left(\left(\mathbb{1}_i\mathbb{1}_i^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_j^t\right)+\left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_i\mathbb{1}_j^t\right)+\left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_i^t\right)\right)$
Función característica	$\Phi_X(\omega) = \dfrac{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}+1\right)}{(\nu+2)^{\frac{\nu}{4}}}\,e^{\imath\omega^t m}\,\left(\omega^t\Sigma\omega\right)^{-\frac{\nu}{4}}\,J_{\frac{\nu}{2}}\left(\sqrt{(\nu+2)\,\omega^t\Sigma\omega}\right)$

Nota: nuevamente se puede escribir $X \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}}R+m$ con $R\sim\mathcal{R}_\nu(0,I)$ donde R es dicha *Student-*r *estandar* y las características de X son vínculadas a las de R (y vice-versa) por transformación lineal (ver secciones anteriores).

Unas densidades de probabilidad Student-r estandares y las funciones de repartición asociadas en el caso escalar son representadas en la figura Fig. 1-32-(a) y (b) para varios grado de libertad, y una densidad en un contexto bi-dimensional figura Fig. 1-32(c).

Nota: en el caso $\nu = d$ la ley es uniforme adentro del dominio de definición $X \sim \mathcal{U}\left(m+\Sigma^{\frac{1}{2}}\,\mathbb{B}_d\left(0,\sqrt{d+2}\right)\right)$. Para $d-2<\nu<d$ la densidad diverge en los bordes del dominio de definición (divergencia integrable).

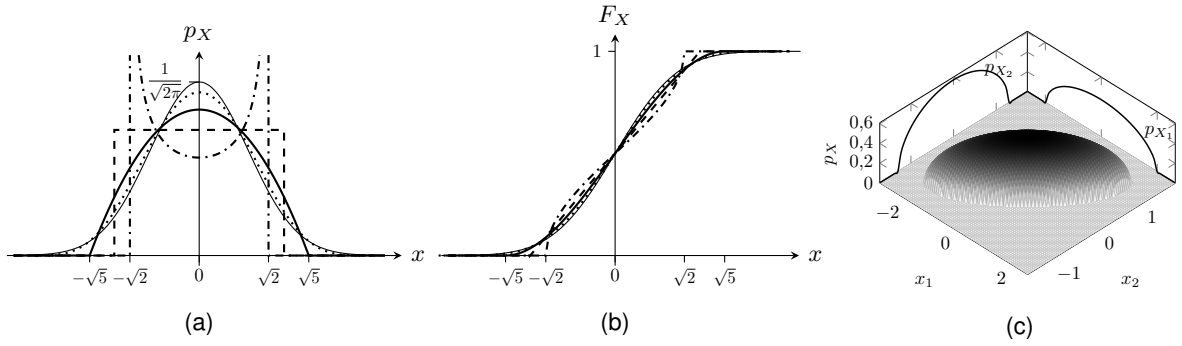


Figura 1-32: Ilustración de una densidad de probabilidad Student-r escalar estandar (a), y la función de repartición asociada (b) con $\nu = 0$ (linea mixta), $\nu = 1$ (linea guionada), $\nu = 3$ (linea llena), $\nu = 11$ (linea punteada) y $\nu \to +\infty$ (linea llena fina; ver más adelante) grado de libertad, así que una densidad de probabilidad Student-r bi-dimensional con $\nu = 3$ grado de libertad, centrada, y de matriz característica $\Sigma = R(\theta)\Delta^2 R(\theta)^t$ con $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ matriz de rotación y $\Delta = \operatorname{diag}\left(\begin{bmatrix} 1 & a \end{bmatrix}\right)$ matriz de cambio de escala, y sus marginales $X_1 \sim \mathcal{R}_\nu\left(0, \cos^2\theta + a^2\sin^2\theta\right)$ y $X_2 \sim \mathcal{R}_\nu\left(0, \sin^2\theta + a^2\cos^2\theta\right)$ (ver más adelante). En la figura, $a = \frac{1}{3}$ y $\theta = \frac{\pi}{6}$.

Contrariamente al caso gaussiano, de la forma de la densidad de probabilidad, es claro que aún si la matriz Σ es diagonal, la densidad no factoriza, así que las componentes del vector no son independientes. Este ejemplo muestra claramente y nuevamente que la reciproca del lema 1-6 es falsa en general.

Sin embargo, las distribuciones Student-r tienen varias propiedades notables.

Lema 1-52 (Stabilidad por transformación lineal)**.** *Sean $X \sim \mathcal{R}_\nu(m, \Sigma)$, A matriz de $\mathcal{M}_{d',d}(\mathbb{R}))$ con $d' \leq d$, y de rango lleno y $b \in \mathbb{R}^{d'}$. Entonces*

$$AX + b \sim \mathcal{R}_\nu(Am + b, A\Sigma A^t)$$

En particular los componentes de X son student-r,

$$X_i \sim \mathcal{R}_\nu(m_i, \Sigma_{i,i})$$

Demostración. La prueba es inmediata usando la función característica y sus propiedades por transformación lineal. La condición sobre A es necesaria y suficiente para que $A\Sigma A^t \in P_{d'}^+(\mathbb{R})$. □

Lema 1-53 (Límite gaussiana)**.** *Sea $X_\nu \sim \mathcal{R}_\nu(m, \Sigma)$ vector Student-r parametizado por ν su grado de libertad. Entonces*

$$X_\nu \xrightarrow[\nu\to\infty]{\mathrm{d}} = X \sim \mathcal{N}(m, \Sigma)$$

con $\xrightarrow{\mathrm{d}}$ límite es en distribución.

Demostración. La prueba es inmediata tomando el logaritmo de la densidad de probabilidad, usando la formula de Stirling $\log\Gamma(z) = \left(z - \frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + o(1)$ en $z \to +\infty$ (Stirling,

1730; Abramowitz & Stegun, 1970; Gradshteyn & Ryzhik, 2015) y $\frac{\nu-d}{2}\log\left(1-\frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu}\right) = -\frac{\nu-d}{2}\left(\frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu+2}+o\left(\nu^{-1}\right)\right) = -\frac{(x-m)^t\Sigma^{-1}(x-m)}{2}+o(1)$. Además, se nota que $\mathcal{X}\to\mathbb{R}^d$. □

Lema 1-54 (Relación con la distribución gamma y la ley gaussiana)**.** *Sea* $V\sim\mathcal{G}\left(\frac{\nu-d}{2}+1\,,\,\frac{1}{2}\right)$ *y* $G\sim\mathcal{N}(0,I)$ d*-dimensional e independientes de* V*. Entonces, para* $\Sigma\in P_d^+(\mathbb{R})$ *y* $m\in\mathbb{R}^d$*,*

$$\frac{\sqrt{\nu+2}\,\Sigma^{\frac{1}{2}}G}{\sqrt{V+\|G\|^2}}+m\;\sim\;\mathcal{R}_\nu(m,\Sigma)$$

Demostración. Sea $X=\frac{\sqrt{\nu+2}\,G}{\sqrt{V+\|G\|}}$. De la nota siguiendo la tabla de características es necesario y suficiente probar que $X\sim\mathcal{R}_\nu(0,I)$. Probaremos en la sección 1.10.4 que $G\overset{\mathrm{d}}{=}RU$ con $R>0$ de densidad de probabilidad $p_R(r)\propto r^{d-1}e^{-\frac{r^2}{2}}$, independiente de U, vector de distribución uniforme sobre la esfera $\mathbb{S}_d$. Probaremos también que $Y\sim\mathcal{R}_\nu(0,I)$ se escribe de la misma manera $Y\overset{\mathrm{d}}{=}SU$ ahora con $S>0$ de densidad de probabilidad $p_S(s)\propto s^{d-1}\left(1-\frac{s^2}{\nu+2}\right)_+^{\frac{\nu-d}{2}}$. Ahora, del teorema 1-37, y de la escritura estocástica de G tenemos

$$X\overset{\mathrm{d}}{=}\frac{\sqrt{d+2}\,R}{\sqrt{V+R^2}}\,U$$

Sea

$$S=\frac{\sqrt{d+2}\,R}{\sqrt{V+R^2}}$$

Claramente, R,V siendo independientes de U, la variable S es independiente de U en esta escritura estocastica. Entonces, sufice probar que $p_S(s)\propto s^{d-1}\left(1-\frac{s^2}{\nu+2}\right)_+^{\frac{\nu-d}{2}}$. Por eso, sea

$$T=\sqrt{\frac{V+R^2}{\nu+2}}\qquad\text{y}\qquad g:(r,v)\mapsto(s,t)$$

con ambas $(r,v)\in\mathbb{R}_+^2$ y $(s,t)\in\mathbb{R}_+^2$. Se calcula sencillamente la jacobiana de $g^{-1}:(s,t)\mapsto\left(st,t^2(\nu+2-s^2)\right)$ y a continuación el valor absoluto de su determinente que vale

$$\left|\mathrm{J}_{g^{-1}}\right|=2\,(\nu+2)\,t^2$$

Del teorema de transformación 1-12 y de la independencia de R y V tenemos

$$\begin{aligned}p_{S,T}(s,t) &= 2\,(\nu+2)\,t^2\,p_R(st)\,p_V\left(t^2(\nu+2-s^2)\right)\\ &\propto t^2\,(st)^{d-1}\,e^{-\frac{s^2t^2}{2}}\,t^{\nu-d}\,\left(\nu+2-s^2\right)_+^{\frac{\nu-d}{2}}\,e^{-\frac{t^2(\nu+2-s^2)}{2}}\end{aligned}$$

es decir

$$p_{S,T}(s,t)\propto s^{d-1}\left(1-\frac{s^2}{\nu+2}\right)_+^{\frac{\nu-d}{2}}\,t^{\nu+1}\,e^{-\frac{1}{2}t^2}$$

Inmediatamente, de la forma separable de $p_{S,T}$, vemos que S y T son independientes, y sobre todo se reconoce en el primer factor que la ley de S es dada por $p_S(s)\propto s^{d-1}\left(1-\frac{s^2}{\nu+2}\right)_+^{\frac{\nu-d}{2}}$, lo que cierra la prueba. Pasando, de la ley de T se notará que $T^2\sim\mathcal{G}\left(\frac{\nu}{2}+1,\frac{1}{2}\right)$, como suma de dos variables gamma independientes respectivamente $V\sim\mathcal{G}\left(\frac{\nu-d}{2}+1,\frac{1}{2}\right)$ y $R^2=\|G\|^2\sim\mathcal{G}\left(\frac{d}{2},\frac{1}{2}\right)$ (ver lemma 1-33). T es de distribución raiz-gamma. □

Más propiedades de esta distribución se encuentran en libros especializados, por ejemplo (Fang et al., 1990; Kotz et al., 2000) o en (Zozor, 2012, Sec. 3.2.1).

La distribución Student-r se generaliza al caso complejo Z definido sobre $\mathbb{C}^d$; se denota $Z \sim \mathcal{CR}_\nu(m, \Sigma)$ donde $m \in \mathbb{C}^d$, $\Sigma \in P_d^+(\mathbb{C})$, $\nu > 2d-2$ y la densidad es dada por

$$p_Z(z) = \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\pi^d(\nu+2)^d\Gamma\left(\frac{\nu}{2}-d+1\right)\,|\Sigma|}\left(1-\frac{(z-m)^\dagger\,\Sigma^{-1}\,(z-m)}{\nu+2}\right)_+^{\frac{\nu}{2}-d}$$

Generalización matriz variada?

---

von Mises en el circulo? Y multivariate (von Mises-Fisher)? Cantor = singular? A ver con la idea de mixta

Gaussian ensembles? Wigner ensembles? Invariant matrix ensembles? Tracy-Widom

1.10.3 Familia exponencial

Muchas de las leyes que hemos visto, que sean discreta o continuas, partenecen a una clase que comparte propriedades particulares, y que juega un rol particular que sea en física (ej. en problema de maximización de entropía de Shannon como lo vamos a ver en el cápitulo 2) o en el marco de la inferencia bayesiana. Esta clase es la familia dicha exponencial (Darmois, 1935; Koopman, 1936; Andersen, 1970; Lehmann & Casella, 1998; ?, ?; Mukhopadhyay, 2000; Kotz et al., 2000; Robert, 2007; van den Bos, 2007; Cencov, 1982; Kay, 1993; ?, ?). En inferencia bayesiana, cuando una distribución de sampleo cae en esta familia, se puede por ejemplo deducir a priori conjugados, es decir tales que la distribución dicha a posteriori caiga en la misma familia (*i. e.,* tenga la misma forma paramétrica pero con parámetros dependientes de los datos) [74]. Notar que esta familia es conocido o a veces *familia de Koopman-Darmois* debido a la introducción por Koopman (Koopman, 1936) o Darmois (Darmois, 1935) en los años 1935-1936 (ver también Pitman (Pitman, 1936)) y es definida de la manera siguiente:

Definición 1-49 (Familia exponencial y exponencial natural)**.** *Sea X vector aleatorio definido sobre $\mathcal{X} \subset \mathbb{R}^d$, de densidad de probabilidad p_X con respecto a una medida μ (discreta, continua o mixta).*

---

[74] Ver nota de pie [51]. Recordamos que si observaciones tienen una distribución $p_{X|\Theta=\theta}(x)$ con una distribución a priori del parámetro $p_\Theta(\theta)$, por la regla de Bayes el a posteriori, es decir la ley del parametro dados las observaciones es dada por $p_{\Theta|X=x}(\theta) \propto p_{X|\Theta=\theta}(x)p_\Theta(\theta)$ con $\propto$ significando "proporcional a". Si $p_{\Theta|X=x}$ tiene la misma forma paramétrica que $p_\Theta(\theta)$ inferir θ con datos que se observan se reduce a actualizar el parámetro de la ley a posteriori.

La distribución de probabilidad p_X *es dicha de la* familia exponencial de orden k, $k \in \mathbb{N}^*$, *si se escribe de la forma*

$$p_X(x) = C(\theta)\, h(x)\, \exp\left(\eta(\theta)^t S(x)\right)$$

donde

$$C : \Theta \subset \mathbb{R}^m \mapsto \mathbb{R}_+, \qquad h : \mathcal{X} \mapsto \mathbb{R}_+, \qquad \eta : \Theta \mapsto \mathbb{R}^k, \qquad S : \mathcal{X} \mapsto \mathbb{R}^k$$

con $m \in \mathbb{N}^*$. *En otras palabras, la familia exponencial es una familia parametrica de la forma así definida, con* Θ *el espacio de los parámetros. La familia es dicha* exponencial natural *si tiene la forma*

$$p_X(x) = \frac{1}{Z(\eta)}\, h(x)\, \exp\left(\eta^t S(x)\right) = h(x)\, \exp\left(\eta^t S(x) - \varphi(\eta)\right)$$

con

$$\eta \in \mathrm{N} \subset \mathbb{R}^k, \qquad Z : \mathbb{R}^k \mapsto \mathbb{R}_+, \qquad \varphi = \log(Z), \qquad h : \mathcal{X} \mapsto \mathbb{R}_+, \qquad S : \mathcal{X} \mapsto \mathbb{R}^k$$

donde $\mathrm{N} = \left\{ \eta \in \mathbb{R}^k \;\middle|\; \int_{\mathcal{X}} h(x)\, \exp\left(\eta^t S(x) - \varphi(\eta)\right)\, d\mu(x) < +\infty \right\}$ *es (convexo y) llamado* espacio de parámetros naturales.

Se notara que con la reparametrización $\eta(\theta)$ se puede siempre (por lo menos formalmente) escribir una ley de la familia exponencial bajo su forma natural. Con respeto a cada termino:

- $S(X)$ es llamada *estadística suficiente* o *estadística exaustiva*. Esta denominación viene del hecho que el conocimiento de $S(X)$ es suficiente para estimar η, o es un resumen exaustivo de η. En particular, la estimación de verosimilitud [75], o cualquier estimador Bayesiano [76] de η depende solamente de $S(X)$ (Kay, 1993; Lehmann & Casella, 1998; Robert, 2007; Cencov, 1982; Ibarrola & Pérez, 2012; Mukhopadhyay, 2000). Formalmente, una estadistica $T(X)$ es suficiente para un parámetro θ si la distribución de X condicionalmente a $S(X) = s$ no depende más de θ. Por ejemplo, para la familia exponencial, en el caso discreto, tenemos, $p_{X|S(X)=s}(x) = \frac{P\big((X=x) \cap (S(X)=s)\big)}{P(S(X)=s)} = \frac{h(x) \exp\left(\eta^t s - \varphi(\eta)\right) \mathbb{1}_{\{S(x)\}}(s)}{\sum_{x \in \mathcal{X}} h(x) \exp\left(\eta^t s - \varphi(\eta)\right) \mathbb{1}_{\{S(x)\}}(s)} = \frac{h(x) \mathbb{1}_{\{S(x)\}}(s)}{\sum_{x \in \mathcal{X}} h(x) \mathbb{1}_{\{S(x)\}}(s)}$ no depende de η. Veremos en el capítulo 2, sección 2.4.6 que la información de Fisher en η, medida

[75] El estimador del máximo de verosilitud η_{mv} es el η que máximiza p_X o cualquier función creciente como el logaritmo por ejemplo. Para la familia exponencial, eso da sencillamente η_{mv} satisficiendo $S(x) = \nabla\varphi(\eta)$, dependiente solamente de $S(x)$.

[76] Ver nota de pie **??**. Recuerdense que se modeliza el parámetro como aleatorio, así que la distribución que se considera se ve como la distribución condicional $p_{X|N=\eta}(x) = h(x) \exp\left(\eta^t S(x) - \varphi(\eta)\right)$, con la distribución a priori $p_N(\eta)$. De la regla de Bayes, la distribución a posteriori, *i. e.,* del parámetro dadas las observaciones x se escribe $p_{N|X=x}(\eta) \propto p_{X|N=\eta}(x) p_N(\eta)$. Si se da un costo de estimación $C(\widehat{\eta}, N) > 0$ caracterizando la "distancia" entre la estimación y el parametro "verdadero", se busca la función $\widehat{\eta}$ que minimiza el costo promedio, lo que es equivalente para cada x a buscar el valor $\widehat{\eta}$ que minimiza $\int_N C(\widehat{\eta}, \eta)\, \exp\left(\eta^t S(x) - \varphi(\eta)\right)\, p_N(\eta)\, d\mu(\eta)$ (Robert, 2007): claramente, el mínimo es función, solamente de $S(x)$, cualquieras sean el costo C y el a priori p_N (estos solamente determinan cual función de $S(X)$ vamos a obtener).

informacional y apareciendo en la cota del error cuadratico mínimo posible de un estimador de η, es la covarianza de S, mostrando de nuevo que S es sufiente para caracterizar η. En este mismo capítulo 2, sección 2.4.1, veremos que, sujeto a $\mathrm{E}[S(X)]$, la distribución que maximiza la entropía, medida de incerteza, es una distribución exponencial, enfatizando el rol de esta familia en física cuando se tiene la media de una estadística fija (ej. energía fija).

- θ es el parámetro, escalar o multivariado, y η es llamado *parámetro natural*.
- La función Z es llamada *función de partición*; A veces, φ es así llamada *log-función de partición*; no solo juegan un rol en la normalización de la ley, pero tienen una significación física como lo vamos a evocar. Apareció Z en física estadística por ejemplo en trabajos de Gibbs (?, ?; Gibbs, 1902).

Se notará, en particular, que Z es relacionada a los momentos de S, o φ a los cumulantes:

Teorema 1-54 (Función de partición y generadoras)**.** *Sea X de distribución exponencial natural de parámetro natural η y estadística suficiente S y denotamos Z la función de partición y φ su logaritmo. Entonces, las funciónes generadoras de los momentos y de los cumulantes $M_{S(X)}$ y $C_{S(X)}$ de la estadística suficiente $S(X)$ son relacionadas a Z y a $\varphi = \log Z$ por, $\forall\, u$ tal que $u + \eta \in N$,*

$$M_{S(X)}(u) = \frac{Z(u+\eta)}{Z(\eta)} \qquad \textit{y} \qquad C_{S(X)}(u) = \varphi(u+\eta) - \varphi(\eta)$$

En particular, si φ es diferenciable tenemos

$$\nabla\varphi(\eta) = \mathrm{E}\left[S(X)\right]$$

y si φ es dos veces diferenciable tenemos

$$\mathcal{H}\varphi(\eta) = \mathrm{Cov}\left[S(X)\right]$$

Pasando, de $\mathrm{Cov}\left[S(X)\right] \geq 0$ tenemos que, hessiana de φ siendo positiva, φ es convexa (Cambini & Martein, 2009).

Demostración. De la definición de la función generadora de $S(X)$, tenemos

$$\begin{aligned} M_{S(X)}(u) &= \mathrm{E}\left[\exp\left(u^t S(X)\right)\right] \\ &= \int_{\mathcal{X}} \frac{1}{Z(\eta)}\, h(x)\, \exp\left((u+\eta)^t S(x)\right)\, d\mu(x) \\ &= \frac{Z(u+\eta)}{Z(\eta)} \int_{\mathcal{X}} \frac{1}{Z(u+\eta)}\, h(x)\, \exp\left((u+\eta)^t S(x)\right)\, d\mu(x) \\ &= \frac{Z(u+\eta)}{Z(\eta)} \end{aligned}$$

La secunda relación es inmediata de $C_X = \log M_X$ conjuntamente a $\varphi = \log Z$.

A continuación, los cumulantes y momentos coinciden hasta el orden 3, dando inmediatamenta

$$\begin{aligned}\mathrm{E}\left[S(X)\right] &= \nabla_u C_{S(X)}\big|_{u=0} \\ &= \nabla_u \left(\varphi(u+\eta)-\varphi(\eta)\right)\big|_{u=0} \\ &= \nabla\varphi(\eta)\end{aligned}$$

y

$$\begin{aligned}\mathrm{Cov}\left[S(X)\right] &= \mathcal{H}_u C_{S(X)}\big|_{u=0} \\ &= \mathcal{H}_u \left(\varphi(u+\eta)-\varphi(\eta)\right)\big|_{u=0} \\ &= \mathcal{H}\varphi(\eta)\end{aligned}$$

□

En problemas de estimación, frecuentemente, se tiene n vectores aleatorios independiente de misma ley que se usan para estimar un párametro. En la familia exponencial, resuelte que la distribución conjunta en tal situación queda en la familia exponencial:

Lema 1-55. *Sean $X_1,\ldots,X_n$ vectores aleatorios, independientes, de misma ley de la familia exponencial natural, de estadistica suficiente S y de log-función de partición φ. Entonces la ley conjunta de los X_i cae el la familia exponencial de mismo orden que el de p_{X_i}, con el mismo parámetro, de estadistica suficiente $(X_1,\ldots,X_n)\mapsto\sum_{i=1}^{n}S(X_i)$ y de log-función de partición $n\varphi$.*

Demostración. El resulta es inmediato siendos los X_i independientes, dando la ley conjunta como el producto de las leyes de los X_i. □

Muchas distribuciones caen en la familia exponencial, que sean discretas o continuas, como lo vamos a ver en dos ejemplos.

Ejemplo 1-19. *Sea p_X distribución de Bernoulli $\mathcal{B}(p)$. Esta distribución partenece a la familia exponencial de orden 1. De hecho, se puede escribir la ley $p_X(x)=p^x(1-p)^{1-x}$ bajo la forma*

$$p_X(x)=\exp\left(x\log\left(\frac{p}{1-p}\right)+\log(1-p)\right)$$

Aparece que el parámetro natural es $\eta=\log\left(\frac{p}{1-p}\right)$ y la estadistica suficiente correspondiente es $S(X)=X$. En particular, para X_i, $i=1,\ldots,n$ independientes tales que $X_i\sim\mathcal{B}(p)$, una estadistica sufficiente de la ley conjunta es la media empírica $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$. Aparece que el estimador de verosimilitud máxima (Ver nota de pie 75) es precisamente la media empírica; es el estimador $\widehat{p}$ de error cuadratico $\mathrm{E}\left[(\widehat{p}-p)^2\right]$ mínimo (ver ej. (Kay, 1993)).

Ejemplo 1-20. *Sea p_X distribución gamma $\mathcal{G}(a,b)$. Esta distribución partenece a la familia exponencial de orden 2. De hecho, se puede escribir la ley bajo la forma*

$$p_X(x) = \frac{1}{x}\,\exp\left(\begin{bmatrix} a & b \end{bmatrix}\begin{bmatrix} \log x \\ -x \end{bmatrix} - \log\left(\frac{\Gamma(a)}{b^a}\right)\right)$$

El parámetro natural es así $\eta = \begin{bmatrix} a & b \end{bmatrix}^t$ y la estadistica suficiente correspondiente es $S(X) = \begin{bmatrix} \log x & -x \end{bmatrix}^t$.

Si muchas distribuciones partenecen a la familia exponencial, no todas caen en esta familia:

Ejemplo 1-21. *Sea p_X distribución Student-t $\mathcal{T}_\nu(m,\Sigma)$. Esta distribución no cae el la familia exponencial. De hecho, se puede todavía escribir la ley bajo la forma*

$$p_X(x) = C(\nu,\Sigma)\exp\left(-\frac{d+\nu}{2}\log\left(1+\frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu}\right)\right)$$

A pesar de que la ley parece tener la forma de la definición 1-49 con $\eta(\nu) = -\frac{d+\nu}{2}$, su factor $\log\left(1+\frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu}\right)$ no es función únicamente de x; se quedan todo los parámetros ν, m, Σ. Aún si un de esos es fijo, y no visto como parámetro, quedará un parámetro en este término.

Se puede también que una ley partenece o no a la familia exponencial, según que unos parámetros sean fijos (de hecho, no son parámetros más entonces), o no:

Ejemplo 1-22. *Sea p_X distribución binomial $\mathcal{B}(n,p)$. Si n es fijo, es decir no visto como parámetro, la distribución cae el la familia exponencial de orden 1. Si se consideran ambos n y p como parámetros, no cea más en la familia exponencial. Para ver eso, se escribir la ley bajo la forma*

$$p_X(x) = \frac{n!}{x!(n-x)!}\exp\left(x\log\left(\frac{p}{1-p}\right) + n\,\log(1-p)\right)$$

Entonces, si n es fijo, se concluye que p_X es en la familia exponencial de orden 1, de parámetro $\eta = \log\left(\frac{p}{1-p}\right)$ y estadística suficiente correspondiente $S(x) = x$. Al revés, si n es un parámetro (que p sea fijo o no), en $(n-x)!$, no se puede "separar" x de n y tampoco escribir este termino de la forma $\exp(f(n)g(x))$: la ley no es de la familia exponencial mas.

De las distribuciones que hemos visto:

- Caen en la familia exponencial la leyes: de Bernoulli, binomial cuando n es fijo, binomial negativa cuando r es fijo, multinomial cuando n es fijo, geometrica, de Poisson, gaussiana [77], gamma, Wishart [78], beta, de Dirichlet.

---

[77] En este caso, se puede ver el parámetro natural como $\left(\Sigma^{-1}m, -\frac{1}{2}\Sigma^{-1}\right)$ en lugar del vector formado de los $\left(\Sigma^{-1}m\right)_i$ y $-\frac{1}{2}\left(\Sigma^{-1}\right)_{i,j}$, $1 \leq i \leq j \leq d$ y la estadística suficiente como $\left(x, xx^t\right)$ siendo $x^t\Sigma^{-1}x = \mathrm{Tr}\left(\Sigma^{-1}xx^t\right)$, en lugar del vector formado de los x_i y x_ix_j, $1 \leq i \leq j \leq d$ (de la simetría). El orden es $d + \frac{d(d+1)}{2}$.

[78] De nuevo, en este caso se puede ver el parámetro natural formalmente como $\left(\frac{\nu-d-1}{2}, -\frac{1}{2}V^{-1}\right)$ y la estadística suficiente como $(\log|x|, x)$. El orden es $1 + \frac{d(d+1)}{2}$.

- No partenecen a la familia exponencial las leyes: binomial cuando n es un parámetro, binomial negativa cuando r es un parámetro, multinomial cuando n es un parámetro, hipergeometrica, hipergeometrica negativa, hipergeometrica multivaluada [79], Student-t, Student-r y uniformas [80].

Las distribuciones exponenciales aparecen frecuentemente en física estádistica a través de la teoría de Boltzmann (Boltzmann, 1896, 1898; Gibbs, 1902; Landau & Lifshitz, 1980; Mézard & Montanari, 2009; Merhav, 2010, 2018). Además, cantidades físicas se derivan de la log-función partición (Maxwell, 1867; Gibbs, 1902; Landau & Lifshitz, 1980; Mézard & Montanari, 2009; Merhav, 2010, 2018):

Ejemplo 1-23. *En física estadistica, se enfrente al problema de descripción macroscopico de un sistema de muchas partículas (ej. hirviente de un liquido). Hay tantas partículas que no se puede estudiar tales sistemas con las leyes usuales de la mecánica, así que se usa un enfoque probabilístico. Por eso, se considera un espacio $\mathcal{X}$ d-dimensional dicho* espacio de configuraciones *(puede ser discreto o continuo). En $x = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}$, cada x_i representa el* estado *de la i-ésima partícula (posición, velocidad, espín,...). Lo importante es que a un tipo de sistema se asocia una función energía $\mathcal{E}(x)$. Por ejemplo, en un sistema sin interacciones, $\mathcal{E}(x) = \sum_{i=1}^{d} \mathcal{E}_i(x_i)$. El el caso del gaz perfecto, $\mathcal{E}(x) = \frac{1}{2} m \sum_{i=1}^{d} x_i^2$ donde m es la masa de cada particula y x_i la velocidad de la i-ésima particula (espacio de configuraciones continuo). En el* modelo ferromagnético de Ising*, se consideran partículas en una retícula y $x_i = \pm 1$ es el espín de la partícula i (espacio de configuraciones discreto). Sometido a un campo magnético B, la energía es dada por $\mathcal{E}(x) = -\sum_{(i,j) \text{ vecinos}} x_i x_j - B \sum_{i=1}^{d} x_i$ (Lenz, 1920; Ising, 1925; Onsager, 1944; Landau & Lifshitz, 1980; Mézard & Montanari, 2009; Merhav, 2010, 2018). Se puede poner pesos $J_{i,j}$ en cada vecinos, positivos para interracciones ferromagnéticas, y negativos para interacciones antiferromagnéticas (modelos* vidrio de espín*, o más exactamente de Edwards-Anderson (?, ?; Landau & Lifshitz, 1980; Mézard & Montanari, 2009; Merhav, 2010, 2018)). El modelo de Curie-Weiss [81] se presenta de la misma manera, con la energía $\mathcal{E}(x) = -\frac{1}{d} \sum_{i \neq j) \text{ pares}} x_i x_j + B \sum_{i=1}^{d} x_i$ (Mézard & Montanari, 2009; Merhav, 2010, 2018).*

Según la teoria de Gibbs-Boltzmann, la dicha distribución de Gibbs-Boltzman *asociada a un espacio de configuración y modelo de energía es dada por*

$$p_X(x) = \frac{1}{Z(\beta)} \exp\left(-\beta \mathcal{E}(x)\right), \qquad \beta = \frac{1}{k_B T}$$

---

[79] En los casos hipergeometricos, haría falta que sean fijos respectivamente $n, m, k,\ n, r, k$ y $n, m, k_1, \ldots, k_c$ y la leyes no serían parametricas mas.

[80] Eso viene del hecho de que el soporte depende de los parámetros.

[81] Fue llamado así en relación a los trabajos de P. Curie (Curie, 1895) y P. Weiss (Weiss, 1896, 1907) sobre los materiales ferromagneticos.

donde $k_B \approx 1{,}38 \times 10^{-23}$ *julio por Kelvin es la constante de Boltzmann, y* T *es la temperatura en Kelvin. Esta distribución partenece claramente a la familia exponencial natural de parámetro* β *y de estadística suficiente* $-\mathcal{E}(x)$ *(acá,* $h = 1$*). En física estadística, la log-función de partición aparece en varias cantidades y potenciales físicos:*

$$F(\beta) = -\frac{1}{\beta} \log Z(\beta)$$

es la energía libre *o* energía libre de Helmholtz *del sistema. Es la energía disponible (o que se puede usar) de un sistema aislado.*

Luego, se define

$$U(\beta) = \frac{\partial}{\partial \beta} (\beta F(\beta)) = -\frac{\partial \log Z(\beta)}{\partial \beta} = \mathrm{E}\left[\mathcal{E}(X)\right]$$

donde X *sería el vector aleatorio de distribución* p_X*.* $U(\beta)$ *es la* energía interna *del sistema, promedio estadístico de la energía a través de todas las configuraciones posibles.*

Se define también una medida de incerteza llamada entropía *o* entropía de Gibbs *(Boltzmann, 1877, 1896, 1898; Gibbs, 1902; Jaynes, 1965; Landau & Lifshitz, 1980; Mézard & Montanari, 2009; Merhav, 2010, 2018). Esta medida caracteriza las fluctuaciones de la energía libre* [82]*,*

$$S(\beta) = \beta^2 \frac{\partial}{\partial \beta} F(\beta)$$

Aparece por un lado que $S(\beta) = \log Z(\beta) - \beta \frac{\partial}{\partial \beta} \log Z(\beta)$ *es decir, reconiciendo en el primer término* $-\beta F(\beta)$ *y en el secundo* $\beta U(\beta)$*,*

$$F = U - k_B T S$$

conocido como transformada de Legendre de la energía interna, y consecuencia de la primera ley de la termodynamica. Aparece también que $S(\beta) = \log Z(\beta) + \beta \,\mathrm{E}[\mathcal{E}(X)] = \int_{\mathcal{X}} (\log Z(\beta) + \beta \mathcal{E}(x))\, p_X(x)\, d\mu(x)$ *es decir*

$$S = -\int_{\mathcal{X}} p_X(x)\, \log p_X(x)\, d\mu(x)$$

Volveremos en esta definición de la entropía en el capitulo 2 en un marco más general.

Hablar de estadistica suficiente mínima?

1.10.4 Familia elíptica

1.10.4.1. Caso real

[82] La letra S se usó historicamente. Obviamente no corresponde a la estadística sufficiente que es acá $\mathcal{E}$.

El estudio de estos vectores es bastante antigua. Hace falta volver a trabajos de Maxwell en 1867 sobre la teoría del gas para encontrar unas de las primeras menciones a este formalismo (Maxwell, 1867) o (?, ?, pp. 377–391). El problema de Maxwell era de encontrar una distribución (tridimensional) que sea isotrópica y separable a la vez: monstró que tal distribución es necesariamente gaussiana (es ahora conocido como teorema de Maxwell-Hershel [83]. Volveremos en este teorema. La clase de las distribuciones elíptical, o a simetría elíptica fue estudiada intensivamente formalmente (Bartlett, 1934a, 1934b; Vershik, 1964; McGraw & Wagner, 1968; Cambanis et al., 1981; Eaton, 1981; Kano, 1994; Laurent, 1975; Yao, 1973; ?, ?; Fang et al., 1990; Muirhead, 1982; Bilodeau & Brenner, 1999). Fue también usadas en aplicaciones en estadística (Blake & Thomas, 1968; Chu, 1973; ?, ?; Arellano-Valle, del Pino & Iglesias, 2006; Bausson, Pascal, Forster, Ovarlez & Larzabal, 2007; Chitour & Pascal, 2008), o procesamiento de señal o imagenes (Goldman, 1976; Rangaswamy, Weiner & Öztürk, 1993, 1995; Zozor & Vignat, 2010; Zozor, 2012), entre otros.

Empezamos por la definición, antes de ir más allá estudiando sus propiedades remarcables.

Definición 1-50 (Vector esfericamente invariante)**.** *Sea X vector aleatorio d-dimensional real. X es dicho esfericamente invariante, o rotacionalmente invariante, o a simetría esférica, o de distribución esférica si para cualquier matriz ortogonal (o de rotación)* [84] *O,*

$$OX \overset{\mathrm{d}}{=} X$$

Tales vectores modelizan naturalmente fenómenos isotrópicos. Pero más allá, se puede que haya direcciones privilegiadas ortogonales pero con simetrias, *i. e.,* en lugar de simetrías esféricas, simetrías como en una pelota de rugby. Además, se puede que eso se pasa en torno a un punto no cero.

Definición 1-51 (Vector a simetría elíptica)**.** *Sea X vector aleatorio d-dimensional real. X es dicho a simetría elíptica, o elipticalmente invariante, o de distribución elíptica, en torno a $m \in \mathbb{R}^d$, si existe una matriz $\Sigma \in P_d^+(\mathbb{R})$ tal que para cualquier matriz ortogonal O,*

$$O\,\Delta^{-\frac{1}{2}}\,Q^t\,(X-m) \overset{\mathrm{d}}{=} \Delta^{-\frac{1}{2}}\,Q^t\,(X-m)$$

donde la matriz diagonal $\Delta > 0$ es la matriz de autovalores de Σ y Q la matriz de los autovectores correspondientes (matriz ortogonal (Bhatia, 1997, 2007; Horn & Johnson, 2013)), $\Sigma = Q\Delta Q^t$. Dicho de otra manera, $\Delta^{-\frac{1}{2}}Q^t\,(X-m)$ es a simetría esférica.

m es llamado parámetro de posición *y la matriz Σ es llamada* matriz característica.

---

[83] Ver (Bilodeau & Brenner, 1999, Prop. 4.11). Se notará que no hay muchas menciones de este teorema bajo esta denominación. no sabemos si la razón es que no tienen ni Maxwell, ni Herschel la partenidad o si no lo revendicaron. Sin embargo, ver (Maxwell, 1867)

[84] Recordarse que O es ortogonal o de rotación si $OO^t = O^tO = I$.

Se puede inmediatamente ver que Σ es definida por lo menos mediante un factor escalar. De hecho, si un Σ conviene, cualquier $a\Sigma$ con $a > 0$ conviene también.

Comparativamente a un vector esféricamente invariante, m es el centro de simetría, P contiene las direciones de "estiramientos" y $\Delta^{\frac{1}{2}}$ los factores de estiramientos, $X \overset{\mathrm{d}}{=} m + P\Delta^{\frac{1}{2}}Y$ con Y a simetría esférica. Para $m = 0$ y $\Delta \propto I$, se recupera obviamente un vector a simetría esférica.

Como lo hemos visto, un vector aleatorio es completamente definido por su medida de probabilidad, o equivalentemente por su función carácteristica. La última tiene una forma particular en el contexto elíptico:

Teorema 1-55 (Funciónes y generadoras características)**.** *Sea X vector aleatorio d-dimensional a simetría elíptica en torno a $m \in \mathbb{R}^d$ y de matriz característica $\Sigma \in P_d^+(\mathbb{R})$. Entonces la función caractéristica se escribe bajo la forma*

$$\Phi_X(\omega) = e^{\imath\,\omega^t m}\varphi_X\left(\omega^t\Sigma\omega\right)$$

donde $\varphi_X : \mathbb{R}_+ \mapsto [-1\,;1]$ escalar, es llamado generadora característica*. Tomando el logaritmo, obviamente la secunda función caracteristica se escribe*

$$\Psi_X(\omega) = \imath\,\omega^t m + \psi_X\left(\omega^t\Sigma\omega\right)$$

donde $\psi_X = \log\varphi_X : \mathbb{R}_+ \mapsto \mathbb{C}$ escalar. La llamaremos secunda generadora característica*. Reciprocamente, si la función característica tiene esta forma, X es a simetría elíptica.*

Demostración. Sea $Y = \Delta^{-\frac{1}{2}}Q^t\,(X - m)$ con $\Sigma = Q\Delta Q^t$ diagonalización de Σ. Por definición y del teorema 1-34, para cualquier matriz ortogonal O y cualquier $\omega \in \mathbb{R}^d$

$$\Phi_Y(\omega) = \Phi_{OY}(\omega) = \Phi_Y(O^t\omega)$$

En otros terminos, la función característica queda invariante bajo cualquier transformación ortogonal (rotación) sobre ω, y entonces depende solamente de la norma euclideana de ω. Es decir, existe una función escalar φ_X tal que

$$\Phi_Y(\omega) = \varphi_X(\omega^t\omega)$$

De nuevo, del teorema 1-34,

$$\Phi_X(\omega) \;=\; \Phi_{Q\Delta^{\frac{1}{2}}Y+m}(\omega) \;=\; e^{\imath\,\omega^t m}\,\Phi_Y(\Delta^{\frac{1}{2}}Q^t\omega) \;=\; e^{\imath\,\omega^t m}\,\varphi_x(\omega^t Q\Delta Q^t\omega)$$

lo que cierra la prueba directa. Reciprocamente, si Φ_X tiene la forma dada, para cualquier matriz ortogonal O $\Phi_{OY}(\omega) = \Phi_Y(O^t\omega) = \Phi_Y(\omega)$ y, por relación uno-uno entre la medida de probabilidad de una variable aleatorio y su función característica, $Y \overset{\mathrm{d}}{=} OY$.

Al final, de la simetría hermítica de la función de repartición, y de la simetría elíptica, tenemos $\varphi_X^*\left(\|\omega\|^2\right) = \varphi_X\left(\|-\omega\|^2\right) = \varphi_X\left(\|\omega\|^2\right)$, lo que proba que φ_X es a valores reales, siendo a valores también de modulo menor que $\varphi_X(0) = 1$. □

Se notará que si tomamos una matriz característica Σ y la generadora correspondiente φ, $a\Sigma$ y $\varphi_X\left(\frac{u}{a}\right)$ conviene también, lo que es de acuerdo con la indeterminencia de Σ bajo un factor positivo. Se puede añadir un vínculo, por ejemplo fijando $\operatorname{Tr}\Sigma$ para que Σ y φ_X sean únicamente definidas. Entonces, X será completamente caracterizado por m, Σ y φ_X, y escribiremos

$$X \sim \mathcal{ED}\left(m, \Sigma, \varphi_X\right)$$

y los conjuntos de generadoras características que resultan de la restricción de $\mathfrak{P}_d$ (y de $\mathfrak{P}$) a las funciones caracteristicas esféricamente invariante como

$$\mathfrak{EP}_d = \left\{\varphi : \mathbb{R}_+ \mapsto [-1;1] \text{ continuas con } \varphi(0) = 1 \;\middle|\; \Phi : x \mapsto \varphi\left(\|x\|^2\right) \in \mathfrak{P}_d\right\}$$

y

$$\mathfrak{EP} = \bigcap_{d=1}^{+\infty} \mathfrak{EP}_d = \left\{\varphi : \mathbb{R}_+ \mapsto [-1;1] \text{ continuas con } \varphi(0) = 1 \;\middle|\; \Phi : x \mapsto \varphi\left(\|x\|^2\right) \in \mathfrak{P}\right\}$$

(ver notaciones).

Vamos a ver más adelante varios ejemplos de vectores aleatorios a simetría elíptica que ya hemos vistos en las subsecciones anteriores. Un caso particular que va a jugar un rol importante es el de un vector de distribución uniforme sobre la esfera $\mathbb{S}_d$, $U \sim \mathcal{U}(\mathbb{S}_d)$ (Fang et al., 1990):

Ejemplo 1-24 (Distribución uniforme sobre la esfera unitaria)**.** *Sea* $U \sim \mathcal{U}(\mathbb{S}_d)$. *Entonces* $U \sim \mathcal{ED}\left(0, I, \varphi_U\right)$ *con*

$$\varphi_U(u) = 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right) u^{-\frac{d-2}{4}} J_{\frac{d}{2}-1}\left(\sqrt{u}\right)$$

con J_ν *función de Bessel*[85] *primera especie y de orden* ν *(ver notaciones).*

De hecho, de la definición de la función característica, tenemos

$$\Phi_U(\omega) = \frac{1}{|\mathbb{S}_d|}\int_{\mathbb{S}_d} e^{\imath\,\omega^t s} d\mu_H(s)$$

con μ_H *la medida de Haar*[86] *sobre la esfera y* $|\mathbb{S}_d| = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$ *la superficia de la esfera unitaria (Gradshteyn & Ryzhik, 2015). Ahora, denotanto* $\mathbb{S}_d^+$ *y* $\mathbb{S}_d^-$ *respectivamente la semiesfera superior y inferior, se puede parametrizar* $s \in \mathbb{S}_d^{\pm}$ *bajo la forma* $s = \left[b^t \quad \pm\sqrt{1-\|b\|^2}\right]^t$, $b \in \mathbb{B}_{d-1}$*, lo que da, siguiendo (Gradshteyn & Ryzhik, 2015, ec. 4.644) (cambio de variables) y notando* μ_L *la medida de Lebesgue,*

$$\begin{aligned}\Phi_U(\omega) &= \frac{2\,\Gamma\left(\frac{d}{2}\right)}{\pi^{\frac{d}{2}}}\int_{\mathbb{B}_{d-1}} \frac{e^{\imath\,\omega^t s}}{\sqrt{1-\|b\|^2}} d\mu_L(b) \\ &= \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)}\int_0^\pi e^{\imath\,\|\omega\|\cos\theta}\,\sin^{d-2}\theta\,d\theta\end{aligned}$$

---

[85] Según Schœnberg (Schœnberg, 1938, Nota de pie 9) and Watson (Watson, 1922, p. 24, nota de pie *), debería llamarse integral de Poisson proque fue introducida primariamente por Poisson en 1823 (Poisson, 1823), pero apareció implicitamente aún antes en trabajos de Euler (Euler, 1769, Cap. X, § 1036).

[86] Para $S \subset \mathbb{S}$ $\mu(S) = |S|$.

Al final, de la forma de la función de Bessel (Gradshteyn & Ryzhik, 2015, Ec. 8.411-7) (ver también (Abramowitz & Stegun, 1970; Watson, 1922; Gray & Mathew, 1895)), se obtiene

$$\Phi_U(\omega) = \frac{2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)}{\|\omega\|^{\frac{d}{2}-1}} J_{\frac{d}{2}-1}(\|\omega\|)$$

lo que cierra la prueba. Volveremos a esta función caracteristica tratando de las coordenadas esféricas.

La forma de la función característica tiene varias consecuencias. La primera es que se puede escribir estocaticamente un vector a simetría elíptica a partir de un vector esféricamente invariante de varias maneras, entre otros:

Corolario 1-14. *Sea $Y \sim \mathcal{ED}(0, I, \varphi_Y)$, $m \in \mathbb{R}^d$ y $\Sigma \in P_d^+(\mathbb{R})$. Sean $\Sigma = Q\Delta Q^t$ la descomposición diagonal de Σ, $\Sigma^{\frac{1}{2}} = Q\Delta^{\frac{1}{2}}Q^t$ única matriz de $P_d^+(\mathbb{R})$ raiz cuadrada de Σ, y $\Sigma = LL^t$ descomposición de Cholesky de Σ, con L triangular inferior (Horn & Johnson, 2013; Bhatia, 2007). Entonces*

$$Q\Delta^{\frac{1}{2}}Y + m \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}}Y + m \overset{\mathrm{d}}{=} LY + m \sim \mathcal{ED}(m, \Sigma, \varphi_Y)$$

Demostración. El resultado es consecuencia del teorema 1-34 y de la forma de la función característica del teorema 1-55. □

Una otra consecuencia es que, dado el orden, el tensor de los momentos tiene una estructura dada para cualquier ley, bajo un factor escalar que depende de la ley. Eso vale también para los cumulantes:

Teorema 1-56 (Momentos centrales y cumulantes)**.** *Sea X vector aleatorio d-dimensional de distribución elíptica en torno a un vector m, y de matriz característica $\Sigma \in P_d^+(\mathbb{R})$. Entonces, si estos momentos y cumulantes existen, m es la media de X, i. e.,*

$$\zeta_1[X] = 0, \qquad \kappa_1[X] = m$$

y para cualquier orden superior a 2, los momentos centrales y cumulantes de orden impares son ceros,

$$\zeta_{2k+1}[X] = \kappa_{2k+1}[X] = 0, \qquad k \geq 1$$

y los de orden pare son dados para $k \geq 1$, por

$$\zeta_{2k}[X] = \alpha_k(\varphi_X)\,\mathcal{T}_k(\Sigma) \qquad y \qquad \kappa_{2k}[X] = \alpha_k(\psi_X)\,\mathcal{T}_k(\Sigma)$$

con el coefficient α_k dado por

$$\alpha_k(f) = (-2)^k f^{(k)}(0)$$

y el tensor $\mathcal{T}_k(\Sigma)$ de orden $2k$ de componentes

$$\mathcal{T}_{i_1,\ldots,i_{2k}}(\Sigma) = \sum_{\pi \in \Pi_{2k,2}} \prod_{(l,n)\in\pi} \Sigma_{i_l,i_n}$$

donde $\Pi_{2k,2}$ es el conjunto de particiones por pares[87] *de $\{1, \ldots, 2k\}$.*

---

[87] Por ejemplo, $\Pi_{4,2} = \Big\{\{\{1,2\},\{3,4\}\},\ \{\{1,3\},\{2,4\}\},\ \{\{1,4\},\{2,3\}\}\Big\}$; cuando $\pi = \big\{\{1,3\},\{2,4\}\big\}$, el término del producto es $\Sigma_{i_1,i_3}\Sigma_{i_2,i_4}$.

Una prueba es dada por inducción en (Berkane & Bentler, 1986) para los momentos (ver también (Fang et al., 1990, p. 44) para el resultat con los momentos). Damos una prueba más directa, valida para ambos momentos y cumulantes.

Demostración. Recordamosnos que para $k \in \mathbb{N}^*, \quad (i_1, \ldots, i_k) \in \{1, \ldots, d\}^k$,

$$\zeta_{i_1,\ldots,i_k}[X] = (-\imath)^k \left.\frac{\partial^k \Phi_{X-m_X}}{\partial\omega_{i_1}\cdots\partial\omega_{i_k}}\right|_{\omega=0} \qquad \text{y} \qquad \kappa_{i_1,\ldots,i_k}[X] = (-\imath)^k \left.\frac{\partial^k \Psi_X}{\partial\omega_{i_1}\cdots\partial\omega_{i_k}}\right|_{\omega=0}$$

Por definición de los momentos centrales $\zeta_1 = 0$. Luego, de $\Psi_X(\omega) = \imath\,\omega^t m + \log\left(\varphi_X(\omega^t\omega)\right)$ tenemos

$$\nabla_\omega \Psi_X(\omega) = \imath\, m + \frac{2\,\varphi'_X(\omega^t\omega)}{\varphi'_X(\omega^t\omega)}\,\omega$$

Recordandose que $\Phi_X(0) = \mathrm{E}\left[e^{\imath\, 0^t X}\right] = 1$ y $m = -\imath\,\nabla_\omega \Psi_X(0)$: m es la media de X lo que corresponde a la intuición.

Luego, salimos de la fórmula de Hardy, extensión de la fórmula de Faà di Bruno, que vimos sección 1.8.4 que recordamos: Para $h(\omega) = f\left(g(\omega)\right), \quad \forall\, n \in \mathbb{N}^*, \quad \forall\, (i_1, \ldots, i_n) \in \{1, \ldots, d\}^n$,

$$\frac{\partial^n h}{\partial\omega_{i_1}\cdots\partial\omega_{i_n}} = \sum_{\pi\in\Pi_n} f^{(|\pi|)}\left(g(\omega)\right) \prod_{B\in\pi} \frac{\partial^{|B|} g}{\prod_{j\in B}\partial\omega_{i_j}}$$

con Π_n el conjunto de las particiones de $\{1, \ldots, n\}$ y $f^{(l)}$ la l-ésima derivada de f. En la expresión de los momentos centrales y cumulantes tenemos

$$g(\omega) = \sum_{i,j=1}^{d} \omega_i \omega_j \Sigma_{i,j}$$

así que, por simetría de Σ,

$$\frac{\partial g}{\partial\omega_{j_1}} = 2\sum_{l=1}^{d} \omega_{j_1}\Sigma_{j_1,l}, \qquad \frac{\partial^2 g}{\partial\omega_{j_1}\partial\omega_{j_2}} = 2\Sigma_{j_1,j_2}, \qquad \forall\, n \geq 3, \quad \frac{\partial^n g}{\prod_{l=1}^n \partial\omega_{j_l}} = 0$$

Es decir que, para $n \geq 1$,

$$\left.\frac{\partial^n g}{\prod_{l=1}^n \partial\omega_{j_l}}\right|_{\omega=0} = \begin{cases} 2\Sigma_{j_1,j_2} & \text{si} \quad n = 2 \\ 0 & \text{si} \quad n \neq 2 \end{cases}$$

Entonces, en la formula de Hardy tomada en $\omega = 0$ quedan solas las particiones que contienen unicamente pares de indices. Eso da momentos centrales y cumulantes nulos para k impar (obvio por simétria). Ademaás, siendo $\Pi_{2k,2}$ el conjunto de particiones por pares de $\{1, \ldots, 2k\}$, notando que necesariamente cada partición de $\Pi_{2k,2}$ contiene k pares,

$$\left.\frac{\partial^{2k} h}{\partial\omega_{i_1}\cdots\partial\omega_{i_{2k}}}\right|_{\omega=0} = \sum_{\pi\in\Pi_{2k,2}} f^{(k)}(0) \prod_{(l,n)\in\pi} (2\Sigma_{i_l,i_n})$$

La prueba se cierra tomando respectivamente $f = \varphi_X$ y $f = \psi_X$. □

Veremos más adelante una prueba aún más directa y obvia que, a orden dado, el tensor de los momentos centrales tiene una estructura dada bajo un factor escalar dependiendo de la ley. Lo que es

menos obvio, de la relación momentos-cumulantes, que eso vale para el tensor de los cumulantes, y que, más que eso, estas estructuras son las mismas.

De este resultado se puede también explicitar la matriz de covariancia y el tensor curtosis para un vector a simetría elíptica (Fang et al., 1990, p. 44):

Corolario 1-15. *Sea X vector aleatorio d-dimensional de distribución elíptica de matriz característica $\Sigma \in P_d^+(\mathbb{R})$. Entonces*

$$\Sigma_X = -2\,\varphi_X'(0)\,\Sigma$$

y

$$\kappa_X = \frac{\varphi_X''(0)}{4\left(\varphi_X(0)\right)^2}\sum_{i,j=1}^{d}\Big(\left(\mathbb{1}_i\mathbb{1}_i^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_j^t\right)+\left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_i\mathbb{1}_j^t\right)+\left(\mathbb{1}_i\mathbb{1}_j^t\right)\otimes\left(\mathbb{1}_j\mathbb{1}_i^t\right)\Big)$$

Demostración. El resultado es inmediato por lo de la covarianza. Para la curtosis, momento central de orden cuatro del vector normalizado, es equivalente considerar $\Sigma = -\frac{1}{2\,\varphi_X(0)}I$, lo que da el resultado del corrolario. □

Nota: de $\mathrm{Tr}(\Sigma_X) = \mathrm{Tr}\left(\mathrm{E}\left[(X-m_X)(X-m_X)^t\right]\right) = \mathrm{E}\left[(X-m_X)^t(X-m_X)\right]$ tenemos para $U \sim \mathcal{U}(\mathbb{S}_d)$, $\mathrm{Tr}(\Sigma_U) = 1$, *i. e.,*

$$\text{Para } U \sim \mathcal{U}(\mathbb{S}_d), \quad \Sigma_U = \frac{1}{d}I$$

Se puede ver de este corolario que la covarianza es proporcional a Σ. Es decir que un vínculo que se puede poner también para fijar (Σ, φ_X) es de imponer un homotecía a φ_X para tener $\varphi_X'(0) = -\frac{1}{2}$, para que Σ y Σ_X coincidan.

De la forma de la función característica se proba también que cualquier transformación lineal de un vector a simetría elíptica da un vector a simetría elíptica:

Teorema 1-57. *Sea $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$ d-dimensional, $A \in \mathcal{M}_{d',d}(\mathbb{R})$ de rango lleno tal que $d' \leq d$ y $c \in \mathbb{R}^{d'}$. Entonces*

$$AX + c \sim \mathcal{ED}(Am + c, A\Sigma A^t, \varphi_X)$$

Demostración. El resultado es inmediato de la forma de la función característica, teorema 1-55 y como consecuencia del teorema 1-34. □

En particular, la proyección de un vector a simetría elíptica queda elíptica en torno a la proyección del vector posición, con la misma generadora característica.

Eso víncula también un vector a simetría elíptica con cualquier proyección:

Teorema 1-58 (Proyección y componentes)**.** *Sea X, vector aleatorio d-dimensonal de componentes X_i. Entonces*

$$X \sim \mathcal{ED}(m, \Sigma, \varphi_X) \qquad \Longleftrightarrow \qquad \forall\, a \in \mathbb{R}^d, \quad a^t(X-m) \overset{\mathrm{d}}{=} \sqrt{\frac{a^t\Sigma a}{\Sigma_{i,i}}}(X_i - m_i)$$

Demostración. Sea $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$, entonces

$$\Phi_{a^t(X-m)}(\omega) = \Phi_{X-m}(\omega a) = \varphi_X\left(\omega^2 a^t \Sigma a\right) = \Phi_{X_i-m_i}\left(\omega\sqrt{\frac{a^t\Sigma a}{\Sigma_{i,i}}}\right) = \Phi_{\sqrt{\frac{a^t\Sigma a}{\Sigma_{i,i}}}(X_i-m_i)}(\omega)$$

(se puede sacar el valor absoluto a ω porque, necesariamente, $X_i - m_i$ es a simetría elíptica, es decir $(X_i - m_i) \overset{\mathrm{d}}{=} -(X_i - m_i)$). Reciprocamente, si para cualquier $a \in \mathbb{R}^d$ tenemos $a^t(X - m) \overset{\mathrm{d}}{=} \sqrt{\frac{a^t\Sigma a}{\Sigma_{i,i}}}(X_i - m_i)$, necesariamente

$$\Phi_{X-m}(a) = \mathrm{E}\left[e^{\imath\, a^t(X-m)}\right] = \mathrm{E}\left[e^{\imath\sqrt{\frac{a^t\Sigma a}{\Sigma_{i,i}}}(X_i-m_i)}\right] = \Phi_{X_i-m_i}\left(\sqrt{\frac{a^t\Sigma a}{\Sigma_{i,i}}}\right)$$

Es una función de $a^t\Sigma a$, lo que cierra la prueba por el teorema 1-55. □

Vimos en las secciones 1.10.2.9 y 1.10.2.10, tratando de vectores de distribución Student-t y -r, situaciones en las cuales la matriz de covarianza es (propporcional a) la identidad, mientras que las componentes del vector no son independientes, contrariamente a lo que pasa en el caso gaussiano. De hecho este resultado es general en el marco de vectores a simétria elíptica (Bilodeau & Brenner, 1999; Maxwell, 1867):

Teorema 1-59 (Maxwell-Hershell)**.** *Sea $X \sim \mathcal{ED}(m, I, \varphi_X)$. Las componentes X_i son independientes si y solamente si $X \sim \mathcal{N}(m, \alpha I)$ con $\alpha > 0$. En otros terminos, los solos vectores a simetría esférica en torno a un vector m son gaussianos de covarianza proporcional a la identidad y de media m.*

Demostración. Del teorema (?, ?), $\begin{bmatrix} X_1 & X_2 \end{bmatrix}^t \sim \mathcal{ED}\left(\begin{bmatrix} m_1 & m_2 \end{bmatrix}^t, I, \varphi_X\right)$ y $X_i \sim \mathcal{ED}(m_i, 1, \varphi_X)$. Si estas variables son independientes (condición necesaria), buscamos entonces φ_X tal que $\forall\, \omega \in \mathbb{R}^2, \quad \varphi_X\left(\omega_1^2 + \omega_2^2\right) = \varphi_X\left(\omega_1^2\right)\varphi_X\left(\omega_2^2\right)$, *i. e.,* por reparametrización

$$\forall\, (u, v) \in \mathbb{R}_+^2, \qquad \varphi_X(u + v) = \varphi_X(u)\varphi_X(v)$$

La sola función continua satisfaciendo este morfismo es la función exponencial, es decir $\varphi_X(u) = \exp(au)$. La función debe ser una generadora caractéristica, asi que necesariamente $a < 0$, que podemos escribir $a = -\frac{\alpha}{2}$ con $\alpha > 0$, lo que cierra la prueba. □

Una consecuencia importante de la forma de la función característica es que sirve a probar que un vector a simetría esférica se escribe estocasticamente como una mezcla de escala de un vector uniforme sobre la esfera unitaria $\mathbb{S}_d$. Este resultado es debido a Schœnberg (Schœnberg, 1938; Fang et al., 1990) (ver también (Keilson & Steutel, 1974; Teicher, 1960)) y se enuncia como sigue:

Teorema 1-60 (Mezcla de escala de base uniforme)**.** *Sea X d-dimensional a simetria esférica. Entonces, este vector se escribe estocasticamente como*

$$X \overset{\mathrm{d}}{=} RU \qquad \textit{con} \qquad U \sim \mathcal{U}(\mathbb{S}_d), \quad R > 0 \quad \textit{independientes}$$

Más generalmente, para $Y \sim \mathcal{ED}(m, \Sigma, \varphi_Y)$,

$$Y \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}} RU + m$$

Además, inmediatamente

$$R \stackrel{\mathrm{d}}{=} \|X\|$$

(obviamente, $\frac{X}{\|X\|} \stackrel{\mathrm{d}}{=} U$*).*

Demostración. Escribimos $\omega = wu$ con $w \geq 0$ y $u \in \mathbb{S}_d$. Entonces, con μ_H la medida de Haar sobre la esfera y P_X la medida de probabilidad de X, tenemos

$$\begin{aligned}
\Phi_X(\omega) &= \varphi_X\left(w^2\right) \\
&= \frac{1}{|\mathbb{S}_d|}\int_{\mathbb{S}_d} \varphi_X\left(w^2\right)\, d\mu_H(v) \\
&= \frac{1}{|\mathbb{S}_d|}\int_{\mathbb{S}_d} \Phi_X(wv)\, d\mu_H(v) \\
&= \frac{1}{|\mathbb{S}_d|}\int_{\mathbb{S}_d} \left(\int_{\mathbb{R}^d} e^{\imath\, wv^t x}\, dP_X(x)\right) d\mu_H(v) \\
&= \int_{\mathbb{R}^d} \left(\int_{\mathbb{S}_d} e^{\imath\, wv^t x}\, \frac{1}{|\mathbb{S}_d|}\, d\mu_H(v)\right) dP_X(x) \\
&= \int_{\mathbb{R}^d} \Phi_U(wx)\, dP_X(x) \\
&= \int_{\mathbb{R}^d} \varphi_U\left(w^2\|x\|^2\right)\, dP_X(x)
\end{aligned}$$

con Φ_U función caracteristica de un vector $U \sim \mathcal{U}(\mathbb{S}_d)$ y φ_U la generadora característica correspondiente. Sea $F_R(r) = 0$ para $r \leq 0$ y

$$F_R(r) = \int_{\mathbb{B}(0,r)} dP_X(x)$$

si no. Claramente F_R es creciente de 0 a 1: es una función de repartición. Notando R la variable aleatoria positiva de función de repartición F_R y P_R la medida de probabilidad asociada,

$$\begin{aligned}
\Phi_X(\omega) &= \int_{\mathbb{R}_+} \varphi_U(w^2 r^2)\, dP_R(r) \\
&= \int_{\mathbb{R}_+} \Phi_U(r\omega)\, dP_R(r) \\
&= \mathrm{E}\left[\mathrm{E}\left[e^{\imath\, \omega^t RU}\middle| R\right]\right]
\end{aligned}$$

el paso de la anteúltima a la última linea siendo valido para R independiente de U. En otros terminos, con R de medida de probabilidad P_R y $U \sim \mathcal{U}(\mathbb{S}_d)$ independiente de R, tenemos del teorema de esperanza total 1-23,

$$\Phi_X(\omega) = \Phi_{RU}(\omega)$$

La prueba se cierra de la relación uno-uno entre la medida de probabilidad de un vector aleatorio y la función característica. De $X \stackrel{\mathrm{d}}{=} RU$ y $\|U\| = 1$ viene $R \stackrel{\mathrm{d}}{=} \|X\|$. □

Se puede referirse también a (Bilodeau & Brenner, 1999, Prop. 4.10) para tener una prueba alternativa basado sobre el teorema de Cramér-Wold, Teo. 1-35.

De esta escritura, se puede ir un paso más allá (Fang et al., 1990, Teo 2.3):

Corolario 1-16. *Sea X d-dimensional a simetria esférica. Entonces, este vector se escribe estocasticamente como*

$$X = \|X\| \, \frac{X}{\|X\|}$$

tales que $\|X\| \overset{\mathrm{d}}{=} R$ y $\frac{X}{\|X\|} \overset{\mathrm{d}}{=} U \sim \mathcal{U}(\mathbb{S}_d)$ son independientes.

Demostración. Se aplica el teorema 1-37 a $f(x) = \begin{bmatrix} \|x\| \\ \frac{x}{\|x\|} \end{bmatrix}$ con $X \overset{\mathrm{d}}{=} Y = R\,U$. □

.

De la escritura $X \overset{\mathrm{d}}{=} RU$, se puede ver $X \sim \mathcal{ED}(0, I, \varphi_X)$ como muñecas rusas: a cada escala o capa $R = r$ tenemos una distribución uniforma sobre la esfera de este rayo r. Por eso, X es también dicho *mezcla de escala* de una *base* uniforme y R es llamada *variable generadora* con respecto a esta base. En esta situación, llamaremos también la variable R *rayo*.

De esta escritura, queda ahora claro que cada tensor de momentos centrales tienen una estructura fija: de $X = R\Sigma U$ tenemos $\zeta_l[X] = \mathrm{E}\left[R^l\right] \zeta_l[\Sigma U]$ (ver ej. (Fang et al., 1990, teo. 2.8) para $\Sigma \propto I$): la estructura, común a cualquier vector aleatorio de $\mathcal{ED}(m, \Sigma, \varphi_X)$ es dada por $\zeta_k[\Sigma U]$ (cero cuando $l = 2k + 1$, por simetría central) y el factor, dependiente de la ley es dada por el momento del "rayo" R. Además, se puede dar una otra forma del coefficiente α_k:

Lema 1-56. *El coeficiente $\alpha_k(\varphi_X)$ del tensor de los momentos centrales del teorema 1-56 es también dado por*

$$\alpha_k(\varphi_X) = \frac{\Gamma\left(\frac{d}{2}\right)}{2^k \Gamma\left(\frac{d}{2} + k\right)} \, \mathrm{E}\left[\left((X - m)^t \Sigma^{-1} (X - m)\right)^k\right]$$

Demostración. Por ejemplo, del desarollo de Taylor de la función de Bessel dado en (Gradshteyn & Ryzhik, 2015), aplicado a $\varphi_U(u) = 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) u^{-\frac{d-2}{4}} J_{\frac{d}{2}-1}\left(\sqrt{u}\right)$ se obtiene

$$\varphi_U^{(k)}(0) = \frac{(-1)^k \Gamma\left(\frac{d}{2}\right)}{4^k \Gamma\left(\frac{d}{2} + k\right)}$$

A continuación, para $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$, de $X - m = R\,\Sigma\,U$ (R es escalar) y de la formula del tensor de momentos centrales se obtiene

$$\zeta_{2k}[X] = \mathrm{E}\left[R^{2k}\right] \zeta_{2k}[\Sigma U] = \mathrm{E}\left[R^{2k}\right] \alpha_k(\varphi_U) \mathcal{T}_k(\Sigma)$$

es decir

$$\alpha_k(\varphi_X) = (-2)^k \varphi_U^{(k)}(0) \, \mathrm{E}\left[R^{2k}\right]$$

Ahora, $(X - m)^t \Sigma^{-1} (X - m) = R^2 U^t U = R^2$, lo que cierra la prueba. □

Fijense que un vector a simetría elíptica no admite necesariamente una densidad con respecto a la medida de Lebesgue, como por ejemplo en el caso de un vector uniforme sobre $\mathbb{S}_d$ (pero esa tiene una densidad, constante, con respecto a la medida de Haar sobre la esfera). Un otro ejemplo puede ser $X = BG$ con $B \sim \mathcal{B}(p)$ y $G \sim \mathcal{N}(0, I)$ independiente de B. $\Phi_X(\omega) = p\, e^{-\frac{\|\omega\|^2}{2}} + 1 - p$ no es L_1, *i. e.,* no tiene una transformada de Fourier inversa usual. Sin embargo, cuando un vector elíptico admite una densidad, esa tiene propiedades remarcables también.

Teorema 1-61. *Sea $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$. Si admite una densidad*[88]*, entonces esta densidad tiene la forma*

$$p_X(x) = |\Sigma|^{-\frac{1}{2}} d_X\left((x-m)^t \Sigma^{-1}(x-m)\right)$$

donde $d_X : \mathbb{R}_+ \mapsto \mathbb{R}_+$ escalar, es llamada generadora de densidad*. Reciprocamente, si la densidad tiene esta forma, X es a simetría elíptica.*

Demostración. Sin perdida de generalidad, se puede considerar $X \sim \mathcal{ED}(0, I, \varphi_X)$ y recuperar el caso general por cambio de variables. Entonces, por cambio de variables (ver sección 1.4) tenemos para cualquier matriz ortogonal O y cualquier x

$$p_X(x) = p_{OX}(x) = |O|^{-\frac{1}{2}} p_X\left(O^{-\frac{1}{2}} x\right) = p_X\left(O^{-\frac{1}{2}} x\right)$$

siendo $|O| = 1$ (Bhatia, 1997; Horn & Johnson, 2013). $O^{-\frac{1}{2}}$ es también una matriz de rotación, probando que p_X queda invariante bajo cualquier rotación de su argumento, es decir que depende solamente de la norma de x. □

De este resultado, cuando X admite una densidad de probablidad, se escribe tambien

$$X \sim \mathcal{ED}(m, \Sigma, d_X)$$

aún que puede ser confuso.

Claramente, para $X \sim \mathcal{ED}(m, \Sigma, d_X)$, los niveles de probabildad $\{x \mid p_X(x) = c\} = \left\{x \mid (x-m)^t \Sigma^{-1}(x-m) = c\right\}$ son elipsiodes centrados en m, de direcciones y estiramientos respectivamente dados por los autovectores y autovalores de Σ. Eso justifica también la denominación "X a simetría elíptica".

Como lo vimos, de una forma, todas las características estádisticas de X es en el "rayo" $R \overset{\mathrm{d}}{=} \sqrt{(X-m)^t \Sigma^{-1}(X-m)}$. En lo que sigue, sin perdida de generalidad, se puede concentrarse en $X \sim \mathcal{ED}(0, I, d_X)$.

Primero, se puede escribir la ley de R a partir de d_X:

---

[88] De hecho, sufice que admita una densidad con respecto a una medida que depende solamente del volumen, como la de Lebesgue pero también, de Haar.

Teorema 1-62 (Densidad del rayo)**.** *Sea* $X \sim \mathcal{ED}(0, I, d_X)$ *y* $R \overset{\mathrm{d}}{=} \|X\|$. R *admite también una densidad que se escribe*

$$p_R(r) = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} r^{d-1} d_X\left(r^2\right)$$

Demostración. Como lo hemos visto en la prueba del teorema 1-60, la función de repartición de R se escribe

$$F_R(r) = \int_{\mathbb{B}_d(0,r)} dP_X(x)$$

es decir, P_X admitiendo una densidad

$$F_R(r) = \int_{\mathbb{B}_d(0,r)} d_X\left(x^t x\right)\, dx$$

lo que da, de (Gradshteyn & Ryzhik, 2015, Ec. 4.642)

$$F_R(r) = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^r \rho^{d-1} d_X\left(\rho^2\right)\, d\rho$$

F_R es continua y diferenciable, llegando al resultado del teorema. Volveremos a este resultado más adelante tratando de las coordenadas esféricas. □

De este resultado, se puede ahora dar la relación que existe entre φ_X y d_X siendos Φ_X y p_X relacionados por transformada de Fourier.

Teorema 1-63. *Sea* $X \sim \mathcal{ED}(m, \Sigma, \varphi_X) \equiv \mathcal{ED}(m, \Sigma, d_X)$ *las generadoras características y de densidad son relacionadas por*

$$\varphi_X\left(w^2\right) = (2\pi)^{\frac{d}{2}} w^{1-\frac{d}{2}} \int_{\mathbb{R}_+} r^{\frac{d}{2}} d_X\left(r^2\right)\, J_{\frac{d}{2}-1}(rw)\, dr$$

y reciprocamente

$$d_X\left(r^2\right) = (2\pi)^{-\frac{d}{2}} r^{1-\frac{d}{2}} \int_{\mathbb{R}_+} w^{\frac{d}{2}} \varphi_X\left(w^2\right)\, J_{\frac{d}{2}-1}(rw)\, dw$$

La transformación dando $w^{\frac{d}{2}-1}\varphi_X\left(w^2\right)$ *a partir de* $r^{\frac{d}{2}-1} d_X\left(r^2\right)$ *es conocida como* transformada de Hankel *de orden* $\frac{d}{2}-1$ *(Schœnberg, 1938; Schwartz, 1969, 1971; Lord, 1954; Poularikas, 1999; Poularikis, 2010). A veces, la transformación dando* $\varphi_X\left(w^2\right)$ *a partir de* $d_X\left(r^2\right)$ *es llamada* transformada de Hankel modificada *de orden* $\frac{d}{2}-1$.

Más generalmente, φ_X *es relacionada a la medida de probabilidad* P_R *del rayo por transformación dicha de Hankel-Stieltjes (modificada) (Schœnberg, 1938; Nussbaum, 1973; Cholewinski, Haimo & Nussbaum, 1970; Schwartz, 1971)*

$$\varphi_X(w^2) = 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right) w^{1-\frac{d}{2}} \int_{\mathbb{R}_+} r^{1-\frac{d}{2}} J_{\frac{d}{2}-1}(rw)\, dP_R(r)$$

Demostración. Hay varias pruebas de este resultado. Una prueba elegante se basa sobre la generadora caracteristica de la variable $U \sim \mathcal{U}(\mathbb{S}_d)$. Primero, sin perdida de generalidad, consideramos

$m = 0,\ \Sigma = I$. Entonces, de la escritura $X \stackrel{\mathrm{d}}{=} RU$ tenemos

$$\begin{aligned}\Phi_X(\omega) &= \mathrm{E}\left[e^{\imath\, R\omega^t U}\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[e^{\imath\, R\omega^t U}\middle| R\right]\right] \\ &= \mathrm{E}\left[\Phi_U\left(R^2\|\omega\|^2\right)\right]\end{aligned}$$

Es decir, del ejemplo 1-24 se obtiene

$$\varphi_X\left(w^2\right) = 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)\int_{\mathbb{R}_+}(rw)^{-\frac{d-2}{2}}J_{\frac{d}{2}-1}\left(rw\right)\,dP_R(r)$$

transformada de Hankel-Stieltes (modificada) de P_R. A continuación del teorema 1-62 se obtiene

$$\varphi_X\left(\|\omega\|^2\right) = 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)\int_{\mathbb{R}_+}(r\|\omega\|)^{-\frac{d-2}{2}}\,J_{\frac{d}{2}-1}\left(r\|\omega\|\right)\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}r^{d-1}\,d_X\left(r^2\right)\,dr$$

Eso da la expresión de φ_X como transformada de Hankel de d_X. Para la transformación inversa, sufice recordarse que Φ_X siendo transformada de Fourier de p_X, tenemos p_X tranformada inversa de Φ_X. Luego, por simetría (cambio de variables $w \to -w$, esta transformada inversa es nada mas que la transformada directa, por un factor $(2\pi)^{-d}$ (ver teoremas 1-32). □

Tratando de un vector a simetría esférica, resuelte frecuentemente más comodo tratarlo en su representación en coordenadas hiperesféricas, es decir, para $d \geq 2$

$$x_i = r\left(\prod_{k=1}^{i-1}\sin\theta_k\right)\cos\theta_i, \quad 1 \leq i \leq d$$

con

$$(r, \theta_1, \ldots, \theta_{d-1}) \in \mathbb{R}_+ \times [0\,;\,\pi)^{d-2} \times [0\,;\,2\pi)$$

y las convenciones

$$\theta_d = 0, \quad \prod_{k=1}^{0} = 1$$

Los parámetros de las coordenadas hiperesféricas (rayo, angulos) son representadas en las figuras 1-33(a) para $d = 2$, y 1-33(b) para $d = 3$.

En el caso de vector a simetría esférica, aparece que el rayo R y los angulos son independientes y se puede calcular cada distribución (Fang et al., 1990; Lord, 1954; ?, ?, ?, ?):

Teorema 1-64. *Sea $X \sim \mathcal{ED}(0, I, d_X)$ y su representación en coordenadas hiperesféricas $X_i = R\left(\prod_{k=1}^{i-1}\sin\Theta_k\right)\cos\Theta_i, \quad 1 \leq i \leq d$. Entonces R y los $\Theta_i,\ 1 \leq i \leq d-1$ son independientes y*

$$p_R(r) = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}r^{d-1}d_X\left(r^2\right)$$

$$p_{\Theta_i}(\theta_i) = \frac{\Gamma\left(\frac{d-j+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-j}{2}\right)}\left(\sin\theta_i\right)^{d-i-1}\mathbb{1}_{[0\,;\,\pi)}(\theta_i), \quad 1 \leq i \leq d-2, \qquad p_{\Theta_{d-1}}(\theta_{d-1}) = \frac{1}{2\pi}\,\mathbb{1}_{[0\,;\,2\pi)}(\theta_{d-1})$$

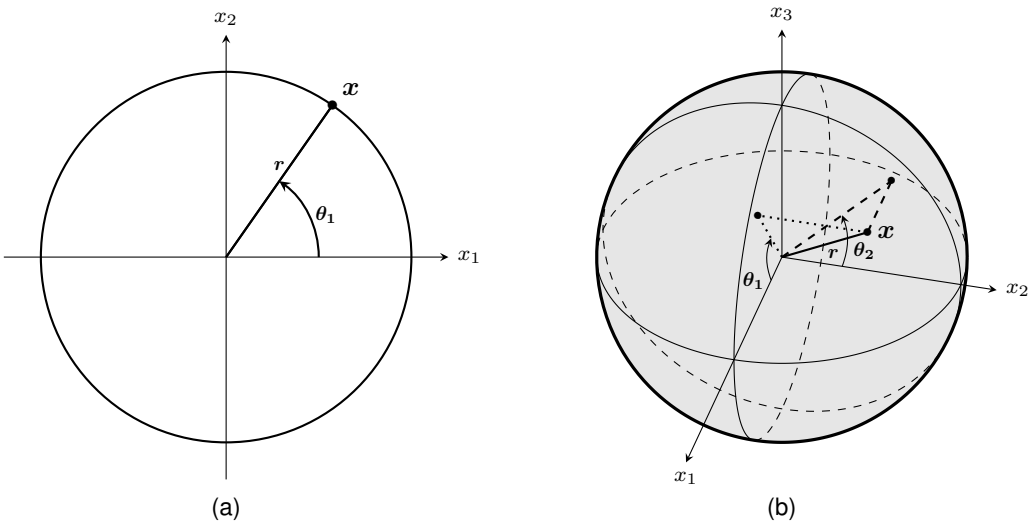


Figura 1-33: Coordenadas hiperesféricas para un punto dado de $\mathbb{R}^d$. (a) caso $d = 2$ y (b) caso $d = 3$.

Demostración. Sea la transformación $g : (x_1, \ldots, x_d) \mapsto (r, \theta_1, \ldots, \theta_{d-1})$. La jacobiana de g^{-1} tiene la forma

$$\mathrm{J}_{g^{-1}} = \begin{bmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta_1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ \frac{\partial x_{d-1}}{\partial r} & \frac{\partial x_{d-1}}{\partial \theta_1} & \cdots & & \frac{\partial x_{d-1}}{\partial \theta_{d-1}} \\ \frac{\partial x_d}{\partial r} & \frac{\partial x_d}{\partial \theta_1} & \cdots & \cdots & \frac{\partial x_d}{\partial \theta_{d-1}} \end{bmatrix}$$

Desarrolando el determinante por la última columna por ejemplo (ver (Bhatia, 1997; Horn & Johnson, 2013)), y así por inducción se obtiene

$$\left|\mathrm{J}_{g^{-1}}\right| = r^{d-1} \prod_{j=1}^{d-2} (\sin\theta_i)^{d-j-1}$$

Entonces, por transformación (ver sección 1.4) tenemos

$$f_{R,\Theta_1,\ldots,\Theta_{d-1}}(r, \theta_1, \ldots, \theta_{d-1}) = d_X\left(r^2\right) r^{d-1} \prod_{j=1}^{d-2} \left((\sin\theta_i)^{d-j-1} \, \mathbb{1}_{[0\,;\,\pi)}(\theta_i) \right) \, \mathbb{1}_{[0\,;\,2\pi)}(\theta_{d-1})$$

Claramente se factoriza probando la independencia y la forma de cada distribución marginal. El factor viene de la normalización (ver (Gradshteyn & Ryzhik, 2015, Ec. 8.380-2) para los angulos, lo que determina necesariamente el del rayo). □

Obviamente, se recupera la ley de $R \overset{\mathrm{d}}{=} \|X\|$ que encontramos. Adémas, estas coordenadas hiperesféricas, a $r = 1$, parametriza $\mathbb{S}_d$ y permite por ejemplo de probar la independencia entre $\|X\|$ y $\frac{X}{\|X\|}$ (y entonces la escritura en mezcla), de calcular directamente $\Phi_U(\omega)$ para $U \sim \mathcal{U}(\mathbb{S}_d)$, o de vincular las generadoras característica y de densidad entre otros.

En la familia elíptica, hay una subfamilia particular que queda invariante por marginalización, y extensión, como la gaussiana por ejemplo:

Definición 1-52 (Consistencia (Yao, 1973; Kano, 1994))**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}$ *a simetría elíptica de generadora de densidad* d_X*. Este vector es dicho* consistente *si la generadora de densidad* $d_{X'}$ *de* $X' = \begin{bmatrix} X_1 & \cdots & X_{d'} \end{bmatrix}$ *tiene la misma forma que* d_X *donde el parámetro dimensional* d *es reemplazado por* d'*. En otros termoinos tenemos invarianza por marginalización si* $d' < d$*, pero se puede ver* X *como marginal de un vector más grande con la misma generadora de probabilidad reemplazando* d *por* d'*.*

Se proba que, equivalentemente, la generadora caracteristica φ_X *no es relacionada a* d *(ver (Kano, 1994; Fang et al., 1990) para la prueba). Entonces,* φ_X *va a ser una generadora característica de un vector aleatorio de cualquier dimensión* $d \in \mathbb{N}^*$*, o, con la caracterisación por functiones definida no negativa (ver teorema de Bocher),* $\varphi_X \in \mathfrak{EP}$ *(ver notaciones, y las a continuación del teorema 1-55).*

Si X a simetrá esférica (y por transformación afín a simetría elíptica) se escribe siempre como mezcla de escala de base uniforme, esta escritura estocástica no es única. Por ejemplo, si consideramos $X = A\,G$, con A variable aleatoria positiva y $G \sim \mathcal{N}(0, I)$ independiente de G, queda claro que X es a simetría esférica. Estos vectores, llamados *mezcla de escala de base gaussiana*, o simplemente *mezcla de escala gaussiana* (GSM para gaussian scale mixture en inglés) fueron estudiados intensivamente de manera formal (Kano, 1994; Yao, 1973; Vershik, 1964; Picinbono, 1970; Kelker, 1971; Kingman, 1972; Keilson & Steutel, 1974; Teicher, 1960; Andrews & Mallows, 1974). Ecuentra aplicaciones en varias area, modelizando la textura de imagenes o datos de radar (ej. suma aleatoria de vectores gaussianos) (Portilla, Strela, Wainwright & Simoncelli, 2003; Bombrun & Beaulieu, 2008; Selesnick, 2008; Shi & Selesnick, 2007; Zozor & Vignat, 2010; Tison, Nicolas, Tupin & Maître, 2004; Todros & Tabrikian, 2007).

Una pregunta natural es de saber si se puede escribir cualquier vector a simetría esférica como mezcla de escala de base gaussiana. La respuesta es negativa. Por ejemplo, del dominio de definición de la variable, queda claro que un vector uniforme sobre la esfera no puede ser mezcla de escala de base gaussiana (ver un otro contra-ejemplo en (Picinbono, 1970)). Escribir un vecor como tal mezcla resuelte posible solamente bajo restricciones. Más precisamente, una condición necesaria y suficiente es que la variable debe ser consistente, como lo probó Schœnberg (Schœnberg, 1938, Teo. 2) (ver también (Steerneman & van Perlo-ten-Kleij, 2005, Teo. 2), (Yao, 1973, Lem. 2.2) o (Kano, 1994, Teo. 1)).

Teorema 1-65 (Schœberg'38 – a partir de la generadora característica)**.** *Sea* $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$*. Entonces* X *es una mezcla de gaussiana si y solamente si* X *es consistente, es decir*

$$X \overset{\mathrm{d}}{=} A\,\Sigma^{\frac{1}{2}}\,G + m \quad \Leftrightarrow \quad \varphi_X \in \mathfrak{EP}$$

con $A > 0$ *independiente de* $G \sim \mathcal{N}(0, I)$*.*

Demostración. La directa es inmediata de la formula de esperanza total

$$\varphi_X(w) = \mathrm{E}\left[\, \mathrm{E}\left[\varphi_G\left(Aw\right) \mid A\right] \,\right]$$

conjuntamente a $\varphi_G : w \mapsto e^{-\frac{w}{2}} \in \mathfrak{EP}$.

La recíproca es más dificil a probar. Para los detalles, dejamos el lector a (Schœnberg, 1938). Los elementos de prueba son los siguientes. De la escritura como mezcla de escala uniforme tenemos entonces para cualquier d

$$\varphi_X\left(w^2\right) = 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)\int_{\mathbb{R}_+}(wr)^{1-\frac{d}{2}}\,J_{\frac{d}{2}-1}(wr)\,dP_{R_d}(r)$$

con R_d la variable generadora de base uniforme correspondiente a la dimension d (ver prueba del teorema 1-60 y ejemplo 1-24). Por cambio de variables se escribe también

$$\varphi_X\left(w^2\right) = \int_{\mathbb{R}_+} 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)\left(wr\sqrt{d}\right)^{1-\frac{d}{2}}\,J_{\frac{d}{2}-1}\left(wr\sqrt{d}\right)\,dP_{R_d}\left(r\sqrt{d}\right)$$

Eso siendo valid para cualquier orden, se nota P_R el límite de P_{R_d} cuando d tiende al infinito. Luego, el desarollo de Taylor de la función de Bessel y de la fórmula de Stirling (?, ?, Ec. 8.402 y 8.327) se proba que

$$\lim_{d\to+\infty} 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)\left(wr\sqrt{d}\right)^{1-\frac{d}{2}}\,J_{\frac{d}{2}-1}\left(wr\sqrt{d}\right) \;=\; e^{-\frac{w^2r^2}{2}}$$

Todo el juego consiste a probar que la convergencia es uniforme, para poder intercambiar límite e integral. Dio una prueba Schœnberg en (Schœnberg, 1938), y luego propuso una "más moderna" Steerneman y van Perlo-ten-Kleij en (Steerneman & van Perlo-ten-Kleij, 2005). Basicamente se reconoce en $w \mapsto 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)\left(w\sqrt{d}\right)^{1-\frac{d}{2}}\,J_{\frac{d}{2}-1}\left(w\sqrt{d}\right)$ la función característica de $q_d(x) \propto \left(1-\frac{x^2}{d}\right)_+^{\frac{d-3}{2}}$ que tiende a la gaussiana (ver por ejemplo sección 1.10.2.10). Se usa el lema de Scheffé o de Riez (ver (Riesz, 1928; Scheffe, 1947; Novinger, 1972; Kusolitsch, 2010) o (Athreya & Lahiri, 2006; Bogachev, 2007a; Billingsley, 2012)) para probar la convergencia de la integral. □

Se encuentra una prueba alternativa también en (Fang et al., 1990; Kingman, 1972).

Este teorema pone en juega la condición necesaria y suficiente sobre la función característica para tener una mezcla de escala gaussiana, así que no es necesario que X admita una densidad de probabilidad. Sin embargo, al imagen de la consistencia que se exprime también a través de la generadora de densidad, el teorema de Schœnberg tiene una versión usando esta generadora. Por eso, introducimos el conjunto de funciones complatemente monotonas sobre $\mathbb{R}_+$ hasta un cierto orden, y para cualquier orden [89] (ver notaciones):

$$\mathfrak{CM}_n = \left\{ f \in C^{n-1}(\mathbb{R}_+) \;\middle|\; \forall\, k = 0,\ldots,n-1,\;\; (-1)^k f^{(k)} \geq 0 \quad\wedge\quad (-1)^k f^{(k)} \text{ es decreciente} \right\}$$

y

$$\mathfrak{CM} = \bigcap_{n=0}^{+\infty} \mathfrak{CM}_n = \left\{ f \in C^{\infty}(\mathbb{R}_+) \;\middle|\; \forall\, k \in \mathbb{N},\;\; (-1)^k f^{(k)} \geq 0 \right\}$$

[89] Una función $f : I \subset \mathbb{R} \mapsto \mathbb{R}$ es dicha *absolutamente monotona* si es continua, diferenciable a todos los ordenes en el interior de I, y todas sus derivadas son positivas, $\forall\, k \in \mathbb{N},\quad f^{(k)} \geq 0$. f es dicha *completamente monotona* si $f(-x)$ es absolutamente monotona (Bernstein, 1929).

Teorema 1-66 (Schœberg'38 – a partir de la generadora de densidad)**.** *Sea* $X \sim \mathcal{ED}(m, \Sigma, d_X)$. *Entonces* X *es una mezcla de gaussiana si y solamente si* d_X *es completamente monotona sobre* $\mathbb{R}_+$

$$X \overset{\mathrm{d}}{=} A\,\Sigma^{\frac{1}{2}}\,G + m \quad \Leftrightarrow \quad d_X \in \mathfrak{CM}$$

con $A > 0$ *independiente de* $G \sim \mathcal{N}(0, I)$.

Demostración. La prueba se apoya sobre el teorema de Haussdorf-Bernstein-Widder que proba la equivalencia entre la clase de funciones completamente monotonas y la de las funciones f que se escriben como una transformada de Laplace-Stieltjes [90] (en el eje real)

$$f(t) = \int_{\mathbb{R}_+} e^{-tu} d\mu(u)$$

donde μ es una medida finita (ver (Schœnberg, 1938, Teo. 3), (Bernstein, 1929; ?, ?, ?, ?), (Widder, 1946, § 12), (Feller, 1971, § XIII.4)). Con el cambio de variables $u = \frac{1}{2a^2}$ y $\mu\big((0\,;\,u)\big) = (2u)^{\frac{d}{2}} P_A\left(\left(\frac{1}{\sqrt{2u}}\,;\,+\infty\right)\right)$, se aplica entonces este teorema a (formula de probabilidad total)

$$d_X(r) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}_+} a^{-d} e^{-\frac{r}{2a^2}}\, dP_A(a)$$

Basicamente la directa viene de esta formula de probabilidad total y de $\dfrac{d^k e^{-\frac{r}{2a^2}}}{dr^k} = (-1)^k \dfrac{e^{-\frac{r}{2a^2}}}{2^k a^{2k}}$ (se usa para poder intercambiar derivada e integral el teorema de convergencia dominada sec. 1.3.1).

De nuevo, la recíproca es más díficil a probar y es detallada en (Widder, 1946, § 12) por ejemplo. Basicamente, se muestra que si d_X es completamente monotona, se puede escribir $d_X(n) = \displaystyle\int_{\mathbb{R}_+} v^n d\mu(v) = \int_{\mathbb{R}_+} e^{-nu} d\mu\left(e^{-u}\right)$, momento de orden n con respecto a una medida de probabilidad μ (Hausdorff, 1921a, 1921b), visto como transformada de Laplace-Stieltjes en $t = n$. Se define la función $f(r) = \displaystyle\int_{\mathbb{R}_+} e^{-ru} d\mu\left(e^{-u}\right)$ que coincide con d_X sobre $\mathbb{N}$. Se proba finalmente que $d_X(r) - f(r)$ se extiende analiticamente en el semi plano complejo $\Re e\,\{s\} \geq 0$ y de la analiticidad que d_X y f coinciden entonces en todo este semi-plano (Carlson, 1921; Carrier et al., 2005). □

Se notará que la consistencia se manifesta también a través de la generadora de la mezcla gaussiana:

---

[90] Por definición, la transformada de Laplace-Stieltjes de una medida μ definida sobre $\mathbb{R}_+$ es una función de un número complejo $\mathcal{LS}[\mu](s) = \displaystyle\int_{\mathbb{R}_+} e^{-st}\, d\mu(t)$. Se proba que un real x_0, llamado abscisa de convergencia tal que la convergencia es uniforma en el semi-plano $\Re e\,\{s\} > x_0$, y a continuación tal que $\mathcal{LS}[\mu]$ es analítica en este semi-plano. Si $\mu = \displaystyle\sum_i \alpha_i \delta_{t_i}$ es una medida discreta se obtiene $\mathcal{LS}[\mu](s) = \displaystyle\sum_i \alpha_i e^{-st_i}$ conocido como serie de Dirichlet, y si μ admite una densidad g, $\mathcal{LS}[\mu](s) \equiv \mathcal{L}[g](s) = \displaystyle\int_{\mathbb{R}_+} e^{-st}\, g(t)\, dt$ es dicha transformada de Laplace ordinaria de g. Para $s = \imath\omega$ con $\omega \in \mathbb{R}$, las transformaciones de Laplace-Stieltjes y de Laplace coinciden con las transformaciones $\mathcal{FS}$ de Fourier-Stieltjes y $\mathcal{F}$ de Fourier ordinaria (ver también sección **??** para esta transformación). Al imagen de la transformada de Fourier inversa que vimos brevemente sección 1.8.3, existen fórmulas de inversión de las transformadas de Laplace-Stieltjes y de Laplace ordinaria, denotadas $\mathcal{LS}^{-1}$ y $\mathcal{L}^{-1}$ respectivamente. Se referirá por ejemplo a (Widder, 1946) para tener más detalles.

Lema 1-57. *Sea $\varphi_X \in \mathfrak{EP}$ (o $d_X \in \mathfrak{CM}$). Entonces, para cualquier dimension d y $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$ d-dimensional, en la escritura $X \overset{\mathrm{d}}{=} A\,\Sigma^{\frac{1}{2}} G + m$ con $A > 0$ independiente de $G \sim \mathcal{N}(0, I)$ la ley de A no depende de la dimension d.*

Demostración. La prueba es inmediata del hecho que, para cualquier matriz $M \in \mathcal{M}_{d,n}(\mathbb{R})$ de rango lleno con $d \leq n$, $X \overset{\mathrm{d}}{=} A\Sigma^{\frac{1}{2}} G$ con $G \sim \mathcal{N}(0, I)$ n-dimensional, tenemos $X' = MX \overset{\mathrm{d}}{=} AM\Sigma^{\frac{1}{2}} G \overset{\mathrm{d}}{=} A\,(M\Sigma M^t)^{\frac{1}{2}}\, G'$ con $G' \sim \mathcal{N}(0, I)$ d-dimensional (ver teorema 1-42). Una mezcla de gaussiana queda mezcla por proyección, con la misma generadora A, y al revés puede ser vista como proyección de una mezcla de dimensión más grande (no va a cambiar la generadora) [91]. □

Si en la escritura como mezcla de escala de base uniforme se escribe sencillamente la ley del rayo R, se puede también caracterizar la generadora en el caso de mezcla de base gaussiana. Viene de la escritura como mezcla, donde se reconoce une transformada de Laplace-Stieltjes (ver nota de pie 90).

Lema 1-58. *Sea $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$ mezcla de escala gaussiana, con A la variable aleatoria generadora. φ_X admite una continuación analítica $s \mapsto \varphi_X(s)$ en el semi-plano complejo $\Re e\,\{s\} > 0$. Notando $\mathcal{LS}$ la transformada de Laplace-Stieltjes y $\mathcal{LS}^{-1}$ la transformada inversa (ver nota de pie 90), la función de repartición de A es dada por*

$$F_A(a) = \mathcal{LS}^{-1}[\varphi_X]\left(\left(0\,;\,\frac{a^2}{2}\right)\right)$$

Si la medida P_A admite una densidad p_A, se obtiene inmediatamente

$$p_A(a) = a\,\mathcal{L}^{-1}[\varphi_X]\left(\frac{a^2}{2}\right)$$

Similarmente, si X admite una generadora de densidad d_X, d_X se extiende también analíticamente en el semi-plano complejo $\Re e\,\{s\} > 0$ y se obtiene

$$\int_0^a u^{-d} dP_A(u) = (2\pi)^{\frac{d}{2}}\,\mathcal{LS}^{-1}[d_X]\left(\left(\frac{1}{2a^2}\,;\,+\infty\right)\right) \qquad y \qquad f_A(a) = (2\pi)^{\frac{d}{2}}\,a^{d-3}\,\mathcal{L}^{-1}[d_X]\left(\frac{1}{2a^2}\right)$$

respectivamente.

Demostración. Para U uniforme sobre la esfera, del desarollo en serie de la función de Bessel (Gradshteyn & Ryzhik, 2015, Ec. 8.402), es claro que $\varphi_U(w) = 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right) w^{-\frac{d-2}{4}} J_{\frac{d}{2}-1}(\sqrt{w})$ admite una continuación analítica en el semi-plano complejo $\Re e\,\{s\} > 0$. De la mezcla uniforme, también $\varphi_X(w) = \int_{\mathbb{R}_+} \varphi_X(rw)\, dP_R(r)$ admite una continuación analítica en el semi-plano complejo $\Re e\,\{s\} > 0$. Luego, de la escritura de mezcla gaussiana tenemos

$$\varphi_X(w) = \int_{\mathbb{R}_+} e^{-\frac{wa^2}{2}}\, dP_A(a) = \int_{\mathbb{R}_+} e^{-wt}\, dP_A(\sqrt{2t})$$

[91] Cuidense de que no anda si salimos a partir de una mezcla de base uniforme: como lo vamos a ver, las proyecciones no son uniforme más si $d < n$, lo que se intuite por razones de dominio imagen (bolas por proyección de la esfera).

por cambio de variables. Se reconoce la transformada de Laplace-Stieltjes $\mathcal{L}$ de la medida μ (definida sobre $\mathbb{R}_+$) dada por $\mu\big((0\,;\,t)\big) = P_A\left((0\,;\,\sqrt{2t})\right) = F_A(\sqrt{2t})$. La función de repartición en $\sqrt{2t}$ es entonces dada por la transformada inversa de Laplace-Stieljes de $s \mapsto \varphi_X(s)$, lo que cierra la prueba para F_A. Ahora, para escribir p_A se puede diferenciar F_A obtenido, o simplemente escribir $dP_A(a) = p_A(a)\,da$, *i. e.,* $dP_A(\sqrt{2t}) = \frac{p_A(\sqrt{2t})}{\sqrt{2t}}\,dt$ y reconocer en $\varphi(s)$ la transformada ordinaria de Laplace de $t \mapsto \frac{p_A(\sqrt{2t})}{\sqrt{2t}}$.

La prueba que d_X se extiende analiticamente en el semi-plano $\Re e\,\{s\} > 0$ es dada en (Widder, 1946, Cap. IV, Teo. 3a) [92]. Basicamente, siendo d_X C^∞, se escribe su desarollo de Taylor en torno a un $r_0 > 0$ y se proba que sobre $r \in (0\,;\,r_0)$, esta serie de terminos todos positivos converge uniformamente. Por consecuencia, se tiene la convergencia uniforme también para $r = s$ en la bola compleja $|s - r_0| < r_0$. Se cierra la prueba de la validez para cualquier $r_0 > 0$. A continuación, la forma de P_A y p_A a partir de d_X viene de los mismos pasos, saliendo de

$$d_X(r) = \int_{\mathbb{R}_+} (2\pi)^{-\frac{d}{2}}\, a^{-d} e^{-\frac{r}{2}}\, dP_A(a) = (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}_+} (2t)^{\frac{d}{2}}\, e^{-wt}\, dP_A\left((2t)^{-\frac{1}{2}}\right)$$

y introduciendo la medida μ_A definida por $\mu_A(B) = \displaystyle\int_B (2t)^{\frac{d}{2}}\, dP_A\left((2t)^{-\frac{1}{2}}\right)$ por un lado, y escribiendo $dP_A(a) = p_A(a)da$ por el otro lado. □

Cuando se puede extender φ_X o d_X al eje complejo imaginario puro, $w = \imath\,\omega$ o $r = \imath\,\omega$ se reemplazan las transformadas de Laplace-Stieltjes y Laplace ordinarias (y sus inversas) por las de Fourier-Stieltjes $\mathcal{FS}$ y Fourier ordinaria $\mathcal{F}$ (y sus inversas) [93] (Zolotarev, 1957; Poularikas, 1999; Poularikis, 2010; Widder, 1946; Paris & Kaminski, 2001).

Varios vectores aleatorios que hemos visto caen en la familia elíptica. No admiten todas una densidad con respecto a la medida de Lebesgue, como lo hemos visto con la ley uniforme sobre la esfera

---

[92] Se proba de hecho que una función absolutamente monotona sobre $\mathbb{R}_-$, como lo es $r \mapsto d_X(-r)$, se extiende analiticamente en el semi-plano $\Re e\,\{s\} < 0$.

[93] Se puede también pensar a la transformación de Mellin-Stieltjes y la de Mellin ordinario dada respectivamente por $\mathcal{MS}[\mu](s) = \displaystyle\int_{\mathbb{R}_+} t^s\, d\mu(t)$ y $\mathcal{M}[g](s) = \displaystyle\int_{\mathbb{R}_+} t^{s-1}\, g(t)\, dt$, definidas sobre una franja del plano complejo del tipo $\Re e\,\{s\} \in (x_m\,;\,x_M)$ (abscisas de convergencia), y usar sus propiedades remarcables (Zolotarev, 1957; Poularikas, 1999; Poularikis, 2010; Widder, 1946; Paris & Kaminski, 2001). En el caso con densidad por ejemplo, se reconoce en $g_X(r) = r^{d-1} d_X\left(r^2\right)$ (ley del rayo, bajo un factor) lo que se llama convolución afín, entre $g_G(r)$ (ley del rayo de la gaussiana) y la medida de A, $g_X(r) = \displaystyle\int_{\mathbb{R}_+} \frac{1}{a} g_G\left(\frac{r}{a}\right)\, p_A(a)\, da$. Al imagen de las propiedades de la transformada de Laplace, la transformación de Mellin de una convolución afín es el producto de las transformaciones de Mellin, y por un lado $\mathcal{M}[t^\alpha g(t)](s) = \mathcal{M}[g](s+\alpha)$ y por el otro lado $\mathcal{M}[g(r^\alpha)](s) = |\alpha|^{-1}\mathcal{M}[g](s/\alpha)$. Así, se obtiene sencillamente, de $\mathcal{M}[d_G](s) = 2^s\Gamma(s)$ (Poularikis, 2010, Cap. 12, Tabla 12.1) o (Poularikas, 1999, Cap. 18, Tabla 18.1) $\mathcal{M}[p_A](s) = \frac{\pi^{\frac{d}{2}}\mathcal{M}[d_X]\left(\frac{s+d-1}{2}\right)}{2^{\frac{s-1}{2}}\Gamma\left(\frac{s+d-1}{2}\right)}$. Similarmente, se muestra en el caso con densidad que $\mathcal{M}[p_A](s) = \frac{\mathcal{M}[\varphi_X]\left(\frac{1-s}{2}\right)}{2^{\frac{1-s}{2}}\Gamma\left(\frac{1-s}{2}\right)}$, y obviamente $\mathcal{M}[\varphi_X] = \frac{\pi^{\frac{d}{2}} 4^s \Gamma(s)}{\Gamma\left(\frac{d}{2}-s\right)}\mathcal{M}[d_X]\left(\frac{d}{2}-s\right)$ (lo que se obtiene tambien a partir de la relación via transformada de Hankel entre esta generadoras). Se puede referirse también a (Zozor, 2012, § 3.2.1)).

(admite una con respecto a la medida de Haar), tampoco no son todas consistentes (y entonces no se escriben todas como mezcla de escala de base gaussiana).

Ejemplo 1-25 (Distribución gaussiana)**.** *Sea* $X \sim \mathcal{N}(m, \Sigma)$*. Entonces,* $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$ *o equivalentemente* $X \sim \mathcal{ED}(m, \Sigma, d_X)$ *con*

$$\varphi_X(u) = e^{-\frac{u}{2}} \qquad \textit{y} \qquad , \qquad d_X(r) = (2\pi)^{-\frac{d}{2}} e^{-\frac{r}{2}}$$

Eso da la ley del rayo

$$p_R(r) = \frac{1}{2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)} r^{d-1} e^{-\frac{r^2}{2}}$$

Aparece que $R^2 \sim \mathcal{G}\left(\frac{d}{2}, \frac{1}{2}\right)$ *distribución gamma.*

Además, la generadora característica siendo independiente de la dimensión d*,* $\Phi_X \in \mathfrak{P}$*, o equivalentemente la generadora de densidad* d_X *teniendo la forma que damos para cualquier dimensión (cambiar de dimensión es equivalente a cambiar* d*), la gaussiana es consistente; Las marginales de una gaussiana son gaussianas, y cualquier gaussiana puede ser vista como marginal de gaussiana de cualquier dimensión más grande.*

Obviamente, se escribe como mezcla de escala de gaussiana con $A = 1$ *variable cierta.*

Ejemplo 1-26 (Distribución Student-t)**.** *Sea* $X \sim \mathcal{T}_\nu(m, \Sigma)$*. Entonces,* $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$ *o equivalentemente* $X \sim \mathcal{ED}(m, \Sigma, d_X)$ *con*

$$\varphi_X(u) = \frac{\nu^{\frac{\nu}{4}}}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{\nu}{2}\right)} u^{\frac{\nu}{4}} K_{\frac{\nu}{2}}\left(\sqrt{\nu\, u}\right) \qquad \textit{y} \qquad d_X(r) = \frac{\Gamma\left(\frac{d+\nu}{2}\right)}{(\pi\nu)^{\frac{d}{2}}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{r}{\nu}\right)^{-\frac{d+\nu}{2}}$$

Se obtiene de d_X *la ley de rayo*

$$p_R(r) = \frac{2}{\nu^{\frac{d}{2}} B\left(\frac{\nu}{2}, \frac{d}{2}\right)} \frac{r^{d-1}}{\left(1 + \frac{r^2}{\nu}\right)^{\frac{d+\nu}{2}}}$$

Con $F = \frac{R^2}{d}$ *se obtiene por transformación*

$$p_F(f) \propto \frac{f^{\frac{d}{2}-1}}{\left(1 + \frac{d}{\nu} f\right)^{\frac{d+\nu}{2}}}$$

conocido como ley de Fisher-Snedecor (o F*), de grados de libertad* (d, ν)*, o tipo Pearson VI, o beta de secunda especie (Johnson et al., 1995b; Mukhopadhyay, 2000; Brémaud, 1988; Ibarrola & Pérez, 2012) (ratio de gamma independientes).*

La Student-t *es también consistence de la independencia de* φ_X *con la dimensión* d*,* $\Phi_X \in \mathfrak{P}$*, o equivalentemente del hecho de que la generadora de densidad* d_X *teniendo la forma que damos para cualquier dimensión (cambiar de dimensión es equivalente a cambiar* d*); Las marginales de una Student-*t *con* ν *grado de libertad queda Student-*t *con* ν *grado de libertad y cualquier Student-*t *con* ν *grado de libertad pued ser vista como marginal de un Student-*t *con* ν *grado de libertad de cualquier dimensión más grande.*

Vímos en la sección 1.10.2.9, lema 1-49, que se escribe una Student-t como mezcla de escala gaussiana donde $A \overset{\mathrm{d}}{=} \frac{1}{\sqrt{V}}, V \sim \mathcal{G}\left(\frac{\nu}{2}\,,\,\frac{\nu}{2}\right)$ ley gamma, i. e., A es la raiz cuadrada de una gamma inversa (ver también (Fang et al., 1990; Kotz & Nadarajan, 2004)). Se puede re-obtener este resultado buscando directamente p_A por transformación de Laplace inversa de d_X o de φ_X, lema 1-58 y (Poularikis, 2010, Cap. 5, Tab. A.5.1, Ec. 14) o (Poularikas, 1999, Cap. 2, Tab. 2.3, Ec. 14).

Ejemplo 1-27 (Distribución student-r)**.** *Sea $X \sim \mathcal{R}_\nu(m,\Sigma)$. Entonces, $X \sim \mathcal{ED}(m,\Sigma,\varphi_X)$ o equivalentemente $X \sim \mathcal{ED}(m,\Sigma,d_X)$ con*

$$\varphi_X(u) = \frac{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}+1\right)}{(\nu+2)^{\frac{\nu}{4}}}\, u^{-\frac{\nu}{4}} J_{\frac{\nu}{2}}\left(\sqrt{(\nu+2)\,u}\right) \qquad \text{y} \qquad d_X(r) = \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\pi^{\frac{d}{2}}(\nu+2)^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}\right)}\left(1-\frac{r}{\nu+2}\right)_+^{\frac{\nu-d}{2}}$$

Se obtiene así la ley del rayo bajo la forma

$$p_R(r) = \frac{2}{(\nu+2)^{\frac{d}{2}}}\, r^{d-1}\left(1-\frac{r^2}{\nu+2}\right)^{\frac{\nu-d}{2}} \mathbb{1}_{(0\,;\,1)}\left(\frac{r^2}{\nu+2}\right)$$

Se concluye que $\frac{R^2}{\nu+2} \sim \beta\left(\frac{d}{2}, \frac{\nu-d}{2}+1\right)$ distribución beta.

Pero $X \sim \mathcal{R}_\nu(m,\Sigma)$ no es consistente. Puede parecer contradictorio porque φ_X parece no depender de la dimensión y d_X tiene una forma tal que cambiar de dimensión parece equivalente a cambiar d. Sin embargo, la dependencia de φ_X en d es escondida en el vínculo sobre el grado de libertad $\nu > d-2$. Dicho de otra manera, a ν dado, si aumentamos la dimensión, se lo puede hacer solamente a condición que $d < \nu+2$, i. e., $\Phi_X \notin \mathfrak{P}$ ($\Phi_X \in \mathfrak{P}_n,\ n < \nu+2$). Tratando de d_X, las marginales de X son Student-r con grado de libertad ν, pero una Student-r no puede ser vista como marginales de cualquier dimensión más grande; es posible hasta la dimensión máxima $\lceil \nu+2 \rceil - 1$.

Obviamente, sin calculos, se ve del soporte acotado de la densidad que X no puede ser mezcla de escala de gaussiana. Pero recordar de la sección 1.10.2.10, lema 1-54, que $X \overset{\mathrm{d}}{=} \frac{\sqrt{\nu+2}\,\Sigma^{\frac{1}{2}}G}{\sqrt{V+\|G\|^2}} + m \sim \mathcal{R}_\nu(m,\Sigma)$ con $V \sim \mathcal{G}\left(\frac{\nu-d}{2}+1\,,\,\frac{1}{2}\right)$ y $G \sim \mathcal{N}(0,I)$ d-dimensional e independientes de V. Parece una mezcla de gaussiana, pero con la generadora que sería dependiente de G en este caso.

Ejemplo 1-28 (Distribución uniforme sobre la esfera - continuación)**.** *Claramente, $U \sim \mathcal{U}(\mathbb{S}_d)$ tampoco es consistente: las marginales de U no pueden ser uniformes sobre $\mathbb{S}_{d'}$; sin calculo, se queda claro que para $d' = d-1$, el vector es definido en la bola $\mathbb{B}_{d-1}(\mathbb{R})$, y, por inducción es lo mismo para cualquier dimensión más baja.*

En calculos, recuendense del ejemplo 1-24 que la generadora característica toma la forma

$$\varphi_X\left(w^2\right) = 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right) w^{1-\frac{d}{2}}\, J_{\frac{d}{2}-1}(w)$$

depende explicitamente de d. Más allá, para $d' < d$ y $X' = \begin{bmatrix} X_1 & \cdots & X_{d'} \end{bmatrix}^t$, por transformada inversa de Hankel y (Gradshteyn & Ryzhik, 2015, Ec. 6.575-1), se obtiene

$$d_{X'}\left(r^2\right) = \frac{\Gamma\left(\frac{d}{2}\right)}{\pi^{\frac{d'}{2}}\Gamma\left(\frac{d-d'}{2}\right)}\left(1-r^2\right)^{\frac{d-2-d'}{2}} \mathbb{1}_{(0\,;\,1)}(r)$$

y no existe para $d' > d$. Claramente lar marginales depende de ambas d y d'; no son uniforma, como lo vimos sin calculos y no se puede ver la uniforma sobre la esfera como marginal de un vector aleatorio de dimensión más alta. De hecho, se puede notar que las marginales de U son Student-r con $\nu = d-2$ el grado de libertad, a condición que $\nu > d' - 2$, es decir solamente cuando $d' < d$ (ver (?, ?, ?, ?)). Cuando $d' = d - 1$, la distribución diverge en los bordes del dominio de definición. Notar que si no cambia el grado de libertad para cualquier marginales, depende de la dimensión d: la generadora de densidad no depende solamente de d'. Obviamente, del soporte acotado de la densidad, se vio que X no podía ser mezcla de escala de gaussiana.

Si un vector a simetría elíptica no se escribe siempre como mezcla de escala de base gaussiana, existe una situación que podemos ver como "intermediaria": a veces se escribe como mezcla de escala de base Student-r. Este resultado fue probado por ejemplo en el caso escalar $d = 1$ en (Williamson, 1956; Kemperman, 1971; Keilson & Steutel, 1974) (ver unos elementos en (van Dantzig, 1951)), pero se extiende sencillamente al caso multivariado:

Teorema 1-67 (Mezcla de escala de base Student-r – a partir de la generadora característica)**.** *Sea $X \sim \mathcal{ED}(m, \Sigma, \varphi_X)$ d-dimensional. Si $\varphi_X \in \mathfrak{EP}_n$, $n > d$, entonces se puede escribir X como mezcla de escala de Student-r con $\nu = n - 2$ el grado de libertad, y reciprocamente*

$$\varphi_X \in \mathfrak{EP}_n,\ n > d \quad \Leftrightarrow \quad X \overset{\mathrm{d}}{=} B\,\Sigma^{\frac{1}{2}}\,S_{n-2} + m$$

con $B > 0$ independiente de $S_{n-2} \sim \mathcal{R}_{n-2}(0, I)$.

Demostración. La prueba la más sencilla se apoya sobre el hecho que si $\varphi_X \in \mathfrak{EP}_n$, se puede considerar un vector n-dimensional $Y \sim \mathcal{ED}(0, I, \varphi_X)$ de generadora característica φ_X y escribir $Y \overset{\mathrm{d}}{=} RU$ con $U \sim \mathcal{U}(\mathbb{S}_n)$, y reciprocamente. Ahora, del ejemplo 1-28, la proyección d-dimensional de U es Student-r con $n-2$ el grado de libertad, y reciprocamente se puede ver cualquier Student-r con $n-2$ el grado de libertad como proyección de un uniforme más grande. Además, U siendo independiente de R, la proyección es también independiente de R. Eso cierra la prueba.

Alternativamente, escribiendo la generadora característica a partir de la mezcla uniforme n-dimensional como en el teorema 1-65, en

$$\varphi_X\left(w^2\right) = \int_{\mathbb{R}_+} 2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right)\left(wr\sqrt{n}\right)^{1-\frac{n}{2}} J_{\frac{n}{2}-1}\left(wr\sqrt{n}\right)\, dP_{R_d}\left(r\sqrt{n}\right)$$

se reconoce en $w \mapsto 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)\left(\sqrt{dw}\right)^{1-\frac{d}{2}} J_{\frac{d}{2}-1}\left(\sqrt{dw}\right)$ la generadora característica de la Student-r con $\nu = n-2$ el grado de libertad. □

Pasando, se notará que para buscar la ley de B, sufice ver la Student-r como proyección de una uniforme de dimensión $d+2$, *i. e.,* $X \sim \mathcal{ED}(m, \Sigma, \varphi_X) \overset{\mathrm{d}}{=} B\Sigma^{\frac{1}{2}}S_{n-2} + m$, d-dimensional:

$$\text{Construimos } Y \sim \mathcal{ED}(0, I, \varphi_X) \quad (d+2)\text{-dimensional} \qquad \text{y} \qquad Y \overset{\mathrm{d}}{=} RU \ \Leftrightarrow\ B \overset{\mathrm{d}}{=} R$$

De nuevo, existe un teorema equivalente tratando de la generadora de densidad:

Teorema 1-68 (Mezcla de escala de base Student-r – a partir de la generadora de densidad)**.** *Sea* $X \sim \mathcal{ED}(m, \Sigma, d_X)$ d*-dimensional. Si* $d_X \in \mathfrak{CM}_n$, $n \in \mathbb{N}^*$*, entonces se puede escribir* X *como mezcla de escala de Student-*r *con* $\nu = 2n + d - 2$ *el grado de libertad, y reciprocamente,*

$$d_X \in \mathfrak{CM}_n,\ n \in \mathbb{N}^* \quad \Leftrightarrow \quad X \overset{\mathrm{d}}{=} B\,\Sigma^{\frac{1}{2}}\,S_{2n+d-2} + m$$

con $B > 0$ *independiente de* $S_{2n+d-2} \sim \mathcal{R}_{2n+d-2}(0, I)$.

Demostración. La recíproca es inmediata de

$$d_X(r) = \frac{\Gamma\left(\frac{d}{2}+n\right)}{\pi^{\frac{d}{2}}(2n+d)^{\frac{d}{2}}\Gamma(n)} \int_{\mathbb{R}_+} b^{-d}\left(1 - \frac{r}{(2n+d)b^2}\right)_+^{n-1} dP_B(b)$$

Si $n = 1$, claramente $d_X(r) = \dfrac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{\frac{d}{2}}(d+2)^{\frac{d}{2}}} \displaystyle\int_{\left(\sqrt{\frac{r}{2n+d}}\,;\,+\infty\right)} b^{-d} dP_B(b)$ representa una medida de $\left(\sqrt{\frac{r}{2n+d}}\,;\,+\infty\right)$: es decreciente con r. Si $n \geq 2$, se muestra que se puede usar el teorema de convergencia dominada para intercambiar derivada e integral, y eso hasta el orden $n-1$, dando, para $0 \leq k \leq n-1$

$$(-1)^k d_X^{(k)}(r) = \frac{\Gamma\left(\frac{d}{2}+n\right)}{\pi^{\frac{d}{2}}(2n+d)^{\frac{d}{2}+k}\Gamma(n-k)} \int_{\mathbb{R}_+} b^{-d-2k}\left(1 - \frac{r}{(2n+d)b^2}\right)_+^{n-k-1} dP_B(b) \ \geq\ 0$$

y de nuevo $(-1)^{n-1} d_X^{(n-1)}(r) = \dfrac{\Gamma\left(\frac{d}{2}+n\right)}{\pi^{\frac{d}{2}}(2n+d)^{\frac{d}{2}+n-1}} \displaystyle\int_{\left(\sqrt{\frac{r}{2n+d}}\,;\,+\infty\right)} b^{-d-2n+2} dP_B(b)$ es decreciente.

Una prueba detallada de la propiedad directa se encuentra en (Williamson, 1956) en el contexto escalar, pero se extiende al caso multivariado sin costo adicional. Basicamente, se proba que para $f \in \mathfrak{CM}_n$, $\forall k = 1, \ldots, n-1$ y $\alpha > 0$, (i) $t^{k-1} f^{(k)}(t)$ es integrable sobre $(\alpha\,;\,+\infty)$ y (ii) $\lim\limits_{t \to +\infty} t^k f^{(k)}(t) = 0$. **Luego, sea** P_n **tal que**

$$P_n\big((a\,;\,b)\big) = (-1)^n \Big(f^{(n-1)}(b) - f^{(n-1)}(a)\Big)$$

Es una medida por la decrecencia de $(-1)^{n-1} f^{(n-1)}$ **y finita del resultado (i)**. Entonces, se escribe el desarollo de Taylor en torno a α, hasta el orden $n-1$ con resto de la forma integral, es decir cuando $\alpha \geq t$

$$f(t) = \sum_{k=0}^{n-1} \frac{(t-\alpha)^k}{k!} f^{(k)}(\alpha) - (-1)^n \int_{(t\,;\,\alpha)} \frac{(t-u)^{n-1}}{(n-1)!}\, dP_n(u)$$

Escribiendo $(t-u)^{n-1} = (-1)^{n-1} u^{n-1} \left(1 - \frac{t}{u}\right)^{n-1}$ y aplicando el desarollo anterior a $d_X(r)$, con el cambio de variables $u = (2n+d)b^2$ se obtiene

$$d_X(r) = \sum_{k=0}^{n-1} \frac{(r-\alpha)^k}{k!}\, d_X^{(k)}(\alpha) + \frac{(2n+d)^{n-1}}{\Gamma(n)} \int_{\left(\sqrt{\frac{t}{2n+d}}\,;\,\sqrt{\frac{\alpha}{2n+d}}\right)} \left(1 - \frac{r}{(2n+d)b^2}\right)^{n-1} dP_n\big((2n+d)b^2\big)$$

Ahora, dejando α tender al infinito, se nota que: $d_X(\alpha) \to 0$ (generadora de densidad d-dimensional, $\alpha^{d-1} d_X(\alpha^2)$ debe tender a 0); del punto (ii), cada termino de la suma tiende a 0. Queda solo el termino integral, que se escribe también

$$d_X(r) = \frac{(2n+d)^{n-1}}{\Gamma(n)} \int_{\mathbb{R}_+} b^{-d}\left(1 - \frac{r}{(2n+d)b^2}\right)_+^{n-1} b^{2n+d-2} dP_n\big((2n+d)b^2\big)$$

Tiene precisamente la forma de mezcla de Student-r d-dimensional con grado de libertad $2n+d-2$. **Bien fijar que todo listo.** □

Pasando, se notará que la ley de B es también dada por

$$P_B(A) = \frac{\pi^{\frac{d}{2}}(2n+d)^{\frac{2n+d-2}{2}}}{\Gamma\left(\frac{d}{2}+n\right)} \int_A b^{2n+d-2}\, dP_n\big((2n+d)b^2\big) \qquad \text{con} \qquad \mathbf{P_n\big((a\,;\,b)\big) = (-1)^n \left(f^{(n-1)}(b) - f^{(n-1)}(a)\right)}$$

Nota: obviamente, si denotamos $\mathfrak{S}_{n,d}$ el conjunto de vectores aleatorios d-dimensjonal, mezcla de Student-r con $n-2$ el grado de libertad, $\mathfrak{S}_{d+1,d} \supset \mathfrak{S}_{d+2,d} \supset \cdots$ y $\mathfrak{S}_{+\infty,d} = \bigcap_{n=d+1}^{+\infty} \mathfrak{S}_{n,d}$ es el conjunto de mezcla de gaussiana d-dimensional. Así, se puede imaginar también buscar la ley de una mezcla de gaussiana pasando por la de mezcla de Student-r, tomando el límite cuando n tiende al infinito de la ley de B.

1.10.4.2. Caso complejo en unas palabras

Vímos en la sección 1.9 el caso de vectores aleatorio reales, y la noción de circularidad. No es equivalente a la de esféricidad, aún que hay vínculos, como lo vamos a ver. De hecho, la necesidad de trabajar con vector a simetría elíptica se justifica con los mismos argumentos que llevaron al estudio de vectores aleatorios complejos. Se encontrará estudios de esta familia en referencias tales que (Krishnaiah & a & J. Lin, 1986; Micheas, Dey & Mardia, 2006; Ollila, Eriksson & Koivunen, 2011; Ollila, Tyler, Koivunen & Poor, 2012; Fang et al., 1990; Besson & Abramovich, 2013; Bausson et al., 2007; Chitour & Pascal, 2008) por ejemplo. Damos en esta sección lo esencial.

Antes de ir más allá, empezamos con la definición de tales vectores a simetría elíptica.

Definición 1-53 (Vector complejo esféricamente invariante)**.** *Sea Z vector aleatorio d-dimensional complejo. X es dicho esfericamente invariante, o rotacionalmente invariante, o a simetría esférica, o de distribución esférica si para cualquier matriz unitaria*[94] *V,*

$$VZ \overset{\mathrm{d}}{=} Z$$

A continuación, como en el caso real, se extiende haciendo estiramientos y transformación unitaria (equivalente de una rotación) de manera siguiente:

Definición 1-54 (Vector complejo a simetría elíptica)**.** *Sea Z vector aleatorio d-dimensional complejo. Z es dicho a simetría elíptica, o elipticalmente invariante, o de distribución elíptica, en torno a $m \in \mathbb{C}^d$, si existe una matriz $\Sigma \in P_d^+(\mathbb{C})$ tal que para cualquier matriz unitaria V,*

$$V\,\Delta^{-\frac{1}{2}}\,Q^\dagger\,(X-m) \overset{\mathrm{d}}{=} \Delta^{-\frac{1}{2}}\,Q^\dagger\,(X-m)$$

[94] Recordarse que V es unitaria si $VV^\dagger = V^\dagger V = I$.

donde la matriz real diagonal $\Delta > 0$ *es la matriz de autovalores de* Σ *y* Q *la matriz de los autovectores correspondientes (matriz unitaria (Bhatia, 1997, 2007; Horn & Johnson, 2013)),* $\Sigma = Q\Delta Q^\dagger$*. Es decir,* $\Delta^{-\frac{1}{2}} Q^\dagger (X - m)$ *es a simetría esférica.*

De nuevo, m *es un parámetro de posición y* Σ *matriz característica.*

Obviamente, se queda en el caso complejo la indeterminación mediante un factor escalar.

El vínculo entre la simetría elíptica y la circularidad es también obvia:

Lema 1-59. *Sea* Z *vector complejo* d*-dimensional a simetría elíptica en torno a* $m \in \mathbb{C}^d$*. Entonces* $Z - m$ *es circular.*

Demostración. La prueba es inmediata de $Z - m \overset{\mathrm{d}}{=} Q\,\Delta^{\frac{1}{2}}\,Y$ con Y a simetría esférica. Se cierra la prueba de, $e^{\imath\theta}(Z - m) \overset{\mathrm{d}}{=} Q\,\Delta^{\frac{1}{2}}\,\left(e^{\imath\theta} I\right) Y$ conjuntamente a $e^{\imath\theta} I$ unitaria. □

Es importante notar las inmersiones biyectivas de $\mathbb{C}^d$ en $\mathbb{R}^{2d}$ y de $\mathcal{M}_{d',d}(\mathbb{C})$ en un subconjunto de $\mathcal{M}_{2d',2d}(\mathbb{R})$ siguientes:

$$\forall\, z \in \mathbb{C}^d \qquad \text{en biyección con} \qquad \widetilde{z} = \begin{bmatrix} \Re e\,\{z\} \\ \Im m\,\{z\} \end{bmatrix} \in \mathbb{R}^{2d}$$

$$\forall\, M \in \mathcal{M}_{d',d}(\mathbb{C}) \qquad \text{en biyección con} \qquad \overline{M} = \begin{bmatrix} \Re e\,\{M\} & -\Im m\,\{M\} \\ \Im m\,\{M\} & \Re e\,\{M\} \end{bmatrix} \in \mathcal{M}_{2d',2d}(\mathbb{R})$$

Claramente para matrices M, N y un vector z,

$$Mz \quad \text{es en biyección con} \quad \overline{M}\,\widetilde{z}, \qquad MN \quad \text{es en biyección con} \quad \overline{M}\,\overline{N}$$

Llamaremos *forma real* ambas inmersiones y las escrituras precedientes. Además, se ve sencillamente que

- $\overline{M^\dagger} = \overline{M}^{\,t}$,
- $M \in H_d(\mathbb{C}) \;\Leftrightarrow\; \overline{M} \in S_{2d}(\mathbb{R}) \;\Leftrightarrow\; \Re e\,\{M\}^t = \Re e\,\{M\} \;\wedge\; \Im m\,\{M\}^t = -\Im m\,\{M\}$,
- $\forall\, M \in H_d(\mathbb{C}),\ z \in \mathbb{C}^d, \quad z^\dagger M z = \widetilde{z}^t\,\overline{M}\,\widetilde{z}$,
- M unitaria $\Leftrightarrow$ $\overline{M}$ ortogonal.

Ahora, se víncula naturalmente la noción de elipticidad del caso complejo al caso real:

Lema 1-60. *Sea* Z *vector complejo* d*-dimensional a simetría elíptica en torno a* m*. Entonces,* $\widetilde{Z}$ *vector real* $2d$*-dimensionales es a simetría elíptica en torno a* $\widetilde{m}$*.*

Demostración. Se obtiene la forma real de la descomposición $\Sigma = Q\Delta Q^\dagger$ bajo la forma $\overline{\Sigma} = \overline{Q}\,\overline{\Delta}\,\overline{Q}^t$ con $\overline{\Delta} = \begin{bmatrix} \Delta & 0 \\ 0 & \Delta \end{bmatrix} \in P_{2d}^+(\mathbb{R})$ diagonal, $\overline{Q}$ ortogonal. La prueba se cierra saliendo de la definición 1-54 escrita bajo su forma real, notando que $\overline{\Delta^{-\frac{1}{2}}} = \overline{\Delta}^{-\frac{1}{2}}$. □

De esta equivalencia, todos los resultados anteriores se extienden naturalmente al caso complejo. Para Z a simetrá elíptica en torno a m y de matriz característica Σ:

- De la sección 1.9.1 se obtiene $\Phi_Z(\omega) = \Phi_{\widetilde{Z}}(\widetilde{\omega}) = e^{\imath\widetilde{\omega}^t\widetilde{m}}\varphi_{\widetilde{Z}}\left(\widetilde{\omega}^t\,\overline{\Sigma}\,\widetilde{\omega}\right) = e^{\imath\Re e\left\{\omega^\dagger m\right\}}\varphi_{\widetilde{Z}}\left(\omega^\dagger\,\Sigma\omega\right)$;
  Denotaremos $\varphi_{\widetilde{Z}} \equiv \varphi_Z$ quien, con m y Σ caracteriza completamente Z;
  Escribiremos $Z \sim \mathcal{CED}\,(m, \Sigma, \varphi_Z)$.

- Similarmente, si Z admite una densidad, se obtiene la densidad bajo la forma [95] $p_Z(z) = |\Sigma|^{-1} d_Z\left((z-m)^\dagger\Sigma^{-1}(z-m)\right)$ con $d_Z \equiv d_{\widetilde{Z}} : \mathbb{R}_+ \mapsto \mathbb{R}_+$.

- De eso, se extiende naturalmente la mayoria de los teoremas:

  - **Momentos y cumulantes 1-56, y corrolario 1-15** (Krishnaiah, 1976)
  - Teorema 1-57 en: $Z \sim \mathcal{CED}(m, \Sigma, \varphi_Z)$ d-dimensional, $A \in \mathcal{M}_{d',d}(\mathbb{C})$ de rango lleno tal que $d' \leq d$ y $c \in \mathbb{C}^{d'} \quad \Rightarrow \quad AX + c \sim \mathcal{CED}(Am + c, A\Sigma A^\dagger, \varphi_Z)$;
  - Teorema 1-58 en: Sea Z, vector aleatorio complejo d-dimensonal de componentes Z_i. Entonces $Z \sim \mathcal{CED}(m, \Sigma, \varphi_Z) \quad \Leftrightarrow \quad \forall\, a \in \mathbb{C}^d, \quad a^\dagger(Z-m) \overset{\mathrm{d}}{=} \sqrt{\dfrac{a^\dagger\Sigma a}{\Sigma_{i,i}}}(Z_i - m_i)$;
  - Teorema 1-59 en: $Z \sim \mathcal{CED}(m, I, \varphi_Z)$ tiene sus componentes independientes si y solamente si $Z \sim \mathcal{CN}(m, \alpha I)$ con $\alpha > 0$;
  - Teorema 1-60 en: $Z \sim \mathcal{CED}(0, I, \varphi_Z) \;\Leftrightarrow\; Z \overset{\mathrm{d}}{=} RU$ con $U \sim \mathcal{U}\,(\mathbb{SC}_d)$ uniforme sobre la esfera d-dimensional compleja y más generalmente, para $Z \sim \mathcal{ED}(m, \Sigma, \varphi_Z), \quad Y \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}} RU + m$.
  - Corolario 1-16 en: $Z \sim \mathcal{CED}(0, I, \varphi_Z)$ da $\|Z\| \overset{\mathrm{d}}{=} R$ y $\frac{Z}{\|Z\|} \overset{\mathrm{d}}{=} U \sim \mathcal{U}(\mathbb{SC}_d)$ son independientes;
  - **Lema para alpha momentos 1-56**;
  - Teorema 1-62 en: $Z \sim \mathcal{CED}(0, I, d_Z)$ y $R \overset{\mathrm{d}}{=} \|Z\|$ admite una densidad que se escribe $p_R(r) = \dfrac{2\pi^d}{\Gamma(d)}\, r^{2d-1}\, d_X\left(r^2\right)$;
  - Teorema 1-63 en: para $Z \sim \mathcal{CED}(m, \Sigma, \varphi_Z) \equiv \mathcal{CED}(m, \Sigma, d_Z)$ las generadoras características y de densidad son relacionadas por $\varphi_Z\left(w^2\right) = (2\pi)^d\, w^{1-d} \displaystyle\int_{\mathbb{R}_+} r^d d_Z\left(r^2\right)\, J_{d-1}(rw)\, dr$ y $\quad d_Z\left(r^2\right) = (2\pi)^{-d}\, r^{1-d} \displaystyle\int_{\mathbb{R}_+} w^d \varphi_Z\left(w^2\right)\, J_{d-1}(rw)\, dw$;
  - Teorema 1-65 en: $Z \sim \mathcal{CED}\,(m, \Sigma, \varphi_Z) \overset{\mathrm{d}}{=} A\,\Sigma^{\frac{1}{2}}\,G + m \quad \Leftrightarrow \quad \varphi_Z \in \mathfrak{EP}$, con $A > 0$ independiente de $G \sim \mathcal{CN}(0, I)$;
  - Teorema 1-66 en: $Z \sim \mathcal{CED}\,(m, \Sigma, d_Z) \overset{\mathrm{d}}{=} A\,\Sigma^{\frac{1}{2}}\,G + m \quad \Leftrightarrow \quad d_Z \in \mathfrak{EM}$, con $A > 0$ independiente de $G \sim \mathcal{CN}(0, I)$;

---

[95] Se notará que, escribiendo las formas diagonales $\Sigma = Q\Delta Q^\dagger$ y $\overline{\Sigma} = \overline{Q}\,\overline{\Delta}\,\overline{Q}^t$ con Q y $\overline{Q}$ respectivamente unitaria y ortogonal, tenemos $|\overline{\Sigma}| = |\overline{\Delta}| = |\Delta|^2 = |\Sigma|^2$.

- Lema 1-57 en: $\varphi_Z \in \mathfrak{EP}$ (o $d_Z \in \mathfrak{CM}$) y para cualquier dimension d y $Z \sim \mathcal{CED}(m, \Sigma, \varphi_Z)$ d-dimensional, en la escritura $X \overset{\mathrm{d}}{=} A\,\Sigma^{\frac{1}{2}} G + m$ con $A > 0$ independiente de $G \sim \mathcal{N}(0, I)$ la ley de A no depende de la dimension d;
- Teorema 1-67 en: $Z \sim \mathcal{CED}(m, \Sigma, \varphi_Z)$ d-dimensional con $\varphi_z \in \mathfrak{EP}_n$, $n > 2d$, dando la escritura estocástica Z como mezcla de escala de Student-r compleja con $\nu = n - 2$ el grado de libertad, y reciprocamente: $\varphi_Z \in \mathfrak{EP}_n$, $n > 2d \quad \Leftrightarrow \quad X \overset{\mathrm{d}}{=} B\,\Sigma^{\frac{1}{2}}\,S_{n-2} + m$ con $B > 0$ independiente de $S_{n-2} \sim \mathcal{CR}_{n-2}(0, I)$ (ver sección 1.10.2.10);
- Teorema 1-68 en: $Z \sim \mathcal{CED}(m, \Sigma, d_Z)$ d-dimensional con $d_X \in \mathfrak{CM}_n$, $n \in \mathbb{N}^*$, dando la escritura estocástica Z como mezcla de escala de Student-r con $\nu = 2n + 2d - 2$ el grado de libertad, y reciprocamente: $d_Z \in \mathfrak{CM}_n$, $n \in \mathbb{N}^* \quad \Leftrightarrow \quad Z \overset{\mathrm{d}}{=} B\,\Sigma^{\frac{1}{2}}\,S_{2n+2d-2} + m$ con $B > 0$ independiente de $S_{2n+2d-2} \sim \mathcal{CR}_{2n+2d-2}(0, I)$.

1.10.4.3. Caso matriz variado en unas palabras

La extensión matriz variada de vector a simetría elíptica no es trivial. De hecho, hay pocas resultados en la literatura y aparece esencialmente en contexto de estimación de matriz de covarianza (Bilodeau & Brenner, 1999, §. 13.2) o (Tyler, 1982; Grübel & Rocke, 1990; ?, ?; Gupta & Varga, 1995; Fang & Li, 1999; Caro-Lopera, Farías & Balakrishnan, 2016). Tipicamente, si uno sale de $X_1, \ldots, X_n$ vectores aleatorios d-dimensionales independientes, de misma ley, y suponemos la media cero para simplificar el ejemplo, un estimador natural de la matriz de covarianza $\Sigma_W X$ es $\widehat{\Sigma}_X = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i X_i^t$, matriz de $S_d(\mathbb{R})$ (y aún de $P_d(\mathbb{R})$, y de $P_d^+(\mathbb{R})$ casi siempre si $n \leq d$). Ahora, si los X_i son a simétria eliptíca, W va a tener simetría por cambio de escala y rotaciones. Más precisamente, si se supone X_i esférica, para cualquier rotación O se tiene $OX_i \overset{\mathrm{d}}{=} X_i$, dando $O\widehat{\Sigma}_X O^t \overset{\mathrm{d}}{=} \widehat{\Sigma}_X$. Es precisamente lo que va a dar la noción de matriz a simetría esférica, y luego elíptica.

Definición 1-55 (Matriz esfericamente invariante)**.** *Sea X matriz aleatoria definida sobre $\mathcal{X} = S_d(\mathbb{R})$. X es dicha esfericamente invariante, o rotacionalmente invariante, o a simetría esférica, o de distribución esférica si para cualquier matriz ortogonal O se tiene*

$$OXO^t \overset{\mathrm{d}}{=} X$$

Naturlamente, se extiende como en el caso de vectores con estiramientos y una translación:

Definición 1-56 (Vector a simetría elíptica)**.** *Sea X matriz aleatoria d-dimensional definida sobre $\mathcal{X} = S_d(\mathbb{R})$. X es dicha a simetría elíptica, o elipticalmente invariante, o de distribución elíptica, en torno a $M \in \mathcal{M}_{d,d}(\mathbb{R})$, si existe una matriz $\Sigma \in P_d^+(\mathbb{R})$ tal que para cualquier matriz ortogonal O,*

$$O\,\Delta^{-\frac{1}{2}}\,Q^t\,(X - m)\,Q\,\Delta^{-\frac{1}{2}}\,O \overset{\mathrm{d}}{=} \Delta^{-\frac{1}{2}}\,Q^t\,(X - m)\,Q\,\Delta^{-\frac{1}{2}}$$

donde la matriz diagonal $\Delta > 0$ es la matriz de autovalores de Σ y Q la matriz de los autovectores correspondientes, $\Sigma = Q\Delta Q^t$. Dicho de otra manera, $\Delta^{-\frac{1}{2}} Q^t\,(X - m)\,Q\Delta^{-\frac{1}{2}}$ es a simetría esférica.

Llamaremos también m parámetro de posición *y la matriz* Σ matriz característica.

- **Volver a los cumulantes $2k$: se puede decir más?**
- **Citar mi HDR (Zozor, 2012, Sec. 3.2.1)**
- **Ver BilBre def 13 o (Krishnaiah, 1976): extension matricial**
- **Ver Shiryayev, Fang and Anderson, "Statistical inference in elliptically contoured and related distributions" a buscar**

---

hablar de simulación? Metoto inverso, mezcla (aparece en la concavidad de la entropia), rejeccion, a traves de la condicional para el caso vectorial?

CAPÍTULO 2
Nociones de teoría de la información

"Deberías llamarla 'entropía', por dos motivos.
En primer lugar su función de incerteza
ha sido usada en la mecánica estadística
bajo ese nombre, y por ello, ya tiene un nombre.
En segundo lugar, y lo que es más importante,
nadie sabe lo que es realmenta la entropía,
por ello, en un debate, siempre llevará la ventaja.
VON NEUMANN TO SHANNON (TRIBUS & MCIRVINE, 1971)

2.1 Introducción

La noción de información encuentra su origen con el desarrollo de la comunicación moderna, por ejemplo a través del telégrafo siguiendo la patente de Morse en 1840. La idea de asignar un código (punto o barra, más espacio entre letras y entre palabras) a las letras del alfabeto es la semilla de la codificación entrópica, la que se basa precisamente sobre la asignación de un código a símbolos de una fuente (codificación de fuente) según las frecuencias (o probabilidad de aparición) de cada símbolo en una cadena. De hecho, el principio de codificar un mensaje y mandar la versión codificada por un canal de transmisión es mucho más antiguo, a pesar de que no había ninguna formalización matemática ni siquiera explícitamente una noción de información. Entre otros, se puede mencionar el fotofone de A. G. Bell en 1880 (?, ?; Bruce, 1990) (sistema de comunicación con luz), el telégrafo óptico de Claude Chappe (1794), experimentos con luces por Guillaume Amontons (en los años 1690 en Paris), o aún más antiguamente la transmisión de mensaje con antorchas en la Grecia antigua, con humo por los indios o chiflando en la prehistoria (Montagné, 2008) o (Arndt, 2001, Cap. 3). Cada forma es una instancia práctica del esquema de comunicación de Shannon (Shannon, 1948; Shannon & Weaver, 1964), es decir la codificación de la información, potencialmente de la manera más económica

que se puede, su transmisión a un “receptor” (por un canal ruidoso) que la interpreta/lee/decodifica. Implícitamente, la noción de información es al menos tan antigua como la humanidad.

A pesar de que la idea de codificar y transmitir “información” sea tremendamente antigua, la formalización matemática de la noción de incerteza o falta de información, íntimamente vinculada a la noción de información, nació bajo el impulso de Claude Shannon y la publicación de su papel seminal, “A mathematical theory of communication” en 1948 (Shannon, 1948), o un año después en su libro re-titulado “The mathematical theory of communication” reemplazando el “A” (Una) por un “The” (La). Desde estos años, las herramientas de dicha teoría de la información dieron lugar a muchas aplicaciones especialmente en comunicación (Cover & Thomas, 2006; Verdu, 1998; Gallager, 2001, y ref.), pero también en otros campos muy diversos tal como la estimación o la discriminación (Cover & Thomas, 2006; Kay, 1993; van den Bos, 2007; Lehmann & Casella, 1998, y ref.), la inferencia estadística (Robert, 2007; Pardo, 2006), el procesamiento de señal o de datos (Phillips & Rousseau, 1992; Ebeling, Molgedey, Kurths & Schwarz, 2000; Basseville, 2013, y Ref.), en ciencias de la ingenería (Arndt, 2001; Kapur, 1989; Kapur & Kesavan, 1992; Phillips & Rousseau, 1992), física (Arndt, 2001; Ohya & Petz, 1993; Merhav, 2018, y Ref.) entre muchas otras (ver por ejemplo el esquema pagina 2 de (Cover & Thomas, 2006)).

La meta de este capítulo es describir las ideas y los pasos dando lugar a la definición de la entropía, como medida de incerteza o (falta de) información. En este capítulo, se empieza con la descripción intuitiva que subyace a la noción de información contenida en una cadena de símbolos, lo que condujo a la definición de la entropía. Esta definición puede ser deducida también de un conjunto de propiedades “razonables” que debería cumplir una medida de incerteza (enfoque axiomático). Se continuará con la descripción de tal noción de entropía, pasando del mundo discreto (símbolos, alfabeto) al mundo continuo, lo que no es trivial ni siquiera intuitivo. Se adelantará presentando el concepto de entropía condicional, lo que va a dar lugar a la noción de información compartida entre dos sistemas o variables aleatorias, concepto fundamental en el marco de la transmisión de información o de mensajes. A continuación, se presentará la noción de entropía relativa a una distribución de probabilidad de referencia, así que el concepto de distancia estadística o divergencia de una distribución con respecto a una referencia. En este capítulo veremos como estas medidas informacionales son entrelazadas a través varias identidades y desigualdades, así que varias relaciones con medidas del mundo de la estimación. Al final, se darán ejemplos y aplicaciones, así que varias generalizaciones de las medidades informacionales.

En todo este capítulo, hablaremos de “variable” aleatoria, que sea escalar o multivariata (vector, matriz).

2.2 Entropía como medida de incerteza

2.2.1 Entropía de Shannon, propiedades

Uno de los primeros trabajos tratando de formalizar la noción de información de una cadena de símbolos es debido a Ralph Hartley (Hartley, 1928). En su papel, Hartley definió la información de una secuencia como siendo proporcional a su longitud. Más precisamente, para símbolos de un alfabeto de cardinal α, existen α^n cadenas distintas de longitud n. Se definió la información de tales cadenas como siendo Kn (K dependiente de α). Para ser consistente, dos conjuntos del mismo tamaño $\alpha_1^{n_1} = \alpha_2^{n_2}$ deben llegar a la misma información, así que la información de Hartley es definida como $H = \log(\alpha^n)$ donde la base del logaritmo es arbitraria. Dicho de otra manera, tomando un logaritmo de base 2, esta información es nada más que los números de bits (0-1) necesarios para codificar todas las cadenas de longitud n de símbolos de un alfabeto de cardinal α. La información de Hartley es el equivalente de la entropía de Boltzmann de la mecánica estadística, la famosa fórmula $S = k_B \log W$ (Boltzmann, 1896, 1898; Jaynes, 1965; Merhav, 2010, 2018).

Una debilidad del enfoque de Hartley es que considera implícitamente que en un mensaje, cada cadena de longitud dada puede aparecer con la misma frecuencia, o probabilidad $1/\alpha^n$ (en Boltzmann, misma probabilidad de cada configuración), siendo la información menos el logaritmo de estas probabilidades. Al contrario, parece más lógico considerar que secuencias muy frecuentes no llevan mucha información (se sabe que aparecen), mientras que las que aparecen raramente llevan más información (hay más sorpresa, más incerteza en observarlas). Volviendo a los símbolos elementales x, vistos como aleatorios (o valores, o estados que puede tomar una variable aleatoria), la (falta de) información o incerteza va a estar íntimamente vinculada a la probabilidad de aparición de estos símbolos x. Siguiendo la idea de Hartley, la información elemental asociada al estado x va a ser $-\log p(x)$ donde $p(x)$ es la probabilidad de aparición de x. Se define la incerteza asociada a la variable aleatoria como el promedio estadístico sobre todos los estados posibles x (Shannon, 1948; Shannon & Weaver, 1964) [96].

Definición 2-57 (Entropía de Shannon)**.** *Sea X una variable aleatoria definida sobre un alfabeto discreto $X(\Omega) = \mathcal{X} = \{x_1, \ldots, x_\alpha\}$ de cardinal $\alpha = |\mathcal{X}| < +\infty$ finito. Sea p_X la distribución de probabilidad de X, i. e., $\forall\, x \in \mathcal{X}, \quad p_X(x) = P(X = x)$. La entropía de Shannon de la variable X está definida por*

$$H(p_X) = H(X) = -\sum_{x \in \mathcal{X}} p_X(x)\, \log p_X(x),$$

[96] En la misma época que Shannon, independientemente, medidades informacionales aparecieron en cálculos de capacidad de canal en varios trabajos como los de los ingenieros franceses André Clavier (Clavier, 1948) o Jacques Laplume (Laplume, 1948), o en el libro del estadounidense Norbert Wiener (Wiener, 1948, Cap. III) entre varios otros (ver (Verdu, 1998; Lundheim, 2002; Rioul & Magossi, 2014; Flandrin & Rioul, 2016; Rioul & Flandrin, 2017; Chenciner, 2017, y Ref.)).

con la convención $0 \log 0 = 0$ *(*$\lim_{t \to 0} t \log t = 0$*).*

La base del logaritmo es arbitraria; si es $\log_2$ el logaritmo de base 2, H está en unidades binarias o bits (se encuentra tambíen la denominación Shannons), si se usa el logaritmo natural $\ln$, H está en unidades naturales o nats, si es el de base 10, H se da en dígitos decimales o dits (se encuentra tambíen la denominación bans o Hartleys). *En este capítulo, se usará* H *sin especificar la base del logaritmo. Si es necesario que tenga una base* a *dada, se denotará la entropía correspondiente* H_a *y se especificará la base del logaritmo* $\log_a$. Notar que $\log_a x = \frac{\log x}{\log a}$, dando

$$H_a(X) = H_b(X) \log_a b.$$

En lo que sigue, aún que, rigurosamente, H sea una función de la distribución de probabilidad p_X y no de la variable X, se usará indistintamente tanto la notación $H(p_X)$ como $H(X)$ según lo más conveniente. Además, p_X podrá denotar indistintamente la distribución de probabilidad, o el vector de probabilidad $p_X \equiv \begin{bmatrix} p_X(x_1) & \cdots & p_X(x_\alpha) \end{bmatrix}^t$. En lo que sigue, de vez a cuando, usaremos $p_i \equiv p_X(x_i)$ por simplificación de escritura.

H es el equivalente de la entropía de Gibbs en mecánica estadística (denotata S en este marco), como lo hemos visto en la sección 1.10.3. Precisamente, se reconoce en la fórmula de le entropía de Shannon la forma ya vista tratando de la familia exponencial, sección 1.10.3, como fluctuaciones de la energía libre. En la sección mencionada, la vimos en el contexto de la ley de Boltzmann, dada una función energía. En el contexto de este capítulo, quitamos el contexto de la mecánica estadística y la definición es dada para cualquier densidad de probabilidad. Sin embargo, los vínculos entre ambas son impotante, como lo vamos a ver también en el problema de entropía máxima (ver también el epígrafo de este capítulo). La letra H viene del teorema-H debido a... Ludwig Boltzmann (Jaynes, 1965; Merhav, 2010, 2018), a pesar de que en mecánica estadística se encuentra frecuentemente la letra... S (sin relación con Shannon) y también en mecánica cúantica tratando de la entropía de von Neuman (ver más adelante).

La entropía de Shannon H tiene propiedades notables que corresponden a las que se puede exigir a una medida de incerteza (Shannon, 1948; Shannon & Weaver, 1964; Cover & Thomas, 2006; Rioul, 2007; Dembo, Cover & Thomas, 1991; Johnson, 2004).

[P1] *Continuidad:* vista como una función de α variables $p_i = p_X(x_i)$, H es continua con respeto a los p_i.

[P2] *Invariance bajo una permutación:* obviamente, la entropía es invariante bajo una permutación de las probabilidades, *i. e.,*

$$\text{para cualquiera permutación } \sigma : \mathcal{X} \to \mathcal{X}, \quad H(p_{\sigma(X)}) = H(p_X) \quad \text{con} \quad p_{\sigma(X)}(x) = p_X(\sigma(x)),$$

lo que se escribe también $H(\sigma(X)) = H(X)$. En particular, denotando $p_X^{\downarrow}$ el vector de probabilidades obtenido a partir de p_X, clasificando las probabilidades en orden decreciente, $p_1^{\downarrow} \geq p_2^{\downarrow} \geq$

$\cdots \geq p^{\downarrow}_{\alpha}$ donde $p^{\downarrow}_{i}$ es la i-ésima componente de $p^{\downarrow}_{X}$ (ver sección 1.3.3, definición 1-23),

$$H(p^{\downarrow}_{X}) = H(p_X).$$

[P3] *Invariance bajo una transformación biyectiva:* la entropía es invariante bajo cualquiera transformación biyectiva, *i. e.,*

$$\text{para cualquiera función biyectiva } g : \mathcal{X} \to g(\mathcal{X}), \quad H(g(X)) = H(X).$$

A través tal transformación los estados cambian, pero no cambia la distribución de probabilidad vinculada al alfabeto transformado. Tomando el ejemplo de un dado, la incerteza vinculada al dado no debe depender de los símbolos escritos sobre las caras, sean enteras o cualesquiera letras.

[P4] *Positividad:* la entropía es acotada por debajo,

$$H(X) \geq 0,$$

con igualdad si y solamente si existe un $x_j \in \mathcal{X}$ tal que $p_X(x_j) = 1$ y $p_X(x) = 0$ para $x \neq x_j$, *i. e.,* $p_X = \mathbb{1}_j$,

$$H(X) = 0 \quad \text{ssi} \quad X \text{ es determinista.}$$

En otras palabras, cuando X no es aleatoria, *i. e.,* $X = x_j$, no hay incerteza, o la observación no lleva información (se sabe lo que va a salir, sin duda): $H = 0$. La positividad es consecuencia de $p_X(x) \leq 1$, dando $-p_X(x) \log p_X(x) \geq 0$. Además, la suma de términos positivos vale cero si y solamente si cada término de la suma vale cero, dando $p_X(x) = 0$ o $p_X(x) = 1$. Se concluye p_X siendo una distribución de probabilidad, sumando a 1.

[P5] *Maximalidad:* la entropía es acotada por arriba,

$$H(X) \leq \log \alpha,$$

con igualdad si y solamente si X es uniforme sobre $\mathcal{X}$, *i. e.,*

$$H(X) = \log \alpha \quad \text{ssi} \quad \forall\, x \in \mathcal{X},\ p_X(x) = \frac{1}{\alpha}.$$

En otras palabras, la incerteza es máxima cuando cualquier estado x puede aparecer con la misma probabilidad; cada observación lleva una información importante sobre el sistema que genera X. La cota máxima resuelta de la maximización de H sujeto a $\sum_x p_X(x) = 1$, es decir, con la técnica del Lagrangiano para tomar en cuenta el vínculo (Miller, 2000; Cambini & Martein, 2009), notando $p_i = p_X(x_i)$, hay que minimizar $\sum_i (-p_i \log p_i + \eta\, p_i)$ donde el factor de Lagrange η se determinará para satisfacer el vínculo. Se obtiene sencillamente que $\log p_i = -\eta$, dando la distribución uniforma.

La figura Fig. 2-34 representa la entropía de un sistema a dos estados, de probabilidades $p_X =$

$\begin{bmatrix} 1-p & p \end{bmatrix}^t$ (ley de Bernoulli de parametro p), entropía a veces dicha *entropía binaria*, en función de p. Esta figura ilustra ambas cotas ($p = 1$ o 1, $p = \frac{1}{2}$) así que la invariancia bajo una permutación ($h(p) = H(1-p, p) = H(p, 1-p) = h(1-p)$).

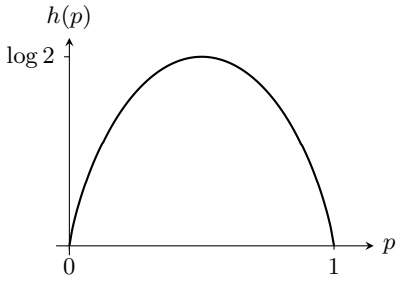


Figura 2-34: Entropía binaria (de una variable de Bernoulli) $h(p) = H(1-p, p)$ en función de $p \in [0\,;\,1]$.

[P6] *Expansibilidad:* Añadir un estado de probabilidad 0 no cambia la entropía, *i. e.,* sean X definido sobre $\mathcal{X}$ y $\widetilde{X}$ sobre $\widetilde{X} = \mathcal{X} \cup \{\widetilde{x}_0\},\ \widetilde{x}_0 \notin \mathcal{X}$, con

$$p_{\widetilde{X}}(x) = p_X(x) \quad \text{si} \quad x \in \mathcal{X}, \quad p_{\widetilde{X}}(\widetilde{x}_0) = 0, \qquad \text{entonces} \quad H(p_{\widetilde{X}}) = H(p_X).$$

Esta propiedad es obvia, consecuencia de $\lim\limits_{t \to 0} t \log t = 0$.

[P7] *Recursividad:* Juntar dos estados baja la entropía de una cantidad igual a la entropía interna de los dos estados por la probabilidad de ocurrencia de este conjunto de estados, y vice-versa. De la invarianza de la entropía por permutación, sin perdida de generalidad se puede considerar que los estados que se juntan son los dos últimos, *i. e.,* sean X definido sobre $\mathcal{X}$ y $\breve{X}$ sobre $\breve{\mathcal{X}}$ tales que,

$$\begin{cases} \breve{\mathcal{X}} = \{x_1, \ldots, x_{\alpha-2}, \breve{x}_{\alpha-1}\} \quad \text{con el estado interno} \quad \breve{x}_{\alpha-1} = \{x_{\alpha-1}\,,\, x_\alpha\}, \\ p_{\breve{X}}(x_i) = p_X(x_i), \quad 1 \le i \le \alpha - 1 \quad \text{y} \quad p_{\breve{X}}(\breve{x}_{\alpha-1}) = p_X(x_{\alpha-1}) + p(x_\alpha) \quad \text{distribución sobre } \breve{\mathcal{X}} \\ \breve{q}(x_j) = \dfrac{p_X(x_j)}{p_X(x_{\alpha-1}) + p_X(x_\alpha)}, \quad j = \alpha - 1, \alpha \quad \text{distribución del estado interno} \end{cases}$$

entonces,

$$H(p_X) = H(p_{\breve{X}}) + p_{\breve{X}}(\breve{x}_{\alpha-1})\, H(\breve{q}),$$

lo que se escribe también

$$H(p_1, \ldots, p_\alpha) = H(p_1, \ldots, p_{\alpha-2}, p_{\alpha-1} + p_\alpha) + (p_{\alpha-1} + p_\alpha)\, H\left(\frac{p_{\alpha-1}}{p_{\alpha-1} + p_\alpha}, \frac{p_\alpha}{p_{\alpha-1} + p_\alpha}\right).$$

Esta relación viene de $a \log a + b \log b = (a+b)\left(\frac{a}{a+b} \log\left(\frac{a}{a+b}\right) + \frac{b}{a+b} \log\left(\frac{a}{a+b}\right) - \log(a+b)\right)$ es ilustrada en la figura Fig. 2-35.

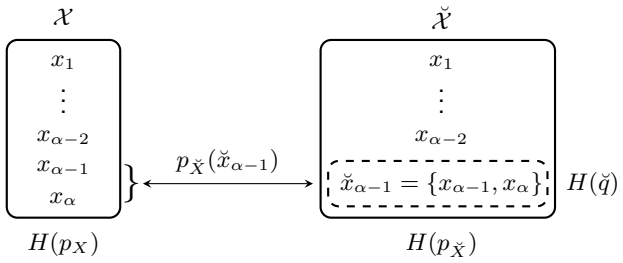


Figura 2-35: Ilustración de la propiedad de recursividad, que cuantifica como decrece la entropía en un conjunto cuando se juntan dos estados, relacionando la entropía total, la entropía después del la agrupación y la entropía interna a los dos estados juntados.

[P8] *Concavidad:* la entropía es cóncava ($-H$ es convexa, ver definición 1-36), en el sentido de que la entropía de una combinación convexa de distribuciones (mezcla) de probabilidades es siempre mayor o igual a la combinación convexa de entropías:

$$\forall\, \{\pi_i\}_{i=1}^n, \quad 0 \leq \pi_i \leq 1, \quad \sum_{i=1}^n \pi_i = 1 \quad \text{and cualquier conjunto de distribuciones} \quad \{p_{(i)}\}_{i=1}^n,$$

$$H\left(\sum_{i=1}^n \pi_i\, p_{(i)}\right) \geq \sum_{i=1}^n \pi_i H(p_{(i)}).$$

Esta relación es conocida también como desigualdad de Jensen (Jensen, 1906). Es una consecuencia directa de la convexidad de la función $\phi : t \mapsto t\log t$, como ilustrado en la figura Fig. 2-36-(a). La figura Fig. 2-36-(b) ilustra como se puede obtener una mezcla de distribuciones de dos probabilidad $p_{(1)}$ (dado izquierda) y $p_{(2)}$ (dado derecho) haciendo una elección aleatoria a partir de una moneda en este ejemplo (probabilidad $\pi_1 = 1 - \pi_2$ de elegir el dado izquierda).

[P9] *Schur-concavidad:* como se lo puede querrer, lo más "concentrado" es una distribución de probabilidad, lo menos hay incerteza, y entonces lo más pequeño debe ser la entropía. Esta propiedad intuitiva se resuma a partir de la noción de mayorización vista en la definición 1-23, (recuerdense que si los vectores no tienen el mismo tamaño, el más pequeño es completado por ceros; es equivalente a añadir estados fictivos de probabilidad nula, lo que no cambia la entropía). La Schur-concavidad se traduce por la relación

$$p \prec q \quad \Rightarrow \quad H(p) \geq H(q).$$

Fijense de que las cotas sobre H pueden ser vistas como consecuencias de esta desigualdad: la distribución cierta mayoriza cualquiera distribución y cualquiera distribución mayoriza la distribución uniforme (Marshall et al., 2011, p. 9, (6)-(8)). Además, de la Schur-concavidad se obtiene que

$$H\left(\begin{bmatrix}\frac{1}{\alpha} & \cdots & \frac{1}{\alpha}\end{bmatrix}^t\right) \quad \text{es una función creciente de } \alpha.$$

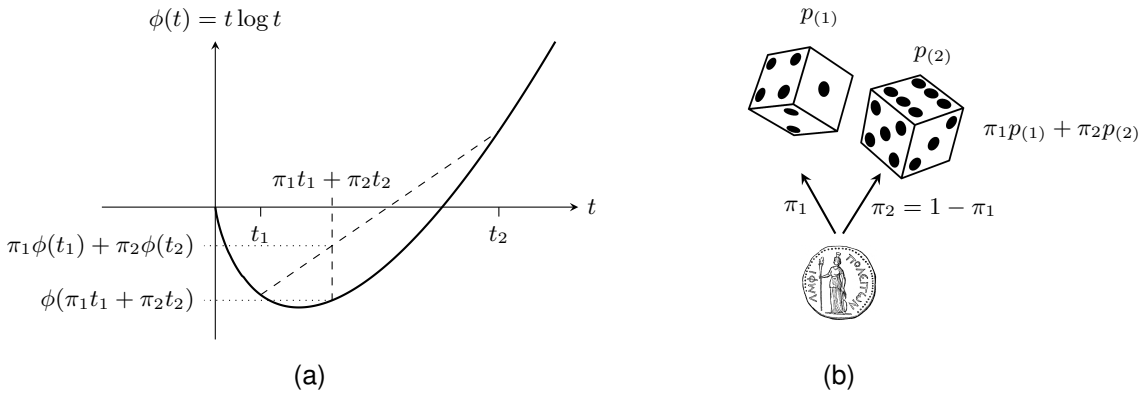


Figura 2-36: (a) $\phi(t) = t \log t$ es convexa: la curva es siempre debajo de sus cuerdas; entonces, cada promedio de $\phi(t_1)$ y $\phi(t_2)$ estando en la cuerda juntando estos puntos, queda arriba de la función tomada en el promedio de t_1 y t_2. Escribiendo eso para (más de dos puntos) sobre los $\sum_i \pi_i \, p_{(i)}(x)$ y sumando sobre los x da la desigualdad de Jensen. (b) Ilustración de una distribución de mezcla, acá mezclando $p_{(1)}$ y $p_{(2)}$ a partir de una tercera variable aleatoria (acá de Bernoulli).

> La prueba de la Schur-concavidad se apoya sobre la desigualdad de Schur o Hardy-Littlewood-Pólya o Karamata (Schur, 1923; Hardy, Littlewood & Pólya, 1929; Karamata, 1932; Hardy, Littlewood & Pólya, 1952), (Marshall et al., 2011, Cap. 3, Prop. C.1) o (Bhatia, 1997, Teorema II.3.1): $t \prec t' \;\Rightarrow\; \sum_i \phi(t_i) \leq \sum_i \phi(t'_i)$ para cualquiera función ϕ convexa. Basta considerar $\phi(t) = t \log t$ para concluir.

En muchos casos, uno tiene que trabajar con varias variables aleatorias. Para simplificar las notaciones, consideramos un par de variables X y Y definidas respectivamente sobre los alfabetos $\mathcal{X}$ y $\mathcal{Y}$ de cardinal $\alpha = |\mathcal{X}|$ y $\beta = |\mathcal{Y}|$. Tal par de variables puede ser vista como una variable $\begin{bmatrix} X \\ Y \end{bmatrix}$ definida sobre el alfabeto $\mathcal{X} \times \mathcal{Y}$ de cardinal $\alpha\beta$ tal que se define naturalmente la entropía para esta variable; tal entropía es llamada *entropía conjunta* de X y Y:

Definición 2-58 (Entropía conjunta)**.** *Sean X e Y dos variables aleatorias definidas sobre los alfabetos discretos $\mathcal{X}$ y $\mathcal{Y}$, de cardinal $\alpha = |\mathcal{X}| < +\infty$ y $\beta = |\mathcal{Y}| < +\infty$ respectivamente. Sea $p_{X,Y}$ la distribución de probabilidad conjunta de X e Y, i. e., $\forall\,(x,y) \in \mathcal{X} \times \mathcal{Y}, \quad p_{X,Y}(x,y) = P((X = x) \cap (Y = y))$. La entropía conjunta de Shannon de las variables X e Y es definida por*

$$H(p_{X,Y}) = H(X,Y) = -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{X,Y}(x,y) \, \log p_{X,Y}(x,y),$$

con la convención $0 \log 0 = 0$.

A partir de esta definición, aparecen otras propiedades importantes, sino que fundamentales, de la entropía de Shannon.

[P10] *Aditividad:* la entropía conjunta de dos variables aleatorias X e Y independientes se suma, y

recíprocamente:

$$X \text{ e } Y \text{ independientes} \quad \Leftrightarrow \quad H(X,Y) = H(X) + H(Y).$$

Dicho de otra manera, para dos variables aleatorias, la incerteza global es la suma de las incertezas de cada variable individual. La propiedad "⇒" es consecuencia directa de $p_{X,Y}(x,y) = p_X(x)p_Y(y)$. Se va a probar en la sección siguiente la recíproca. Esta propiedad se escribe también

$$H\left(p_X \otimes p_Y\right) = H\left(p_X\right) + H\left(p_Y\right),$$

donde $\otimes$ es el producto externo (ver notaciones) Se generaliza sencillamente a un conjunto de variables aleatorias $\{X_i\}_{i=1}^n$ (o, equivalentemente a un producto externo de un conjunto de vectores de probabilidades).

[P11] *Sub-aditividad:* la entropía conjunta de n variables aleatorias $\{X_i\}_{i=1}^n$ es siempre menor que la suma de cada entropía individual:

$$H(X_1,\ldots,X_n) \leq \sum_{i=1}^{n} H(X_i) \qquad \textit{i. e.,} \qquad H\left(p_{X_1,,\ldots,X_n}\right) \leq H\left(p_{X_1} \otimes \cdots \otimes p_{X_n}\right) = \sum_{i=1}^{n} H\left(p_{X_i}\right).$$

Dicho de otra manera, las variables aleatorias pueden compartir información, de tal manera que la entropía global sea menor que la suma de cada entropía. De la propiedad anterior, se obtiene la igualdad si y solamente si los X_i son independientes.

[P12] *Super-aditividad:* la entropía conjunta de n variables aleatorias $\{X_i\}_{i=1}^n$ es siempre mayor que cualesquiera de las entropías individuales

$$H(X_1,\ldots,X_n) \geq \max_{1\leq i\leq n} H(X_i).$$

Es importante notar que existen varios enfoques basados sobre una serie de axiomas, dando lugar a la definición de la entropía tal como definida. Estos axiomas son conocidos como axiomas de Shannon-Khinchin y son la continuidad [P1], la maximalidad [P5], la expansabilidad [P6] y la aditividad [P10]. Existen varios otros conjuntos de axiomas, conduciendo también a la entropía de Shannon (ver (Shannon, 1948, Sec. 6) o (Shannon & Weaver, 1964; Fadeev, 1956, 1958; Khinchin, 1957; Rényi, 1961), entre otros).

Para una serie de variables aleatorias, $X_1, X_2, \ldots$, representando símbolos, se puede definir una entropía por símbolo como una entropía conjunta divido por el número de símbolos, $\frac{H(X_1,\ldots,x_n)}{n}$, así que una tasa de entropía cuando n va al infinito.

Definición 2-59 (Tasa de entropía)**.** *Sea $X \equiv \{X_i\}_{i\in\mathbb{N}^*}$ una serie de variables aleatorias, o proceso estocástico. La tasa de entropía del proceso es definida por*

$$\mathcal{H}(X) = \lim_{n\to\infty} \frac{H(X_1,\ldots,X_n)}{n}.$$

Esta cantidad siempre existe porque $H(X_1,\ldots,X_n) \leq \sum_{i=1}^{n} H(X_i) \leq \sum_{i=1}^{n} \log \alpha_i \leq n \max_{1\leq i\leq n} \alpha_i$ donde los α_i son los cardinales de los alfabetos de definición de los X_i.

Se termina esta subsección con el caso de variables discretas definidas sobre un alfabeto $\mathcal{X}$ de cardinal infinito $|\mathcal{X}| = +\infty$, por ejemplo $\mathcal{X} = \mathbb{N}$. Por analogía, se puede siempre definir la entropía como en la definición Def. 2-57. Esta extensión resuelta delicada dando de que unas propiedades se perdien. Por ejemplo, la entropía no queda acotada por arriba como se lo puede probar para la distribución de probabilidad $p(x) \propto \dfrac{1}{(x+2)\left(\log(x+2)\right)^2}$, $x \in \mathbb{N}$, correctamente normalizada ($\propto$ significa "proporcional a"): $\dfrac{\log\log(x+2)}{(x+2)\left(\log(x+2)\right)^2} \geq 0$ y la serie $\sum_x \dfrac{1}{(x+2)\log(x+2)}$ es divergente, así que la serie $-\sum_x p(x)\log p(x)$ diverge.

2.2.2 Entropía diferencial

Volviendo a la definición Def. 2-57 de la entropía de Shannon, usando el operador E promedio estadístico o esperanza matemática, se puede reescribir la entropía de Shannon como $H(X) = \mathrm{E}\left[-\log p_X(X)\right]$. Con este punto de vista, es fácil extender la definición de la entropía para variables aleatorias continuas admitiendo una densidad de probabilidad. Eso da lugar a lo que es conocido como la *entropía diferencial*:

Definición 2-60 (Entropía diferencial)**.** *Sea X una variable aleatoria continua admitiendo una densidad de probabilidad p_X, definida sobre $\mathbb{R}^d$ y $X(\Omega) = \mathcal{X} = \{x \in \mathbb{R}^d : p_X(x) > 0\} \subseteq \mathbb{R}^d$ el soporte de $p_X(x)$. La entropía diferencial de la variable X es definida por*

$$H(p_X) = H(X) = -\int_{\mathcal{X}} p_X(x)\,\log p_X(x)\,dx$$

(con la convención $0\log 0 = 0$, se puede escribir la integración en $\mathbb{R}^d$).

Como en el caso discreto, para $X = (X_1,\ldots,X_d)$, esta entropía de X es dicha entropía conjunta de los X_i.

Como se lo va a ver, la entropía diferencial no tiene la misma significación de incerteza, siendo de que depende no solamente de la distribución de probabilidad, sino que de los estados también. Más allá, no se la puede ver como límite continua de un caso discreto: a través de tal límite, se va a ver que se llama diferencial, a causa del efecto de la "diferencial dx". Para ilustrar este hecho, consideramos una variable aleatoria escalar X y p_X su densidad de probabilidad de soporte $\mathbb{R}$. Sea $\Delta > 0$ y sea el alfabeto $\mathcal{X}^\Delta = \{x_k\}_{k\in\mathbb{Z}}$ donde los x_k se definen tal que $p_X(x_k)\Delta = \int_{k\Delta}^{(k+1)\Delta} p_X(x)\,dx$, como ilustrado en la figura Fig. 2-37. Se define la variable aleatoria discreta $X^\Delta = \sum_k x_k \mathbb{1}_{(X\in[k\Delta\,;\,(k+1)\Delta))}$ sobre $\mathcal{X}^\Delta$ tal que $P(X^\Delta = x_k) = p_{X^\Delta}(x_k) = p_X(x_k)\Delta$. Se puede ver X^Δ como la versión cuantificada de X,

con $X^\Delta = x_k$ cuando $X \in [k\Delta\,;\,(k+1)\Delta)$. Al revés, aún que sea delicado, se puede interpretar X como el "límite" de X^Δ cuando Δ tiende a 0. Ahora, es claro que

$$\begin{aligned} H(X^\Delta) &= -\sum_k p_{X^\Delta}(x_k)\log p_{X^\Delta}(x_k) \\ &= -\log\Delta - \sum_k \Big(p_X(x_k)\log p_X(x_k)\Big)\,\Delta \end{aligned}$$

lo que se escribe también

$$H(X^\Delta) + \log\Delta = -\sum_k \Big(p_X(x_k)\log p_X(x_k)\Big)\,\Delta.$$

Entonces, de la integración de Riemann sale que

$$\lim_{\Delta\to 0}\big(H(X^\Delta) + \log\Delta\big) = H(X).$$

Dicho de otra manera, la entropía diferencial de X no es el límite de la entropía de su versión cuantificada: aparece con la entropía el término "diferencial" $\log\Delta$.

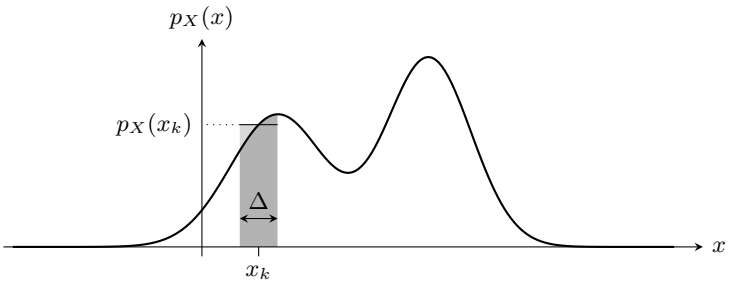


Figura 2-37: Densidad de probabilidad p_X de X, construcción del alfabeto $\mathcal{X}^\Delta$ donde se define la versión cuantificada X^Δ de X con su distribución discreta de probabilidad p_{X^Δ}. La superficia en gris oscuro es igual a la superficia definida por el rectángulo en gris claro.

Más allá de esta notable diferencia entre la entropía y la entropía diferencial, la última depende de los estados, es decir que si $Y = g(X)$ con g biyectiva, no se conserva la entropía, *i. e.,* se pierde la propiedad [P3] del caso discreto:

$$\begin{aligned} H(Y) &= -\int_{\mathbb{R}^d} p_Y(y)\log p_Y(y)\,dy \\ &= -\int_{\mathbb{R}^d} p_Y(g(x))\log p_Y(g(x))\,|\,\mathrm{J}_g(x)|\,dx \\ &= -\int_{\mathbb{R}^d} p_Y(g(x))\Big(\log\big(p_Y(g(x))\,|\,\mathrm{J}_g(x)|\big) - \log|\mathrm{J}_g(x)|\Big)\,|\,\mathrm{J}_g(x)|\,dx \end{aligned}$$

donde J_g es la matriz Jacobiana de la transformación $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ y $|\cdot|$ representa el valor absoluto del determinante de la matriz (ver notaciones). Recordando que $p_X(x) = p_Y(g(x))|\,\mathrm{J}_g(x)|$ (ver subsección **??**), se obtiene la propiedad siguiente:

[P'3] Para cualquiera biyección $g : \mathbb{R}^d \mapsto \mathbb{R}^d$

$$H(g(X)) = H(X) + \int_{\mathbb{R}^d} p_X(x) \log |\,\mathrm{J}_g(x)|\, dx,$$

donde el último término, $\mathrm{E}\left[\log |\,\mathrm{J}_g(X)|\right]$ no vale cero en general. En particular, si H es invariante bajo una translación,

$$H(X + b) = H(X) \quad \forall\, b \in \mathbb{R}^d,$$

no es invariante por cambio de escala,

$$H(aX) = H(X) + \log |a| \quad \forall\, a \in \mathbb{R}^*.$$

Esta última relación queda válido para a matriz invertible. Por esta última relación, se puede ver que, dado X, cuando a tiende a 0, la entropía de aX tiende a $-\infty$. Es decir que, para a suficientemente pequeño, se puede tener $H(aX) < 0$, así que se pierde también la positividad [P4]. Por esta perdida, se quita definitivamente la interpretación de incerteza/información que hubiera podido tener la entropía diferencial.

A veces, se usa lo que es llamado potencia entrópica:

Definición 2-61 (Potencia entrópica)**.** *Sea X una variable aleatoria d-dimensional. La potencia entrópica de X es definida por*

$$N(X) = \frac{1}{2\pi\,\mathrm{e}} \exp\left(\frac{2}{d} H(X)\right).$$

Por construcción, $N(X) \geq 0$. Además, en el caso continuo, $N(aX + b) = |a|^2 N(X)$ (queda válido para una matriz a invertible): esta propiedad puede justificar la idea de "potencia"; además $N(aX + b)$ tiende naturalmente a cero cuando a tiende a cero. Se recupera así la noción informacional a través de N en este contexto ($aX + b$ "tiende" a b, variable determinista).

Si se pierde la propiedad de invarianza bajo una biyección, sorprendentemente, se conserva la entropía bajo el equivalente continuo del rearreglo.

[P'2] *invarianza bajo un rearreglo:* Sea p_X densidad de probabilidad sobre un abierto de $\mathbb{R}^d$,

$$H\left(p_X^{\downarrow}\right) = H(p_X).$$

donde $p_X^{\downarrow}$ es el rearreglo símetrico de p_X definido Def. **??**.

Esta propiedad es probada para funciones convexas de la densidad de probabilidad por ejemplo en (Lieb & Loss, 2001) o (Wang & Madiman, 2004, Lema 7.2) [97], y entonces para el caso particular $\phi(t) = t \log t$.

[97] En (Lieb & Loss, 2001, Sec. 3.3) lo muestran para $\phi(p_X)$ donde ϕ es la diferencia de dos funciones monotonas, siendo $\phi(t) = t \log t$ un caso particular.

Una pregunta natural es de saber lo que pasa en término de mayorización en el contexto continuo d-dimensional. Aparece que la Schur-concavidad [P9] se conserva en el caso continuo, *i. e.,*

$$p \prec q \quad \Rightarrow \quad H(p) \geq H(q).$$

con la relación de mayorización continua vista Def. **??**. La desigualdad inversa es probada para cualquier función ϕ convexa de la densidad (Chong, 1974) o (Wang & Madiman, 2004, Prop. 7.3), en particular para $\phi(t) = t \log t$.

Como se lo ha visto, la entropía diferencial no es siempre positiva, como consecuencia de la propiedad [P'3]. También, la propiedad de cota superior [P5] se pierde en general, salvo si se ponen vínculos:

[P'5] *a*) Si $\mathcal{X}$ es de volumen finito $|\mathcal{X}| < +\infty$, la entropía es acotada por arriba,

$$H(X) \leq \log |\mathcal{X}|,$$

con igualdad si y solamente si X es uniforme.

b) Si $\mathcal{X} = \mathbb{R}^d$ y X tiene una matriz de covarianza dada $\Sigma_X = \mathrm{E}\left[XX^t\right] - m_X m_X^t$ $(m_X = \mathrm{E}[X])$, la entropía es también acotada por arriba,

$$H(X) \leq \frac{d}{2} \log(2\pi\,\mathrm{e}) + \frac{1}{2} \log |\Sigma_X|,$$

con igualdad si y solamente si X es gaussiana. En particular, la potencia entrópica de la gaussiana vale $N(X) = |\Sigma_X|^{\frac{1}{d}}$, dando de nuevo un "sabor" de potencia a N. Como se lo va a ver en este capítulo, la gaussiana juega un rol central en la teoría de la información.

En ambos casos, estas desigualdades con la distribución maximizante se obtienen resolviendo el problema de maximización de la entropía sujeto a vínculos. Se trata del caso más general en la subsección Sec. 2.4.1.

Al final, se conservan obviamente las propiedades de concavidad [P8], de aditividad [P10] y de sub-aditividad [P11]. Es interesante notar que de la desigualdad [P11], puramente entrópica, se puede deducir la desigualdad de Hadamard, desigualdad puramente matricial: $|R| \leq \prod_i R_{i,i}$ para cualquiera matriz simétrica definida positiva (viene de la propiedad [P11] escrita para una gaussiana de covarianza R y tomando una exponencial de la desigualdad).

Como lo hemos visto, la entropía y su versión diferencial no tienen ni las mismas propiedades, ni completamente la misma interpretación. Sin embargo, varias propiedades se comparten y se proban de la misma manera. De las escrituras, con una suma o una integral, a veces se encuentra en la literatura la escritura única $\sum\!\!\!\!\!\!\int$ para significar que se usa la suma en el caso discreto, y la integración en el caso continuo con densidad (Rioul, 2007). Sin embargo, volviendo al fin de la subsección 1.3.4 y a la definición 1-15 de una medida discreta sobre $\mathcal{X} = \{x_j\}_j$ dada por $\mu_{\mathcal{X}} = \sum_j \delta_{x_j}$, vimos de que en el caso discreto p_X es la densidad de la medida de probabilidad con respeto a $\mu_{\mathcal{X}}$. Además, de la

propiedad $\int_{\mathbb{R}} f(x)\, d\delta_{x_j} = f(x_j)$ se puede ver una suma como una integral con respeto a une medida discreta. De estas observacione, se puede escribir de la misma forma la entropía discreta y diferencial:

Definición 2-62 (Escritura única de la entropía)**.** *Sea X variable aleatoria definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$, admitiendo una densidad de probabilidad p_X con respeto a una medida μ (ej. $\mu_{\mathcal{X}}$ en el caso discreto $\mu = \mu_L$ en el caso diferencial). La entropía de X con respeto a μ se escribe como*

$$H(X) \equiv H(p_X) = -\int_{\mathcal{X}} p_X(x)\, \log(p_X)\, d\mu(x)$$

Insistamos en el hecho de que se puede entender esta definición para cualquier μ y densidad con respeto a μ, que sea discreta, de Lebesgue, o cualquiera.

2.3 Entropía condicional, información mutua, entropía relativa

Tratando de un par de variables aleatorias X e Y, una cuestión natural que ocurre es de cuantificar la incerteza que queda sobre una de las variables cuando se observa la otra. Dicho de otra manera, si se mide $Y = y$, ¿que información lleva sobre X? La respuesta a esta interogación se encuentra en la noción de entropía condicional. Si uno mide $Y = y$, la descripción estadística de X conociendo este $Y = y$ se resuma a la distribución condicional de probabilidad $p_{X|Y=y} = \frac{p_{X,Y}(\cdot,y)}{p_Y(y)}$. Con esta restricción, se puede evaluar una incerteza sobre X, sabiendo que $Y = y$,

$$H(X|Y = y) = H\left(p_{X|Y=y}\right).$$

Entonces, condicionalmente a la variable aleatoria Y, la incerteza va a ser el promedio estadístico sobre todos los estados Y es decir $H(X|Y) = \int_{\mathbb{R}^d} p_Y(y) H(X|Y = y)\, d\mu(y)$ (con la medida μ adecuada).

Definición 2-63 (Entropía condicional)**.** *Sean X e Y dos variables aleatorias, respectivamente d y d'-dimensionales. La entropía condicional de X, con respeto a Y, es definida por*

$$H(X|Y) = -\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} p_{X,Y}(x,y) \log p_{X|Y=y}(x)\, d\mu(x,y),$$

con μ medida adecuada (ej. discreta en el caso discreto o de Lebesgue en el caso diferencial).

Si X e Y son independientes, $p_{X|Y=y}$ se reduce a p_X, así que obviamente,

[P13]

$$X \text{ e } Y \text{ independientes} \quad \Leftrightarrow \quad H(X|Y) = H(X).$$

Esta propiedad se interpreta como el hecho que Y no lleva ninguna información sobre X, y entonces ninguna medición de Y va a cambiar la incerteza sobre X.

Siendo $H(X|Y = y)$ una entropía, va a heredar de todas las propiedades de la entropía (o entropía diferencial). Además, de $p_{X,Y}(\cdot, y) = p_{X|Y=y} p_Y(y)$ se deduce la propiedad siguiente

[P14] *Regla de cadena*

$$H(X,Y) = H(X|Y) + H(Y).$$

Esta regla se generaliza sencillamente a

$$H(X_1,\ldots,X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_{i-1},\ldots,X_1).$$

De esta regla de cadena se recupera la propiedad [P13] a partir de la propiedad [P10].

Siendo $H(X|Y=y)$ una entropía, en el caso discreto esta cantidad es positiva. Entonces, en el caso discreto, $H(X|Y)$ es positiva, lo que prueba la super-aditividad [P12].

De la regla de cadena $H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$ aparece que las cantidades $H(X|Y) - H(X)$, $H(Y|X) - H(Y)$ y $H(X,Y) - H(X) - H(Y)$ son todas iguales. Estas cantidades definen lo que se llama la información mutua entre X e Y:

Definición 2-64 (Información mutua)**.** *Sean X e Y dos variables aleatorias, la información mutua entre X e Y es la cantidad simétrica*

$$I(X;Y) = H(X|Y) - H(X) = H(Y|X) - H(Y) = H(X,Y) - H(X) - H(Y);$$

Se expresa

$$I(X;Y) = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right) d\mu(x,y)$$

con μ medida adecuada (discreta en el caso discreto o de Lebesgue en el caso diferencial).

Las diferentes cantidades pueden ser vistas a través de una visión ensemblista, como descrita en la figura Fig. 2-38, diagrama de Venn o de Euler (ver nota de pie 9.

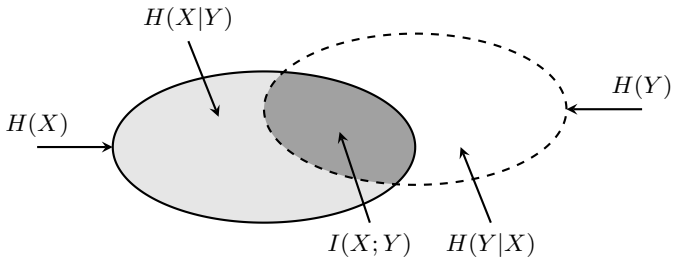


Figura 2-38: Diagrama de Venn: Ilustración de la definición de la entropía condicional, de la información mutua, y de las relaciones entre cada medida. La superficia del elipse en linea llena (parte grise) representa $H(X)$ y el interior de la en linea discontinua representa $H(Y)$. La parte grise clara representa $H(X|Y)$ superficia del "conjunto $H(X)$" quitando la parte que partenece a $H(Y)$. La parte blanca representa $H(Y|X)$ superficia del "conjunto $H(Y)$" quitando la parte que partenece a $H(X)$. La parte en grise oscuro es entonces lo que X e Y comparten, es decir $I(X;Y)$.

Como se lo va a probar, I es positiva; representa realmente una información, la compartida entre X e Y: Si de la incerteza de X se quita la incerteza de X una vez que Y es medida, lo que queda tiene la

significación de la información que estas variables tienen en común. En particular, de $I(X;X) = H(X)$ se denoma a veces $H(X)$ *auto información* de X.

Para probar la positividad de I, se introduce de manera más general la noción de entropía relativa, conocida también como divergencia de Kullback-Leibler (Kullback & Leibler, 1951; Kullback, 1968; Cover & Thomas, 2006; Rioul, 2007):

Definición 2-65 (Entropía relativa)**.** *La entropía relativa, o divergencia de una medida de probabilidad Q, con respeto a una medida de probabilidad de referencia P tal que $Q \ll P$, ambas definidas sobre $\mathbb{R}^d$, es definida como*

$$D_{\mathrm{kl}}(Q\,\|P) = \int_{\mathbb{R}^d} \frac{dQ}{dP}(x)\,\log\left(\frac{dQ}{dP}(x)\right)\,dP(x) = \int_{\mathbb{R}^d} \log\left(\frac{dQ}{dP}(x)\right)\,dQ(x).$$

Si P y Q admiten una densidad con respeto a una medida μ (basicamente nos interesamos a μ_L y $\mu_{\mathcal{X}}$), se escribe a través de las densidades como [98]

$$D_{\mathrm{kl}}(q\,\|p) \equiv D_{\mathrm{kl}}(Q\,\|P) = \int_{\mathbb{R}^d} \log\left(\frac{q(x)}{p(x)}\right)\,q(x)\,d\mu(x).$$

Inicialmente, esta medida fue introducida por Kullback y Leibler en la misma linea que Shannon, interpretando $\log\left(\frac{dQ}{dP}(x)\right)$ como una información de discriminación entre dos hipótesis de distribuciones Q y P a partir de la observación x, la divergencia siendo la información de discriminación promedia. Introdujeron también una versión simétrica, que veremos más adelante. Se notará que, $D_{\mathrm{kl}}(Q\,\|P) = -\int_{\mathbb{R}^d} q(x)\,\log(q(x))\,d\mu(x) + \int_{\mathbb{R}^d} p(x)\,\log(q(x))\,d\mu(x)$. El primer termino es nada mas que la entropía de q, que se puede ver como distribución "presupuesta". Menos el secundo termino se interpreta como el promedio de la incerteza elemental $\log(q)$ con respeto a la distribución de referencia ("verdadera"), a veces llamado *entropía cruzada*. Por eso, esta divergencia es una entropía relativamente a la distribución p. En la misma linea, se puede inmediatamente ver de la definición general Def. 2-62, que $D_{\mathrm{kl}}(Q\,\|P) \equiv -H\left(\frac{dQ}{dP}\right)$ con respeto a la medida $\mu = P$. Por ejemplo, en el caso discreto finito, si p es la distribución uniforme sobre un alfabeto de cardinal α, $D_{\mathrm{kl}}(q\,\|p) = \log\alpha - H(q)$, lo que representa una desviación de la entropía de su valor máximo. La misma interpretación queda en el caso continuo con la ley uniforme (p y q definidas sobre el mismo espacio de volumen finito) o con la gaussiana (p y q teniendo la misma matriz de covarianza). *Como para la entropía, cuando se necesitará un logaritmo especificamente de base a, se notará la divergencia $D_{\mathrm{kl},a}$.*

Lema 2-61 (Positividad de la entropía relativa)**.**

$$D_{\mathrm{kl}}(Q\,\|P) \geq 0 \quad \textit{con igualdad ssi} \quad P = Q.$$

---

[98] En el caso discreto, esta cantidad depende solamente de p y q y no de los estados. La condición necesaria es que p y q tienen los mismos números de componentes (se completa el vector lo más corto) y si la i-esima componente de q vale cero, entonces la de p vale cero también. Además, con p y q de mismo tamaño, se puede poner en biyección los alfabetos asociados a p y q, sin perdida de generalidad. En el caso continuo, esta razonamiento no vale más, esta cantidad dependiendo en general de los estados...

Demostración. Existen varias pruebas, pero la más linda puede ser la usando la desigualdad de Jensen [99], teorema 1-21: para ϕ convexa e Y variable aleatoria escalar, $\mathrm{E}[\phi(Y)] \geq \phi(\mathrm{E}[Y])$ con igualdad ssi Y es determinista (casi siempre) si ϕ es estrictamente convexa. Sea X de medida de probabilidad P. Se escribe la entropía relativa $D_{\mathrm{kl}}(Q\,\|P) = \mathrm{E}\left[\frac{dQ}{dP}(X)\log\left(\frac{dQ}{dP}(X)\right)\right]$. Sea $Y = \frac{dQ}{dP}(X)$ y $\phi(u) = u\log u$, función estrictamente convexa. Entonces $D_{\mathrm{kl}}(Q\,\|P) = \mathrm{E}[\phi(Y)] \geq \phi(\mathrm{E}[Y])$. Pero $\mathrm{E}[Y] = \mathrm{E}\left[\frac{dQ}{dP}(X)\right] = \int_{\mathbb{R}^d}\frac{dQ}{dP}(x)\,dP(x) = 1$ según el lema 1-2. Se cierra la prueba con el hecho que $\phi(1) = 0$ El caso de igualdad apareciendo si y solamente si Y es determinista, es decir $\frac{dP}{dQ}(X)$ determinista, es equivalente a $\frac{dP}{dQ} = 1 \quad (P$-c.s.) (constante, la constante siendo igual a 1 del lema 1-2 y P y Q siendo medidas de probabilidad). □

Esta propiedad tiene consecuencias fijandose que

$$I(X;Y) = D_{\mathrm{kl}}(P_{X,Y}\,\|P_X P_Y)\,,$$

i. e., la información mutua es la divergencia de Kullback-Leibler de la distribución conjunta relativa al producto de las marginales (obviamente $P_{X,Y} \ll P_X P_Y$). Con este enfoque, no se necesita que (X,Y) sea discreta o continua admitiendo una densidad.

[P15] *I es positiva, como medida de independencia:*

$$I(X;Y) \geq 0 \quad \text{con igualdad ssi } X \text{ e } Y \text{ son independientes.}$$

[P16] *Condicionar reduce la entropía*

$$H(X|Y) \leq H(X) \quad \text{con igualdad ssi } X \text{ e } Y \text{ son independientes.}$$

Esta desigualdad, con la regla de cadena, prueba la sub-aditividad [P11]. Esta reducción de incerteza vale en promedio, pero el conocimiento de un valor particular puede ser tal que $H(X|Y=y) > H(X)$, *i. e.,* ¡un conocimiento particular puede aumentar la entropía! (ver ejemplos en (Rioul, 2007, p. 59)).

Fijense que si D_{kl} es positiva, no es simétrica y tampoco satisface la desigualdad triangular. Por eso, no es una distancia y tiene el nombre de *divergencia*. La distribución de referencia P juega un rol fundamental.

Al final, se mencionará las propiedades adicionales siguientes:

1. $D_{\mathrm{kl}}(q\,\|p)$ queda invariante bajo una misma transformación biyectiva sobre ambos p y q. Es trivial en el caso discreto y si no se prueba sencillamente por un cambio de variables en la forma integral.

---

[99] En el caso discreto, se puede usar también la desigualdad $\sum t_i \log t_i \geq \sum t_i \log t'_i$ una instancia de la desigualdad conocida como desigualdad log-sum, o conocida también como desigualdad de Gibbs (debido a J. W. Gibbs mismo) (Gibbs, 1902; Cover & Thomas, 2006; Rioul, 2007; Merhav, 2010, 2018).

2. D_{kl} es convexa con respecto al par (P, Q) en el sentido que, para $\pi_i \geq 0$, $\sum_i \pi_i = 1$, y dos conjuntos $\{P_{(i)}\}_i$, $\{Q_{(i)}\}_i$ de medidas de probabilidades tales que $Q_{(i)} \ll P_{(i)}$,

$$D_{\mathrm{kl}}\left(\sum_i \pi_i\, Q_{(i)} \middle\| \sum_i \pi_i\, P_{(i)}\right) \geq \sum_i \pi_i D_{\mathrm{kl}}\left(Q_{(i)} \middle\| P_{(i)}\right).$$

La prueba de esta desigualdad es dada subsección **??** en un contexto más general.

3. Para Q fijo, $D_{\mathrm{kl}}\left(Q \middle\| P\right)$ es convexa con respecto a P en el sentido que, para $\pi_i \geq 0$, $\sum_i \pi_i = 1$, y un conjunto $\{P_{(i)}\}_i$ de medidas de probabilidades tales que $Q \ll P_{(i)}$,

$$D_{\mathrm{kl}}\left(Q \middle\| \sum_i \pi_i\, P_{(i)}\right) \geq \sum_i \pi_i\, D_{\mathrm{kl}}\left(Q \middle\| P_{(i)}\right).$$

Eso es la consecuencia obvia de la concavidad de $u \mapsto \log u$ (escribiendo la divergencia como densidad con respeto a una medida dada). Es sencillo ver que si D_{kl} siendo convexa con respecto a P (Q dada) y al par (P, Q), no puede ser convexa con respecto a Q (con P dada).

2.4 Unas identidades y desigualdades

Desigualdades de Fano? Rioul p. 78, Cover P. 663, Sanov? Pythagorean? Gene: cf Zyc p60

2.4.1 El principio de entropía máxima

En la termodinámica, el estudio de las características macroscópicas (dinámica de las moléculas) es prohibitivo tan el número de moléculas es importante. Por ejemplo, un litro del gas que respiramos contiene $2{,}7 \times 10^{22}$ moléculas. De esta constatación se desarolló la física estadísticas bajo el impulso de Boltzmann (Boltzmann, 1896, 1898), Maxwell (Maxwell, 1867), Gibbs (Gibbs, 1902), Planck (Planck, 2015) entre otros (ver también (Jaynes, 1965; Merhav, 2010, 2018, y ref.)), considerando el sistema macroscópico a través de lo que llamaron ensembles estadísticos: el sistema global (macroscópico) es al equilibrio pero las configuraciones (micro-estados) son fluctuantes. De una forma, se puede asociar a una configuración su frecuencia de ocurrencia (imaginando tener une infinidad de copias del sistema en el mismo estado macroscópico), es decir su probabilidad de occurencia. En este marco, la entropía, describiendo la falta de información, juega un rol fundamental. Un sistema sujeto a vínculos, como por ejemplo teniendo una energía dada, debe estar en sus estado lo más desorganizado dados los vínculos. En su marco, se introdujo la noción de entropía termodinámica, pero la misma es tremendamente vinculada a la entropía de Shannon [100] (claramente, identificando las frecuencias a probabilidades de

[100] Ver epígrafe del capítulo...

ocurrencia). En otro terminos, la distribución describiendo los micro-estados debe ser de entropía máxima, dados los vínculos. Por ejemplo, en un gas perfecto, donde las partículas no interactuan (aparte chocandose), la energía es dada por las velocidades (suma de las energías cinéticas individuales). Dada una energía fija, la distribución de las velocidad debe ser de entropía máxima sujeto a la energía dada (nada más que la energía va a "organizar" las configuraciones posibles). Intuitivamente, en un sistema aislado de N partículas, las configuraciones van a ser equiprobables, precisamente la distribución maximizando la entropía. **En la subsección Sec. 2.5.4 se va a desarollar un poco más este ejemplo.**

De manera general, el problema se formaliza como la búsqueda de la entropía máxima sujeto a vínculos. Si este principio nació en mecánica estadística (ver también (Jaynes, 1957a, 1957b, 1965; Merhav, 2010, 2018)), encontró un echo en varios campos: en inferencia bayesiana para elegir distribuciones del a priori [101] conociendo unos momentos de la ley (Robert, 2007; Jaynes, 1968, 1982; Csiszàr, 1991), hacer estimación espectral o de procesos estocasticos autoregresivos (Burg, 1967, 1975; Jaynes, 1982) o (Cover & Thomas, 2006, cap. 12), entre otros (Arndt, 2001; Kapur, 1989; Kapur & Kesavan, 1992, & ref.).

Sea X variable aleatoria viviendo sobre (de distribución de probabilidad de soporte) $X(\Omega) = \mathcal{X} \subseteq \mathbb{R}^d$ con $K \geq 0$ momentos $\mathrm{E}\left[M_k(X)\right] = m_k$ fijos, con $M_x : \mathcal{X} \to \mathbb{R}$. Suponemos de que X tiene una densidad p con respeto a una medida μ, basicamente μ_L en el contexto continuo y $\mu_{\mathcal{X}}$ en el caso discreto. El problema de entropía máxima se formula de la manera siguiente: sean $M(x) = \begin{bmatrix} M_1(x) & \cdots & M_K(x) \end{bmatrix}^t$ y $m = \begin{bmatrix} m_1 & \cdots & m_K \end{bmatrix}^t$ (si $K = 0$ M desaparece)

$$p^* = \underset{p}{\operatorname{argmáx}}\, H(p) \qquad \text{sujeto a} \qquad p \geq 0, \quad \int_{\mathcal{X}} p(x)\, d\mu(x) = 1 \quad \text{y} \quad \int_{\mathcal{X}} M(x)\, p(x)\, d\mu(x) = m,$$

donde los dos primeros vínculos (positividad, normalización) aseguran de que p^* sea una distribución de probabilidad. En el ejemplo del gas perfecto, $K = 1$, $M_1(x) = \sum_i x_i^2$ (los x_i son las velocidades). Introduciendo factores de Lagrange η_0, $\eta = \begin{bmatrix} \eta_1 & \cdots & \eta_K \end{bmatrix}^t$ para tener en cuenta los vínculos, el problema variacional consiste a resolver (Gelfand & Fomin, 1963; van Brunt, 2004; Miller, 2000; Cambini & Martein, 2009; Cover & Thomas, 2006)

$$p^* = \underset{p}{\operatorname{argmáx}} \int_{\mathcal{X}} \left(-p(x) \log p(x) + \eta_0\, p(x) + \eta^t M(x)\, p(x)\right) d\mu(x),$$

donde η_0, η serán determinados para satisfacer a los vínculos. En el caso continuo $\mu = \mu_L$ se usa la ecuación de Euler-Lagrange (Gelfand & Fomin, 1963; van Brunt, 2004), esquematicamente anulando la "derivada" del integrande con respeto a p; en el caso discreto, se anula realmente un gradiente con

[101] Ver nota de pie **??**. En la inferencia bayesiana de un parametro modelizado como aleatorio θ, para eligir la ley a priori $p_\Theta(\theta)$, si se conocen momentos por una razón o una otra, se puede elegir esta distribución como la "menos informativa" posible, *i. e.,* de entropía máxima dados los momentos.

respeto a los componentes de p. Reparametrizando los factores de Lagrange, se obtiene así

$$p^*(x) = \frac{1}{Z(\eta)} e^{\eta^t M(x)} = e^{\eta^t M(x) - \varphi(\eta)},$$

con η tal que se satisfacen los vínculos de momentos y $Z(\eta) = e^{\varphi(\eta)} = \int_{\mathcal{X}} e^{\eta^t M(x)} d\mu(x)$ coeficiente de normalización. Esta distribución cae precisamente en la familia exponencial que hemos visto sección 1.10.3, donde M es la estadística suficiente y η el parámetros naturale (en este caso, $h(x) = 1$).

Un problema que puede aparecer es que no se puede determinar η tal que se satisfacen todos los vínculos, en particular se puede que no se normaliza la ley. Por ejemplo, si $\mathcal{X} = \mathbb{R}$ y $K = 0,\ p$ debería ser constante (ley uniforme) sobre... $\mathbb{R}$, lo que no es normalizable. De la misma manera, si $K = 3$ y $M_k(x) = x^k$, tampoco es normalizable la función obtenida [102]. En otros terminos, en este caso, el problema no tiene solución [103].

Existe una prueba informacional de este resultado, saliendo de la solución.

Lema 2-62. *Sea* $\mathcal{P}_m = \left\{ p \geq 0 \;\middle|\; \int_{\mathcal{X}} p(x)\, d\mu(x) = 1, \quad \int_{\mathcal{X}} M(x)\, p(x)\, d\mu(x) = m \right\}$ *conjunto de densidades con respeto a* μ*, con los mismos momentos* M*, y sea* $p^* \in \mathcal{P}_m$ *que sea de la forma* $p^*(x) = \mathrm{e}^{\eta^t M(x) - \varphi(\eta)}$*. Entonces*

$$\forall\, p \in \mathcal{P}_m, \quad H(p) \leq H(p^*) \qquad \textit{con igualdad ssi} \quad p = p^* \quad \mu\textit{-(c. s.)},$$

donde H *es de la definición general Def. 2-62, pagina 220, de la entropía con respeto a* μ*.*

Demostración.

$$\begin{aligned} H(p) &= -\int_{\mathcal{X}} p(x)\, \log p(x)\, d\mu(x) \\ &= -\int_{\mathcal{X}} p(x)\, \log\left(\frac{p(x)}{p^*(x)}\right) d\mu(x) - \int_{\mathcal{X}} p(x)\, \log p^*(x)\, d\mu(x) \end{aligned}$$

De $\log p^* = \eta^t M - \varphi(\eta)$ se obtiene

$$\begin{aligned} H(p) &= -D_{\mathrm{kl}}\left(p \,\|p^*\right) - \int_{\mathcal{X}} \left(\eta^t M(x) - \varphi(\eta)\right) p(x)\, d\mu(x) \\ &= -D_{\mathrm{kl}}\left(p \,\|p^*\right) - \int_{\mathcal{X}} \left(\eta^t M(x) - \varphi(\eta)\right) p^*(x)\, d\mu(x) \\ &= -D_{\mathrm{kl}}\left(p \,\|p^*\right) - \int_{\mathcal{X}} p^*(x) \log p^*(x)\, d\mu(x) \\ &= -D_{\mathrm{kl}}\left(p \,\|p^*\right) + H(p^*) \end{aligned}$$

---

[102] En el enfoque Bayesiano se puede que no sea problemático, si el a posteriori es normalizable (Robert, 2007), pero va más allá de la meta de esta sección.

[103] Más precisamente, existen casos en los cuales se puede acotar la entropía por arriba por un $H^{\sup}$, tal que $\sup_p H(p) \leq H^{\sup}$ pero no se puede alcanzar esta cota, *i. e.,* es un supremum, no un máximo (Cover & Thomas, 2006, sec. 12.3).

porque $p, p^* \in \mathcal{P}_m$ (segunda linea) y $\eta^t M - \varphi(\eta) = \log p^*$ (tercera linea). La prueba se cierra notando que $D_{\mathrm{kl}}(p \,\|p^*) \geq 0$ con igualdad si y solamente si $p = p^*$ (μ-c.s.). □

Este lema prueba que, dados vínculos "razonables", la entropía es acotada por arriba, y que se alcanza la cota para una distribución de la familia exponencial. Por ejemplo,

- Con $K = 0$ y $\mathcal{X}$ de volumen finito $|\mathcal{X}| < +\infty$, la distribución de entropía máxima es la distribución uniforme de la propiedad [P'5]*a* de la subsección Sec. 2.2.2 en el caso continuo, o propiedad [P5] de la sección Sec. 2.2.1 en el caso discreto.

- Con $K = 1$, $\mathcal{X} = \mathbb{R}^d$ y $M(x) = xx^t$ (visto como $K = d^2$ momentos), la distribución de entropía máxima es la distribución gausiana de la propiedad [P'5]*b* de la sección Sec. 2.2.2.

Con el enfoque del lema 2-62, se necesita solamente que p sea una densidad con respeto a una medida μ fija, cualquiera. En particular, si es una medida de probabilidad (de referencia) $\widetilde{P}$, el problema de entropía máxima vuelve ser un problema de minimización de la divergencia de Kullback-Leibler entre P y la medida de referencia, siendo p la densidad con respeto a esta medida (ver definición Def. 2-65). Es decir, tomando $\mu = \widetilde{P}$ medida de probabilidad aparece inmediatamente

$$P^* = \operatorname*{argmáx}_P D_{\mathrm{kl}}\left(P \,\middle\|\, \widetilde{P}\right) \quad \text{sujeto a} \int_{\mathcal{X}} M(x)\, dP(x) = m \qquad \Leftrightarrow \qquad \frac{dP^*}{d\widetilde{P}}(x) = \mathrm{e}^{\eta^t x - \varphi(\eta)}\,.$$

2.4.2 Desigualdad de la potencia entrópica

Sean X e Y dos variables independientes. Si se conoce las relaciones vinculando $H(X,Y)$, $H(X)$, $H(Y)$, una pregunta natural concierna la relación que podría tener $X+Y$ con cada variable en término de entropía. La respuesta no es trivial, y el resultado general concierna el caso de variables continuas sobre $\mathbb{R}^d$. Es conocido como desigualdad de la potencia entrópica (EPI para entropy power inequality en inglés). No vincula las entropías, sino que las potencias entrópicas.

Teorema 2-69 (Desigualdad de la potencia entrópica)**.** *Sean X e Y dos variables d-dimensionales continuas independientes. Entonces*

$$N(X+Y) \geq N(X) + N(Y),$$

con igualdad si y solamente si X e Y son gaussianas con matrices de covarianza proporcionales, $\Sigma_Y \propto \Sigma_X$.

Existen varias formulaciones alternativas a esta desiguladad (Shannon, 1948; Lieb, 1978; Cover & Thomas, 2006; Dembo et al., 1991; Rioul, 2007):

1. Sean $\widetilde{X}$ y $\widetilde{Y}$ gaussianas independientes de matrices de covarianza proporcionales y tal que $H(\widetilde{X}) = H(X)$ y $H(\widetilde{Y}) = H(Y)$. Entonces

$$N(X+Y) \geq N\left(\widetilde{X}+\widetilde{Y}\right),$$

con igualdad si y solamente si X y Y son gaussianas. De hecho, la primera formulación es equivalente a $N(X+Y) \geq N\left(\widetilde{X}\right) + N\left(\widetilde{Y}\right) = \frac{1}{2\pi\,\mathrm{e}}\left(\left|\Sigma_{\widetilde{X}}\right|^{\frac{1}{d}} + \left|\Sigma_{\widetilde{Y}}\right|^{\frac{1}{d}}\right) \geq \frac{1}{2\pi\,\mathrm{e}}\left|\Sigma_{\widetilde{X}} + \Sigma_{\widetilde{Y}}\right|^{\frac{1}{d}} = N\left(\widetilde{X}+\widetilde{Y}\right)$ (la última desigualdad viniendo de la desigualdad matricial de Minkowski (Hardy et al., 1952; Minkowski, 1910)). Se notará que, de la relación uno-uno entre H y N la desigualdad se escribe también

$$H(X+Y) \geq H\left(\widetilde{X}+\widetilde{Y}\right).$$

2. *Desigualdad de preservación de covarianza:*

$$\forall\, 0 \leq a \leq 1, \quad H\left(\sqrt{a}\,X + \sqrt{1-a}\,Y\right) \geq a\,H(X) + (1-a)\,H(Y),$$

con igualdad si y solmente si X e Y son gaussianas con matrices de covarianza proporcionales. Claramente, se cumple la igualdad para $\widetilde{X}$ e $\widetilde{X}$, entonces $H\left(\sqrt{a}\,X + \sqrt{1-a}\,Y\right) \geq a\,H(X) + (1-a)\,H(Y) \quad \Leftrightarrow \quad H\left(\sqrt{a}\,X + \sqrt{1-a}\,Y\right) \geq H\left(\sqrt{a}\,\widetilde{X} + \sqrt{1-a}\,\widetilde{Y}\right)$ lo que es nada más que la desigualdad anterior reemplazando X por $\sqrt{a}\,X$ e Y por $\sqrt{1-a}\,Y$ (y vice-versa).

La prueba de esta(s) desigualdad(es) no es trivial. Númeras versiones existen, dadas por ejemplo en las referencias (Blachman, 1965; Stam, 1959; Shannon & Weaver, 1964; Rioul, 2007, 2011, 2017; Cover & Thomas, 2006; Dembo et al., 1991; Lieb, 1978; Verdú & Guo, 2006) (ver tambien teorema 6 de (Lieb, 1975)). Como se lo puede ver, la gaussiana juega un rol particular en esta desigualdad, saturandola.

Una gracia de la desigualdad de la potencia entrópica es que puede dar lugar a pruebas informacionales de desigualdades matriciales, como por ejemplo la desigualdad de Minkowsky de los determinentes $|R_1 + R_2|^{\frac{1}{d}} \geq |R_1|^{\frac{1}{d}} + |R_2|^{\frac{1}{d}}$ para cualquieras matrices R_1, R_2 simétricas definidas positivas, con igualdad si y solamente si $R_2 \propto R_1$ (viene de X e Y gaussianas de covarianza R_1 y R_2). Aparece también para acotar la información mutua entre variables y calcular la capacidad de un canal de comunicación como se lo va a ver más adelante (Cover & Thomas, 2006; Dembo et al., 1991; Rioul, 2007; Johnson, 2004).

Se mencionará que existe una versión de la desigualdad de la potencia entrópica con rearreglo (Wang & Madiman, 2004):

Teorema 2-70 (Desigualdad de la potencia entrópica con rearreglo)**.** *Sean X e Y dos variables d-dimensionales continuas independientes de densidades de probabilidades p_X y p_Y respectivamente. Sean $p_X^{\downarrow}$ y $p_Y^{\downarrow}$ los rearreglos de p_X y p_Y respectivamente y denotamos $X^{\downarrow}$ y $Y^{\downarrow}$ vectores independientes de distribución de probabilidad $p_X^{\downarrow}$ y $p_Y^{\downarrow}$ respectivamente. Enconces, Entonces*

$$N(X+Y) \geq N(X^{\downarrow} + Y^{\downarrow}).$$

Se referirá a (Madiman & Barron, 2007, y Ref.) por ejemplo para varias generalizaciones de la desigualdad de la potencia entrópica.

Para cerrar esta secciíon, se mencionará de que en el caso discreto, no hay un resultado general y aún existen contra-ejemplos (Johnson & Yu, 2010, Sec. IV). Existen solamente resultados para variables particulares como para variables binarias (Shamai & Wyner, 1990), leyes binomiales (Harremoës & Vignat, 2003; Sharma, Das & Muthukrishnan, 2011) (ver también (Johnson & Yu, 2010; Haghighatshoar, Abbe & Telatar, 2014)).

2.4.3 Desigualdad de procesamiento de datos

Esta desigualdad traduce que procesando datos, no se puede aumentar la información disponible sobre una variable. Se basa sobre una desigualdad que satisface la información mutua aplicada a un proceso de Markov.

Definición 2-66 (Proceso de Markov)**.** *Una secuencia $X_1 \mapsto X_2 \mapsto \ldots \mapsto X_n$ es dicha de Markov si para cualquier $i > 1$,*

$$\forall\, x_i, \quad P_{X_{i-1},X_{i+1}|X_i=x_i} = P_{X_{i-1}|X_i=x_i}\, P_{x_{i+1}|X_i=x_i}.$$

Dicho de otra manera, condicionalmente a $(X_i = x)$, las variables X_{i-1} y X_{i+1} son independientes. Eso es equivalente a

$$P_{X_{i+1}|(X_i,X_{i-1},\ldots)=(x_i,x_{i-1},\ldots)} = P_{X_{i+1}|X_i=x_i}.$$

Si i representa un tiempo, significa que la estadística de X_{i+1} conociendo todo el pasado se reduce a esa conociendo el pasado inmediato (las probabilidades dichas de transición $P_{X_{i+1}|X_i=x_i}$ y la distribución inical P_{X_1} caracterizan completamente el proceso). Es sencillo fijarse de que $X_n \mapsto X_{n-1} \mapsto \ldots \mapsto X_1$ es tambien un proceso de Markov.

Teorema 2-71 (Desigualdad de procesamiento de datos)**.** *Sea $X \mapsto Y \mapsto Z$ un proceso de Markov. Entonces,*

$$I(X;Y) \geq I(X;Z),$$

con igualdad si y solamente si $X \mapsto Z \mapsto Y$ es también un proceso de Markov. En particular, es sencillo ver que para cualquiera función g, $X \mapsto Y \mapsto g(Y)$ es un proceso de Markov, lo que da

$$\forall\, g, \quad I(X;Y) \geq I(X;g(Y)).$$

La última desigualdad se escribe también $H(X|g(Y)) \geq H(X|Y)$ y significa que procesar Y no aumenta la información que Y da sobre X (la incerteza condicional es más importante).

Demostración. Por definición de la información mutua, considerando X y la variable conjunta (Y, Z),

$$\begin{aligned} I(X;Y,Z) &= H(X) - H(X|Y,Z) \\ &= H(X) - H(X|Y) + H(X|Y) - H(X|Y,Z) \end{aligned}$$

Por la propiedad de que $Z \mapsto Y \mapsto X$ sea también un proceso de Markov, es sencillo probar que $H(X|Y,Z) = H(X|Y)$ (conciendo Y sufice para caracterizar completamente X), lo que da

$$I(X;Y,Z) = I(X;Y).$$

También,

$$\begin{aligned} I(X;Y,Z) &= H(X) - H(X|Z) + H(X|Z) - H(X|Y,Z) \\ &= I(X;Z) + H(X|Z) - H(X|Y,Z) \end{aligned}$$

Además, escribiendo $\frac{p_{X|(Y,Z)=(y,z)}(x)}{p_{X|Z=z}(x)} = \frac{p_{X|(Y,Z)=(y,z)}(x)\, p_{Y|Z=z}(y)}{p_{X|Z=z}(x) p_{Y|Z=z}(y)} = \frac{p_{X,Y|Z=z}(x,y)}{p_{X|Z=z}(x) p_{Y|Z=z}(y)}$ se nota de que $H(X|Z) - H(X|Y,Z)$ es la divergencia de Kullback-Leibler de $p_{X,Y|Z=z}$ relativamente a $p_{X|Z=z} p_{Y|Z=z}$, o información mutua $I(X;Y|Z)$ entre X e Y, condicionalmente a Z. Entonces, de las dos formas de $H(X;Y,Z)$ viene

$$I(X;Y) = I(X;Z) + I(X;Y|Z).$$

La desigualdad del teorema viene de la positividad de $I(X;Y|Z)$. Además, se obtiene la igualdad si y solamente si $I(X;Y|Z) = 0$, es decir X e Y independientes condicionalmente a Z, lo que es la definición de que $X \mapsto Z \mapsto Y$ sea un proceso de Markov. □

2.4.4 Segunda ley de la termodinámica

Tratando de procesos de Markov, aparece el equivalente de la segunda ley de la termodinámica: un sistema aislado evolua hasta llegar su estado lo más desorganizado (ver ej. (Cover & Thomas, 2006; Merhav, 2010, 2018, y ref.)).

Lema 2-63 (Versión informacional de la segunda ley de la termodinámica)**.** *Sea $X_1 \mapsto X_2 \mapsto \cdots \mapsto X_n \mapsto \cdots$ un proceso de Markov, con probabilidades de transición $r_{X_{n+1}|X_n=x_n}$ dadas (independientemente de la condición inicial). Estas últimas modelizan el sistema, independiente de las condiciones iniciales. Sean dos distribuciones (condiciones) iniciales diferentes p_{X_1} y q_{X_1}, conduciendo a las distribuciones p_{X_n} y q_{X_n} (con respeto a una medida μ dada) para X_n. Entonces:*

- *Para cualquier $n \geq 1$,*

$$D_{\mathrm{kl}}\left(p_{X_{n+1}} \middle\| q_{X_{n+1}}\right) \leq D_{\mathrm{kl}}\left(p_{X_n} \| q_{X_n}\right);$$

  las distribuciones p_{X_n} y q_{X_n} no se "alejan" (tiende a acercarse);

- *Si p^* es una distribución estacionaria,*

$$D_{\mathrm{kl}}\left(p_{X_{n+1}}\,\|p^*\right) \leq D_{\mathrm{kl}}\left(p_{X_n}\,\|p^*\right);$$

la distribución no se aleja de la distribución estacionaria.

- *Además, si los X_n tienen K momentos fijos $m = \mathrm{E}[M(X_n)] \quad \forall\, n$ y si p^* es la densidad (con respeto a la medida μ dada) de entropía máxima tiendo los mismos momentos como descrito subsección Sec. 2.4.1, (ej. $K = 0$, $\mathcal{X}$ de cardinal o volumen finito y ley uniforme, $K = 2$, $M_k(x) = x^k$ y ley gausiana),*

$$H(X_{n+1}) \geq H(X_n);$$

el sistema tiende a desorganizarse (dando los vinculos/momentos).

Demostración. Se muestra sencillamente que $D_{\mathrm{kl}}\left(p_{X_{n+1},X_n}\,\|q_{X_{n+1},X_n}\right) = D_{\mathrm{kl}}\left(p_{X_{n+1}}\,\|q_{X_{n+1}}\right) + \int_{\mathcal{X}_{n+1}} D_{\mathrm{kl}}\left(p_{X_n|X_{n+1}=x_{n+1}}\,\|q_{X_n|X_{n+1}=x_{n+1}}\right) d\mu(x_{n+1}) = D_{\mathrm{kl}}\left(p_{X_n}\,\|q_{X_n}\right) + \int_{\mathcal{X}_n} D_{\mathrm{kl}}\left(p_{X_{n+1}|X_n=x_n}\,\|q_{X_{n+1}|X_n=x_n}\right) d\mu(x_n)$. Además, $p_{X_{n+1}|X_n=x_n} = r_{X_{n+1}|X_n=x_n} = q_{X_{n+1}|X_n=x_n}$ (transición independiente de la condición inicial), conduciendo a $D_{\mathrm{kl}}\left(p_{X_{n+1}|X_n=x_n}\,\|q_{X_{n+1}|X_n=x_n}\right) = 0$ con consecuencia de que $D_{\mathrm{kl}}\left(p_{X_n}\,\|q_{X_n}\right) = D_{\mathrm{kl}}\left(p_{X_{n+1}}\,\|q_{X_{n+1}}\right) + \int_{\mathcal{X}_{n+1}} D_{\mathrm{kl}}\left(p_{X_n|X_{n+1}=x_{n+1}}\,\|q_{X_n|X_{n+1}=x_{n+1}}\right) d\mu(x_{n+1})$. $p_{X_n|X_{n+1}=x_{n+1}}$ no es necesariamente igual a $q_{X_n|X_{n+1}=x_{n+1}}$, pero la divergencia siendo no negativa, se obtiene la primera desigualdad. La segunda desigualdad se obtiene tomando $q_{X_n} = p^*$. Además, si p^* es le entropía máxima de mismos momentos $m = \mathrm{E}[M(X_n)]$ que los X_n, hemos visto de que $p^*(x) = \mathrm{e}^{\eta^t M(x)}$ ley de la familia exponencial, dando $D_{\mathrm{kl}}\left(p_{X_n}\,\|p^*\right) = -H(X_n) - \eta^t m$, conduciendo a la última desigualdad. □

2.4.5 Principio de incerteza entrópico

Bourret 58, Leipnik 59, Stam59, entre otros que ya citamos un par de veces

2.4.6 Un foco sobre la información de Fisher

Si la entropía y las heramientas relacionadas son naturales como medidas de información, no se puede resumir una distribución a una medida escalar. En el marco de la teoría de la estimación, R. Fisher introdujo una noción de información intimamente relacionada al error cuadrático en la estimación de un parámetro a partir de una variable parametrizado por este parámetro (Fisher, 1922, 1925b; Kay, 1993; van den Bos, 2007; Cover & Thomas, 2006; Frieden, 2004).

Definición 2-67 (Matriz información de Fisher parámetrica)**.** *Sea X una variable aleatoria parametrizada por un parámetro m-dimensional, $\theta \in \Theta \subseteq \mathbb{R}^m$, de distribución de probabilidad $p_X(\cdot;\theta)$ (con respeto a una medida μ dada) sobre $\mathcal{X} \subseteq \mathbb{R}^d$ su soporte. Suponga que p_X sea diferenciable con respeto a θ sobre Θ. La matriz de Fisher, de tamaño $m \times m$, es definida por*

$$J_\theta(X) = \mathrm{E}\left[\Big(\nabla_\theta \log p_X(X;\theta)\Big)\Big(\nabla_\theta \log p_X(X;\theta)\Big)^t\right],$$

donde $\nabla_\theta = \left[\cdots \frac{\partial}{\partial\theta_i} \cdots\right]^t$ es el gradiente en θ y $\log$ el logaritmo natural. Es la matriz de covarianza del score paramétrico $S_\theta(X) = \nabla_\theta \log p_X(X;\theta)$ notando que su media es igual a cero (escribiando el promedio y intercambiando integral y gradiente), siendo $\log p_X$ la log-verosimilitud. Bajo condiciones de regularidad, se puede mostrar [104] *que $J_\theta(X) = -\mathrm{E}[\mathcal{H}_\theta \log p_X(X;\theta)]$ con $\mathcal{H}_\theta$ la Hessiana* [105] *$\mathcal{H}_\theta$ de $\log p_X(X;\theta)$. Nota: a veces se define la información de Fisher como $\mathrm{Tr}(J)$, traza de la matriz información de Fisher.*

Como para la entropía, la matriz de Fisher se escribe generalmente $J_\theta(X)$, a pesar de que no sea función de X pero de la densidad de probabilidad. Se la notará también $J_\theta(p_X)$ según la escritura la más conveniente.

Notar que para une distribución de la familia exponencial natural vista sección 1.10.3, $p_X(x;\eta) = \exp\left(\eta^t S(x) - \varphi(\eta)\right)$ tenemos $S_\eta(x) = \nabla_\eta \log p_X(x;\eta) = S(x) - \nabla\varphi(\eta) = S(x) - \mathrm{E}[S(X)]$ así que

$$J_\eta(X) = \mathrm{Cov}[S(X)] = \mathcal{H}\varphi(\eta)$$

(ver por ejemplo (Lehmann & Casella, 1998; van den Bos, 2007)).

En el caso continuo, $\mu = \mu_L$, con una densidad diferenciable, tomando el gradiente en x en lugar de θ da la matriz de información de Fisher no paramétrica,

Definición 2-68 (Matriz información de Fisher no paramétrica)**.** *Sea X una variable aleatoria continua admitiendo una densidad de probabilidad p_X definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$ su soporte. Suponga que p_X sea diferenciable con respeto a x. La matriz de Fisher no paramétrica, $d \times d$, es definida por*

$$J(X) = \mathrm{E}\left[\Big(\nabla_x \log p_X(X)\Big)\Big(\nabla_x \log p_X(X)\Big)^t\right].$$

Es la matriz de covarianza de la función score $\nabla_x \log p_X(X)$ (escribiendo la media y p_X siendo cero el los bordes de $\mathcal{X}$, se ve que el promedio de $\nabla_x \log p_X(X)$ también vale cero) o, bajo condiciones de regularidad, $J(X) = -\mathrm{E}[\mathcal{H}_x \log p_X(X;\theta)]$.

Es interesante notar que:

---

[104] Es una consecuencia del teorema de la divergencia, suponiendo que los bordes del soporte $\mathcal{X}$ no dependen de θ y que la función score se cancela en estos bordes.

[105] Recordamos de que, para $f : \mathbb{R}^m \mapsto \mathbb{R}$, $\mathcal{H}_\theta f$ es la matriz de componentes $\frac{\partial^2 f}{\partial\theta_i \partial\theta_j}$.

- Cuando θ es un parámetro de posición, $p_X(x;\theta) = p(x-\theta)$, $\nabla_\theta \log p_X = -\nabla_x \log p_X$ tal que la información paramétrica se reduce a la información no paramétrica.

- Si X es gaussiano de matriz de covarianza Σ_X, entonces se muestra sencillamente de que $J(X) = \Sigma_X^{-1}$ (o, de una forma, inversa de la dispersión o incerteza en término de estadísticas de orden 2).

- Es sencillo ver que, por definición $J_\theta(X)$ y $J(X)$ son simétricas y que $J_\theta(X) > 0$ y $J(X) > 0$ (matrices definidas positivas). Además,

$$\forall\ A \text{ matrix no singular}, \quad J(AX) = A^{-t} J(X) A^{-1},$$

con $A^{-t} = \left(A^{-1}\right)^t = (A^t)^{-1}$ (Cover & Thomas, 2006; Dembo et al., 1991; Barron, 1986). Esta relación da a $J(X)$ un sabor de información en el sentido de que, cuando A es real y tiende al infinito, $J(AX)$ tiende a 0; AX tiende a ser muy dispersada así que no hay información sobre su posición.

- J_θ y J son convexas en el sentido de que para cualquier conjunto de $\pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$ y cualquier conjunto de distribuciones $p_{(k)},\ k = 1, \ldots, K$ (Cohen, 1968; Frieden, 2004),

$$J_\theta\left(\sum_{k=1}^K \pi_k p_{(k)}\right) < \sum_{k=1}^K \pi_k\, J_\theta\left(p_{(k)}\right) \qquad \text{y} \qquad J\left(\sum_{k=1}^K \pi_k p_{(k)}\right) < \sum_{k=1}^K \pi_k J\left(p_{(k)}\right),$$

donde $A < B$ significa que $B - A > 0$. La prueba es dada por Cohen en el caso escalar, pero se extiende sin costo adicional en el caso multivariado. Hace falta probarlo para $K = 2$ y, por recurrencia, se extiende para cualquier K. En este caso, observando que $\left(\nabla \log p\right)\left(\nabla \log p\right)^t p = \frac{\left(\nabla p\right)\left(\nabla p\right)^t}{p}$, considerando el gradiente con respeto a θ (resp. a x) tratando de J_θ (resp. J), se obtiene $\sum_k \pi_k \frac{\left(\nabla p_{(k)}\right)\left(\nabla p_{(k)}\right)^t}{p_{(k)}} - \frac{\left(\nabla \sum_k \pi_k p_{(k)}\right)\left(\nabla \sum_k \pi_k p_{(k)}\right)^t}{\sum_k \pi_k p_{(k)}} = \frac{1}{\sum_k \pi_k p_{(k)}} \sum_{k,l} \pi_k \pi_l \left(\frac{p_l}{p_{(k)}}\left(\nabla p_{(k)}\right)\left(\nabla p_{(k)}\right)^t - \left(\nabla p_{(k)}\right)\left(\nabla p_l\right)^t\right)$, lo que vale, tratando del caso $K = 2$, $\frac{\pi_1 \pi_2}{p_{(2)} p_{(2)} \left(\pi_1 p_{(1)} + \pi_2 p_{(2)}\right)}\left(p_{(2)} \nabla p_{(1)} - p_{(1)} \nabla p_{(2)}\right)\left(p_{(2)} \nabla p_{(1)} - p_{(1)} \nabla p_{(2)}\right)^t \geq 0$. No puede ser idénticamente cero (salvo si $\pi_1 \pi_2 = 0$ o $p_{(1)} = p_{(2)} \ldots$) así que se obtiene la desigualdad sobre la matriz de Fisher integrando esta última desigualdad.

2.4.6.1. Desigualdad de Cramér-Rao

Una otra interpretación de J como información es debido a la desigualdad de Cramér-Rao que la relaciona a la covarianza de estimación [106] (Rao, 1945, 1992; Rao & Wishart, 1947; Cramér, 1946;

[106] De hecho, pareció esta formula también en los papeles de Fréchet y de Darmois (Fréchet, 1943; Darmois, 1945). Como citado por Fréchet, aparece que la primera versión de esta formula es mucho más vieja y debido a K. Pearson & L. N. G Filon (Pearson & Filon, 1898) en 1898; luego fue extendida por Edgeworth (Edgeworth, 1908), Fisher (Fisher, 1925b) o Doob (Doob, 1936).

Rioul, 2007; Cover & Thomas, 2006; Frieden, 2004; Kay, 1993; van den Bos, 2007). Sea X parametrizada por θ. La meta es estimar θ a partir de X. Tal estimador va a ser una función únicamente de X, lo que se escribe usualmente [107] $\widehat{\theta}(X)$ (la función no depende explícitamente de θ). Las características de la calidad de un estimator es naturalmente su sesgo $b(\theta) = \mathrm{E}\left[\widehat{\theta}(X)\right] - \theta$ y su matriz de covarianza $\Sigma_{\widehat{\theta}}$ (la varianza da la dispersión alrededor de su promedio). La desigualdad de Cramér-Rao acota por debajo esta covarianza.

Teorema 2-72 (Desigualdad de Cramér-Rao)**.** *Sea X parametrizada por θ, de densidad de soporte $\mathcal{X} \subseteq \mathbb{R}^d$ independiente de θ, y $\widehat{\theta}(X)$ un estimador de θ. Sea $b(\theta)$ su sesgo y $\Sigma_{\widehat{\theta}}$ su matriz de covarianza. Sea $\mathrm{J}_b(\theta)$ la matriz Jacobiana del sesgo b. Entonces,*

$$\Sigma_{\widehat{\theta}} - (I + \mathrm{J}_b(\theta))\, J_\theta(X)^{-1}\, (I + \mathrm{J}_b(\theta))^t \geq 0.$$

En particular, en el caso θ escalar,

$$\sigma^2_{\widehat{\theta}} \geq \frac{(1 + b'(\theta))^2}{J_\theta(X)},$$

donde b' es la derivada de b.

Tomando θ parámetro de posición y $\widehat{\theta} = X$, estimador sin sesgo ($b = 0$), da lo que es conocido como la desigualdad no paramétrica de Cramér-Rao y toma la expresión

$$\Sigma_X - J(X)^{-1} \geq 0,$$

o, en el caso escalar,

$$\sigma^2_X \geq \frac{1}{J(X)}.$$

Además, en el caso no paramétrico, se alcanza la cota si y solamente si X es un vector gaussiano.

Esta desigualdad acota la varianza de cualquier estimador, *i. e.,* da la varianza o error mínimo(a) que se puede esperar. Esta cota es el inverso de la información de Fisher, *i. e.,* $J_\theta(X)$ caracteriza la información que X tiene sobre θ.

Demostración. Sea $S_\theta = \nabla_\theta \log p_X$ y $\theta_0 = \mathrm{E}\left[\widehat{\theta}(X)\right] = \theta + b(\theta)$. Fijandose que $p_X\, \nabla_\theta \log p_X = \nabla_\theta p_X$, que $\widehat{\theta}$ no es función de θ, y que el soporte $\mathcal{X}$ no depende de θ, se obtiene [108]

$$\begin{aligned}
\mathrm{E}\left[S_\theta(X)\left(\widehat{\theta}(X) - \theta_0\right)^t\right] &= \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\widehat{\theta}(x)^t\, dx - \left(\int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\, dx\right)\theta_0^t \\
&= \nabla_\theta \int_{\mathbb{R}^d} p_X(x;\theta)\widehat{\theta}(x)^t\, dx - \left(\nabla_\theta \int_{\mathbb{R}^d} p_X(x;\theta)\, dx\right)\theta_0^t \\
&= \nabla_\theta\left(\theta + b(\theta)\right) - \left(\nabla_\theta 1\right)\theta_0^t \\
&= (I + \mathrm{J}_b(\theta))^t
\end{aligned}$$

[107] Por ejemplo, si θ es un promedio común a los componentes de X, un estimador podría ser $\widehat{\theta} = \frac{1}{d}\sum_i X_i$.

[108] Se supone que los integrandes sean θ-localmente integrables, tal que se puede invertir derivada en θ e integración; ver también teorema 1-7, pagina 41.

Además, fijandose que $\mathrm{E}\left[S_\theta(X)S_\theta(X)^t\right] = J_\theta(X)$ y $\mathrm{E}\left[\left(\widehat{\theta}(X)-\theta_0\right)\left(\widehat{\theta}(X)-\theta_0\right)^t\right] = \Sigma_{\widehat{\theta}}$, la desigualdad de Cauchy-Bunyakovsky-Schwarz (ver corolario **??**, pagina **??**) conduce a

$$\left(u^t\left(I+\mathrm{J}_b(\theta)\right)^t v\right)^2 \;=\; \mathrm{E}\left[u^t S_\theta(X)\left(\widehat{\theta}(X)-\theta_0\right)^t v\right]^2 \;\leq\; u^t J_\theta(X) u\, v^t \Sigma_{\widehat{\theta}} v.$$

La prueba se termina tomando $u = J_\theta(X)^{-1}\left(I+\mathrm{J}_b(\theta)\right)^t\, v$ (recordandose que J_θ es simétrica).
Para θ parametro de posición, $\widehat{\Theta} = X$, con la elección de u, en la desigualdad de Cauchy-Bunyakovsky-Schwarz, se obtiene la igualdad cuando $v^t J(X)^{-1} S_\theta(x) \propto v^t(x-\theta)$ para cualquier v y x (c.s.), est decir $\nabla_x p_X(x) \propto J(X)(x-\theta)p_X(x)$, lo que es la ecuación diferencial que satisface (solamente) la gaussiana: en este caso, se verifica a posteriori que $J(X) = \Sigma_X^{-1}$, y entonces que se alcanza la cota de la desigualdad de Cramér-Rao no paramétrica. □

En el caso paramétrico, no se puede estudiar el caso de igualdad del hecho de que $\widehat{\Theta}$ no es algo dado. Además, aún dado un estimador (explícitamente independiente de θ), no hay garantía de que existe una densidad parametrizada por θ que alcanza la cota, o al revés, dada una familia de densidades, tampoco hay garantía que existe un estimador que permite alcanzar la cota (Cover & Thomas, 2006; Kay, 1993; van den Bos, 2007).

Fijense de que, nuevamente, la gaussiana juega un rol particular en la desigualdad de Cramér-Rao no paramétrica, permitiendo de alcanzar la cota.

Nota: para dos matrices $A \geq 0$ y $B \geq 0$, si $A - B \geq 0$ entonces $|A| \geq |B|$, con igualdad si y solamente si $A = B$ (Magnus & Neudecker, 1999, cap. 1, teorema 25). Entonces, de las desigualdades de Cramér-Rao se deducen desigualdades de Cramér-Rao escalares

$$\left|\Sigma_{\widehat{\theta}}\right| \;\geq\; \frac{\left|I+\mathrm{J}_b(\theta)\right|^2}{|J_\theta(X)|} \qquad \text{y} \qquad |\Sigma_X| \;\geq\; \frac{1}{|J(X)|}.$$

Obviamente, en la segunda, se alcanza la igualdad si y solamente si X es gaussiano. Además, para una matriz $A \geq 0$, existe la "relación determinente-traza" $|A|^{\frac{1}{d}} \leq \frac{1}{d}\operatorname{Tr}(A)$, con igualdad si y solamente si $A = I$ (Magnus & Neudecker, 1999, cap. 11, sec. 4), dando otras versiones escalares de la desigualdad de Cramér-Rao, por ejemplo, de la versión no paramétrica,

$$|\Sigma_X|^{\frac{1}{d}} \;\geq\; \frac{d}{\operatorname{Tr}\left(J(X)\right)}, \qquad \operatorname{Tr}\left(\Sigma_X\right) \;\geq\; \frac{d}{|J(X)|^{\frac{1}{d}}} \qquad \text{o} \qquad \operatorname{Tr}\left(\Sigma_X\right) \;\geq\; \frac{d^2}{\operatorname{Tr}\left(J(X)\right)}.$$

En estos casos, se obtiene la igualdad si y solamente si X es gaussiano (igualdad de la Cramér-Rao no paramétrica) y además de covarianza proporcional a la identidad (igualdad en la relación determinente-traza).

Se notará que, al imagen de las leyes de entropía máxima, la información de Fisher juega también un rol particular en la inferencia bayesiana a través del prior de Jeffrey (Jeffrey, 1946, 1948; Lehmann & Casella, 1998; Robert, 2007) [109].

---

[109] Ver nota de pie **??** pagina **??**. A veces, se toma como distribución a priori $p_\Theta(\theta) \propto |J_\theta(X)|^{\frac{1}{2}}$ por su invarianza por

2.4.6.2. Fisher como curvatura de la entropía relativa

Si la desigualdad de Cramér-Rao da a la matriz de Fisher un sabor de información, aparece que J_θ es también relacionada a la entropía relativa (Cover & Thomas, 2006; Frieden, 2004):

Teorema 2-73 (Fisher como curvatura de la entropía relativa)**.** *Sea X parametrizado por $\theta_0 \in \mathring{\Theta}$ interior de Θ (Θ contiene un vecinaje de θ_0). Siendo $D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right)$ función de $\theta \in \Theta$, aparece que*

$$D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right) = \frac{1}{2}\left(\theta-\theta_0\right)^t J_{\theta_0}(X)\left(\theta-\theta_0\right) + o\left(|\theta-\theta_0|\right),$$

donde $o(\cdot)$ es un resto pequeño con respecto a su argumento. En otros términos, $J_{\theta_0}(X)$ es la curvatura de la entropía relativa en θ_0.

Demostración. La relación es consecuencia de un desarrollo de Taylor al orden 2 de la función $D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right)$ de θ, tomada en $\theta=\theta_0$. Por propiedad de D_{kl}, la divergencia es positiva y se cancela cuando $\theta=\theta_0$. Entonces, el primer término del desarrollo vale cero y el segundo también, D_{kl} siendo mínima en $\theta=\theta_0$. Además,

$$\begin{aligned}
\nabla_\theta D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right) &= \nabla_\theta \int_{\mathcal{X}} p_X(x;\theta)\log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right)dx \\
&= \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right)dx + \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\,dx \\
&= \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right)dx + \nabla_\theta \int_{\mathcal{X}} p_X(x;\theta)\,dx \\
&= \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right)dx
\end{aligned}$$

la último ecuación como consecuencia de que p_X suma a 1. Entonces,

$$\mathcal{H}_\theta D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right) = \int_{\mathcal{X}} \mathcal{H}_\theta p_X(x;\theta)\log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right)dx + \int_{\mathcal{X}} \frac{\nabla_\theta p_X(x;\theta)\,\nabla_\theta^t p_X(x;\theta)}{p_X(x;\theta)}\,dx.$$

Tomado en $\theta=\theta_0$ el primer término vale cero. En el segundo se reconoce $J_\theta(X)$, lo que termina la prueba. □

Este teorema, ilustrado en la figura Fig. 2-39, relaciona claramente dos objectos viniendo de la teoría de la estimación y de la teoría de la información, mundos a priori diferentes. Como se lo puede ver en la figura, cuando $J_\theta(X)$ tiene pequeños autovalores (figura (a)), p_θ se "aleja" lentamente de θ_0 cuando θ se aleja de θ_0: hay una alta incerteza o pequeña información sobre θ_0. Y vice-versa (figura (b)).

---

reparametrización $\eta=\eta(\theta)$, *i. e.,* el prior de Jeffrey en η es unívocamente obtenido con la Fisher en η o por cambio de variables saliendo de p_Θ.

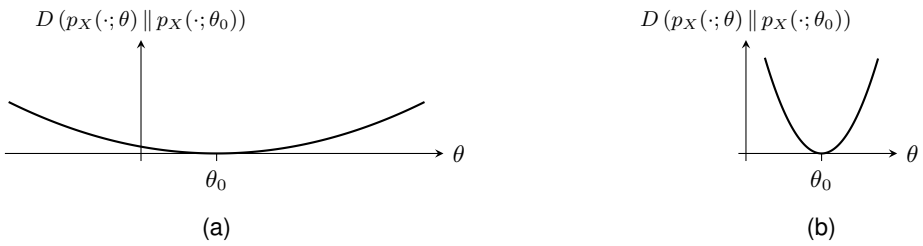


Figura 2-39: Ilustración del comportamiento local de $D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right)$ en función de θ en θ_0 en el contexto escalar $\Theta \subseteq \mathbb{R}$. (a) Caso con $J_{\theta_0}(X)$ "pequeño" y (b) caso con $J_{\theta_0}(X)$ "grande". En el caso (b), la determinación de θ usando D_{kl} va a ser más "sencillo" que en el caso (a) porque el mínimo es más "picado".

2.4.6.3. Identidad de de Bruijn

Un otro vínculo entre el mundo de la información y el de la estimación aparece a través de la identidad de de Bruijn [110] (Stam, 1959; Cover & Thomas, 2006; Johnson, 2004; Barron, 1984, 1986; Palomar & Verdú, 2006; Toranzo, Zozor & Brossier, 2018). Esta identidad caracteriza lo que es conocido como canal gaussiano de la figura Fig 2-40-(a), *i. e.,* la salida Y es una versión ruidosa de la entrada. La identidad vincula las variaciones de entropía de salida con respeto al nivel de ruido, y la información de Fisher.

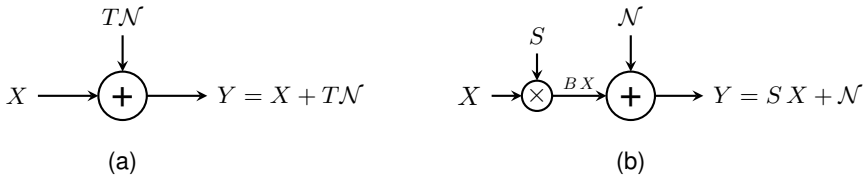


Figura 2-40: Canal de comunicación gaussiano de entrada X. (a) Canal gaussiano usual, donde T maneja los parámetros (nivel) del ruido. (b) canal gaussiano con un preprocesamiento S de la entrada.

Teorema 2-74 (Identidad de de Bruijn)**.** *Sea X un vector aleatorio continuo sobre un abierto de $\mathbb{R}^d$ y admitiendo una matriz de covarianza, y sea $Y = X + T\mathcal{N}$ donde T es determinista, $d \times d'$ con $d \leq d'$, de rango máximo, y $\mathcal{N}$ un vector gaussiano centrado y de covarianza $\Sigma_{\mathcal{N}}$, independiente de X (ver figura Fig. 2-40-(a)). Entonces, la entropía de Shannon y la información de Fisher de Y satisfacen*

$$\nabla_T H(Y) = J(Y)\,T\,\Sigma_{\mathcal{N}},$$

[110] A pesar de que tomó este nombre, esta identidad en su primera versión fue publicada por Stam. En su papel (Stam, 1959), menciona que esta identidad fue comunicada al Profesor van Soest por el Profesor de Bruijn.

donde $\nabla_T \cdot$ *es la matriz de componentes* $\frac{\partial \cdot}{\partial T_{i,j}}$. *Si* $T = T(\theta)$ *depende de un parámetro escalar*[111] θ,

$$\frac{\partial}{\partial \theta} H(Y) = \operatorname{Tr}\left(J(Y)\, T\, \Sigma_{\mathcal{N}}\, \frac{\partial T^t}{\partial \theta} \right).$$

Demostración. La clave de este resultado viene del hecho de que la densidad p de $T\mathcal{N}$ satisface una ecuación diferencial particular. La distribución de $T\mathcal{N}$ se escribe $p(x) = (2\pi)^{-\frac{d}{2}} \left|T\Sigma_{\mathcal{N}} T^t\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^t \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} x\right)$ (el rango máximo de T asegura que $T\Sigma_{\mathcal{N}} T^t$ sea invertible). Para una matriz invertible R, desarollando $|R|$ con respecto a su línea i, se obtiene que $\frac{\partial |R|}{\partial R_{i,j}} = R^*_{i,j}$ cofactor de $R_{i,j}$, dando por la regla de Cramér $\nabla_R |R| = |R|\, R^{-t}$ (ver también (Magnus & Neudecker, 1999, cap. 1 & 9)), es decir $\nabla_R |R|^{-\frac{1}{2}} = -\frac{1}{2} |R|^{-\frac{1}{2}} R^{-t}$. De $\frac{\partial |R|^{-\frac{1}{2}}}{\partial T_{i,j}} = \sum_{k,l} \frac{\partial |R|^{-\frac{1}{2}}}{\partial R_{k,l}} \frac{\partial R_{k,l}}{\partial T_{i,j}} = -\frac{1}{2} |R|^{-\frac{1}{2}} \sum_{k,l} \left(R^{-1}\right)_{l,k} \frac{\partial R_{k,l}}{\partial T_{i,j}}$ con $R = T\Sigma_{\mathcal{N}} T^t$ (simétrica) y cálculos básicos se obtiene finalmente

$$\nabla_T \left|T\Sigma_{\mathcal{N}} T^t\right|^{-\frac{1}{2}} = -\left|T\Sigma_{\mathcal{N}} T^t\right|^{-\frac{1}{2}} \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} T\Sigma_{\mathcal{N}}.$$

Además, de $\left(T\Sigma_{\mathcal{N}} T^t\right)\left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} = I$ viene $\frac{\partial \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1}}{\partial T_{i,j}} = -\left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} \frac{\partial \left(T\Sigma_{\mathcal{N}} T^t\right)}{\partial T_{i,j}} \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1}$. Usando le vector $\mathbb{1}_i$ con 1 en su i-ésima componente, y cero si no, se obtiene

$$\begin{aligned} \frac{\partial \left(x^t \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} x\right)}{\partial T_{i,j}} &= -x^t \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} \left(\mathbb{1}_i \mathbb{1}_j^t \Sigma_{\mathcal{N}} T^t + T\Sigma_{\mathcal{N}} \mathbb{1}_j \mathbb{1}_i^t\right) \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} x \\ &= -2\, \mathbb{1}_i^t \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} x x^t \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} T\Sigma_{\mathcal{N}} \mathbb{1}_j \end{aligned}$$

usando la relación $x^t A \mathbb{1}_k \mathbb{1}_l^t B x = \mathbb{1}_l^t B x x^t A \mathbb{1}_k = \mathbb{1}_k^t A^t x x^t B^t \mathbb{1}_l$ (escalares comutan y un escalar es igual a su transpuesta) y usando la simetría de $T\Sigma_{\mathcal{N}} T^t$. Eso significa que

$$\nabla_T \left(x^t \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} x\right) = -2 \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} x x^t \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} T\Sigma_{\mathcal{N}},$$

dando

$$\nabla_T p(x) = \left(-\left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} + \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1} x x^t \left(T\Sigma_{\mathcal{N}} T^t\right)^{-1}\right) T\Sigma_{\mathcal{N}}\, p(x).$$

Tomando la Hessiana de p con respeto a x se obtiene sencillamente que p satisface la ecuación diferencial

$$\nabla_T\, p = \mathcal{H}_x\, p\, T\, \Sigma_{\mathcal{N}}.$$

Suponiende que se puede intervertir derivadas y integrales (ver (Barron, 1984, 1986) donde se dan condiciones rigurosas, y el teorema 1-7, pagina 41), $p_Y(y) = \int_{\mathbb{R}^d} p_X(x)\, p(y - x)\, dx$ (ver ejemplo 1-8,

[111] Si el parámetro es multivariado, hace falta entender la desigualdad a través de derivas parciales con respeto a los componentes de θ.

pagina 64) satisface también esta ecuación diferencial, y además

$$\begin{aligned}
\nabla_T H(Y) &= -\int_{\mathbb{R}^d} \nabla_T\, p_Y(y) \log p_Y(y)\, dy - \int_{\mathbb{R}^d} \nabla_T\, p_Y(y)\, dy \\
&= -\left(\int_{\mathbb{R}^d} \mathcal{H}_y\, p_Y(y) \log p_Y(y)\, dy\right) T\, \Sigma_{\mathcal{N}} - \nabla_T \int_{\mathbb{R}^d} p_Y(y)\, dy \\
&= -\left(\int_{\mathbb{R}^d} \left(\mathcal{H}_y\Big(p_Y(y)\log p_Y(y)\Big) - \mathcal{H}_y\, p_Y(y) - \frac{\nabla_y\, p_Y(y)\, \nabla_y\, p_Y(y)^t}{p_Y(y)}\right) dy\right) T\, \Sigma_{\mathcal{N}} \\
&= -\left(\int_{\mathbb{R}^d} \mathcal{H}_y\Big(p_Y(y)\log p_Y(y)\Big)\, dy - \int_{\mathbb{R}^d} \mathcal{H}_y p_Y(y)\, dy\right) T\, \Sigma_{\mathcal{N}} \;+\; J(Y)\, T\, \Sigma_{\mathcal{N}}
\end{aligned}$$

usando la ecuación diferencial en la segunda línea, el hecho de que p_Y suma a 1 en la tercera línea (su gradiente es cero entonces), y la definición de la matriz de Fisher en la última línea. Usando el teorema de la divergencia (intergración por partes) aplicada respectivamente a los componentes de $\nabla_y p_Y \log p_Y$ y $\nabla_y p_Y$, suponiendo que estos gradientes se cancelan en el borde del dominio de integración, los dos términos integrales valen cero, lo que cierra la prueba de la desigualdad general. Además, si $T = T(\theta)$, la segunda desigualdad sigue de $\frac{\partial \cdot}{\partial \theta} = \sum_{i,j} \frac{\partial \cdot}{\partial T_{i,j}} \frac{\partial T_{i,j}}{\partial \theta} = \mathrm{Tr}\left(\nabla_T \cdot \frac{\partial T^t}{\partial \theta}\right)$. □

La versión inicial de la identidad de de Bruijn, con $\Sigma_{\mathcal{N}} = I$, que se escribe

$$\frac{d}{d\theta} H(X + \sqrt{\theta}\mathcal{N}) = \frac{1}{2}\,\mathrm{Tr}\left(J(X + \sqrt{\theta}\mathcal{N})\right),$$

se recupera en el caso particular $T = \sqrt{\theta} I$. En este caso, la ecuacíon diferencial satisfecha por la densidad de probabilidad p es la *ecuación del calor*. Esta desigualdad cuantifica las variaciones de entropías bajo variaciones de "niveles" del ruido del canal de comunicación. De una forma, caracteriza la robustez del canal con respeto al nivel de ruido gaussiano (la gaussiana juega de nuevo un rol central acá).

Existe una otra forma muy similar de esta desigualdad debido a Guo, Shamai, Verdú, Palomar (Guo, Shamai & Verdú, 2005; Palomar & Verdú, 2006; Toranzo et al., 2018). Esta versión vincula aún más el mundo de la información y el de la estimación. Del lado de la comunicación, consiste a caracterizar la información mutua entre la entrada X de un canal ruidoso y su salida, $Y = SX + \mathcal{N}$ donde S coresponde a un pre-tratamiento antes de la salida. Eso es ilustrado en la figura Fig. 2-40-(b). Del lado de la estimación, uno puede querer estimar X observando solamente Y. Es conocido que el estimador que minimiza el error cuadrático promedio $\mathrm{E}\left[\left|\widehat{X}(Y) - X\right|^2\right]$ es la esperanza condicional $\widehat{X}(Y) = \mathrm{E}[X|Y]$ (Kay, 1993; Robert, 2007; Lehmann & Casella, 1998). Una característica de un estimador siendo su matriz de covarianza, se notará $\mathcal{E}(X|Y) = \mathrm{E}\left[(X - \mathrm{E}[X|Y])\,(X - \mathrm{E}[X|Y])^t\right]$ esta matriz. Sorpredentemente, existe también una identidad entre $I(X;Y)$ y $\mathcal{E}(X|Y)$:

Teorema 2-75 (Identidad de Guo–Shamai–Verdú)**.** *Sea X un vector aleatorio continuo sobre un abierto de $\mathbb{R}^{d'}$ y admitiendo una matriz de covarianza, y sea $Y = SX + \mathcal{N}$ donde S es determinista,*

$d \times d'$, y $\mathcal{N}$ un vector gaussiano centrado y de covarianza $\Sigma_{\mathcal{N}}$, independiente de X (ver figura Fig. 2-40-(b)). Entonces, la información mutua entre X e Y y la matriz de covarianza del estimador de error cuadrático mínimo satisfacen

$$\nabla_S \, I(X;Y) = \Sigma_{\mathcal{N}}^{-1} S \, \mathcal{E}(X|Y).$$

Si $S = S(\mathfrak{s})$ depende de un parámetro escalar $\mathfrak{s}$,

$$\frac{\partial}{\partial \mathfrak{s}} I(X;Y) = \operatorname{Tr}\left(\Sigma_{\mathcal{N}}^{-1} \, S \, \mathcal{E}(X|Y) \, \frac{\partial S^t}{\partial \mathfrak{s}} \right).$$

Demostración. Notando que $p_{Y|X=x}(y) = (2\pi)^{-\frac{d}{2}} \, |\Sigma_{\mathcal{N}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-Sx)^t \Sigma_{\mathcal{N}}^{-1}(y-Sx)\right)$ viene $\nabla_S \, p_{Y|X=x}(y) = p_{Y|X=x}(y) \, \Sigma_{\mathcal{N}}^{-1}(y-Sx)x^t$ (ver unos pasos de la prueba de la identidad de de Bruijn) así que $\nabla_y \, p_{Y|X=x}(y) = p_{Y|X=x}(y) \, \Sigma_{\mathcal{N}}^{-1}(y-Sx)$, dando

$$\nabla_S \, p_{Y|X=x}(y) = \nabla_y \, p_{Y|X=x}(y) x^t \qquad \text{y} \qquad \nabla_S \, p_{X,Y}(x,y) = \nabla_y \, p_{X,Y}(x,y) x^t$$

(multiplicando ambos lados por p_X). Ahora, $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\mathcal{N})$ (de la independencia, cuando $X = x$, $Y = Sx + \mathcal{N}$ gaussiana de misma convarianza que $\mathcal{N}$ y de promedio Sx (ver ejemplo 1-9, pagina 69), así que

$$\begin{aligned}
\nabla_S I(X;Y) &= \nabla_S H(Y) \\
&= -\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \nabla_S \Big(p_{X,Y}(x,y) \, \log p_Y(y) \Big) \, dx \, dy \\
&= -\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \nabla_S \, p_{X,Y}(x,y) \, \log p_Y(y) \, dx \, dy - \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} p_{X|Y=y}(x) \, \nabla_S \, p_Y(y) \, dx \, dy \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \nabla_y \, p_{X,Y}(x,y) \, x^t \log p_Y(y) \, dx \, dy - \int_{\mathbb{R}^d} \nabla_S \, p_Y(y) \, dy \\
&= -\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \nabla_y \, p_Y(y) x^t p_{X|Y=y}(x) \, dx \, dy \\
&= -\int_{\mathbb{R}^d} \nabla_y \, p_Y(y) \, \mathrm{E}\left[X^t | Y = y \right] \, dy
\end{aligned}$$

La segunda línea viene de la escritura de $H(Y)$ usando p_Y como marginale de $p_{X,Y}$ en x y intercambiando gradiente e integral (ver pasos de la prueba de la desigualdad de de Bruijn); la tercera de $\frac{p_{X,Y}(x,y)}{p_Y(y)} = p_{X|Y=y}(x)$; en la cuarta se usa la ecuación diferencial satisfecha por $p_{X,Y}$ en la primera integral y integrando en x en la segunda integral; la quinta línea se obtiene usando el teorema de la divergencia (intergración por partes) en la integración en y de la primera integral, e intercambiando

gradiente e integral el la segunda (p_Y sumando a 1, el término se cancela). Además,

$$\begin{aligned}
\nabla_y p_Y(y) &= \int_{\mathbb{R}^{d'}} \nabla_y \, p_{Y|X=x}(y)\, p_X(x)\, dx \\
&= -\Sigma_{\mathcal{N}}^{-1} \int_{\mathbb{R}^{d'}} (y - Sx)\, p_{Y|X=x}(y)\, p_X(x)\, dx \\
&= -\Sigma_{\mathcal{N}}^{-1} \left(y - S \int_{\mathbb{R}^{d'}} x\, p_{X|Y=y}(x)\, dx \right) p_Y(y) \\
&= -\Sigma_{\mathcal{N}}^{-1} \Big(y - S\, \mathrm{E}\left[X|Y=y\right] \Big)\, p_Y(y)
\end{aligned}$$

escribiendo $p_{Y|X=x}(y)\, p_X(x) = p_{X|Y=y}(x)\, p_Y(y)$ en la tercera línea. Esta ecuación permite escribir

$$\begin{aligned}
\nabla_S I(X;Y) &= \Sigma_{\mathcal{N}}^{-1} \int_{\mathbb{R}^d} \Big(y - S\, \mathrm{E}\left[X|Y=y\right] \Big)\, \mathrm{E}\left[X^t|Y=y\right]\, p_Y(y)\, dy \\
&= \Sigma_{\mathcal{N}}^{-1} \left(\mathrm{E}\left[Y\, \mathrm{E}\left[X^t|Y\right] \right] - S\, \mathrm{E}\left[\mathrm{E}\left[X|Y\right] \mathrm{E}\left[X|Y\right]^t \right] \right) \\
&= \Sigma_{\mathcal{N}}^{-1} \left(\mathrm{E}\left[Y X^t \right] - S\, \mathrm{E}\left[\mathrm{E}\left[X|Y\right] \mathrm{E}\left[X|Y\right]^t \right] \right) \\
&= \Sigma_{\mathcal{N}}^{-1} S \left(\mathrm{E}\left[X X^t \right] - \mathrm{E}\left[\mathrm{E}\left[X|Y\right] \mathrm{E}\left[X|Y\right]^t \right] \right)
\end{aligned}$$

la última línea viniendo de $Y = SX + \mathcal{N}$ con $\mathcal{N}$ independiente de X y de promedio 0. La prueba se cierra notando que $\mathrm{E}\left[\mathrm{E}[X|Y]\right] = \mathrm{E}[X]$ y por la formula de König-Huyggens (ver capítulo **??**, subsección 1.6.2, pagina 72).

La segunda identidad viene de $\frac{\partial \cdot}{\partial \mathfrak{s}} = \mathrm{Tr}\left(\nabla_S \frac{\partial S^t}{\partial \mathfrak{s}} \right)$ (ver prueba de la identidad de de Bruijn). □

La primera versión de esta identidad se recupera con $S = \sqrt{\mathfrak{s}}$, $\Sigma_{\mathcal{N}} = I$ y X de covarianza la identidad; s es conocido como relación señale/ruido en este caso.

Existen versiones aún más completas (con gradientes con respeto a la matriz $\Sigma_{\mathcal{N}}$ por ejemplo) que se pueden consultar en (Johnson, 2004; Palomar & Verdú, 2006; Payaró & Palomar, 2009).

2.4.6.4. Desigualdad de Stam

De la desigualdad de la potencia entrópica y de la identidad de de Bruijn surge una otra desigualdad implicando la potencia entrópica N y la información de Fisher J. Esta desigualdad es conocida como desigualdad de Stam [112] (Cover & Thomas, 2006; Rioul, 2007; Stam, 1959), o a veces "desigualdad isoperimetrica para la entropía" (Wang & Madiman, 2004).

Teorema 2-76 (Desigualdad de Stam)**.** *Sea X una variable aleatoria continua sobre $\mathcal{X} \subseteq \mathbb{R}^d$. Entonces,*

$$N(X)\, \mathrm{Tr}\left(J(X)\right) \geq d,$$

[112] Como para la identidad de de Bruijn, Stam mencionó que esta desigualdad fue comunicada al Profesor van Soest por el Profesor de Bruijn quien da una prueba variacional de la desigualdad.

con igualdad si y solamente si X es gaussiano de covarianza proporcional a la identidad.

Demostración. De la desigualdad de la potencia entrópica se obtiene $N(X+\sqrt{\theta}\mathcal{N}) \geq N(X)+\theta\,|\Sigma_{\mathcal{N}}|^{\frac{1}{d}}$. Tomando $\Sigma_{\mathcal{N}} = I$, se obtiene $\forall\,\theta > 0,\ \frac{N(X+\sqrt{\theta}\mathcal{N})-N(X)}{\theta} \geq 1$. Entonces, tomando el límite $\theta \to 0$, aparece que $\frac{d}{d\theta}N(X+\sqrt{\theta}\mathcal{N})\Big|_{\theta=0} \geq 1$. La prueba se cierra con $\frac{d}{d\theta}N(X+\sqrt{\theta}\mathcal{N}) = \frac{1}{2\pi\,\mathrm{e}}\frac{d}{d\theta}\exp\left(\frac{2}{d}H(X+\sqrt{\theta}\mathcal{N})\right) = \frac{2}{d}N(X+\sqrt{\theta}\mathcal{N})\frac{d}{d\theta}H(X+\sqrt{\theta}\mathcal{N}) = dN(X+\sqrt{\theta}\mathcal{N})\,\mathrm{Tr}\left(J(X+\sqrt{\theta}\mathcal{N})\right)$ (por la identidad de de Bruijn). Además, la igualdad se obtiene cuando se alcanza la cota de la desigualdad de la potencia entrópica, es decir cuando X es gaussiano de varianza proporcional a la del ruido, que es la identidad en este caso. □

Se puede ver de nuevo el rol central que juega la gaussiana en esta desigualdad. Además, de la desigualdad de Stam se puede deducir tambíen las versiones escalares de la desigualdad de Cramér-Rao. Viene del hecho de que, dada una matriz de covarianza, la entropía $H(X)$ es máxima cuando X es gaussiano. Entonces, para cualquier X de covarianza Σ_X, $N(X) \leq |\Sigma_X|^{\frac{1}{d}}$, dando de la desiguldad de Stam, $|\Sigma_X|^{\frac{1}{d}}\,\mathrm{Tr}\,(J(X)) \geq d$ (y las otras versiones escalares de la relación determinente-traza). Como se lo puede esperar, se obtiene la igualdad si y solamente X es gaussiano (potencia entrópica alcanzando su cota superior) y de matriz la identidad (desiguladad de Stam se saturando).

Varias otras pruebas de la desigualdad de Stam pueden provenir de generalizaciones (Bercher, 2012, 2013; Lutwak, Yang & Zhang, 2005; Lutwak, Lv, Yang & Zhang, 2012; Zozor, Puertas-Centeno & Dehesa, 2017). **La sección ZZZ lo va a rapidamente evocar. Ver caso discreto Kagan (Kagan, 2001).**

2.4.6.5. Fisher aditividad, procesamiento de datos y convolución

Además del grán número de relaciones entre la información de Fisher y otras medidas informacionales, la información de Fisher satisface también desigualdades en si mismo, muy parecidas a las satisfechas por la entropía o información mutua.

Primero, al imagen de la entropía condicional, se puede definir una información condicional al imagen de la definición Def. 2-63,

Definición 2-69 (Matriz información de Fisher parámetrica condicional)**.** *Sean X e Y dos variables aleatoria parametrizada por el mismo parámetro m-dimensional, $\theta \in \Theta \subseteq \mathbb{R}^m$, de distribución de probabilidad conjunta $p_{X,Y}(\cdot,\cdot;\theta)$ continua sobre $\mathcal{X}\times\mathcal{Y}$ su soporte, $p_{X|Y=y}(\cdot;\theta)$ la distribución condicional de X connociendo $Y=y$ y p_Y la distribución marginal. Suponga que estas distribuciones sean diferenciable en θ sobre Θ. La matriz de Fisher de X condicionalmente a Y es el promedio estadístico sobre p_Y de la matriz de Fisher de $p_{X|Y}(\cdot;\theta)$, es decir*

$$J_\theta(X|Y) = \mathrm{E}\left[\Big(\nabla_\theta \log p_{X|Y}(X;\theta)\Big)\Big(\nabla_\theta \log p_{X|Y}(X;\theta)\Big)^t\right].$$

donde $p_{X|Y}(\cdot;\theta) = \frac{p_{X,Y}(\cdot,Y;\theta)}{p_Y(Y)}$ es acá una variable aleatoria.

De esta definición, es sencillo probar de que la matriz de Fisher paramétrica sigue una regla de cadena al imagen de la propiedad [P14],

$$J_\theta(X,Y) = J_\theta(X|Y) + J_\theta(Y).$$

Además, si X e Y son independientes, la información de Fisher es aditiva de la misma manera que H satisface las propiedades [P10] y [P13], *i. e.,*

$$J_\theta(X|Y) = J_\theta(X) \quad \Leftrightarrow \quad J_\theta(X,Y) = J_\theta(X) + J_\theta(Y) \quad \Leftrightarrow \quad X \,\&\, Y \text{ son independientes.}$$

En particular, tratando de una secuencia $X = \{X_i\}_{i=1}^n$ de vectores aleatorias independientes parametrizados por θ, $J_\theta(X) = nJ_\theta(X_i)$, lo que significa que estimando θ a partir de la secuencia se baja a la taza $1/n$ la cota de Cramér-Rao. Se referirá a (Fisher, 1925b; Stam, 1959; Kay, 1993; Kagan & Smith, 1999; Johnson, 2004; Cover & Thomas, 2006; Rioul, 2007) entre otros para estas propiedades.

De la regla de cadena, viene obviamente la desigualdad siguiente, parecida a la propiedad de superaditividad [P12],

$$J_\theta(X_1, \ldots, X_n) \;\geq\; J_\theta(X_i) \quad \forall\, 1 \leq i \leq n,$$

y una desigualdad de procesamiento de datos via la información de Fisher (Zamir, 1998; Rioul, 2007; Cover & Thomas, 2006; Frieden, 2004; Kagan & Smith, 1999):

Teorema 2-77 (Desigualdad de procesamiento de datos tipo Fisher)**.** *Sea $\theta \mapsto X \mapsto Y$ un proceso de Markov con θ determinista y $p_{X,Y}$ parametrizado por θ, es decir en este contexto que, $p_{Y|X=x}$ no es parametrizado por θ. Entonces*

$$J_\theta(X) \geq J_\theta(Y),$$

con igualdad si y solamente si $\theta \mapsto Y \mapsto X$ es también de Markov. En particular,

$$\forall\, g, \quad J_\theta(X) \geq J_\theta(g(X)).$$

Demostración. De la regla de cadena tenemos

$$J_\theta(Y|X) + J_\theta(Y) = J_\theta(X|Y) + J_\theta(X).$$

Del hecho de que $p_{Y|X=x}$ no es parametrizado por θ es sencillo ver que $J_\theta(Y|X) = 0$, la prueba se cerrando de $J_\theta(X|Y) \geq 0$. Además se obtiene la igualdad si y solamente si $J_\theta(X|Y) = 0$, es decir de la "positividad" del integrante dando la matrix de Fisher, si y solamente si $p_{X|Y=y}$ no es parametrizado por θ. □

Mencionamos de que existe también una desigualdad parecida a la de la potencia entrópica, teorema 2-69, dada en el caso escalar en (Johnson, 2004; Blachman, 1965; Zamir, 1998; Dembo et al., 1991; Kagan & Yu, 2008):

Teorema 2-78 (Desigualdad convolucional de Fisher)**.** *Sean X e Y dos variables d-dimensionales continuas independientes parametrizadas. Entonces*

$$\forall\, a \in [0\,;\,1], \quad J\left(\sqrt{a}\,X + \sqrt{1-a}\,Y\right) \,\leq\, a\,J(X) + (1-a)\,J(Y),$$

con igualdad si y solamente si X e Y son gaussianas con matrices de covarianza proporcionales, $\Sigma_Y \propto \Sigma_X$.

Demostración. X e Y siendo independientes, tenemos para $W = X + Y$, $p_W(w) = \int_{\mathcal{X}} p_X(x) p_Y(w - x)\,dx$ convolución de las distribuciones de X y de Y (ver ejemplo 1-8, pagina 64). Escribiendo $S_X = \nabla_x \log p_X$ el score de X y lo mismo para Y y W,

$$\begin{aligned}
S_W(w) &= \int_{\mathcal{X}} \frac{p_X(x)}{p_W(w)}\,\nabla_w p_Y(w-x)\,dx \\
&= -\int_{\mathcal{X}} \frac{p_X(x)}{p_W(w)}\,\nabla_x p_Y(w-x)\,dx \\
&= \int_{\mathcal{X}} \frac{p_Y(w-x)}{p_W(w)}\,\nabla_x p_X(x)\,dx \\
&= \int_{\mathcal{X}} \frac{p_Y(w-x)p_X(x)}{p_W(w)}\,\nabla_x \log p_X(x)\,dx \\
&= \int_{\mathcal{X}} p_{X|W=w}(w)\,\nabla_x \log p_X(x)\,dx \\
&= \mathrm{E}\left[\,S_X(X)\,|\,W = w\right]
\end{aligned}$$

Intercambiando los roles de X e Y, tenemos también $S_W(w) = \mathrm{E}\left[\,S_Y(Y)\,|\,W = w\right]$, así que, para cualquier $0 \leq a \leq 1$,

$$S_W(w) = \mathrm{E}\left[\,aS_X(X) + (1-a)S_Y(Y)\,|\,W = w\right].$$

A continuación, de la formula de König-Huyggens (ver capítulo **??**, subsección 1.6.2, pagina 72),

$$S_W(w)S_W(w)^t \,\leq\, \mathrm{E}\left[\left(aS_X(X) + (1-a)S_Y(Y)\right)\left(aS_X(X) + (1-a)S_Y(Y)\right)^t \middle|\, W = w\right],$$

es decir, tomando el promedio en W,

$$J(X+Y) \,\leq\, a^2 J(X) + (1-a)^2 J(Y) + a(1-a)\left(\mathrm{E}\left[S_X(X)S_Y(Y)^t\right] + \mathrm{E}\left[S_Y(Y)S_X(X)^t\right]\right).$$

Luego, X e Y siendo independientes, $S_X(X)$ y $S_Y(Y)$ son también independientes. Además son centradas, probando que el término en $a(1-a)$ vale cero, dando una versión equivalente del teorema; La versión dada se recupera re-emplazando X por $\sqrt{a}\,X$ e Y por $\sqrt{1-a}\,Y$.

Escribiendo la desigualdad viniendo de la formula de König-Huyggens, se nota de que la igualdad es satisfecha si y solamente si $aS_X(w-x) + (1-a)S_Y(x) = S_W(w)$ para cualquier x, w. Integrando en x se obtiene $-a\log p_X(w-x) + (1-a)\log p_Y(x) = xS_W(w) + g(w)$. Derivando en w obtenemos $-a\nabla_w \log p_X(w-x) = S_W(w) + \nabla_w g(w)$, es decir, en $w = 0$, notando de que $\nabla_w \log p_X(w-x) =$

$-\nabla_x \log p_X(w-x)$, se nota de que $\nabla_x \log p_X(x)$ es constante, *i. e.,* X es necesariamente gausiana. Similarmente, Y es necesariamente gausiana también. Además, calculando las informaciones de Fisher en el caso gausiano, obtenemos $(\Sigma_X + \Sigma_Y)^{-1} = \Sigma_X^{-1} + \Sigma_Y^{-1}$ lo que es posible si y solamente si Σ_X y Σ_Y son proporcionales. □

Este teorema tiene varias consecuencias. En particular, interviene en la prueba de la desigualdad de la potencia entrópica.

(2) ver MinFisher Frieden p. 235, Berchet Vignat 2009, Ernst 2017; cf. travaux rederivant MQ de Frieden-Plastino-Soffer (1999, 2002), Reginato 98, Bickel 81

2.5 Unos ejemplos y aplicaciones

2.5.1 Canal de transmisión y su capacidad

Siguiendo el esquema de comunicación de Shannon, un mensaje que se modeliza como un vector aleatorio [113] X pasa por un canal de comunicación y se recibe un mensaje Y, vector aleatorio. En el trabajo de Shannon, el canal es supuesto a ruido aditivo, es decir que se añade un ruido a X. De manera general, para conocer la información de X que se recibe, se calcula la información mutua $I(X;Y)$, es decir la cantidad de información que comparten la entrada y la salida del canal. Lo más I es grande, lo más de información se transmite. Dado el canal, se puede arreglar X (su distribución) de manera a maximizar $I(X;Y)$, es decir la cantidad máxima que se puede transmitir en este canal. Es lo que es conocido como capacidad del canal (Shannon, 1948, part. II & III) (ver también (Cover & Thomas, 2006; Rioul, 2007) entre otros):

Definición 2-70 (Capacidad de canal)**.** *Sea un canal de transmisión, X su entrada e Y su salida, como ilustrado figura Fig. 2-41. Sea p_X la distribución de probabilidad de X. La capacidad C del canal es definida por*

$$C = \max_{p_X} I(X;Y).$$

2.5.1.1. Canal binario

[113] De punto de vista de un receptor, este mensaje es desconocido. Además, se lo puede ver como una instancia de una clase importante de posibles mensajes, justificando la modelización aleatoria.

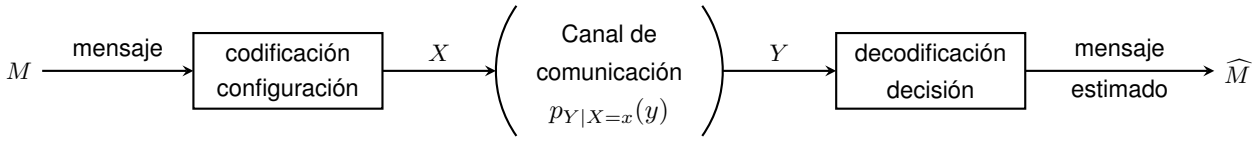


Figura 2-41: Esquema de comunicación de Shannon. En una primera etapa, un mensaje M a transmitir es códificado (ej. código binario) o puesto en forma (ej. símbolos modulando una función para que sea analógica y en una banda frecuencial dada). Sea X este mensaje codificado o puesto en forma. A la recepción, se mide Y (ej. versión ruidosa de X), antes de ser decodificado o usado para tomar una decisión, $\widehat{M}$ siendo la estimación de M (ej. símbolos estimados a partir de Y). Una etapa importante es el vínculo entre la entrada X y la salida Y del canal, es decir la cantidad de información que tienen en común. La capacidad del canal es la información $I(X;Y)$ máxima con respeto a su entrada.

Suponiendo que el mensaje mandado en un canal es una cadena de símbolos, variables aleatorias independientes, se puede concentrarse sobre cada símbolo. En este marco, un canal de comunicación lo más simple es conocido como *canal binario* (Shannon, 1948, Sec. 15): X es una variable definida sobre $\mathcal{X} = \{0\,,\,1\}$; tal tipo de entrada es natural, pensando a la codificación binaria. La salida Y es también definida sobre $\mathcal{X}$; se puede imaginar medir y tomar una decisión binaria usando la medida. Tal canal es definido por sus probabilidaded de transición $p_{Y|X=x}(y)$, *i. e.,* las probabilidades que un 0 (resp. un 1) se transmite correctamente o cambia en un 1 (resp. 0), *i. e.,*

$$\varepsilon = PY = 1|X = 0) = 1 - P(Y = 0|X = 0) \qquad \text{y} \qquad \vartheta = P(Y = 0|X = 1) = 1 - P(Y = 1|X = 1).$$

ε y ϑ representan errores de comunicación. Tal canal es descrito figura Fig. 2-42-(a). La figura Fig. 2-42-(b) da un esquema "práctico" que podría ser al origen de un tal canal. Cuando $\varepsilon = \vartheta$, el canal es conocido como *canal binario simétrico*. Cuando $\varepsilon = 0$ y $\vartheta \in (0\,;\,1)$, el canal es conocido como *canal binario en Z*.

En este caso, trabajando con bits, aparece legítimo usar el logaritmo de base 2. Luego, sean

$$r = P(X = 0),$$

dando la distribución de la entrada. La distribución de la salida va a ser dada a partir de $s = P(Y = 0) = P(Y = 0|X = 0)P(X = 0) + P(Y = 0|X = 1)P(X = 1)$ es decir

$$s = P(Y = 0) = \vartheta + r(1 - \varepsilon - \vartheta).$$

La información mutua se escribe $I_2(X;Y) = H_2(Y) - H_2(Y|X) = H_2(Y) - H_2(Y|X = 0)P(X = 0) + H_2(Y|X = 1)P(X = 1)$, lo que toma la expresión

$$I_2(X;Y) = h_2(s) - rh_2(\varepsilon) - (1 - r)h_2(\vartheta),$$

donde $h_2(u) = -u\log_2 u - (1 - u)\log_2(1 - u)$ es la entropía binaria en bits. Para calcular la capacidad C_2 en bits, hace falta maximizar I_2 con respeto a r. Diferenciando I_2 en r, *i. e.,* $\frac{\partial I_2(X;Y)}{\partial r} = \frac{\partial h_2(s)}{\partial s}\frac{\partial s}{\partial r} -$

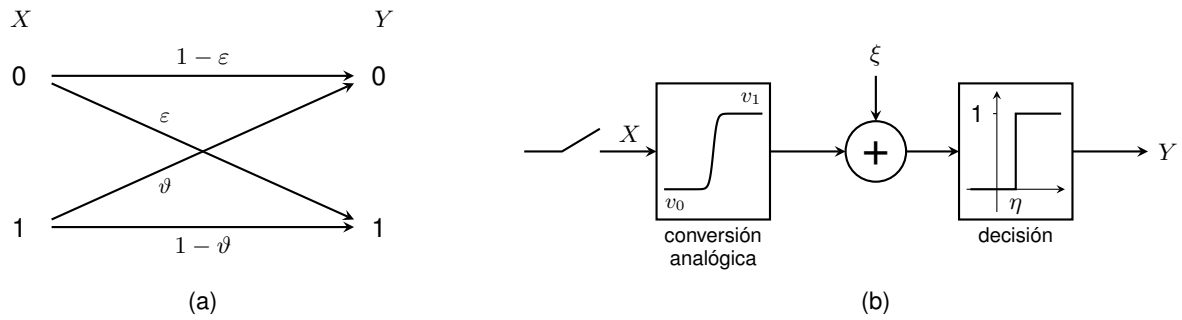


Figura 2-42: (a): Canal binario. La entrada X definida sobre $\mathcal{X} = \{0\,,\,1\}$ pasa por este canal e Y definida sobre $\mathcal{Y} = \mathcal{X}$ es recibido. Este canal es caracterizado por las probabilidades de transición $p_{Y|X=x}(y)$. (b): Esquema que puede conducir al canal binario; una variable puede ser la salida de una puerta lógica, con niveles v_0 (nivel bajo, codificando 0) y v_1 (nivel alto, codificando 1). Se puede imaginar que este voltaje es transmitido por un canal añandiendo un ruido ξ. En la recepción, se toma una decisión, por ejemplo 0 (resp. 1) si la medida es mayor (resp. menor) que $\eta = \frac{v_0+v_1}{2} + \mathrm{E}[\xi]$. En este ejemplo, ε y ϑ van a ser caracterizados completamente por la distribución del ruido (y de los dos niveles posibles de la entrada), pero no de la distribución p_X.

$h_2(\varepsilon) + h_2(\vartheta)$, es decir

$$\frac{\partial I_2(X;Y)}{\partial r} = (1-\varepsilon-\vartheta)\log_2\left(\frac{1-s}{s}\right) - h_2(\varepsilon) + h_2(\vartheta).$$

- Claramente,

$$\vartheta = 1-\varepsilon \quad \Rightarrow \quad C_2 = 0.$$

  Viene del hecho de que para $\vartheta = 1-\varepsilon$, de $h_2(\varepsilon) = h_2(1-\varepsilon)$ se deduce que $I_2(X;Y) = 0$ constante. De hecho, en este caso, un 0 en la salida puede venir de un 0 o 1 con probabilidades iguales, y lo mismo para un 1 en la salida; en otros términos, la salida aparece ser independiente de la entrada. Eso se verifica formalmente con $s = \vartheta$, dando $p_{Y|X=x} = p_Y$, dando una información mutua nula, y entonces una capacidad nula.

- Si $\vartheta \neq 1-\varepsilon$, la derivada de I_2 con respeto a r se anula para $s = s^{\mathrm{opt}}$ ($r = r^{\mathrm{opt}}$),

$$s^{\mathrm{opt}} = \frac{1}{1+2^{\frac{h(\varepsilon)-h(\vartheta)}{1-\varepsilon-\vartheta}}} \qquad \text{siendo} \qquad r^{\mathrm{opt}} = \frac{s^{\mathrm{opt}}-\vartheta}{1-\varepsilon-\vartheta},$$

  y dando un extremo para I_2. A continuación, $\frac{\partial^2 I_2}{\partial r^2} = \frac{(1-\varepsilon-\vartheta)^2}{s(1-s)} > 0$ (en particular para el s "óptimo"), probando de que el extremo es un máximo. Poniendo la expresión de r^{opt} en la formula de $I_2(X;Y)$, luego de muchos cálculos (básicos), se obtiene

$$C_2 = \log_2\left(1+2^{\frac{h_2(\varepsilon)-h_2(\vartheta)}{1-\varepsilon-\vartheta}}\right) - \frac{(1-\vartheta)\,h_2(\varepsilon) - \varepsilon\,h_2(\vartheta)}{1-\varepsilon-\vartheta}.$$

  Cuando $\vartheta \to 1-\varepsilon$, notando que $h_2(\varepsilon) = h_2(1-\varepsilon)$ y tomando el límite de esta formula, se recupera que $C_2 \to 0$.

De $I_2(X;Y) = H_2(Y) - H_2(Y|X) \leq H_2(Y) \leq 1$ bit (Y es binario, de entropía máxima en el caso uniforme), aparece sin cálculos que

$$C_2 \leq 1 \text{ bit},$$

i. e., la capacidad es menor que 1 bit [114]: para transmitir información en este canal, hace falta introducir redundancia en el mensaje. Se alcanza $C_2 = 1$ bit si, (i) por un lado $H_2(Y|X) = 0$, es decir $rh_2(\varepsilon) + (1-r)h_2(\vartheta) = 0$ y además (ii) $h_2(s) = 1$. Estudiando cada caso (ej. con $r = 0$ y $\vartheta = 0$ se satisface (i) pero no (ii) porque $s = 0$), se obtiene que

$$C_2 = 1 \qquad \Leftrightarrow \qquad r = \frac{1}{2} \quad \text{y} \quad \varepsilon = \vartheta = \frac{1 \pm 1}{2}.$$

Para $\varepsilon = \vartheta = 0$ el canal es perfecto, mientras que para $\varepsilon = \vartheta = 1$ el canal es llamado *canal volteando*; en ambos casos, se recupera la entrada (o directamente, o tomando el opuesto) “sin perdida”.

La figura Fig. 2-43 representa la información mutua $I(X;Y)$ para unos canales (ε y ϑ dados) en función de r. Se nota que la curva es cóncava y tiene un máximo, capacidad del canal. La figura Fig. 2-44 representa la capacidad del canal en función de ε y ϑ así que unos casos particulares/cortes.

En el caso particular $\varepsilon = \vartheta$, conocido como *canal simético*, la capacidad es

$$C_2 = 1 - h_2(\varepsilon)$$

(alcanzada con una entrada uniforme). Como visto en el caso general, la capacidad vale 1 bit si y solamente si $h_2(\varepsilon) = 0$, es decir $\varepsilon = 0$ o $\varepsilon = 1$. Al revés, la capacidad es mínima cuando H_2 est máximo, es decir para $\varepsilon = \vartheta = \frac{1}{2}$, y $C_2 = 0$ (instancia particular de $\vartheta = 1 - \varepsilon$). $h_2(\varepsilon)$ es la perdida en bit para cada bit transmitido. La capacidad C_2 en función de ε es dada figura Fig. 2-44-(b).

En el caso particular $\varepsilon = 0$, conocido como *canal en Z*, la capacidad es

$$C_2 = \log_2\left(1 + 2^{-\frac{h_2(\vartheta)}{1-\vartheta}}\right).$$

Se nota en este caso también que la capacidad alcanza 1, su máximo, si y solamente si $\vartheta = 0$ (canal perfecto). Al revés, cuando $\vartheta \to 1$, $C \to 0$, instancia particular de $\vartheta = 1 - \varepsilon$. La capacidad C_2 en función de ϑ es dada figura Fig. 2-44-(c).

En (Cover & Thomas, 2006; Rioul, 2007) entre otros, se estudian diversos otros canales discretos, binarios o con más estados. Unos son representados en la figura Fig. 2-45 (ver también (Shannon, 1948; Elias, 1957) o (Arimoto, 1972) para el cálculo numérico de la capacidad en el caso general).

---

[114] De manera general, de la escritura de I con entropías condicionales, para X definido sobre $\mathcal{X}$ e Y sobre $\mathcal{Y}$, da $0 \leq C \leq \min(\log|\mathcal{X}|, \log|\mathcal{Y}|)$. Además, $p_{Y|X=x}$ depende solo del canal y no de la entrada, así que para $p_X = \pi_1 p_{(1)} + \pi_2 p_{(2)}$ ($\pi_2 = 1 - \pi_1$) se obtiene $p_Y = \pi_1 q_{(1)} + \pi_2 q_{(2)}$ con $q_{(i)}$ distribución de la salida corespondiente a una entrada de distribución $p_{(i)}$. Ahora, de $I(X;Y) = H(Y) - H(Y|X)$, el segundo término siendo dependiente solamente del canal, de la concavidad de H se obtiene de que I es cóncava con respeto a p_X. A continuación, p_X parteneciendo a un convexo, I tiene un máximo que es único.

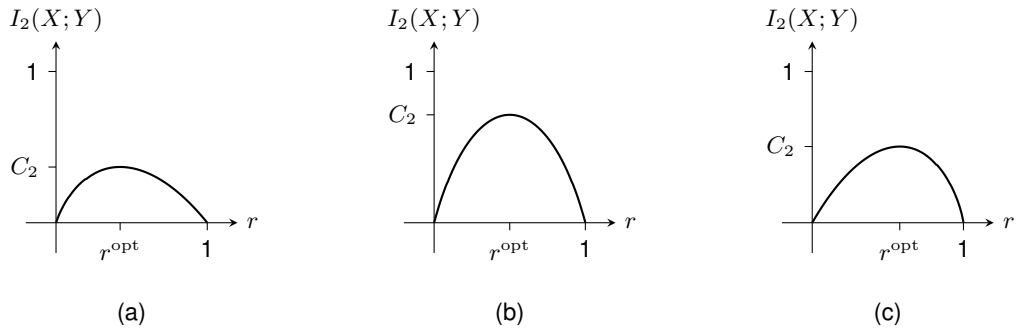


Figura 2-43: Información mutua (en bits) entrada-salida $I_2(X;Y)$ del canal binario en función de $r = P(X = 0)$. (a): $\varepsilon = 0{,}4$ y $\vartheta = 0{,}01$; (b): $\varepsilon = \vartheta = 0{,}05$ (canal simético); (c): $\varepsilon = 0$ y $\vartheta = 0{,}05$ (canal en Z).

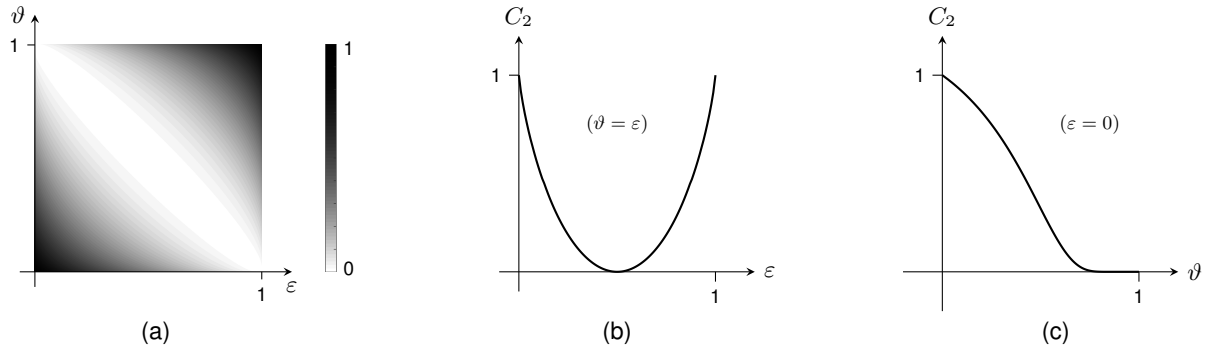


Figura 2-44: Capacidad C_2 del canal binario. (a): en función de ε y ϑ. (b): en función de ε para el canal simético ($\varepsilon = \vartheta$); (c): en función de ϑ para $\varepsilon = 0$ (canal en Z).

2.5.2 Canal de transmisión continuo gaussiano y su capacidad

Un canal de comunicación continuo relativamente simple es conocido como *canal gaussiano* (Shannon, 1948, Sec. 25), (Cover & Thomas, 2006; Rioul, 2007): X es una variable continua definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$ y la salida Y es una versión ruidosa de X, *i. e.,* $Y = X + \xi$ con el ruido ξ independiente de X. En el canal gaussiano, $\xi \equiv \mathcal{N}$ es un vector gaussiano. Este canal es también definido por su densidad de probabilidad "de transición" $p_{Y|X=x}(y)$, *i. e.,* por la distribución del ruido. Tal canal es descrito figura Fig. 2-46. Se supone conocida la matriz de covarianza $\Sigma_{\mathcal{N}}$ del ruido, y se nota Σ_X la de la entrada. En práctica, no se puede mandar un mensaje a una potencia tan alta que se quiere, lo

[115]Se mencionará de que en toda esta subsección, no se necesita de que X y/o Y sean variables aleatorias reales, *i. e.,* pueden tomar sus valores sobre cualquier espacio discreto (por ejemplo de letras).

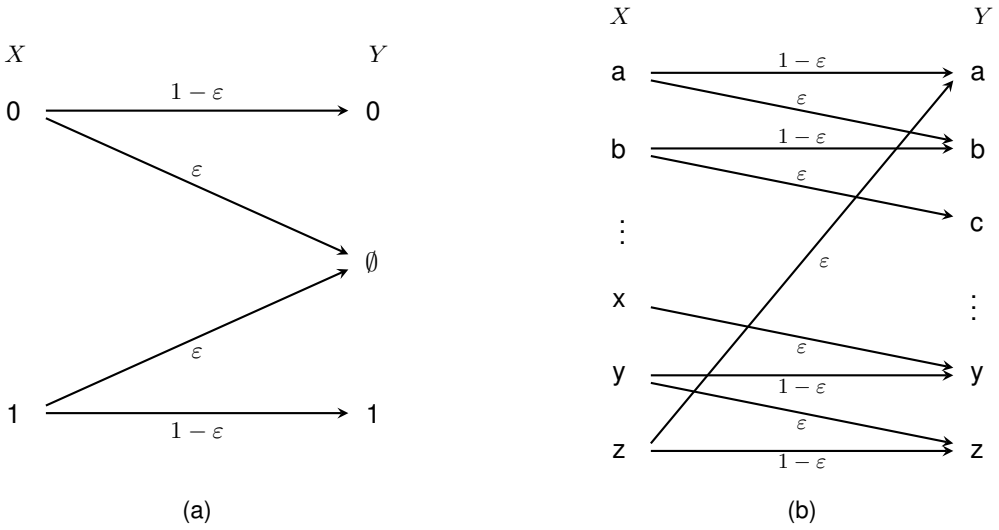


Figura 2-45: Ejemplos de canales discretos usuales. (a): canal borrador, donde un 0 (de probabilidad de ocurrencia r) o 1 (de probabilidad de occurrencia $1-r$) puede transmitirse correctamente or ser borado/perdido (estado $\emptyset$) con una probabilidad ε. Se calcula $I_2(X;Y) = (1-\varepsilon)h_2(r)$, dando la capacidad $C_2 = 1-\varepsilon$, alcanzada para una entrada uniforme. (b): canal tipo machina de escribir [115], donde cada letra de un ensemble de n letras (acá con $n = 26$) se transmite correctamente con una probabilidad $1-\varepsilon$ o a la letra siguiente (de manera cíclica) con una probabilidad ε. De $I_n(X;Y) = H_n(Y) - H_n(Y|X) = H_n(Y) - h_n(\varepsilon)$ se deduce que I_n es máxima si Y es uniforme, lo que es posible si X es uniforma, dando $C_n = 1 - h_n(\varepsilon)$.

que se traduce por una limitación

$$\mathrm{Tr}\,(\Sigma_X) \leq \mathcal{P},$$

potencia límite permitida por componente (sampleo).

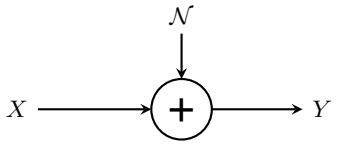


Figura 2-46: Canal gaussiano. La entrada X, modelizada por un vector aleatorio, es corrupta aditivamente por un ruido gaussiano $\mathcal{N}$ independiente de X. La salida es entonces $Y = X + \mathcal{N}$ y el canal es completamente descrito por $p_{Y|X=x}(y) = p_{\mathcal{N}}(y-x)$ (obviamente independiente de la distribución de la entrada).

Por definición, la información mutua $I(X;Y)$ entrada-salida es dada por $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\mathcal{N})$. Maximizar $I(X;Y)$ es equivalente a maximizar $H(Y) = H(X+\mathcal{N})$ sujeto a $\mathrm{Tr}\,(\Sigma_X) \leq \mathcal{P}$. Fijando un Σ_X, la propiedad [P'5]*b* de la entropía diferencial implica que $H(Y)$ sea máxima si y solamente si Y es gaussiana, es decir si y solamente si X est gaussiana, dando $I(X;Y) = \frac{1}{2}\log|\Sigma_X + \Sigma_{\mathcal{N}}| - \frac{1}{2}\log|\Sigma_{\mathcal{N}}|$. Tomando en cuenta el límite de potencia, hace falta maximi-

zar $|\Sigma_X + \Sigma_\mathcal{N}|$ sujeto a $\operatorname{Tr}\Sigma_X \leq \mathcal{P}$ y $\Sigma_X \geq 0$ simética lo que no es trivial. Se encuentra el enfoque permitiendo solucionar el problema en (Cover & Thomas, 2006, Sec. 9.4). Sea U, matriz ortogonal ($UU^t = U^tU = I$) de los autovectores de la matriz $\Sigma_\mathcal{N} \geq 0$ simética [116], de columnas u_i ordenadas tal que los autovalores corespondientes $\lambda_i^\mathcal{N}$ sean en orden creciente, *i. e.,*

$$\Sigma_\mathcal{N} = U \operatorname{diag}\left(\begin{bmatrix}\lambda_1^\mathcal{N} & \cdots & \lambda_d^\mathcal{N}\end{bmatrix}^t\right) U^t \qquad \text{con} \qquad 0 \leq \lambda_1^\mathcal{N} \leq \cdots \leq \lambda_d^\mathcal{N},$$

donde diag es la matriz diagonal teniendo los λ_i en su diagonal (ver notaciones). Sea $R_X = U^t\Sigma_X U$. Es sencillo ver que $|\Sigma_X + \Sigma_\mathcal{N}| = |R_X + \Lambda_\mathcal{N}|$ (de $|AB| = |A|\,|B|$) y que $\operatorname{Tr}\Sigma_X = \operatorname{Tr} R_X$ (de $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$). Entonces, el problema se reduce a maximizar $|R_X + \Lambda_\mathcal{N}|$ sujeto a $\operatorname{Tr} R_X \leq \mathcal{P}$ y $R_X \geq 0$ simética. La desigualdad de Hadamard ya evocada da $|R_X + \Lambda_\mathcal{N}| \leq \prod_i (R_X + \Lambda_\mathcal{N})_{i,i} = \prod_i \left((R_X)_{i,i} + \lambda_i^\mathcal{N}\right)$ donde $(\cdot)_{i,i}$ denota la componente i,i de la matriz, con igualdad si y solamente si R_X es diagonal: para maximizar $|R_X + \Lambda_\mathcal{N}|$, R_X debe ser diagonal (dada una diagonal, se alcanza el máximo si los otros términos son nulos). Es decir que la base que diagonaliza $\Sigma_\mathcal{N}$ debe diagonalizar también Σ_X. Sean λ_i^X los términos diagonales de R_X: queda que maximizar $\prod_i \left(\lambda_i^X + \lambda_i^\mathcal{N}\right)$ sujeto a $\sum_i \lambda_i^X \leq \mathcal{P}$ y $\lambda_i^X \geq 0$. Este problema de optimización sujeto a una desigualdad se resuelva con el enfoque de Karush-Kuhn-Tucker [117] (KKT) (Miller, 2000; Cambini & Martein, 2009), dando $\lambda_i^X = \left(\lambda - \lambda_i^\mathcal{N}\right)_+$ con $(\cdot)_+ = \max(\cdot, 0)$ y λ tal que $\sum_i \left(\lambda - \lambda_i^\mathcal{N}\right)_+ = \mathcal{P}$. En conclusión, la capacidad es dada por

$$C = \frac{1}{2}\log\left(\frac{|\Sigma_\mathcal{N} + \Sigma_X|}{|\Sigma_\mathcal{N}|}\right) \qquad \text{con} \qquad \Sigma_X = U \operatorname{diag}\left(\begin{bmatrix}\left(\lambda - \lambda_1^\mathcal{N}\right)_+ & \cdots & \left(\lambda - \lambda_d^\mathcal{N}\right)_+\end{bmatrix}^t\right) U^t,$$
$$\lambda \text{ tal que } \sum_i \left(\lambda - \lambda_i^\mathcal{N}\right)_+ = \mathcal{P}$$

alcanzada por X gaussiano de matriz de covarianza Σ_X así construida.

La última condición se resuelva a través de lo que es conocido como “llenado de agua” (water-filling en inglés), illustrado figura Fig. 2-47. El principio es parecido a tener niveles $\lambda_i^\mathcal{N}$ representando las potencias del ruido (en la base que diagonaliza la matriz de covarianza), y de “llenar con agua” hasta un nivel λ tal que el “volumen” añadido vale $\mathcal{P}$; en cada $\lambda_i^\mathcal{N}$ se ha añadido el λ_i^X (Cover & Thomas, 2006, Sec. 9.4).

En el caso escalar, se obtiene

$$C = \frac{1}{2}\log\left(1 + \frac{\mathcal{P}}{\sigma_\mathcal{N}^2}\right),$$

---

[116] Se recordará de que $A \geq 0$ significa que A es definida no negativa.

[117] Se introduce el factor de Lagrange y se maximiza $\prod_i \left(\lambda_i^X + \lambda_i^\mathcal{N}\right) + \eta \sum_i \lambda_i^X$. Eso da $\lambda_i^X + \lambda_i^\mathcal{N} = \lambda$ constante si λ es tal que se satisfaga la positividad de λ_i^X, y $\lambda_i^X = 0$ sino. En otras palabras, $\lambda_i^X = \left(\lambda - \lambda_i^\mathcal{N}\right)_+$ con λ el factor de Lagrande después de una reescritura. Queda que maximizar los λ_i^X para maximizar $|R_X + \Lambda_\mathcal{N}|$, es decir tomar λ lo más grande que se puede, pero satisfaciendo $\sum_i \lambda_i^X \leq \mathcal{P}$, *i. e.,* alcanzando la igualdad.

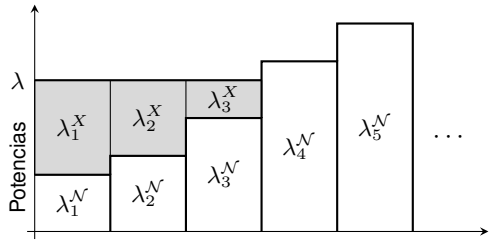


Figura 2-47: Principio del "water-filling" para obtener los λ_i^X satisfaciendo el vínculo de potencia límite y permitiendo de construir Σ_X a partir de la matriz diagonal de los λ_i^X y la base que diagonaliza la covarianza $\Sigma_{\mathcal{N}}$ del ruido. La zona en grise representa esquemáticamente $\mathcal{P}$.

donde $\frac{\mathcal{P}}{\sigma^2_{\mathcal{N}}}$ es conocido como relación señale-ruido [118]

En (Cover & Thomas, 2006; Rioul, 2007) por ejemplo, se dan otros ejemplos de canal de comunicación en el contexto continuo (entrada X_t siendo una señal/proceso, canal filtrando, canal con retroacción (o feedback), etc.).

2.5.3 Codificación entrópica sin perdida

El problema de codificación de fuente puede presentarse de la manera siguiente (Cover & Thomas, 2006, cap. 5) o (Rioul, 2007, cap. 13). Sea un proceso aleatorio $\{X_t\}_{t\in\mathbb{Z}}$, supuesto estacionario, llamado *fuente*, donde los X_t toman sus valores sobre un alfabeto discreto finito no necesariamente real (X puede tomar cualquier etiqueta)

$$\mathcal{X} = \{x_1\,,\,\ldots\,,\,x_\alpha\} \qquad \text{alfabeto fuente,}$$

de distribución p_X. A cada posible secuencia [119] $s_1\cdots s_n \in \mathcal{X}^n$ de letras de $\mathcal{X}$, se quiere asignar un código $c(s_1\cdots s_n)$ de letras de un alfabeto discreto finito,

$$\mathcal{C} = \{\mathfrak{c}_1\,,\,\ldots\,,\,\mathfrak{c}_d\} \qquad \text{alfabeto código.}$$

El código es dicho d-*ario*. Por ejemplo, se puede asignar un código $c(x_i) = \mathfrak{c}_{i,1}\cdots\mathfrak{c}_{i,l_i} \in \mathcal{C}^{l_i}$ a cada símbolo x_i, código llamado *palabras códigos*, y a secuencias $s_1\cdots s_n$ la concatenación de las palabras

---

[118] Esta formula es muy parecida a la de Shannon, Laplume, o Clavier (Shannon, 1948; Laplume, 1948; Clavier, 1948) (ver también (Cover & Thomas, 2006, Sec. 9.3) o (Rioul, 2007, Sec. 11.2)). De hecho, si se considera símbolos mandados durante T segundos cada uno (símbolos puestos en forma para dar una señal analógica) usando una banda de transmisión B, por el teorema de Nyquist $B = \frac{1}{2T}$ (caso límite). Si el ruido es blanco en la banda B, de densidad espectral de potencia por unidad de frecuencia igual a N_0, para un símbolo la relación señal-ruido se escribe $\frac{\mathcal{P}}{N_0 B}$. Además, se calcula en general la capacidad por unidad de tiempo es decir la capacidad por símbolo divido por $T = \frac{1}{2B}$, *i. e.,* $C = B\log\left(1+\frac{\mathcal{P}}{N_0 B}\right)$ por segundos, lo que es precisamente la capacidad calculdada por Shannon. Esta es a veces conocida como formula de Shannon-Hartley.

[119] Por abuso de escritura una cadena de n símbolos puede ser vista como un n-uplet.

códigos correspondiente a cada símbolo, *i. e.,* el código $c(s_1)\cdots c(s_n)$. En el sistema Moorse por ejemplo, $\mathcal{C}$ consiste en un punto, una barra, una espacio entre letras, un espacio entre palabras. En una computadora en general todo se codifica en bits $\mathcal{C}=\{0\,,\,1\}$. Más formalmente, sean

$$F_{\mathcal{X}}=\bigcup_{k=0}^{\infty}\mathcal{X}^k \qquad \text{y} \qquad F_{\mathcal{C}}=\bigcup_{k=0}^{\infty}\mathcal{C}^k,$$

unión de secuencias de k letras de $\mathcal{X}$ y $\mathcal{C}$ respectivamente. Una codificación de fuente consiste en una función de $F_{\mathcal{X}}$ dentro de $F_{\mathcal{C}}$. En lo que sigue, nos concentramos en códigos definidos para bloques de símbolos de tamaño $m\geq 1$:

$$\begin{array}{rccl} c_m: & \mathcal{X}^m & \to & F_{\mathcal{C}} \\ & x & \mapsto & c_m(x)\in\mathcal{C}^{l_{c_m}(x)} \end{array},$$

donde $l_{c_m}(x)\in\mathbb{N}^*$ es el *largo* de la palabra código $c_m(x)$, y

$$\forall\, n\geq 1,\quad \forall\, s_1\cdots s_n\in\mathcal{X}^{nm},\quad c_m(s_1\cdots s_n)\equiv c_m(s_1)\cdots c_m(s_n),$$

lo que es llamado *extensión del código*. En lo que sigue, se escribirá $c\equiv c_1$.

Una manera ingenua de codificar consiste a apoyarse sobre la descomposición de base d de un entero, *i. e.,* para $1\leq i\leq\alpha$ se puede escribir de manera única $i-1=(i_0-1)+(i_1-1)d+\cdots+(i_K-1)d^K$ donde $K=\left\lceil\log_d|\mathcal{X}|\right\rceil$ y $1\leq i_k\leq\alpha$. Entonces, se puede asignar la palabra código $c(x_i)=\mathfrak{c}_{i_0}\cdots\mathfrak{c}_{i_k}$ al símbolo x_i. Haciendo eso, cada palabra código tieno el mismo largo. Pero, es más económico hacer una codificación dicha de largos variables, teniendo en cuenta las probabilidades de aparición de cada x_i. Implicitamente, es la idea del código de Moorse, que asigna un punto o series de puntos o código pequeño a las letras muy frecuentes (ej. un punto para el 'e', dos puntos para el 'i', etc.), y barras o combinaciones largas a las letras que son raras (ej. bara-bara-punto-bara para el 'q' o cinco baras para el '0'). Dicho de otra manera, el código ingenuo sería "eficaz" para x_i apareciendo con las mismas frecuencias/probabilidades.

En los códigos de largos variables (incluyendo el código ingenuo), volviendo a c_m, existen varios tipos de códigos. Un código es dicho *no singular* si c_m es inyectiva: a cada $x\in\mathcal{X}^m$ corresponde una palabra código única. Esta propiedad es un requisito que parece obvio querer para un código. Pero no es suficiente para poder decodificar un mensaje, compuesta por una secuencia de palabras código. Lo importante en este caso es poder decodificar la secuencia sin ambigüedad: un código est dicho *descifrable* o *a decodificación única* (o sin perdida) si todas las extensiones son no singulares.

Ejemplo 2-29 (Código no singular, pero no decifrable)**.** *Sean, sean* $\mathcal{X}=\{\aleph\,,\,\beth\,,\,\gimel\,,\,\daleth\}$, $\mathcal{C}=\{0\,,\,1\}$ *y* $c(\aleph)=0$, $c(\beth)=00$, $c(\gimel)=1$, $c(\daleth)=01$ $(m=1)$*. El código es no singular, pero no descifrable. La secuencia* 0010 *puede provenir de* $\aleph\gimel\aleph\aleph$, *de* $\aleph\daleth\aleph$ *o de* $\aleph\gimel\beth$.

Obviamente, se requiere en general de un código que sea descifrable. Frecuentemente, se requiere también poder decodificar sobre la marcha, sin esperar de medir toda la secuencia codificada: es lo que se llama *código instantaneo*.

Ejemplo 2-30 (Código decifrable, pero no instantaneo)**.** *Sea el código* $c(\aleph) = 00,\ c(\beth) = 10,\ c(\gimel) = 11,\ c(\daleth) = 110$*. Este código es descifrable, pero no instantaneo. Considera la secuencia* 0011011 *y marcha sobre ella.* 0 *no es una palabra código;* 00 *es y sin ambigüedad proviene de un* $\aleph$ *(no hay otras palabras empezando por* 00*); luego* 1 *no es una palabra, y* 11 *es una palabra código, pero se necesita adelantar para saber si viene de un* $\gimel$ *o de un* $\daleth$*; la letra siguiente siendo un* 0*, todavía no se puede concluir si* 110 *vino de* $\gimel$ *y algó o* $\daleth$*. Al final, con* 1101*, se sabe que se tuvimos un* $\daleth$ *porque ninguna palabra código empieza por* 01*. Al final, sin ambigüedad el antecedente de la secuencia binaria era* $\aleph\daleth\gimel$*. Pero se necesitó marchar sobre toda la secuencia antes de decodificar.*

Obviamente, un código instantaneo es tal que ninguna palabra código es prefijo de una otra, *i. e.,* si $c_m(x)$ es una palabra código, las otras palabras código no pueden empezar con $c_m(x)$; el código es también dicho *libre de prefijo*. Estas distinciones estan ilustradas en la figura Fig. 2-48 (ver (Cover & Thomas, 2006, cap. 5)).

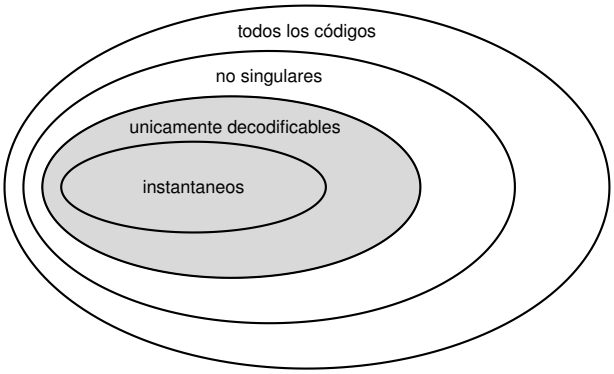


Figura 2-48: Clases de códigos. Los códigos contienen la clase de los no singulares. La misma contiene la clase de los códigos descifrables. Ella contiene los códigos instantaneos. En grise se representan las clases de códigos sin perdida a lo cuales se dedica esta sección.

Además de la decodificación sin ambigüedad, una caracterización importante del código es la taza de codificación [120]

$$R_{c_m} = \frac{\log_d \left(\sum_{x\in\mathcal{X}^m} l(x)P(X=x)\right)}{m},$$

donde X representa una secuencia de m variables X_t. El argumento del logaritmo (de base adecuada al cardinal de $\mathcal{C}$) es el *largo promedio* del código. Por ejemplo, para $d = 2$, R_{c_m} es el número de bits promedio del código por símbolo.

En general, se quiere minimizar R_{c_m} (compresar el mensaje a mandar), lo que puede ser contradictorio con la necesidad de añadir redundancia para no perder información durante una transmisión. En lo que sigue, nos concentramos en el problema de compresión, sin tener en cuenta el paso de

[120] En (Rioul, 2007) por ejemplo, se define esta taza suponiendo que cada secuencia fuente es codificada por el mismo número de bits. La taza es entonces el número de bits por símbolo.

transmisión de mensajes codificados en un canal. Minimizar la taza es equivalente a minimizar el largo promedio. Además, se puede focalisarse en $m = 1$; todo se extiende sencillamente a $m > 1$.

La meta de la compresión es entonces construir un código c, descifrable, que minimizar el largo promedio

$$L(c) = \sum_{x \in \mathcal{X}} p_X(x)\, l(x).$$

Antes de ir más adelante, hace falta traducir en ecuación el vínculo de que c sea descifrable. Eso es dado por la desigualdad de Kraft-McMillan (Kraft Jr, 1949; McMillan, 1956; Karush, 1961) [121]

Teorema 2-79 (Desigualdad de Kraft-McMillan)**.** *Los largos $l_c(x)$ de las palabras código de un código c descifrable deben satisfacer la desigualdad*

$$\sum_{x \in \mathcal{X}} d^{-l_c(x)} \leq 1.$$

Recíprocamente, para cada conjunto de enteros $\{\ell_x\}_{x \in \mathcal{X}}$ satisfaciendo esta desigualdad, es posible de construir un código descifrable con $l_c(x) = \ell_x$.

Demostración. Para cualquier $k \geq 1$ y cualquiera cadena $s = s_1 \cdots s_k \in \mathcal{X}^k$, la extensión del código, $c_k(s_1 \cdots s_k) = c(s_1) \cdots c(s_k)$ satisface $l_{c_k}(s) = \sum_{i=1}^{k} l_c(s_i)$. Entonces

$$\left(\sum_{x \in \mathcal{X}} d^{-l_c(x)} \right)^k = \sum_{\bar{x} \in \mathcal{X}^k} d^{-l_{c_k}(\bar{x})} = \sum_{m=1}^{k\, l_c^{\max}} \#(m)\, d^{-m},$$

re-escribiendo la segunda suma, agrupando los términos de mismo largos, donde $\#(m)$ es el número de códicos de $\mathcal{X}^k$ teniendo el largo m y $l_c^{\max} = \max_{x \in \mathcal{X}} l_c(x)$ es el largo mayor. c siendo descifrable, c_k debe ser inyectiva, imponiendo $\#(m) \leq d^m$ (no hay más palabras de largo m que el cardinal de $\mathcal{C}^m$), dando inmediatemente que necesariamente

$$\forall\, k \in \mathbb{N}^*, \quad \sum_{x \in \mathcal{X}} d^{-l_c(x)} \leq \left(k\, l_c^{\max}\right)^{\frac{1}{k}} \quad \Leftrightarrow \quad \sum_{x \in \mathcal{X}} d^{-l_c(x)} \leq \min_{k \in \mathbb{N}^*} \left(k\, l_c^{\max}\right)^{\frac{1}{k}}.$$

Un estudio rápido de $u \mapsto \left(u\, l_c^{\max}\right)^{\frac{1}{u}}$ para $u \geq 1$ y teniendo en cuenta de que $l_c^{\max} \leq 1$ permite concluir que el mínimo es igual a 1, terminando la parte directa del teorema.

Recíprocamente, sea $\{\ell_x\}_{x \in \mathcal{X}}$ un conjunto de enteros satisfaciendo la desigualdad de Kraft-McMillan. Se puede agrupar los largos iguales y clasificarlos. Sea n_ℓ los números de largos igual a $\ell = 1, \ldots, \ell^{\max} \leq \alpha$. Consideramos ahora un arbol empezando con una raíz, correspondiente a un largo 0, que se divide en d ramas, correspondiente a los largos iguales a 1; a cada nudo se asocian las letras $\mathfrak{c}_1, \ldots, \mathfrak{c}_d$. Esto nudos se dividen cada uno en d otras ramas, y los nudos de "padre" $\mathfrak{c}_i$ se

[121] Esta desigualdad fue probabada por L. G. Kraft para códigos instantaneos en su tesis de maestria (Kraft Jr, 1949). Luego, fue extendida a los códigos descifrables por B. McMillan (McMillan, 1956) (en una nota de pie de pagina de su papel, atribua esta observación a J. L. Doob hecha oralemente durante una escuela de verano en Ann Arbor, MI en agosto 1955).

va a asociar las palabras códigos $\mathfrak{c}_i\mathfrak{c}_1, \ldots, \mathfrak{c}_i\mathfrak{c}_\alpha$, etc. Este arbol, conocido como arbol de Kraft, es ilustrado en la figura Fig. 2-49 para $d = 2$ y $\mathcal{C} = \{0\,,\,1\}$. Claramente, $n_1 \leq d$ si no $n_1\, d^{-1} > 1$ y los largos no podrían satisfacer la desigualdad de Kraft-McMillan. El principio es entonce de asociar a los n_1 (posiblemente igual a 0) largos iguales a 1 unos nudos con las palabras código asociadas de largo 1 (primera profundez de ramas) y de prohibir todos las ramas de padre los nudos seleccionados (lineas punteadas en la figura Fig. 2-49). Estos nudos son llamados *hojas* (no hay ramas). En la capa de "hijos" de profundez/largos 2, quedan $d^2 - n_1\, d$ nudos (accessibles) que se pueden dividir en ramas. Nuevamente, $n_2 \leq d^2 - n_1\, d$ sino tendríamos $n_1\, d^{-1} + n_2\, d^{-2} > 1$, incompatible con la desigualdad de Kraft-McMillan. Se puede asociar a los n_2 largos iguales a 2 unos nudos con las palabras código asociadas de largo 2 (segunda profundez), y de prohibir que salen de estos nudos nuevas ramas (son entonces hojas en la segunda profundez), etc. Haciendo así, se asocia un código c de largos $l_c(x) = \ell_x$ que aparece libre de prefijo, es decir instantaneo. Entonces, este código es también descifrable. □

A este punto, se menciona los hechos siguientes

- Los largos de un código descifrable satisfacen la desigualdad de Kraft-McMillan, pero con el conjunto de largos correspondientes se puede siempre construir un código instantaneo. Claramente, se puede buscar un código de largo promedio mínimo en los códigos instantaneo, sin perdida de optimalidad (buscar en la clase más amplia de los descifrable no permite bajar el largo promedio).
- En los códigos libres de prefijo, si se fija el número de hojas (última profundez) borradas contruyendo un código, este vale $\sum_{i=1}^{\ell^{\text{máx}}} n_i\, d^{\ell^{\text{máx}}-i} = \sum_{x\in\mathcal{X}} d^{\ell^{\text{máx}}-l_c(x)}$. Es necesariamente menor que el número total $d^{\ell^{\text{máx}}}$ de hojas, lo que prueba el teorema para los códigos instantaneos (Kraft Jr, 1949; Karush, 1961).
- El teorema se generaliza obviamente para codificar una fuente (discreta) con un número infinito de estados, tomando el límite $\alpha \to \infty$.
- Si se conocen los largos óptimos, es suficiente para poder construir un código libre de prefijo.

El formalismo dado, se va a ver ahora reaparecer la entropía de Shannon como cota de la codificación de fuente sin perdida:

Teorema 2-80 (Cota inferior de códigos descifrables)**.** *Para cualquier código c descifrable de la fuente X, su largo promedio es acotado por debajo por la entropía de Shannon de base d de X,*

$$L(c) = \sum_{x\in\mathcal{X}} p_X(x)\, l_c(x) \;\geq\; H_d(X).$$

Demostración. Sea $q(x) = \frac{d^{-l_c(x)}}{\sum_{x\in\mathcal{X}} d^{-l_c(x)}}$, siendo una distribución de probabilidad por construcción. Escribiendo $l_c(x) = \log_d d^{-l_c(x)}$, se puede expresar el largo promedio de la forma

$$L(c) = -\sum_{x\in\mathcal{X}} p_X(x)\, \log_d d^{-l_c(x)} = -\sum_{x\in\mathcal{X}} p_X(x)\, \log_d q(x) - \log_d \sum_{x\in\mathcal{X}} d^{-l_c(x)}.$$

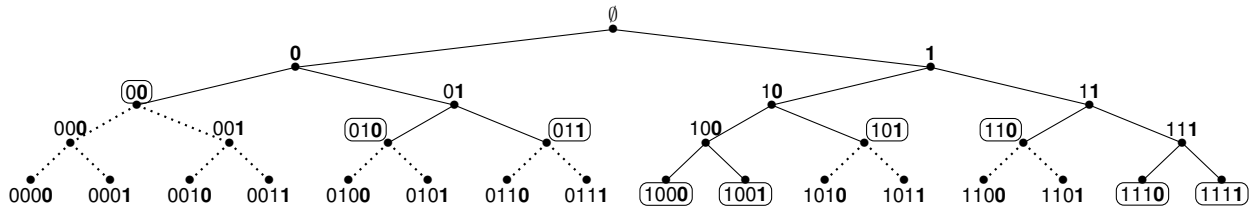


Figura 2-49: Arbol de Kraft en el caso binario ($d = 2$). La raíz, de código $\emptyset$ de largo 0, se divide en dos ramas, de códigos respectivamente 0 y 1 (profundez 1). Cada nudo de esta profundez se divide en dos ramas (profundez dos), dando cuatros nuevos nudos con los códigos 00 y 01 de padre 0, y 10 y 11 de padre 1. Etc. En cada nudo de esta figura, en el código, se marca en negrita la letra correspondiente al bit añadido al código padre. Para hacer un código libre de prefijo, una vez que un nudo es seleccionado para ser una palabra código (encuadrado en la figura), no puede tener nudos "hijos" siendo también una palabra código: se boran las ramas saliendo de un nudo-palabra código (ramas punteadas).

Notando que $-\log_d q = \log_d\left(\frac{p_X}{q}\right) - \log_d p_X$ se obtiene

$$L(c) = H_d(X) + D_{\mathrm{kl},d}\left(p_X \| q\right) - \log_d \sum_{x\in\mathcal{X}} d^{-l_c(x)}.$$

El resultado proviene de la positividad de la divergencia de Kullback-Leibler y de la desigualdad de Kraft-McMillan (el argumento del logaritmo siende menor que 1). □

Este resultado significa que la taza de compresión sin perdida no puede ser más bajo que el contenido informacional de la fuente. En este sentido, H tiene realmente un sabor de información sobre la fuente X.

La entropía aparece también en la cota superior del código óptimo:

Teorema 2-81 (Cota superior del código descifrable óptimo)**.** *El largo promedio L^{opt} del código c^{opt} descifrable, de largo promedio mínimo es acotado por arriba por la entropía de Shannon de base d de X más un dit (1 símbolo de $\mathcal{C}$),*

$$L^{\mathrm{opt}} \;<\; H_d(X) + 1.$$

Demostración. Por eso, empezamos por buscar los largos óptimos, solución de la optimización

$$\min \sum_{x\in\mathcal{X}} p_X(x)\, l(x) \qquad \text{sujeto a} \qquad \sum_{x\in\mathcal{X}} d^{-l(x)} \;\leq\; 1.$$

Escribiendo $l_c(x) = \log_d d^{-l_c(x)}$, se puede expresar el largo promedio de la forma

$$L(c) = -\sum_{x\in\mathcal{X}} p_X(x)\, \log_d d^{-l_c(x)} = -\sum_{x\in\mathcal{X}} p_X(x)\, \log_d q(x) - \log_d \sum_{x\in\mathcal{X}} d^{-l_c(x)}.$$

Olvidando que los $l_i \equiv l_c(x_i)$ son enteros, $L(c)$ es convexa con respecto a los l_i así que el vínculo, garantizando que el mínimo existe y es único. El problema se resuelva con el enfoque KKT [122], opti-

[122] Ver nota de pie 117 pagina 251.

mización con vínculos tipo desigualdades (Miller, 2000; Cambini & Martein, 2009), conduciendo a los "largos"

$$\widetilde{l}(x) = -\log_d p_X(x).$$

$\widetilde{l}(x)$ no es necesariamente entero, así que una posibilidad para volver a largos enteros puede ser de tomar la parte entera superior de $\widetilde{l}(x)$,

$$l(x) = \left\lceil -\log_d p_X(x) \right\rceil.$$

Obviamente el conjunto de largos satisface la desigualdad de Kraft-McMillan, así que se puede construir un código c^{sh} descrifrable con estos largos. De $l(x) < -\log_d p_X(x) + 1$ se obtiene

$$L^{\mathrm{opt}} \leq L\left(c^{\mathrm{sh}}\right) < H_d(X) + 1.$$

□

De

$$H_d(X) \leq L^{\mathrm{opt}} < H_d(X) + 1$$

se revela el rol fundamental de la entropía en la codificación de fuente sin perdida. La codificación es a veces dicha *codificación entrópica* y da un rol operacional a la entropía de Shannon. Se notará también que de la demostración precediente de que aparece un código particular a través de los $\left\lceil -\log_d p_X(x) \right\rceil$:

Definición 2-71 (Código de Shannon)**.** *Un código c^{sh} de una fuente X, de largos $l^{\mathrm{sh}}(x) = \left\lceil -\log_d p_X(x) \right\rceil$, libre de prefijo (construido sobre el arbol de Kraft) es llamado código de Shannon.*

Obviamente, también

$$H_d(X) \leq L\left(c^{\mathrm{sh}}\right) < H_d(X) + 1.$$

Al lo contrario de primer vista, un código de Shannon no es óptimo, como se lo puede ver con el ejemplo siguiente.

Ejemplo 2-31. *Sea $\mathcal{X} = \mathcal{C} = \{0\,,\,1\}$ y una fuente X tal que $p_X(0) = 0{,}999 = 1 - p_X(1)$. Los largos de Shannon van a ser $l^{\mathrm{sh}}(0) = 1$ y $l^{\mathrm{sh}}(1) = 10$, y el largo promedio vale $L\left(c^{\mathrm{sh}}\right) = 1{,}009$. Obviamente, un código óptimo es $c(x) = x$ de largos $l_c(x) = 1$ dando $L^{\mathrm{opt}} = 1$ bit.*

De hecho, volviendo al problema con largos virtualmente no enteros, el mínimo se alcanza para $\widetilde{l}(x) = -\log_d p_X(x)$, es decir que, los largos siendo enteros, se alcanza la cota mínima del código óptimo si y solamente si $-\log_d p_X(x)$. Una distribución satifaciendo esta condición es dicha d-ádica. Sin embargo, el código de Shannon es "competitivo" en el sentido de que:

Teorema 2-82 (Competitividad del código de Shannon)**.** *Sea X fuente sobre $\mathcal{X}$, de distribución p_X y c^{sh} el código de Shannon asociado sobre el alfabeto código $\mathcal{C} = \{\mathfrak{c}_1\,,\,\ldots\,,\,\mathfrak{c}_d\}$, de largos $l^{\mathrm{sh}}(x) = \left\lceil -\log_d p_X(x) \right\rceil$. Para cualquier código c descifrable y cualquier $k \geq 1$,*

$$P\Big(l^{\mathrm{sh}}(X) \geq l_c(X) + k\Big) \leq \frac{1}{d^{k-1}}.$$

Demostración. Por definición de la parte entera superior, $a+1 > \lceil a \rceil$, así que $\lceil a \rceil \geq b \Rightarrow a > b-1$. A continuación, de la implicación de eventos $(Y \geq a) \subset (Y \geq b-1)$ dando $P(A \geq a) \leq P(Y \geq b-1)$ y de la definición de un código de Shannon se obtiene

$$\begin{aligned} P\Big(l^{\mathrm{sh}}(X) \geq l_c(X)+k\Big) &\leq P\Big(-\log_d p_X(X) \geq l_c(X)+k-1\Big) \\ &= P\Big(p_X(X) \leq d^{-l_c(X)-k+1}\Big) \\ &= \sum_{x\in\mathcal{X}: p_X(x) \leq d^{-l_c(x)-k+1}} p_X(x) \end{aligned}$$

Pero, sumando sobre lo x tal que $p_X(x) \leq d^{-l_c(x)-k+1}$, se obtiene

$$P\Big(l^{\mathrm{sh}}(X) \geq l_c(X)+k\Big) \leq \frac{1}{d^{k-1}} \sum_{x\in\mathcal{X}: p_X(x) \leq d^{-l_c(x)-k+1}} d^{-l_c(x)} \leq \frac{1}{d^{k-1}} \sum_{x\in\mathcal{X}} d^{-l_c(x)}$$

(añadiendo términos positivos en la suma). La prueba se cierra notando que c siendo descifrable, l_c satisface la desigualdad de Kraft-McMillan. □

Este teorema traduce el hecho de que a pesar de que c^{sh} no sea óptimo, tomando cualquier otro código, incluyendo el óptimo, la probabilidad que $c^{\mathrm{sh}}(X)$ tenga un largo más grande que $c(X)+k$ decrece exponencialmente con k.

Ejemplo 2-32. *De manera general, con $d=2$ y para $k=9$, $P\Big(l^{\mathrm{sh}}(X) \geq l_c(X)+9\Big) \leq 0{,}391\,\%$. En particular, en el ejemplo 2-31, notando que sólo en la codificación de $x=1$ se puede tener $l^{\mathrm{sh}}(x) \geq l_c(x)+9$, si no se usa 1 bit, este resultado significa que la probabilidad de usar más de 1 bit con el código de Shannon es menor que $0{,}391\,\%$. De hecho, una palabra código de largo 10 aparece con una probabilidad $0{,}1\,\%$...*

En el problema de minimización, el hecho de que los largos deben ser enteros no permite solucionar explicitamente el problema de busqueda del código óptimo. Números investigadores contruyeron códigos, intentando probar de que eran óptimos (ver ej. (Shannon, 1948; Shannon & Weaver, 1964; Fano, 1949) por los primeros, y (Cover & Thomas, 2006, & ref.)). El código conocido como *código de Fano* [123] c^{fa} se basa sobre el hecho de que se alcanza la cota mínima para una distribución d-ádica.

Definición 2-72 (Código de Fano)**.** *El principio es de clasificar los estados de $\mathcal{X}$ para obtener las probabilidades clasificadas en orden decreciences ($p_X^{\downarrow}$). Luego, se divide $\mathcal{X}$ en d ensembles a lo más equiprobables que se puede (i. e., de probabilidad a lo más cerca de d^{-1}) y de asignar $\mathfrak{c}_i$ al conjunto i. Luego, se repite el proceso a cada sub-conjunto (para tener sub-conjuntos de probabilidades a lo más cerca de d^{-2}) y al subconjunto j del conjunto i se va a asignar le código $\mathfrak{c}_i\mathfrak{c}_j$, etc. Eso es ilustrado en la figura Fig. 2-50-(a).*

[123] A pesar de que sea diferente del de Shannon y que cada uno fueron hechos independientemente, a veces es conocido como código de Fano-Shannon, o aun Shannon-Fano-Elias (Cover & Thomas, 2006; Krajči, Liu, Mikeš & Moser, 2015).

Probar/mencionar que también

$$H(X) \leq L\left(c^{\mathrm{fa}}\right) < H(X) + 1.$$

Fijense de que no hay un único código de Fano o de Shannon (tal como no hay un óptimo único). Por exemple, hacer una permutacion de los $\mathfrak{c}_i$ da los mismos largos y el mismo largo promedio sin cambiar el aspecto libre de prefijo. De la misma manera, en el arbol de Kraft, en cada profundez se puede permutar los símbolos asociados a las hojas de esta profundez sin cambiar el aspecto libre de prefijo y sin que cambien los largos $l(x_i)$ (y entonces con el mismo largo promedio).

Una solución para construir un código óptima fue propuesta por Huffman en 1951-1952 (Huffman, 1952; Pigeon, 2003) [124]

Definición 2-73 (Código de Huffman)**.** *Suponemos que existe un* $\beta \in \mathbb{N}$ *tal que* [125] $\alpha = |\mathcal{X}| = d + \beta(d-1)$*. El algoritmo de Huffman consiste a construir un arbol donde cada nudo es asociado a un conjunto de símbolos fuente y una letra de* $\mathcal{C}$ *de la manera siguiente:*

1. *Clasificar las probabilidades en orden decrecientes: por cambio de escritura, llamamos* p_i *las probabilidades rearregladas y* x_i *los símbolo fuente correspondientes.*
2. *A cada* $x_i,\ \alpha - d + 1 \leq i \leq \alpha$*, associar un nudo y la letra "hijo"* $\mathfrak{c}_i$*.*
3. *Crear* d *ramas saliendo de un nudo padre hasta los* d *nudos* $x_i,\ \alpha - d + 1 \leq i \leq \alpha$*.*
4. *Crear un nuevo conjunto de símbolos fuente* $\widetilde{x}_i = x_i, 1 \leq i \leq \alpha - d$ *de probabilidades respectivas* $\widetilde{p}_i = p_i$ *y* $\widetilde{x}_{\alpha-d+1} = \{x_j,\ \alpha - d + 1 \leq j \leq \alpha$ *de probabilidad* $\widetilde{p}_{\alpha-d+1} = p_{\alpha-d+1} + \ldots + p_\alpha$*. El útimo "super-símbolo" fuente es asociado al nudo padre de la etapa 3.*
5. *Si quedan más de un (super-)símbolo fuente, volver a la etapa 1 con* $p \equiv \widetilde{p}$ *y* $x \equiv \widetilde{x}$*.*

Como descrito tratando del código usando el arbol de Kraft, $c^{\mathrm{huf}}(x_i)$ *se construye saliendo de la raíz del arbol así construido, agregando las letras del camino que llega hasta la hoja* x_i*. Eso es ilustrado en la figura Fig. 2-50-(b) en el caso binario.*

Se mencionará que a cada etapa, el nuevo conjunto de super-símbolos fuente contiene exactamente $d-1$ símbolos menos que a la etapa precediente. Así, con $\alpha = d + \beta(d-1)$ el algoritmo tiene exactamente $\beta + 1$ bucles y en cada profundez no hay ningún nudo vacio en el sentido que o es una

[124] De hecho, Huffman fue estudiantes de maestria de Fano, trabajando en el MIT. Su tesis era de probar que el código de Fano era óptimo: Huffman propusó su propio código, andando al revés del enfoque de Fano, y demostró que era óptimo (Stix, 1991).

[125] Si no, se puede elegir $\beta = \left\lceil \frac{\alpha - d}{d-1} \right\rceil$, y completar $\mathcal{X}$ con $d + \beta(d-1) - \alpha$ símbolos fuente fictivos de probabilidades ceros, lo que no va a cambiar ni la entropía, ni el largo promedio del códico aferente.

hoja, o es un nudo padre/prefijo (quedarán exactamente d nudos a agregar a la raíz en la última etapa). Por exemplo, con $d = 3$ si tuvieramos $\alpha = 4$, en la segunda etapa tendriamos 2 estados a juntar, dando un código de largos $2, 2, 2, 1$. Empezando la primera etapa con la asociaciación de 2 estados, es decir 3 teniendo en cuenta un estado fictivo ($\alpha = 5$, $\beta = 1$) van a quedar 3 estados en la segunda etapa, dando un código de largos $2, 2, 1, 1$, es decir de largo promedio más pequeño.

Teorema 2-83 (Óptimalidad del código de Huffman)**.** *El algoritmo de Huffman da un código c^{huf} de largo promedio mínimo en la clase de los códigos descrifrables y los libre de prefijo (se recordará que con los largos de códigos descifrables, siempre se puede construir un código libre de prefijo), es decir $L^{\text{opt}} = L\left(c^{\text{huf}}\right)$.*

Demostración. Una prueba es dada por ejemplo en (Cover & Thomas, 2006, Sec. 5.8) en el caso binario, pero la extensión para $d > 2$ es un poco más sutil. La prueba más general es dada por Huffman (Huffman, 1952) y se consigue también por parte en (Pigeon, 2003). Suponemos que $\beta \geq 1$ (sino, el resultado es obvio). Las etapas son

- Sean j, k dos indices. Si c^{opt} es un código óptimo, y c un código tal que $l(x_i) = l_i^{\text{opt}}, \quad i \neq j, k, \quad l_j = l_k^{\text{opt}} \quad \& \quad l_k = l_j^{\text{opt}}$, se obtiene $0 \leq L(c) - L^{\text{opt}} = \sum_i p_i \left(l_i - l_i^{\text{opt}}\right) = (p_j - p_k)\left(l_k^{\text{opt}} - l_j^{\text{opt}}\right)$. Entonces $p_j > p_k \Rightarrow l_j^{\text{opt}} \leq l_k^{\text{opt}}$.

- Sea η el número de símbolos fuente con un código de largo máximo $l_{\text{máx}}$ y $\eta' = \text{mín}(\eta, d)$. Del punto anterior, los η símbolos con palabra código de largo máximo son los de probabilidades más pequeñas.

- Como descrito antes, se puede permutar las letras códigos de una profundez del arbol de Kraft sin cambiar ni el aspecto libre de prefijo, ni el largo promedio. Se puede entonces considerar el código óptimo tal que los η' símbolos de probabilidades las más pequeñas tienen el mismo nudo padre, *i. e.,* solamente la última letra código cambia entre ellos.

- Suponemos que $\eta' = \eta < d$. Sea una "super-fuente" $\mathcal{X}^{(2)} = \left\{x_i^{(2)}\right\}_{i=1}^{\alpha-\eta'+1}$ con $x_i^{(2)} = x_i, \quad 1 \leq i \leq \alpha - \eta'$ de probabilidades respectivas $p(x_i)$ y $x_{\alpha-\eta'+1}^{(2)} \equiv \{x_i\}_{i=\alpha-\eta'+1}^{\alpha}$ de probabilidad $p_{\alpha-\eta'+1} + \cdots + p_\alpha$ (se "plegan" las η' hojas en un super-símbolo). La códificación óptima es entonces una codificación libre de prefijo de $\mathcal{X}^{(2)}$, "arbol raíz" del código óptimo, a la cual se añade una letra código $\mathfrak{c}_j$ diferente a cada símbolo del super-símbolo $x_{\alpha-\eta'+1}^{(2)}$. La profundez máxima del código arbol es $l_{\text{máx}} - 1$ y debe ser llena, en el sentido de que no debe tener un nudo que sea ni una hoja, ni un prefijo. En el caso contrario, se podría desplazar un símbolo de $x_{\alpha-\eta'+1}^{(2)}$ al nudo "vacio" de la profundez $l_{\text{máx}} - 1$, sin cambiar el aspecto libre de prefijo, pero ganando una letra código sobre un símbolo, *i. e.,* hacer un código libre de prefijo con un largo promedio menor. Sería contradictorio con la optimalidad del código inicial.

- Para códificar $\mathcal{X}^{(2)}$, se necesita por lo menos $\lceil \log_d(\alpha - \eta' + 1) \rceil$ profundez en el arbol raíz. En esta profundez (máxima en el caso optimista), hay $d^{\lceil \log_d(\alpha-\eta'+1) \rceil} \geq \alpha - \eta' + 1$ nudos. En la última

profundez pueden ser todos ocupados si y solamente si $d^{\lceil \log_d(\alpha-\eta'+1)\rceil} = \alpha - \eta' + 1$. En otras palabras, es posible si y solamente si existe un entero k tal que $\alpha - \eta' + 1 = d^k$, es decir, con $\alpha = d+\beta(d-1)$, que teniamos el entero $\beta = \frac{d^k-d}{d-1} + \frac{\eta'-1}{d-1}$. La primera fracción $\frac{d^k-d}{d-1} = d^{k-1}+\cdots+1$ siendo entera, β no puede ser entero con $\eta' < d$. En otros términos, necesariamente $\eta' = d$, *i. e.,* los d símbolos de probabilidad más debiles son el la última profundez y se puede elegir que compartent el mismo nudo padre.

- Sea $c^{\mathrm{opt},(1)}$ el código óptimo correspondiente a $\mathcal{X}$ y $c^{(2)}$ el código "padre" sobre $\mathcal{X}^{(2)}$ ($c^{\mathrm{opt},(1)}$ quitando la último letra código de los símbolos juntados, *i. e.,* con la raíz común de estos). De la misma manera, sea $c^{\mathrm{opt},(2)}$ un código óptimo sobre $\mathcal{X}^{(2)}$ y $c^{(1)}$ el que se obtiene deplegando el super-símbolo $x^{(2)}_{\alpha-d+1}$ en d hojas. De $L^{\mathrm{opt},(1)} = L\left(c^{(2)}\right) + p_{\alpha-d+1} + \cdots + p_\alpha$ (pasar de $\mathcal{X}^{(2)}$ a $\mathcal{X}$ se añade solo una letra palabra a los símbolos del super-símbolo) y $L\left(c^{(1)}\right) = L^{\mathrm{opt},(2)} + p_{\alpha-d+1} + \cdots + p_\alpha$ se obiene $\left(L^{\mathrm{opt},(1)} - L\left(c^{(1)}\right)\right) + \left(L^{\mathrm{opt},(2)} - L\left(c^{(2)}\right)\right) = 0$. Cada término entre parentesis siendo positivo, valen necesariamente cero (la suma de términos positivos vale cero si y solamente si todos son nulos). En conclusión, $c^{(2)}$ padre de $c^{\mathrm{opt},(1)}$ queda óptimo, $c^{(2)} \equiv c^{\mathrm{opt},(2)}$ (y $c^{(1)} \equiv c^{\mathrm{opt},(1)}$).

- Notando que $|\mathcal{X}^{(2)}| = \alpha - (\beta-1)(d-1)$, el razonamiento se propaga por inducción, pasando de $c^{\mathrm{opt},(k)}$ a $c^{\mathrm{opt},(k+1)}$ juntando los d super-símbolos de probabilidades más debiles, hasta tener un super-símbolo tendiendo todos los símbolos, $\left|\mathcal{X}^{(K)}\right| = 1$, raíz del arbol.

□

De esta prueba, se puede ver que

- Cada profundez siendo llena, los largos obtenidos van a saturar la desigualdad de Kraft-McMillan.

- Si $\frac{\alpha-d}{d-1}$ no es entero, en lugar de completar $\mathcal{X}$ con símbolos fictivos se puede empezar el algoritmo de Huffman juntando los $\alpha - d - \left\lfloor \frac{\alpha-d}{d-1} \right\rfloor (d-1) + 1$ símbolos fuentes de probabilidades más debiles en un super-símbolo, y luego hacer el bucle descrito (juntando por super-símbolos de d símbolos en cada bucle); en este caso, no se satura más la desigualdad de Kraft-McMillan, pero no es contradictorio con el punto anterior que contaba los largos de estados fictivos(que no codificamos en realidad).

- Obviamente, en el caso binario $d = 2$, no es necesario completar $\mathcal{X}$ por estados fuentes, o empezar con menos de d símbolos juntados (α es necesariamente de la forma $\alpha = d+\beta(d-1) = 2+\beta$).

- El algoritmo no permite conocer los largos de manera analítica en función de p_i, y tampoco el largo promedio. Se los pueden deducir solamente implementando el algoritmo (una vez que es construido el código). Era el caso también con el enfoque de Fano.

Volviendo al código ingenuo, sería óptimo (y equivalente a los de Fano y de Shannon) para una distribución uniforme. En este contexto, la entropía es $H_d(X) = \log_d |\mathcal{X}|$, precisamente la incerteza del enfoque de Hartley que corresponde a los números de dits necesarios para codificar (ingenuosamente) la fuente.

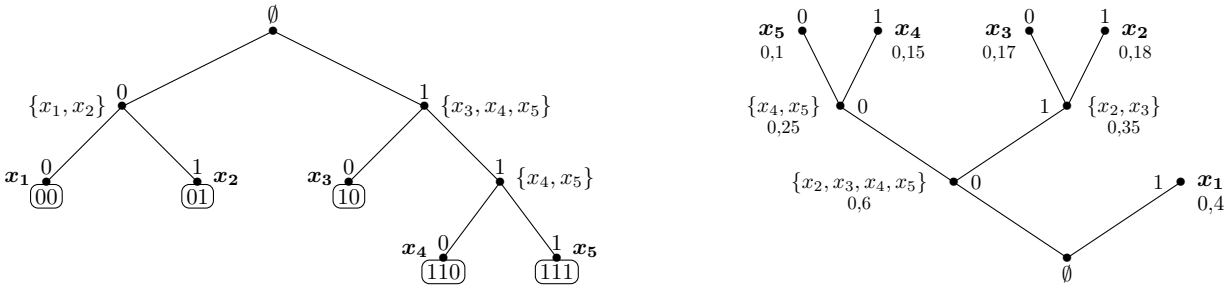


Figura 2-50: Construcción de un código binario sobre $\mathcal{C} = \{0\,,\,1\}$ asociado al vector de probabilidad $p_X = \begin{bmatrix} 0{,}4 & 0{,}18 & 0{,}17 & 0{,}15 & 0{,}1 \end{bmatrix}^t$ sobre el arbol de Kraft. En este caso, $H_2(X) \approx 2{,}1514$ (a): Enfoque de Fano, saliendo de la raíz. En cada nudo, se menciona el conjunto de símbolos que va a tener el código correspondiente (en negro cuando es un solo símbolo). Se pasa de una profundez a la otra dividiendo los conjunto en sub-conjuntos a lo más equiprobables. Esta construcción da el código $c^{\mathrm{fa}}(x_1) = 00$, $c^{\mathrm{fa}}(x_2) = 01$, $c^{\mathrm{fa}}(x_3) = 10$, $c^{\mathrm{fa}}(x_4) = 110$, $c^{\mathrm{fa}}(x_5) = 111$ de largo promedio $L\left(c^{\mathrm{fa}}\right) = 2{,}25$. (b): Enfoque de Huffman, saliendo de las hojas. En cada nudo, se menciona el correspondiente (i) conjunto de símbolos, (ii) $\mathfrak{c}_i$ de esta profundez/posición, (iii) la probabilidad asociada al conjunto. Se pasa de una profundez a la otra juntando los conjuntos menos probables en sobre-conjuntos. En negro son indicados los símbolos simples: van a tener el código agregando los de los nudos yendo de la raíz hasta las hojas. Esta construcción da el código $c^{\mathrm{huf}}(x_1) = 1$, $c^{\mathrm{huf}}(x_2) = 011$, $c^{\mathrm{huf}}(x_3) = 010$, $c^{\mathrm{huf}}(x_4) = 001$, $c^{\mathrm{huf}}(x_5) = 000$ de largo promedio $L^{\mathrm{opt}} = 2{,}2$.

Se notará de que, tratando de una fuente $\{X_t\}_{t\in\mathbb{Z}}$ de variables independientes, se puede codificar la fuente con un largo promedio arbitrariamente cerca de $H_d(X)$. El principio es de considerar vectores $\begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^t$ viviendo sobre $\mathcal{X}^n$, llamado *extensión de orden n de la fuente*, con un código descifrable (o libre de prefijo) de esta extensión; es llamado *codificación de la extensión de la fuente* pero no es necesariamente una extension de c. Así, $H_d(X_1, \ldots, X_n) \leq L^{\mathrm{opt},n} < H_d(X_1, \ldots, X_n) + 1$, es decir, de la independencia,

$$H_d(X) \;\leq\; \frac{L^{\mathrm{opt},n}}{n} \;<\; H_d(X) + \frac{1}{n} \quad \text{por símbolo}$$

(ver también (Rioul, 2007, cap. 13, teorema de Shannon)). Fijense que si $\lim_{n\to\infty} \dfrac{L^{\mathrm{opt},n}}{n} \to H(X)$, $\frac{L^{\mathrm{opt},n}}{n}$ no es necesariamente decreciente con respeto a n. Eso es descrito figura Fig. 2-51. Lo mismo puede ocurir con el código de Shannon **y lo de Fano**. Además, el cardinal del alfabeto extendido $\mathcal{X}^n$ crece exponencialmente con n, lo que no permite elegir un n muy grande.

Para codificar una fuente, que se haga el código óptimo de Fano, o de Shannon, hace falta usar la distribución de probabilidad de la fuente X. Prácticamente, es usual que no se la tiene. Frecuentemente,

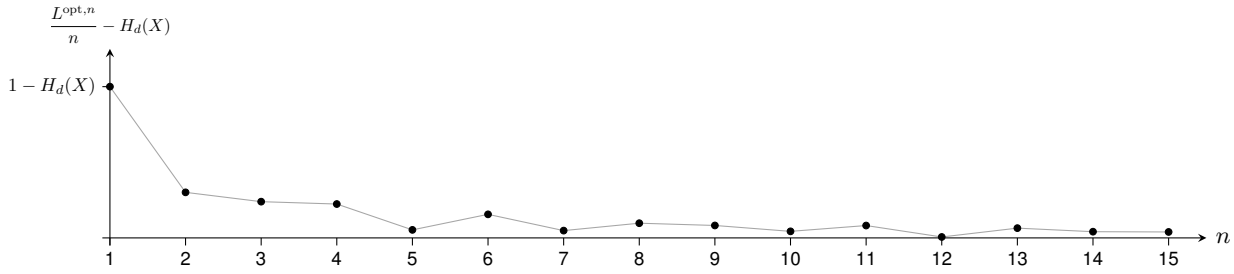


Figura 2-51: $\frac{L^{\mathrm{opt},n}}{n} - H_d(X)$ (puntos), diferencia entre el largo promedio óptimo por símbolo de las extensiones $\mathcal{X}^n$ de orden n de la fuente $\mathcal{X}$ y la cota inferior en función de n. La linea llena en grise sirve como guía. En esta ilustración se usa el ejemplo lo más simple con $d = 2$ y $p = \begin{bmatrix} 0{,}33 & 0{,}67 \end{bmatrix}^t$.

es estimada a partir de datos, o, dicho de otra manera, se códifica con una distribución que no es la distribución verdadera de la fuenta. Una pregunta que surge es de conocer lo que se pierde usando una distribución no adaptada (o "falsa"). La respuesta general no es obvia, pero tratando del código de Shannon se puede contestar:

Teorema 2-84 (Código falso de Shannon)**.** *Sea $c^{\mathrm{sh}}(p)$ el código de Shannon sobre el alfabeto código $\mathcal{C} = \{\mathfrak{c}_1, \ldots, \mathfrak{c}_d\}$ asociado a la distribución p. Sea X fuente sobre $\mathcal{X}$, de distribución p_X y q una distribución cualquiera (ej. estimada de p_X presupuesta...). Entonces el largo promedio $L_{c^{\mathrm{sh}}(q)}$ del código $c^{\mathrm{sh}}(q)$ aplicado a la fuente X satisface las desigualdades siguientes*

$$H_d(p_X) + D_{\mathrm{kl},d}\left(p_X \| \, q\right) \;\leq\; L_{c^{\mathrm{sh}}(q)} \;<\; H_d(p_X) + D_{\mathrm{kl},d}\left(p_X \| \, q\right) + 1.$$

Demostración. Por definición,

$$L_{c^{\mathrm{sh}}}(q) = \sum_{x \in X} p_X(x) \left\lceil -\log_d q(x) \right\rceil.$$

La desigualdad viene de $a \leq \lceil a \rceil < a + 1$ y escribiendo $-p_X \log_d q = -p_X \log_d p_X + p_X \log\left(\frac{p_X}{q}\right)$. □

Olvidando el posible extra dit (pensar a la códification por bloques), este teorema da una interpretación operacional a la entropía relativa, o divergencia de Kullback-Leibler. Esta cantidad cuantifica la perdida en término de largo promedio codificando con una distribución falsa. Dicho de otra manera, usando q en lugar de p_X, se usa la información de p_X porque se códifica la fuente X, pero suponiendo la distribucíon q, se piedre lo que representa la información relativa de p_X con respeto a la referencia (distribución supuesta) q.

Existen varios otros modos de codificar símbolos. En particular, con la meta de transmitir los símbolos codificados en un canal de comunicación, a veces no es oportuno de compresar drasticamente el mensaje. Existen por ejemplo codificaciones que permiten una corrección de error en la recepción. Pueden tomar en cuenta las caracteristicas del canal de transmisión. Estas consideraciones van más

allá de la ilustración de esta sección. El lector puede referirse a (Berlekamp, 1974; Gallager, 1978; Sayood, 2003; Cover & Thomas, 2006; Rioul, 2007) entre otros para tener más detalles sobre varios esquemas de codificación/compresión.

2.5.4 Gas perfecto

En el marco del gas perfecto

Va donner un lien avec Boltzmann

Feder Merhav IT'94 et lien avec discrimination; Vacisek en test de Gaussianite et cf plus loin avec generalises Gok75 etc

2.6 Entropías y divergencias generalizadas

Partout, voir convexite stricte, en 1, idem pour h monotone, etc. vis a vis des cas d'egalite avec les divergences.

A pesar de que la entropía de Shannon y sus cantidades asociadas demostraron sus potencias tan de un punto de vista descriptivo que en término de aplicaciones en la transmisión de la información y la compresión, varias nociones informacionales, tipo entropías o divergencias, aparecieron luego. En esta sección no se desarollará todos los enfoques ni todas las aplicaciones tan la literatura es importante. La meta es dar los caminos conduciendo a las generalizaciones de la entropía de Shannon por un lado, y de la divergencia de Kullback-Leibler por el otro lado. No son siempre vinculados, a pesar de que sea deseable que a cada entropía sean asociados nociones de entropías condicionales y relativas.

2.6.1 Entropías y propiedades

Si la entropía de Shannon fue el punto de salida fundamental en todo el desarollo de la teoría de la información, un poco más de una decada después de su papel clave y muy completo, Rényi propuso una medida generalizada (Rényi, 1961). Su punto de vista fue más matemático que físico o ingeniero. Retomó los axiomas de Fadeev (Fadeev, 1956, 1958; Khinchin, 1957) para probabilidades incompletas [126] $p = \begin{bmatrix} p_1 & \cdots & p_n \end{bmatrix}^t$, $p_i \geq 0$, $w_p = \sum_i p_i \leq 1$: (i) la invarianza de $H(p)$ por

[126] En esta sección, los p_i son componentes del vector p no necesariamente asociado a una variable aleatoria; Hay que entender de que $p_i = p_X(x_i)$ si son asociados a una variable aleatoria X.

permutación de os p_i, (ii) la continuidad de la incerteza elemental $H(p_i)$ (p_i visto como probabilidad incompleta), (iii) $H\left(\frac{1}{2}\right) = 1$, (iv) la aditividad $H(p \otimes q) = H(p) + H(q)$ donde $p \otimes q$ es el producto externo (ver notaciones), *i. e.,* probabilidad conjunta de dos variables independientes, y consideró en lugar de la recursividad un axioma dicho de valor promedio, axioma muy parecido a la recursividad. Para p y q probabilidades incompletas tales que $p \cup q = \begin{bmatrix} p_1 & \cdots & p_n & q_1 & \cdots & q_m \end{bmatrix}^t$ sea incompleta ($w_p + w_q \leq 1$), el axioma (v) es $H(p \cup q) = \frac{w_p\, H(p) + w_q\, H(q)}{w_p + w_q}$. Demostró que con (v) en lugar de la recursividad, el conjunto de axiomas conduce de nuevo a la entropía de Shannon. La generalización propuesta por Rényi era de generalizar el axioma (v) reemplazando la media aritmética por una media generalizada (v') $H^{\mathrm{r}}(p \cup q) = g^{-1}\left(\frac{w_p\, g\left(H^{\mathrm{r}}(p)\right) + w_q\, g\left(H^{\mathrm{r}}(q)\right)}{w_p + w_q}\right)$ con g estrictamente monótona y continua, llamado media *cuasi-aritmética, o cuasi-lineal, o de Kolmogorov-Nagumo*. De las propiedades de la media cuasi-aritmética (Nagumo, 1930; Kolmogorov, 1930, 1991; Hardy et al., 1952), eso es equivalente a buscar una entropía elemental $H^{\mathrm{r}}(p_i)$ y reemplazar la media aritmética $\sum_i p_i H^{\mathrm{r}}(p_i)$ por una media de Kolmogorov-Nagumo, $g^{-1}\left(\sum_i p_i g\left(H^{\mathrm{r}}(p_i)\right)\right)$. Rényi propusó la función de Kolmogorov-Nagumo $g_\lambda(x) = 2^{(\lambda-1)x}, \quad \lambda > 0, \quad \lambda \neq 1$, probando de que los axiomas (i)-(ii)-(iii)-(iv)-(v') se cumplen, conduciendo a la entropía de Rényi de un vector de probabilidad p,

$$H^{\mathrm{r}}_\lambda(p) = \frac{1}{1-\lambda} \log_2 \left(\sum_{i=1}^{n} p_i^\lambda \right).$$

Relaxando el axioma (iii), se puede elegir $g_\lambda(x) = a^{(\lambda-1)x}, \quad a > 0, \quad a \neq 1$; el logaritmo será de la base a cualquiera. En lo que sigue, usaremos $\log$ sin precisar la elección de base. Rényi nombró esta medida de incerteza *entropía de orden* λ. Notablemente,

$$H^{\mathrm{r}}_1(p) \equiv \lim_{\lambda \to 1} H^{\mathrm{r}}_\lambda(p) = H(p) \quad \text{entropía de Shannon.}$$

En otros términos, la clase de Rényi contiene como caso particular la entropía de Shannon. En su papel, Rényi introdujo una ganancia de información, parecida a una entropía relativa, probando que las solas entropías admisibles son la de Shannon y la que introdujo. Volveremos en la sección siguiente sobre esta entropía relativa, o divergencia de Rényi. Por axiomas, las propiedades [P1] (continuidad), [P2] (invarianza por permutación) y [P10] (additividad) de la entropía de Shannon se conservan entonces en el marco de Rényi y se pierde [P7] (recursividad), todavía por axiomas. Veremos luego la otras que se conservan o modifican en un marco más general.

Unos años después de Rényi, de la famosa escuela matemática checa, J. Havrda & F. Charvát en (Havrda & Charvát, 1967) (ver también (Vajda, 1968, en checo)) volvieron a los axiomas de Khintchin, para extender la entopia de Shannon, *i. e.,* considerando (i) la invarianza por permutación, (ii) la continuidad, (iii) la expansividad, (iv) $H^{\mathrm{hc}}(1) = 0$ y $H^{\mathrm{hc}}\left(\frac{1}{2}, \frac{1}{2}\right) = 1$, pero generalizando la recursividad por (v) $H^{\mathrm{hc}}(p_1, \ldots, p_n) = H^{\mathrm{hc}}(p_1, \ldots, p_{n-2}, p_{n-1} + p_n) + \lambda (p_{n-1} + p_n)^\lambda H^{\mathrm{hc}}\left(\frac{p_{n-1}}{p_{n-1}+p_p}, \frac{p_n}{p_{n-1}+p_p}\right), \quad \lambda > 0$ [127]. Con $\lambda = 1$ se recupera la recursividad estandar, pero con $\lambda \neq 1$ eso permite dar un peso di-

[127] En sus papel, lo imponen para cualquier par (p_i, p_j) sin imponer la invarianza por permutación, pero es equivalente a la

ferente a la incerteza del estado interno, *i. e.,* a las probabilidades que se juntan (la describen como clasificación refinada). Estos axiomas conducen necesariamente a la entropía (teorema 1 del papel)

$$H_\lambda^{\mathrm{hc}}(p) = \frac{1}{1-2^{1-\lambda}}\left(1-\sum_i p_i^\lambda\right)$$

que nombraron λ-*entropía structural*. De nuevo, relaxando el axioma (iv), se puede reemplazar en el coeficient $2^{1-\lambda}$ por $a^{1-\lambda}, \quad a>0, \quad a\neq 1$. De nuevo, aparece que la entropía de Shannon es un caso particular,

$$H_1^{\mathrm{hc}}(p) \equiv \lim_{\lambda\to 1} H_\lambda^{\mathrm{hc}}(p) = H(p) \quad \text{entropía de Shannon.}$$

Por axioma, se conservan las propiedades [P1] (continuidad) y [P6] (expansabilidad) de Shannon en este marco. Se probó también que se conserva la propiedad de concavidad con respecto a los p_i [P8], la de maximalidad [P5] alcanzada para una distribución uniforma (teorema 2). Aun que no aparece así en el papel, satisface la propiedad de Schur-concavidad [P9] (teorema 3). A pesar de que mencionan que H_λ^{hc} sea diferente de H_λ^{r}, es sencillo ver que hay un mapa uno-uno entre las dos entropías. Se mencionarán en un marco más general otras propiedades de esta entropía.

Independiente de Havrda & Charvát, de la escuela húgara de la teoría de la información, Z. Daróczy en (Daróczy, 1970) definó la entropía H^f a partir de una *función información* f satifaciendo (i) $f(0) = f(1)$, (ii) $f\left(\frac{1}{2}\right)=1$ y la ecuación funcional (ii) $f(x)+(1-x)f\left(\frac{y}{1-x}\right) = f(y)+(1-y)f\left(\frac{x}{1-y}\right)$ sobre $\{(x,y)\in[0\,;1)^2, \quad x+y\leq 1\}$, siendo $H^f(p)=\sum_{i=2}^n s_i f\left(\frac{p_i}{s_i}\right), \quad s_i=\sum_{j=1}^{i-1} p_j$. Daróczy mostró que si f es medible, o continua en 0, o no negativa y acotada, necesariamente $f(x)=h_2(x)=-x\log_2 x-(1-x)\log_2(1-x)$, conduciendo a la entropía de Shannon (teorema 1; ver también (Lee, 1964; Tverberg, 1958; Kendall, 1964)). En otros términos, su axioma (v) es alternativa a la recursividad. Para extender la entropía de Shannon, propuso extender este axioma (v) por la ecuación funcional $f_\lambda(x)+(1-x)^\lambda f_\lambda\left(\frac{y}{1-x}\right) = f_\lambda(y)+(1-y)^\lambda f_\lambda\left(\frac{x}{1-y}\right)$, lo que condujo necesariamente a la entropía (teoremas 2 y 3)

$$H_\lambda^{\mathrm{d}}(p) = \frac{1}{1-2^{1-\lambda}}\left(1-\sum_i p_i^\lambda\right),$$

es decir nada más que la entropía introducida por Havdra & Charvát. En lo que sigue, se la denotará H_λ^{hcd}. Sin embargo, el estudio de Daróczy fue más intensivo que el de Havdra & Charvát. Primero, notó el mapa entre su entropía y la de Rényi. Adicionalmente a Havdra-Charvát probó que se conserva la propiedad [P2] (invarianza por permutación, que no era un axioma en su enfoque), $H_\lambda^{\mathrm{hcd}}\left(\frac{1}{2},\frac{1}{2}\right)=1$ (lo llama normalización), la expansividad [P6], una aditividad extendida, una recursividad extendida precisamente del modelo de Havrda-Charvát (teorema 4). Probó también [P4], positividad alcanzado en el caso determinista y la maximalidad [P5] en el caso uniforme (teorema 6), que incidentalmente $H_\lambda^{\mathrm{hcd}}\left(\frac{1}{\alpha},\ldots,\frac{1}{\alpha}\right)$ crece con el cardinal $|\mathcal{X}|=\alpha$. Muy interesante también es que se puede definir una entropía condicional en el mismo modelo que en el caso

---

exposición de este párrafo.

de Shannon, $H_\lambda^{\mathrm{hcd}}(X|Y) = \sum_y \left[p_{X|Y=y}(x)\right]^\lambda H_\lambda^{\mathrm{hcd}}(p_{X|Y=y})$, que existe una regla de cadena [P14], $H_\lambda^{\mathrm{hcd}}(X,Y) = H_\lambda^{\mathrm{hcd}}(Y) + H_\lambda^{\mathrm{hcd}}(X|Y)$ y que condicionar reduce la entropía $H_\lambda^{\mathrm{hcd}}(X|Y) \leq H_\lambda^{\mathrm{hcd}}(X)$ (teorema 8) [P16]. Mostró también que si se pierde la aditividad, se obiene para X e Y independientes $H_\lambda^{\mathrm{hcd}}(X,Y) = H_\lambda^{\mathrm{hcd}}(X) + H_\lambda^{\mathrm{hcd}}(Y) + \left(2^{1-\lambda}-1\right) H_\lambda^{\mathrm{hcd}}(X) H_\lambda^{\mathrm{hcd}}(Y)$. La propiedades de regla de cadena le permitió revisitar la caracterisación de un canal de transmisión y redefinir una capacidad canal extendidas (capacidad tipo λ; basicamente se usa el mismo enfoque que Shannon, pero usando H_λ^{hcd} en lugar de H, ver sección 6 del papel).

Las entropías tipo Havdra-Charvát-Daróczy fueron (re)descubiertos varias otras veces y/o estudiadas más detenidamente en varios campos y varios extensiones fueron introducidas (Varma, 1966; Onicescu, 1966; Kapur, 1967; Vajda, 1968; Lindhard & Nielsen, 1971; Arimoto, 1971; Burg, 1972; Aczél & Daróczy, 1975; Sharma & Mittal, 1975, 1975; Sharma & Taneja, 1975; Mittal, 1975; Boekee & van der Lubbe, 1980; Ferreri, 1980; Tsallis, 1988; Rathie, 1991; Kaniadakis, 2001; Beck, 2009, entre otros). Un primer enfoque más general es debido a S. Arimoto en los primeros años de la decada 1970 (Arimoto, 1971). Fue rediscubierto y estudiado con más detalles una decada después por J. Burbea y C. R. Rao (Burbea & Rao, 1982) y luego por M. Salicrú (Salicrú, 1987) o M. Teboulle (Teboulle, 1992) entre otros. La medida propuesta, llamada ϕ-*entropía*, es definida por

$$H_\phi(p) = -\sum_i \phi(p_i) \qquad \text{con} \qquad \phi \text{ estrictamente convexa.}$$

Burbea y Rao asociaron una medida de divegencia a esta entropía. Las ϕ-entropías contienen Shannon como caso particular ($\phi(x) = x\log x$), así que la clase de Havdra-Charvát-Daróczy ($\phi(x) = \frac{x-x^\lambda}{2^{1-\lambda}-1}$) como mencionado, pero no la clase de Rényi. De hecho, las ϕ-entropías se enmarcan en una clase un poco más amplia, llamada (h,ϕ)-*entropías* (Salicrú, Menéndez, Morales & Pardo, 1993; Menéndez, Morales, Pardo & Salicrú, 1997). Cambiamos acà substancialmente su escritura de la literatura por razones de homogeneidad con la ϕ-entropía (y las divergencias que se introducirán luego) [128]

Definición 2-74 ((h,ϕ)-entropía)**.** *La (h,ϕ)-entropía de una distribución de probabilidad p_X definida sobre $\mathcal{X}$ de cardinal finito $|\mathcal{X}| = \alpha$ es definida por*

$$H_{(h,\phi)}(X) = H_{(h,\phi)}(p_X) = h\left(-\sum_{x\in\mathcal{X}} \phi\left(p_X(x)\right)\right),$$

donde o

- ϕ *es estrictamente convexa y* h *creciente, o*
- ϕ *es estrictamente cóncava y* h *decreciente*

[128] En la literatura, no hay el signo $-$, y hay que invertir cóncava y convexa.

Frecuentemente, se supone adicionalmente que ϕ y h son de clase C^2, que $\phi(0) = 0$ (la incerteza elemental asociada a un estado de probabilidad nula vale cero) y, sin perdida de generalidad, que $h(-\phi(1)) = 0$.

(ver también (Esteban, 1997) para una generalización aun más amplia). Cuándo $h(x) = x$ se recupera la ϕ-entropía, incluyendo la de Shannon y las de Havdra-Charvát-Daróczy. Además, la familia de Rényi cae tambíen en esta familia ($\phi(x) = -x^\lambda$ y $h(x) = \frac{\log x}{1-\lambda}$) así que todas las entropías evocadas en el párrafo anterior.

Como en el caso de Shannon, para $X = (X_1, \ldots, X_d)$, la (h, ϕ)-entropía de X es una (h, ϕ)-entropía conjunta de los X_i.

Obviamente, de las propiedades de la entropía de Shannon, se conservan las propiedades [P1] (continuidad), [P2] (invariaza por permutación), [P3] (invarianza por transformación biyectiva de X), [P6] (expansabilidad, debido a $\phi(0) = 0$).

Además se conserva la Schur-concavidad con una reciproca:

[P$_\phi$9] Schur-convavidad:

$$p \prec q \quad \Longleftrightarrow \quad H_{(h,\phi)}(p) \geq H_{(h,\phi)}(q) \quad \forall\, (h, \phi).$$

En otros términos, se obtiene la relación de mayorisación si se cumple la relación de ordre entrópicas para todos los pares de funciones entrópicas (h, ϕ). La Schur-concavidad (y su reciproca) es consecuencia de la desigualdad de Schur (Schur, 1923) o Hardy-Littlewood-Pólya (Hardy et al., 1929, 1952) o Karamata (Karamata, 1932) (ver también (Marshall et al., 2011, Cap. 3, Prop. C.1 & Cap. 4, Prop. B.1) o (Bhatia, 1997, Teorema II.3.1)): $p \prec q \;\Rightarrow\; \sum_i \phi(p_i) \leq \sum_i \phi(q_i)$ para toda función ϕ convexa, y reciprocamente.

Como consecuencia, se conservan la positividad [P4] gracia a $\phi(0) = 0$ y $h(-\phi(1)) = 0$ (alcanzado en el caso determinista), la maximalidad [P5] (caso uniforme),

$$0 \leq H_{(h,\phi)}(p_X) \leq h\left(-\alpha\, \phi\left(\frac{1}{\alpha}\right)\right),$$

así que

$$H_{(h,\phi)}\left(\begin{bmatrix} \frac{1}{\alpha} & \cdots & \frac{1}{\alpha} \end{bmatrix}^t\right) \quad \text{función creciente de } \alpha.$$

Con respecto a la concavidad [P8], no se conserva en general:

[P$_\phi$8] Si h es cóncava, entonces $H_{(h,\phi)}(p)$ es cóncava con respecto a p. Eso es una consecuencia de la concavidad de ϕ y decrecencia de h (resp. convexidad/crecencia) conjuntamente a la concavidad de h. La reciproca no es verdad. Por ejemplo, se puede ver que si $\lambda < 1$, la entropía de Rényi es cóncava, pero se proba que existe un $\lambda^*(\alpha) > 1$ tal que para cualquier $\lambda \leq \lambda^*(\alpha)$ se conserva la concavidad, a pesar de que h no sea necesariamente cóncava (Bengtsson & Życzkowski, 2006, p. 57).

Se pierde la propiedad de recursividad [P7], pero se puede vincular la entropía total con la obtenida juntando dos estados por una desigualdad:

[P$_\phi$7] Sean X definido sobre $\mathcal{X}$ y $\breve{X}$ sobre $\breve{\mathcal{X}}$,

$$\begin{cases} \breve{\mathcal{X}} = \{x_1\,,\,\ldots\,,\,x_{\alpha-2}\,,\,\breve{x}_{\alpha-1}\} \quad \text{con el estado interno} \quad \breve{x}_{\alpha-1} = \{x_{\alpha-1}\,,\,x_\alpha\}, \\ p_{\breve{X}}(x_i) = p_X(x_i), \quad 1 \leq i \leq \alpha-1 \quad \text{y} \quad p_{\breve{X}}(\breve{x}_{\alpha-1}) = p_X(x_{\alpha-1}) + p(x_\alpha) \quad \text{distribución sobre } \breve{\mathcal{X}} \\ \breve{q}(x_j) = \dfrac{p_X(x_j)}{p_X(x_{\alpha-1}) + p_X(x_\alpha)}, \quad j = \alpha-1, \alpha \quad \text{distribución del estado interno} \end{cases},$$

$$H_{(h,\phi)}\left(p_X\right) \geq H_{(h,\phi)}\left(p_{\breve{X}}\right).$$

Esta desigualdad es consecuencia de la desigualdad de Petrović (Kuczma, 2009, 43, Teorema 8.7.1), $\phi(a+b) \geq \phi(a) + \phi(b)$ para ϕ convexa y que se cancela en 0 (y la conversa en el caso cóncavo), conjuntamente con h creciente (resp. decreciente). Aparte en el caso de Shannon y el de Havdra-Charvát-Daróczy, no hay ningún vínculo explicito general entre $H_{(h,\phi)}\left(p_X\right)$ y $H_{(h,\phi)}\left(p_{\breve{X}}\right)$.

Se conserva la super-aditividad [P12]. De hecho, si ϕ es convexa (resp. cóncava) con $\phi(0) = 0$, $\forall\ 0 \leq a \leq 1$, $\phi(au) = \phi(au + (1-a)0) \leq a\phi(u)$ (resp. desigualdad reversa). Entonces, $\phi\left(p_{X,Y}(x_i, y_j)\right) = \phi\left(p_{X|Y=y_j}(x_i)p_Y(y_j)\right) \leq p_{X|Y=y_j}(x_i)\phi\left(p_Y(y_j)\right)$, *i. e.,* $\sum_{i,j}\phi\left(p_{X,Y}(x_i, y_j)\right) \leq \sum_{i,j} p_{X|Y=y_j}(x_i)\phi\left(p_Y(y_j)\right) = \sum_i \phi\left(p_Y(y_j)\right)$ (resp. desigualdad reversa). Se cierra la prueba con la crecencia (resp. decrecencia) de h.

Sin embargo, en general, se pierden las propiedades [P10] (aditividad), y [P11] (sub-aditividad). En particular, se conserva solamente en el caso Shannon:

Teorema 2-85. *Sea $p_{X,Y}$ distribución conjunta de variables aleatorias discretas X y Y y p_X y p_Y las de X y de Y (marginales).*

$$H_{(h,\phi)}\left(p_{X,Y}\right) \leq H_{(h,\phi)}\left(p_X \otimes p_Y\right) \quad \forall\, p_{X,Y} \qquad \Longleftrightarrow \qquad \phi(x) = x\log x,$$

i. e., $H_{(h,\phi)}$ es una función creciente de la entropía de Shannon.

Demostración. La reciproca de este teorema es nada más que la propidad [P11] con el hecho de que h es creciente en este caso.

A continuación, la parte directa se demuestra en dos etapas:

- Con un caso particular sobre $\mathcal{X}$ e $\mathcal{Y}$ de cardenal 3 cada unos se proba de que la desigualdad no se puede cumplir, salvo si la derivada ϕ' de la función entrópica satisface a una ecuación funcional.

- la sola solución admisible de esta ecuación se reduce a $\phi(x) = -x\ln x$.

Etapa 1: Sea el vector de probabildad

$$p_{X,Y} = p_X \otimes p_Y - c \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \qquad \text{con} \qquad p_X = \begin{bmatrix} a \\ \tilde{a} \\ 1-a-\tilde{a} \end{bmatrix} \quad \text{y} \quad p_Y = \begin{bmatrix} b \\ \tilde{b} \\ 1-b-\tilde{b} \end{bmatrix}$$

donde $(a, \tilde{a}, b, \tilde{b}) \in D$,

$$D = \{(u, \tilde{u}, v, \tilde{v}) \in [0\,;\,1]^4 : \quad 0 < \tilde{u} \leq 1-u \quad \& \quad 0 < \tilde{v} \leq 1-v\},$$

y $c \in C_{a,\tilde{a},b,\tilde{b}}$,

$$C_{a,\tilde{a},b,\tilde{b}} = \big[-1 + \text{máx}\left\{ab, \tilde{a}\tilde{b}, 1-a\tilde{b}, 1-\tilde{a}b\right\}, \text{mín}\left\{ab, \tilde{a}\tilde{b}, 1-a\tilde{b}, 1-\tilde{a}b\right\}\big].$$

Ahora, si ϕ es convexa (resp. cóncava)

$$\forall\, u, v \quad \phi(v) - \phi(u) \;\geq\; (v-u)\,\phi'(u),$$

i. e., la variación (cuerda) es mayor que la derivada en u, como ilustrado figura Fig. 2-52 (desigualdad reversa para ϕ cóncava). Aplicamos esta desigualdad a $u = p_{X,Y}(x,y)$ y $v = p_X(x)p_Y(y)$ y sumamos

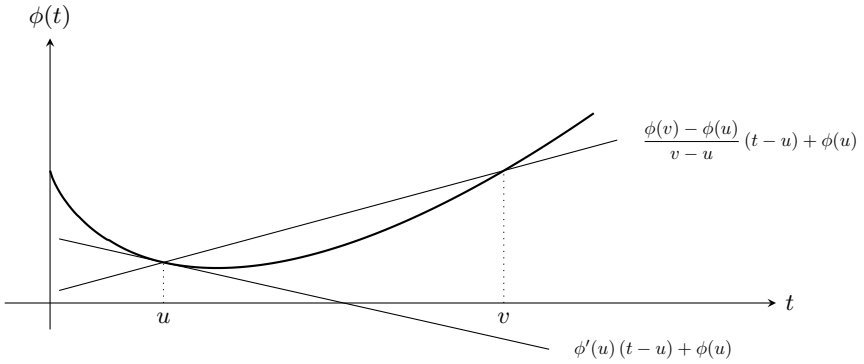


Figura 2-52: ϕ estrictamente convexa: la variación (cuerda) $\frac{\phi(v)-\phi(u)}{v-u}$ es mayor que la derivada $\phi'(u)$. Aplicado a dos distribuciones p y q, de componentes p_i y q_i, con $u = p_i$ y $v = q_i$ y sumando, se obtiene $H_\phi(q) - H_\phi(p) \geq \sum_i (p_i - q_i)\phi'(p_i)$ con $H_\phi \equiv H_{(\text{id},\phi)}$, id siendo la identidad.

en x, y, para $(a,b) \in (0\,;\,1)^2$ (para que $C_{a,\tilde{a},b,\tilde{b}}$ no sea reducido a $\{0\}$), y $c \in \mathring{C}_{a,\tilde{a},b,\tilde{b}}$ donde $\mathring{\cdot}$ denota el interior de un conjunto, se obtiene para ϕ convexa,

$$H_\phi\,(p_X \otimes p_Y) - H_\phi\,(p_{X,Y}) \;\leq\; c\, g(a, \tilde{a}, b, \tilde{b}, c)$$

(para ϕ cóncava se reemplaza H_ϕ por $-H_{-\phi}$ con la igualdad inversa), donde

$$g(a, \tilde{a}, b, \tilde{b}, c) = \phi'\big(ab + c\big) + \phi'\big(\tilde{a}\tilde{b} + c\big) - \phi'\big(a\tilde{b} - c\big) - \phi'\big(\tilde{a}b - c\big). \tag{1}$$

Supongamos que existe un $(s, \tilde{s}, t, \tilde{t}) \in D$, con $(s,t) \in (0\,;\,1)^2$, tal que $g(s, \tilde{s}, t, \tilde{t}, 0) \neq 0$. De la continuidad de ϕ', la función g es continua, entonces existe un vecinaje $V_0 \subset \mathring{C}_{s,\tilde{s},t,\tilde{t}}$ de 0 tal que la función

$c \mapsto g(s, \tilde{s}, t, \tilde{t}, c)$ tiene un signo constante sobre V_0. Eso permite concluir que $c \mapsto c\, g(s, \tilde{s}, t, \tilde{t}, c)$ no tiene un signo constante sobre V_0, y entonces de concluire que, de la desigualdad debido a la concavidad de ϕ (resp. convexidad), $H_\phi\left(p_{X,Y}\right)$ puede ser major (resp. menor) que $H_\phi\left(p_X \otimes p_Y\right)$, y entonces, con la crecencia (resp. decrecencia) de h que si $g(a, \tilde{a}, b, \tilde{b}, 0)$ no es identicamente cero sobre D, $H_{(h,\phi)}$ no puede ser sub-aditiva (distribución conjunta vs producto de las marginales).

Etapa 2. De la etapa 1, se sabe que la sub-aditividad es potencialmente posible solamente si $g(a, v, s, t, 0) = 0$ sobre $D_{a,\tilde{a},b,\tilde{b}} \cap (0\,;\,1)^4$. Eso significa que ϕ' debe necesariamente satisfacer la ecuación funcional

$$\phi'\big(ab\big) + \phi'\big(\tilde{a}\tilde{b}\big) - \phi'\big(a\tilde{b}\big) - \phi'\big(\tilde{a}b\big) = 0,$$

así que no se puede usar el argumento de la etapa 1 para concluir. Sin embargo, se puede solucionar esta ecuación funcional, siguiendo (Daróczy & Járai, 1979, § 6) donde una ecuación funcional muy similar es estudiada. Por eso, se fija $(a, b) \in (0\,;\,1)^2$, se deriva la identidad precediente con respecto a $\tilde{a}$ se multiplica el resultado por $\tilde{a}$ para obtener

$$\tilde{a}\tilde{b}\,\phi''(\tilde{a}\tilde{b}) = \tilde{a}b\,\phi''(\tilde{a}b) \qquad \text{para} \qquad (\tilde{a}, \tilde{b}) \in (0\,;\,1-a) \times (0\,,\,1-b).$$

Eso significa de que $x\,\phi''(x)$ es constante sobre $x \in (0\,;\,(1-a)\,\text{máx}\{b, 1-b\})$, y para cualquier par $(a, b) \in (0\,;\,1)^2$. Entonces, $x\,\phi''(x)$ es constante sobre $x \in (0\,;\,1)$, es decir que ϕ es necesariamente de la forma $\phi(x) = \eta\, x \ln x + \theta x + \vartheta$. Debido a la continuidad de ϕ, queda válido sobre el cerrado $[0\,;\,1]$. De que se aplica a un vector de probabilidad, sumando a uno, se puede reducir el problema a $\theta = 0$ (poniendo θ adentro de ϑ sin cambiar el valor de entropía obtenida). Además, del requisito $\phi(0) = 0$ tenemos $\vartheta = 0$. Para que ϕ sea convexa (resp. cóncava) hace falta tener $\eta > 0$ (resp. $\eta < 0$) así que, sin perdida de generalidad, η puede ser puesta también en h. Tomar $\phi(x) = x \ln x$ con h creciente o $\phi(x) = -x \ln x$ con h decreciente es completamente equivalente, así que se puede fijar $\phi(x) = x \ln x$ satisfaciendo la ecuación funcional, y h creciente. En conclusión, $g = 0$ sobre $D \cap (0\,;\,1)^2$ es decir, por continuidad, sobre D se reduce a necesitar tener $H_\phi = H$. Esta entropía siendo sub-aditiva (propidad [P11]), cualquiera función creciente de H va obviamente quedar sub-aditiva, lo que cierra la prueba. □

Al revés, a partir de $p_{XY} = \frac{1}{2}\begin{bmatrix}1 & 0\end{bmatrix}^t \otimes \begin{bmatrix}1 & 0\end{bmatrix}^t + \frac{1}{2}\begin{bmatrix}0 & 1\end{bmatrix}^t \otimes \begin{bmatrix}0 & 1\end{bmatrix}^t$ se obtiene $p_X = p_Y = \frac{1}{2}\begin{bmatrix}1 & 1\end{bmatrix}^t$ y entonces (i) $H_{(h,\phi)}\left(p_{XY}\right) = h\left(-2\,\phi\left(\frac{1}{2}\right)\right)$, $H_{(h,\phi)}\left(p_X \otimes p_Y\right) = h\left(-4\,\phi\left(\frac{1}{4}\right)\right)$ y $H_{(h,\phi)}\left(p_X\right) + H_{(h,\phi)}\left(p_Y\right) = 2\,h\left(-2\,\phi\left(\frac{1}{2}\right)\right)$, así que, en este ejemplo $H_{(h,\phi)}\left(p_{XY}\right) > H_{(h,\phi)}\left(p_X \otimes p_Y\right)$ (consecuencia de la Schur-concavidad) y $H_{(h,\phi)}\left(p_{XY}\right) > H_{(h,\phi)}\left(p_X\right) + H_{(h,\phi)}\left(p_Y\right)$: tampoco las (h, ϕ)-entropía son super-aditivas.

La definición de entropías generalizadas condicionales aparece mucho más problemático. Por ejemplo, si se define a la Shannon, es decir definiendo $H_{(h,\phi)}\left(X|Y\right)$ tomando $\sum_{y\in\mathcal{Y}} p_Y(y) H_{(h,\phi)}\left(p_{X|Y=y}\right)$ se pierde la regla de cadena [P14]. Como se lo ha visto, en el marco de la entropía de Havdra-Charvát-Daróczy se conserva la regla de cadena si se reemplaza p_Y por su potencia p_Y^λ. Sin embargo, generali-

zar este esquema en el caso general falla (la gracia en Havdra-Charvát-Daróczy viene de la propiedad de morfismo de la exponencial y del logaritmo). Como consecuencia, generalizar la noción se vuelve problemático también. Por ejemplo se pierde el diagrama de Venn aparte si se define la entropía condicional a partir de la regla de cadena. Pero en este caso, si la super-aditividad garantiza la positividad de la entropía condicional, se pierde la propiedad [P13] por perdida de la aditividad, y por consecuencia la propiedad de positividad/independencia [P15] de una información mutua construida sobre un modelo diagrama de Venn. Veremos en la sección siguiente que un tercero camino puede ser usar divergencias.

Como en el caso de Shannon, se puede extender la generalización de la entropía al caso de vectores aleatorios discretos sobre de cardenal infinito, con las mismas debilidades que en el caso de Shannon. A continuación, se puede también extenderla a vectores aleatorios admitiendo una densidad de probabilidad, reemplazando la suma por una integración.

Definición 2-75 ((h,ϕ)-entropía diferencial)**.** *Sea X una variable aleatoria continua sobre $\mathbb{R}^d$ y sea $p_X(x)$ la densidad (distribución) de probabilidad de X de soporte $\mathcal{X}$. La (h,ϕ)-entropía diferencial de la variable X es definida por*

$$H_{(h,\phi)}\left(p_X\right) = H_{(h,\phi)}\left(X\right) = h\left(-\int_{\mathcal{X}} \phi\left(p_X(x)\right)\, dx\right),$$

con h y ϕ cumpliendo los requisitos de la definición discreta 2-74 (de $\phi(0)$, se puede escribir la integración sobre $\mathbb{R}^d$).

De nuevo para $X = (X_1, \ldots, X_d)$, la (h,ϕ)-entropía diferencial de X es una (h,ϕ)-entropía diferencial conjunta de los X_i.

La versión diferencial de la (h,ϕ)-entropía comparte obviamente las mismas debilidades del caso particular de Shannon: se pierden la propiedad de invarianza por transformación biyectiva [P3], *i. e.,* independencia con respecto a los estados, la positividad [P4], la de cota superior [P5] (salvo si se pone vínculos, ver más adelante), en adición de las que ya la versión discreta perdió.

Sin embargo, se conservan unas propidedades, y entre otros si h es cóncava, la (h,ϕ)-entropía diferencia es cóncava [P$_\phi$8]. Más sorprendentemente a primer vista, se conserva la (h,ϕ)-entropía diferencial bajo un rearreglo [P'2],

$$H_{(h,\phi)}\left(p_X^{\downarrow}\right) = H_{(h,\phi)}\left(p_X\right).$$

De hecho, como evocado en el caso de Shannon, eso fue probado entre otros en (Lieb & Loss, 2001) o (Wang & Madiman, 2004, Lema 7.2) [129].

[129] Recuerdense que en (Lieb & Loss, 2001, Sec. 3.3) lo muestran para ϕ diferencia de dos funciones monótonas, siendo una función convexa un caso particular.

Se probó en (Chong, 1974) o (Wang & Madiman, 2004, Prop. 7.3) que se conserva la Schur-concavidad [P9] para las ϕ-entropías. Entonces, de h creciente (para ϕ cóncava desigualdad reversa para la integral, pero h es decreciente), se generaliza a las (h,ϕ)-entropías,*i. e.,*

$$p \prec q \quad \Rightarrow \quad H_{(h,\phi)}(p) \geq H_{(h,\phi)}(q) \quad \forall\,(h,\phi).$$

Quide de la reciproca? Quid sub-aditividad ssi fct creciente de Shannon?

Terminamos esta subsección notando de que, como para la entropía de Shannon, el enfoque discreto y diferencial son contenido en la forma general usando densidades con respeto a una medida (respectivamente discreta y de Lebesgue en estos casos).

Definición 2-76 (Escritura única de las (h,ϕ)-entropías)**.** *Sea X variable aleatoria definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$, admitiendo una densidad de probabilidad p_X con respeto a una medida μ (ej. $\mu_{\mathcal{X}}$ en el caso discreto $\mu = \mu_L$ en el caso diferencial). La (h,ϕ)-entropía de X con respeto a μ se escribe como*

$$H_{(h,\phi)}(X) \equiv H_{(h,\phi)}(p_X) = h\left(-\int_{\mathcal{X}} \phi\left(p_X(x)\right)\, d\mu(x)\right),$$

con h y ϕ cumpliendo los requisitos de la definición discreta 2-74 (de $\phi(0)$, se puede escribir la integración sobre $\mathbb{R}^d$).
Insistamos de nuevo en el hecho de que se puede entender esta definición para cualquier μ y densidad con respeto a μ, que sea discreta, de Lebesgue, o mixta.

2.6.2 Divergencias y propiedades

En esta sub-sección vamos a ver que la literatura trató casi conjuntamente de tres enfoques dando lugar a generalizaciones de la divergencia de Kullback-Leibler. Lamentablemente, ninguna generalización contiene las otras, a pesar de que divergencias conocidas pueden partenecer a varias clases distinctas. Practicamente, cada clase tiene sus ventajas y justificación en termino de aplicaciones.

2.6.2.1. Clase de Jensen

Como se lo ha visto tratando de la entropía relativa, la divergencia de Kullback-Leibler no define una distancia entre distribuciones de probabilidades, siendo no simétrica entre otros. Un primer paso para recuperar la simetría sin perder la positividad de esta medida informacional fue simetrizarla, definiendo lo que es conocido como *J-divergencia* (Kullback & Leibler, 1951; Kullback, 1968; Lin, 1991) [130],

$$D_J(Q\|P) = D_{\mathrm{kl}}(P\,\|Q\,) + D_{\mathrm{kl}}(Q\,\|P\,).$$

[130] Esta expresión apareció en (Jeffrey, 1946, Ec. (1)) o en (Jeffrey, 1948), antes de la introduccíon de la divergencia de Kullback-Leibler en el campo de la estimación Bayesiana, Jeffrey siendo citado por Kullback y Leibler.

Esta versión simetrizada de la divergencia queda naturalmente positiva, pero sufre todavía de unas debilidades de D_{kl}. Esta bien definida siempre que $P \ll Q$ conjuntamente a $Q \ll P$ (las medidas son dichas *medidas equivalentes* en este caso). Además, no cumple tampoco la desigualdad triangular. A pesar de sus debilidades, se usó bastante en problemas de discriminación, debido a su positividad con igualdad si y solamente si $P = Q$ (propiedad herida del hecho de que la suma de términos positivos es nula si y solamente si cada uno vale cero).

Unas decadas después, Lin introdujo lo que llamó K-divergencia directada, $K(P,Q) = D_{\mathrm{kl}}\left(P \left\| \frac{P+Q}{2}\right.\right)$, su versión simetrizada, antes de generalizarla bajo la terminologia de *divergencia de Jensen* (Lin, 1991) [131].

$$\begin{aligned} D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)}) &= \pi_1 D_{\mathrm{kl}}\left(P_{(1)} \left\| \pi_1 P_{(1)} + \pi_2 P_{(2)}\right.\right) + \pi_2 D_{\mathrm{kl}}\left(P_{(2)} \left\| \pi_1 P_{(1)} + \pi_2 P_{(2)}\right.\right) \\ &= H(\pi_1 p_{(1)} + \pi_2 p_{(2)}) - \pi_1 H(p_{(1)}) - \pi_2 H(p_{(2)}) \qquad \pi = [\pi_1 \quad \pi_2], \quad 0 \leq \pi_1 = 1 - \pi_2 \leq 1 \end{aligned}$$

con $p_{(i)}$ densidades de $P_{(i)}$ con respeto a una misma medida μ (puede ser discreta, de Lebesgue, o mixta).

D^{π}_{js} heride obviamente de D_{kl} su positividad con igualdad si y solamente si $P_{(1)} = P_{(2)}$. La misma propiedad puede ser vista a través de la desigualdad de Jensen, dando este nombre a la medida. Además, se quita el problema de definición, siendo de que $P_{(i)} \ll \pi_1 P_{(1)} + \pi_2 P_{(2)}$. No es simétrica en general, pero se obtiene esta propiedad cuando $\pi = \pi_{\mathrm{u}} \equiv [\frac{1}{2} \quad \frac{1}{2}]^t$. Además, en este caso, a pesar de que la divergencia no cumpla la desigualdad triangular, aparece que $\left(J^{\pi_{\mathrm{u}}}_{\mathrm{js}}(P_{(1)}, P_{(2)})\right)^s, \quad 0 < s \leq \frac{1}{2}$ es una metrica [132] (Österreicher & Vajda, 2003, Teorema 1 & Nota 2) o (Endres & Schindelin, 2003; Kafka, Österreicher & Vincze, 1991; Osán, Bussandri & Lamberti, 2018). Si puede parecer más lógico definir tal divergencia con a priori/proporciones π_i iguales, de hecho la versión no simétrica, con pesos π_i se vuelve natural en el marco de la discriminación donde apareció implicitamente esta cantidad. En particular, cuando estamos frente a dos hypotesis $i = 1, 2$ o clases, a las cuales la distribución de las observaciones es $P_{(i)}$, con probabilidad a priori π_i. A partir de observaciones x hay que elegir si eran sorteando de $P_{(1)}$ o $P_{(2)}$ (distribuciones de sampleos, *i. e.,* condicionalmente a la hypotesis). El enfoque Bayesiano más natural consiste maximizar la probabilidad a posteriori (probabilidad de estar en hypotesis i condicionalmente a la observación), y se prueba que la probabilidad de error es dada por $P_e = \int_{\mathcal{X}} \mín(\pi_1 p_{(1)}(x)\,,\, \pi_2 p_{(2)}(x))\, d\mu(x)$ con $p_{(i)}$ densidad con respeto a μ (Kay, 1993). Probó Lin de que

$$\frac{1}{4}\left(H(\pi) - D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)})\right)^2 \leq P_e \leq \frac{1}{2}\left(H(\pi) - D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)})\right),$$

---

[131] De hecho, apareció implicitamente en varios trabajos anteriores, por ejemplo en mecanica cuántica (Holevo, 1973, 2011) o en reconocimiento de patrones (Wong & You, 1985)

[132] Se necesita soló de que los $P_{(i)}$ admiten una densidad con respeto a una medida σ-finita; nos referiremos al resultado 1-9, pagina 53.

lo que da naturalmente un rol operacional a esta divergencia. Incidentalmente, de esta desigualdad es inmediato ver de que $D^{\pi}_{\text{js}}(P_{(1)}, P_{(2)}) \leq H(\pi) - 2P_e$. P_e siendo positivo, da

$$0 \leq D^{\pi}_{\text{js}}(P_{(1)}, P_{(2)}) \leq H(\pi) \leq \log(2)$$

Usando el logaritmo de base 2, adaptado a este caso de dos distribuciones, la cota vale 1: D^{π}_{js} es dicha *normalizada*.

Un otro vínculo natural entre la divergencia de Jensen-Shannon y las medidas informacionales a la Shannon viene todavía del campo de la clasificación. Si unos datos pueden provenir de una distribución $P_{(i)}$, $i = 1, 2$, con una probabilidad π_i, la variable aleatoria X dada por los datos tiene la distribución de mezcla $P = \sum_i \pi_i P_{(i)}$ como ilustrado figura Fig. 2-36-(b), pagina 214. Sea Z la variable aleatoria binaria sobre $\{1\,,\,2\}$ tal que $P(Z = i) = \pi_i$, variable de selección entre las distribuciones $P_{(i)}$ (ej. la moneda de la figura). Por definición de la entropía condicional, $H(X|Z) = \sum_i \pi_i H(X|Z = i) = \sum_i \pi_i H(p_{(i)})$. De $D^{\pi}_{\text{js}}(P_{(1)}, P_{(2)}) = H(p) - \sum_i \pi_i H(p_{(i)})$ viene $D^{\pi}_{\text{js}}(P_{(1)}, P_{(2)}) = H(X) - H(X|Z)$, es decir

$$D^{\pi}_{\text{js}}(P_{(1)}, P_{(2)}) = I(X; Z).$$

La divergencia de Jensen-Shannon mide la información mutua entre la observación X y la variable de selección Z, justificando aun más su uso natural en problemas de clasificación o selección de modelos. Incidentalmente, de $I(X; Z) = H(Z) - H(Z|X) \leq H(Z) \leq \log(2)$ (Z siendo discreta) se recupera las cotas mayor de D^{π}_{js}.

Se encuentran otras desigualdades implicando D^{π}_{js} y D_J o D^{π}_{js} y la distancia L^1 entre distribuciones o divergencia de variación total en (Lin, 1991).

Más allá, en el campo de la clasificación, se puedre tratar de más de dos clases, dando lugar a la generalización de la divergencia de Jensen-Shannon a n distribuciones de probabilidad y π un n-componentes vector de probabilidad,

$$\begin{aligned} D^{\pi}_{\text{js}}(P_{(1)}, \ldots, P_{(n)}) &= H\left(\sum_i \pi_i p_{(i)}\right) - \sum_i \pi_i H(p_{(i)}) \\ &= \sum_i \pi_i D_{\text{kl}}\left(P_{(i)} \middle\| \sum_j \pi_j P_{(j)}\right). \end{aligned}$$

De la desigualdad de Jensen, esta cantitad queda positiva con igualdad si y solamente si todos los $P_{(i)}$ son iguales. Se conserva una cota superior

$$D^{\pi}_{\text{js}}(P_{(1)}, \ldots, P_{(n)}) \leq H(\pi) \leq \log(n),$$

así que $D^{\pi}_{\text{js}}(P_{(1)}, P_{(2)}) = I(X; Z)$ con X de distribución la mezcla $\sum_i \pi_i P_{(i)}$ y Z definida sobre $\{1\,,\,\ldots\,,\,n\}$ variable de selección de distribución π.

convexidad?

Un punto clave que dio lugar a la definición de la divergencia de Jensen-Shannon es la concavidad de la entropía de Shannon. Naturalmente, el mismo enfoque se generaliza a cualquier entropía cóncava de un vector de probabilidad. Tal generalización fue propuesta de maner formal por Burbea-Rao e (Burbea & Rao, 1982), y luego generalizado y estudiado más detenidamente por Nielsen et al. (Nielsen & Boltz, 2011; Nielsen & Nock, 2017). A pesar que que apareció ya en el papel de Burbea & Rao, Nielsen llamó tal generalización "divergencia de Burbea-Rao asimetrizada". Más formalemente, se puede definir una divergencia de Jensen de la manera siguiente:

Definición 2-77 (Divergencias de Jensen)**.** *Sea* $f : \mathcal{U} \subset \mathbb{R}^m \mapsto \mathbb{R}$ *convexa y de clase* C^1 *sobre* $\mathcal{U}$*, un cerrado convexo de* $\mathbb{R}^m$ *y* $\pi = \begin{bmatrix}\pi_1 & \pi_2\end{bmatrix}^t$ *con* $0 \leq \pi_1 = 1 - \pi_2 \leq 1$*. Las divergencias de Jensen entre dos puntos* $u_1, u_2 \in \mathcal{U}$ *son definidas por*

$$J_f^{\pi}(u_1, u_2) = \pi_1 f(u_1) + \pi_2 f(u_2) - f(\pi_1 u_1 + \pi_2 u_2).$$

Se ilustra a que corresponde esta cantidad con respecto a f *en la figura Fig. 2-53 más adelante.*

Esta definición se generalizada a densidad de probabilidad, donde f es a valor reales, actuando sobre el convexo de las medidas de probabilidades (o tomando densidades en un x e integrando sobre $\mathcal{X}$) (Nielsen & Boltz, 2011; Nielsen & Nock, 2017).

Definición 2-78 (Divergencia (h, ϕ)-Jensen entrópica)**.** *Para* (h, ϕ)*-entropías cóncavas (ej. con* h *cóncava), siendo* $-H_{(h,\phi)}$ *convexa, se puede entonces asociar una divergencia de Jensen*

$$D^{\mathrm{j},\pi}_{(h,\phi)}\left(p_{(1)}, p_{(2)}\right) \equiv J^{\pi}_{-H_{(h,\phi)}}(p_{(1)}, p_{(2)}) = H_{(h,\phi)}\left(\pi_1 p_{(1)} + \pi_2 p_{(2)}\right) - \pi_1 H_{(h,\phi)}\left(p_{(1)}\right) - \pi_2 H_{(h,\phi)}\left(p_{(2)}\right).$$

con $p_{(i)}$ *densidad con respeto a una medida* μ*. Cuando* $h \equiv \mathrm{id}$*, se notará* $D^{\mathrm{j},\pi}_{\phi}$*.*

La definición se generaliza a cualquier conjunto $\{p_{(i)}\}_{i=1}^n$ *de distribuciones de probabilidades y* π *vector de probabilidad* n*-dimensional,*

$$D^{\mathrm{j},\pi}_{(h,\phi)}(\{p_{(i)}\}) = H_{(h,\phi)}\left(\sum_i \pi_i p_{(i)}\right) - \sum_i \pi_i H_{(h,\phi)}\left(p_{(i)}\right).$$

Por analogía a la información mutua, Burbea y Rao llamarón esta medida "información mutua generalizada". Eso viene de que si se define una información condicional en el mismo esquema que el de Shannon, *i. e.,* para Y discreta, $H_{(h,\phi)}(X|Y) = \sum_y p_Y(y) H_{(h,\phi)}\left(p_{X|Y=y}\right)$, entonces, con $\pi \equiv p_Y$ y $\{p_{(i)}\}_i \equiv \{p_{X|Y=y}\}_y$ aparece de que $D^{\mathrm{j},p_Y}_{(h,\phi)}(\{p_{X|Y=y}\}_y) = H_{(h,\phi)}(X) - H_{(h,\phi)}(X|Y)$. Esta expresión es parecida a una de las formas de la información mutua de Shannon, justificando la terminilogía de Burbea-Rao. Sin embargo, hay que tener conciencia de que no todo se translata obviamente del mundo Shannon al mundo generalizado. Por ejemplo, con tal definicón de la entropía condicional, se pierde la regla de cadena, y por consecuencia la simetría de tal información mutua generalizada o la forma usando la entropía conjunta y las marginales.

Se notará de que Nielsen propusó generalizaciones mas avanzadas, usando generalizaciones de la noción de convexidad. Estas generalizaciones van más allá de la meta del capítulo y el lector so puede referir a (Nielsen & Nock, 2017).

Las divergencias de Jensen tiene las propiedades siguientes

1. Positividad:

$$J_f^\pi(P,Q) \geq 0 \quad \text{con igualdad si y solamente si} \quad P = Q.$$

   Esta propiedad es la consecuencia directa de la convexidad estricta de f, como ilustrado figura Fig. 2-53.

2. Pensando a J_f^π con respecto a f, es lineal en el sentido de que $J^\pi_{a_1 f_1 + a_2 f_2} = a_1 J^\pi_{f_1} + a_2 J^\pi_{f_2}$ (con f_i convexas y $a_i \geq 0$).

Desgraciamente, las divergencias de Jensen no cumplen la desigualdad triangular en general, y entonces no son métricas entre distribuciones de probabilidad. Se refierá a (Burbea & Rao, 1982; Nielsen & Boltz, 2011; Nielsen & Nock, 2017) para tener más propiedades.

Se notará que la clase de las divergencias de Jensen contiene el cuadrado de la distancia de Mahalanobis (por un factor), *i. e.,* con $f(u) = u^t K u$ con $K > 0$ simétrica se obtiene $J_f(u,v) = \pi_1\pi_2(v-u)^t K(v-u)$ (siendo la distancia L^2 un caso particular). Se generaliza al caso continuo y distancias L^2 con un nucleo.

2.6.2.2. Clase de Bregman

Estas divergencias fueron intoducidos en el campo de la programación lineal convexa, para resolver problemas de minimización convexa [133] (Bregman, 1967), pero con aplicaciones en varios campos (Basseville, 1989, 2013, y ref.):

Definición 2-79 (Divergencias de Bregman)**.** *Sea $f : \mathcal{U} \subset \mathbb{R}^m \mapsto \mathbb{R}$ convexa y de clase C^1 sobre $\mathcal{U}$, un cerrado convexo de $\mathbb{R}^m$. Las divergencias de Bregman de un punto $v \in \mathcal{U}$ relativamente a un punto $u \in \mathcal{U}$ son definidas por*

$$B_f(v\|u) = f(v) - f(u) - (v-u)^t \nabla f(u).$$

Dicho de otra manera, B_f corresponde al desarollo de Taylor al orden 1 de f en la referencia u. Se ilustra a que corresponde esta cantidad con respecto a f en la figura Fig. 2-53 más adelante.

Esta definición fue generalizada a funciones actuando sobre espacios más generales (ej. actuando sobre matrices o operadores en espacios de Hilbert de dimensión infinita) (Petz, 2007). En lo que nos concierna en este capitulo, tratando posiblemente de densidad de probabilidades, nos interesamos a funciones de funciones (Frigyik, Srivastava & Gupta, 2008; Nielsen & Nock, 2017):

Definición 2-80 (Divergencias de Bregman funcional)**.** *Sea $f : \mathcal{U} \mapsto \mathbb{R}$ convexa y de clase C^1 sobre $\mathcal{U}$, un cerrado convexo de un espacio de Banach. Las divergencias de Bregman de un "punto" (una*

[133] Aún que aparece en una revista de matemática y física matemática, una gracia del papel de Bregman es que toma el ejemplo de maximización de la entropía de Shannon sujeto a momentos...

función) $v \in \mathcal{U}$ *relativamente a un "punto"* $u \in \mathcal{U}$ *son definidas por*

$$B_f(v\|u) = f(v) - f(u) - \lim_{t\to 0} \frac{f(u + t(v-u)) - f(u)}{t}.$$

El último término de esta formula es connocida como derivada de Gâteau (o derivada direccional) de f *en* u *en la dirección* $v - u$ *(siendo* u *una función)* [134]*.*

En el caso de que $\mathcal{U} \subset \mathbb{R}^m$ se recupera sencillamente la definición original.

Definición 2-81 (Divergencia (h,ϕ)-Bregman entrópica)**.** *Para* (h,ϕ)*-entropías cóncavas (ej. con* h *cóncava), se puede entonces asociar una divergencia de Bregman*

$$D^{\mathrm{b}}_{(h,\phi)}(q\|p) = H_{(h,\phi)}(p) - H_{(h,\phi)}(q) - h'\big(H_\phi(p)\big)\int_{\mathcal{X}} (q(x) - p(x))\phi'(p(x))\, d\mu(x).$$

Cuando $h \equiv \mathrm{id}$*, se notará* D^{b}_{ϕ} *y es equivalente a salir de la definición inicial* $u = p(x),\ v = q(x)$ *y sumar la divergencia obtenida sobre* $\mathcal{X}$ *con respeto a la medida* μ*. En el caso particular discreto toma la expresión*

$$\begin{aligned} D^{\mathrm{b}}_{(h,\phi)}(q\|p) \equiv B_{-H_{(h,\phi)}}(q\|p) &= H_{(h,\phi)}(p) - H_{(h,\phi)}(q) - (p-q)^t \nabla H_{(h,\phi)}(p) \\ &= H_{(h,\phi)}(p) - H_{(h,\phi)}(q) - h'\big(H_\phi(p)\big)(q-p)^t \phi'(p) \end{aligned}$$

Aparece de que las divergencias de Jensen se escriben a partir de divergencias de Bregman, y vice-versa:

Lema 2-64. *De la definiciones 2-78, 2-79 y 2-80, se muestra sencillamente de que las divergencias de Jensen se escriben como combinaciones convexas de divergencias de Bregman,*

$$J^\pi_f(P_{(1)}, P_{(2)}) = \pi_1 B_f(P_{(1)}\|\pi_1 P_{(1)} + \pi_2 P_{(2)}) + \pi_2 B_f(P_{(1)}\|\pi_1 P_{(1)} + \pi_2 P_{(2)}),$$

Al revés, las divergencias de Bregman se escriben como limites de divergencias de Jensen,

$$B_f(P_{(2)}\|P_{(1)}) = \lim_{\pi_2 \to 0} \frac{J^\pi_f(P_{(1)}, P_{(2)})}{\pi_1 \pi_2}$$

(o similarmente con densidad $p_{(i)}$ *con respeto a una medida* μ*).*

(ver (Zhang, 2004; Nielsen & Boltz, 2011; Nielsen & Nock, 2017)).

La figura Fig. 2-53 ilustra a que corresponden D_f y J_f con respecto a la función convexa f.

La divergencia de Bregman tiene las propiedades siguientes

[134] De hecho, en la extensión de Frigyik et al. (Frigyik et al., 2008), se usa la derivada de Féchet, que es más general. Viene de un límite identica independientemente de la dirección. Entonces, si una función tiene una derivada de Fréchet, tiene necesariamente derivadas de Gâteau, pero no es reciproca. Esta subtileza va más allá de la meta de esta sección.

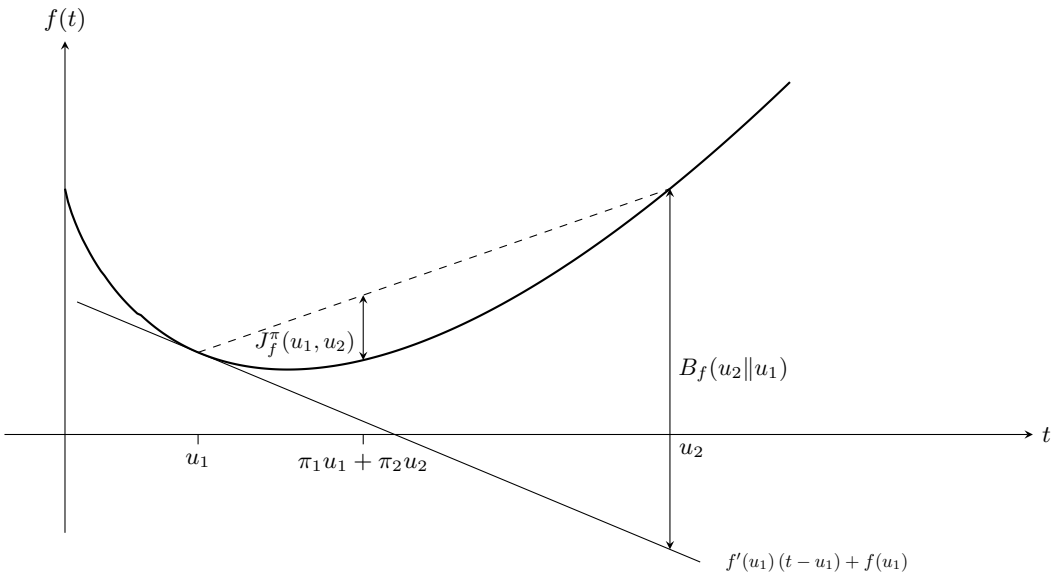


Figura 2-53: f estrictamente convexa. Las cantidad positiva marcada por la dupla-flecha representan respectivamente la divergencia de f-Jensen $J_f^\pi(u_1, u_2)$, diferencia entre la combinaci¨on convexa de los $f(u_i)$ y f de la combinación convexa de los u_i, y la divergencia de Bregman $B_f(u_2\|u_1)$ diferencia entre el valor en u_2 (punto de evaluación) y la tangente en u_1 (punto referencia). Para J_f^π, se toma como referencia $\pi_1 u_1 + \pi_2 u_2$, se calcula D_f en los u_i y se toma la combinación convexa.

1. Positividad:

   $$B_f(Q\|P) \geq 0 \quad \text{con igualdad si y solamente si} \quad P = Q.$$

   Esta propiedad es la consecuencia directa de la convexidad estricta de f, como ilustrado figura Fig. 2-53.

2. $B_f(Q\|P)$ es convexa con respecto a Q, pero no necesariamente con respecto a P. Es también consecuencia directa de la convexidad de f.

3. Pensando a B_f con respecto a f, es lineal en el sentido de que $B_{a_1 f_1 + a_2 f_2} = a_1 B_{f_1} + a_2 B_{f_2}$ (con f_i convexas y $a_i \geq 0$).

Ver (Frigyik et al., 2008; Nielsen & Boltz, 2011; Nielsen & Nock, 2017) para tener más propiedades.

Se notará que la clase de las divergencias de Bregman contiene el cuadrado de la distancia de Mahalanobis con $f(u) = u^t K u$ con $K > 0$ simétrica (siendo la distancia L^2 un caso particular), el cuadrado de la distancia L^1 con $f(u) = \left(\sum_i u_i\right)^2$, la distancia de Itakura-Saito cuando $f(u) = -\log u$ (asociado a la entropía de Burg), entre otros. Unas se exteinden sencillamente al caso continuo (Frigyik et al., 2008).

Como en el caso de divergencias de Jensen, Nielsen propusó generalizaciones más avanzadas, usando las generalizaciones de la noción de convexidad usada para generalizar las divergencias de

Jensen. Estas generalizaciones también van más allá de la meta del capítulo y el lector so puede referir a (Nielsen & Nock, 2017).

También, varias aplicaciones se encuentran en la literatura (Basseville, 1989; Csiszár, 1995; Csiszàr & Matúš, 2012; Basseville, 2013, y ref.) en adición de las referencias de esta sección, para dar unas.

2.6.2.3. Clase de Csiszár o Ali-Silvey

Un primer paso generalizando la noción de entropía relativa o divergencia, siguiendo el enfoque de Kullback y Leibler y sus versiones tipo J-divergencia o divergencia de Jensen-Shannon fue debido a Rényi. En su papel (Rényi, 1961), A. Rényi introdujo una noción de ganancia o perdida de información de una distribucíon (incompleta) de probabilidad q relativa a una referencia p, $I^{\mathrm{r}}(q\|p)$, teniendo un enfoque axiomático similar al que uso para definir su entropía: (i) la medida sea invariante a una misma permutación de los componentes de p y de q, (ii) si $\forall\, i,\ p_i \leq q_i$ entonces $I^{\mathrm{r}}(q\|p) \geq 0$ y vice versa $I(p\|q) \leq 0$, (iii) $I([1]\|[1/2]) = 1$, (iv) $I(q_{(1)} \otimes q_{(2)}\|p_{(1)} \otimes p_{(2)}) = I(q_{(1)}\|p_{(1)}) + I(q_{(2)}\|p_{(2)})$ y (v) una propiedad de media generalizada $I(q_{(1)} \cup q_{(2)}\|p_{(1)} \cup p_{(2)}) = g^{-1}\left(\frac{w_{q_1} I^{\mathrm{r}}(q_{(1)}\|p_{(1)}) + w_{q_2} I^{\mathrm{r}}(q_{(2)}\|p_{(2)})}{w_{q_1}+w_{q_2}}\right)$ conduciendo a

$$I^{\mathrm{r}}_{\lambda}(q\|p) = \frac{1}{\lambda - 1}\log_2\left(\sum_i p_i \left(\frac{q_i}{p_i}\right)^{\lambda}\right).$$

Unos años después, se introdujo una clase más general debido a I. Csiszár (Csiszár, 1963; Csiszàr, 1967; Csiszár & Shields, 2004), T. Morimoto (Morimoto, 1963) o S. M. Ali & S. D. Silvey (Ali & Silvey, 1966), clase que llamaremos *ϕ-divergencias de Csiszàr*. De manera general, con $Q \ll P$, toma la forma

$$D^{\mathrm{c}}_{\phi}\left(Q\|\,P\right) = \int_{\mathcal{X}} \phi\left(\frac{dQ}{dP}(x)\right)\, dP(x),$$

donde ϕ es una función convexa. Estas divergencias o casos particulares fueron muy estudiadas las decadas que siguieron, dando también lugar a varias aplicaciones (Gupta & Sharma, 1976; Burbea & Rao, 1982; Cressie & Read, 1984; Ben-Tal, Charnes & Teboulle, 1989; Teboulle, 1992; Ben-Tal, Bornwein & Teboulle, 1992; Salicrú et al., 1993; Salicrú, 1994; Csiszár, 1995; Cressie & Pardo, 2000; Liese & Vajda, 2006).

Como para el caso de las ϕ-entropías, esta clase se enmarca dentro de una clase un poco más general (Ali & Silvey, 1966, Secs. 4.5 & 5) (ver también (Orsak & Paris, 1995, Sec. I)):

Definición 2-82 ((h,ϕ)-divergencia)**.** *La (h,ϕ)-divergencia de una distribución de probabilidad Q con respeto a una distribución de referencia P tal que $Q \ll P$ es definida por*

$$D^{\mathrm{c}}_{(h,\phi)}\left(Q\|\,P\right) = h\left(\int_{\mathcal{X}} \phi\left(\frac{dQ}{dP}(x)\right)\, dP(x)\right)$$

Si P y Q admiten una densidad con respeto a una medida μ,

$$D^{\mathrm{c}}_{(h,\phi)}\left(q\|\,p\right) = h\left(\int_{\mathcal{X}} p(x)\phi\left(\frac{q(x)}{p(x)}\right) d\mu(x)\right)$$

en su versión diferencial, donde o

- ϕ *es estrictamente convexa y* h *creciente, o*
- ϕ *es estrictamente cóncava y* h *decreciente*

y $\mathcal{X} = X(\Omega)$ *para* X *de distribución* [135] P*. Frecuentemente, se supone adicionalmente que* ϕ *y* h *son de clase* C^2 *y sin perdida de generalidad, que* $h(\phi(1)) = 0$*.*

Se notará de que, obviamente, $D^{\mathrm{c}}_{\phi} = D^{\mathrm{c}}_{(\mathrm{id},\phi)}$.

Notablemente, cuando $\phi(u) = u \log u$ y $h = \mathrm{id}$ se recupera de nuevo la divergencia de Kullback-Leibler: esta última partenece simultaneamente a la clase de Csiszàr y a la de Bregman y es la sola en este caso (Csiszàr, 1991). Cuando $\phi(u) = \pi_2 u \log u - (\pi_1 + \pi_2 u) \log(\pi_1 + \pi_2 u)$ y $h = \mathrm{id}$ se recupera la divergencia de Jensen-Shannon **sola de la clase de Jensen en este caso?**. Además de D_{kl}, la clase de Csiszár contiene la ganancia de información de Rényi para $\phi(u) = u^{\lambda}$ y $h(u) = \frac{\log u}{\lambda - 1}$ apareciendo tambien en una forma muy parecida en (Hellinger, 1909; Chernoff, 1952; Cressie & Read, 1984; Liese & Vajda, 2006) y conocida como divergencia de Chernoff o de Hellinger. Contiene varias otra como la J-divergencia para $\phi(u) = u \log u - \log u$ y $h = \mathrm{id}$, la distancia de Bhattacharyya (Bhattacharyya, 1943, 1946) $-\log \int_{\mathcal{X}} \sqrt{\frac{dQ}{dP}(x)}\, dP(x)$ para $\phi(u) = \sqrt{u}$ y $h(u) = -\log u$, instancia particular de la de Rényi ($\lambda = \frac{1}{2}$), la divergencia de variación total (o L^1 distancia) para $\phi(u) = |u - 1|$ y $h = \mathrm{id}$, la divergencia de Pearson o divergencia χ^2 para $\phi(u) = (u-1)^2$ o $u^2 - 1$ y $h = \mathrm{id}$, para mencionar unas.

Las divergencias de Csiszár tienen las propiedades siguientes

1. Positividad:

   $$D^{\mathrm{c}}_{(h,\phi)}(Q \| P) \geq 0 \quad \text{con igualdad si y solamente si} \quad P = Q.$$

   Esta propiedad es la consecuencia directa de la convexidad estricta de ϕ conjuntamente a $h(\phi(1)) = 0$. De hecho, de la desigualdad de Jensen con X de distribución P tenemos en el caso ϕ convexa y h creciente $D^{\mathrm{c}}_{(h,\phi)}(Q \| P) = h\left(\mathrm{E}\left[\phi\left(\frac{dQ}{dP}(X)\right)\right]\right) \geq h\left(\phi\left(\mathrm{E}\left[\frac{dQ}{dP}(X)\right]\right)\right) = h(\phi(1))$ (y similarmente en el caso ϕ cóncava y h decreciente). Fijense de que la positividad no es en contradicción con el enfoque de Rényi en su caso, porque conideró el caso discreto finito con probabilidades incompletas, *i. e.,* su axioma (ii) se cumpla potencialemente solamente para los vectores de probabilidades incompletos.

2. $D^{\mathrm{c}}_{(h,\phi)}$ satisface un teorema de procesamiento de datos (o secunda ley de la termodinámica) en el sentido de que si dos distribuciones son consecuencias de la misma probabilidad de transición (condicional) $p_{X_{n+1}|X_n=x_n} = q_{X_{n+1}|X_n=x_n}$ (densidades con respeto a una medida μ), entonces

   $$D^{\mathrm{c}}_{(h,\phi)}\left(p_{X_{n+1}} \middle\| q_{X_{n+1}}\right) \leq D^{\mathrm{c}}_{(h,\phi)}\left(p_{X_n} \| q_{X_n}\right).$$

   Probar

[135] En general, por convención, $0\,\phi\left(\frac{0}{0}\right) = 0$. Además, se requiere de que $0\,\phi\left(\frac{a}{0}\right) = \lim_{\varepsilon \to 0^+} \varepsilon\, \phi\left(\frac{a}{\varepsilon}\right) = a \lim_{u \to +\infty} \frac{\phi(u)}{u}$.

3. $D^{\mathrm{c}}_{(h,\phi)}$ es convexa con respecto al par (P,Q), pero no necesariamente con respecto a p solamente y/o q. En el caso ϕ convexa, es consecuencia directa de la convexidad [136] (resp. cóncavidad) de $(u,v)\mapsto u\,\phi\left(\frac{v}{u}\right)$ sobre $\mathbb{R}^2_+$ conjuntamente a la crecencia (resp. decrecencia) de h.

4. Pensando a D^{c}_{ϕ} con respecto a ϕ, es lineal en el sentido de que $D^{\mathrm{c}}_{a_1\phi_1+a_2\phi_2} = a_1 D^{\mathrm{c}}_{\phi_1} + a_2 D^{\mathrm{c}}_{\phi_2}$ (con ϕ_i convexas y $a_i \geq 0$).

5. Sea $\phi^*(u) = u\phi\left(\frac{1}{u}\right)$. Es sencillo ver de que si ϕ es convexa (resp. cóncava), ϕ^* es también convexa (resp. cóncava). ϕ^* es llamada **-conjugada convexa (resp. cóncava)* de ϕ. Luego,

$$D^{\mathrm{c}}_{(h,\phi)}\left(Q\|\,P\right) = D^{\mathrm{c}}_{(h,\phi^*)}\left(P\|\,Q\right).$$

6. $D^{\mathrm{c}}_{(h,\phi)}$ es simétrica si y solamente si $\phi = \phi^* + c(\mathrm{id}-1)$; sin perdida de generalidad, consideramos $c=0$. Sin embargo, en el caso general, se puede definir una versión simetrizada al imagen de la J-divergencia, considerando $D^{\mathrm{c}}_{(h,\phi)} + D^{\mathrm{c}}_{(h,\phi^*)}$. En particular, cuando $h=\mathrm{id}$, tenemos

$$D^{\mathrm{c}}_{\phi} + D^{\mathrm{c}}_{\phi^*} = D^{\mathrm{c}}_{\phi+\phi^*}$$

   que es simétrica ($(\phi^*)^* = \phi$).

7. Cota superior:

$$D^{\mathrm{c}}_{(h,\phi)} \leq h\left(\phi(0)+\phi^*(0)\right)$$

   posiblemente infinita [137].

Estas propiedades con varias otras se encuentran por ejemplo en (Vajda, 1972; Csiszár, 1974; Liese & Vajda, 1987; Kafka et al., 1991; Österreicher, 1996; Österreicher & Vajda, 2003; Vajda, 2009; Kumar & Chhina, 2005; Liese & Vajda, 2006). Como en el caso de las divegencias de Jensen, en general las divergencias simétricas ($\phi=\phi^*$) no satisfacen en general a la desigualdad triangular, y entonces no dan lugar a una distancia entre distribuciones de probabilidad, aparte en casos particulares (ej. divergencia de la variación total, divergencia de Hellinger o Rényi con $\lambda=\frac{1}{2}$). Sin embargo, se probó en (Kafka et al., 1991, Teoremas 1 & 2, Remark 6) y (Österreicher, 1996; Österreicher & Vajda, 2003; Vajda, 2009) el lema siguiente, condición suficiente para que una potencia de la divergencia satisfaga a la desigualdad triangular:

Lema 2-65. *Sea ϕ una función convexa tal que $\phi^*=\phi$, $\phi(0)\neq 0$ y D^{c}_{ϕ} su divergencia de Csiszár asociada ($h=\mathrm{id}$). Adicionalmente de supone de que $\phi(1)=0$ (ver definición Def. 2-82) y de que ϕ es*

---

[136] Con la hipotesis de que ϕ sea de clase C^2, es sencillo ver de que la Hessiana de la función $(u,v)\mapsto u\,\phi\left(\frac{v}{u}\right)$ con respeto a (u,v) es no negativa, implicando la convexidad de esta función bi-variada (Cambini & Martein, 2009).

[137] Por ejemplo, para la divergencia de Kullback-Leibler, $\phi(u) = u\log u$, dando $\phi^*(u) = -\frac{\log u}{u}$, tales que $\phi(0)=0$ y $\phi^*(0)=+\infty$: no es acotada por arriba.

estrictamente convexa en 1. *Si existe* $\kappa \in \mathbb{R}_+^*$ *tal que*

$$h(u) = \frac{(1-u^\kappa)^{\frac{1}{\kappa}}}{\phi(u)}, \quad u \in [0\,;\,1) \quad \textit{es no decreciente,}$$

entonces

$$\left(D_\phi^{\mathrm{c}}\left(\cdot\|\cdot\right)\right)^s \quad s \in (0\,;\,\kappa] \quad \textit{satisface la desigualdad triangular}$$

y entonces es una métrica entre dos discribuciones de probabilidades.

Además, en (Kafka et al., 1991, Sec. 3) se dan condiciones necesarias que debe cumplir κ cuando ϕ tiene un comportamiento particular en $u \to 0$ y/o $u \to 0$; eso va más allá de la meta de esta sección y el lector se podrá referir a (Kafka et al., 1991; Österreicher, 1996; Österreicher & Vajda, 2003) para tener más detalles.

Este lema se usó para probar el caracter métrico de $\left(J_{\mathrm{js}}^{\pi_{\mathrm{u}}}(p_{(1)}, p_{(2)})\right)^s, \quad 0 < s \leq \frac{1}{2}$ (Österreicher & Vajda, 2003; Osán et al., 2018, y ref.), siendo J_{js} una divergencia de Csiszár particular. Se usó también para probar de que $\left(D_\phi^{\mathrm{c}}\right)^s$ con $\phi(u) = \frac{\lambda}{\lambda-1}\left(\left(1+u^\lambda\right)^{\frac{1}{\lambda}} - 2^{\frac{1}{\lambda}-1}(1+u)\right)$ y $\kappa = \min\left(\lambda\,,\,\frac{1}{2}\right)$ es una métrica (Österreicher & Vajda, 2003).

Para cerrar esta sección, se mencionará de que varias aplicaciones se encuentran en la literatura que sea en estimación, discriminación, reconocimiento de patrones, pruebas de adecuación o inferencia estadística, entre otros (Kailath, 1967; Boekee & van der Lubbe, 1979; Poor, 1988; Basseville, 1989; Csiszár, 1995; Orsak & Paris, 1995; Menéndez, Morales, Pardo & Vajda, 1977; Pardo, 1999; Liese & Vajda, 2006; Pardo, 2006; Nielsen & Boltz, 2011; Csiszàr & Matúš, 2012; Basseville, 2013, y ref.) en adición de las referencias de esta sección, para dar unas.

cf Bha 43 egalement

2.6.3 ¿Como se generalizan las identidades y desigualdades?

Principio de entropía máxima Si este principio nació en el marco de la termodynamica o física, con la entropía de Shannon (Boltzman), tratando de las nociones generalizadas de incertas, vuelve natural preguntarse sobre la extensión de este problema en el marco general. **Tal estudio fue hecho en varios trabajos (?, ?; Ben-Tal et al., 1992) nous, Kesavan, Kagan 63**.

El problema se formaliza como en el caso Shannon, buscando la entropía máxima sujeto a vínculos: sea X variable aleatoria viviendo sobre $\mathcal{X} \subset \mathbb{R}^d$ con K momentos $\mathrm{E}\left[M_k(X)\right] = m_k$ fijos, con $M_x : \mathcal{X} \to \mathbb{R}$, el problema de (h,ϕ)-entropía máxima se formula de la manera siguiente: sean $M(x) = \begin{bmatrix} M_1(x) & \cdots & M_K(x) \end{bmatrix}^t$ y $m = \begin{bmatrix} m_1 & \cdots & m_K \end{bmatrix}^t$, se busca,

$$p^* = \operatorname*{argmáx}_p H_{(h,\phi)}(p) \qquad \text{sujeto a} \qquad p \geq 0, \quad \int_{\mathcal{X}} p(x)\,d\mu(x) = 1, \quad \int_{\mathcal{X}} M(x)\,p(x)\,d\mu(x) = m.$$

donde los dos primeros vínculos aseguran de que p^* (positividad, normalización) sea una distribución de probabilidad. Si ϕ es convexa (resp. cóncava), h es creciente (resp. decreciente) así que maximizar $H_{(h,\phi)}$ es equivalente a maximizar H_ϕ (resp. $H_{-\phi}$). Sin perdida de generalidad, se puede considerar la situación ϕ convexa. Como en el caso de Shannon, introduciendo factores de Lagrange $\eta_0,\ \eta = \begin{bmatrix}\eta_1 & \cdots & \eta_K\end{bmatrix}^t$ para tener en cuenta los vínculos, el problema variacional consiste a resolver (Gelfand & Fomin, 1963; van Brunt, 2004; Miller, 2000; Cambini & Martein, 2009; Cover & Thomas, 2006)

$$p^* = \operatorname*{argmáx}_p \int_{\mathcal{X}} \left(-\phi\left(p(x)\right) + \eta_0\, p(x) + \eta^t M(x)\, p(x)\right) d\mu(x)$$

donde η_0, η serán determinados para satisfacer a los vínculos. De nuevo, de la ecuación de Euler-Lagrange (Gelfand & Fomin, 1963; van Brunt, 2004) en el caso continuo, o derivando con respeto a las probabilidades en el caso discreto, se obtiene la ecuación $-\phi'(p(x)) + \eta_0 + \eta^t M(x) = 0$. La función entrópica ϕ es cóncava y de clase C^2, así que ϕ' es continua decreciente, y de la monotonicidad es invertible. Entonces,

$$p^*(x) = \phi'^{-1}\left(\eta^t M(x) - \varphi(\eta)\right)$$

con η tal que se satisfacen los vínculos de momentos y $\varphi(\eta)$ tal que se satisface el vínculo de normalización. Si el resultado no es positivo en $\mathcal{X}$, de las condiciones KKT, $p^*(x) = \left(\phi'^{-1}\left(\eta^t M(x) - \varphi(\eta)\right)\right)_+$. Estas distribuciónes no caen en general en la familia exponencial. De una forma, usando entropía generales permite escaparse de esta familia.

Como en el caso de Shannon, queda obviamente el hecho de que no se puede determinar η tal que se satisfacen todos los vínculos (y en particular la de normalización).

Tal como en el caso Shannon, existe una prueba informacional:

Lema 2-66. *Sea* $\mathcal{P}_m = \left\{p \geq 0 \;\middle|\; \int_{\mathcal{X}} p(x)\, d\mu(x) = 1, \quad \int_{\mathcal{X}} M_k(x)\, p(x)\, d\mu(x) = m\right\}$ *y* $p^* \in \mathcal{P}_m$ *que satisfaga* $\phi'(p^*(x)) = \eta^t M(x) + \eta_0$. *Entonces*

$$\forall\, p \in \mathcal{P}_m, \quad H_{(h,\phi)}\left(p\right) \leq H_{(h,\phi)}\left(p^*\right) \qquad \textit{con igualdad ssi} \quad p = p^* \quad (\mu\textit{s. s.}).$$

Demostración. Sin perdida de generalidad, consideramos ϕ convexa. Calcuando la divergencia de Bregman asociado a ϕ de p relativamente a p^* da

$$\begin{aligned} D_\phi^{\mathrm{b}}\left(p \| p^*\right) &= H_\phi\left(p^*\right) - H_\phi\left(p\right) - \int_{\mathcal{X}} \left(p(x) - p^*(x)\right) \phi'\left(p^*(x)\right)\, d\mu(x) \\ &= H_\phi\left(p^*\right) - H_\phi\left(p\right) - \eta_0 \int_{\mathcal{X}} \left(p(x) - p^*(x)\right) d\mu(x) - \eta^t \int_{\mathcal{X}} \left(p(x) - p^*(x)\right) M(x)\, d\mu(x) \\ &= H_\phi\left(p^*\right) - H_\phi\left(p\right) \end{aligned}$$

siendo p y p^* en $\mathcal{P}_m$. El resulta proviene entonces de la positividad de la divergencia de Bregman, con igualdad si y solamente si $p = p^*$ conjuntamente a la crecencia de h. □

Este lema prueba que, dando vínculos "razonables", la (h, ϕ)-entropía es acotada por arriba, y que se alcanza la cota. Por ejemplo,

- Con $K = 0$ y $\mathcal{X}$ de volumen finito $|\mathcal{X}| < +\infty$, la distribución de (h, ϕ)-entropía máxima es la distribución uniforme en el caso discreto tal como en el caso continuo.

- **Con $K = 1$, $\mathcal{X} = \mathbb{R}^d$ y $M(x) = xx^t$ (visto con d^2 vínculos), y $\phi(u) = u^\lambda$ (Rényi o Havrda-Charvát-Daróczy), la distribución de entropía máxima es Student; Costa and son on. Gausiana se recupera caso límite.**

Como en el caso de Shannon, si $\mu = Q$ es una medida de referencia, el problema vuelev ser un problema de minimización de la (h, ϕ)-divergencia de $P \ll Q$ con respeto a Q. La densidad obtenida es $p^* = \frac{dP^*}{dQ}$.

reapparition Fisher comme courbure, cf Varma, Jizba, MenMor97...

EPI variaciones Madiman Barron MadBar07

On the theory of Fisher's amount of information Sov. Math. Dokl., 4 (1963), pp. 991-993, etc, la codificación a la Renyi (Cambell, Hooda 2001, Bercher)

y la cuantificacion fina; EPI generalizada por Madiman, etc. Lutwak, Bercher etc., Kagan; Boeke 77 An extension of the Fisher information measure I. Csiszár, P. Elias (Eds.), Topics in Information Theory, North-Holland, Berlin/New York (1977), pp. 113-123 o Hammad o Vajda 73 o Ferentinos81 en el marco Fisher; Kesavan gene MaxEnt

Revisite capacite a la Daroczy? codage; parler de la quantification fine et HCD

2.7 Entropias cuanticas discretas

Mas alla caso de informaciones a partir de medida; caso infinito, continuo queda en discusiones

CAPÍTULO 3
Elementos de geometría diferencial

Pedro Walter Lamberti

$\acute{\alpha}\gamma\epsilon\omega\mu\acute{\epsilon}\tau\rho\eta\tau o\varsigma\,;\;;\,\mu\eta\delta\epsilon\iota\varsigma\,;\;;\,\epsilon\iota\sigma\iota\tau\omega$
Que no ingrese nadie que no sepa geometría.
FRASE GRABADA EN LA ENTRADA DE LA ACADEMIA DE PLATÓN

3.1 Estructuras

Una de las nociones más elementales de la matemática es la de *conjunto*. Un conjunto es una colección de elementos perfectamente caracterizados. Los elementos pueden ser de cualquier tipo: números, funciones, personas, autos, etc. El enfoque matemático moderno es ir montando estructuras de distinta naturaleza sobre un dado conjunto. En este capítulo comenzaremos con la noción de espacio topológico y llegaremos al concepto de variedad Riemanniana. Este procedimiento ha mostrado ser de utilidad en el marco de la física, que es nuestro principal ámbito de interés. El mapa de ruta de las distintas estructuras que veremos en este capítulo es el siguiente:

- Espacio topológico (continuidad)
- Espacio métrico (distancia)
- Variedad topológica (coordenadas)
- Variedad diferenciable (diferenciabilidad)
- Estructura afin (paralelismo)
- Estructura métrica (Finsler y Riemann)

Si bien existe una estructura intermedia entre la topológica y la diferenciable, que se conoce como *estructura lineal a trozos*, aquí prescindiremos de su estudio. A su vez, hay otras estructuras matemáticas que son usadas en el marco de las teorías físicas. Se destacan la estructura de producto interno sobre un espacio vectorial complejo, la cual conduce a la noción de espacio de Hilbert, de fundamental importancia en mecánica cuántica; la estructura simpléctica, útil en mecánica clásica y la estructura de Kähler, de relevancia en teoría de cuerdas.

3.2 Espacio Topológico

Un conjunto arbitrario X está desprovisto de toda estructura que permita definir nociones tales como la *convergencia* de una sucesión de elementos de X, la *proximidad* de dos elementos de X, etc. En principio se dispone sólo de las operaciones elementales de *unión* $\bigcup$ e *intersección* $\bigcap$ de subconjuntos. Estas operaciones también pueden realizarse entre distintos conjuntos. Denotaremos con $\emptyset$ al conjunto vacío. Surge entonces el desafío de construir alguna estructura matemática definida sobre X que permita definir, de manera precisa las nociones de proximidad, continuidad, convergencia, etc. Esto se logra a través de la idea de una **topología** sobre X.

Definición 3-83 (Topología)**.** *Una topología Υ sobre el conjunto X es una familia de subconjuntos de X que cumple con las siguientes condiciones:*

1. *X y $\emptyset$ están en Υ: $X, \emptyset \in \Upsilon$*

2. *La intersección de cualquier colección finita de elementos de Υ está en Υ:*

$$A_i \in \Upsilon, \quad \forall\, i = 1, \ldots, n \quad \Rightarrow \quad \bigcap_{i=1}^{n} A_i \in \Upsilon$$

3. *La unión de una colección arbitraria –finita o no– de elementos de Υ, pertenece a Υ:*

$$A_i \in \Upsilon \quad \Rightarrow \quad \bigcup_{i} A_i \in \Upsilon$$

Definición 3-84 (Espacio topológico y abiertos)**.** *Al par (X, Υ) lo llamaremos espacio topológico. Los conjuntos que están en Υ se llaman abiertos.*

Ejemplos:

- *Topología trivial*. Es la que consta de sólo dos elementos, el conjunto vacío y el conjunto total $X : \Upsilon = \{\emptyset, X\}$.

- *Topología discreta*. Es la que en todo subconjunto de X está en Υ, es decir $\Upsilon = \mathcal{P}(X)$ donde $\mathcal{P}(X)$ representa a las partes de X.

- En los cursos elementales de análisis matemático hemos estudiado en $\mathbb{R}^n$, es decir el conjunto de n-tuplas de números reales, la noción de bolas abiertas. Más precisamente, una bola abierta en $\mathbb{R}^n$ centrada en el punto $p = (p_1, ..., p_n) \in \mathbb{R}^n$ y de radio $r > 0$ es el conjunto

$$\mathcal{B}_{r,p} = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : 0 \leq \sqrt{\sum_i (x_i - p_i)^2} < r \right\}$$

  La colección de todas las bolas abiertas en $\mathbb{R}^n$ constituyen una topología para $\mathbb{R}^n$. Se conoce como la *topología usual* de $\mathbb{R}^n$.
  Obsérvese que un subconjunto A de $\mathbb{R}^n$ es abierto (en el sentido usual), cuando para todo $x \in A$, existe un $\varepsilon > 0$ tal que $\mathcal{B}_{\varepsilon,x} \subset A$.

Definición 3-85 (Entorno)**.** *Un entorno de un punto $x \in X$ es un conjunto U que contiene a x y tal que existe un abierto V contenido en U: $x \in V \subseteq U$ con $V \in \Upsilon$.*

Definición 3-86 (Función continua)**.** *Sea $f : X \to Y$ una función entre dos espacios topológicos (X, Υ) e (Y, ω). f es una* **función continua** *en $x \in X$ sii dado cualquier entorno abierto $U \subset Y$ de $f(x)$, existe un entorno de x, $V \subset X$ tal que $f(V) \subset U$. Equivalentemente se puede definir una función continua de la siguiente manera: f es una función continua sii la imagen inversa de cada conjunto abierto es un abierto.*

Es fácil demostrar la equivalencia entre ambas definiciones, y hacerlo queda como ejercicio para el lector.

Definición 3-87 (Homomorfismo)**.** *Un homomorfismo Ψ entre dos espacios topológicos (X, Υ) e (Y, ω) es una función $\Psi : X \to V \subseteq Y$ biyectiva, continua y con inversa continua.*

Definición 3-88 (Sucesión)**.** *Una sucesión en un conjunto X es una aplicación $s : \mathbb{N} \to X$ donde $\mathbb{N}$ es el conjunto de los números naturales. Denotaremos a la sucesión por $\{x_n\}_{n \in \mathbb{N}}$.*

En un espacio topológico podemos introducir la noción de convergencia de una sucesión. Obsérvese que ésto es posible gracias a que disponemos de la noción de conjunto abierto.

Definición 3-89 (Límite)**.** *Sea (X, Υ) un espacio topológico y $\{x_n\}_{n \in \mathbb{N}}$ una sucesión en X. Diremos que x es el límite de x_n si para todo entorno V de x, existe un $n_0 \in \mathbb{N}$ tal que $\forall n \geq n_0$ se tiene que $x_n \in V$.*

Los límites de las sucesiones no tienen porque ser únicos. Una condición que debe cumplir el espacio topológico (X, Υ) para que las sucesiones tengan un único límite es que dados dos puntos distintos $x \neq y$,con $x, y \in X$ existen entornos disjuntos de x e y. A los espacios topológicos que cumplen con esta condición se los llama espacios de Hausdorff o espacios T_2.

3.3 Espacios métricos

En el tercer ejemplo de espacio topológico, usamos la noción de métrica euclídea para definir las bolas abiertas en $\mathbb{R}^n$. El disponer de una métrica no es algo que ocurre en todo conjunto. Eso motiva la siguiente definición:

Definición 3-90 (Espacio métrico)**.** *Un espacio métrico en un conjunto X munido de una función $d : X \times X \to \mathbb{R}_+$ tal que se cumplen las condiciones:*

1. $d(x, y) \geq 0 \quad \forall\, x, y \in X$ *y la igualdad se cumple sii* $x = y$,
2. $d(x, y) = d(y, x)$ *simetría.*
3. $d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z \in X$.

La última condición se conoce como *desigualdad triangular*. Mas adelante en este libro veremos funciones $d : X \times X \to \mathbb{R}_+$ que no satisfacen ni la condición 2 ni la condición 3, pero que sin embargo sirven para medir cuán separados están dos puntos de X. En ese caso diremos que d es una *distancia* definida sobre X.

3.4 Variedad Topológica

Nuestra experiencia cotidiana de percibir que estamos inmersos en un espacio de 3 dimensiones, en el cual podemos medir ángulos y determinar distancias entre dos puntos, ha hecho que usemos estas caracteristicas de nuestro habitat, como motivación de la defición de ciertas estructuras matemáticas en espacios abstractos.

En primer lugar, con la noción de una variedad topológica buscaremos simular en un conjunto cualquiera, la noción de cercanía y dimensionalidad que tenemos en $\mathbb{R}^n$.

Definición 3-91 (Variedad topológica n-dimensional)**.** *Una Variedad topológica n-dimensional es un espacio topológico $\mathcal{M}$ tal que es localmente euclídeo, es decir que para cada $x \in \mathcal{M}$ existe un entorno abierto U de x, homeomorfo a un abierto V de $\mathbb{R}^n$: $\phi : U \subseteq \mathcal{M} \to \mathbb{R}^n$ tal que $\phi : U \to V$ y ϕ es un homeomorfismo. También pediremos que $\mathcal{M}$, como espacio topológico, sea un espacio Hausdorff.*

A los pares (U, ϕ) se los denominan *cartas sobre* $\mathcal{M}$. Se supone que la colección de todas las cartas cubren completamente a $\mathcal{M}$. Las cartas permiten asignar *coordenadas* a $\mathcal{M}$:

$$\text{Si} \quad p \in U \subseteq \mathcal{M} \quad \text{entonces} \quad \phi : p \to (p_1, \ldots, p_n) \in \mathbb{R}^n$$

A la colección de números reales $(p_1, \ldots, p_n)$ se llaman las coordenadas de p de acuerdo a la carta (U, ϕ). La existencia de coordenadas, es el aspecto fundamental por el que el concepto de variedad es tan útil en física.

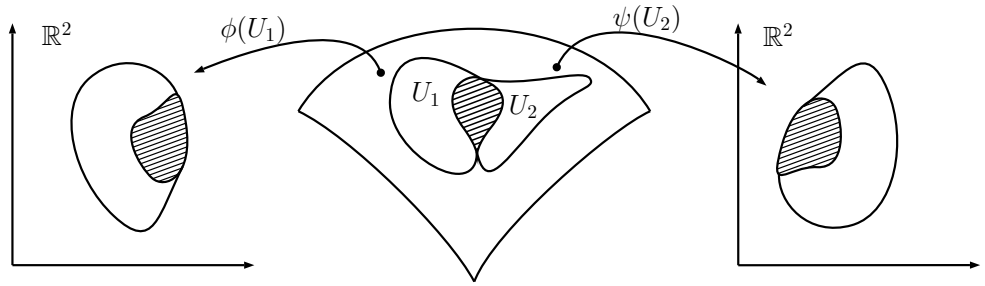


Figura 3-54: Cartas coordenadas usadas en la definición de una variedad topológica.

Podría suceder que un mismo punto p pertenezca a más de una carta, digamos (U_1, ϕ_1) y (U_2, ψ_2). En ese caso hablaremos de un cambio de coordenadas:

$$\psi \circ \phi^{-1} : \phi(U_1 \cap U_2) \to \psi(U_1 \cap U_2) \tag{2}$$

Si denotamos por $(p_1, \ldots, p_n)$ a las coordenadas correspondientes a la carta (U_1, ϕ_1) y por $(\tilde{p}_1, \ldots, \tilde{p}_n)$ a las correspondientes a la carta (U_2, ψ_2), entonces las funciones $\tilde{p}_i = \tilde{p}_i(p_1, \ldots, p_n)$ son funciones continuas, y dan el cambio de coordenadas. Estas funciones son invertibles con inversa continua.

Ejemplos de variedades topológicas son:

- $\mathbb{R}^n$. En este caso hay una carta coordenada global que cubre toda la variedad y donde el homeomorfismo es la identidad.
- $\mathbb{S}^n$, la esfera de dimensión n. Ella está definida como el conjunto:

$$\mathbb{S}^n = \left\{(x_1, \ldots, x_{n+1}),\ x_i \in \mathbb{R} : \quad x_1^2 + \cdots + x_{n+1}^2 = 1\right\}$$

Se debe observar que al definir $\mathbb{S}^n$ no estamos pensando que está inmersa en $\mathbb{R}^n$. En este caso podemos usar las siguientes cartas: (U_N, ϕ_N) y (U_S, ϕ_S), donde $U_N = \mathbb{S}^n - \{(0, 0, \ldots, 1)\}$, $U_S = \mathbb{S}^n - (1, 0, \ldots, 0)$ y los mapas

$$\phi_N : U_N \to \mathbb{R}^n / (\phi_N(x_1, \ldots, n_{n+1}))_i = \frac{x_i}{1 - x_{n+1}}$$

y

$$\phi_S : U_S \to \mathbb{R}^n / (\phi_N(x_1, \ldots, n_{n+1}))_i = \frac{x_i}{1 + x_{n+1}}$$

Ambos mapas son homeomorfismos. Observemos que $\phi_N(x_1, \ldots, x_{n+1}) = (tx_1, \ldots, tx_n)$ y $\phi_S(x_1, \ldots, x_{n+1}) = (ux_1, \ldots, ux_n)$ con $t = \frac{1}{1 - x_{n+1}}$ y $u = \frac{1}{1 + x_{n+1}}$, respectivamente. Es directo verificar la inyectividad pues si $(tx_1, \ldots, tx_n) = (ty_1, \ldots, ty_n) \Rightarrow x_i = y_i \quad \forall\, i$. Entonces los puntos x e y son idénticos. Para ver la suryectividad consideremos el punto $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Si tomamos $x = \left(t^{-1}y_1, \ldots, t^{-1}y_n, y_{n+1}\right)$ con $t \neq 0$ e $y_{n+1} = t\sqrt{1 - (t^{-1}y_1)^2 - \cdots - (t^{-1}y_n)^2}$ vemos que para cada $y \in \mathbb{R}^n$ existe un $x \in \mathbb{S}^n$ tal que $\phi(x) = y$. Usando las expresiones explćitas de ϕ_N y ϕ_S es directo verificar que se trata de funciones continuas.

Nota: Hay propiedades de las variedades topológicas que no tienen que ver con sus características locales, las que hemos dicho son similares a las de $\mathbb{R}^n$, sino con sus propiedades globales. Por ejemplo una esfera 2-dimensional es homeomorfa a la superficie de una pelota de futbol, aún cuando pensemos en una pelota de futbol verdadera, la cual es una colección de parches hexagonales o pentagonales, unidos unos con otros. Ambos objetos, la esfera y la pelota de futbol, son objetos compactos, cerrados y simplemente conexos. Sin embargo un toro y una esfera no comparten todas estas características: un toro es cerrado, compacto pero no simplemente conexo, es decir no todo lazo sobre él puede contraerse continuamente a un punto. Por ello diremos que un toro y una esfera son localmente homeomorfos, pero no lo son globalmente. Este tipo de situaciones ha llevado a introducir cantidades que de alguna manera caractericen a las propiedades globales de una variedad topológicas. Un ejemplo muy conocido es la característica de Euler. Para un poliedro de tres dimensiones la caracteristica de Euler Ξ está definida por

$$\Xi = V - A + C$$

donde V, A y C son el número de vertices, de aristas y de caras del poliedro, respectivamente. Para un cubo, por ejemplo, $\Xi = 2$. Supongamos que el cubo está hecho en un material elástico, apoyado sobre un armazón (las aristas) de metal. Si inflamos ese cubo, obtenemos una esfera. Matemáticamente eso significa que el cubo y la efera son globalmente homeomorfos entre si, y por lo tanto topológicamente equivalentes. Es posible extender el concepto de característica de Euler a la superficie de una esfera, a través de la triangularización de la superficie esférica, es decir cubriendo la esfera por triángulos. En ese caso la característica de Euler se calcula como el número de triángulos menos el número de aristas más el número de vértices. Haciéndo ese cálculo para la esfera resulta el valor 2. Lo mismo sucede con cualquier otro poliedro que se pueda deformarse continuamente a una esfera. Hay maneras de definir la característica de Euler para una variedad topológica arbitraria y esa cantidad es un invariante topológico, es decir una cantidad que no cambia entre variedades homeomórficos. Para un toro la característica de Euler vale 0.

3.5 Variedad Diferenciable

Sobre una variedad topológica se puede "montar" una nueva estructura. Es posible hacer eso imponiendo condiciones de diferenciabilidad a los mapas coordenados de la definición de una variedad topológica. Sin embargo, no tenemos definida la noción de diferenciablidad sobre una variedad cualquiera. Por ello, para definir una estructura diferenciable sobre una variedad topológica arbitraria, recurrimos a $\mathbb{R}^n$ donde si está definida la noción de diferenciabilidad. Por ello hacemos la siguiente:

Definición 3-92 (C^r-compatibilidad)**.** *Diremos que dos cartas coordenadas (U, ϕ) y (V, ψ) sobre una variedad $\mathcal{M}$ son C^r-compatibles si cuando $U \bigcap V \neq \emptyset$ entonces $\phi \circ \psi^{-1}$ y $\psi \circ \phi^{-1}$ son de clase C^r sobre los subconjuntos $\phi(U \bigcap V)$ y $\psi(U \bigcap V)$ de $\mathbb{R}^n$, respectivamente.*

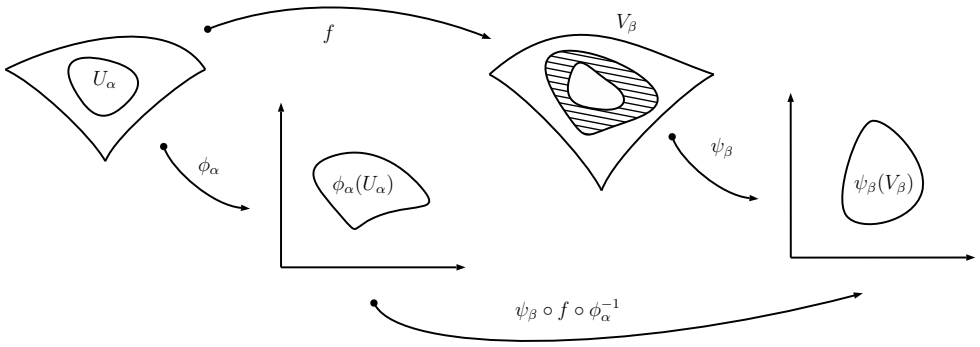


Figura 3-55: Cartas coordenadas usadas en la definición de una función diferenciable.

Con esto podemos avanzar en la siguiente:

Definición 3-93 (Variedad diferenciable)**.** *Una Variedad diferenciable n-dimensional de clase C^r, $\mathcal{M}$, es una variedad topológica y una familia de cartas coordenadas $\mathcal{B} = (U_\alpha, \phi_\alpha)$, tales que:*

1. *los U_α cubren $\mathcal{M}$,*
2. *para cualquier par α, β, los entornos (U_α, ϕ_α) y (U_β, ϕ_β) son C^r - compatibles,*
3. *Cualquier entorno coordenado (V, ψ) C^r-compatible con cualquiera de los $(U_\alpha, \phi_\alpha) \in \mathcal{B}$ está en $\mathcal{B}$.*

Cualquier superficie "suave" en $\mathbb{R}^3$ es un ejemplo de (sub) variedad diferenciable. Este ejemplo no debe conducir a la confusión de pensar que una variedad debe estar inmersa en $\mathbb{R}^n$. Otro ejemplo de variedad diferenciable de dimensión n es la esfera $\mathbb{S}^n$, definida previamente.

Definición 3-94 (Diferenciabilidad de clase C^k)**.** *Dadas dos variedades $\mathcal{M}$ y $\mathcal{M}'$ de clase C^r, una aplicación $f : \mathcal{M} \to \mathcal{M}'$, se dice diferenciable de clase C^k, $\quad k \leq r$ si para toda carta (U_α, ϕ_α) de $\mathcal{M}$ y toda carta de (V_β, ψ_β) de $\mathcal{M}'$ tal que $f(U_\alpha) \subset V_\beta$, la aplicación $\psi_\beta \circ f \circ \phi_\alpha^{-1}$ de $\phi_\alpha(U_\alpha)$ en $\psi_\beta(V_\beta)$, es diferenciable de clase C^k.*

El disponer de la noción de función diferenciable, permite asignar a cada punto de una variedad diferenciable, un espacio vectorial. Éste estará dado por operadores lineales que actúan sobre funciones diferenciables y dan por resultado un número. Antes de ir a la definición de ese espacio vectorial, introducimos el concepto de curva suave sobre una variedad.

Definición 3-95 (Curva de clase C^k sobre una variedad)**.** *Sea $\mathcal{M}$ una variedad de clase C^r. Una curva λ en $\mathcal{M}$ de clase C^k, $\quad k \leq r$ es una función del intervalo real $[a\,,\,b]$ en $\mathcal{M}$ tal que para toda carta (U_α, ϕ_α) en $\mathcal{M}$ la composición*

$$\phi_\alpha \circ \gamma : [a\,,\,b] \to \phi_\alpha(U_\alpha)$$

es de clase C^k. En coordenadas

$$\phi_\alpha \circ \gamma(t) = \{x^1(t), \ldots, x^n(t)\}$$

Con esto podemos ahora dar la noción de vector tangente a una variedad:

Definición 3-96 (Tangente a una variedad)**.** *Sea $\mathcal{F}(p)$ el conjunto de funciones diferenciables de clase C^1 definidas en un entorno del punto p. Sea $\gamma(t)$ una curva de clase C^1, $a \leq t \leq b$ tal que $\gamma(t_0) = p$. El vector tangente a la curva $\gamma(t)$ en el punto p es una aplicación $\mathbb{X}_p : \mathcal{F}(p) \to \mathbb{R}$ cuyo efecto es*

$$\mathbb{X}_p f = \frac{df(\gamma(t))}{dt}|_{t_0}$$

El vector $\mathbb{X}_p$ satisface las siguientes propiedades

- $\mathbb{X}_p$ es una aplicación lineal de $\mathcal{F}(p)$ en $\mathbb{R}$,
- $\mathbb{X}_p(fg) = (\mathbb{X}_p f)\, g(p) + f(p)\, (\mathbb{X}_p g)$ para $f, g \in \mathcal{F}(p)$.

Dejamos para el lector demostrar estas propiedades.

Sean $(u^1, \ldots, u^n)$ coordenadas locales en un entorno U de p. Para cada j, $\left(\frac{\partial}{\partial u^j}\right)|_p$ es una aplicación de $\mathcal{F}(p)$ en $\mathbb{R}$ la cual satisface las propiedades (i) e (ii). Veremos a continuación que el conjunto de todas las aplicaciones $\mathbb{X}$ de $\mathcal{F}(p)$ en $\mathbb{R}$ es un espacio vectorial n-dimensional, siendo n la dimensión de la variedad diferenciable $\mathcal{M}$.

Dada una curva $\gamma(t)$ con $\gamma(t_0) = p$, sean $u^j(t) = \gamma^j(t), \quad j = 1, \ldots, n$ las coordenadas locales de esa curva. Entonces $\frac{df(\gamma(t))}{dt}|_{t_0} = \sum_j \left(\frac{\partial f}{\partial u^j}|_p\right)\left(\frac{d\gamma^j(t)}{dt}\right)|_{t_0}$. Extra expresión indica que todo vector en p es una combinación lineal de los vectores (operadores).

$$\left(\frac{\partial}{\partial u^1}|_p\right), \ldots, \left(\frac{\partial}{\partial u^n}|_p\right) \tag{3}$$

Sea la combinación lineal $\sum_j \xi^j \frac{\partial}{\partial u^j}|_p$ y sea la curva definida por

$$u^j(t) = u^j(p) + \xi^j t \quad j = 1, \ldots, n$$

El vector tangente a esta curva en $t = 0$ es $\sum \xi^j \frac{\partial}{\partial u^j}|_p$. Además si

$$\sum \xi^j \frac{\partial}{\partial u^j}|_p = 0,$$

entonces

$$0 = \sum \xi^j \left(\frac{\partial u^k}{\partial u^j}\right)|_p = \xi^k \quad k = 1, \ldots, n$$

Esto demuestra la independencia lineal de los vectores (3).

Definición 3-97 (Espacio tangente)**.** *El conjunto de vectores tangentes en $p \in \mathcal{M}$, es llamado el espacio tangente de $\mathcal{M}$ en p, y lo denotaremos por $T_p(\mathcal{M})$.*

La colección de todos los espacios tangentes, $\bigcup_{p\in\mathcal{M}} T_p(\mathcal{M})$ se llama *fibrado tangente*.

Al fibrado tangente se le puede dar la estructura de un álgebra (álgebra de Lie). Esta surge de calcular el conmutador $[\mathbb{X}, \mathbb{Y}]$ entre dos campos vectoriales $\mathbb{X}$ e $\mathbb{Y}$:

$$[\mathbb{X}, \mathbb{Y}]f \equiv (\mathbb{X}\mathbb{Y} - \mathbb{Y}\mathbb{X})\, f$$

Si los vectores se escriben en término de los vectores de la base coordenada $(\frac{\partial}{\partial x^a})$, el conmutador entre ellos resulta ser el vector:

$$\sum_{ab} X^a \frac{\partial Y^b}{\partial x^a}\frac{\partial}{\partial x^b} - \sum_{ab} Y^a \frac{\partial X^b}{\partial x^a}\frac{\partial}{\partial x^b}$$

A cada espacio tangente $T_p(\mathcal{M})$ podemos asignar su dual, $T_p^*(\mathcal{M})$, es decir el conjunto de todos los operadores lineales y homogéneos que actúan sobre $T_p(\mathcal{M})$. A un elemento del espacio dual lo llamaremos 1-*forma*. Denotaremos a la acción de un elemento de $T_p^*(\mathcal{M})$, digamos ω_p, por:

$$\omega_p(\mathbb{X}_p) = \langle \omega_p, \mathbb{X}_p \rangle .$$

Para cada función $f \in \mathcal{F}(p)$, el *diferencial de* f, denotado por $(df)_p$, es el elemento de $T_p^*(\mathcal{M})$ que tiene por acción:

$$\langle (df)_p, \mathbb{X}_p \rangle = \mathbb{X}_p f, \quad \mathbb{X}_p \in T_p(\mathcal{M})$$

Cada función coordenada u^j es una función de $\mathcal{M}$ sobre $\mathbb{R}$. Entonces podemos calcular el diferencial de u^j, cuya acción sobre un vector $\mathbb{X}_p \in T_p(\mathcal{M})$ está dada por

$$\left\langle (du^j)_p, \mathbb{X}_p \right\rangle = \mathbb{X}_p^j$$

En particular, si $\mathbb{X}_p = \left(\frac{\partial}{\partial u^k}\right)$ resulta

$$\left\langle (du^j)_p, \left(\frac{\partial}{\partial u^k}\right) \right\rangle = \delta_k^j;$$

es decir $\left\{(du^j)_p\right\}_{j=1}^n$ es la base dual de $\left\{\left(\frac{\partial}{\partial u^j}\right)_p\right\}_{j=1}^n$. Toda 1-forma ω se puede escribir en término de esta base:

$$\omega = \sum_a \omega_a dx^a$$

Con los espacios $T_p(\mathcal{M})$ y $T_p^*(\mathcal{M})$ podemos construir el espacio producto cartesiano

$$(T_p(\mathcal{M}))_s^r = T_p(\mathcal{M}) \times T_p(\mathcal{M}) \ldots T_p(\mathcal{M}) \times T_p^*(\mathcal{M}) \times T_p^*(\mathcal{M}) \ldots \times T_p^*(\mathcal{M})$$

con r factores de $T_p(\mathcal{M})$ y s factores de $T_p^*(\mathcal{M})$.

Definición 3-98 (Tensor de tipo (r,s))**.** *Un tensor de tipo* (r,s) *es un operador* S*,*

$$S : (T_p(\mathcal{M}))_s^r \to \mathbb{R}$$

que es lineal y homogéneo en cada uno de sus argumentos.

Definición 3-99 (Campo tensorial)**.** *Un campo tensorial* S *de clase* C^k *de tipo* (r,s) *sobre* $V \subseteq \mathcal{M}$ *es un mapa* C^k *que asigna un tensor de tipo* (r,s) *a cada punto* $p \in V$*.*

En término de las bases $\left\{\left(\frac{\partial}{\partial u^j}|_p\right)\right\}_{j=1}^n$ y $\left\{(du^j)_p\right\}_{j=1}^n$, el campo tensorial S se puede escribir:

$$S(p) = S_{b_1 \ldots b_s}^{a_1 \ldots a_r}(p) \frac{\partial}{\partial x^{a_1}} \bigotimes \ldots \bigotimes \frac{\partial}{\partial x^{a_r}} \bigotimes dx^{b_1} \bigotimes \ldots \bigotimes dx^{b_s}$$

donde las funciones $S_{b_1 \ldots b_s}^{a_1 \ldots a_r}$ son de clase C^k y $\bigotimes$ es el producto tensorial.

Entre los campos tensoriales que se pueden definir sobre una variedad $\mathcal{M}$, hay uno particularmente importante, y es conocido como el tensor métrico. Éste se define por medio de un producto escalar:

Definición 3-100 (Producto escalar)**.** *Un* **producto escalar** *sobre* $T_p(\mathcal{M})$ *es una función*

$$g : T_p(\mathcal{M}) \times T_p(\mathcal{M}) \to \mathbb{R}$$

que satisface

1. $g(\mathbb{X}, \mathbb{Y}) = g(\mathbb{Y}, \mathbb{X}), \quad$ *para* $\quad \mathbb{X}, \mathbb{Y} \in T_p(\mathcal{M})$
2. $g(\mathbb{X}, a\mathbb{Y} + b\mathbb{Z}) = ag(\mathbb{X}, \mathbb{Y}) + bg(\mathbb{X}, \mathbb{Z})$

El producto escalar se dice *no degenerado* si $g(\mathbb{X}, \mathbb{Y}) = 0 \quad \forall\, \mathbb{Y} \in T_p(\mathcal{M})$ implica $\mathbb{X} = 0$. Obviamente el producto escalar es un tensor de tipo $(0, 2)$. Como campo tensorial $g(\ ;\ ,\ ;\ )$ se puede expresar en términos de la base $\left(\frac{\partial}{\partial u^1}|_p\right), \ldots, \left(\frac{\partial}{\partial u^n}|_p\right)$

$$g(\mathbb{X}, \mathbb{Y}) = \sum_{ab} X^a Y^b g\left(\frac{\partial}{\partial u^a}, \frac{\partial}{\partial u^b}\right)$$

o, de manera equivalente

$$g(\mathbb{X}, \mathbb{Y}) = \sum_{ab} g_{ab} X^a Y^b$$

donde $g_{ab} = g\left(\frac{\partial}{\partial u^a}, \frac{\partial}{\partial u^b}\right)$. Si el producto escalar es no degenerado, entonces existe la matriz inversa de la matriz g_{ab}, a cuyos elementos los denotaremos por g^{ab}, de modo que

$$\sum_c g_{ac} g^{cb} = \delta_a^b$$

La existencia de un campo tensorial métrico (o producto escalar definido localmente), permite introducir la idea de *longitud de una curva*. En efecto, sea $\gamma(t), \quad t \in [a\,,\, b]$ una curva de clase C^1 sobre $\mathcal{M}$, que une los puntos p y q: $\gamma(a) = p, \quad \gamma(b) = q$. En el punto $\gamma(t)$ tenemos el vector tangente a la curva γ dado por

$$\left(\frac{\partial}{\partial t}\right)_\gamma = \sum_j \frac{d\gamma^j}{dt} \frac{\partial}{\partial x^j}$$

Definición 3-101 (Longitud de una curva)**.** *La longitud de la curva* γ *entre los puntos* p *y* q *está dada por la cantidad*

$$L = \int_a^b |g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)|^{\frac{1}{2}} dt \tag{4}$$

O, equivalentemente

$$L = \int_a^b \left| \sum_{ij} g_{ij}(x) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} \right|^{\frac{1}{2}} dt \tag{5}$$

3.6 Estructura Afin

En el espacio euclídeo n-dimensional (pensado aquí como una variedad diferenciable), cuando usamos coordenadas cartesianas, caracterizamos a dos vectors paralelos como aquellos que tienen

iguales componentes. Si reemplazamos las coordenadas cartesianas por las polares, por ejemplo, esta caracterización deja de ser válida. Veamos cómo podemos introducir la noción de paralelismo de vectores, usando cualquier sistema de coordenadas. Sea $\{x^a\}$ el sistema de coordenadas cartesiano del espacio. En este sistema, hemos dicho que dos vectores paralelos, por ejemplo $\mathbb{V}$ y $\tilde{\mathbb{V}}$ tienen iguales componentes:

$$V^a = \tilde{V}^a$$

Si el vector $\mathbb{V}$ es tangente al espacio en el punto p con coordenadas $\{x^a\}$ y el vector paralelo $\tilde{\mathbb{V}}$ es tangente al punto q con coordenadas $x^a + \delta x^a$, vale

$$\tilde{V}^a(q) - V^a(p) = 0$$

Dado un vector $\mathbb{V}$ en p, podemos definir un campo de vectores paralelos a $\mathbb{V}$ en un entorno de p. Denotemos a este campo por $\tilde{\mathbb{V}}$. Este campo cumple que en el punto p coincide con $\mathbb{V}$ y con la condición:

$$\tilde{V}^a(x + \delta x) - V^a(x) = \frac{\partial \tilde{V}^a}{\partial x^b}(p)\delta x^b$$

Sea ξ^a otro sistema de coordenadas para el espacio euclídeo, vinculado con x^a mediante las relaciones

$$\xi^a = \xi^a(x^b), \quad x^b = x^b(\xi^a) \tag{6}$$

A partir de ellas, resulta

$$\delta\xi^a = \frac{\partial \xi^a}{\partial x^b}\delta x^b, \qquad \delta x^b = \frac{\partial x^b}{\partial \xi^a}\delta\xi^a \tag{7}$$

Las componentes de $\tilde{\mathbb{V}}$ se transforman de acuerdo con

$$\tilde{V}^a = \frac{\partial x^a}{\partial \xi^b}\tilde{V'}^b$$

donde $\tilde{V'}^a$ son las componentes de $\tilde{\mathbb{V}}$ en las coordenadas $\{\xi^a\}$. Entonces, podemos escribir

$$\begin{aligned} \frac{\partial \tilde{V}^a}{\partial x^b} &= \frac{\partial}{\partial \xi^c}\left(\frac{\partial x^a}{\partial \xi^d}\tilde{V'}^d\right)\frac{\partial \xi^c}{\partial x^b} \\ &= \frac{\partial^2 x^a}{\partial \xi^c \partial \xi^d}\tilde{V'}^d\frac{\partial \xi^c}{\partial x^a} + \frac{\partial x^a}{\partial \xi^d}\frac{\partial \tilde{V'}^d}{\partial \xi^c}\frac{\partial \xi^c}{\partial x^b} \end{aligned} \tag{8}$$

Si definimos la cantidad $\delta\tilde{V'}^d = \frac{\partial \tilde{V'}^d}{\partial x^e}\delta\xi^e$ y después de un poco de álgebra, llegamos a la relación

$$\delta\tilde{V'}^n = -\frac{\partial^2 x^a}{\partial \xi^e \partial \xi^d}\frac{\partial \xi^n}{\partial x^a}\tilde{V'}^d\delta\xi^e \tag{9}$$

Esta expresión puede reescribirse de la siguiente forma:

$$\delta\tilde{V'}^n = -\Gamma'^n_{ed}\tilde{V'}^d\delta\xi^e \tag{10}$$

en donde los coeficientes Γ' están definidos en la expresión (9). De su definición resulta que las cantidades Γ' se anulan para cambios *lineales* de coordenadas (6).

Obsérvese que al haber arribado a la definición de los coeficientes Γ no hemos hecho uso de ninguna propiedad especial del espacio euclídeo. Es por ello que la expresión (9) es válida para cualquier variedad n-dimensional. Es fácil ver que frente a un cambio de coordenadas

$$x^a \to x'^a = x'^a\left(x^b\right) \tag{11}$$

las cantidades Γ cambian según la expresión

$$\Gamma^f_{de} = \Gamma'^a_{mn}\frac{\partial x^f}{\partial x'^a}\frac{\partial x'^m}{\partial x^d}\frac{\partial x'^n}{\partial x^e} + \frac{\partial x^f}{\partial x'^a}\frac{\partial^2 x'^a}{\partial x^e \partial x^d} \tag{12}$$

Debemos remarcar que esta ley de transformación es lineal y homogénea (tensorial) sólo cuando el cambio de coordenadas (6) es lineal. Esta propiedad de los coeficientes Γ nos permite generalizar la idea de paralelismo en una variedad arbitraria:

Definición 3-102 (Conexión afín)**.** *Cuando en una variedad n-dimensional arbitraria $\mathcal{M}$ se introducen n^3 coeficientes Γ que se transforman de acuerdo con la ley* (12)*, diremos que sobre esa variedad se ha definido una conexión afín*

A partir de los coeficientes Γ es posible definir una nueva derivada para un campo vectorial arbitrario, digamos $V^a(x)$:

Definición 3-103 (Derivada covariante de campo)**.** *Sea un campo vectorial V definido en un entorno del punto x. La derivada covariante del campo V está dado por las componentes de un tensor de tipo* $(1,1)$

$$V^a_{;c} = V^a_{,c} + \Gamma^a_{bc}V^b$$

Definición 3-104 (Derivada covariante en una dirección)**.** *Dados dos campos vectoriales $U(x)$ y $V(x)$, la derivada covariante de V en la dirección de U es el campo vectorial definido por*

$$U(x)\cdot\nabla V(x) \equiv \sum_{ab} V^a_{;b}(x)U^b(x)\mathbb{E}_a \equiv \nabla_U V$$

donde $\mathbb{E}^a$ es el campo de vectores coordenados asociados con las coordenadas x^a. Esta última definición permite trasladar paralelamente a un vector a lo largo de una curva. Basta con tomar como $\mathbb{U}$ al campo tangente a la curva.

3.7 Variedad Riemmanniana

Sea $\mathcal{M}$ una variedad diferenciable n-dimensional. Si $\mathcal{M}$ tiene definida una métrica no singular sobre ella, recibe el nombre de *variedad Riemanniana*. La existencia de una métrica sobre $\mathcal{M}$ permite introducir una conexión afín particular, conocida como la conexión de Levi-Civita. Sean g_{ab} y g^{ab} los coeficientes de la métrica g y su inversa, en las coordenadas $\{x^a\}$, respectivamente.

Para dos puntos próximos, la separación entre ellos viene dada por la expresión:

$$ds^2 = \sum_{ab} g_{ab} dx^a dx^b \tag{13}$$

Además de definir una distancia entre puntos próximos, la existencia de una métrica permite definir una conexión particular sobre una variedad riemanniana:

Definición 3-105 (Conexión de Levi-Civita)**.** *La conexión de Levi-Civita en las coordenadas x^a está dada por:*

$$\Gamma^a_{bc} = \frac{1}{2} \sum_d g^{ad} \left(g_{bd,c} + g_{cd.b} - g_{bc.d} \right) \tag{14}$$

La existencia de esta particular conexión no imposibilita la existencia de otras conexiones definidas sobre $\mathcal{M}$.

Como hemos visto más arriba, el tener definida una métrica permite definir la longitud de una curva. Bajo ciertas condiciones, que supondremos que se satisfacen, podemos plantearnos el problema de determinar la curva que minimiza (en realidad extremiza) su longitud al unir dos puntos fijos sobre la variedad. Esto se puede tratar resolviendo el problema variacional asociado con el funcional (5). La ecuación de Euler-Lagrange conduce en este caso a:

$$\frac{d^2 x^d}{dt^2} + \Gamma^d_{ca} \frac{dx^c}{dt} \frac{dx^a}{dt} = 0 \tag{15}$$

donde $x^a(t)$ son las coordenadas de la curva y t es un parámetro adecuadamente elegido. Una curva que satisface (15), se llama una *curva geodésica*. Es posible caracterizar a una curva geodésica de otro modo. Sea $\mathbb{U}(t)$ el vector tangente a una curva $\gamma(t)$ definida sobre $\mathcal{M}$. La curva γ se dice una geodésica si su vector tangente es trasladado paralelamente a lo largo de ella:

$$\mathbb{U} \cdot \nabla \mathbb{U} = f(t) \mathbb{U}$$

Siempre es posible elegir al parámetro t de forma tal que $f(t) = 0$, con lo cual reobtenemos la ecuación (15).

El disponer de geodésicas, permite dar a una variedad riemanniana el carćter de espacio métrico. En efecto, podemos definir la distancia entre dos puntos p y q de la variedad $\mathcal{M}$ a través de la expresión:

$$d(p, q) = \min_{\gamma} L(\gamma) \tag{16}$$

donde el mínimo se evalúa entre todas las curvas que unen los puntos p y q, y L es la longitud (5). Como siempre, todo esto es posible ser realizado localmente. Las geodésicas son las curvas que localmente minimizan la distancia entre dos puntos. La distancia definida por (16) verifica la desigualdad triangular, y por eso es una métrica.

Dada una conexión ∇ se define un tensor de tipo $(1, 3)$, llamado *tensor de curvatura* asociado a la conexión ∇, cuya espresión es:

$$\mathcal{R}(\mathbb{X}, \mathbb{Y}) \mathbb{Z} = \nabla_{\mathbb{X}} (\nabla_{\mathbb{Y}} \mathbb{Z}) - \nabla_{\mathbb{Y}} (\nabla_{\mathbb{X}} \mathbb{Z}) - \nabla_{[\mathbb{X}, \mathbb{Y}]} \mathbb{Z}$$

Si los vectores $\mathbb{X}$, $\mathbb{Y}$ y $\mathbb{Z}$ son reemplazados por los vectors coordenados $\frac{\partial}{\partial x^a}$, $\frac{\partial}{\partial x^b}$ y $\frac{\partial}{\partial x^c}$, respectivamente, resulta

$$\mathcal{R}\left(\frac{\partial}{\partial x^a}, \frac{\partial}{\partial x^b}\right)\frac{\partial}{\partial x^c} = R^d_{cab}\frac{\partial}{\partial x^d}$$

con

$$R^d_{cba} \equiv \left(\frac{\partial \Gamma^d_{cb}}{\partial x^a} + \Gamma^d_{ra}\Gamma^r_{cb} - \frac{\partial \Gamma^d_{ca}}{\partial x^b} - \Gamma^d_{rb}\Gamma^r_{ca}\right)\frac{\partial}{\partial x^d}$$

Nota: Si bien existe una motivación geométrica para introducir el tensor de curvatura, aquí no la hemos dado. Ella tiene que ver con la idea de cuánto cambia un vector al desplazarlo paralelamente a lo largo de una curva cerrada. En general diremos que una variedad es plana, si todas las componentes de su tensor de curvatura, se anulan.

Concluimos este capítulo con una breve nota histórica. En sus trabajos originales sobre geometría, B. Riemann introdujo el elemento de línea entre dos puntos vecinos p y q por medio de la expresión

$$ds = F(p, \mathbb{X})dt \tag{17}$$

con $F(p, \mathbb{X})$ una función homogénea de grado 2 en la segunda variable. Aquí estamos suponiendo que los puntos p y q tienen coordenadas x^a y $x^a + X^a dt$, respectivamente. La geometría basada sobre el elemento de línea se conoce como geometría de Finsler. Obsérvese que el elemento (13) (de Riemann) es un caso particular de la geometría de Finsler.

CF libro de Bullet et al. (Bullet, Fearn & Smith, 2017), Cencov (Cencov, 1982), Amari (Amari & Nagaoka, 2000).

EPÍLOLOGO

Este libro surgio de la experiencia de los autores en el dictado del curso semestral “Métodos de geometría diferencial en teoría de la información”, que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. **. . . acabar** *Los autores*

Referencias

Ablowitz, M. J. & Fokas, A. S. (2003). *Complex Variables: Introduction and Applications* (2nd ed.). New-York: Cambridge University Press.

Abramowitz, M. & Stegun, I. A. (1970). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 9th printing. New-York: Dover.

Aczél, J. & Daróczy, Z. (1975). *On Measures of Information and Their Characterizations*. New-York: Academic Press.

Akemann, G., Baik, J., & Di Francesco, P. (2015). *The Oxford Handbook of Random Matrix Theory*. Oxford, UK: Oxford University Press.

Akritas, A. A., Akritas, E. K., & Malaschonok, G. I. (1996). Various proofs of sylvester's (determinant) identity. *Mathematics and Computers in Simulation*, *42*(4-6), 585–593.

Ali, S. M. & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society B*, *28*(1), 131–142.

Amari, S.-I. & Nagaoka, H. (2000). *Methods of Information Geometry*. Providence, Rhode Island, USA: Oxford University Press.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, *65*(331), 1248–1255.

Andersen, E. S. & Larsen, M. E. (1994). Combinatorial summation identities. In *Analysis: Algebra and Computers in Mathematical Research: Proceedings of the Twenty-first Nordic Congress of Mathematicians*, Lecture Notes in Pure and Applied Mathematics, (pp. 1–23). CRC Press.

Anderson, G. W., A., Guionnet, & Zeitouni, O. (2010). *An Introduction to Random Matrices*. New-York, USA: Cambridge University Press.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Hoboken, New-Jersey, USA: John Wiley & Sons.

Ando, A. & Kaufman, G. M. (1965). Bayesian analysis of the independent multinormal process – neither mean nor precision known. *Journal of the American Statistical Association*, *60*(309), 347–358.

Andrew, G. E. & Berndt, B. C. (2013). *Ramanujan's Lost Notebook-Part IV*. New-York: Springer.

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, *36*(1), 99–102.

Andrews, G. E., Askey, R., & Roy, R. (1999). *Special Functions*. Cambridge, UK: Cambridge University

Press.

Appell, P. (1925). *Sur les fonctions hypergéométriques de plusieurs variables, les pôlynomes d'Hermite et autres fonctions hypersphériques dans l'hyperespace*, volume fascicule 3. Paris: Mémorial des Sciences Mathématiques; Gauthier-Villars.

Arellano-Valle, R. B., del Pino, G., & Iglesias, P. (2006). Bayesian inference in spherical linear models: robustness and conjugated analysis. *Journal of Multivariate Analysis*, *97*(1), 179–197.

Arimoto, S. (1971). Information-theoretical considerations on estimation problems. *Information and control*, *19*(3), 181–194.

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20.

Arndt (2001). *Information Measures: Information and its Description in Sciences and Ingineering*. Berlin: Springer Verlag.

Ash, R. B. & Doléans-Dade, C. A. (1999). *Probability and Measure Theory* (2nd ed.). San Diego, CA, USA: Academic Press.

Askey, R. A. (1975). *Orthogonal Polynomials and Special Functions*. SIAM.

Athreya, K. B. & Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. New-York: Springer.

Balakrishnan, N. & Basu, A. P. (1995). *The Exponential Distribution: theory, Methods and Applications*. Amsterdam, The Netherlands: Gordon an Breach Publishers.

Balakrishnan, N., Brito, M. R., & Quiroz, A. J. (2007). A vectorial notion of skewness and its use in testing for multivariate symmetry. *Communications in Statistics - Theory and Methods*, *36*(9), 1757–1767.

Balakrishnan, N. & Scarpa, B. (2012). Multivariate measures of skewness for the skew-normal distribution. *Journal of Multivariate Analysis*, *104*(1), 73–87.

Barnard, G. A. (1958). Studies in the history of probability and statistics: IX. Tomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, *45*(3-4), 293–295.

Barone, J. & Novikoff, A. (1978). A history of the axiomatic formulation of probability from Borel to Kolmogorov: Part I. *Archive for History of Exact Sciences*, *18*(2), 123–190.

Barron, A. R. (1984). Monotonic central limit theorem for densities. Technical report no. 50, Department of Statistics, Stanford University.

Barron, A. R. (1986). Entropy and the central limit theorem. *The Annals of Probability*, *14*(1), 336–342.

Bartlett, M. S. (1934a). On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, *53*, 260–283.

Bartlett, M. S. (1934b). The vector representation of a sample. *Proceedings of the Cambridge Philosophical Society*, *30*(3), 327–340.

Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Processing*, *18*(4), 349–369.

Basseville, M. (2013). Divergence measures for statistical data processing – an annotated bibliography.

Signal Processing, *93*(4), 621–633.

Bausson, S., Pascal, F., Forster, P., Ovarlez, J.-P., & Larzabal, P. (2007). First- and second-order moments of the normalized sample covariance matrix of spherically invariant random vectors. *IEEE Signal Processing Letters*, *14*(6), 425–428.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics*, *50*(4), 495–510.

Bell, E. T. (1927-1928). Partition polynomials. *The Annals of Mathematics*, *29*(1/4), 38–46.

Bellhouse, D. (2005). Decoding cardano's liber de ludo aleae. *Historia Mathematica*, *32*(2), 180–202.

Ben-Tal, A., Bornwein, J. M., & Teboulle, M. (1992). Spectral estimation via convex programming. In F. Y. Phillips & J. J. Rousseau (Eds.), *Systems and Management Science by Extremal Methods* chapter 18, (pp. 275–290). Springer.

Ben-Tal, A., Charnes, A., & Teboulle, M. (1989). Entropic means. *Journal of Mathematical Analysis and Applications*, *139*(2), 537–551.

Bengtsson, I. & Życzkowski, K. (2006). *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge: Cambridge University Press.

Bercher, J.-F. (2012). On a (β, q)-generalized Fisher information and inequalities invoving q-Gaussian distributions. *Journal of Mathematical Physics*, *53*(6), 063303.

Bercher, J.-F. (2013). On multidimensional generalized Cramér-Rao inequalities, uncertainty relations and characterizations of generalized q-Gaussian distributions. *Journal of Physics A*, *46*(9), 095303.

Berkane, M. & Bentler, P. (1986). Moments of elliptically distributed random variates. *Statistics & Probability Letters*, *4*(6), 333–335.

Berlekamp, E. R. (Ed.). (1974). *Key Papers in the Development of Coding Theory*. IEEE Press.

Bernoulli, J. (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilæ reticularis*. Basel, Switzeland: Thurneysen Brothers.

Bernstein, S. (1929). Sur les fonctions absolument monotones. *Acta Mathematica*, *52*(1), 1–66.

Besson, O. & Abramovich, Y. I. (2013). On the Fisher information matrix for multivariate elliptically contoured distributions. *IEEE Signal Processing Letters*, *20*(11), 1130–1133.

Bhatia, R. (1997). *Matrix Analysis*. New-York: Springer Verlag.

Bhatia, R. (2007). *Positive Definite Matrices*. Princeton: Princeton University Press.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99–109.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, *7*(4), 401–406.

Bienaymé, I.-J. (1853a). Considérations à l'appui de la découverte de Laplace sur la loi de probabilité

dans la méthode des moindres carrées. *Comptes Rendus de l'Académie des Sciences.*, *37*, 158–176.

Bienaymé, J. (1838). *Mémoires présenté à l'Académie Royale des Sciences de l'Institut de France*, volume 5, chapter Mémoire sur la probabilité des résultats moyens des observations ; démonstation directe de la règle de Laplace, (pp. 513–558). Paris: Imprimerie Royale.

Bienaymé, J. (1853b). Remarques sur les différences qui distinguent l'interpolation de M. Cauchy de la méthode des moindres carrés et qui assurent la supériorité de cette méthode. *Journal de Mathématiques Pures et Appliquées*, *18*, 299–308.

Billingsley, P. (2012). *Probability and Measure* (3rd ed.). Hoboken, NJ, USA: John Willey & Sons.

Bilodeau, M. & Brenner, D. (1999). *Theory of Multivariate Statistics*. New-York: Springer.

Blachman, N. M. (1965). The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, *11*(2), 267–271.

Blake, I. F. & Thomas, J. B. (1968). On a class of processes arising in linear estimation theory. *IEEE Transactions on Information Theory*, *14*(1), 12–16.

Bochner, S. (1932). Ein konvergenzsatz für mehrvariablige fouriersche integrale. *Mathematische Zeitschrift*, *34*(1), 440–447.

Bochner, S. (1959). Monotonic functions, Stieltjes integrals and harmonic analysis. In *Lectures on Fourier Integrals* (pp. 292–331). Princeton University Press.

Boekee, D. E. & van der Lubbe, J. C. A. (1979). Some aspects of error bounds in feature selection. *Pattern Recognition*, *11*(5-6), 353–360.

Boekee, D. E. & van der Lubbe, J. C. A. (1980). The R-norm information measure. *Information and Control*, *45*(2), 136–155.

Bogachev, V. I. (2007a). *Measure Theory*, volume I. Berlin: Springer.

Bogachev, V. I. (2007b). *Measure Theory*, volume II. Berlin: Springer.

Boltzmann, L. (1877). Über die beziehung zwischen dem zweiten hauptsatze der mechanischen wärmetheorie und der wahrscheinlichkeitsrechnung resp. den sätzen über das wärmegleichgewicht. *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien*, *76*, 373–435.

Boltzmann, L. (1896). *vorlesungen über Gastheorie - I*. Leipzig, Germany: Verlag von Johann Ambrosius Barth.

Boltzmann, L. (1898). *vorlesungen über Gastheorie - II*. Leipzig, Germany: Verlag von Johann Ambrosius Barth.

Bombrun, L. & Beaulieu, J.-M. (2008). Fisher distribution for texture modeling of polarimetric SAR data. *IEEE Geoscience and Remote Sensing Letters*, *5*(3), 512–516.

Borel, E. (1898). *Le0.98628cons sur la théorie des fonctions*. Paris: Gauthier-Villars et fils.

Borel, E. (1909). *Éléments de la théorie des probabilités*. Paris: A. Hermann & fils.

Bouniakowsky, V. (1859). Sur quelques inégalités concernant les intégrales ordinaires et les intégrales aux différences finies. *Mémoires de l'Académie Impériale des Sciences de Saint-Petersbourg*,

I(9).

Boutet de Monvel, A., Pastur, L., & Shcherbina, M. (1995). On the statistical mechanics approach in the random matrix theory: Integrated density of states. *Journal of Statistical Physics*, *79*(3-4), 585–611.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problem in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*(3), 200–217.

Breit, G. & Wigner, E. (1936). Capture of slow neutrons. *Physical Review*, *49*(7), 519–531.

Brémaud, P. (1988). *An Introduction to Probabilitic Modeling*. New-York: Springer.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory*. Philadelphia, USA.

Brockmeyer, E., Halstrøm, H. L., & Jensen, A. (1948). *The life and works of A. K. Erlang (Transactions of the Danish Academy of Technical Sciences)*, chapter The Theory of Probabilities and Telephone Conversations, (pp. 131–137). Number 2. Akademiet foe de Tekniske Videnskaber.

Brockwell, P. J. & Davis, R. A. (1987). *Time Series: Theory and Methods* (2nd ed.). New-York: Springer Verlag.

Bruce, R. V. (1990). *Bell: Alexander Bell and the Conquest of Solitude*. Ithaca, New York, USA: Cornell University Press.

Bullet, S., Fearn, T., & Smith, F. (2017). *Analysis and Mathematical Physics*. London: World Scientific.

Burbea, J. & Rao, C. R. (1982). On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, *28*(3), 489–495.

Burg, J. P. (1967). Maximum entropy spectral analysis. In *Proceedings of the 37th Meeting of the Society of Exploration Geophysicists*, Oklahoma City, Oklahoma.

Burg, J. P. (1972). The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, *37*(2), 375–376.

Burg, J. P. (1975). *Maximum entropy spectral analysis*. PhD thesis, Department of Geophysics, Stanford University, Stanford University, Stanford, CA.

Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, *11*(3), 368–385.

Cambini, A. & Martein, L. (2009). *Generalized Convexity and Optimization: Theory and Applications*. Heidelberg: Springer Verlag.

Cardano, J. (1663). *Liber de ludo aleae, en "Opera Omnia"*, volume 1, (pp. 262–276). Lyon: Cura Caroli Sponii.

Carlson, F. (1921). Über ganzwertige funktionen. *Mathematische Zeitschrift*, *11*(1-2), 1–23.

Carmeli, M. (1983). *Statistical Theory and Random Matrices*. New-York, USA: Marcel Dekker Inc.

Caro-Lopera, F. J., Farías, G. G., & Balakrishnan, N. (2016). Matrix-variate distribution theory under elliptical models-4: Joint distribution of latent roots of covariance matrix and the largest and smallest latent roots. *Journal of Multivariate Analysis*, *145*, 224–235.

Carrier, G. F., Krook, M., & Pearson, C. E. (2005). *Function of a Complex Variable: Theory and Technique*. Philadelphia: SIAM.

Cauchy, A.-L. (1821). *Cours d'analyse de l'école royale polytechnique*, volume 1: analyse algébrique. Paris: Imprimerie royale (digital version, Cambrige, 2009).

Cauchy, A.-L. (1853a). Mémoire sur l'interpolation, ou remarques sur les remarques de m. jules bienaymé. *Comptes Rendus de l'Académie des Sciences de Paris*, *37*, 64–68.

Cauchy, A.-L. (1853b). Sur la nouvelle méthode d'interpolation comparée à la méthode des moindres carrés. *Comptes Rendus de l'Académie des Sciences de Paris*, *37*, 100–109.

Cencov, N. N. (1982). *Statistical Decision Rules and Optimal Inference*. Providence, Rhode Island, USA: American Mathematical Society.

Chenciner, A. (2017). La force d'une idée simple. *Gazette de la Société de Mathématiques Fran0.98628caise*, *152*, 16–22.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, *23*(4), 493–507.

Chitour, Y. & Pascal, F. (2008). Exact maximum likelihood estimates for SIRV covariance matrix: Existence and algorithm analysis. *IEEE Transactions on Signal Processing*, *56*(10), 4563–4573.

Cholesky, A.-L. (1910 (2005)). Sur la résolution numérique des systèmes d'équations linéaires. *Bulletin de la société des amis de la bibliothèque de l'École polytechnique (SABIX)*, (39). Manuscript notes.

Cholewinski, F., Haimo, D., & Nussbaum, A. (1970). A necessary and sufficient condition for the representation of a function as a Hankel-Stieltjes transform. *Studia Mathematica*, *36*(3), 269–274.

Chong, K. M. (1974). Some extensions of a theorem of Hardy, Littlewood and Pólya and their applications. *Journal canadien de mathématiques*, *26*, 1321–1340.

Chu, K.-C. (1973). Estimation and decision for linear systems with elliptical random processes. *IEEE Transactions on Automatic Control*, *18*(5), 499–505.

Clavier, A. G. (1948). Evaluation of transmission efficiency according to Hartley's expression of information content. *Technical Journal of the International Telephone and Telegraph Corporation and Associate Companies*, *25*(4), 414–420.

Cohen, M. (1968). The Fisher information and convexity. *IEEE Transactions on Information Theory*, *14*(4), 591–592.

Cohn, D. L. (2013). *Measure Theory* (2nd ed.). New-York: Springer.

Costa, J. A., Hero III, A. O., & Vignat, C. (2003). On solutions to multivariate maximum α-entropy problems. In Rangarajan, A., Figueiredo, M. A. T., & Zerubia, J. (Eds.), *4th International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, volume 2683 of *Lecture Notes in Computer Sciences*, (pp. 211–226)., Lisbon, Portugal. Springer Verlag.

Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Cox, D. R. (1962). *Renewal Theory*. London: Methuen & Co.

Cramér, H. (1946). *Mathematical Methods of Statistics*. New-York: Princeton University Press.

Cressie, N. & Pardo, L. (2000). Minimum ϕ-divergence estimator and hierarchical testing in loglinear models. *Statistica Sinica*, *10*(3), 867–884.

Cressie, N. & Read, L. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B*, *46*(3), 440–464.

Csiszár, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, *8*(1-2), 85–108.

Csiszàr, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, *2*, 299–318.

Csiszár, I. (1974). Information measures: A critical survey. In *Transactions of the Seventh Prague Conference on Information Theory*, volume B, (pp. 73–86)., Prague, 18-23 august.

Csiszàr, I. (1991). Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, *19*(4), 2031–2066.

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, *68*(1-2), 161–186.

Csiszàr, I. & Matúš, F. (2012). Generalized minimizers of convex integral functionals, Bregman distance, Pythgorean identities. *Kybernetica*, *48*(4), 637–689.

Csiszár, I. & Shields, P. C. (2004). Information theory and statistics: A tutorial. *Foundations and Trends^{TM} in Communications and Information Theory*, *1*(4), 417–528.

Curie, P. (1895). Les propriétés magnétiques des corps à diverses températures. *Annales de Chimie et de Physique*, *7*(V), 289–405.

Darmois, G. (1935). Sur les lois de probabilités à estimation exhaustive. *Comptes rendus de l'Acadéie des Sciences*, *200*, 1265–1966.

Darmois, G. (1945). Sur les limites de la dispersion de certaines estimations. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *13*(1/4), 9–15.

Daróczy, Z. (1970). Generalized information functions. *Information and Control*, *16*(1), 36–51.

Daróczy, Z. & Járai, A. (1979). On the measurable solution of a functional equation arising in information theory. *Acta Mathematica Academiae Scientiarum Hungaricae*, *34*(1-2), 105–116.

David, H. A. & Edwards, A. W. F. (2001). *Annotated Readings in the History of Statistics*. New York, USA: Springer Verlag.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, *68*(1), 265–274.

de Laplace, P. S. (1820). *Théorie analytique des Probabilités* (3ème ed.). Paris: Mme Ve Courcier, Imprimeur-Libraire pour les Mathématiques.

de Moivre, A. (1710). De mensura sortis, seu, de probabilitate eventuum in ludis a casu fortuito pen-

dentibus. *Philosophical Transactions of the Royal Society of London*, *27*(329), 213–264.

de Moivre, A. (1730). *Miscellanea analytica de seriebus et quadraturis*. London: Londini: J. Tonson & J. Watts.

de Moivre, A. (1733). Approximatio ad summam terminorum binomii $(a + b)^n$ in seriem expansi. Very few copies published privately in London (see also The Doctrine of Chance).

de Moivre, A. (1756). *The Doctrine of Chances : or, a method for calculating the probabilities of events in play* (3rd ed.). London: AMS Chelsea Publishing.

de Montmort, P. R. (1713). *Essay d'analyse sur les jeux de hazard* (2nd ed.). Paris: Jacque Quillau, Imprimeur Juré Libraire de l'Université.

de Morgan, A. (1838). *An Essay on Probabilities and on their application to Life Contingencies and Insurance Offices*. London, UK: Longman, Orme, Brown, Green & Longmans.

Dembo, A., Cover, T. M., & Thomas, J. A. (1991). Information theoretic inequalities. *IEEE Transactions on Information Theory*, *37*(6), 1501–1518.

Deming, W. E. (1933). De moivre's "miscellanea analytica", and the origin of the normal curve. *Nature*, *132*(3340), 713–713.

Dickey, J. M. (1967). Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *The Annals of Mathematical Statistics*, *38*(2), 511–518.

Doob, J. L. (1936). Statistical estimation. *Transactions of the American Mathematical Society*, *39*(3), 410–421.

Dutka, J. (1991). The early history of the factorial function. *Archive for History of Exact Sciences*, *43*(3), 225–249.

(E. D. Sylla, Translator), J. B. (1713). *The Art of Conjecturing - Together with a "Letter to a Friend on Set in Court Tennis"*. Johns Hopkins University Press.

Eaton, M. L. (1981). On the projections of isotropic distributions. *Annals of Statistics*, *9*(2), 391–400.

Ebeling, W., Molgedey, L., Kurths, J., & Schwarz, U. (2000). Entropy, complexity, predictability and data analysis of time series and letter sequences. In *Theory of Disaster* (A. Bundle and H.-J. Schellnhuber ed.). Berlin: Springer Verlag.

Edelman, A. & Rao, N. R. (2005). Random matrix theory. *Acta Numerica*, *14*, 233–297.

Edgeworth, F. Y. (1908). On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, *71*(3, 6 & 7), 381–397, 499–512 & 499–512.

Elias, P. (1957). List decoding for noisy channels. Technical Report 335, Research Laboratory of Electronics, MIT, MIT, Cambridge, MA.

Endres, D. & Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, *49*(7), 1858–1860.

Erdélyi, A. (1940). Integration of a certain system of linear partial differential equations of hypergeometric type. *Proceedings of the Royal Society of Edinburgh*, *59*, 224–241.

Eriksson, J. & Koivunen, V. (2006). Complex random vectors and ICA models: Identifiability, uniqueness,

and separability. *IEEE Transactions on Information Theory*, *52*(3), 1017–1029.

Eriksson, J., Ollila, E., & Koivunen, V. (2009). Statistics for complex random variables revisited. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 3565–3568)., Taipei, Taiwan.

Erlang, A. K. (1909). Sandsynlighedsregning og telefonsamtaler. *Nyt Tidsskrift for Matematik*, *20*(B), 33–39.

Erlang, A. K. (1925). Calcul des probabilités et conversations téléphoniques. *Revue générale de l'Electricité*, *18*(8), 305–309.

Esteban, M. D. (1997). A general class of entropy statistics. *Applications of Mathematics*, *42*(3), 161–169.

Euler, L. (1741). Observationes analyticae varias de combinationibus. *Commentarii academiae scientiarum Petropolitanae*, *13*, 64–93.

Euler, L. (1750). De partitione numerorum. *Novi Commentarii academiae scientiarum Petropolitanae*, *3*, 125–169.

Euler, L. (1768). *Lettres à une princesse d'Allemagne sur divers sujets de physique & de philosophie*, volume 2. Saint Petersbourg, Rusia: Académie Impériale des Sciences de Saint Petersbourg.

Euler, L. (1769). *Institutiones Calculi Integralis. Volulen Secundum.* Saint Petersburg, Russia: Petropoli Impensis Academiae Imperialis Scientiarum.

Faà de Bruno, F. (1857). Note sur une nouvelle formule de calcul differentiel. *The Quarterly Journal of Pure and Applied Mathematics*, *1*, 359–360.

Faà di Bruno, F. (1855). Sullo sviluppo delle funzioni. *Annali di Scienze Matematiche e Fisiche*, *6*, 479–480.

Fadeev, D. K. (1956). On the concept of entropy of a finite probabilistic scheme (russian). *Uspekhi Matematicheskikh Nauk*, *11*(1(67)), 227–231.

Fadeev, D. K. (1958). *Fundations in Information Theory*, chapter On the concept of entropy of a finite probabilistic scheme (English traduction). New-York: McGraw-Hill.

Fang, K. T., Kotz, S., & Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Monographs on statistics and probability 36. London: Chapman & Hall.

Fang, K.-T. & Li, R. (1999). Bayesian statistical inference on elliptical matrix distributions. *Journal of Multivariate Analysis*, *70*(1), 66–85.

Fano, R. M. (1949). The transmission of information. Technical Report 65, Research Laboratory of Electronics, MIT, MIT, Cambridge, MA.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (3 ed.)., volume 1. New-York: John Wiley & Sons, Inc.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2. New-York: John Wiley & Sons, Inc.

Ferentinos, K. (1982). On Tchebycheff's type inequalities. *Trabajos de Estadística e Investigación*

Operativa, *33*(1), 125–132.

Ferreri, C. (1980). Hypoentropy and related heterogeneity, divergence and information measures. *Statistica*, *2*, 155–167.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, *222*(594-604), 309–368.

Fisher, R. A. (1925a). Applications of “Student’s” distributions. *Metron*, *5*(3), 90–104.

Fisher, R. A. (1925b). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725.

Fisher, R. A. (1930). Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society*, *s2-30*(1), 199–238.

Flandrin, P. & Rioul, O. (2016). Laplume, sous le masque.

Forcina, A. (2017). Gini’s early investigation about the distribution of sexes in human births. *Statistica Applicata - Italian Journal of Applied Statistics*, *28*(2-3), 223–238.

Fourier, J. (1822). *Théorie Analitique de la Chaleur*. Paris: Firmin Didot, père et fils.

Fréchet, M. (1943). Sur l’extension de certaines evaluations statistiques au cas de petits echantillons. *Revue de l’Institut International de Statistique / Review of the International Statistical Institute*, *11*(3/4), 182–205.

Frieden, B. R. (2004). *Science from Fisher Information: A Unification*. Cambridge, UK: Cambridge University Press.

Frigyik, B. A., Srivastava, S., & Gupta, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Infomation Theory*, *54*(11), 5130–5139.

Gallager, R. (1978). Variations on a theme by Huffman. *IEEE Transactions on Information Theory*, *24*(6), 668–674.

Gallager, R. (2001). Claude E. Shannon: a retrospective on his life, work, and impact. *IEEE Transactions on Information Theory*, *47*(7), 2681–2695.

Galton (1877a). Typical laws of heredity. *Journal of the Royal Institution of Great Britain*, *8*, 282–301.

Galton, F. (1877b). Typical laws of heredity. *Nature*, *15*, 492–495.

Galton, F. (1889). *Natural Inherence*. London, UK: McMillan.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hambourg, Germany: Perthes et Besser.

Gauss, C. F. (1810). *Disquisitio de elementis ellipticis Palladis, ex oppositionibus annorum 1803, 1804, 1805, 1807, 1808, 1809*. Göttingen, Germany: Gottingae : Apud H. Dieterich.

Gelfand, I. M. & Fomin, S. V. (1963). *Calculus of Variations*. Englewood Cliff, NJ, USA: Prentice Hall.

Gel’fand, I. M. & Shilov, G. E. (1964). *Generalized Functions*, volume 1: Properties and Operations. New-York: Academic Press.

Gel’fand, I. M. & Shilov, G. E. (1968). *Generalized Functions*, volume 2: Spaces of Fundamental and Generalized Functions. New-York: Academic Press.

Gibbs, J. W. (1902). *Elementary Principle in Statistical Mechanics*. Cambridge, USA: University Press - John Wilson and son.

Gini, C. (1911). Considerazioni sulle probabilità posteriori e applicazioni al rapporto dei sessi nelle nascite umane. *Studi Economico-Giuridici della Università de Cagliari. Anno III (reproduced in Metron, 15(1-4): 133-171, 1949)*, 5–41.

Golberg, R. R. (1961). *Fourier Transforms*. Cambridge University Press.

Goldman, J. (1976). Detection in the presence of spherically symmetric random vectors. *IEEE Transactions on Information Theory*, *22*(1), 52–59.

Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of Mathematical Statistics*, *34*(1), 152–177.

Gradshteyn, I. S. & Ryzhik, I. M. (2015). *Table of Integrals, Series, and Products* (8th ed.). San Diego: Academic Press.

Graham, R. L., Knuth, D. E., & Patashnik, O. (1994). *Concrete Mathematics* (2nd ed.). Reading: Addison Wesley Longman.

Gray, A. & Mathew, G. B. (1895). *A treatise on Bessel functions and their application to physics*. Macmillan and co.

Grübel, R. & Rocke, D. M. (1990). On the cumulants of affine equivariant estimators in elliptical families. *Journal of Multivariate Analysis*, *35*(2), 203–222.

Guo, D., Shamai, S., & Verdú, S. (2005). Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, *51*(4), 1261–1282.

Gupta, A. K. & Nagar, D. K. (1999). *Matrix Variate Distributions*. Chapman and Hall.

Gupta, A. K. & Varga, T. (1995). Some inference problems for matrix variate elliptically contoured distributions. *Statistics*, *26*(3), 219–229.

Gupta, H. C. & Sharma, B. D. (1976). On non-additive measures of inaccuracy. *Czechoslovak Mathematical Journal*, *26*(4), 584–595.

Gupta, R. D. & Richards, D. S. P. (2001). The history of the Dirichlet and Liouville distributions. *International Statistical Review*, *69*(3), 433–446.

Haas, A. (1929). *Introduction to theoretical physics* (2nd ed.)., volume 1. London, UK: Constable and Co Limited.

Hadamard, J. (1893). Etude sur les propriétés des fonctions entières et en particulier d'une function considérée par Riemann. *Journal de Mathématiques Pures et Appliquées*, *58*(9), 171–215.

Haghighatshoar, S., Abbe, E., & Telatar, I. E. (2014). A new entropy power inequality for integer-valued random variables. *IEEE Transactions on Information Theory*, *60*(7), 3787–3796.

Hald, A. (1984). Nicholas Bernoulli's theorem. *International Statistical Review*, *52*(1), 93–99.

Hald, A. (1990). *History of Probability and Statistics and Their Applications before 1750*. John Wiley & Sons, Inc.

Hald, A. (2006). *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. New-

York, USA: Springer.

Halmos, P. R. (1950). *Measure Theory*. New-York: Springer.

Hardy, G., Littlewood, J. E., & Pólya, G. (1952). *Inequalities* (2nd ed.). Cambridge, UK: Cambridge University Press.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1929). Some simple inequalities satisfied by convex functions. *Messenger of Mathematics*, *58*, 145–152.

Hardy, M. (2006). Combinatorics of partial derivatives. *Electronic Journal of Combinatorics*, *13*(1), R1.

Harremoës, P. & Vignat, C. (2003). An entropy power inequality for the binomial family. *Journal of Inequalities in Pure and Applied Mathematics*, *4*(5), 93.

Hartley, R. V. L. (1928). Transmission of informations. *The Bell System Technical Journal*, *7*(3), 535–563.

Harville, D. A. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag.

Hausdorff, F. (1901). Beiträge zur wahrscheinlichkeitsrechnung. *Berichte über die Verhandlungen der Köeniglich Sächsische Akademie der Wissenschaften zu Leipzig*, *53*(1), 152–178.

Hausdorff, F. (1921a). Summationsmethoden und momentfolgen. I. *Mathematische Zeitschrift*, *9*(1-2), 74–109.

Hausdorff, F. (1921b). Summationsmethoden und momentfolgen. II. *Mathematische Zeitschrift*, *9*(3-4), 280–299.

Havrda, J. & Charvát, F. (1967). Quantification method of classification processes: Concept of structural α-entropy. *Kybernetika*, *3*(1), 30–35.

Hellinger, E. (1909). Neue bergründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die riene und angewandte Mathematik*, 210–271.

Helmert, F. R. (1875). Über die bestimmung des wahrscheinlichen fehlers aus einer endlichen anzahl wahrer beobachtungsfehler. *Zeitschrift für Mathematik und Physik*, (8), 300–303.

Helmert, F. R. (1876). Die genauigkeit der formel von peters zur berechnung des wahrscheinlichen beobachtungsfehlers directer beobachtungen gleicher genauigkeit. *Astronomische Nachrichten*, *88*(8-9), 113–131.

Hogg, R. V., McKean, J. W., & Craig, A. (2013). *Introduction to Mathematical Statistics* (7th ed.). Boston: Pearson.

Hölder, O. (1889). Ueber einen mittelwerthabsatz. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, *2*, 38–47.

Holevo, A. (2011). *Probabilistic and statistical aspects of quantum theory* (2nd ed.)., volume 1 of *Quaderni Monographs*. Pisa: Edizioni Della Normale.

Holevo, A. S. (1973). Bounds for the quantity of information transmitted by a quantum communication channel. *Problems of Information Transmission*, *9*(3), 177–183.

Horn, R. A. & Johnson, C. R. (2013). *Matrix Analysis* (2nd ed.). Cambridge University Press.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of*

the IRE, *40*(9), 1098–1101.

Humbert, P. (1922). The confluent hypergeometric functions of two variables. *Proceedings of the Royal Society of Edinburgh*, *41*, 73–96.

Hurst, S. (1995). The characteristic function of the student-t distribution. Technical Report Financial Mathematics Research Report No. FMRR006-95, Statistics Research Report No. SRR044-95, Center for Financial Mathematics, School of Mathematical Sciences, Australian National University, Canberra, Australia.

Huygens, C. (1657). De ratiociniis in ludo aleae. In *printed in Exercitationum Mathematicarum by F. Van Schooten*. Leiden, The Netherland: Elsevirii.

Ibarrola, P., Pardo, L., & Quesada, V. (1997). *Teoría de la Probabilidad*. Madrid: Síntesis.

Ibarrola, R. V. & Pérez, A. G. (2012). *Principios de Inferencia Estadística*. Madrid: Universidad Nacional de Educación a Distancia.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, *31*(1), 253–258.

Isogai, T. (1982). On a measure of multivariate skewness and a test for multivariate normality. *Annals of the Institute of Statistical Mathematics*, *34*(3), 531–541.

Jacob, J. & Protters, P. (2003). *Probability Essentials* (2nd ed.). Berlin: Springer.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical Review*, *108*(2), 171–190.

Jaynes, E. T. (1965). Gibbs vs Boltzmann entropies. *American Journal of Physics*, *33*(5), 391–398.

Jaynes, E. T. (1968). Prior probabilities. *IEEE transactions on systems science and cybernetics*, *4*(3), 227–241.

Jaynes, E. T. (1982). On the rational of maximum-entropy methods. *Proceedings of the IEEE*, *70*(9), 939–952.

Jeffrey (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A*, *186*(1007), 453–461.

Jeffrey, H. (1948). *Theory of Probability* (2nd ed.). Oxford: Clarendon.

Jeffrey, H. (1973). *Scientific Inference* (3rd ed.). Cambridge: Cambridge University Press.

Jensen, J. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, *30*(1), 175–193.

Jessen, B. (1931a). Bemærkninger om konvekse funktioner og uligheder imellem middelærdier. I. *Matematisk Tidsskrift. B*, 17–28.

Jessen, B. (1931b). Bemærkninger om konvekse funktioner og uligheder imellem middelærdier. II. *Matematisk Tidsskrift. B*, 84–95.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995a). *Continuous Univariate Distributions* (2nd ed.)., volume 1. New-York: John Wiley & Sons.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995b). *Continuous Univariate Distributions* (2nd ed.).,

volume 2. New-York: John Wiley & Sons.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New-York: John Wiley & Sons.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate Discrete Distributions* (2nd ed.). New-York: John Wiley & Sons.

Johnson, O. (2004). *Information Theory and The Central Limit Theorem*. London: Imperial college Press.

Johnson, O. & Vignat, C. (2007). Some results concerning maximum Rényi entropy distributions. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, *43*(3), 339–351.

Johnson, O. & Yu, Y. (2010). Monotonicity, thinning, and discrete versions of the entropy power inequality. *IEEE Transactions on Information Theory*, *56*(11), 5387–5395.

Kafka, P., Österreicher, F., & Vincze, I. (1991). On powers of f-divergences defining a distance. *Studia Scientiarum Mathematicarum Hungarica*, *24*(4), 415–422.

Kagan, A. (2001). A discrete version of the stam inequality and a characterization of the Poisson distributions. *Journal of Statistical Planning and Inference*, *92*(1-2), 7–12.

Kagan, A. & Smith, P. J. (1999). A stronger version of matrix convexity as applied to functions of Hermitian matrices. *Journal of Inequalities and Applications*, *3*(2), 143–152.

Kagan, A. & Yu, T. (2008). Some inequalities related to the Stam inequality. *Applications of Mathematics*, *53*(3), 195–205.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communications*, *15*(1), 52–60.

Kaniadakis, G. (2001). Non-linear kinetics underlying generalized statistics. *Physica A*, *296*(3-4), 405–425.

Kano, Y. (1994). Consistency property of elliptic probability density functions. *Journal of Multivariate Analysis*, *51*(1), 139–147.

Kapur, J. N. (1967). Generalized entropy of order α and type β. *The Mathematical Seminar*, *4*, 78–94.

Kapur, J. N. (1989). *Maximum Entropy Model in Sciences and Engineering*. New-Dehli: Wiley Eastern Limited.

Kapur, J. N. & Kesavan, H. K. (1992). *Entropy Optimization Principle with Applications*. San Diego: Academic Press.

Karamata, J. (1932). Sur une inegalité relative aux fonctions convexes. *Publications Mathématiques de l'Université de Belgrade*, *1*, 145–148.

Karush, J. (1961). A simple proof of an inequality of McMillan. *IEEE Transactions on Information Theory*, *7*(2), 118–118.

Kay, S. M. (1993). *Fundamentals for Statistical Signal Processing: Estimation Theory*. vol. 1. Upper Saddle River, NJ: Prentice Hall.

Keilson, J. & Steutel, F. W. (1974). Mixture of distributions, moment inequalities and measures of expo-

nentiality and normality. *Annals of Probability*, *2*(1), 112–130.

Kelker, D. (1971). Infinite divisibility and variance mixture of normal distributions. *The Annals of Mathematical Statistics*, *42*(2), 802–808.

Kemperman, J. H. B. (1971). Moment problem with convexity conditions I. In Rustagi, J. S. (Ed.), *Proceedings of the symposium on Optimizing Methods in Statistics*, (pp. 115–178)., Center for Tomorrow, the Ohio State University, USA.

Kendall, D. G. (1964). Functional equations in information theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *2*(3), 225–229.

Khinchin, A. I. (1957). *Mathematical foundations of information theory*. New-York: Dover Publications.

Kibria, B. M. G. & Joarder, A. H. (2006). A short review of multivariate t-distribution. *Journal of Statistical Research*, *40*(1), 59–72.

Kingman, J. F. C. (1972). On random sequences with spherical symmetry. *Biometrika*, *59*(2), 492–494.

Knuth, D. E. (1997). *The Art of Computer Programming* (3rd ed.)., volume 1 / fundamental algorithms. Reading: Addison Wesley Longman.

Koenigio, S. (1751). De universali principio æquilibrii & motus, in vi viva reperto, deque nexu inter vim vivam & actionem, utriusque minimo, dissertatio. *Nova acta eruditorum*, 125–135, 162–176.

Kollo, T. (2008). Multivariate skewness and kurtosis measures with an application in ICA. *Journal of Multivariate Analysis*, *99*(10), 2328–2338.

Kolmogorov, A. N. (1930). Sur la notion de la moyenne. *Atti della Reale Accademia Nazionale dei Lincei*, *12*, 388–391.

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability* (2nd ed.). New-York: Chelsea Publishing Company.

Kolmogorov, A. N. (1991). On the notion of mean. In V. M. Tikhomirov (Ed.), *Selected Works of A. N. Kolmogorov*, volume I: Mathematics and Mechanics (pp. 144–146). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kolmogorov, A. N. & Fomin, S. V. (1957). *Elements of the Theory of Function and Functional Analysis*, volume 1: Metric and Normed Spaces. Rochester, NY, USA: Graylock Press.

Kolmogorov, A. N. & Fomin, S. V. (1961). *Elements of the Theory of Function and Functional Analysis*, volume 2: Measure. The Lebesgue Integral. Hilbert Space. Rochester, NY, USA: Graylock Press.

Kondo, T. (1930). A theory of the sampling distribution of standard deviations. *Biometrika*, *22*(1-2), 36–64.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, *39*(3), 399–399.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous Multivariate Distributions* (2nd ed.)., volume 1: Models and Applications. New-York: John Wiley & Sons.

Kotz, S. & Nadarajan, S. (2004). *Multivariate t-Distributions and Their Applications*. Cambridge University Press.

Kraft Jr, L. G. (1949). A device for quantizing, grouping, and coding amplitude-modulated pulses. Master's thesis, Department of Electrical Engineering, MIT, Massachusetts Institute of Technology.

Krajči, S., Liu, C.-F., Mikeš, L., & Moser, S. M. (2015). Performance analysis of Fano coding. In *2015 IEEE International Symposium on Information Theory (ISIT)*, (pp. 1746–1750)., Hong-Kong, China.

Krishnaiah, P. R. (1976). Some recent developments on complex multivariate distributions. *Journal of Multivariate Analysis*, *6*(1), 1–30.

Krishnaiah, P. R. & a & J. Lin, K. (1986). Complex elliptically symmetric distributions. *Communications in Statistics*, *15*(12), 3693–3718.

Kuczma, M. (2009). *An Introduction to the Theory of Functional Equations and Inequalities: Cauchy's Equation and Jensen's Inequality* (2nd ed.). Basel: Birkhäuser.

Kullback, S. (1968). *Information Theory and Statistics*. Dover Publications.

Kullback, S. & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Kumar, P. & Chhina, S. (2005). A symmetric information divergence measure of the Csiszár's f-divergence class and its bounds. *Computers and Mathematics with Applications*, *49*(4), 575–588.

Kusolitsch, N. (2010). Why the theorem of Scheffé should be rather called a theorem of Riesz. *Periodica Mathematica Hungarica*, *61*(1-2), 225–229.

Lancaster, H. O. (1966). Forerunners of the Pearson χ2. *Australian Journal of Statistics*, *8*(3), 117–126.

Landau, L. D. & Lifshitz, E. M. (1976). *Mechanics* (3rd ed.)., volume 1 of *Course of theorical physics*. Oxford, UK: Butterworth-Heinemann.

Landau, L. D. & Lifshitz, E. M. (1980). *Statistical Physics* (3rd ed.). Part 1, volume 5. Oxford: Butterworth Heinemann.

Langius, J. C. (1712). *Nvclevs Logicae Weisianae*. Giessen: Henningius Müllerus.

Lapidoth, A. (2017). *A Foundation in Digital Communication* (2nd ed.). Cambridge, UK: Cambridge University Press.

Laplace, P. S. (1809a). Mémoire sur les approximations des formules qui sont fonctions de très grand nombres et sur leur application aux probabilités. *Mémoires de l'académie des sciences de Paris*, *Ière série T. X.*, 353–415.

Laplace, P. S. (1809b). Supplément aux mémoire sur les approximations des formules qui sont fonctions de très grand nombres et sur leur application aux probabilités. *Mémoires de l'académie des sciences de Paris*, *Ière série T. X.*, 559–565.

Laplace, P. S. (1812). *Théorie analytique des Probabilités*. Paris: Mme Ve Courcier, Imprimeur-Libraire pour les Mathématiques.

Laplace, P. S. (1814). *Théorie analytique des Probabilités* (2nd ed.). Paris: Mme Ve Courcier, Imprimeur-Libraire pour les Mathématiques.

Laplume, J. (1948). Sur le nombre de signaux discernables en présence de bruit erratique dans un sys-

tème de transmission à bande passance limitéee. *Comptes Rendus de l'Academie des Sciences*, *226*, 1348–1349. Séance du 26 avril.

Laurent, A. G. (1975). Applications of fractional calculus to spherical (radial) probability models and generalizations. In Dold, A. & Eckmann, B. (Eds.), *Fractional Calculus and Its Applications*, volume 457 of *Lecture Notes in Mathematics*, (pp. 256–266). Springer-Verlag.

Le Cam, L. (1986). The central limit theorem around 1935. *Statistical Science*, *1*(1), 78–91.

Lebesgue, H. (1904). *Le0.98628cons sur l'Intégration et la recherche des Fonctions Primitives*. Paris: Gauthier-Villars et fils.

Lebesgue, H. (1918). Remarques sur les théories de la mesure et de l'intégration. *Annales Scientifiques de l'Ecole Normale Supérieure*, *35*, 191–250.

Lee, P. M. (1964). On the axioms of information theory. *The Annals of Mathematical Statistics*, *35*(1), 415–418.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris, France: Firmin Didot, Libraie pour les Mathématiques, la Marine, L'architecture, et les Édition stéréotypes.

Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). New-York: Springer-Verlag.

Lejeune-Dirichlet (1839). Sur une nouvelle méthode pour la détermination des intégrales multiples. *Journal de Mathématiques Pures et Appliquées*, 164–168.

Lenz, W. (1920). Beiträge zum verständnis der magnetischen eigenschaften in festen körpern. *Physikalische Zeitschrift*, *21*, 613–615.

Leonov, V. P. & Shiryaev, A. N. (1959). On a method of calculation of semi-invariants. *Theory of Probability & Its Applications*, *4*(3), 319–329.

Lieb, E. H. (1975). Some convexity and subadditvity properties of entropy. *Bulletin of the American Mathematcal Society*, *81*(1), 1–13.

Lieb, E. H. (1978). Proof of an entropy conjecture of Wehrl. *Communications in Mathematical Physics*, *62*(1), 35–41.

Lieb, E. H. & Loss, M. (2001). *Analysis* (2nd ed.). Providence, Rhode Island: American Mathematical Society.

Liese, F. & Vajda, I. (1987). *Convex Statistical Distances*. Leipzig, Germany: Teubner.

Liese, F. & Vajda, I. (2006). On divergence and informations in statistics and information theory. *IEEE Transactions on Information Theory*, *52*(10), 4394–4412.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Lindeberg, J. W. (1922). Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung Mathematische Zeitschrift*, *15*, 211–225.

Lindhard, J. & Nielsen, V. (1971). Studies in statistical mechanics. *Det Kongelige Danske Videnskaber-*

nes Selskab Matematisk-fysiske Meddelelser, *38*(9), 1–42.

Liouville (1839). Note sur quelques intégrales définies. *Journal de Mathématiques Pures et Appliquées*.

Livan, G., Novaes, M., & Vivo, P. (2018). *Introduction to Random Matrices. Theory and Practice*. Cham, Switzerland: Springer-Verlag.

Lord, R. (1954). The use of the Hankel transform in statistics I. General theory and examples. *Biometrika*, *41*(1/2), 44–55.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209–219.

Lukacs, E. (1961). Recent developments in the theory of characteristic functions. In *Proceeding of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 2: Contributions to Probability Theory, (pp. 307–335). University of California Press, Berkeley, CA.

Lundheim, L. (2002). On Shannon and "Shannon's formula". *Telektronikk*, *98*(1), 20–29.

Lüroth, J. (1876). Vergleichung von zwei werthen des wahrscheinlichen fehlers. *Astronomische Nachrichten*, *87*(14), 209–220.

Lutwak, E., Lv, S., Yang, D., & Zhang, G. (2012). Extension of Fisher information and Stam's inequality. *IEEE Transactions on Information Theory*, *58*(3), 1319–1327.

Lutwak, E., Yang, D., & Zhang, G. (2005). Cramér-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information. *IEEE Transactions on Information Theory*, *51*(2), 473–478.

MacDonald, H. M. (1898). Zeroes of the Bessel functions. *Proceedings of the London Mathematical Society*, *s1-30*(1), 165–179.

Madiman, M. & Barron, A. (2007). Generalized entropy power inequalities and monotonicity properties of information. *IEEE Transactions on Information Theory*, *53*(7), 2317–2329.

Magnus, J. R. & Neudecker, H. (1979). The commutation matrix: Some properties and applications. *The Annals of Statistics*, *7*(2), 381–394.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (3rd ed.). New-York: John Wiley & Sons.

Malkovich, J. F. & Afifi, A. A. (1973). On tests for multivariate normality. *Journal of the American Statistical Association*, *68*(341), 176–179.

Mandel, L. & Wolf, E. (1995). *Optical coherence and quantum optics*. Cambridge University Press.

Marčenko, V. A. & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*(4), 457–483.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530.

Markov, A. (1884). *On certain applications of algebraic continued fractions*. PhD thesis, University of Saint Petersburg, St. Petersburg, Russia.

Marquis de Condorcet (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues*

à la pluralité des voix. Paris, France: Imprimerie Royale de Paris.

Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of Majorization and Its Applications* (2nd ed.). New-York: Springer Verlag.

Maxwell, J. C. (1867). On the dynamical theory of gases. *Philosophical Transactions of the Royal Society of London*, *157*, 49–88.

McGraw, D. K. & Wagner, J. F. (1968). Elliptically symmetric distributions. *IEEE Transactions on Information Theory*, *14*(1), 110–120.

McMillan, B. (1956). Two inequalities implied by unique decipherability. *IEEE Transactions on Information Theory*, *2*(4), 115–116.

Mehta, M. L. (2004). *Random Matrices* (3rd ed.). Amsterdam, The Netherlands: Elsevier Academic Press.

Menéndez, M. L., Morales, D., Pardo, L., & Salicrú, M. (1997). (h, ϕ)-entropy differential metric. *Applications of Mathematics*, *42*(1-2), 81–98.

Menéndez, M. L., Morales, D., Pardo, L., & Vajda, I. (1977). Testing in stationary models based on divergences of observed and theoretical frequencies. *Kybernetica*, *33*(5), 465–475.

Merhav, N. (2010). Statistical physics and information theory. *Foundations and Trends® in Communications and Information Theory*, *6*(1-2), 1–212.

Merhav, N. (2018). *Statistical Physics for Electrical Engineering*. Springer.

Mézard, M. & Montanari, A. (2009). *Information, Physics, and Computation*. New-York: Oxford University Press.

Mezzadri, F. & Snaith, N. C. (Eds.). (2008). *Recent Perspectives in Random Matrix Theory and Number Theory*. Cambridge, UK: Cambridge University Press.

Micheas, A. C., Dey, D. K., & Mardia, K. V. (2006). Complex elliptical distributions with application to shape analysis. *Journal of Statistical Planning and Inference*, *136*(9), 2961–2982.

Miller, R. E. (2000). *Optimization: Foundations and Applications*. New-York: John Wiley & Sons, inc.

Minkowski, H. (1910). *Geometrie der Zahlen*. Leipzig, Germany: Teubner.

Mittal, D. P. (1975). On additive and non-additive entropies. *Kybernetica*, *11*(4), 271–276.

Montagné, J.-C. B. (2008). *Transmissions. L'histoire des moyens de communication à distance depuis l'Antiquité jusqu'au milieu du xxe siècle*. Bagneux, JCB Montagné.

Móri, T. F., Rohatgi, V. K., & Székely, G. J. (1994). On multivariate skewness and kurtosis. *Theory of Probability & Its Applications*, *38*(3), 547–551.

Morimoto, T. (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, *18*(3), 328–331.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Mukhopadhyay, N. (2000). *Probability and Statistical Inference* (5th ed.)., volume 162 of *"Statistics: textbooks and monographs"*. New-York: Marcel Dekker.

Nagumo, M. (1930). Über eine klasse der mittelwerte. *Japanese journal of mathematics: transactions and abstracts*, *7*, 71–79.

Navarro, J. (2013). A very simple proof of the multivariate Chebyshev's inequality. *Communications in Statistics - Theory and Methods*, *45*(12), 3458–3463.

Neudecker, H. & Wansbeek, T. (1983). Some results on commutation matrices, with statistical applications. *Canadian Journal of Statistics*, *11*(3), 221–231.

Nielsen, F. & Boltz, S. (2011). The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, *57*(8), 5455–5466.

Nielsen, F. & Nock, R. (2017). Generalizing skew Jensen divergences and Bregman divergences with comparative convexity. *IEEE Signal Processing Letters*, *24*(8), 1123–1127.

Nikodym, O. (1930). Sur une généralisation des intégrales de M. J. Radon. *Fundamenta Mathematicae*, *15*(1), 131–179.

Novinger, W. P. (1972). Shorter notes: Mean convergence in L^p spaces. *Proceedings of the American Mathematical Society*, *34*(2), 627–628.

Nussbaum, A. E. (1973). On functions positive definite relative to the orthogonal group and the representation of functions as Hankel-Stieltjes transforms. *Transactions of the American Mathematical Society*, *175*, 389–389.

Ohya, M. & Petz, D. (1993). *Quantum Entropy and Its Use*. Berlin: Springer Verlag.

Olkin, I. & Pratt, J. W. (1958). A multivariate tchebycheff inequality. *The Annals of Mathematical Statistics*, *29*(1), 226–234.

Olkin, I. & Rubin, H. (1964). Multivariate Beta distributions and independence properties of the Wishart distribution. *The Annals of Mathematical Statistics*, *35*(1), 261–269.

Ollila, E., Eriksson, J., & Koivunen, V. (2011). Complex elliptically symmetric random variables – generation, characterization, and circularity tests. *IEEE Transactions on Signal Processing*, *59*(1), 58–69.

Ollila, E., Tyler, D. E., Koivunen, V., & Poor, H. V. (2012). Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Transactions on Signal Processing*, *60*(11), 5597–5625.

Onicescu, O. (1966). Energie informationnelle. *Comptes rendus de l'académie des Sciences. série 1, mathématiques*, *263*(3), 841–842.

Onsager, L. (1944). Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review*, *65*(3-4), 117–149.

Orsak, G. C. & Paris, B.-P. (1995). On the relationship between measures of discrimination and the performance of suboptimal detectors. *IEEE Transactions on Information Theory*, *41*(1), 188–203.

Osán, T. M., Bussandri, D. G., & Lamberti, P. W. (2018). Monoparametric family of metrics derived from classical Jensen-Shannon divergence. *Physica A*, *495*, 336–344.

Österreicher, F. (1996). On a class of perimeter-type distances of probability distributions. *Kybernetika*,

32(4), 389–393.

Österreicher, F. & Vajda, I. (2003). A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, *55*(3), 639–653.

Palomar, D. P. & Verdú, S. (2006). Gradient of mutual information in linear vector Gaussian channels. *IEEE Transactions on Information Theory*, *52*(1), 141–154.

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. Boca Raton, FL, USA: Chapman & Hall.

Pardo, M. C. (1999). On Burbea-Rao divergence based goodness-of-fit tests for multinomial models. *Journal of Multivariate Analysis*, *69*(1), 65–87.

Paris, R. B. & Kaminski, D. (2001). *Asymptotics and Mellin-Barnes integrals*. Cambridge, UK: Cambridge University Press.

Park, K. I. (2018). *Fundamentals of Probability and Stochastic Processes with Applications to Communications*. Cham, Switzerland: Springer-Verlag.

Pascal, B. (1679). *Varia opera Mathematica D. Petri de Fermat Senatoris Tolosae*, chapter Lettre de Monsieur Pascal à M. de Fermat, (pp. 184–188). Toulouse, France: Joannem Pech, Comitiorum Fuxenfiul Typographum, juxta Collegium PP. Societatis Jesu.

Payaró, M. & Palomar, D. P. (2009). Hessian and concavity of mutual information differential entropy, and entropy power in linear vector gaussian channels. *IEEE Transactions on Information Theory*, *55*(8), 3613–3628.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *186*, 343–414.

Pearson, K. (1905). “das fehlergesetz und seine verallgemeinerungen durch fechner und pearson.”. A rejoinder. *Biometrika*, *4*(1/2), 169–212.

Pearson, K. (1916). Mathematical contributions to the theory of evolution. XIX. second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society A*, *216*(538-548), 429–457.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25–45.

Pearson, K. (1924). Historical note on the origin of the normal curve of errors. *Biometrika*, *16*(3/4), 402–404.

Pearson, K., de Moivre, A., & Archibald, R. C. (1926). A rare pamphlet of Moivre and some of its discoveries. *ISIS A Journal of The History of Science Society*, *8*(4), 671–683.

Pearson, K. & Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. IV. on the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London A*, *191*, 229–311.

Peddada, S. D. & Richards, D. S. P. (1991). Proof of a conjecture of M. L. Eaton on the characteristic function of the Wishart distribution. *The Annals of Probability*, *19*(2), 868–874.

Perlman, M. D. (1974). Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. *Journal of Multivariate Analysis*, *4*(1), 52–65.

Petz, D. (2007). Bregman divergence as relative operator entropy. *Acta Mathematica Hungarica*, *116*(1-2), 127–131.

Pfanzagl, J. (1996). Studies in the history of probability and statistics XLIV. a forerunner of the t-distribution. *Biometrika*, *83*(4), 891–898.

Phillips, F. Y. & Rousseau, J. J. (Eds.). (1992). *Systems and Management Science by Extremal Methods*. Springer.

Phillips, P. C. B. (1988). The characteristic function of the Dirichlet and multivariate F distribution. discussion paper 985, Cowles Foundation for Research in Economics Yale University.

Picinbono, B. (1970). Spherically invariant and compound Gaussian stochastic processes. *IEEE Transactions on Information Theory*, *17*(1), 77–79.

Picinbono, B. (1996). Second-order complex random vectors and normal distributions. *IEEE Transactions on Signal Processing*, *44*(10), 2637–2640.

Pigeon, S. (2003). Huffman coding. In K. Sayood (Ed.), *Lossless Compression Handbook* chapter 4, (pp. 79–99). San Diego, CA: Academic Press.

Pinsky, M. A. (2009). *Introduction to Fourier Analysis and Wavelets*, volume 102. Providence, Rhode Island, USA: American Mathematical Society.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society*, *32*(4), 567–579.

Planck, M. (2015). *Eight Lectures on Theoretical Physics*. New-York: Columbia University Press.

Poincaré, H. (1899). Sur les propriétés du potentiel et sur les fonctions Abéliennes. *Acta Mathematica*, *22*(0), 89–178.

Poisson, M. (1823). Second mémoire sur la distribution de la chaleur dans les corps solides. *Journal de l'École Royale Polytechnique*, *19*(XII), 249–403.

Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Paris, France: Bachelier, Imprimeur-Librairie pour les mathématiques, la physique, etc.

Polya, G. (1920). Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, *8*(3-4), 171–181.

Poor, H. V. (1988). Fine quantization in signal detection and estimation. *IEEE Transactions on Information Theory*, *34*(5), 960–972.

Portilla, J., Strela, V., Wainwright, J., & Simoncelli, R. P. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, *12*(11), 1338–1351.

Poularikas, A. D. (1999). *The Handbook of Formulas and Table for Signal Processing*. Boca Raton: CRC Press.

Poularikis, A. D. (2010). *The Transforms and Applications Handbook* (3nd ed.). Boca Raton: CRC

Press.

Rangaswamy, M., Weiner, D., & Öztürk, A. (1993). Non-Gaussian random vector identification using spherically invariant random processes. *IEEE Transactions on Aerospace and Electronics Systems*, *26*(1), 111–124.

Rangaswamy, M., Weiner, D., & Öztürk, A. (1995). Computer generation of correlated non-Gaussian radar clutter. *IEEE Transactions on Aerospace and Electronics Systems*, *31*(1), 106–1116.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, *37*(3), 81–91.

Rao, C. R. (1992). Information and the accuracy attainable in the estimation of statistical parameters. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Foundations and Basic Theory*, volume I (pp. 235–247). New York: Springer.

Rao, C. R. & Wishart, J. (1947). Minimum variance and the estimation of several parameters. *Mathematical Proceedings of the Cambridge Philosophical Society*, *43*(2), 280–283.

Rathie, P. N. (1991). Unified (r, s)-entropy and its bivariate measures. *Information Sciences*, *54*(1-2), 23–39.

Remmert, R. (1991). *Theory of Complex Functions*. New York, USA: Springer.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Contributions to the Theory of Statistics, (pp. 547–561). University of California Press, Berkeley, CA.

Rényi, A. (2007). *Probability Theory*. Mineola, New-York: Dover Publications INC.

Riesz, F. (1928). Sur la convergence en moyenne. *Acta Scientiarum Mathematicarum (Szeged)*, *4*, 58–64.

Rioul, O. (2007). *Théorie de l'information et du codage*. Paris: Lavoisier.

Rioul, O. (2011). Information theoretic proofs of entropy power inequalities. *IEEE Transactions on Information Theory*, *57*(1), 33–55.

Rioul, O. (2017). Yet another proof of the entropy power inequality. *IEEE Transactions on Information Theory*, *63*(6), 3595–3599.

Rioul, O. & Flandrin, P. (2017). Le dessein de laplume. In *Colloque GRETSI*, Juan-les-Pins, France.

Rioul, O. & Magossi, J. (2014). On Shannon's formula and Hartley's rule: Beyond the mathematical coincidence. *Entropy*, *16*(12), 4892–4910.

Robert, C. P. (2007). *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation* (2nd ed.). New-York: Springer.

Rudin, W. (1991). *Functional Analysis* (2nd ed.). New-York: McGraw-Hill.

Salicrú, M. (1987). Funciones de entropía asociada a medidas de Csiszár. *Qüestiió*, *11*(3), 3–12.

Salicrú, M. (1994). Measures of information associated with Csiszár's divergences. *Kybernetica*, *30*(5), 563–573.

Salicrú, M., Menéndez, M. L., Morales, D., & Pardo, L. (1993). Asymptotic distribution of (h, ϕ)-entropies.

Communications in Statistics – Theory and Methods, *22*(7), 2015–2031.

Samorodnitsky, G. & Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes. Stochastic Models with infinite Variance*. New-York: Chapman & Hall.

Sasvári, Z. (2013). *Multivariate Characteristic Functions and Correlations Functions*. Berlin, Germany: Walter De Gruyter GmbH.

Sayood, K. (Ed.). (2003). *Lossless Compression Handbook*. San Diego, CA: Academic Press.

Scheffe, H. (1947). A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, *18*(3), 434–438.

Schœnberg, I. J. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics*, *39*(4), 811–841.

Schreier, P. & Scharf, L. (2003). Second-order analysis of improper complex random vectors and processes. *IEEE Transactions on Signal Processing*, *51*(3), 714–725.

Schur, I. (1923). Über eine klasse von mittelbildungen mit anwendungen auf die determinantentheorie. *Sitzungsberichte der Berliner Mathematischen Gesellschaft*, *22*, 9–20.

Schwartz, A. (1971). The structure of the algebra of Hankel transforms and the algebra of Hankel-Stieltjes transforms. *Canadian Journal of Mathematics*, *23*(2), 236–246.

Schwartz, A. L. (1969). The smoothness of Hankel transforms. *Journal of Mathematical Analysis and Applications*, *28*(3), 500–507.

Schwartz, L. (1966). *Théorie des distributions*. Paris: Hermann.

Schwarz, H. A. (1888). Ueber ein die flächen kleinsten flächeninhalts betreffendes problem der variationsrechnung. *Acta societatis scientiarum Fennicæ*, *15*, 315–362.

Seber, G. (2004). *Multivariate Observations*. Hoboken, New-Jersey, USA: John Wiley & Sons.

Selesnick, I. W. (2008). The estimation of Laplace random vectors in additive white Gaussian noise. *IEEE Transactions on Signal Processing*, *56*(8), 3482–3496.

Serrano Marugán, E. (2000). Etimología de algunos términos matemáticos. *Suma*, *35*, 87–96.

Shafer, G. & Vovk, V. (2006). The sources of Kolmogorov's grundbegriffe. *Statistical Science*, *21*(1), 70–98.

Shamai, S. & Wyner, A. (1990). A binary analog to the entropy-power inequality. *IEEE Transactions on Information Theory*, *36*(6), 1428–1430.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(4), 623–656.

Shannon, C. E. & Weaver, W. (1964). *The Mathematical Theory of Communication*. Urbana, USA: The University of Illinois Press.

Sharma, B. D. & Mittal, D. P. (1975). New non-additive measures of entropy for discrete probability distributions. *Journal of Mathematical Sciences*, *10*, 28 –40.

Sharma, B. D. & Taneja, I. J. (1975). Entropy of type (α, β) and other generalized measures in information theory. *Metrika*, *22*(1), 205–215.

Sharma, N., Das, S., & Muthukrishnan, S. (2011). Entropy power inequality for a family of discrete random variables. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, (pp. 1945–1949)., Saint Petersburg, Russia.

Sheynin, O. (1995). Helmert's work in the theory of errors. *Archive for History of Exact Sciences*, *49*(1), 73–104.

Shi, F. & Selesnick, I. W. (2007). An elliptically contoured exponential mixture model for wavelet based image denoising. *Applied and Computational Harmonic Analysis*, *23*(1), 131–151.

Shiryayev, A. N. (1984). *Probability*. New York, USA.

Shohat, J. (1929). Inequalities for moments of frequency functions and for various statistical constants. *Biometrika*, *21*(1-4), 361–375.

Sierpiński, W. (1918). Sur les définitions axiomatiques des ensembles mesurables. *Bulletin international de l'Académie des sciences de Cracovie: Série A. Classe des sciences mathématiques et naturelles – Sciences mathématiques*, 29–34.

Sierpiński, W. (1975). *Oeuvres choisies, Tome II: Théorie des ensembles et ses applications*. Warszawa, Poland: PWM Éditions scientifiques de Pologne.

Sierpiński, W. (1976). *Oeuvres choisies, Tome III: Théorie des ensembles et ses applications*. Warszawa, Poland: PWM Éditions scientifiques de Pologne.

Simpson, T. (1740). *The Nature and Laws of Chance. The Whole afetr a new, general and conspicuous Manner , and illustrated with a great Variety of Examples*. London, UK: Edward Cave.

Song, D.-K., Park, H.-J., & Kim, H.-M. (2014). A note on the characteristic function of multivariate t distribution. *Communications for Statistical Applications and Methods*, *21*(1), 81–91.

Spiegel, M. (1976). *Probabilidad y Estadística*. México: McGrew Hill.

Srivastava, H. M. & Karlsson, P. W. (1985). *Multiple Gaussian Hypergeometric Series*. Chichester: John John Wiley & Sons.

Srivastava, M. (1984). A measure of skewness and kurtosis and a graphical method for assessing multivariate normality. *Statistics & Probability Letters*, *2*(5), 263–267.

Stam, A. J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, *2*(2), 101–112.

Stanley, R. P. (1999). *Enumerative Combinatorics*, volume 2. New-York, USA.

Steele, J. M. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge: Cambridge University Press.

Steerneman, A. G. M. & van Perlo-ten-Kleij, F. (2005). Spherical distribution: Schœnberg (1938) revisited. *Expositiones Mathematicæ*, *23*(3), 281–287.

Stein, E. M. & Weiss, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press.

Stellato, B., Van Parys, B. P. G., & Goulart, P. J. (2017). Multivariate Chebyshev inequality with estimated mean and variance. *The American Statistician*, *71*(2), 123–127.

Stigler, S. M. (1974). Studies in the history of probability and statistics. XXXIII Cauchy and the witch of Agnesi: An historical note on the Cauchy distribution. *Biometrika*, *61*(2), 375–380.

Stirling, J. (1730). *Methodus Differentialis sive Tractatus de Summatione et Interpolatione Serierum Infinitarum*. Londini: Typis Gul. Bowyer; impensis G. Strahan.

Stix, G. (1991). Profile: Davis a. Huffman. *Scientific American*, *265*(3), 54–58.

Student (1907). On the error of counting with a haemacytometer. *Biometrika*, *5*(3), 351–360.

Student (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25.

Sultan, S. & Tracy, D. S. (1996). Moments of Wishart distribution. *Stochastic Analysis and Applications*, *14*(2), 237–243.

Sutradhar, B. C. (1986). On the characteristic function of multivariate Student t-distribution. *Canadian Journal of Statistics*, *14*(4), 329–337.

Sylvester, J. (1851). On the relation between the minor determinants of linearly equivalent quadratic functions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *1*(4), 295–305.

Tchébichev, P. (1867). Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées*, *12*, 177–184.

Teboulle, M. (1992). On Φ-divrgence and its applications. In F. Y. Phillips & J. J. Rousseau (Eds.), *Systems and Management Science by Extremal Methods* chapter 17, (pp. 255–273). Springer.

Teicher, H. (1960). On the mixture of distributions. *Annals of Mathematical Statistics*, *31*(1), 55–73.

Teodorescu, P. P. (2007). *Mechanical Systems, Classical Models*, volume 1: Particle Mechanics. New-York, USA: Springer-Verlag.

Thiele, T.N. (1889). Forelæsninger over Almindelig Iagttagelseslære. Reitzel, C. (1889). *Forelæsninger over Almindelig Iagttagelseslære*. Copenhagen, Denmark.

Thiele, T. N. (1903). *Theory of Observations*. London, UK.

Tison, C., Nicolas, J.-M., Tupin, F., & Maître, H. (2004). A new statistical model for Markovian classification of urban areas in high-resolution SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(10), 2046–2057.

Todros, K. & Tabrikian, J. (2007). Blind separation of independent sources using Gaussian mixture model. *IEEE Transactions on Signal Processing*, *55*(7), 3645–3658.

Toranzo, I. V., Zozor, S., & Brossier, J.-M. (2018). Generalization of the de Bruijn identity to general ϕ-entropies and ϕ-fisher informations. *IEEE Transactions on Information Theory*, on press.

Tribus, M. & McIrvine, E. C. (1971). Energy and information. *Scientific American*, *225*(3), 179–188.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, *52*(1-2), 479–487.

Tsallis, C. (1999). Nonextensive statistics: theoretical, experimental and computational evidences and connections. *Brazilian Journal of Physics*, *29*(1), 1–35.

Tulino, A. M. & Verdu, S. (2004). *Random Matrix Theory and Wireless Communications*. Now Publishers

Inc.

Tverberg, H. (1958). A new derivation of the information function. *Mathematica Scandinavica*, *6*, 297–298.

Tyler, D. E. (1982). Radial estimates and the test for sphericity. *Biometrika*, *69*(2), 429–436.

Vajda, I. (1968). Axioms for a-entropy of a generalized probability scheme. *Kybernetika*, *4*(2), 105–112.

Vajda, I. (1972). On the f-divergence and singularity of probability measures. *Periodica Mathematica Hungarica*, *2*(1-4), 223–234.

Vajda, I. (2009). On metric divergences of probability measures. *Kybernetika*, *45*(6), 885–900.

van Brakel, J. (1976). Some remarks on the prehistory of the concept of statistical probability. *Archive for History of Exact Sciences*, *16*(2), 119–136.

van Brunt, B. (2004). *The Calculus of Variations*. New-York: Springer Verlag.

van Dantzig, D. (1951). Une nouvelle généralisation de l'inégalité de Bienaymé. *Annales de l'institut Henri Poincaré*, *12*(1), 31–43.

van den Bos, A. (1995). The multivariate complex normal distribution-a generalization. *IEEE Transactions on Information Theory*, *41*(2), 537–539.

van den Bos, A. (2007). *Parameter Estimation for Scientists and Engineers*. Hoboken, New Jersey: John Wiley & Sons.

Varma, R. S. (1966). Generalization of Rényi's entropy of order α. *Journal of Mathematical Sciences*, *1*, 34–48.

Venn, J. M. A. (1880). I. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *10*(59), 1–18.

Verdu, S. (1998). Fifty years of Shannon theory. *IEEE Transactions on Information Theory*, *44*(6), 2057–2078.

Verdú, S. & Guo, D. (2006). A simple proof of the entropy-power inequality. *IEEE Transactions on Information Theory*, *52*(5), 2165–2166.

Vershik, A. M. (1964). Some characteristic properties of Gaussian stochastic processes. *Theory of Probability and its Applications (SIAM)*, *9*(2), 353–356.

Vignat, C., Hero III, A. O., & Costa, J. A. (2004). About closedness by convolution of the Tsallis maximizers. *Physica A*, *340*(1-3), 147–152.

von Mises, R. (1932). Théorie des probabilités. fondements et applications. *Annales de l'institut Henri Poincaré*, *3*(2), 137–190.

von Neumann, J. & Goldstine, H. H. (1947). Numerical inverting of matrices of high order. *Bulletin of the American Mathematical Society*, *53*(11), 1021–1100.

von Plato, J. (2005). A.N. Kolmogorov, Grundbegriffe der wahrscheinlichkeitsrechnung (1933). In *Landmark Writings in Western Mathematics 1640-1940* chapter 75, (pp. 960–969). Elsevier.

Wang, L. & Madiman, M. (2004). Beyond the entropy power inequality via rearrangements. *IEEE*

Transactions on Information Theory, *60*(9), 5116–5137.

Watson (1922). *A Treatise on the Theory of Bessel Functions*. Cambridge, UK: Cambridge University Press.

Weiss, P. (1896). Recherches sur l'aimantation de la magnétite cristallisée et de quelques alliages de fer et d'antimoine. *L'éclairage électrique*, *8*, 56–68, 105–110, 248–254.

Weiss, P. (1907). L'hypothèse du champ moléculaire et la propriété ferromagnétique. *Journal de Physique Théorique et Appliquée*, *6*(1), 661–690.

Wendl, M. C. (2016). Pseudonymous fame. *Science*, *351*(6280), 1406.

Westfall, P. H. (2014). Kurtosis as peakedness, 1905-2014 R.I.P. *The American Statistician*, *68*(3), 191–195.

Widder, D. V. (1946). *The Laplace Transform*. Princeton Mathematical Series. Princeton University Press.

Wiener, N. (1948). *Cybernetics: or Control and Communication in the Animal end the Machine* (2nd ed.). Cambridge, MA: MIT Press.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *The Annals of Mathematics*, *62*(3), 548–564.

Williamson, R. E. (1956). Multiply monotone functions and their Laplace transforms. *Duke Mathematical Journal*, *23*(2), 189–207.

Wirtinger, W. (1927). Zur formalen theorie der funktionen von mehr komplexen veränderlichen. *Mathematische Annalen*, *97*(1), 357–375.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, *20A*(1/2), 32–52.

Wong, A. K. C. & You, M. (1985). Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *7*(5), 599–609.

Yao, K. (1973). A representation theorem and its applications to spherically-invariant random processes. *IEEE Transactions on Information Theory*, *19*(5), 600–608.

Zamir, R. (1998). A proof of the Fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory*, *44*(3), 1246–1250.

Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation*, *16*(1), 159–195.

Zolotarev, V. M. (1957). Mellin-Stieltjes transforms in probability theory. *Theory of Probability & Its Applications*, *2*(4), 433–460.

Zolotarev, V. M. (1986). *One-dimensional Stable Distributions*. vol. 65. Providence: American Mathematical Society.

Zozor, S. (2012). *Bruit, Non-linéaire et Information : quelques résultats*. Habilitation à Diriger des Recherches, Institut National Polytechnique de Grenoble, Grenoble, France.

Zozor, S., Puertas-Centeno, D., & Dehesa, J. S. (2017). On generalized Stam inequalities and Fisher–Rényi complexity measures. *Entropy*, *19*(9), 493.

Zozor, S. & Vignat, C. (2010). Some results on the denoising problem in the elliptically distributed context. *IEEE Transactions on Signal Processing*, *58*(1), 134–150.

Los autores

Lamberti, Pedro Walter

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea).

Portesi, Mariela Adelina

Obtuvo el título de Licenciada en Física en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata, y el grado de Doctora en Física en la misma casa de altos estudios. Es Investigador Independiente del Consejo Nacional de Investigaciones Científicas y Técnicas, con lugar de trabajo en el Instituto de Física La Plata. Su especialidad es la teoría y geometría de la información en mecánica cuántica. Posee cargo docente de Profesor Adjunto en el Departamento de Matemática de la Facultad de Ciencias Exactas de la UNLP, desempeñándose desde 2013 como integrante del Equipo Coordinador de la asignatura Análisis Matemático II (CiBEx). cursos de grado avanzados y de posgrado en la Facultad de Ciencias Exactas de la UNLP y en la Facultad de Matemática, Astronomía, Física y Computación de la Universidad Nacional de Córdoba. También ha participado en el dictado del curso de grado “Probabilidades” como Profesor Visitante de la Université Grenoble-Alpes en Francia.

Zozor, Steeve

Nació en 1972 en Colmar, Francia. Obtuvo el título de Ingeniero, de Licenciada, el grado de Doctor y la “Habilitation à diriger de Recherches”, respectivamente en 1995, 1999 y 2012, ambos del Instituto Nacional Politécnico de Grenoble (Grenoble INP), Francia. En 2001, paso varios meses en el Laboratorio de Procesamiento de Señales de la Escuela Politénica Federal de Lausanne (EPFL), Suiza como postdoctorante. Pasó un año en el Instituto de Física de La Plata (IFLP) de la Universidad Nacional de La Plata (UNLP), Argentina (2012-2013) así que varios estancias desde 2010 como profesor visitante. En 2001 ingresó al Centro National de la Investigación Cientifica (CNRS), equivalente Francés del CONICET, como “Chargé de Recherche” (cargado de investigación) y es “Directeur de Recherches” (director de investigación) desde 2017, ambos en el Laboratorio de Imagenes, Palabras, Señales y Automática de Grenoble (GIPSA-Lab), Francia. Desde 2015 es editor asociado de la revista IEEE Signal Processing Letters. Sus temas de investigación incluyen el procesamiento no lineal de señales, el estudio del efecto de resonancia estocástica, el estudio de procesamiento de datos en contextos α-estables y/o

de distribuciones de probabilidad elípticas, la téoria de la información (medidades informacionales generalizadas clásicas y cúanticas) con aplicaciones en procesamientos de datos, mecanica cúantica o ingeniería biomédica. Es a cargo de docencia en varias escuelas de Grenoble-INP de matemática para el ingeniero, probabilidades aplicadas, procesamiento estadístico de señales, métodos bayesianos. Da regularmente un mini-curso sobre los básicos de la teoría de la información en la Facultad de Ciencias Exactas de la UNLP.