# Introducción al reconocimiento de patrones: Repaso de Algebra Lineal y Probabilidades

M. Sc. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Computación
PAttern Recongition and MAchine Learning Group (PARMA-Group)

5 de agosto de 2021

**Resumen**

Este material está basado en el repaso de algebra lineal del profesor Andrew NG, de la Universidad de Stanford (`http://cs229.stanford.edu/materials.htm`), y el repaso de conceptos básicos de probabilidad, del libro *Pattern recognition* de Cristopher Bishop[1]. Además se incluye una corta definición de los sistemas lineales, tomada del libro *Análisis de señales*, de Pablo Irarrazabal[3], y un repaso del cálculo multivariable, del libro *Cálculo en varias variables* cuyas figuras también fueron tomadas de tal libro [2].

## 1. Sistemas lineales

Gran parte del curso se basará en el estudio de sistemas o modelos lineales para realizar desde el filtrado de una señal (ya sea para eliminar aspectos indeseados, o mejorar cualidades de importancia), hasta la construcción de modelos de clasificación.

Aunque gran parte de los sistemas reales son no lineales, modelos aproximados lineales de tales sistemas facilitan su análisis. Se presentan entonces el concepto básico de linealidad, fundamental en el desarrollo del curso.

### 1.1. Linealidad

Sea $L\{\cdot\}$ un operador, $f(x)$, $f(x_1)$ y $f(x_2)$ funciones de una variable $x \in \mathbb{R}$ (que en señales unidimensionales corresponde usualmente al tiempo), con los escalares $\alpha \in \mathbb{R}$ y $\beta \in \mathbb{R}$. Se dice que el operador $L$ es lineal si cumple con las propiedades de homogeneidad (también conocida como escalamiento) y superposición, que respectivamente corresponden a:

$$L\{\alpha f(x)\} = \alpha L\{f(x)\}$$

$$L\left\{f_1\left(x\right)+f_2\left(x\right)\right\}=L\left\{f_1(x)\right\}+L\left\{f_2(x)\right\}$$

Lo cual se puede resumir en una sola ecuación como:

$$L\left\{\alpha f_1\left(x\right)+\beta f_2\left(x\right)\right\}=\alpha L\left\{f_1(x)\right\}+\beta L\left\{f_2(x)\right\}$$

**Ejemplos** Sean los siguientes sistemas $L$ cuya entrada es la función $u(t)$ y la salida es $g(t)$ con $h(t)$ cualquiera.

- $g(t)=5\,u(t)$. Con una entrada dada por $\alpha u_1(t)+\beta u_2(t)$, se tiene que:

  $L\left\{\alpha u_1(t)+\beta u_2(t)\right\}=5\left(\alpha u_1(t)+\beta u_2(t)\right)=\alpha\,5\,u_1(t)+\beta\,5\,u_2(t)=\alpha L\left\{u_1(t)\right\}+\beta L\left\{u_2(t)\right\}$

  por lo tanto el sistema es lineal.

- $g(t)=\sqrt{u(t)}$. Con una entrada dada por $\alpha u_1(t)+\beta u_2(t)$, se tiene que:

  $L\left\{\alpha u_1(t)+\beta u_2(t)\right\}=\sqrt{\left(\alpha u_1(t)+\beta u_2(t)\right)}\neq\alpha\sqrt{u_1(t)}+\beta\sqrt{u_2(t)}=\alpha L\left\{u_1(t)\right\}+\beta L\left\{u_2(t)\right\}$

  por lo tanto el sistema no es lineal.

- $g(t)=u(t)\,\cos\left(\omega t\right)$. Con una entrada dada por $\alpha u_1(t)+\beta u_2(t)$, se tiene que:

  $L\left\{\alpha u_1(t)+\beta u_2(t)\right\}=\alpha u_1(t)\cos\left(\omega t\right)+\beta u_2(t)\cos\left(\omega t\right)=\alpha L\left\{u_1(t)\right\}+\beta L\left\{u_2(t)\right\}$

  por lo tanto el sistema es lineal.

- $g(t)=\frac{1}{1+\exp(-u(t))}$. Con una entrada dada por $\alpha u_1(t)+\beta u_2(t)$, se tiene que:

  $$L\left\{\alpha u_1(t)+\beta u_2(t)\right\}=\frac{1}{1+\exp\left(-\alpha u_1(t)-\beta u_2(t)\right)}=\frac{1}{1+\exp\left(-\alpha u_1(t)\right)\exp\left(-\beta u_2(t)\right)}$$

  y dado que

  $$\alpha L\left\{u_1(t)\right\}+\beta L\left\{u_2(t)\right\}=\frac{\alpha}{1+\exp\left(-u_1(t)\right)}+\frac{\beta}{1+\exp\left(-u_2(t)\right)}$$

  $$=\frac{\alpha\left(1+\exp\left(-u_2(t)\right)\right)+\beta\left(1+\exp\left(-u_1(t)\right)\right)}{\left(1+\exp\left(-u_1(t)\right)\right)\left(1+\exp\left(-u_2(t)\right)\right)}$$

  por lo que entonces en este caso $L\left\{\alpha u_1(t)+\beta u_2(t)\right\}\neq\alpha L\left\{u_1(t)\right\}+\beta L\left\{u_2(t)\right\}$, por lo que el sistema es no lineal.

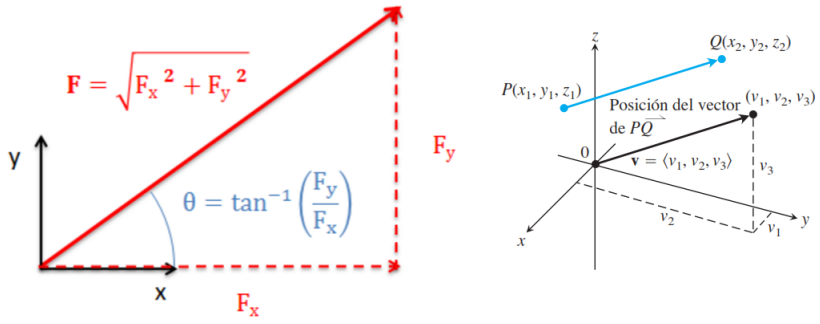



Figura 1: Vector con magnitud y dirección en $\mathbb{R}^2$ y $\mathbb{R}^3$, tomado de [2].

# 2. Algebra lineal

## 2.1. Vectores

Tal como se mencionó, un vector de dimensionalidad $n$ o con $n$ componentes se define de la siguiente manera:

$$\overrightarrow{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

donde se dice que el vector está definido en un espacio $\mathbb{R}^n$. Presenta un punto de origen $A = (a_1, a_2, \ldots, a_n)$ y un punto de destino o final $B = (b_1, b_2, \ldots, b_n)$ y viene entonces dado por:

$$\overrightarrow{v} = \overrightarrow{AB} = (b_1, b_2, \ldots, b_n) - (a_1, a_2, \ldots, a_n) = \begin{bmatrix} b_1 - a_1 \\ b_2 - a_2 \\ \vdots \\ b_n - a_n \end{bmatrix}$$

### 2.1.1. Ilustración de conceptos con vectores en $\mathbb{R}^2$

Un vector tiene una dirección y una magnitud asociados, como lo sugiere el siguiente diagrama para un vector $\overrightarrow{v} \in \mathbb{R}^2$:

El **ángulo** por ejemplo de un vector $\overrightarrow{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ en un espacio $\mathbb{R}^2$, respecto al eje $x$, está dado por:

$$\theta = \arctan\left(\frac{v_2}{v_1}\right)$$

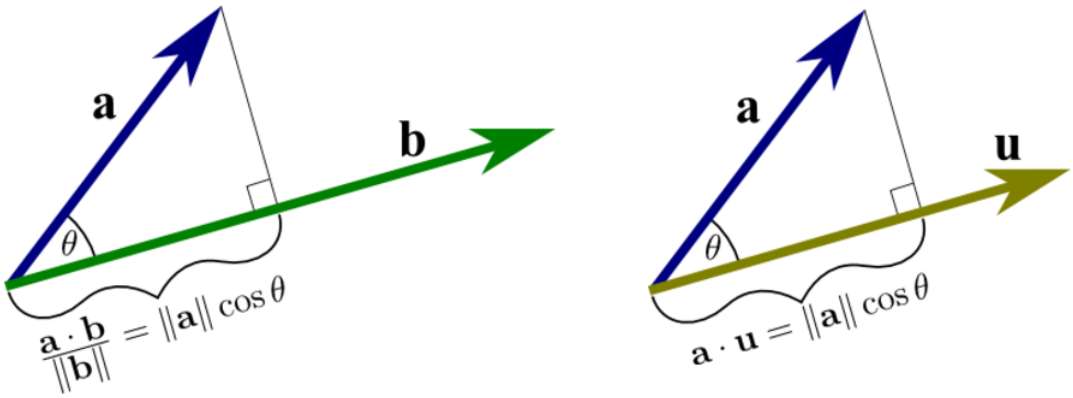



Figura 2: Magnitud de la proyección de los vectores $\overrightarrow{a} \cdot \overrightarrow{b} = \|\overrightarrow{a}\| \left\|\overrightarrow{b}\right\| \cos(\theta)$, y los vectores $\overrightarrow{a} \cdot \overrightarrow{u} = \|\overrightarrow{a}\| \cos(\theta)$, con $\|\overrightarrow{u}\| = 1$, tomado de `http://mathinsight.org/dot_product`.

La **magnitud** se define, para un vector $\overrightarrow{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ en un espacio $\mathbb{R}^2$ como :

$$\|\overrightarrow{v}\| = \sqrt{v_1^2 + v_2^2}$$

y en general para un vector en un $\mathbb{R}^n$ como:

$$\|\overrightarrow{v}\| = \sqrt{v_1^2 + \ldots v_n^2}$$

recordemos además, que es un vector unitario todo aquel vector $\hat{v}$ que cumpla con $\|\hat{v}\| = 1$.

**Producto punto o producto interno o producto escalar para un vector**: la función producto punto, para dos vectores $\overrightarrow{w}$ y $\overrightarrow{v}$ de dimensión $n$ está dada por:

$$s = \overrightarrow{v} \cdot \overrightarrow{w} = \overrightarrow{v}^T \overrightarrow{w} = v_1 w_1 + v_2 w_2 + \ldots + v_n w_n = \sum_{i=1}^{n} v_i w_i$$

donde se dice que $s$ es un escalar que pues $s \in \mathbb{R}^1$.

En el espacio euclidiano, el **producto punto** tiene la siguiente equivalencia geométrica:

$$\overrightarrow{v} \cdot \overrightarrow{w} = \|\overrightarrow{v}\| \|\overrightarrow{w}\| \cos(\theta)$$

donde el ángulo entre los vectores $\overrightarrow{v}$ y $\overrightarrow{w}$ está dado por $\theta$. El producto punto, gráficamente se refiere a la noción de la *sombra* o magnitud de la proyección del vector $\overrightarrow{v}$ en $\overrightarrow{w}$, como muestra la Figura 2.

Esto quiere decir que si los dos vectores son **co-direccionales** $\theta = 0 \Rightarrow \cos(\theta) = 1$ por lo que entonces:

$$\overrightarrow{v} \cdot \overrightarrow{w} = \|\overrightarrow{v}\| \|\overrightarrow{w}\|,$$

lo cual significa que si calculamos el producto punto del vector $\overrightarrow{v}$ consigo mismo:

$$\overrightarrow{v} \cdot \overrightarrow{v} = \|\overrightarrow{v}\|^2 ,$$

por lo que podemos llegar a la definición de la **magnitud o norma** en $\ell_2$ puede expresarse entonces en términos del producto punto como:

$$\|\overrightarrow{v}\| = \sqrt{\overrightarrow{v} \cdot \overrightarrow{v}}.$$

La **magnitud de un vector** puede interpretarse **como la proyección en el espacio $\mathbb{R}^1$ en dirección del vector unitario** $\overrightarrow{v}_u$, una operación básica que permite reducir la dimensionalidad de un vector (la reducción de la dimensionalidad es un concepto fundamental en el reconocimiento de patrones).

Por ejemplo, los vectores unitarios $\hat{i} = \begin{bmatrix}1\\0\\0\end{bmatrix}, \hat{j} = \begin{bmatrix}0\\1\\0\end{bmatrix}$ y $\hat{k} = \begin{bmatrix}0\\0\\1\end{bmatrix}$ son vectores unitarios.

Si el ángulo entre los dos vectores u **ortogonales** $\overrightarrow{v}$ y $\overrightarrow{w}$ es de $90^o$, se tiene que entonces por la definición geométrica del producto punto:

$$\overrightarrow{v} \cdot \overrightarrow{w} = \overrightarrow{v}^T \overrightarrow{w} = 0$$

Los vectores pueden dibujarse en MATLAB como sigue (en $\mathbb{R}^2$):

```
M = [-0.4  0.7  0.2  ;  -0.5  0.1  0.5];
plotv(M, '-')
```

**Operaciones básicas en vectores**: En general, la suma y resta de dos vectores $\vec{r} = \vec{a} \pm \vec{b}$ con $\vec{a}, \vec{b} \in \mathbb{R}^n$ y se define como sigue:

$$\vec{r} = \begin{bmatrix} a_1 \pm b_1 \\ \vdots \\ a_n \pm b_n \end{bmatrix}.$$

La Figura 3 muestra la graficación de dos vectores $\overrightarrow{a} \in \mathbb{R}^2$ y $\overrightarrow{b} \in \mathbb{R}^2$ sus sumas y restas respectivas.

Se pueden también definir los operadores de multiplicación y división *por componente* de vectores, denotados respectivamente como $.\cdot$ y $./$ por lo que entonces el producto por componente de dos vectores $\vec{a}, \vec{b} \in \mathbb{R}^n$ se define como:

$$\vec{a}. \cdot \vec{b} = \begin{bmatrix} a_1 \cdot b_1 \\ \vdots \\ a_n \cdot b_n \end{bmatrix}$$

y de manera similar, la división por componente está dada por:

$$\vec{a}./\vec{b} = \begin{bmatrix} a_1/b_1 \\ \vdots \\ a_n/b_n \end{bmatrix}.$$

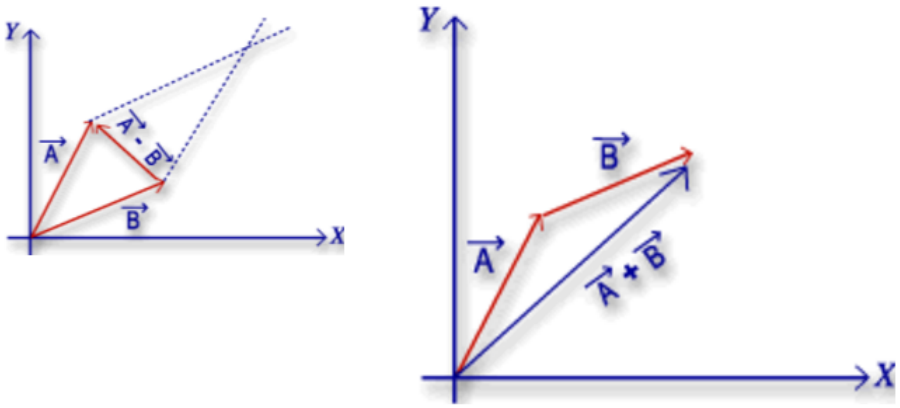



Figura 3: Suma y resta de vectores en $\mathbb{R}^2$.

Propiedades de las operaciones con los vectores, con $\vec{u}, \vec{v}, \vec{w} \in \mathbb{R}^n$ y $a, b \in \mathbb{R}$:

- $\vec{u} + \vec{v} = \vec{v} + \vec{u}$
- $\vec{u} + \vec{0} = \vec{u}$
- $\vec{0} \cdot \vec{u} = 0$
- $a\,(b\,\vec{u}) = (a\,b)\,\vec{u}$
- $(a + b)\,\vec{u} = a\,\vec{u} + b\,\vec{u}$
- $(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$
- $\vec{u} + (-\vec{u}) = 0$
- $1\vec{u} = \vec{u}$
- $a\,(\vec{u} + \vec{v}) = a\,\vec{u} + a\,\vec{v}$

### 2.1.2. Normas:

El concepto de magnitud o norma visto anteriormente, se conoce como la distancia Eucliciana o norma $\ell_2$, la cual se refiere al largo de un vector, como vimos. La norma euclidiana se puede reescribir como:

$$\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{1}$$

y ya se demostró la equivalencia $\|\vec{x}\|_2^2 = \vec{x}^T\vec{x}$. Formalmente, la norma es cualquier función $f : \mathbb{R}^n \to \mathbb{R}$ que satisface las siguientes 4 propiedades, para todo $\vec{x} \in \mathbb{R}^n$, $\vec{y} \in \mathbb{R}^n$ y $t \in \mathbb{R}$:

- **No negatividad:** , $f(\vec{x}) \geq 0$.
- **Nulidad:** $f(\vec{x}) = 0$ si y solo si $\vec{x} = 0$ (vector nulo).
- **Homogeneidad absoluta**: $f(t\,\vec{x}) = |t|\, f(\vec{x})$.
- **Desigualdad triangular**: $f(\vec{x} + \vec{y}) \leq f(\vec{x}) + f(\vec{y})$ . Ver la Figura 3.

Generalizando la ecuación 1 como una norma $\ell_p$, con $p \geq 1$, se tiene que:

$$\|\vec{x}\|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{2}$$

A partir de tal definición general, se tiene la norma $\ell_1$ también conocida como **Manhattan o distancia de bloques**:

$$\|\vec{x}\|_1 = \left( \sum_{i=1}^{n} |x_i| \right) \tag{3}$$

La norma $\ell_\infty$ se define entonces como:

$$\|\vec{x}\|_\infty = \left( \sum_{i=1}^{n} |x_i|^\infty \right)^{1/\infty} . \tag{4}$$

Esta definición parece un tanto confusa. Sin embargo, se puede notar que la máxima entrada o componente denotado por $x_m$ del arreglo $\vec{x}$ viene a hacer que, al elevarse al infinito sea, por mucho, el mayor componente del vector:

$$x_m^\infty \gg x_i^\infty,\ \forall i \neq m$$

por lo que entonces se puede decir que la sumatoria de los componentes del vector $\vec{x}$ tiende al valor $x_m^\infty$ (en términos de aproximación numérica), con ello se tiene que:

$$\sum_{i=1}^{n} |x_i|^\infty \to x_m^\infty. \tag{5}$$

por ello se puede reescribir la ecuación de la norma como:

$$\|\vec{x}\|_\infty = (x_m^\infty)^{1/\infty} = |x_m| = \text{máx}\left(|x_i|\right). \tag{6}$$

La norma de tipo $\ell_p$ del vector diferencia $\vec{d} = \vec{v} - \vec{w}$ entre dos vectores, $\vec{v}$ y $\vec{w}$ se conoce como la distancia $\ell_p$, por ejemplo se definen la distancia Euclidiana y la distancia Manhattan.

- Existen también distintas normas definidas para matrices, como por ejemplo, la **norma de Frobenius,** la cual se define como sigue, para una matriz $A \in \mathbb{R}^{m \times n}$ :

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j}^2} = \sqrt{\textbf{tr}\left(A^T A\right)}$$

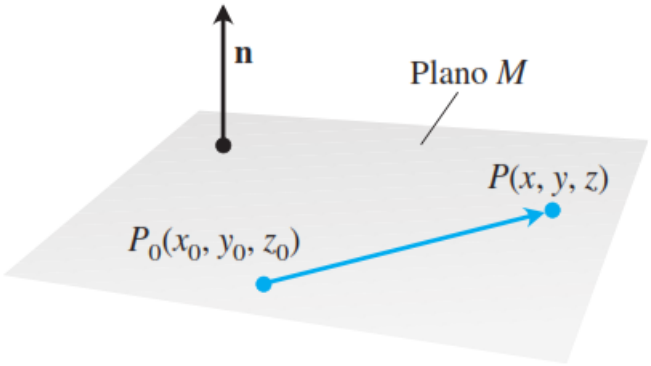



Figura 4: Plano en $\mathbb{R}^3$. Tomado de [2]

## 2.2. Funciones y cálculo multivariable

Antes de pasar a las funciones de $n$ variables o con dominio definido en $\mathbb{R}^n$ examinaremos con más detalle los *planos* o en general *hiperplanos*, correspondientes a las líneas o pendientes definidas como $y = f(x) = m\,x + b$ que conocemos para funciones en una variable.

## 2.3. Rectas

Una recta $L$ en un espacio $\mathbb{R}^n$ que pasa por un punto $P_0 = (x_1, \ldots, x_n)$ y paralela al vector $\vec{v} \in \mathbb{R}^n$, con lo que $L$ está compuesta por todo punto $P$ que haga que el vector $\overrightarrow{P_0P}$ sea paralelo al vector $\vec{v}$, lo que implica que se tiene que cumplir lo siguiente:

$$\overrightarrow{P_0P} = t\,\vec{v} \Rightarrow P - P_0 = t\,\vec{v},$$

para algún escalar $t \in \mathbb{R}$, con lo que el valor de $t$ depende de la posición del punto $P$ en el espacio. Una recta se extiende de forma infinita, por lo que entonces se cumple que $-\infty < t < \infty$. Despejando la ecuación anterior se obtiene la **ecuación paramétrica** de una recta:

$$P = r(t) = P_0 + t\,\vec{v}, \tag{7}$$

se le llama ecuación paramétrica pues el parámetro de tal ecuación es el escalar $t$.

## 2.4. Planos e hiperplanos

Un plano (llamado así en un espacio $\mathbb{R}^3$) o hiperplano (para cualquier espacio $\mathbb{R}^n$) corresponde a una **superficie** completamente planar que se extiende hacia el infinito en todas las dimensiones.

Observe la Figura 4 donde se grafica un plano en un espacio $\mathbb{R}^3$ con un vector normal $\vec{n} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ y sobre el cual existen un punto cualquiera (desconocido) $P = (x, y, z)$ y un punto conocido $P_0 = (x_0, y_0, z_0)$. Ya concluimos que cuando dos vectores son perpendiculares, su producto punto es cero, por lo que entonces podemos escribir la **ecuación vectorial** de un hiper-plano como:

$$\vec{n} \cdot \overrightarrow{P_0 P} = 0,$$

si establecemos los vectores con origen en $(0, 0, 0)$ hacia los puntos $P_0$ y $P$, respectivamente, como $\vec{P_0}$ y $\vec{P}$ y desarrollando la ecuación vectorial se tiene que:

$$\Rightarrow \vec{n} \cdot (P - P_0) = 0 \Rightarrow \vec{n} \cdot \vec{P} = \vec{n} \cdot \vec{P_0}$$
$$\Rightarrow ax + by + cz = ax_0 + by_0 + bz_0,$$

donde dado que conocemos el vector normal y el punto $P_0$, podemos hacer $d = ax_0 + by_0 + bz_0$, con lo que entonces obtenemos la **ecuación cartesiana de un plano**, compuesto por todo punto $P = (x, y, z)$ que haga cumplir:

$$a\,x + b\,y + c\,z = d, \tag{8}$$

lo cual se puede reescribir como la ecuación de una pendiente en un espacio de dimensionalidad mayor:

$$y = \vec{m}^T \vec{x} + k, \tag{9}$$

con $\vec{m}^T = \begin{bmatrix} a & b & c \end{bmatrix}$ y $\vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$. Observe que a diferencia de la ecuación de la recta, que extiende el vector en una sola dirección, la ecuación de un plano está dada por todos los puntos satisfacen la ecuación del plano.

## 2.5. Funciones multivariable

Un plano en $\mathbb{R}^3$ puede conceptualizarse como una función $z = f(x, y)$, con dominio en $\mathbb{R}^2$ y codominio en $\mathbb{R}$, por lo que basados en la ecuación 8, la función vendría dada por:

$$z = \frac{d}{c} - \frac{a}{c}\,x - \frac{b}{c}\,y,$$

donde en general la ecuación de una función plano está entonces dada por:

$$z = f(x, y) = a_1 x + a_2 y + a_3$$

El siguiente código dibuja dos funciones o planos $f(x, y) = 2{,}1\,x + y$ y $g(x, y) = 0$ en MATLAB, mostrados en la Figura 5.

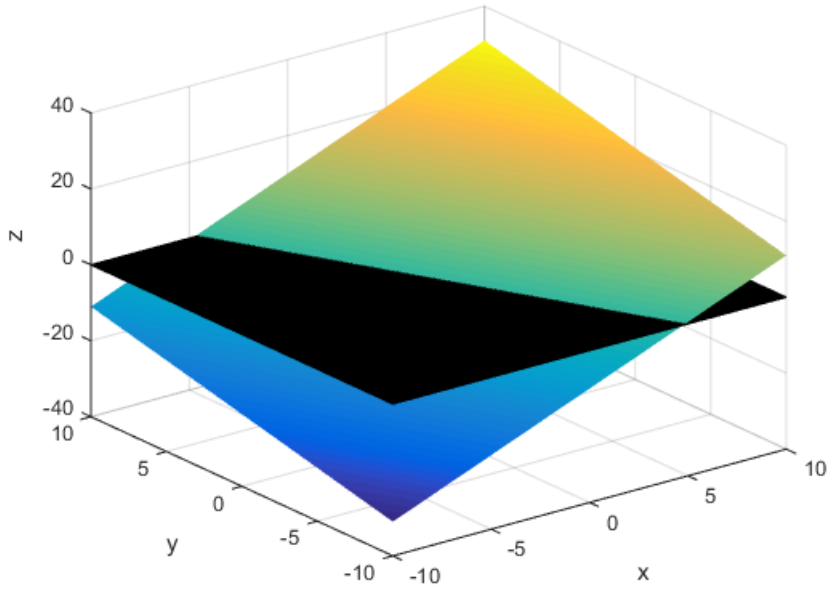



Figura 5: Graficación de dos planos.

```
x=-10:.1:10;
[X,Y] = meshgrid(x);
Z = 2.1*X + Y;
Z1 = Z.*0;
figure; surf(X,Y,Z);
shading flat
xlabel('x');
ylabel('y');
zlabel('z')
hold on;
surf(X,Y,Z1);
hold on;
```

En general, una función con múltiples variables de entrada y una de salida, correspondiente a un dominio $\mathbb{R}^n$ y un codominio en $\mathbb{R}$ generan lo que se llama superficies en un espacio $\mathbb{R}^{n+1}$.

Las siguientes son algunas superficies conocidas (observe que para expresar tales superficies en términos de una función $z = f(x, y)$ con dominio en $\mathbb{R}^2$ y codominio en $\mathbb{R}$, es necesario despejar $z$) y se ilustran en la Figura 6:

- Paraboloide hiperbólica $\frac{y^2}{b^2} - \frac{x^2}{a^2} = \frac{z}{c},\ c > 0$
- Paraboloide elíptica $\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z}{c},\ c > 0$

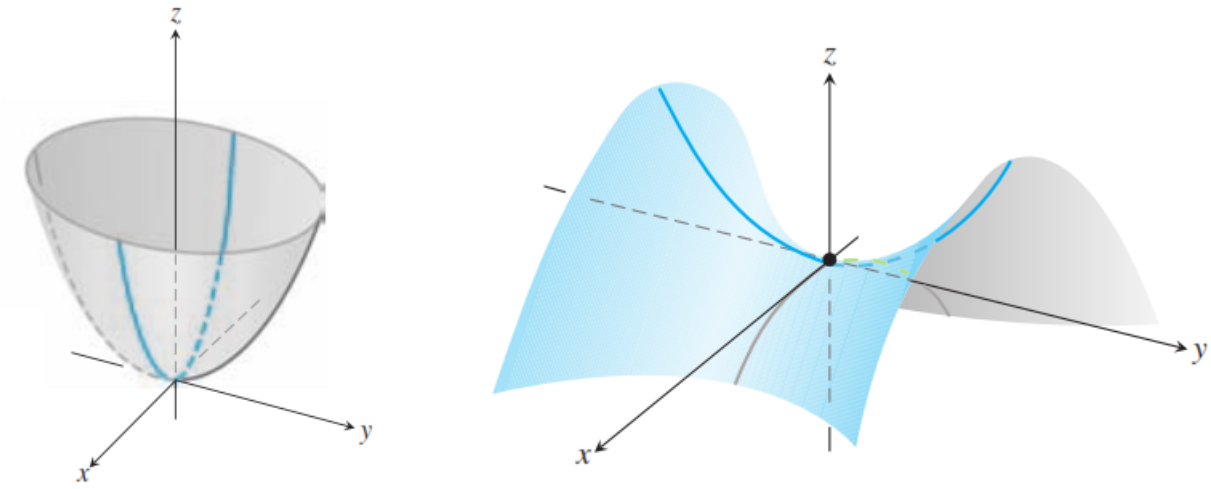



Figura 6: Superficies cuadricas de ejemplo, paraboloide .[2]

## 2.6. Curvas de nivel

Una curva de nivel en $z_0$ es un corte o intersección con un plano con un valor de $z$ constante $z_0$, es decir, un plano con todos sus puntos con dominio o preimágenes en $x, y$ con imagen $z_0$. La Figura 7 muestra el ejemplo de una función multivariable $g(x, y)$ que forma una parabolide hiperbólica, interseca-do con un plano en $z_0 = 40$. Observe que la intersección o **curva de nivel** en este caso corresponde a una parábola. Además la Figura 7 muestra otro ejemplo de curva de nivel en una superficie cuádrica, e ilustra el concepto con las curvas de nivel que se pueden encontrar en los mapas geográficos, para indicar la forma de montañas en la dimensión $z$.

Finalmente, observe la Figura 5, donde se intersecan dos planos correspondiente a una superficie del funcional $f(x, y) = 2{,}1x + y$ y un plano constante $g(x, y) = 0$. Es fácil notar que la intersección entre ambos planos es una línea recta cuya ecuación se puede calcular de la siguiente manera. Para el plano $f$ tómese un valor en $z = 0$ (donde ambos planos coinciden) y despeje función $f$ en $z = 0$ se tiene que $-y = 2{,}1x$. Si nos damos dos valores de $x$, $x = 0$ y $x = 1$, se obtienen los puntos $P_0, P \in \mathbb{R}^2$ para los planos $x$ y $y$:

$$\begin{array}{c} P_0 = (0, 0) \\ P = (1, -2{,}1) \end{array} ,$$

por lo que entonces un vector paralelo a la línea de intersección entre los dos planos viene dado por $\overrightarrow{P_0P} = \begin{bmatrix} 1 \\ -2{,}1 \end{bmatrix}$ con la ecuación vectorial de la línea recta dada por $r(t) = (0, 0) + t \begin{bmatrix} 1 \\ -2{,}1 \end{bmatrix}$.

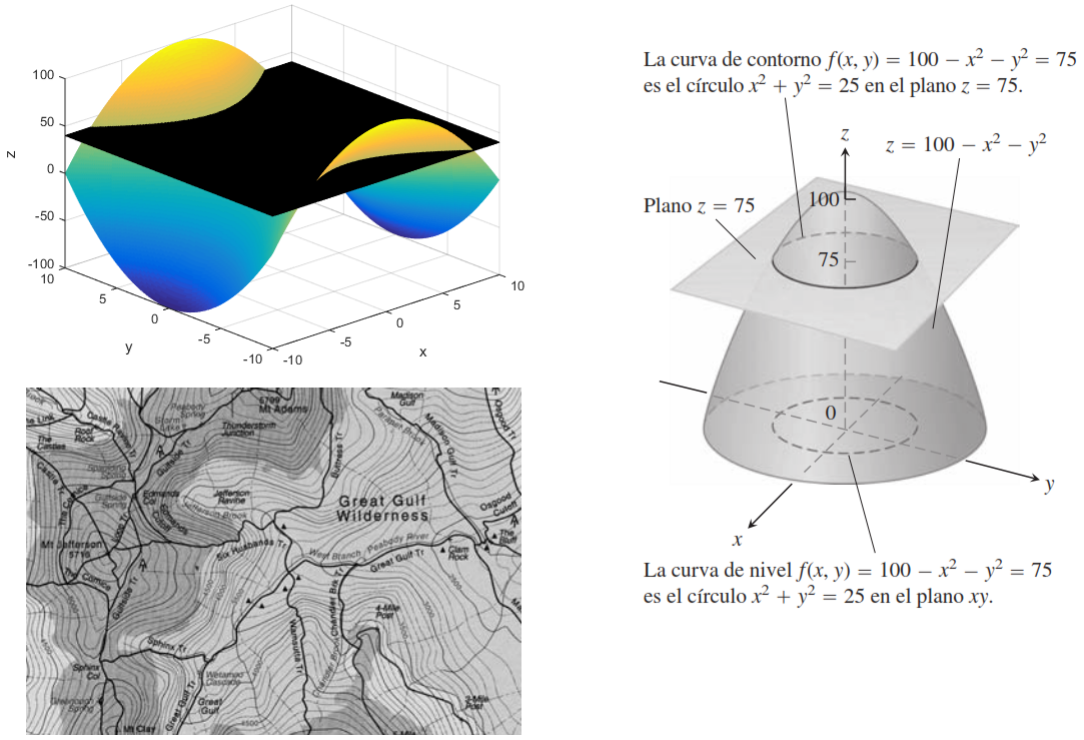



Figura 7: Ejemplos de curvas de nivel. Tomado de [2]

## 2.7. El vector gradiente

A continuación se define la función **derivada parcial** de una función de dos variables $z = f(x, y)$ con dominio en $\mathbb{R}^2$ y codominio en $\mathbb{R}$ respecto a $x$ como:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h},$$

donde se observa que el desplazamiento por $h$ se hace únicamente en el eje $x$, dejando el otro eje intacto. Conceptualmente la derivada parcial respecto a una variable $x$ corresponde al cambio en el funcional en esa dimensión. La evaluación de tal funcional en un punto $(x_0, y_0)$ viene entonces dada por:

$$\frac{\mathrm{d}f}{\mathrm{d}x}_{(x_0,y_0)} = \lim_{h \to 0} \frac{f(x_0+h, y_0) - f(x_0, y_0)}{h}$$

Por ejemplo, para la función $f(x, y) = \frac{y^2}{b^2} - \frac{x^2}{a^2}$ se tiene que $\frac{\mathrm{d}f}{\mathrm{d}x} = -\frac{2}{a^2}x$ y respecto a $y$ como $\frac{\mathrm{d}f}{\mathrm{d}y} = \frac{2}{b^2}y$. La evaluación de ambas derivadas parciales en el punto $(1, 1)$ vendrían a ser $\frac{\mathrm{d}f}{\mathrm{d}x}_{(1,1)} = -\frac{2}{a^2}$ y $\frac{\mathrm{d}f}{\mathrm{d}y}_{(1,1)} = \frac{2}{b^2}$, respectivamente.

En general, para una función con dominio en $\mathbb{R}^n$ y codominio en $\mathbb{R}$, $z = f(x_1, \ldots, x_n)$, la derivada parcial respecto a la variable $x_i$ está dada por:

$$\frac{\mathrm{d}f}{\mathrm{d}x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}.$$

Veamos ahora la definición formal del **vector gradiente**, primero para una función de con dominio en $\mathbb{R}^2$ y codominio en $\mathbb{R}$, $f(x, y)$ evaluado en cualquier punto $(x_0, y_0)$:

$$\nabla f_{(x_0,y_0)} = \frac{\mathrm{d}f}{\mathrm{d}x}_{(x_0,y_0)} \hat{i} + \frac{\mathrm{d}f}{\mathrm{d}y}_{(x_0,y_0)} \hat{j}$$

y en general para una función con dominio en $\mathbb{R}^n$ el vector gradiente en cualquier punto $(u_1, \ldots u_n)$ viene dado por:

$$\nabla f_{(u_1,\ldots u_n)} = \frac{\mathrm{d}f}{\mathrm{d}x_1}_{(u_1,\ldots u_n)} \hat{i_1} + \ldots + \frac{\mathrm{d}f}{\mathrm{d}x_n}_{(u_1,\ldots u_n)} \hat{i_n}.$$

El vector gradiente denota entonces la dirección para la cual una superficie definida por la función $f$ cambia. Observe que ese vector cambia de acuerdo al punto $(u_1, \ldots u_n)$ en el que se evalúe.

Siguiendo el ejemplo de del plano $f(x, y) = 2{,}1x + y$ , calculando su vector gradiente se obtiene que $\nabla f = 2{,}1\,\hat{i} + \hat{j}$. Observe primero que el mismo es constante, sin importar el punto $(x_0, y_0)$ sobre el que se evalúa. Recordemos además que el vector paralelo a la recta que constituye la curva de nivel en $z = 0$ está dado por $\overrightarrow{P_0P} = \begin{bmatrix} 1 \\ -2{,}1 \end{bmatrix}$ . Cual es la relación entre estos vectores? Intiuitivamente, la curva de nivel es un corte que en este caso está dado por una

recta en $\mathbb{R}^2$, al igual que el vector gradiente, el cual indica la dirección hacia la que el plano *crece* o se *orienta,* por lo que entonces es natural pensar que **ambos vectores son ortogonales entre ellos**:

$$\nabla f \cdot \overrightarrow{P_0 P} = \begin{bmatrix} 1 \\ -2{,}1 \end{bmatrix} \begin{bmatrix} 2{,}1 & 1 \end{bmatrix} = 0.$$

Para graficar la superficie y ambos vectores hacemos en MATLAB:

```
x=-10:.1:10;
[X,Y] = meshgrid(x);
Z = 2.1*X + Y;
Z1 = Z.*0;
figure; surf(X,Y,Z);
shading flat
xlabel('x');
ylabel('y');
zlabel('z')
hold on;
surf(X,Y,Z1);
hold on;
M = [1   2.1  ;
 -2.1   1];
plotv(M,'-');
```

La gráfica del vector gradiente (en rojo) y el vector paralelo a la curva de nivel se muestra en la Figura 8.

Ejemplo 2

Tómese la siguiente función multi-variable $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = 3^{2x} + 5^{4y} + 2x + y^4$. El vector gradiente para una función de dos variables está dado en general por:

$$\nabla f = \frac{df}{dx}\hat{i} + \frac{df}{dy}\hat{j}$$

y en este caso cada derivada parcial está dada por (recordando que para una función $f(x) = a^x \Rightarrow f'(x) = a^x x' \ln(a)$):

$$\frac{df}{dx} = 2\,3^{2x}\ln(3) + 2$$

$$\frac{df}{dy} = 4\,5^{4y}\ln(5) + 4y^3$$

$$\Rightarrow \nabla f = \left(2\,3^{2x}\ln(3) + 2\right)\hat{i} + \left(4\,5^{4y}\ln(5) + 4y^3\right)\frac{df}{dy}\hat{j}$$

Como último ejemplo, en la Figura 9 se muestra la superficie correspondiente a la función Gaussiana multivariable $f(x,y) = e^{\left(-x^2-y^2\right)}$, y se grafican los vectores gradientes en varios puntos usando el siguiente código:

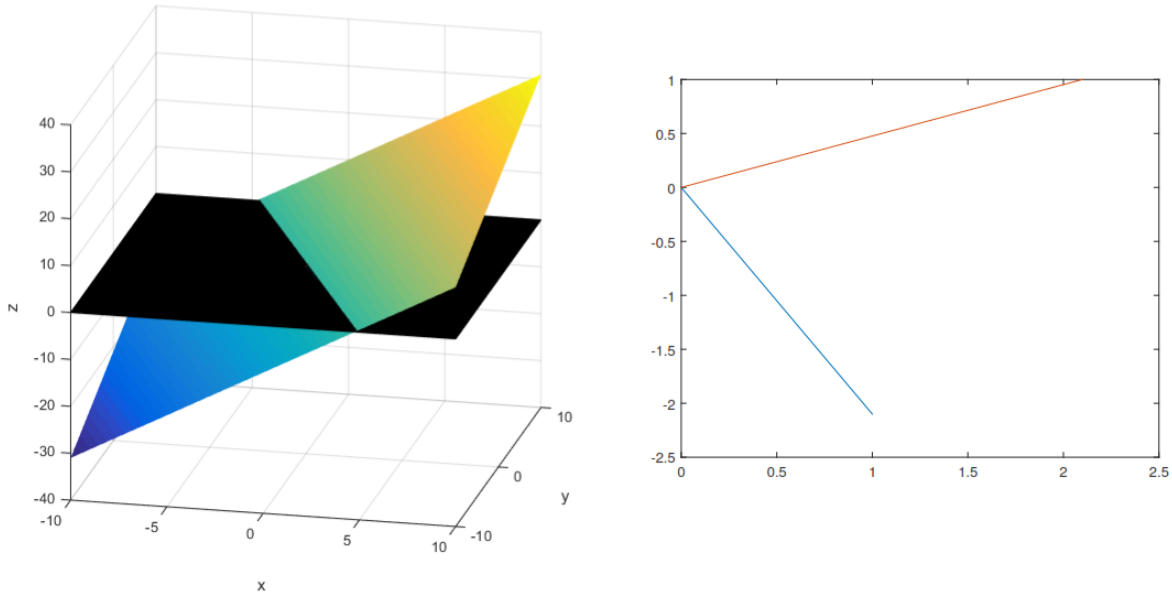



Figura 8: Plano, vectores gradiente (rojo) y curva de nivel (azul).

```
[X,Y] = meshgrid(-2:.2:2);
Z = exp(-X.^2 - Y.^2);
[DX,DY] = gradient(Z);
figure
contour(X,Y,Z)
hold on
quiver(X,Y,DX,DY)
hold off
```

Observe que los vectores gradiente varían en cada punto.

## 2.8. Matrices

La álgebra lineal facilita la expresión de múltiples operaciones, como por ejemplo las operaciones en ecuaciones lineales, como el siguiente sistema de ecuaciones:

$$4x_1 - 5x_2 = -13$$
$$-2x_1 + 3x_2 = 9$$

el sistema de ecuaciones anterior tiene igual número de ecuaciones y variables, por lo que presenta una solución única si las ecuaciones son linealmente independientes (ninguna de las ecuaciones es combinación lineal de otra). En notación matricial, el sistema de ecuaciones anterior se expresa de la siguiente

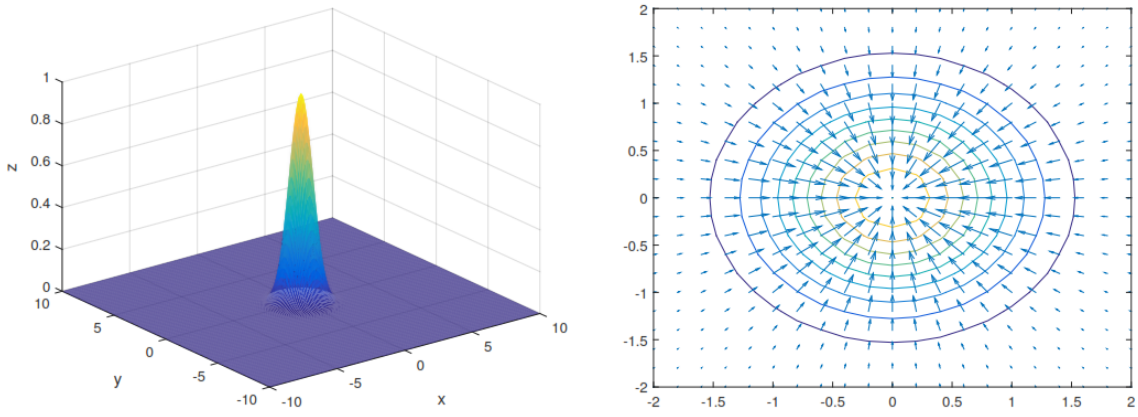



Figura 9: Graficación de la función gaussiana y los vectores gradiente en varios puntos.

forma:

$$A\,\vec{x} = b$$

con

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \qquad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

En el material del curso se utilizará la siguiente notación:

- Con $A \in \mathbb{R}^{m \times n}$ se define una matriz con $m$ filas y $n$ columnas, donde en este caso todas las entradas de $A$ son números reales.
- Con $\vec{x} \in \mathbb{R}^{n \times 1} = \mathbb{R}^n$ se denota un vector con $n$ entradas. Por convención, un vector $n$ dimensional se define como una matriz de $n$ filas y $1$ columna, conocido como el **vector columna**:

$$\overrightarrow{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

  y el elemento $i$ del vector se denota como $x_i$. Un vector fila se define entonces de la siguiente forma (usando la definición de la transpuesta):

$$\overrightarrow{x}^T = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

- Para denotar los elementos de una matriz se usa la notación $a_{i,j}$ o ($A_{ij}$, $A_{i,j}$,$A\,(i,j)$, etc), y para definir una entrada de la matriz $A$ en la fila $i$ y la

columna $j$:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & & a_{m,n} \end{bmatrix}$$

y con la columna $j$ de la matriz $A$ definida como $a_j$ o $A_{:,j}$, de modo que la matriz $A$ está definida en términos de vectores columna por:

$$A = \begin{bmatrix} | & | & \dots & | \\ \overrightarrow{a}_{:,1} & \overrightarrow{a}_{:,2} & \dots & \overrightarrow{a}_{:,n} \\ | & | & \dots & | \end{bmatrix}$$

y se define la fila $i$ de tal matriz como $\vec{a}_{i,:}^T$ o $A_{i,:}$, por lo que en términos de tales vectores fila la matriz $A$ se expresa como:

$$A = \begin{bmatrix} - & \vec{a}_{1,:}^T & - \\ - & \vec{a}_{2,:}^T & - \\ & \vdots & \\ - & \vec{a}_{m,:}^T & - \end{bmatrix}$$

## 2.9. La matriz identidad y diagonal

La matriz identidad, definida como una matriz cuadrada $I \in \mathbb{R}^{n \times n}$ y está formada por una diagonal de unos, y el resto de entradas de la matriz está en cero:

$$I_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

y es el neutro de la multiplicación matricial, por lo que para toda $A \in \mathbb{R}^{m \times n}$ se tiene que:

$$A\, I = A$$

la matriz identidad es un caso particular de una matriz diagonal, donde todos los elementos no diagonales son 0, lo que se denota como: $D = \text{diag}\,(d_1, d_2, \dots, d_n)$ con:

$$D_{i,j} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases}$$

por lo que entonces $I = \text{diag}\,(1, 1, \dots, 1)$.

## 2.10. La matriz transpuesta

La transpuesta de una matriz es el resultado de cambiar las filas a columnas. Sea una matriz $A \in \mathbb{R}^{m \times n}$, su transpuesta se escribe como $A^T \in \mathbb{R}^{n \times m}$ y sus entradas están dadas por:

$$\left(A^T\right)_{i,j} = A_{j,i}.$$

Las siguientes son propiedades de la transpuesta:

- $\left(A^T\right)^T = A$
- $(A\,B)^T = B^T A^T$
- $(A+B)^T = A^T + B^T$.

## 2.11. Matrices simétricas

Una matriz cuadrada $A \in \mathbb{R}^{n\times n}$ es simétrica si $A = A^T$ y es anti simétrica si $A = -A^T$, Para toda matriz $A \in \mathbb{R}^{n\times n}$ es fácil demostrar que la matriz $A + A^T$ es simétrica y la matriz $A - A^T$ es anti-simétrica, por lo que se puede seguir que cualquier matriz cuadrada puede expresarse en términos de una matriz simétrica y anti-simétrica:

$$A = \frac{1}{2}\left(A + A^T\right) - \frac{1}{2}\left(A - A^T\right).$$

Se define entonces el conjunto de matrices simétricas de dimensiones $n \times n$ como $\mathbb{S}^n$ por lo que $A \in \mathbb{S}^n$ si es simétrica. Las matrices simétricas son muy frecuentes en el reconocimiento de patrones, y presentan una serie de propiedades muy útiles que veremos más adelante.

## 2.12. La traza de una matriz

La traza de una matriz cuadrada $A \in \mathbb{R}^{n\times n}$ denotada como $\text{tr}\,(A)$ es la suma de los elementos en la diagonal de una matriz:

$$\text{tr}\,(A) = \sum_{i=1}^{n} A_{i,i}$$

La traza tiene las siguientes propiedades:

- $\text{tr}\,(A) = \text{tr}\left(A^T\right)$
- Superposición $\text{tr}\,(A+B) = \text{tr}\,(A) + \text{tr}\,(B)$
- Homogeneidad: Sea $t \in \mathbb{R}$, $\text{tr}\,(t\,A) = t\,\text{tr}\,(A)$
- Para $A$ y $B$ cuadradas, se tiene que $\text{tr}\,(A\,B) = \text{tr}\,(B\,A)$

## 2.13. Producto de matrices

El producto de dos matrices $A \in \mathbb{R}^{m\times n}$ y $B \in \mathbb{R}^{n\times p}$ es la matriz:

$$C = A \circ B = A\,B \in \mathbb{R}^{m\times p}$$

donde

$$C_{i,j} = A_{i,1}B_{1,j} + \ldots + A_{i,n}B_{n,j} = \sum_{k=1}^{n} A_{i,k}\,B_{k,j}$$

observe que para efectuar el producto matricial la cantidad de columnas en $A$ debe ser igual a la cantidad de filas de la matriz $B$. A continuación se examinan los casos particulares del producto de matrices

## 2.14. Producto vector-vector, producto interno (punto) y producto externo

Sean dos vectores $\overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^n$ el **producto interno** o producto punto se puede definir, en términos del producto entre tales vectores de la siguiente forma:

$$\overrightarrow{x} \cdot \overrightarrow{y} = \overrightarrow{x}^T \overrightarrow{y} \in \mathbb{R}^1 = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \sum_{i=1}^{n} x_i \, y_i$$

Observe entonces que el producto interno es un caso especial de la multiplicación de matrices, y que además, siempre se cumple que

$$\overrightarrow{x}^T \overrightarrow{y} = \overrightarrow{y}^T \overrightarrow{x}.$$

El **producto externo** en cambio, está dado para dos vectores $\overrightarrow{x} \in \mathbb{R}^{m \times 1}$, $\overrightarrow{y} \in \mathbb{R}^{1 \times n}$ (no necesariamente de la misma dimensionalidad) se define como:

$$\overrightarrow{x} \odot \overrightarrow{y} = \vec{x}\vec{y}^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

El producto externo permite, por ejemplo, crear una matriz $A \in \mathbb{R}^{m \times n}$ cuyas columnas sean igual a un vector $x \in \mathbb{R}^m$ usando un vector unitario $\overrightarrow{1} \in \mathbb{R}^n$, como sigue:

$$\overrightarrow{x} \, \overrightarrow{1}^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1_n \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \vec{x} & \vec{x} & \cdots & \vec{x} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \vec{x} & \vec{x} & \cdots & \vec{x} \\ | & | & & | \end{bmatrix}$$

Otra propiedad interesante del producto externo es que el **producto externo de un vector consigo mismo da como resultado una matriz simétrica**: $\vec{x}\,\vec{x}^T = A = A^T$. Se deja al lector como ejercicio.

## 2.15. Producto matriz-vector

Sea una matriz $A \in \mathbb{R}^{m \times n}$ y un vector (columna) $\overrightarrow{x} \in \mathbb{R}^{n \times 1}$ su producto es el vector $\overrightarrow{y} \in \mathbb{R}^{m \times 1}$ .

Si se escribe a la matriz $A$ por columnas, entonces se puede expresar a $A\overrightarrow{x}$ como:

$$\vec{y} = A\overrightarrow{x} = \begin{bmatrix} - & \vec{a}_{1,:}^T & - \\ - & \vec{a}_{2,:}^T & - \\ & \vdots & \\ - & \vec{a}_{m,:}^T & - \end{bmatrix} \overrightarrow{x} = \begin{bmatrix} - & \vec{a}_{1,:}^T & - \\ - & \vec{a}_{2,:}^T & - \\ & \vdots & \\ - & \vec{a}_{m,:}^T & - \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \vec{a}_{1,:}^T\,\vec{x} \\ \vec{a}_{2,:}^T\,\vec{x} \\ \vdots \\ \vec{a}_{m,:}^T\,\vec{x} \end{bmatrix}$$

En otras palabras, la fila $i$ de $y$ , $y_i$ es igual al producto interno de la fila $b_i$ con el vector $\overrightarrow{x}$.

Alternativamente, si se escribe la matriz $A$ en forma de columnas, el producto matriz-vector se puede expresar como:

$$\overrightarrow{y} = A\overrightarrow{x} = \begin{bmatrix} | & | & \cdots & | \\ \vec{a}_{:,1} & \vec{a}_{:,2} & \ldots & \vec{a}_{:,n} \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [\vec{a}_{:,1}]\,x_1 + [\vec{a}_{:,2}]\,x_2 + \ldots + [\vec{a}_{:,n}]\,x_n.$$

ello es fácilmente corroborable si hacemos la multiplicación de sus transpuestas:

$$\overrightarrow{y}^T = \vec{x}^T\,A^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} - & \vec{a}_{:,1}^T & - \\ - & \vec{a}_{:,2}^T & - \\ & \vdots & \\ - & \vec{a}_{:,n}^T & - \end{bmatrix} = x_1\left[\vec{a}_{:,1}^T\right] + x_2\left[\vec{a}_{:,1}^T\right] + \ldots + x_n\left[\vec{a}_{:,n}^T\right].$$

Lo anterior representa el hecho de que el vector $\overrightarrow{y}$ es una **combinación lineal** de las columnas de la matriz $A$, donde los coeficientes están definidos en el vector $\overrightarrow{x}$.

## 2.16. Producto matriz-matriz

El producto matriz-matriz en general de dos matrices $A \in \mathbb{R}^{m\times n}$ y $B \in \mathbb{R}^{n\times p}$ dado por $C \in \mathbb{R}^{m\times p}$ se puede definir en términos de las filas y columnas, **donde para cada entrada $C_{i,j}$ el producto interno de la fila $i$** de $A$ y la **columna $j$ de $B$**, simbólicamente esto se expresa como sigue:

$$C = A\,B = \begin{bmatrix} - & \vec{a}_{1,:}^T & - \\ - & \vec{a}_{2,:}^T & - \\ & \vdots & \\ - & \vec{a}_{m,:}^T & - \end{bmatrix} \begin{bmatrix} | & | & \cdots & | \\ \vec{b}_{1,:} & \vec{b}_{2,:} & \ldots & \vec{b}_{p,:} \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} \vec{a}_{1,:}^T\vec{b}_{1,:} & \vec{a}_{1,:}^T\vec{b}_{2,:} & \cdots & \vec{a}_{1,:}^T\vec{b}_{p,:} \\ \vec{a}_{2,:}^T\vec{b}_{1,:} & \vec{a}_{2,:}^T\vec{b}_{2,:} & \cdots & \vec{a}_{2,:}^T\vec{b}_{p,:} \\ \vdots & \vdots & \ddots & \vdots \\ \vec{a}_{m,:}^T\vec{b}_{1,:} & \vec{a'}_m^T\vec{b}_{2,:} & \cdots & \vec{a}_{m,:}^T\vec{b}_{p,:} \end{bmatrix}$$

$$C = A\,B = \begin{bmatrix} | & | & \cdots & | \\ A\vec{b}_{1,:} & A\vec{b}_{2,:} & \ldots & A\vec{b}_{p,:} \\ | & | & \cdots & | \end{bmatrix}$$

La última igualdad representa el hecho de que la columna $j$ de la matriz $C$ es una combinación lineal de los vectores columna de la matriz $A$ con los coeficientes definidos por el vector columna $\vec{b}_{j,:}$.

Las siguientes son propiedades fácilmente corroborables para el producto matricial:

- Asociatividad: $(A\,B)\,C = A\,(B\,C)$.
- Distributividad: $A\,(B + C) = A\,B + A\,C$.
- No conmutatividad: $A\,B \neq B\,A$.

## 2.17. Independencia lineal y el rango de una matriz

Un conjunto de vectores $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\} \subset \mathbb{R}^m$ se dice que es linealmente independiente, si ningún vector de tal conjunto puede ser representado como una combinación lineal del resto de vectores. De lo contrario, si uno de los vectores en tal conjunto puede ser representado como una combinación lineal del resto de vectores, entonces los vectores son **linealmente dependientes**, lo que se expresa como:

$$\vec{x}_j = \sum_{i=1}^{n-1} \alpha_i \vec{x}_i$$

para cualquier conjunto de valores escalares $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$ se dice que el vector $\vec{x}_j \in \mathbb{R}^m$ es linealmente dependiente de los vectores $\vec{x}_i$.

El **rango de columnas** de la matriz $A \in \mathbb{R}^{m\times n}$ corresponde a la cantidad más grande de columnas en la matriz $A$ linealmente independientes, de manera similar, el **rango de filas** se refiere a la cantidad más grande de filas en tal matriz linealmente independientes.

**El rango** de una matriz $A \in \mathbb{R}^{m\times n}$ es la cantidad máxima de filas l.i si $m < n$, o la cantidad máxima de columnas l.i si $n < m$. Si $m = n$ entonces el mínimo entre el rango de filas y el rango de columnas.

- $\forall A \in \mathbb{R}^{m\times n}$, rango $(A) \leq \text{mín}\,(m, n)$, y si rango $(A) = \min(m, n)$ se dice que $A$ de **rango completo.**
- rango $(A) \leq$ rango $\left(A^T\right)$
- rango $(A\,B) \leq \text{mín}\,(\text{rango}\,(A)\,, \text{rango}\,(B))$
- rango $(A + B) \leq$ rango $(A) +$ rango $(B)$

Ejemplo:

Observe la siguiente matriz:

$$\begin{bmatrix} 1 & 2 & -1 & 3 & -2 \\ 2 & 1 & 0 & 1 & 1 \\ 2 & 4 & -2 & 6 & -4 \\ 0 & 0 & 0 & 0 & 0 \\ 5 & 4 & -1 & 5 & 0 \end{bmatrix}$$

Fácilmente puede notarse que la fila $f_3 = 2f_1$ y además que $f_5 = 2f_2 + f_1$, y que dado que la fila $f_4$ es nula, entonces puede ser expresada en términos de cualquier otra fila en una combinación lineal.

## 2.18. La matriz inversa

La inversa de la matriz cuadrada $A \in \mathbb{R}^{n \times n}$ se denota como $A^{-1}$ es la única matriz que cumple lo siguiente:

$$A^{-1}A = I = A\,A^{-1}$$

Nótese que no todas las matrices tienen inversas, por ejemplo las matrices no cuadradas no tienen inversas por definición, e incluso, pueden existir matrices cuadradas sin inversas.

- Se dice que $A$ es una matriz **invertible** o no singular si $A^{-1}$ existe, si la matriz $A$ presenta **rango completo**, lo que quiere decir que las matrices con filas o columnas que son combinación lineal de otras filas o columnas, no son invertibles.
- Si la matriz $A^{-1}$ no existe, se dice que la matriz es **no invertible** o singular.

Las siguientes son las propiedades de la inversa, asumiendo que $A, B \in \mathbb{R}^{n \times n}$ son no-singulares:

- $\left(A^{-1}\right)^{-1} = A$.
- $(A\,B)^{-1} = B^{-1}A^{-1}$.

  Para demostrar tal propiedad, suponiendo que es cierta, entonces se debe cumplir que $(A\,B)\left(B^{-1}A^{-1}\right) = I$, lo cual es cierto pues $AIA^{-1} = I$.
- $\left(A^{-1}\right)^T = \left(A^T\right)^{-1}$

### 2.18.1. La matriz pseudo inversa

Para una matriz $A \in \mathbb{R}^{m \times n}$, la matriz pseudo inversa $A^+ \in \mathbb{R}^{n \times m}$ es la generalización de una matriz inversa, y se calcula usando métodos como el de Moore-Penrose basado en la descomposición de valores singulares. Básicamente consiste en una aproximación a la inversa de una matriz, usada habitualmente cuando la inversa $A^{-1}$ no existe ($A$ no es invertible). Dado que cuando la matriz es invertible, existe una solución única, el uso de la pseudo inversa cuando la inversa no existe, permite aproximar por ejemplo a la mejor solución posible, y de forma similar para mínimos cuadrados, si existen múltiples soluciones, calcula la mejor solución posible.

La pseudo inversa cumple las siguientes propiedades:

- $AA^+A = A$ lo cual es una versión más relajada de la propiedad $A^{-1}A = A\,A^{-1} = I$

- $A^+AA^+ = A^+$

pero sobre todo:

- $\left(A^TA\right)^{-1}A^T \approx A^+$

## 2.19. Matrices ortogonales

Anteriormente se mencionó que dos vectores $\vec{x}, \vec{y} \in \mathbb{R}^n$ son ortogonales si $\vec{x}^T\vec{y} = 0$. Se dice dice que un vector $\vec{x} \in \mathbb{R}^n$ es normalizado si $\|\vec{x}\|_2 = 1$.

Una matriz cuadrada $U \in \mathbb{R}^{n\times n}$ es **ortogonal** si todas las columnas son ortogonales entre ellas. Si además, todos los vectores están normalizados, se dice que la matriz es **ortonormal.**

Las siguientes son propiedades de las matrices ortogonales:

- Para toda matriz ortonormal $U \in \mathbb{R}^{n\times n}$ , se cumple que:$U^TU = I = U\,U^T$ y sabiendo que $U\,U^{-1} = I$ se arriba a que $U^T = U^{-1}$. Si $U \in \mathbb{R}^{m\times n}$ y $n < m$ pero sus columnas son ortonormales, entonces se cumple que $U^TU = I$ pero $U\,U^T \neq I$.
- Para toda matriz ortonormal $U \in \mathbb{R}^{n\times n}$ y vector $\vec{x} \in \mathbb{R}^n$, se cumple que el operar el vector con una matriz ortonormal, la norma euclidiana no cambia:

$$\|U\,\vec{x}\|_2 = \|\vec{x}\|_2$$

Un ejemplo de una matriz ortonormal

```
U = [0.68567  0.12975  -0.71626; 0.14807  0.93855  0.31176;
     0.71269  -0.31982  0.62433]
```

$$U = \begin{bmatrix} 0{,}68567 & 0{,}12975 & -0{,}71626 \\ 0{,}14807 & 0{,}93855 & 0{,}31176 \\ 0{,}71269 & -0{,}31982 & 0{,}62433 \end{bmatrix}$$

## 2.20. Rango y espacio nulo de la matriz

Un **espacio generado** de un **conjunto de vectores base** $\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}$ $\vec{a}_i \in \mathbb{R}^n$ es el conjunto de vectores que pueden ser expresados como combinación lineal de tales vectores $\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}$:

$$\text{espacioGenerado}\left(\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}\right) = \left\{\vec{v} : \vec{v} = \sum_{i=1}^{m} x_i\vec{a}_i \qquad x_i \in \mathbb{R}^1\right\}.$$

Puede demostrarse que si el conjunto de vectores $\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}$ $\vec{a}_i \in \mathbb{R}^n$ es **linealmente independiente** (con $m \geq n$), el espacio generado por tal conjunto de vectores base es:

$$\text{espacioGenerado}\left(\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}\right) = \mathbb{R}^n.$$

Por ejemplo, los vectores unitarios anteriormente presentados $\hat{i} = \begin{bmatrix}1\\0\\0\end{bmatrix}$, $\hat{j} = \begin{bmatrix}0\\1\\0\end{bmatrix}$ y $\hat{k} = \begin{bmatrix}0\\0\\1\end{bmatrix}$ son linealmente independientes, por lo que entonces es fácil observar que la combinación lineal de tales vectores puede generar cualquier vector en el espacio $\mathbb{R}^3$. Por ejemplo, un vector $\vec{v} = \begin{bmatrix}3\\5\\7\end{bmatrix}$ se puede representar como:

$$\vec{v} = 3\hat{i} + 5\hat{j} + 7\hat{k} = 3\begin{bmatrix}1\\0\\0\end{bmatrix} + 5\begin{bmatrix}0\\1\\0\end{bmatrix} + 7\begin{bmatrix}0\\0\\1\end{bmatrix}$$

por lo que entonces $\vec{v} \in$ espacioGenerado $\left(\left\{\vec{i}, \vec{j}, \vec{k}\right\}\right) = \mathbb{R}^3$, con en este caso $x_1 = 3$, $x_2 = 5$ y $x_3 = 7$.

**La proyección de un vector** $\vec{y} \in \mathbb{R}^n$ en el espacio generado por el conjunto de vectores base $\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}$ $\quad \vec{a}_i \in \mathbb{R}^n$ corresponde al vector $\vec{v} \in$ espacioGenerado $(\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\})$ tal que $\vec{v} \in \mathbb{R}^n$ esté lo más cerca posible del vector $\vec{y} \in \mathbb{R}^n$. Esto medido con por ejemplo una norma euclidiana $\|\vec{v} - \vec{y}\|_2$ y se puede definir formalmente como:

$$\text{proy}\left(\vec{y}; \{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}\right) = \text{argmin}_{\vec{v} \in \text{espacioGenerado}(\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\})} \|\vec{v} - \vec{y}\|_2 .$$

Por otra parte, el **espacio de columnas** de una matriz $A \in \mathbb{R}^{m \times n}$ denotado como $\mathcal{C}(A)$ corresponde al espacio generado por las **columnas de la matriz** $A$, lo cual se representa como sigue:

$$\mathcal{C}(A) = \left\{\vec{v} \in \mathbb{R}^m : \vec{v} = A\,\vec{x},\ \vec{x} \in \mathbb{R}^m,\ A \in \mathbb{R}^{n \times m}\right\},$$

donde recordemos que la multiplicación matricial $A\,\vec{x}$ corresponde a una combinación lineal del vector $\vec{x}$:

$$A\overrightarrow{x} = \begin{bmatrix} | & | & \cdots & | \\ \vec{a}_{:,1} & \vec{a}_{:,2} & \cdots & \vec{a}_{:,n} \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1\left[\vec{a}_{:,1}\right] + x_2\left[a_{:,2}\right] + \ldots + x_n\left[a_{:,n}\right],$$

por lo que entonces el espacio de columnas de la matriz $A$ equivale a:

$$\mathcal{C}(A) = \text{espacioGenerado}\left(\{\vec{a}_{:,1}, \vec{a}_{:,2}, \ldots, \vec{a}_{:,n}\}\right) = \left\{v : v = \sum_{i=1}^{n} x_i \vec{a}_{:,i} \qquad x_i \in \mathbb{R}^1\right\}.$$

Asumiendo que $A$ es de **rango completo** y que $n < m$ se tiene que la proyección del vector $\vec{y} \in \mathbb{R}^n$ en el espacio de columnas de la matriz $A$ está dado por:

$$\text{proy}\left(\vec{y}; A\right) = \text{argmin}_{\vec{v} \in \mathcal{C}(A)} \|\vec{v} - \vec{y}\|_2 = \text{argmin}_{\vec{x}} \sqrt{(A\,\vec{x} - \vec{y}) \cdot (A\,\vec{x} - \vec{y})}$$

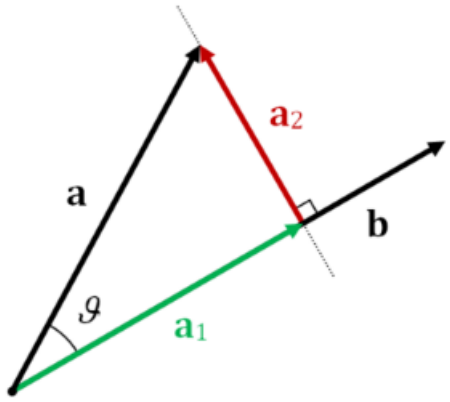



Figura 10: Proyección de vector $\vec{a}$ sobre $\vec{b}$.

$$\Rightarrow \text{proy}\,(\vec{y}; A) = \text{argmin}_{\vec{x}} \sqrt{(A\,\vec{x} - \vec{y})^T\,(A\,\vec{x} - \vec{y})}$$

El encontrar el vector que minimice la ecuación $(A\,\vec{x} - \vec{y})^T\,(A\,\vec{x} - \vec{y})$ se le llama el problema de los **mínimos cuadrados**. Usualmente se eleva el cuadrado la ecuación original de la proyección, dado que tomar su cuadrado no altera el mínimo:

$$\text{argmin}_{\vec{v}\in\mathcal{C}(A)} \|\vec{v} - \vec{y}\|_2^2 = \text{argmin}_{\vec{x}}\,(A\,\vec{x} - \vec{y}) \cdot (A\,\vec{x} - \vec{y})\,.$$

Este tema se retomará al final del presente documento, una vez que se haya definido el concepto de gradiente matricial y se demostrará que:

$$\text{proy}\,(\vec{y}; A) = \text{argmin}_{\vec{v}\in\mathcal{C}(A)} \|\vec{v} - \vec{y}\|_2 = A\,\left(A^T A\right)^{-1} A^T \vec{y}$$

o si $A$ no es invertible, entonces:

$$\text{proy}\,(\vec{y}; A) = \text{argmin}_{\vec{v}\in\mathcal{C}(A)} \|\vec{v} - \vec{y}\|_2 = A\,A^{+}\vec{y}$$

**recuerde además que** $\left(A^T A\right)^{-1} A^T = A^{+}$ **(pseudo inversa)**. Para el caso en que $A$ está formada por una sola columna $\vec{a} \in \mathbb{R}^m$ (correspondiente a un espacio generador de un vector), se tiene el caso especial de la proyección de un vector sobre otro vector:

$$\text{proy}\,(\vec{y}; \vec{a}) = \frac{\vec{a}\,\vec{a}^T}{\vec{a}^T\,\vec{a}}\vec{y}$$

Observe que en tal caso de fijar un conjunto generador de un solo vector, el subespacio generado corresponde únicamente al escalamiento de tal vector, pero la dimensionalidad del vector proyectado tiene la misma dimensionalidad original (por lo que se denomina una proyección a un sub-espacio). La Figura 10 muestra la proyección de un vector sobre otro vector.

```
function proyectar
        v1 = [3; 7];
        v2 = [9; 1];
        proy = proyectarVector(v1, v2);
        figure;
        plotv([proy v1 ]);
        figure;
        plotv([v2 v1 ]);
end
function proyec = proyectarVector(b, a)
    %proyecta b sobre a
        coefMatricial = ((a * a') / (a' * a));
   proyec = coefMatricial * b;
 end
```

El **espacio nulo** de una matriz $A \in \mathbb{R}^{m \times n}$ , se define como el conjunto de todos los vectores que al multiplicarse con la matriz $A$ resultan en $0$, y se denota como

$$\mathcal{N}(A) = \{\vec{x} \in \mathbb{R}^n : A\vec{x} = 0\}$$

**Ejemplo 1 ($m = n$, igual número de vectores en la base que dimensionalidad):**

Sean los vectores $\vec{a}_1 = \begin{bmatrix} 0{,}5 \\ 0 \\ 0 \end{bmatrix}$, $\vec{a}_2 = \begin{bmatrix} 0 \\ 0{,}25 \\ 0 \end{bmatrix}$ y $\vec{a}_3 = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$, los cuales forman la matriz:

$$A = \begin{bmatrix} 0{,}5 & 0 & 0 \\ 0 & 0{,}25 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

determine el vector de proyección $\text{proy}\,(\vec{y}; A) \in \mathbb{R}^3$, para el caso en que $\vec{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. Observe que los vectores son linealmente independientes, y además, la cantidad de vectores $m$ es menor a la dimensionalidad $n$ de $\vec{y}$. La proyección de $\vec{y}$ sobre el espacio de columnas de $A$ está dado por el vector que resulta de:

$$\text{proy}\,(\vec{y}; A) = x_1\vec{a}_1 + x_2\vec{a}_2 + x_3\vec{a}_3$$

$$\Rightarrow \text{proy}\,(\vec{y}; A) = 2\begin{bmatrix} 0{,}5 \\ 0 \\ 0 \end{bmatrix} + 8\begin{bmatrix} 0 \\ 0{,}25 \\ 0 \end{bmatrix} + 1{,}5\begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

**por lo que entonces en este caso**, $x_1 = 2$, $x_2 = 8$, $x_3 = 1{,}5$ son los coeficientes que permiten calcular el vector proyección $\text{proy}\,(\vec{y}; A)$ en el espacio generado

por las columnas de $A$. Si usamos la fórmula para determinar tal vector proyección:

$$\text{proy}\,(\vec{y};A) = \begin{bmatrix}1\\2\\3\end{bmatrix} = \text{argmin}_{\vec{v}\in\mathcal{C}(A)}\,\|\vec{v}-\vec{y}\|_2 = A\,\left(A^TA\right)^{-1}A^T\vec{y}$$

implementando el siguiente código de MATLAB, obtendremos tal resultado:

```
A = [0.5 0 0; 0 0.25 0; 0 0 2];
y = [1; 2; 3];
proyY_A = A*inv(A' * A) * A' * y;
x = inv(A) * proyY_A;
%Otra manera de obtener x1
x1 = (dot(y, A(:, 1)))/norm(A(:,1));
%equivale a
coefMatricial = ((A(:,1) * A(:,1)') / (A(:,1)' * A(:,1)))
    ;
y1 = coefMatricial * y;
x1 = norm(y1, 2);
```

Observe que como los vectores son indepedientes entre si, es posible encontrar un vector proyección que hace que $\|\text{proy}\,(\vec{y};A)-\vec{y}\|_2 = 0$ al resolver $\text{argmin}_{\vec{v}\in\mathcal{C}(A)}\,\|\vec{v}-\vec{y}\|_2$.

**De no conocer los coeficientes** $x_1 = 2$, $x_2 = 8$, $x_3 = 1{,}5$ , los mismos se pueden calcular siguiendo la ecuación:

$$\text{proy}\,(\vec{y};A) = A\,\vec{x} \Rightarrow A^{-1}\,\text{proy}\,(\vec{y};A) = \vec{x}$$

Lo cual para este caso resulta en $\vec{x} = \begin{bmatrix}2 & 8 & 1{,}5\end{bmatrix}^T$. **La magnitud de la proyección en cada uno de los vectores de la base viene dada por:**

$$\|\text{proy}\,(\vec{y};\vec{a_i})\| = \left|\frac{\vec{y}\cdot\vec{a_i}}{\|\vec{a_i}\|^2}\right| = \left\|\frac{\vec{a}_i\,\vec{a}_i^T}{\vec{a}_i^T\,\vec{a}_i}\vec{y}\right\| \neq x_i$$

**Ejemplo 2 (vectores linealmente dependientes):**

Sean los vectores $\vec{a}_1 = \begin{bmatrix}0{,}5\\0\\0\end{bmatrix}$, $\vec{a}_2 = \begin{bmatrix}1\\0\\0\end{bmatrix}$ y $\vec{a}_3 = \begin{bmatrix}0\\1\\2\end{bmatrix}$, los cuales forman la matriz:

$$A = \begin{bmatrix}0{,}5 & 1 & 0\\0 & 0 & 1\\0 & 0 & 2\end{bmatrix}$$

determine el vector de proyección $\text{proy}\,(\vec{y};A) \in \mathbb{R}^3$, para el caso en que $\vec{y} = \begin{bmatrix}1\\2\\3\end{bmatrix}$. Observe que en este caso los vectores $\vec{a}_1$ y $\vec{a}_2$ son combinación lineal uno

del otro, por lo que la matriz $A$ no es de rango completo, y por tanto no invertible, con lo que la ecuación de la proyección $\text{proy}\,(\vec{y};A) = A\left(A^TA\right)^{-1}A^T\vec{y}$ no tiene solución, al no ser posible calcular $A^{-1}$.

**Ejemplo 3 ($m > n$, más vectores en la base que dimensionalidad):**

Sean los vectores $\vec{a}_1 = \begin{bmatrix}0{,}5\\0\\0\end{bmatrix}$, $\vec{a}_2 = \begin{bmatrix}0\\0{,}25\\3\end{bmatrix}$, $\vec{a}_3 = \begin{bmatrix}0\\0\\2\end{bmatrix}$, $\vec{a}_3 = \begin{bmatrix}23\\5\\3\end{bmatrix}$, los cuales forman la matriz: $A = \begin{bmatrix}0{,}5 & 0 & 0 & 23\\0 & 0{,}25 & 0 & 5\\0 & 3 & 2 & 3\end{bmatrix}$ determine el vector de proyección $\text{proy}\,(\vec{y};A) \in \mathbb{R}^3$, para el caso en que $\vec{y} = \begin{bmatrix}1\\2\\3\end{bmatrix}$.

Sabemos que el vector proyección viene dado entonces por $\text{proy}\,(\vec{y};A) = A\left(A^TA\right)^{-1}A^T\vec{y}$. Para calcular lo anterior, se necesita que $A$ sea una matriz cuadrada, por lo que formalmente no se puede calcular la inversa de tal matriz. Es por ello que recurrimos al cálculo de la pseudo-inversa, usando el método de Moore-Penrose, con lo que se obtiene una proyección con error cero. Implementando el siguiente código de MATLAB, se obtiene que $\text{proy}\,(\vec{y};A) = \begin{bmatrix}1 & 2 & 3\end{bmatrix}^T$.

```
A = [0.5 0 0 23; 0 0.25 0 5; 0 3 2 3];
y = [1; 2; 3];
proyY_A = A * pinv(A' * A) * A' * y;
```

**Ejemplo 4 ($m < n$, menos vectores en la base que dimensionalidad):**

Sean los vectores $\vec{a}_1 = \begin{bmatrix}5\\7\\21\end{bmatrix}$ y $\vec{a}_2 = \begin{bmatrix}0\\13\\9\end{bmatrix}$, los cuales forman la matriz: $A = \begin{bmatrix}5 & 0\\7 & 13\\21 & 9\end{bmatrix}$ determine el vector de proyección $\text{proy}\,(\vec{y};A) \in \mathbb{R}^3$, para el caso en que $\vec{y} = \begin{bmatrix}1{,}2\\1{,}3\\1{,}5\end{bmatrix}$.

Sabemos que el vector proyección viene dado entonces por $\text{proy}\,(\vec{y};A) = A\left(A^TA\right)^{-1}A^T\vec{y}$. Para calcular lo anterior, se necesita que $A$ sea una matriz cuadrada, por lo que formalmente no se puede calcular la inversa de tal matriz. Es por ello que recurrimos al cálculo de la pseudo-inversa, usando el método de Moore-Penrose. Implementando el siguiente código de MATLAB, se obtiene que $\text{proy}\,(\vec{y};A) = \begin{bmatrix}1{,}1542 & 13{,}3663 & 14{,}9695\end{bmatrix}^T$. Obsérvese que en el caso en

que se disponen menos vectores en la base respecto a la dimensionalidad del vector a proyectar, por lo que la proyección tiene un error distinto de cero.

```
A = [5 0; 7 13; 21 9];
y = [1.2; 1.3; 1.5];
proyY_A = A * pinv(A' * A) * A' * y;
x = pinv(A) * proyY_A;
u1 = dot(y, A(:, 1)) / norm(A(:, 1));
u2 = dot(y, A(:, 2)) / norm(A(:, 2));
figure;
quiver3(0,0,0, A(1,1), A(2,1), A(3,1));
hold on;
quiver3(0,0,0, A(1,2), A(2,2), A(3,2));
hold on;
quiver3(0,0,0, y(1), y(2), y(3));
hold on;
quiver3(0,0,0, u1, u2, 5);
```

Para reducir la dimensionalidad del vector $\vec{y}$ en una dimensión, en este caso, se construye un vector $\vec{u}$ en $\mathbb{R}^2$ cuyos componentes están definidos por la proyección :

$$u_1 = \frac{\vec{y} \cdot \vec{a_1}}{\|\vec{a_1}\|} = 2{,}0534$$

$$u_2 = \frac{\vec{y} \cdot \vec{a_2}}{\|\vec{a_2}\|} = 1{,}9227$$

Observe que las operaciones anteriores pueden resultar en un número negativo, por lo que preservan la dirección del vector $\vec{u}^T = \begin{bmatrix} 2{,}0534 & 1{,}9227 \end{bmatrix}$ en $\mathbb{R}^2$, a diferencia de usar $|u_i| = \left\| \frac{\vec{a}_i \vec{a}_i^T}{\vec{a}_i^T \vec{a}_i} \vec{y} \right\|$ . El vector $\vec{x}$ en proy $(\vec{y}; A) = A\vec{x}$ nos indica los coeficientes en un espacio expresado en términos de los vectores base de $A$, pero como seguimos dibujando en un espacio $\mathbb{R}^2$ cuya base son los vectores unitarios $\hat{i}, \hat{j}$, la reduccción de dimensionalidad se hace usando la proyección de $\vec{y}$ sobre cada vector base $\vec{a}_i$.

## 2.21. Determinante de una matriz

El determinante de una matriz cuadrada $A \in \mathbb{R}^{n \times n}$ es una función denotada con $\det(A) : \mathbb{R}^{n \times n} \to \mathbb{R}$. Antes de detallar la fórmula que define al determinante, examinaremos la interpretación geométrica del determinante. Sea una matriz compuesta por múltiples filas:

$$A = \begin{bmatrix} - & \vec{a}_{1,:}^T & - \\ - & \vec{a}_{2,:}^T & - \\ & \vdots & \\ - & \vec{a}_{n,:}^T & - \end{bmatrix}$$

considere el conjunto de puntos $S \subset \mathbb{R}^n$ formado al tomar todas las combinaciones lineales posibles de los vectores fila $\vec{a}_{i,:}^T$, donde los coeficientes de tal combinación lineal cumplen que $0 \leq \alpha_i \leq 1, i = 1, \ldots, n$, lo cual formalmente se denota como:

$$S = \left\{ \vec{v} \in \mathbb{R}^n : \vec{v} = \sum_{i=1}^{n} \alpha_i \vec{a}_{i,:}, \qquad 0 \leq \alpha_i \leq 1, i = 1, \ldots, n \right\}$$

El valor absoluto del determinanate de la matriz $A$, $|\det(A)|$, corresponde a una medida del "volumen" de todo el conjunto $S$.

Por ejemplo, para la matriz $A \in \mathbb{R}^{2\times 2}$:

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}$$

cuyos vectores fila están dados por:

$$\vec{a}_{1,:} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \qquad \vec{a}_{2,:} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

se muestra en la Figura 11, sombreado, el conjunto de puntos $S$. Observe que el punto "extremo" $\vec{a}_{1,:} + \vec{a}_{2,:} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$, viene dado cuando $\alpha_1 = \alpha_2 = 1$. El determinante para una matriz de $2 \times 2$ se define como:

$$\det\left( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = a\,d - b\,c$$

y para cualquier matriz de $n \times n$ dimensiones, **el determinante se define recursivamente** como:

$$\det(A) = A_{1,1}\det\left(A_{\setminus f,\setminus 1}\right) - A_{f,2}\det\left(A_{\setminus f,\setminus 2}\right) + \ldots \pm A_{f,n}\det\left(A_{\setminus f,\setminus n}\right)$$

donde el operador $A_{\setminus i,\setminus j}$ denota la eliminación de la fila $i$ y la columna $j$ de la matriz $A$. Se puede escoger cualquier fila a eliminar $f$, o también se puede variar la fila y escoger una columna fija, por lo que:

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} A_{i,j} \left|A_{\setminus i,\setminus j}\right| = \sum_{j=1}^{n} (-1)^{i+j} A_{i,j} \left|A_{\setminus i,\setminus j}\right|$$

Observe que el determinante consiste en la combinación lineal de los determinantes de las submatrices resultantes de eliminar la fila y columna $i$ (denotado como $\det\left(A_{\setminus i,\setminus j}\right)$), multiplicado por el elemento $A_{1,i}$. Con la matriz de ejemplo $A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}$, el determinante viene entonces dado por: $\det(A) = 1 \cdot 2 - 3 \cdot 3 = -7$, y tomando su valor absoluto, se tiene que $|\det(A)| = 7$, lo que corresponde al área del paralelogramo formado por el conjunto de puntos $S$ (en $n$ dimensiones, se refiere como paralelótopo).

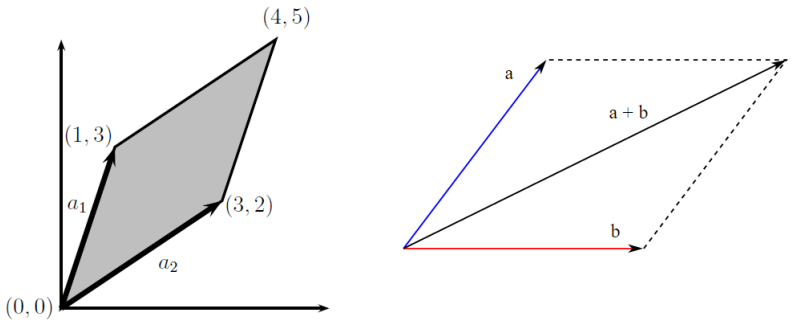



Figura 11: Región $S$ de ejemplo.

Como ejemplo de la recursividad de la definición del determinante, tómese la siguiente matriz:

$$A = \begin{bmatrix} 1 & 4 & 9 \\ 7 & 2 & 5 \\ 6 & 8 & 3 \end{bmatrix}$$

$$\det(A) = 1 \cdot \det\left(\begin{bmatrix} 2 & 5 \\ 8 & 3 \end{bmatrix}\right) - 4 \cdot \det\left(\begin{bmatrix} 7 & 5 \\ 6 & 3 \end{bmatrix}\right) + 9 \cdot \det\left(\begin{bmatrix} 7 & 2 \\ 6 & 8 \end{bmatrix}\right)$$

$$\Rightarrow \det(A) = 1 \cdot (2 \cdot 3 - 5 \cdot 8) - 4 \cdot (7 \cdot 3 - 5 \cdot 6) + 9 \cdot (7 \cdot 8 - 2 \cdot 6)$$

$$\Rightarrow \det(A) = -34 + 36 + 396 = 398$$

Las siguientes son propiedades de la función determinante $\det(A)$ para una matriz cuadrada $A \in \mathbb{R}^{n \times n}$ :

- El volumen de un hipercubo unitario es $\det(I) = 1$.
- Homogeneidad: Sea un escalar $s \in \mathbb{R}$, $\det(s\,A) = s \det(A)$
- $\det(A) = \det\left(A^T\right)$
- $\det(A\,B) = \det(A)\ \det(B)$
- $\det(A) = 0$, **implica que $A$ es una matriz singular (no invertible)**, por lo que entonces no tiene rango completo, y sus columnas son **linealmente dependientes**, por lo que entonces la superficie $S$ no tiene volumen, al un vector no contribuir en cerrar el cuerpo.
- $\det\left(A^{-1}\right) = 1/\det(A)$

## 2.22. Producto cruz

El producto de dos vectores $a$ y $b$ es otro vector:

$$a \times b = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \times \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{bmatrix}$$

El producto cruz $a \times b$ es un vector que es perpendicular al plano en el que los vectores $a$ y $b$ se encuentran.

Su largo es igual al paralelogramo que los vectores $a$ y $b$ contienen.

## 2.23. Autovalores y auto-vectores

Sea una matriz cuadrada $A \in \mathbb{R}^{n \times n}$ , decimos que $\lambda \in \mathbb{C}$ es un **auto-valor** o eigen-valor de $A$ y el vector $\vec{x} \in \mathbb{C}^n$ es su **auto-vector** o eigen-vector si:

$$A\,\vec{x} = \lambda \vec{x}, \qquad \vec{x} \neq 0 \tag{10}$$

Intituivamente, la ecuación anterior significa que la multiplicación de la matriz $A$ por un vector $\vec{x}$ es igual a la multiplicación de tal vector $\vec{x}$ por el escalar $\lambda$ también referido como el **escalamiento del vector** $\vec{x}$.

Los auto-vectores se expresan como vectores normalizados (con norma 1), puesto que cualquier vector escalado de $\vec{x}$, $\vec{v} = c\,\vec{x}$ hace que la ecuación $A\,\vec{v} = \lambda \vec{v}$ se siga cumpliendo, **lo que hace que la ecuación 10 tenga múltiples soluciones**. Siguiendo la ecuación 10, se tiene que:

$$A\,\vec{x} - \lambda \vec{x} = (\lambda I - A)\,\vec{x} = 0, \qquad \vec{x} \neq 0 \tag{11}$$

La ecuación anterior tiene solución no nula o no-cero si y solo sí la matriz $(\lambda I - A)$ tiene un espacio nulo no vacío, lo cual es el caso **si y solo sí tal matriz es singular (no-invertible)**, por lo que en términos del determinante debe cumplir que:

$$\det(\lambda I - A) = 0$$

Esto pues en general, si un sistema de ecuaciones $A\vec{x} = \vec{b}$, se tiene que $A$ es invertible, existe entonces una solución única $\vec{x} = A^{-1}\vec{b}$ . **Si la matriz $A$ no es invertible, existen múltiples soluciones**.

De esta forma, con el cálculo del determinante, se construye el polinomio en términos de la variable $\lambda$ y de grado $n$, para lo cual se encuentran las raíces de tal polinomio. Una vez conocidos los auto-valores $\lambda$, se procede a buscar sus autovectores correspondientes $\vec{x}$, resolviendo la ecuación matricial $(\lambda I - A)\,\vec{x} = 0$.

Un **ejemplo** puede dilucidar mejor el procedimiento, sea la matriz

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

La matriz para cuyo espacio nulo se realizará el cálculo viene dada por:

$$(\lambda I - A) = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} = \begin{bmatrix} \lambda & -1 \\ 2 & \lambda + 3 \end{bmatrix}$$

y su determinante está dado entonces por:

$$\det(\lambda I - A) = \det \begin{bmatrix} \lambda & -1 \\ 2 & \lambda + 3 \end{bmatrix} = \lambda^2 + 3\lambda + 2 = 0$$

Resolviendo tal ecuación cuadrática se obienen las raíces y por ende auto-vectores $\lambda_1 = -1$ y $\lambda_2 = -2$. Se procede entonces a encontrar los auto vectores $\vec{x}_1$ y $\vec{x}_2$. Para el auto-vector $\vec{x}_1$:

$$(\lambda_1 I - A)\,\vec{x}_1 = 0 \Rightarrow \begin{bmatrix} \lambda_1 & -1 \\ 2 & \lambda_1 + 3 \end{bmatrix} \vec{x}_1 = 0 \Rightarrow \begin{bmatrix} -1 & -1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

Resolviendo tal sistema de ecuaciones, se obtiene que el auto-vector $\vec{x}_1$ viene dado por:

$$\vec{x}_1 = k_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

y de manera similar, se obtiene el auto-vector $\vec{x}_2$:

$$\vec{x}_2 = k_2 \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

Sin embargo, la notación anterior de los auto-vectores no está normalizada, por lo que usualmente se expresan de forma normalizada:

$$\widehat{x}_1 = k_1 \frac{\vec{x}_1}{\|\vec{x}_1\|_2} = k_1 \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

$$\widehat{x}_2 = k_2 \frac{\vec{x}_2}{\|\vec{x}_2\|_2} = k_2 \begin{bmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$$

En MATLAB los auto-vectores y auto-valores pueden calcularse como sigue:

```
>> A=[0  1;-2  -3]
>> [v,d] = eig(A)
```

Dada la complejidad de resolver el determinante para matrices grandes, se implementan otros métodos numéricos para calcular los auto-valores y auto-vectores.

Detallando más la interpretación geométrica de los auto-vectores, recordamos la igualdad en términos de la matriz cuadrada $A \in \mathbb{R}^{n \times n}$ , los auto-valores y auto-vectores $A\,\vec{v} = \lambda \vec{v}$, la matriz $A$ actúa como una transformación del auto-vector $\vec{v}$, **la cual "envía" el vector a un nuevo punto del espacio**, como se

observa en la Figura 12. Recuerde que la multiplicación $A\,\vec{x}$ realiza una combinación lineal de los componentes de $\vec{v}$:

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} a_{1,1}v_1 + a_{1,2}v_2 \\ a_{2,1}v_1 + a_{2,2}v_2 \end{bmatrix} = \begin{bmatrix} a_{1,1} \\ a_{2,1} \end{bmatrix} v_1 + \begin{bmatrix} a_{1,2} \\ a_{2,2} \end{bmatrix} v_2$$

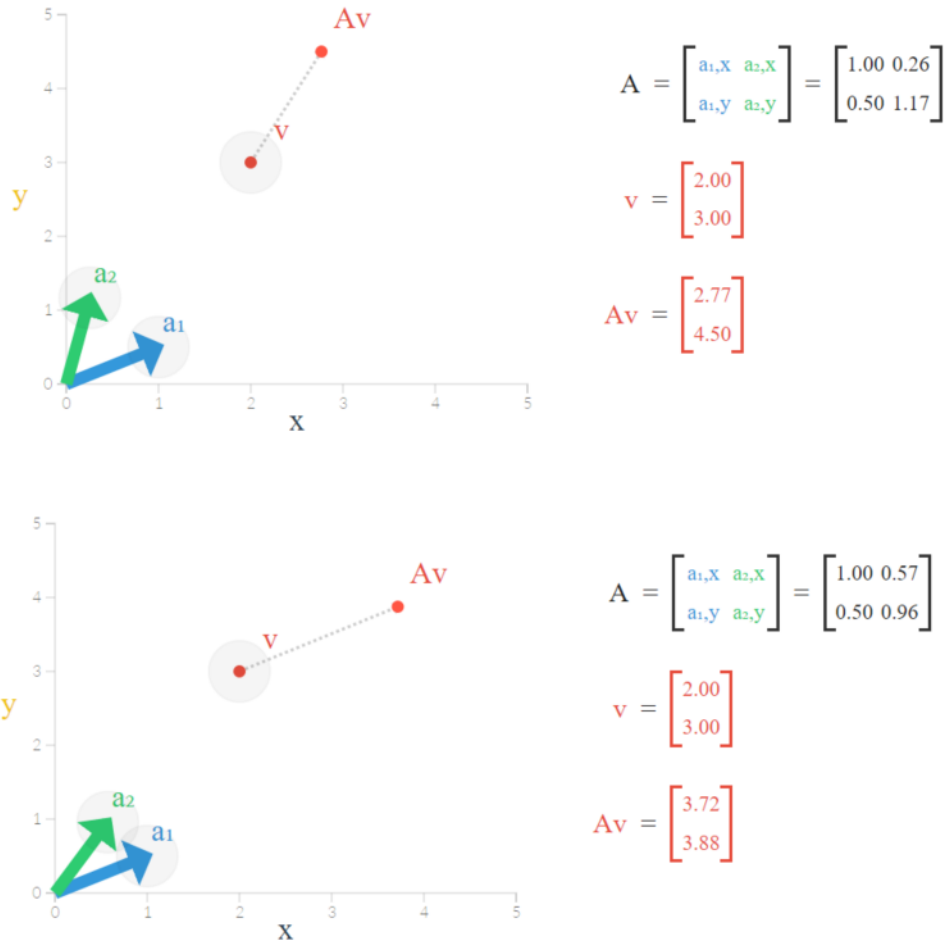



Figura 12: Transformación $A\,\vec{x}$, tomado de `http://setosa.io/ev/eigenvectors-and-eigenvalues/`.

Observe entonces que la ecuación de los auto-vectores $A\,\vec{v} = \lambda\vec{v}$ corresponde a los vectores que transformados por la matriz $A$, son escalados por su auto-valor correspondiente$\lambda$, es decir, son los vectores que **conservan su dirección al ser transformados por la matriz** $A$. Esto se puede verificar fácilmente, si es posible dibujar una línea recta entre los puntos $(0, 0)$, y el final

de los vectores $\vec{v}$ y $A\vec{v}$. En la Figura 12, el **primer caso puede corresponder a un auto-vector, mientras que el segundo, posiblemente no, pues la dirección del mismo es modificada al transformarse por la matriz** $A$. Además, es posible ver que todos los vectores con la misma dirección de un auto-vector $\vec{v}$, son también auto-vectores de tal matriz $A$, como se observa en la Figura 13, donde el auto-vector $\vec{s}_1$ es colinear con el vector $\vec{v}$. Finalmente, es importante notar que un auto-vector $\vec{s}_1$con su auto-valor $\lambda_1 < 1$ denota una transformación $A$ que "encoge" al auto-vector $\vec{s}_1$, y si $\lambda_1 > 1$ más bien lo "alarga", como se ve en la Figura 13. **En la gráfica, en realidad cada recta** $s_1$ **y** $s_2$**contienen una infinidad de auto vectores, por lo que se les llama auto espacios**.

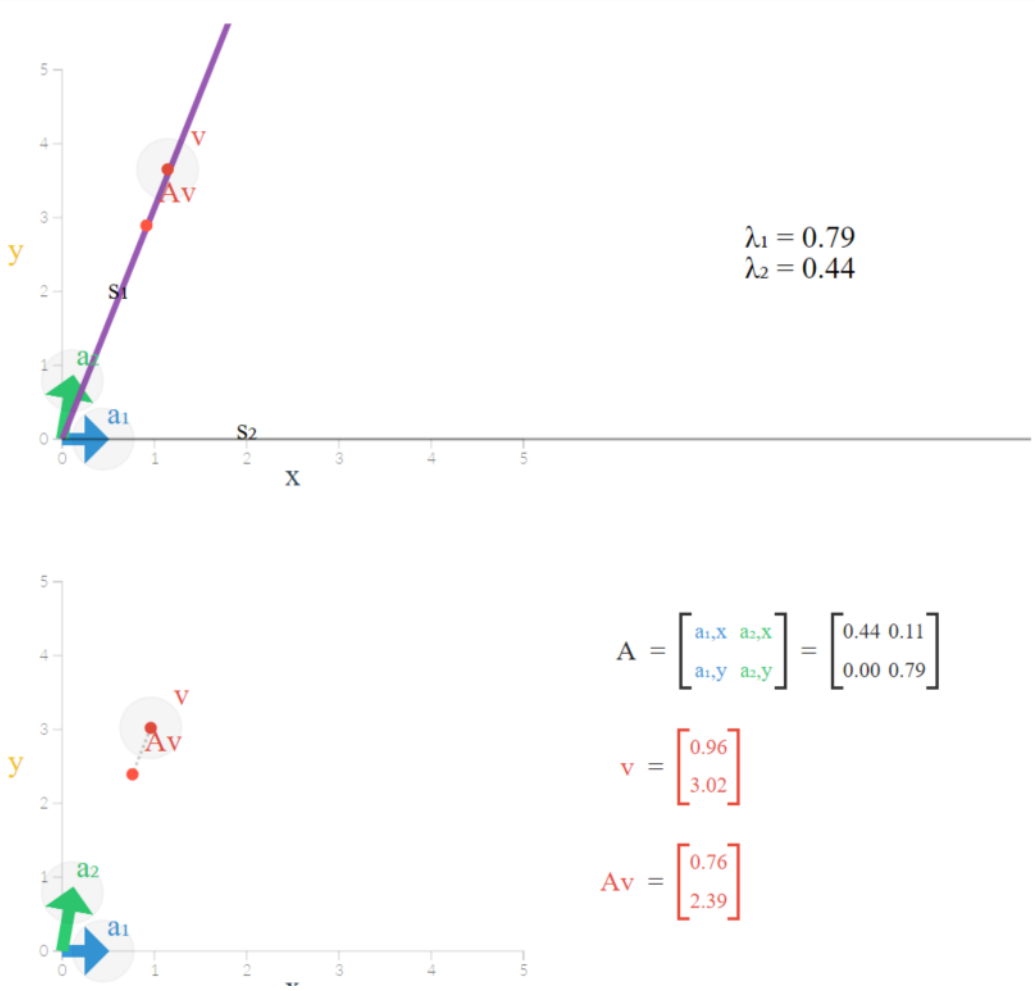



Figura 13: Auto-vectores $\vec{s}_1$ y $\vec{s}_2$, tomado de `http://setosa.io/ev/eigenvectors-and-eigenvalues/`

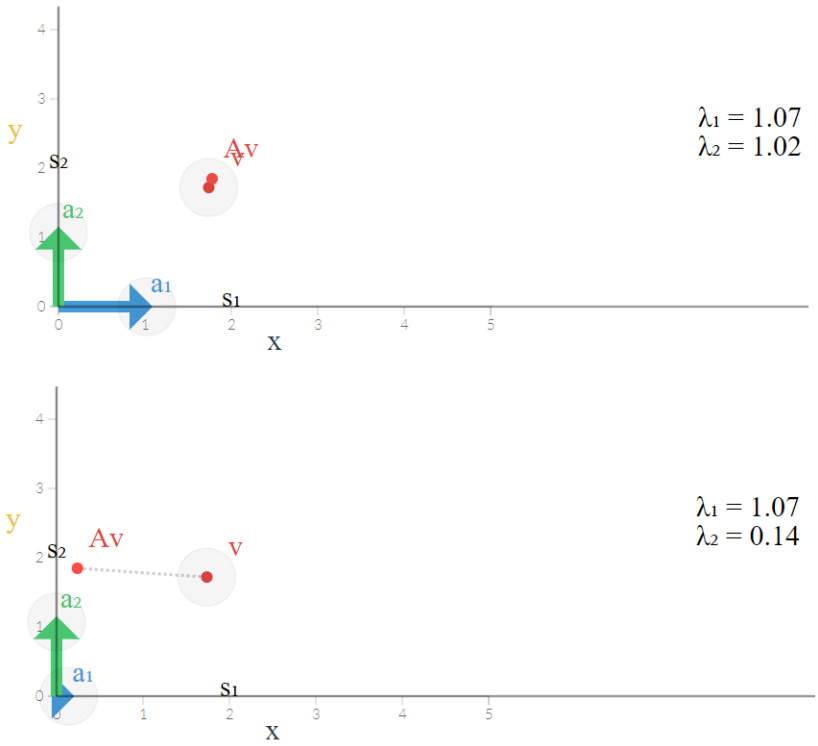



Figura 14: Ejemplo de vectores $\overrightarrow{a}_{:,1}$ y $\overrightarrow{a}_{:,2}$, y sus autoespacios. Tomado de tomado de `http://setosa.io/ev/eigenvectors-and-eigenvalues/`

La **magnitud del auto valor $\lambda_i$ respecto al resto de autovalores, define la contribución de ese autovector en la representación de todos los vectores de** $A$. Por ejemplo, en la Figura 14, si ambos vectores columna dentro de $A$, son perpendiculares, los autovectores describirán tales ejes perpendiculares. Uno de los autovalores tendrá mayor magnitud respecto al otro, en la medida en que una de las columnas en $A$ tenga mayor magnitud respecto a la otra.

Las siguientes son propiedades de los auto-valores y los auto-vectores, donde $A \in \mathbb{R}^{n \times n}$, $\vec{x} \in \mathbb{R}^n$ y $\lambda \in \mathbb{R}$:

- La traza de la matriz $A$ es igual a la suma de sus auto-valores:

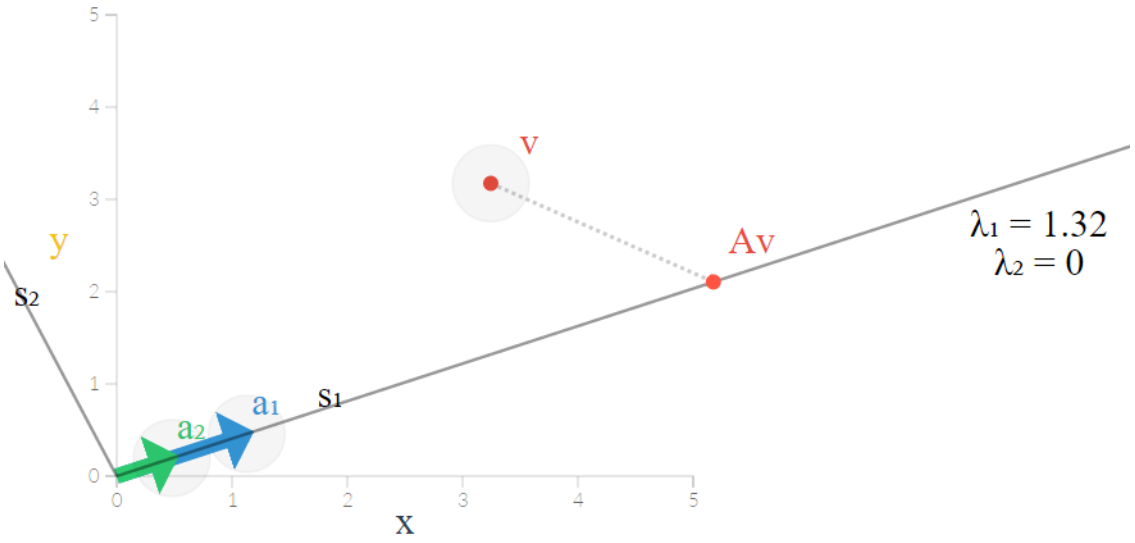
$$\text{tr}\,(A) = \sum_{i=1}^{n} \lambda_i.$$

- El determinante de la matriz $A$ es igual al producto de sus auto-valores:

$$\det(A) = \prod_{i=1}^{n} \lambda_i$$

- **El rango de la matriz $A$ es igual al número de auto-valores no nulos**. Lo anterior tiene mucho sentido, pues si por ejemplo, con una matriz $A \in \mathbb{R}^{2\times 2}$ con una columna combinación lineal de la otra, la contribución independiente de una de ellas en $A$ como transformación lineal es nula, como se ilustra en la Figura 15.

  con




con

Figura 15: Auto-vectores de una matriz con columnas que son combinación lineal de la otra. Tomado de tomado de `http://setosa.io/ev/eigenvectors-and-eigenvalues/`

- Si $A$ es no-singular (invertible), entonces $1/\lambda_i$ es un auto-valor de $A^{-1}$ con su auto-vector asociado $\vec{x}_i$, por lo que entonces $A^{-1}\vec{x}_i = (1/\lambda_i)\,\vec{x}_i$.
- Los auto-valores de una matriz diagonal $D = \text{diag}\,(d_1, \ldots d_n)$ corresponden a tales entradas diagonales $d_1, \ldots d_n$.
- **Para una matriz simétrica $A \in \mathbb{R}^{n\times n}$ existen $\{v_1, v_2, \ldots, v_n\}$ auto-vectores mutuamente ortogonales (importante propiedad, relacionado con la matriz de covarianza)**.

Es usual que para expresar más facilmente los auto-vectores y auto-valores en una sola ecuación:

$$A\,X = X\,\Lambda$$

con la matriz $X \in \mathbb{R}^{n \times n}$ la cual agrupa los auto-vectores por columnas como:

$$X = \begin{bmatrix} | & | & \dots & | \\ \vec{x}_{1,:} & \vec{x}_{2,:} & \dots & \vec{x}_{n,:} \\ | & | & \dots & | \end{bmatrix}$$

y la matriz $\Lambda = \text{diag}\,(\lambda_1, \dots \lambda_n)$ con los auto-valores de la tranformación $A$ en su diagonal. Note que los auto-vectores de $A$ pueden ser linealmente dependientes, por lo que solo si los vectores columna son linealmente independientes se puede escribir:

$$A = X\,\Lambda\,X^{-1}$$

si lo anterior es posible de escribir, se dice que la matriz $A$ es **diagonalizable**.

### 2.23.1. Auto-valores y auto-vectores de matrices simétricas

Las siguientes son propiedades de los auto-valores y auto-vectores para cualquier matriz simétrica $A \in \mathbb{S}^n$ , $A = A^T$ :

- Los auto-valores de la matriz son siempre reales.
- Los auto-vectores son ortonormales, es decir, la matriz $A\,X = X\,\Lambda$ es una matriz ortogonal, por lo que la matriz con los auto-vectores $X$ se denota como $U$, por ello: $A = U\,\Lambda\,U^{-1}$, y recordando que la transpuesta de una matriz inversa equivale a la transpuesta, se tiene que:

$$A = U\,\Lambda\,U^T$$

### 2.23.2. Formas cuadráticas y matrices positivamente definidas

Sea una matriz cuadrada $A \in \mathbb{R}^{n \times n}$, un vector $\vec{x} \in \mathbb{R}^{n \times 1}$, se le llama al escalar $\vec{x}^T\,A\,\vec{x} \in \mathbb{R}$:

$$q = \vec{x}^T\,A\,\vec{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}x_i x_j.$$

Se le llama forma cuadrática, pues suponiendo que el vector $\vec{x}$ corresponde a variables desconocidas, y la matriz $A$ está compuesta por coeficientes conocidos, entonces $\vec{x}^T\,A\,\vec{x}$ corresoponde a un polinomio de forma cuadrática (por ejemplo si $\vec{x} \in \mathbb{R}^2$ un polinomio de forma cuadrática viene dado por $a\,x_1^2 + b\,x_1x_2 + c\,y^2$). Para ilustrar lo anterior, considere la matriz:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$q = \vec{x}^T\,A\,\vec{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 + 2x_2 & 2x_1 + x_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\Rightarrow q = x_1^2 + 2x_1x_2 + 2x_1x_2 + x_2^2 = x_1^2 + 4x_1x_2 + x_2^2,$$

lo cual corresponde a una forma cuadrática.

A continuación se dan las siguientes definiciones, para una matriz $A \in \mathbb{S}^n$ (espacio de las matrices simétricas positivas) y los vectores no nulos $\vec{x} \in \mathbb{R}^n$:

- **Matriz positiva definida**: $A$ es positiva definida si $\vec{x}^T A \vec{x} > 0$.
- **Matriz positiva semidefinida**: $A$ es positiva semidefinida si $\vec{x}^T A \vec{x} \geq 0$.
- **Matriz negativa definida**: $A$ es negativa definida si $\vec{x}^T A \vec{x} < 0$.
- **Matriz negativa semidefinida**: $A$ es negativa semidefinida si $\vec{x}^T A \vec{x} \leq 0$.
- **Matriz indefinida**: $A$ es indefinida si existen al menos dos vectores $\vec{x}_1$ y $\vec{x}_2$ $\vec{x_1}^T A \vec{x_1} < 0$ y $\vec{x_2}^T A \vec{x_2} < 0$.con

Observe que si $A$ es positiva definida, entonces $-A$ es negativa definida.

Una propiedad importante es que las **matrices positivas y negativas definidas son siempre de rango completo**, por lo tanto invertibles y con todas sus columnas independientes linealmente. Probemos lo anterior a través de un contra-ejemplo. Suponga una matriz $A \in \mathbb{R}^{n \times n}$ , $A = \begin{bmatrix} | & | & \dots & | \\ \vec{a}_{1,:} & \vec{a}_{2,:} & \dots & \vec{a}_{n,:} \\ | & | & \dots & | \end{bmatrix}$ la cual no es de rango completo, y por tanto, tiene una columna linealmente dependiente del resto de columnas: =

$$\vec{a}_j = \sum_{i \neq j} x_i \vec{a}_i$$

con los coeficientes de combinación lineal $x_1, \dots, x_n \in \mathbb{R}$. Si fijamos a $x_j = -1$, se tiene que:

$$A \vec{x} = A \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i \vec{a}_i = 0$$

lo que **demuestra que existe un vector no nulo que hace $\vec{x}^T A \vec{x} = 0$, por lo que entonces $A$ no puede ser definida** ni semidefinida, y queda demostrado que para que $A$ sea tanto positiva o negativamente definida, debe ser de rango completo.

## 2.24. Cálculo matricial

A continuación se presentan conceptos básicos del cálculo matricial, el cual consiste en extender en espacios de mayor dimensionalidad los conceptos del cálculo diferencial e integral.

### 2.24.1. El gradiente

Suponga una función multivariable, la cual toma múltiples entradas (representadas en la matriz $A \in \mathbb{R}^{m\times n}$) y retorna una salida escalar $s \in \mathbb{R}$, por lo que $f : \mathbb{R}^{m\times n} \to \mathbb{R}$.

El **gradiente** de la función $f$ con respecto a su entrada $A \in \mathbb{R}^{m\times n}$ es la matriz de derivadas parciales definidas como:

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{1,1}} & \frac{\partial f(A)}{\partial A_{1,2}} & \cdots & \frac{\partial f(A)}{\partial A_{1,n}} \\ \frac{\partial f(A)}{\partial A_{2,1}} & \frac{\partial f(A)}{\partial A_{2,2}} & \cdots & \frac{\partial f(A)}{\partial A_{2,n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m,1}} & \frac{\partial f(A)}{\partial A_{m,2}} & \cdots & \frac{\partial f(A)}{\partial A_{m,n}} \end{bmatrix}$$

en notación compacta, cada entrada viene dada por:

$$(\nabla_A f(A))_{i,j} = \frac{\partial f(A)}{\partial A_{i,j}}$$

en particular, para una entrada definida en un vector $\vec{x} \in \mathbb{R}^n$ el gradiente se define como:

$$\nabla_{\vec{x}} f(\vec{x}) = \begin{bmatrix} \frac{\partial f(A)}{\partial x_1} \\ \vdots \\ \frac{\partial f(A)}{\partial x_n} \end{bmatrix}.$$

Es importante remarcar que el **gradiente sólo está definido si la función retorna un escalar**. Esto quiere decir que por ejemplo, no es posible tomar el gradiente de $A\vec{x}$, pues el resultado de tal producto matricial es un vector, y no un escalar.

La derivada matricial parcial es también un operador lineal, tal como la derivada parcial de una función multivariable, por lo que entonces cumple las propiedades de homogeneidad y superposición:

- $\nabla_{\vec{x}}(f(\vec{x}) + g(\vec{x})) = \nabla_{\vec{x}} f(\vec{x}) + \nabla_{\vec{x}} g(\vec{x})$
- Para un escalar $s \in \mathbb{R}$, $\nabla_{\vec{x}}(s\, f(\vec{x})) = s\, \nabla_{\vec{x}} f(\vec{x})$
- Un ejemplo de una función multidimensional con un vector de entrada es la función $f : \mathbb{R}^n \to \mathbb{R}$

  $$f(\vec{z}) = \vec{z}^T \vec{z} = \sum_{i=1}^{n} z_i^2$$

  la cual como se observa, calcula el producto punto $\vec{z} \cdot \vec{z}$ de su vector entrada $\vec{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}$.

Examinando cada una de las $m$ derivadas parciales $\frac{\partial f(\vec{z})}{\partial z_k}$ ( se puede obviar el hecho de que la entrada está dada por un vector y tratar como cualquier función multivariable) se tiene que:

$$\frac{\partial f(\vec{z})}{\partial z_k} = \cancel{\frac{\partial}{\partial z_k} z_1^2}^{0} + \cancel{\frac{\partial}{\partial z_k} z_2^2}^{0} + \ldots + \cancel{\frac{\partial}{\partial z_k} z_k^2}^{2\,z_k} + \ldots + \cancel{\frac{\partial}{\partial z_i} z_n^2}^{0} = 2\,z_k.$$

Es por ello que el vector gradiente está dado por:

$$\nabla_{\vec{z}} f(\vec{z}) = \begin{bmatrix} \frac{\partial f(\vec{z})}{\partial z_1} \\ \vdots \\ \frac{\partial f(\vec{z})}{\partial z_n} \end{bmatrix} = \begin{bmatrix} 2\,z_1 \\ \vdots \\ 2\,z_n \end{bmatrix} = 2\,\vec{z}.$$

por lo que entonces el equivalente de la derivada de una función cuadrática de una variable es:

$$\nabla_{\vec{z}} f(\vec{z}) = \nabla_{\vec{z}} \left(\vec{z}^T \vec{z}\right) = 2\,\vec{z}.$$

- Generalizando la función anterior, la cual recibe un vector $\vec{x} \in \mathbb{R}^n$ como entrada, y con un vector conocido $\vec{b} \in \mathbb{R}^n$:

$$f(\vec{x}) = \vec{b}^T \vec{x} = \sum_{i=1}^{n} b_i x_i$$

con lo que su derivada parcial viene entonces dada por:

$$\frac{\partial f(\vec{x})}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k$$

Es por ello que se tiene entonce que:

$$\nabla_{\vec{x}} \left(\vec{b}^T \vec{x}\right) = \vec{b}$$

- Considere ahora la función cuadrática (la cual como ya se vió resulta en un escalar), con $A$ una matriz cuadrada y simétrica:

$$f(\vec{x}) = \vec{x}^T A\,\vec{x} = \vec{x}^T \begin{bmatrix} - & \vec{a}_{1,:}^T & - \\ - & \vec{a}_{2,:}^T & - \\ & \vdots & \\ - & \vec{a}_{m,:}^T & - \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix} \begin{bmatrix} \vec{a}_{1,:}^T\,\vec{x} \\ \vec{a}_{2,:}^T\,\vec{x} \\ \vdots \\ \vec{a}_{m,:}^T\,\vec{x} \end{bmatrix} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} x_i x_j$$

Para calcular la derivada parcial $\frac{\partial f(\vec{x})}{\partial x_k}$ para cada componente $x_k$ del vector de entrada $\vec{x}$, se descomponen las sumatorias anidadas en los casos en que no la fila y columna de tal sumatoria es distinta a $k$, en que la fila

es igual a $k$, además del caso en que la columna es igual a $k$, y finalmente, cuando se está en la fila y columna $k$:

$$\frac{\partial f\left(\vec{x}\right)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} x_i x_j$$

$$\Rightarrow \frac{\partial f\left(\vec{x}\right)}{\partial x_k} = \frac{\partial}{\partial x_k} \left[ \sum_{i \neq k}^{n} \sum_{j \neq k}^{n} A_{i,j} x_i x_j + \sum_{i \neq k}^{n} A_{i,k} x_i x_k + \sum_{j \neq k}^{n} A_{k,j} x_k x_j + A_{k,k} x_k^2 \right]$$

$$\Rightarrow \frac{\partial f\left(\vec{x}\right)}{\partial x_k} = \sum_{i \neq k}^{n} A_{i,k} x_i + \sum_{j \neq k}^{n} A_{k,j} x_j + 2 A_{k,k} x_k = \sum_{i=1}^{n} A_{i,k} x_i + \sum_{j=1}^{n} A_{k,j} x_j$$

Dado que se asume que en la forma cuadrática $A$ es simétrica, lo que implica que $A = A^T \Rightarrow A_{i,j} = A_{j,i}$, se tiene que:

$$\Rightarrow \frac{\partial f\left(\vec{x}\right)}{\partial x_k} = \sum_{i=1}^{n} A_{i,k} x_i + \sum_{j=1}^{n} A_{k,j} x_j = 2 \sum_{i=1}^{n} A_{k,i} x_i.$$

Es por ello que se concluye que el gradiente de la forma cuadrática está dado por:

$$\nabla_{\vec{x}} \left( \vec{x}^T A \, \vec{x} \right) = 2 \, A \, \vec{x}.$$

Se concluyen entonces las siguientes derivadas matriciales:

- $\nabla \left( \vec{x}^T \vec{x} \right) = 2 \, \vec{x}$
- $\nabla_{\vec{x}} \left( \vec{b}^T \vec{x} \right) = \vec{b}$
- $\nabla_{\vec{x}} \left( \vec{x}^T A \, \vec{x} \right) = 2 \, A \, \vec{x}$

## 2.25. Mínimos cuadrados

El problema de los mínimos cuadrados en este caso se definirá para encontrar, dadas la matriz de rango completo $A \in \mathbb{R}^{m \times n}$, y por ende, invertible y un vector $\vec{b} \in \mathbb{R}^{m \times 1}$, el vector $\vec{x} \in \mathbb{R}^{n \times 1}$ más cercano al **espacio de columnas** de la matriz $A$, el cual recordamos es denotado como $\mathcal{C}\left(A\right)$ y corresponde al espacio generado por las columnas de la matriz $A$, combinadas linealmente por lo componentes $x_i$ del vector $\vec{x}$:

$$\mathcal{C}\left(A\right) = \left\{ \vec{v} \in \mathbb{R}^m : \vec{v} = A \, \vec{x}, \ \vec{x} \in \mathbb{R}^n, \ A \in \mathbb{R}^{m \times n}, \right\}$$

Asumiendo que $A$ es de rango completo y que $n < m$ se tiene que la proyección del vector $\vec{b} \in \mathbb{R}^{m \times 1}$ **al cuadrado para simplificar su minimización** en el espacio de columnas de la matriz $A$ está dado por:

$$\text{proy}\left(\vec{b}; A\right) = \text{argmin}_{\vec{v} \in \mathcal{C}(A)} \left\| \vec{v} - \vec{b} \right\|_2^2 = \text{argmin}_{\vec{x}} \left( A \, \vec{x} - \vec{b} \right) \cdot \left( A \, \vec{x} - \vec{b} \right)$$

Es necesario entonces encontrar el vector que minimice la ecuación $\left(A\,\vec{x}-\vec{b}\right)^T\left(A\,\vec{x}-\vec{b}\right)$, la cual desarrollandola viene dada por:

$$\left(A\,\vec{x}-\vec{b}\right)^T\left(A\,\vec{x}-\vec{b}\right)=\left(\vec{x}^T\,A^T-\vec{b}^T\right)\left(A\,\vec{x}-\vec{b}\right)=\vec{x}^T\,A^T A\vec{x}-\vec{x}^T\,A^T\vec{b}-\vec{b}^T A\vec{x}+\vec{b}^T\vec{b} \tag{12}$$

Observe con atención los términos $\vec{x}^T\,A^T\vec{b}$ y $\vec{b}^T A\,\vec{x}$. El primer término corresponde a un producto de matrices con dimensiones $1\times n\,n\times m\,m\times 1$ lo cual resulta en un escalar, al igual que el segundo término asociado al producto $1\times m\,m\times n\,n\times 1$. Esto quiere decir que el tomar la transpuesta del primer término por ejemplo, el escalar sigue siendo el mismo, por lo que entonces $\left(\vec{x}^T\,A^T\vec{b}\right)^T=\vec{b}^T A\,\vec{x}$, puesto que $\left(\vec{x}^T\,A^T\vec{b}\right)^T=\vec{b}^T\left(\vec{x}^T\,A^T\right)^T=\vec{b}^T A\,\vec{x}$. Es por esto que la ecuación 12 se simplifica como sigue:

$$\left(A\,\vec{x}-\vec{b}=\right)^T\left(A\,\vec{x}-\vec{b}\right)=\vec{x}^T\,A^T A\,\vec{x}-2\,\vec{b}^T A\,\vec{x}+\vec{b}^T\vec{b}.$$

Para realizar la minimización:

$$f\left(\vec{x}\right)=\text{argmin}_{\vec{x}}\left(A\,\vec{x}-\vec{b}\right)\cdot\left(A\,\vec{x}-\vec{b}\right)$$

se calculará el gradiente de tal producto punto y se igualará a cero, para encontrar su punto mínimo. Recuerde que para toda función que resulta en un escalar $f\left(\vec{x}\right)$, calcular el vector gradiente:

$$\nabla_{\vec{x}}\left(A\,\vec{x}-\vec{b}\right)^T\left(A\,\vec{x}-\vec{b}\right)=0$$

$$\Rightarrow\nabla_{\vec{x}}\left(\vec{x}^T\,A^T A\,\vec{x}-2\,\vec{b}^T A\,\vec{x}+\vec{b}^T\vec{b}\right)=0$$

$$\Rightarrow\nabla_{\vec{x}}\left(\vec{x}^T\,A^T A\,\vec{x}\right)-\nabla_{\vec{x}}\left(2\,\vec{b}^T A\,\vec{x}\right)+\nabla_{\vec{x}}\left(\vec{b}^T\vec{b}\right)=0$$

- Observe que para el gradiente en el primer término se tiene que $\left(\vec{x}^T\,A^T A\,\vec{x}\right)=\left(\vec{x}^T\,K\,\vec{x}\right)$ con $K=A^T A$ una matriz cuadrada y simétrica, lo que corresponde a la forma cuadrática, para la cual ya se había demostrado que el gradiente viene dado por: $\nabla_{\vec{x}}\left(\vec{x}^T\,K\,\vec{x}\right)=2\,K\,\vec{x}$.

- Respecto al segundo término se puede reescribir como $\left(2\,\vec{b}^T A\,\vec{x}\right)=\left(2\,\vec{k}^T\,\vec{x}\right)$, pues observe que del producto $\vec{b}^T A$ resulta un vector con valores conocidos $\vec{b}^T A=\vec{k}^T\in\mathbb{R}^{1\times n}$ dado que los términos del producto se hacen con las dimensiones $1\times m\,m\times n$. Para una expresión similar, dejando fuera el escalar 2, ya demostramos que el gradiente viene dado por $\nabla_{\vec{x}}\left(\vec{k}^T\vec{x}\right)=\vec{k}$. **Por la regla del gradiente** $\nabla_{\vec{x}}\left(2\,\vec{b}^T A\,\vec{x}\right)=2A^T\vec{b}^T$.

- Finalmente, el tercer término corresponde a una constante, por lo que su gradiente es nulo, con lo que se arriba a:

$$\Rightarrow 2\,A^T A\,\vec{x} - \left(2\,\vec{b}^T A\right)^T = 0$$

y tomando la transpuesta del segundo término escalar:

$$\Rightarrow 2\,A^T A\,\vec{x} - 2\,A^T\,\vec{b} = 0$$

$$\Rightarrow A^T A\,\vec{x} = A^T\,\vec{b}$$

$$\Rightarrow \vec{x} = \left(A^T A\right)^{-1} A^T\,\vec{b}$$

$$\Rightarrow \vec{v} = A\left(A^T A\right)^{-1} A^T\,\vec{b}$$

donde recuerde que $\left(A^T A\right)^{-1} A^T = A^+$.

# 3. Probabilidades y la función Gaussiana

La teoría de la probabilidad es de suma importancia para muchos de los algoritmos en el aprendizaje automático y el procesamiento de señales, pues es habitual el lidiar con incertidumbre. El siguiente es un repaso de conceptos básicos de probabilidades. Existen dos ramas o enfoques, la teoría de probabilidades **frecuentista**, basada en frecuencia de ocurrencias de eventos, y la **clásica o axiomática**, como cociente de alternativas equiprobables (segun definicion de Kolgomorov).

## 3.1. Axiomas de la probabilidad

- **Conjunto de muestras** $\Omega$: Conjunto de todos los resultados posibles de un experimento. El **resultado** de un experimento puede conceptualizarse como una descripción completa de un estado del mundo real, al finalizar un experimento, y se denota como $\omega \in \Omega$.
- **Conjunto de eventos (o espacio de eventos)** $\mathcal{F}$: Un conjunto $A$ de posibles salidas (una o más salidas) $\omega$ de un experimento se le llama un **evento,** por lo que entonces $A \subseteq \Omega$**.** $\mathcal{F}$ es un espacio de eventos al que pertenecen uno o más eventos $A_i$, por lo que entonces $A_i \in \mathcal{F}$. El conjunto $\mathcal{F}$ satisface las siguientes propiedades:
  - El conjunto vacío siempre pertenece a $\mathcal{F}$: $\emptyset \in \mathcal{F}$.
  - $A_1, A_2, \ldots A_n \in \mathcal{F} \Rightarrow \cup A_i \in \mathcal{F}$.
- Propiedades básicas de la función de probabilidad: Una función de densidad de probabilidad $p : \mathcal{F} \to \mathbb{R}$:
  - $p\,(A) \geq 0, \forall A \in \mathcal{F}$

- $p(\Omega) = 1$
- Si $A_1, A_2, \ldots A_n$ son eventos disjuntos $A_i \cap A_j$, si $i \neq j$ entonces se tiene que:

$$p(\cup A_i) = \sum_i^n p(A_i)$$

**Ejemplo 1**

Defina el evento de tirar un dado de seis caras. El espacio de muestras en este caso viene dado por $\Omega = \{1, 2, 3, 4, 5, 6\}$ con $\omega_1 = 1, \omega_2 = 1, \ldots$ etc. El espacio de eventos más simple es $\mathcal{F} = \{\emptyset, \Omega\}$, para el cual se define $p(\emptyset) = 0$ y $p(\Omega) = 1$. Otro espacio de eventos posible $\mathcal{F}$ es el conjunto de todos los subconjuntos de $\Omega$. Para este último espacio de eventos, se puede aseginar la probabilidad de cada conjunto $A_i$ en tal espacio de eventos $\mathcal{F}$ como $\frac{k}{n}$ donde $k$ es la cantidad de elementos $|A_i|$ o cardinalidad y $n = |\Omega| = 6$. Por ejemplo $p(A_1 = \{1, 2, 3, 4\}) = \frac{4}{6}$.

**Ejemplo 2**

Imagine dos cajas, una roja y otra azul. En la caja roja existen 2 manzanas y 6 naranjas, mientras que en la azul existen 3 manzanas y una naranja, como se ilustra en la Figura 16. Se definen entonces dos espacios de muestras para dos tipos de eventos distintos: $\Omega_1 = \{r, a\}$ el cual se refiere a la escogencia de la caja azul o roja y $\Omega_2 = \{n, v\}$ el espacio que contiene los resultados experimentales de escoger una bola naranja o verde en cada una de las cajas. El espacio de eventos correspondiente a la escogencia de las cajas se define como $\mathcal{F}_1 = \{\emptyset, A_1 = \{r\}, A_2 = \{a\}\}$, con probabilidades $p(\emptyset) = 0$, $p(A_1) = 0{,}4$ y $p(A_2) = 0{,}6$. Observe que dado que $A_1 \cap A_2 = \emptyset$, entonces $p(\cup A_i) = A_1 + A_2 = 0{,}6 + 0{,}4 = 1$. Más adelante definiremos el espacio de eventos correspondiente a la escogencia de las pelotas, pues observe que ello depende de la caja escogida, asociado con una **probabilidad condicional**.

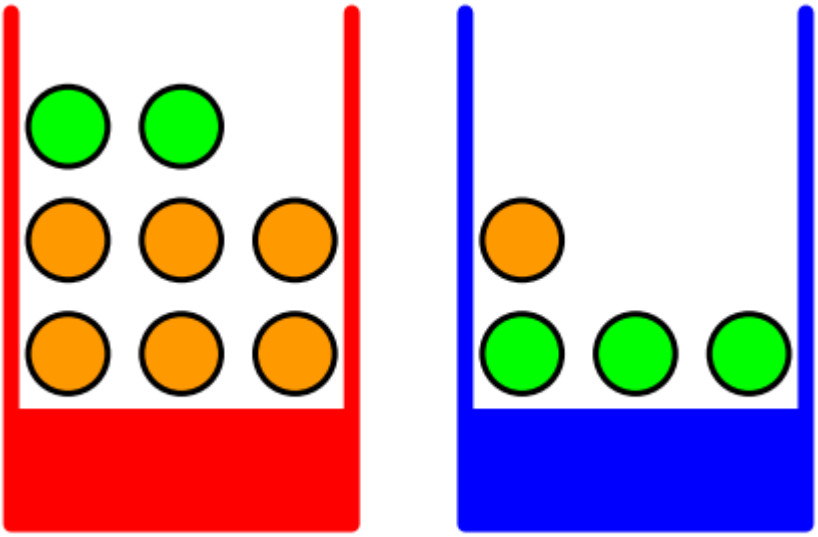


Figura 16: Cajas de naranjas y manzanas. Tomado de (Bishop, 2006)[1]

**Propiedades de los eventos**

- Si $A \subseteq B \Rightarrow p(A) \leq p(B)$
- Cota de la intersección $p(A \cap B) \leq \text{mín}(p(A) + p(B))$
- Cota de la unión $p(A \cup B) \leq p(A) + p(B)$
- Complemento $p(\Omega - A) = p(\Omega \backslash A) = 1 - p(A)$

## 3.2. Variables aleatorias

Considere el experimento de tirar una moneda 10 veces, con el objetivo de saber el número de veces que sale corona. En este caso el conjunto de muestras viene dado por $\Omega = \{c, e\}$. En este caso, el espacio de muestras $\mathcal{F}$ está dado por todas las secuencias posibles de escudos o coronas que salen al tirar la moneda. Sin embargo, en la práctica, no es necesario saber la probabilidad de obtener una secuencia particular de escudos o coronas. En cambio es más útil expresar lo anterior en términos de una función real que denote por ejemplo la cantidad que aparece "cara" después de 10 lanzamientos de la moneda. Tales funciones son conocidas como **variables aleatorias**.

Más formalmente, una variable aleatoria $X$ o $X(\omega)$ para un experimento $\omega$ es una función definida en un conjunto de muestras $\Omega$ $X : \Omega \rightarrow \mathbb{R}$, donde se usan letras minúsculas para los valores que la variable aleatoria puede tomar.

Una **variable aleatoria discreta** $X(\omega)$ es aquella que solo puede tomar un número finito de valores. El ejemplo de las 10 tiradas de la moneda es un caso

en el que se define una variable aleatoria discreta. Formalmente, la probabilidad de que una variable discreta tome un valor $k$ (en el caso de la moneda $c$ o $e$) viene dado por:

$$p\left(X = k\right) := p\left(\omega : X\left(\omega\right) = k\right)$$

Una variable aleatoria continua toma una infinita cantidad de número posible, por lo que usualmente se define la probabilidad de que la variable aleatoria tome un valor en el intervalo de $a \in \mathbb{R}$ a $b \in \mathbb{R}$:

$$p\left(a \leq X \leq b\right) := p\left(\{\omega : a \leq X\left(\omega\right) \leq b\}\right).$$

Un ejemplo de un fenómeno modelado con una variable aleatoria continua, es la probabilidad de que un sensor lumínico reciba una cantidad de lúmenes determinada en un rango $p\left(l_1 \leq X \leq l_2\right)$.

Una notación más resumida permite denotar $p(X)$ como el funcional de la probabilidad de densidad, y $p\left(X = x\right) = p(x)$ como la probabilidad de que se tome la muestra $x$.

**Ejemplo**

Siguiendo el ejemplo de las cajas de naranjas y manzanas ilustrado en la Figura 16, para los espacios $\Omega_1 = \{r, a\}$ y $\Omega_2 = \{n, v\}$ se definen, respectivamente, las variables aleatorias $C$ y $B$ . Así pues, se escriben las probabilidades como $p\left(C = r\right) = 0{,}4$ y $p\left(C = a\right) = 0{,}6$ para la variable aleatoria $C$.

## 3.3. Probabilidad conjunta y condicional

Considere dos variables aleatorias $X$ e $Y$, y para cada una de ellas se definen sus codominios $\Omega_X = \{x_1, x_2, \ldots x_M\}$ y $\Omega_Y = \{y_1, y_2, \ldots y_L\}$. Se realizan un total de $N$ muestras de ambas variables aleatorias donde el número de experimentos en los que la variable aleatoria $X = x_i$ y la variable otra variable aleatoria es $Y = y_j$ está definido por $n_{i,j}$, de modo que:

$$N = \sum_{i}^{M} \sum_{j}^{L} n_{i,j}.$$

Además, se define el número de experimentos en las que la variable aleatoria $X = x_i$ como

$$r_i = \sum_{j}^{L} n_{i,j},$$

y de manera similar, el número de experimentos en los que la variable aleatoria $Y = y_j$ como sigue:

$$c_j = \sum_{i}^{M} n_{i,j}.$$

Lo anterior se ilustra en la Figura 17, para un ejemplo en el que $M = 5$ y $L = 3$.

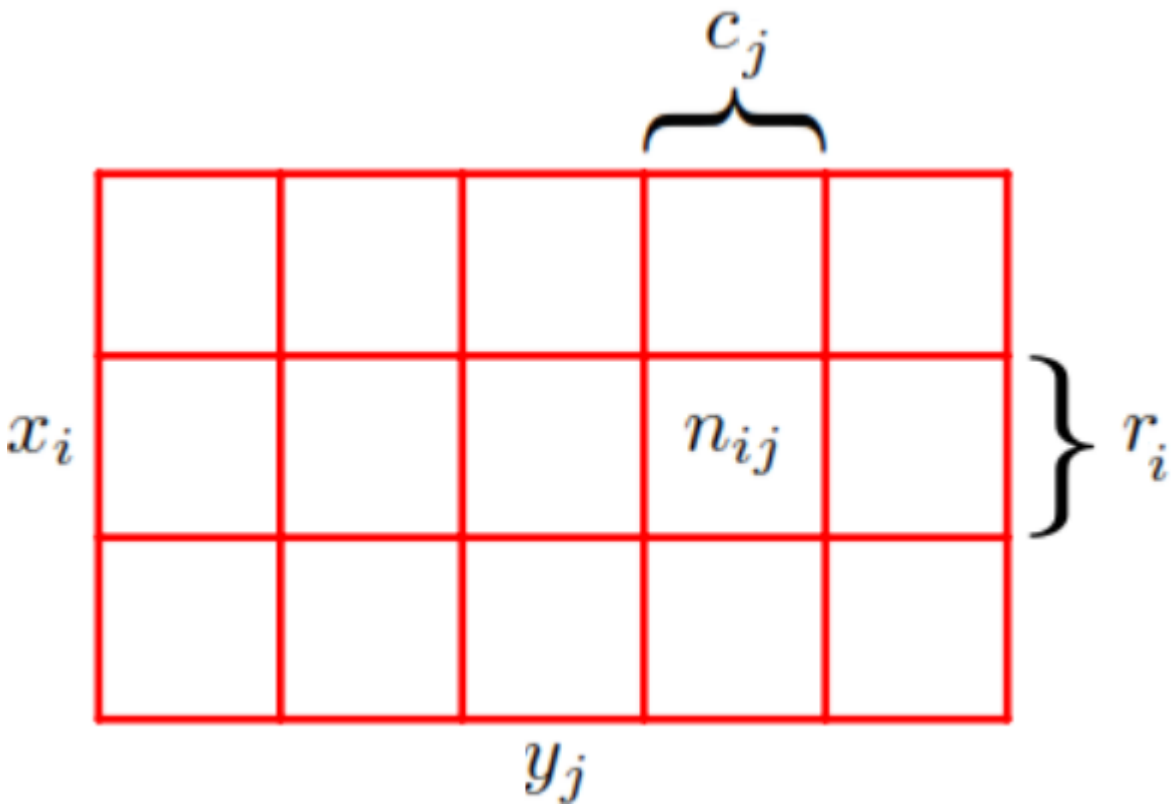



Figura 17: Ejemplo en el que $M = 5$ y $L = 3$, para la definición de la probabilidad conjunta y condicional. Tomado de (Bishop, 2006)[1].

**Probabilidad conjunta**

Desde un punto de vista probabilístico frecuentista en el que se aproxima una función de densidad de probabilidad, a medida que $N \to \infty$, se define la probabilidad conjunta como la probabilidad de que se tome la muestra $x_i$ y la muestra $y_j$, y está dada por la fracción de las muestras en la celda $i, j$ dividida por la cantidad total de muestras $N$ y se denota como:

$$p\left(Y = y_j, X = x_i\right) = p\left(X = x_i, Y = y_j\right) = \frac{n_{i,j}}{N}. \tag{13}$$

**Regla de la suma:** La probabilidad de que $X = x_i$ sin importar el valor de $Y$ viene dado por:

$$p\left(X = x_i\right) = p\left(X = x_i, Y = y_1, \ldots, y_L\right) = \sum_{j=1}^{L} p\left(X = x_i, Y = y_j\right) = \frac{c_i}{N}. \tag{14}$$

similar con la probabilidad de que $Y = y_j$ sin considerar la variable aleatoria $X$, $p\left(Y = y_j\right) = \frac{r_j}{N}$.

**Probabilidad condicional**

La probabilidad condicional se define como la probabilidad de escoger la muestra $Y = y_j$ dado que anteriormente se escogió la muestra $X = x_i$ (en otras palabras, se escogió la columna $i$) se define como:

$$p\left(Y = y_j | X = x_i\right) = \frac{n_{i,j}}{c_i}, \tag{15}$$

observe que la normalización se hace respecto a la cantidad de veces que se tomó la muestra $X = x_i$ sin importar $Y$.

**Regla del producto de probabilidad**: Tomando las ecuaciones 13, 14 y 15 se puede reescribir:

$$p\left(X = x_i, Y = y_j\right) = \frac{n_{i,j}}{N} = \frac{n_{i,j}}{c_i} \cdot \frac{c_i}{N} = p\left(Y = y_j | X = x_i\right) \cdot p\left(X = x_i\right)$$

En resumen se tienen dos reglas fundamentales de la teoría de la probabilidad:

- **Regla de la suma o probabilidad marginal**: $p\left(X\right) = \sum_Y p\left(X, Y\right)$.
- **Regla del producto**: $p\left(X, Y\right) = p\left(Y | X\right) \cdot p\left(X\right)$.
- **Conmutatividad de la probabilidad conjunta**: $p\left(X, Y\right) = p\left(Y, X\right)$. Junto con la regla del producto $p\left(X, Y\right) = p\left(Y, X\right) = p\left(X | Y\right) \cdot p\left(Y\right)$
- **Conjunción de la regla de la suma y del producto**:

$$p\left(X\right) = \sum_Y p\left(Y, X\right) \Rightarrow p\left(X\right) = \sum_Y p\left(X | Y\right) \cdot p\left(Y\right). \tag{16}$$

**Teorema de Bayes**

El teorema de Bayes es de vital importancia para muchos de los algoritmos y técnicas del aprendizaje automático. Para deducirlo, nos basamos en la regla del producto:

$$p(X,Y) = p(Y|X)\,p(X) \tag{17}$$

y dada la conmutatividad de la probabilidad conjunta, además de la regla del producto:

$$p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(Y,X)}{p(X)} = \frac{p(X|Y)\cdot p(Y)}{p(X)}$$

$$\Rightarrow p(Y|X) = \frac{p(X|Y)\cdot p(Y)}{p(X)}$$

En el aprendizaje automático, nos interesa estimar la probabilidad condicional:

$$p(T = t|M = \overrightarrow{m})$$

donde $t$ es una etiqueta de una clase especifica y $\overrightarrow{m}$ es la observacion para la cual queremos estimar la probabilidad de pertenecer a tal clase. Según Bayes, para estimar tal probabilidad, hacemos:

$$p(T = t|M = \overrightarrow{m}) = \frac{p(M = \overrightarrow{m}|T = t)\,p(T = t)}{p(M = \overrightarrow{m})}.$$

Donde las funciones de probabilidad de densidad $p(M = \overrightarrow{m}|T = t)$, $p(T = t)$ y $p(M = \overrightarrow{m})$ pueden estimarse fácilmente a partir del conjunto de datos.

La figura 18 muestra las funciones de densidad de probabilidad, graficadas en cada uno de los ejes, para el caso en el que $M = 9$ y $L = 2$.

- **Fórmula de Bayes**:

$$p(H|D) = \frac{p(D|H)\cdot p(H)}{p(D)}$$

- **Probabilidad a priori o marginal**: Se refiere a la probabilidad $p(H = h)$ de que una hipótesis $H = h$ se de antes de observar uno o más datos $D = d$ al que puede estar condicionado, por lo que se ignora cualquier dato que lo pueda condicionar.

- **Probabilidad a posteriori:** Corresponde a la probabilidad condicional $p(H = h|D = d)$ la cual describe la probabilidad de que la hipótesis $H = h$ se de, dado el suceso anterior del evento o los datos $D = d$.

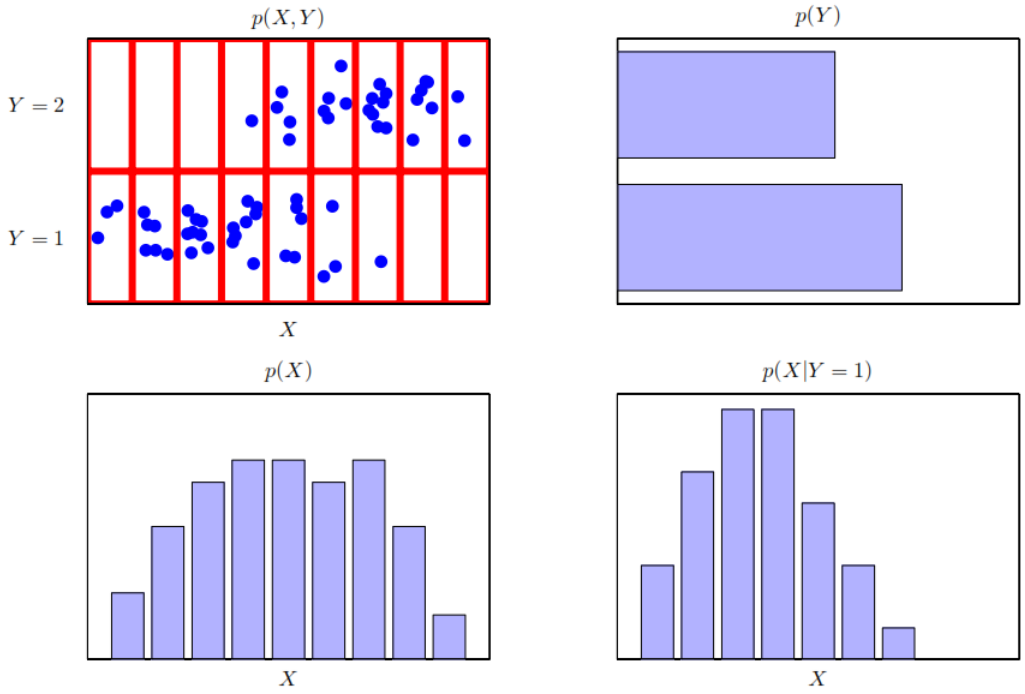



Figura 18: Graficación de $p\,(Y)$, $p\,(X)$ y $p\,(X|Y=1)$, con $N=60$. Tomado de (Bishop, 2006)[1].

**Ejemplo**

Retomando el ejemplo de las cajas de frutas, ya se había fijado que:

$$p\left(C=r\right)=0{,}4$$

$$p\left(C=a\right)=0{,}6$$

Para establecer las probabilidades condicionales, en este caso **no se usa un enfoque frecuentista**, puesto que no tenemos un historial de experimentos previos, en cambio se usa un **enfoque clásico o inferencial** en el que las probabilidades se calculan según características conocidas de antemano para el experimento (en este caso la cantidad de pelotas por caja). Es por esto que ponemos entonces inferir que:

$$\begin{gathered}p\left(B=v|C=r\right)=1/4\\ p\left(B=n|C=r\right)=3/4\\ p\left(B=v|C=a\right)=3/4\\ p\left(B=n|C=a\right)=1/4\end{gathered}$$

Suponga ahora que se desea conocer la probabilidad de obtener una pelota verde $p\left(B=v\right)$, sin importar la caja de la que viene, o la **probabilidad a priori** de que $B=v$. Se le llama **probabilidad a priori, pues ningún evento ha sucedido** (escogencia de la caja o de la pelota). Para ello se usa la regla de la ecuación 16:

$$p\left(B=v\right)=\sum_{C}p\left(B=v|C\right)\cdot p\left(C\right)=p\left(B=v|C=r\right)p\left(C=r\right)+p\left(B=v|C=a\right)p\left(C=a\right) \tag{18}$$

$$p\left(B=v\right)=\frac{1}{4}\times\frac{4}{10}+\frac{3}{4}\times\frac{6}{10}=\frac{11}{20} \tag{19}$$

Finalmente, se desea conocer la probabilidad de que la caja escogida sea roja, dado que se sacó una pelota naranja o la **probabilidad a posteriori** $p\left(C=r|B=n\right)$ (se le llama probabilidad a posteriori, pues se asocia con la probabilidad calculada **después de la experiencia** o evento de sacar una pelota naranja), con Bayes, ello viene dado por:

$$p\left(C=r|B=n\right)=\frac{p\left(B=n|C=r\right)p\left(C=r\right)}{p\left(B=n\right)}=\frac{3}{4}\times\frac{4}{10}\times\frac{20}{9}=\frac{2}{3},$$

y por el complemento podemos calcular $p\left(C=r|B=v\right)=1-2/3$.

**Independencia de variables aleatorias:**

Si dos variables aleatorias $X$ e $Y$ son independientes, la probabilidad conjunta de que $X=x$ y $Y=y$ se puede expresar como:

$$p\left(X,Y\right)=p\left(X\right)p\left(Y\right).$$

Por la regla del producto que establece que $p(X,Y) = p(Y|X) \cdot p(X)$, tenemos para las variables aleatorias $X$ e $Y$ independientes:

$$p(Y) = p(Y|X)$$

lo cual se puede leer como que la probabilidad de que $Y = y$ es la misma de que $Y = y$ suceda dado que anteriormente se dió $X = x$, es decir, es independiente.

En el ejemplo de las cajas con pelotas, si ambas cajas tienen la misma cantidad de pelotas naranja y verdes, por ejemplo en ambas cajas la mitad de las pelotas son naranjas y la mitad son verdes, tendríamos que:

$$\begin{aligned} p(B = v|C = r) &= 0{,}5 \\ p(B = n|C = r) &= 0{,}5 \\ p(B = v|C = a) &= 0{,}5 \\ p(B = n|C = a) &= 0{,}5 \end{aligned}$$

por lo que entonces por ejemplo, la probabilidad conjunta de que la caja sea roja y la pelota verde está dada por:

$$p(B = v, C = r) = p(B = v|C = r) \cdot p(C = r) = 0{,}5 \times 0{,}4 = 0{,}2 \qquad (20)$$

Para calcular las probabilidades marginales $p(B = v)$ y $p(B = n)$ se hace:

$$p(B = v) = p(B = v|C = r)\, p(C = r) + p(B = v|C = a)\, p(C = a) = 0{,}5 \times 0{,}4 + 0{,}5 \times 0{,}6 = 0{,}5$$

y $p(B = n) = 1 - p(B = v) = 0{,}5$. Es por ello que entonces, para verificar la independencia de la escogencia de la caja con el color de la pelota escogida se hace:

$$p(B = v)\, p(C = r) = 0{,}5 \times 0{,}4 = 0{,}2 = p(B = v, C = r)$$

con la última parte de la igualdad deducida en la ecuación 20.

## 3.4. Funciones de densidad de probabilidad y de distribución

Una función de densidad de probabilidad con variable continua $x$, $p_X(x) \equiv p(X = x)$ o más resumido como $p(x)$ se define de tal manera si para un intervalo muy pequeño $\delta x \to 0$, la probabilidad de que la variable aleatoria $x$ esté en un intervalo $(x, x + \delta x)$ es también infinitamente pequeña: $p(x)\, \delta x \to 0$.

Toda función de densidad de probabilidad debe cumplir las siguientes condiciones básicas:

$$0 \leq p(x) \leq 1$$

$$\int_{-\infty}^{\infty} p(x)\, \mathbf{d}x = 1.$$

La Figura 19 muestra algunas funciones de densidad conocidas.

Modelos de funciones de densidad populares:

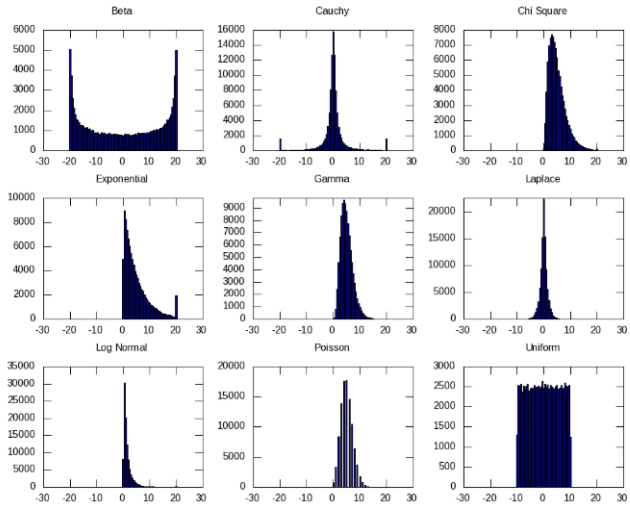



Figura 19: Funciones de densidad continuas.

- Modelo exponencial:

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{sino} \end{cases}$$

- Modelo Gaussiano:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{(x-\mu)^2}{2\sigma}\right)}$$

- Modelo de distribucion uniforme:

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sino} \end{cases}$$

**Función de distribución**

La probabilidad de que $x$ se encuentre en el intervalo $(-\infty, z)$ está dada por la función de distribución acumulativa, o función de distribución $P(z)$ definida como:

$$P_X(z) = P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x = p(x < z)$$

lo que implica que

$$P'(x) = p(x),$$

como se muestra en la Figura 20.

1. **Ejemplo**: Para la función de densidad uniforme

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sino} \end{cases}$$

su función de distribución está dada por:

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x = P(z) = \int_{a}^{z} \frac{1}{b-a}\mathrm{d}x$$

$$|_{a}^{z} \frac{x}{b-a} = \frac{z-a}{b-a}$$

por lo que entonces:

$$P(z) = \begin{cases} 0 & \text{si z<0} \\ \frac{z-a}{b-a} & \text{si a}\leq\text{z}\leq\text{b} \\ 1 & \text{si z>b} \end{cases}$$

Si la función de densidad $p(x)$ se define como discreta también referida como función de , de manera similar se tienen las siguientes propiedades:

$$0 \leq p[x] \leq 1$$

$$\sum_{x=0}^{\infty} p[x] = 1.$$

$$P[z] = \sum_{x=0}^{z} p[x]$$

Las siguientes son propiedades de la función de distribución $P(z)$:

- $0 \leq P(z) \leq 1.$
- $\lim_{z\to\infty} P(z) = 1$
- $\lim_{z\to-\infty} P(z) = 0$

La función de densidad de probabilidad con múltiples variables $x_1, \ldots x_D$ se denota de forma compacta con el vector $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix} = (x_1, \ldots x_D)$ como la función de densidad de probabilidad conjunta $p(\vec{x}) = p(x_1, \ldots x_D)$, por lo que la probablidad de que $\vec{x}$ se encuentre en un volumen infinitesimal $\delta\vec{x}$ está dado por $p(\vec{x})\,\delta\vec{x}$ y cumple también las dos propiedades básicas:

$$p(\vec{x}) \geq 0$$

$$\int_{-\infty}^{\infty} p(\vec{x})\,\mathrm{d}\vec{x} = 1.$$

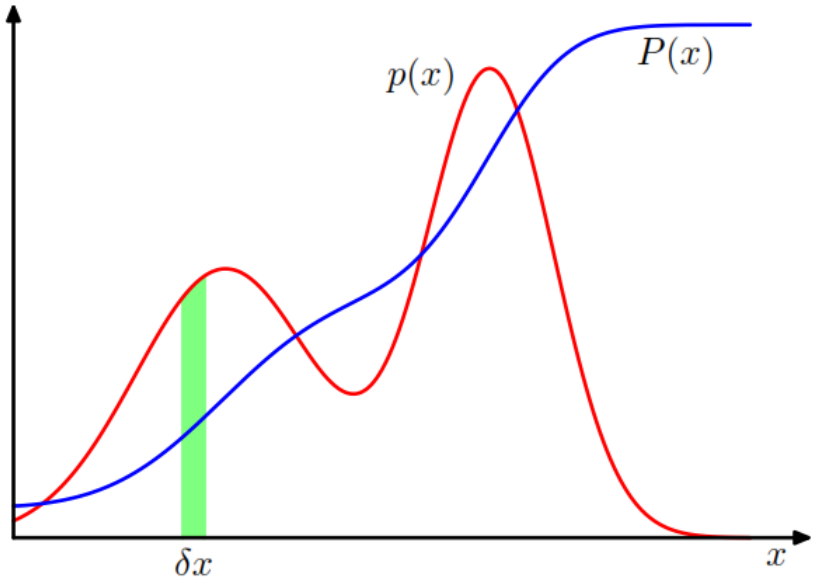



Figura 20: Gráfica de las funciones de densidad y de distribución. Tomado de [1].

**El teorema de Bayes para funciones de densidad**

Las propiedades de suma, del producto y el teorema de Bayes, aplican también a funciones de densidad de probabilidad, por lo que con $x, y \in \mathbb{R}$ se tiene que la probabilidad marginal, conjunta y condicional, respectivamente, están dadas por:

$$p(x) = \int p(x, y)\, \mathrm{d}y$$

$$p(x, y) = p(y|x)\ p(x)$$

$$p(x|y) = \frac{p(y|x)\ p(x)}{p(y)}$$

### 3.4.1. Descriptores de funciones de densidad de probabilidad

**Entropía** La entropía mide el grado de «desorden» en un conjunto de datos, lo cual, según la teoría de la información de Claude Shannon (1948), se corresponde a la cantidad de información en ese conjunto de datos (mucho orden, poca información). Para una función de densidad $p(x)$ se calcula como sigue:

$$H_X = -\sum_{z=0}^{L} p(x) \log_2 (p(x)) .$$

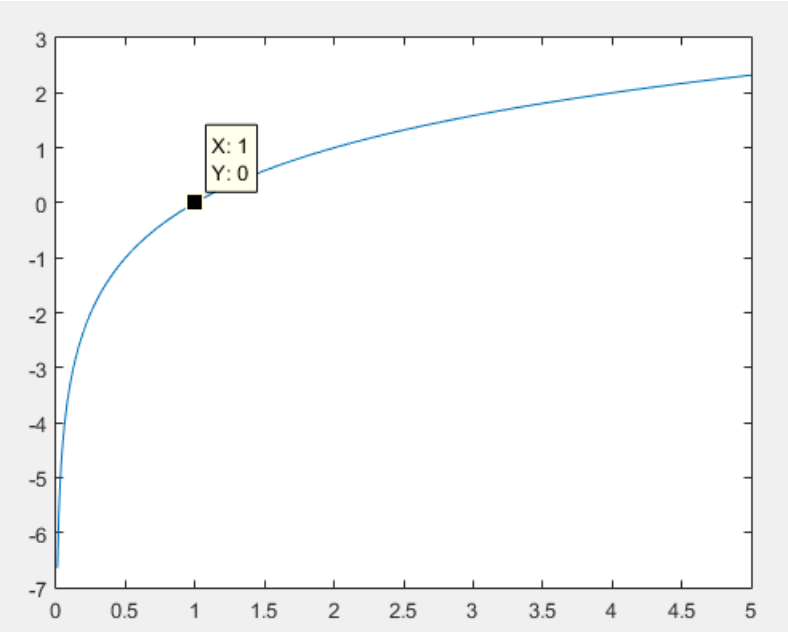



Figura 21: Función logarítmica.

La entropía usa el logaritmo, función monotonicamente creciente, que transforma los valores entre 0 y 1 en valores más grandes. Conforme más pequeño $p(x)$, mucho más grande su logaritmo. De esta forma, la entropía $H_X \rightarrow 0$, cuando hay mucho «orden en los datos», es decir, hay pocos valores distintos, y por ende, con alta probabilidad. Por ejemplo, en el caso de que solo exista un solo valor distinto $x_a$ en todo el conjunto de datos, $\log_2(p(x_a) = 1) = 0$. La Figura 21 muestra la gráfica de la función logarítmica.

**Esperanza, varianza y covarianza de una función de densidad de probabilidad** El cálculo de momentos estadísticos es una operación muy utilizada en el reconocimiento de patrones, pues permite reducir la dimensionalidad de datos, recabando características importantes como el valor esperado o la varianza.

Se define entonces la **esperanza o el valor esperado** de una variable aleatoria $X$ como la sumatoria pesada por la probabilidad de cada valor $x$ que puede tomar tal variable aleatoria (a la izquierda en el caso de ser una variable aleatoria continua, a la derecha discreta):

$$\mu_X = \mathbb{E}[X] = \int x\, p(x)\, \mathrm{d}x \qquad \mu_X = \mathbb{E}[X] = \sum_x x\, p[x]$$

En el caso de una función discreta $h[u]$ la cual acumula el valor $x$ generado registrado para el experimento $u$ para un total de $N$ experimentos, generados

a partir de la variable aleatoria $X$, la esperanza está dada por:

$$\mathbb{E}[X] \cong \frac{1}{N}\sum_{u=1}^{N} h[u]. \tag{21}$$

Desde un punto de vista frecuentista, la aproximación de $\mathbb{E}[X]$ mejora, a medida que $N \to \infty$.

Las siguientes son propiedades de la esperanza:

- Si $a$ es un escalar, tal que $a \in \mathbb{R}$, se tiene que: $\mathbb{E}[a] = a$.
- Homogeneidad: $\mathbb{E}[a\,X] = a\,\mathbb{E}[X]$.
- Superposición: $\mathbb{E}[g(X) + f(X)] = \mathbb{E}[g(X)] + \mathbb{E}[f(X)]$.

**Varianza**

La varianza de una variable aleatoria $X$ es una medida de la "concentración" o dispersión de los datos alrededor de la media. Formalmente la varianza de una variable aleatoria $X$ se define como:

$$\sigma_X^2 = \text{var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right],$$

lo cual equivale a (por la linealidad de la esperanza y su propiedad de equivalencia en constantes):

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[\left(X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right)\right]$$

$$\Rightarrow \text{var}[X] = \left(\mathbb{E}\left[X^2\right] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}\left[\mathbb{E}[X]^2\right]\right)$$

$$\Rightarrow \text{var}[X] = \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2$$

$$\Rightarrow \text{var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

Propiedades de la varianza:

- Si $a$ es un escalar, tal que $a \in \mathbb{R}$, se tiene que: $\text{Var}[a] = 0$.
- Multiplicación por escalar de la entrada: $\text{Var}[a\,X] = a^2\text{Var}[X]$, $\forall a \in \mathbb{R}$.

En el caso de una función discreta $h[u]$ la cual acumula el valor $x$ generado registrado para el experimento $u$ para un total de $N$ experimentos, generados a partir de la variable aleatoria $X$, la varianza está dada por:

$$\sigma_X^2 \cong \frac{1}{N-1}\sum_{u=1}^{N} (h[u] - \mu_X)^2. \tag{22}$$

El estimador anterior se dice que no es "sesgado" si la media real no es conocida. Si la media real de la población es conocida, se normaliza por $N$ ( puede ver

la demostración de lo anterior en `http://www.visiondummy.com/2014/03/divide-variance-n-1/`).

**Ejemplo**

Calcule la media y la varianza de una variable aleatoria uniforme $X$ con una función de densidad $p(x) = 1,\ \forall x \in [0,1],\ 0$ de otro modo:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, p(x)\ \mathrm{d}x = \int_{0}^{1} x\, \mathrm{d}x = \frac{1}{2}$$

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2\, p(x)\ \mathrm{d}x = \int_{0}^{1} x^2\, \mathrm{d}x = \frac{1}{3}$$

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

**Covarianza**

Para dos variables aleatorias $X$ e $Y$ se la covarianza como:

$$\Sigma_{X,Y} = \mathrm{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

y mide la variación conjunta de tales variables aleatorias. Para el caso de contar con arreglos de muestras $h[u]$ y $g[u]$ para las variables aleatorias $X$ e $Y$ respectivamente, se tiene que la covarianza de tales variables aleatorias está dada por:

$$\Sigma_{X,Y} \cong \frac{1}{N-1} \sum_{u=1}^{N} (h[u] - \mu_X)(g[u] - \mu_Y).$$

La matriz de covarianza para $n$ variables aleatorias $X_1, X_2, \ldots, X_n$ se define como:

$$\Sigma = \begin{bmatrix} \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_1 - \mathbb{E}[X_1])] & \ldots & \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_n - \mathbb{E}[X_n])] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n - \mathbb{E}[X_n])(X_1 - \mathbb{E}[X_1])] & \ldots & \mathbb{E}[(X_n - \mathbb{E}[X_n])(X_n - \mathbb{E}[X_n])] \end{bmatrix},$$

observe que en la diagonal de la matriz $\Sigma$ (entrada $\Sigma_{i,i}$) se tiene que

$$\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_i - \mathbb{E}[X_i])] = \sigma_{X_i}^2,$$

por lo que entonces la matriz de covarianza se puede reescribir como:

$$\Sigma = \begin{bmatrix} \sigma_{X_1}^2 & \ldots & \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_n - \mathbb{E}[X_n])] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n - \mathbb{E}[X_n])(X_1 - \mathbb{E}[X_1])] & \ldots & \sigma_{X_n}^2 \end{bmatrix}.$$

Además, la matriz de covarianza $\Sigma$ presenta la propiedad de ser simétrica, puesto que $\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbb{E}[(X_j - \mathbb{E}[X_j])(X_i - \mathbb{E}[X_i])] \Rightarrow \Sigma_{X_i,X_j} = \Sigma_{X_j,X_i}$.

**Ejemplo**

Suponga que se desea encontrar la matriz de covarianza para tres variables aleatorias $X_1, X_2$ y $X_3$, para las cuales se han recabado los siguientes arreglos de muestras para $N = 4$ experimentos, respectivamente:

$$\begin{aligned} h_1 &= \begin{bmatrix} 2 & 4 & 6 & 8 \end{bmatrix} \\ h_2 &= \begin{bmatrix} 4 & 8 & 12 & 16 \end{bmatrix} \\ h_3 &= \begin{bmatrix} 12 & 10 & 5 & 9 \end{bmatrix} \end{aligned}$$

En términos de muestras se tienen 4 muestras

$$U = \{\vec{u}_1, \vec{u}_2, \vec{u}_3, \vec{u}_4\} = \begin{bmatrix} | & | & | & | \\ \vec{u}_1 & \vec{u}_2 & \vec{u}_3 & \vec{u}_4 \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 & 8 \\ 4 & 8 & 12 & 16 \\ 12 & 10 & 5 & 9 \end{bmatrix}$$

con $u_i \in \mathbb{R}^3$, donde cada dimensión es una variable aleatoria, y $U \in \mathbb{R}^{3\times 4}$.

Observe en estos datos, que la dimensión 1 y 2 son combinación lineal para todas las muestras, por lo que la covarianza de ambas dimensiones debe ser alta, no así la dimensión 1 con la 3 o la 2 con la 3. Además

Se procede entonces a calcular las entradas $\Sigma_{X_1,X_2}$, $\Sigma_{X_1,X_3}$ y $\Sigma_{X_2,X_3}$, además de los valores de la diagonal $\sigma^2_{X_1}$, $\sigma^2_{X_2}$ y $\sigma^2_{X_3}$, teniendo en cuenta que $\mu_{X_1} = 5$, $\mu_{X_2} = 10$ y $\mu_{X_3} = 9$:

$$\begin{aligned} \Sigma_{X_1,X_2} &= \tfrac{1}{4-1}\left((5-2)(10-4) + (5-4)(10-8) + (5-6)(10-12) + (5-8)(10-16)\right) \\ \Sigma_{X_1,X_3} &= \tfrac{1}{4-1}\left((5-2)(9-12) + (5-4)(9-10) + (5-6)(9-5) + (5-8)(9-9)\right) \\ \Sigma_{X_2,X_3} &= \tfrac{1}{4-1}\left((10-4)(9-12) + (10-8)(9-10) + (10-12)(9-5) + (10-16)(9-9)\right) \\ \sigma^2_{X_1} &= \tfrac{1}{4-1}\left((5-2)^2 + (5-4)^2 + (5-6)^2 + (5-8)^2\right) \\ \sigma^2_{X_2} &= \tfrac{1}{4-1}\left((10-4)^2 + (10-8)^2 + (10-12)^2 + (10-16)^2\right) \\ \sigma^2_{X_3} &= \tfrac{1}{4-1}\left((9-12)^2 + (9-10)^2 + (9-5)^2 + (9-9)^2\right) \end{aligned}$$

lo cual desarrollado corresponde a:

$$\begin{aligned} \Sigma_{X_1,X_2} &= \tfrac{1}{3}(3\cdot 6 + 1\cdot 2 + -1\cdot -2 + -3\cdot -6) = \tfrac{40}{3} = 13{,}333 \\ \Sigma_{X_1,X_3} &= \tfrac{1}{3}(3\cdot -3 + 1\cdot -1 + -1\cdot 4 + -3\cdot 0) = -\tfrac{14}{3} = -4{,}667 \\ \Sigma_{X_2,X_3} &= \tfrac{1}{3}(6\cdot -3 + 2\cdot -1 + -2\cdot 4 + -6\cdot 0) = -\tfrac{28}{3} = -9{,}333 \\ \sigma^2_{X_1} &= \tfrac{1}{3}(9 + 1 + 1 + 9) = \tfrac{20}{3} = 6{,}667 \\ \sigma^2_{X_2} &= \tfrac{1}{3}(36 + 4 + 4 + 36) = \tfrac{80}{3} = 26{,}667 \\ \sigma^2_{X_3} &= \tfrac{1}{3}(9 + 1 + 16 + 0) = \tfrac{14}{3} = 8{,}667. \end{aligned}$$

Por lo que se obtiene la matriz de covarianza:

$$\Sigma = \begin{bmatrix} \frac{20}{3} & \frac{40}{3} & -\frac{14}{3} \\ \frac{40}{3} & \frac{80}{3} & -\frac{28}{3} \\ -\frac{14}{3} & -\frac{28}{3} & \frac{14}{3} \end{bmatrix} = \begin{bmatrix} 6{,}667 & 13{,}333 & -4{,}667 \\ 13{,}333 & 26{,}666 & -9{,}333 \\ -4{,}667 & -9{,}333 & 8{,}667 \end{bmatrix}.$$

Observe que la covarianza más alta es $\Sigma_{X_1,X_2}$ pues las dimensiones 1 y 2, cuyos datos son generados por las variables aleatorias $X_1$ y $X_2$, pues los datos

siempre covarían positivamente para todas las muestras, en ambas dimensiones. En los otros dos casos, donde la covarianza es negativa, ello denota que cuando en una dimensión los datos varían en una dirección, en la otra los datos varían en dirección contraria.

En MATLAB lo anterior se puede implementar como sigue:

```
U = [2 4 6 8; 4 8 12 16; 12 10 5 9];
%Cada columna son los datos de una dimension
matCov = cov(U');
```

También la matriz de covarianza se puede escribir, para un conjunto de muestras $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m\}$, con $\vec{x}_i \in \mathbb{R}^{n\times 1}$, como:

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu})^T$$

donde $\vec{\mu} \in \mathbb{R}^{n\times 1}$ es la muestra promedio del conjunto de datos $X$ (donde cada componente es el valor medio de cada dimensión).

### Coeficiente de correlación de Pearson

En el ejemplo anterior, se concluyó que las variables aleatorias $X_1$ y $X_2$ covarían fuertemente, lo que sugiere una alta correlación entre ambas variables aleatorias. La matriz de Pearson denotada como $\rho$, permite observar el grado de "correlación", normalizando la covarianza, de modo que el coeficiente cumpla que $-1 \leq \rho \leq 1$. Para dos variables aleatorias $X_i$ y $X_j$, el coeficiente de Pearson normaliza la covarianza usando la desviación estandar de tales variables aleatorias:

$$\rho_{X_i,X_j} = \frac{\text{cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

definiendo así la matriz de correlación de Pearson:

$$\Sigma = \begin{bmatrix} \frac{\mathbb{E}[(X_1-\mathbb{E}[X_1])(X_1-\mathbb{E}[X_1])]}{\sigma_{X_1}\sigma_{X_1}} & \cdots & \frac{\mathbb{E}[(X_1-\mathbb{E}[X_1])(X_n-\mathbb{E}[X_n])]}{\sigma_{X_1}\sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\mathbb{E}[(X_n-\mathbb{E}[X_n])(X_1-\mathbb{E}[X_1])]}{\sigma_{X_n}\sigma_{X_1}} & \cdots & \frac{\mathbb{E}[(X_n-\mathbb{E}[X_n])(X_1-\mathbb{E}[X_1])]}{\sigma_{X_n}\sigma_{X_n}} \cdots \end{bmatrix},$$

donde los valores de la diagonal $\rho_{X_i,X_j} = \frac{\mathbb{E}[(X_i-\mathbb{E}[X_i])(X_j-\mathbb{E}[X_j])]}{\sigma_{X_i}\sigma_{X_j}} = 1$.

La Figura 22, muestra como se comporta el coeficiente de Pearson ante distintas condiciones de dos variables aleatorias $X_1$ y $X_2$ para los cuales se grafican los datos en los arreglos $h_1$ y $h_2$.

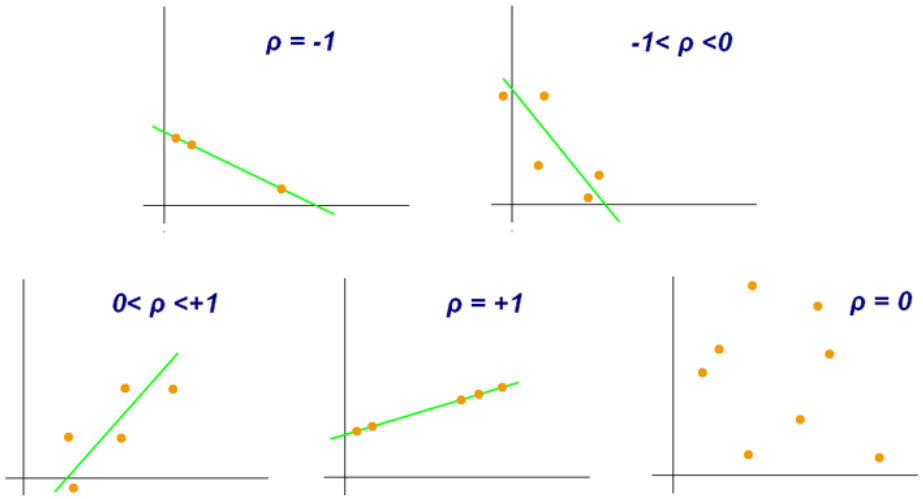



Figura 22: Diagramas de dispersión y coeficiente de Pearson para distintos casos. Observe que el coeficiente se aleja de cero, en cuanto mayor correlación lineal exista en los datos.

**La función de densidad Gaussiana**

La función de densidad Gaussiana es la función de densidad más utilizada en el análisis de datos y el reconocimiento de patrones, pues muchos fenómenos aleatorios naturales (como por ejemplo el peso de las personas en una cierta edad, características físicas en animales y plantas, etc.) se modelan de forma satisfactoria con una función de densidad Gaussiana.

La función de densidad Gaussiana es un modelo con dos parámetros: la media $\mu$ y la dispersión $\sigma$. Para el caso de una dimensión en la que el codominio está dado por $x \in \mathbb{R}$, se tiene que una función de densidad de probabilidad está dada por:

$$p\left(x\right) = \mathcal{N}\left(x|\mu,\sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right),$$

lo cual expresa la probabilidad de que el valor $X = x$ haya sido generado por un modelo Gaussiano con parámetros $\theta = (\mu, \sigma)$. El coeficiente $\frac{1}{\sqrt{2\pi\sigma^2}}$ normaliza la función de densidad y garantiza el área bajo la curva sea de 1:

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma\right) \mathrm{d}x = 1$$

y además:

$$\mathcal{N}\left(x|\mu,\sigma\right) > 0.$$

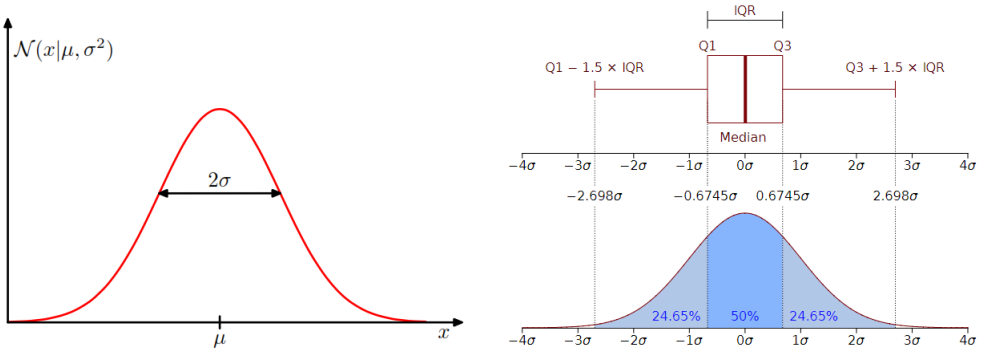



Figura 23: Función de densidad Gaussiana, *box plot*, tomado de [1].

La esperanza de una variable aleatoria $X$ caracterizada por una distribución normal $\mathcal{N}(x|\mu,\sigma)$ viene dada por la media, puesto que:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \mathrm{d}x = \mu,$$

y la esperanza de la variable aleatoria al cuadrado está dada por:

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \mathrm{d}x = \mu^2 + \sigma^2,$$

por lo cual se deduce que:

$$\mathrm{var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \sigma^2.$$

El máximo de la función de densidad de probabilidad Gaussiana, también referido como la **moda**, coincide con la media, como ilustra la Figura 23.

### Función de verosimilitud

Suponga ahora que se cuenta con un arreglo $\vec{h} = [h_1, \ldots, h_M]$ de $M$ observaciones discretas sobre el dominio de valores que puede tomar la variable aleatoria $X$. Se supone además que tales observaciones son generadas independientemente, con una misma distribución Gaussiana de parámetros $\mu$ y $\sigma$. Para denotar la probabilidad conjunta de que todo el arreglo $\vec{h}$ haya sido generada por una variable aleatoria Gaussiana con tales parámetros, se escribe la probabilidad conjunta $p\left(\vec{h}|\mu,\sigma\right) = p\left(h_1, h_2, \ldots, h_m|\mu,\sigma\right)$, la cual, como vimos anteriormente, en el caso de que los datos sean independientes, corresponde a la multiplicación de las probabilidades marginales, a lo que se le conoce como la **función de verosimilitud**:

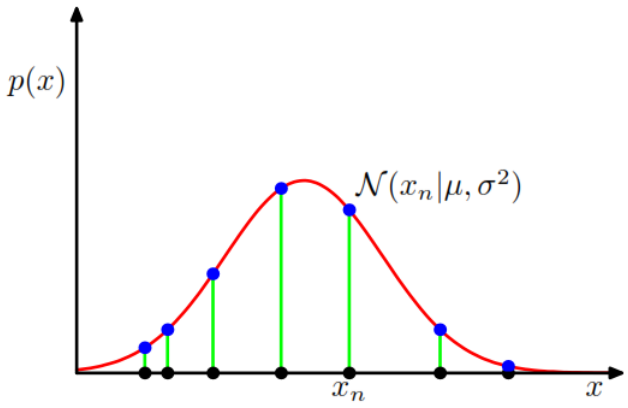



Figura 24: Maximizar la verosimilitud corresponde a ajustar la media y la desviación estándar de modo que la multiplicatoria de la evaulación en la función Gaussiana se maximice según los puntos en el arreglo $\vec{h}$. Tomado [1].

$$p\left(\vec{h}|\mu,\sigma\right) = \prod_{n=1}^{M} \mathcal{N}\left(h_n|\mu,\sigma\right) \tag{23}$$

Dado que lo único que conocemos son los datos en el arreglo $\vec{h}$, es necesario encontrar los parámetros $\mu$ y $\sigma$ que maximicen la función de verosimilitud $p\left(\vec{h}|\mu,\sigma\right)$, por lo que entonces, para cada todos los $h_1,\ldots h_M$ puntos la multiplicatoria debe ser lo máximo posible, lo cual corresponde a lo ilustrado en la Figura 24.

Para facilitar la maximización de la función de verosimilitud se **utiliza el logaritmo natural**, una función monotónicamente creciente como se muestra en la Figura 25, por lo que no altera el sentido de la maximización. El logaritmo natural es usual en cálculos que involucran probabilidades, pues evita el problema del "underflow" que resulta de calcular el producto de muchos números de magnitud menor que uno. Las siguientes son propiedades del logaritmo natural:

1. $\ln\left(x\cdot y\right) = \ln\left(x\right) + \ln\left(y\right)$
2. $\ln\left(e\right) = 1$
3. $\ln\left(x^n\right) = n\,\ln\left(x\right)$
4. $\ln\left(\frac{x}{y}\right) = \ln\left(x\right) - \ln\left(y\right)$
5. $\ln\left(1\right) = 0$
6. $\ln\left(-1\right) = i\pi$

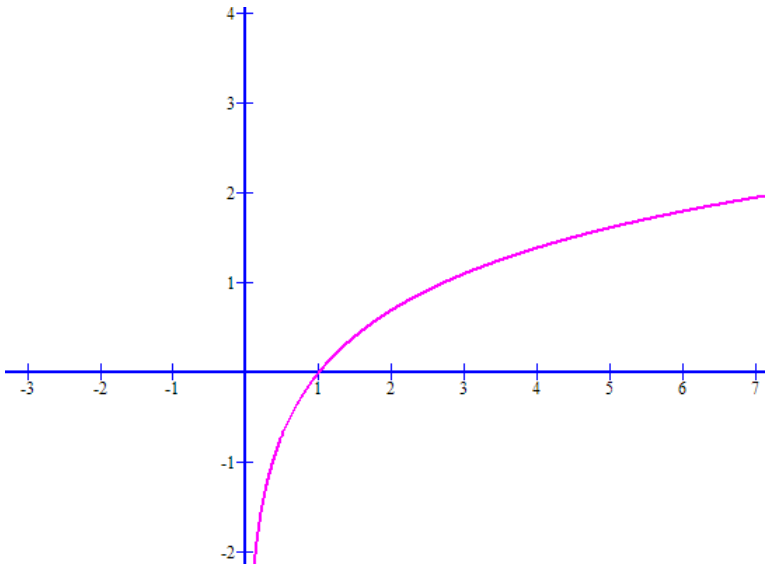



Figura 25: Función logaritmica, con asíntota vertical en el eje $y$, y monotónicamente creciente.

7. $\ln(x) < \ln(y), \forall 0 < x < y$
8. $\frac{d}{dx}\ln(x) = \frac{1}{x}$

Calculando entonces el logaritmo natural de la función de verosimilitud y usando sus propiedades se obtiene la siguiente expresión simplificada:

$$\ln\left(p\left(\vec{h}|\mu,\sigma\right)\right) = \ln\left(\prod_{n=1}^{M}\mathcal{N}\left(h_n|\mu,\sigma\right)\right)$$

$$\Rightarrow \ln\left(p\left(\vec{h}|\mu,\sigma\right)\right) = \sum_{n=1}^{M}\ln\left(\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(h_n-\mu\right)^2\right)\right)$$

$$\Rightarrow \ln\left(p\left(\vec{h}|\mu,\sigma\right)\right) = M\,\ln\left(\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\right) + \sum_{n=1}^{M}\ln\left(\exp\left(-\frac{1}{2\sigma^2}\left(h_n-\mu\right)^2\right)\right)$$

$$\Rightarrow \ln\left(p\left(\vec{h}|\mu,\sigma\right)\right) = -\frac{M}{2}\ln\left(2\pi\sigma^2\right) + -\frac{1}{2\sigma^2}\sum_{n=1}^{M}\left(h_n-\mu\right)^2$$

$$\Rightarrow \ln\left(p\left(\vec{h}|\mu,\sigma\right)\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{M}\left(h_n-\mu\right)^2 - \frac{M}{2}\ln\left(2\pi\right) - M\,\ln\left(\sigma\right).$$

Para obtener los valores de $\mu$ y $\sigma$ que maximicen al logaritmo de la función de verosimilitud, derivamos respecto a $\mu$ y $\sigma$ respectivamente:

$$\frac{d}{d\mu} \ln\left(p\left(\vec{h}|\mu,\sigma\right)\right) = 0$$

$$\Rightarrow \frac{1}{\sigma^2} \sum_{n=1}^{M} (h_n - \mu) - \frac{M}{2} \ln(2\pi) - M \ln(\sigma) = 0$$

$$\Rightarrow \mu = \frac{1}{M} \left( \sum_{n=1}^{M} h_n \right),$$

lo cual coinicide con la fórmula de la esperanza planteada en la ecuación 21.

Ahora, derivando respecto a la desviación estándar e igualando a cero:

$$\frac{d}{d\sigma} \ln\left(p\left(\vec{h}|\mu,\sigma\right)\right) = 0$$

$$\Rightarrow \frac{1}{\sigma^3} \sum_{n=1}^{M} (h_n - \mu)^2 - \frac{M}{2} \ln(2\pi) - M \frac{1}{\sigma} = 0$$

$$\Rightarrow \frac{1}{M} \sum_{n=1}^{M} (h_n - \mu)^2 = \sigma^2,$$

lo cual es distinto a lo planteado en la ecuación 22 (normalizado con $\frac{1}{M-1}$ ) se obtiene lo que se conoce como un estimador de la varianza *sesgado*, el cual se *sobre-ajusta* a los datos (confía mucho en los datos), por lo que con pocas muestras se recomienda usar el estimador sin sesgo de la ecuación 22.

**Función Gaussiana multivariable**

Muchas veces será necesario lidiar con funciones de densidad de probabilidad definidas en un espacio de dimensión $D$, por lo que según lo definido anteriormente, se denota la probabilidad de encontrar $D$ valores de forma compacta con $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix} = (x_1, \dots x_D)$ por lo que $\vec{x} \in \mathbb{R}^D$ y para la función Gaussiana está dada por:

$$p\left(\vec{x}\right) = \mathcal{N}\left(\vec{x}|\mu,\Sigma\right) = \frac{1}{\left(2\pi\right)^{1/D}} \frac{1}{\det\left(\Sigma\right)^{1/2}} \exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \Sigma^{-1} \left(\vec{x}-\vec{\mu}\right)\right), \tag{24}$$

donde:

- El determinante $\det(\Sigma)$ corresponde al volumen del "cuerpo" definido por la matriz de covarianza, y es un escalar parte del coeficiente de normalización.

- El exponente $-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T\Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)$ es una forma cuadrática (escalar), con el vector $\vec{\mu}$ definido como el valor medio en cada dimensión $\vec{\mu} = (\mu_1, \ldots \mu_D)$.
- La **matriz de covarianza** $\Sigma$ debe ser positiviamente definida, es decir, su forma cuadratica para cualquier vector no nulo es positiva: $\overrightarrow{x}^T\Sigma\overrightarrow{x} > 0$.

Para entender mejor el significado de la función Gaussiana multivariable, considere un caso simple en que la función Gausiana está dada por $f\left(\vec{x}\right) = \mathcal{N}\left(\vec{x}|\mu,\Sigma\right)$, con $f : \mathbb{R}^2 \to \mathbb{R}$, y una matriz de covarianza diagonal $\Sigma$, por lo que entonces:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} ,$$

por ello, según la expresión general multi-variable de la función Gaussiana, se tiene para el exponente en forma cuadrática:

$$-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T\Sigma^{-1}\left(\vec{x}-\vec{\mu}\right) = -\frac{1}{2}\left(\begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix}\right)\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1}\left(\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right)$$

Observe que para cualquier matriz diagonal $U = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix}$, su inversa viene dada por: $U^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\lambda_n} \end{bmatrix}$:

$$= -\frac{1}{2}\begin{bmatrix} \frac{1}{\sigma_1^2}\left(x_1-\mu_1\right) & \frac{1}{\sigma_2^2}\left(x_2-\mu_2\right) \end{bmatrix}\left(\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right) = -\frac{1}{2}\left(\frac{1}{\sigma_1^2}\left(x_1-\mu_1\right)^2 + \frac{1}{\sigma_2^2}\left(x_2-\mu_2\right)^2\right).$$

Respecto al coeficiente de normalización se tiene que:

$$\det\left(\Sigma\right)^{1/2} = \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{1/2} = \left(\sigma_1^2\sigma_2^2 - 0*0\right)^{1/2} = \sigma_1\sigma_2$$

por lo que entonces tal coeficiente de normalización en este caso viene dado por $\frac{1}{\sqrt{2\pi}\sigma_1\sigma_2}$, con ello:

$$\frac{1}{\sqrt{2\pi}\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_1^2}\left(x_1-\mu_1\right)^2 + \frac{1}{\sigma_2^2}\left(x_2-\mu_2\right)^2\right)\right)$$

El siguiente código de MATLAB permite graficar una función Gaussiana con una matriz de covarianza $\Sigma$ definida:

```
mu = [0  0];
Sigma = [0.8  -0.3; -0.3  0.5];
```

```
x1 = -3:.2:3;
x2 = -3:.2:3; [X1,X2] = meshgrid(x1,x2);
%Se vectorizan los valores en la matriz
F = mvnpdf([X1(:) X2(:)],mu,Sigma);
%Se redimensionan de nuevos a la matriz
F = reshape(F,length(x2),length(x1));
surf(x1,x2,F);
xlabel('x1');
ylabel('x2');
zlabel('Probability Density');
```

## References


[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] Thomas Finney. Cálculo de una y varias variables, 1998.

[3] Pablo Irarrázaval. *Análisis de señales*. McGraw-Hill Interamericana, 1999.