Desarrollo Matemático PCA

Buscamos una proyección $\tilde{x}_n$ de x_n de menor dimensión

- Dataset i.i.d $\mathcal{X} = \{x_1, \dots, x_N\}$, $x_n \in \mathbb{R}^D$, $\underbrace{\mu = 0}_{\text{centrado}}$

- Matriz de covarianza: $\boxed{S = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^T}$

- Buscamos transformaciones lineales $z_n = B^T x_n \in \mathbb{R}^M$, donde $B = [b_1, b_2, \dots, b_m] \in \mathbb{R}^{D \times M}$ con las columnas de B ortonormales.

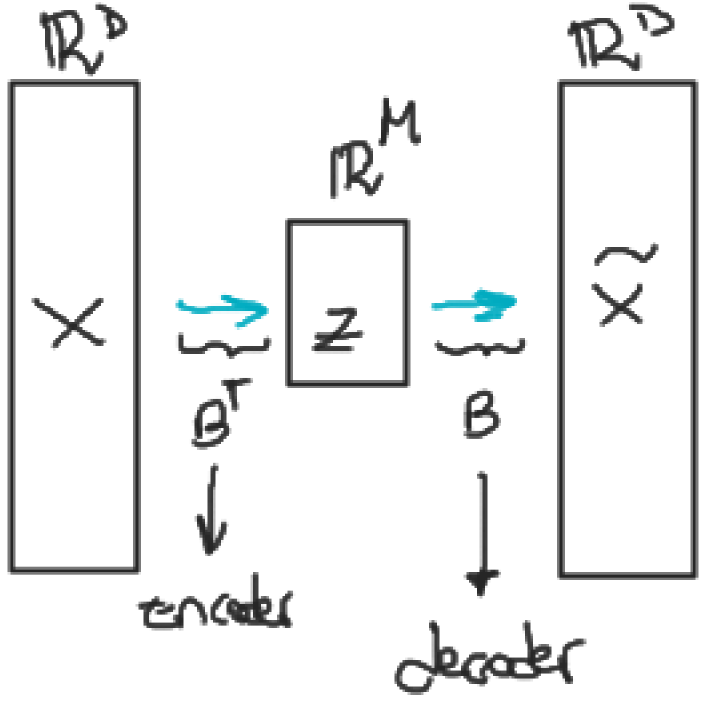
  $\rightarrow$ $b_i^T b_j = 0$ si $i \neq j$

objetivo: encontrar subespacio $U \subseteq \mathbb{R}^D$ / $\dim(U) = M < D$ donde proyectar los datos

$\downarrow$

$\tilde{x}_n$

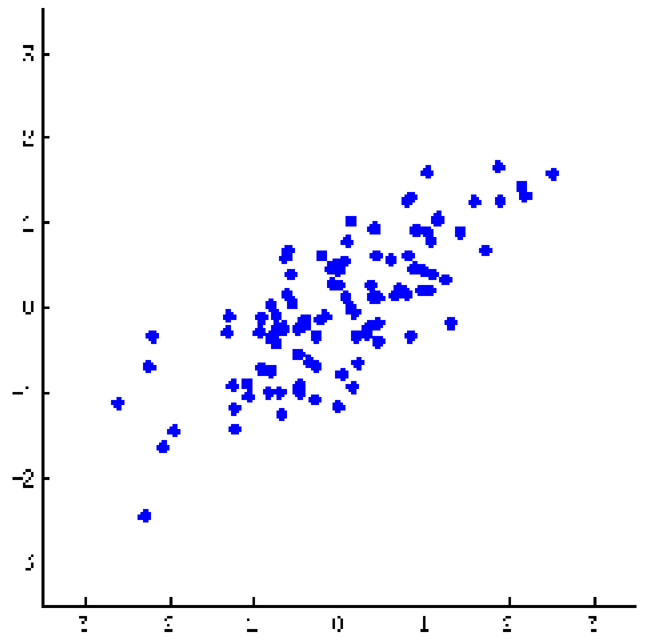
Para encontrar $\tilde{x}_n \in \mathbb{R}^D$ → z_n
→ $\underbrace{b_1, \dots, b_m}_{B}$

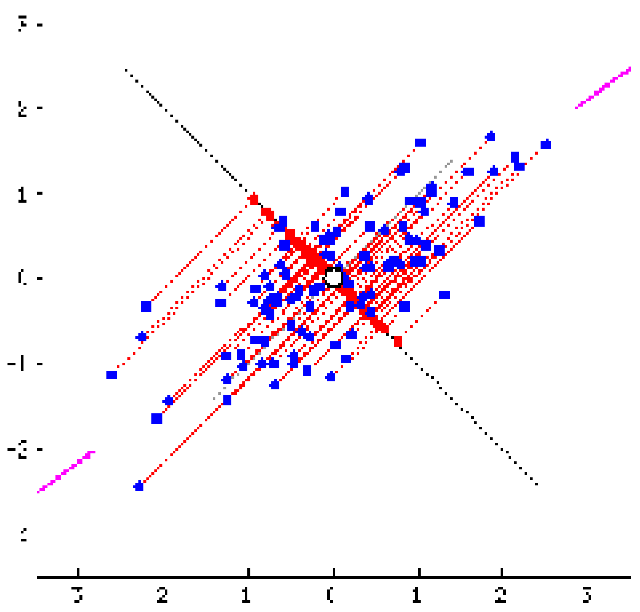


$$z = B^T x$$

$$\tilde{x}_n = B z$$

$$\underbrace{\tilde{x}_n = B B^T x}_{\text{Reconstrucción}}$$





¿Cómo encontramos z_n y B?

- Maximizar Varianza (1)
- Minimizar Error Reconstrucción (2)
- Variables Latentes (3)

Maximización de Varianza

Formulación:

Maximizar la varianza en una representación dimensional inferior $\longrightarrow$ Retener la mayor cantidad de información

$$\mathbb{V}_z[z] = \mathbb{V}_x[B^T(x-\mu)]$$
$$= \mathbb{V}_x[B^T x - B^T\mu]$$
$$= \mathbb{V}_x[B^T x]$$

La varianza no se ve afectada por μ $\downarrow$ Asumimos datos centrados

(1) Partimos con una columna de B, $b_1 \in \mathbb{R}^D$

↳ Maximizamos la varianza z_1 de $z \in \mathbb{R}^M$:

$$V_1 = \mathbb{V}[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2 \;\Big/\; \underbrace{z_{1n} = b_1^T x_n}_{\text{Proyección ortogonal de } x_n \text{ en el subespacio unidimensional formado por } b_1}$$

$$V_1 = \frac{1}{N}\sum_{n=1}^{N}(b_1^T x_n)^2$$
$$= \frac{1}{N}\sum_{n=1}^{N} b_1^T x_n x_n^T b_1$$

$$V_1 = b_1^T \underbrace{\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right)}_{S} b_1 = b_1^T S b_1 \rightarrow$$ Si aumento b_1 se incrementa V_1 $\rightarrow$ buscamos b_1 unitario $\|b_1\| = 1$

S → covarianza

Objetivo : $\max_{b_1} \; b_1^T S b_1 \;, \; \|b_1\|^2 = 1 \rightarrow$ Lagrange (optimización con restricciones

$$\mathcal{L}(b_1, \lambda) = b_1^T S b_1 + \lambda_1 (1 - b_1^T b_1)$$

$$\left.\begin{array}{l} \dfrac{\partial \mathcal{L}}{\partial b_1} = 2 b_1^T S - 2\lambda b_1^T \qquad \dfrac{\partial \mathcal{L}}{\partial \lambda_1} = 1 - b_1^T b_1 \end{array}\right\}$$

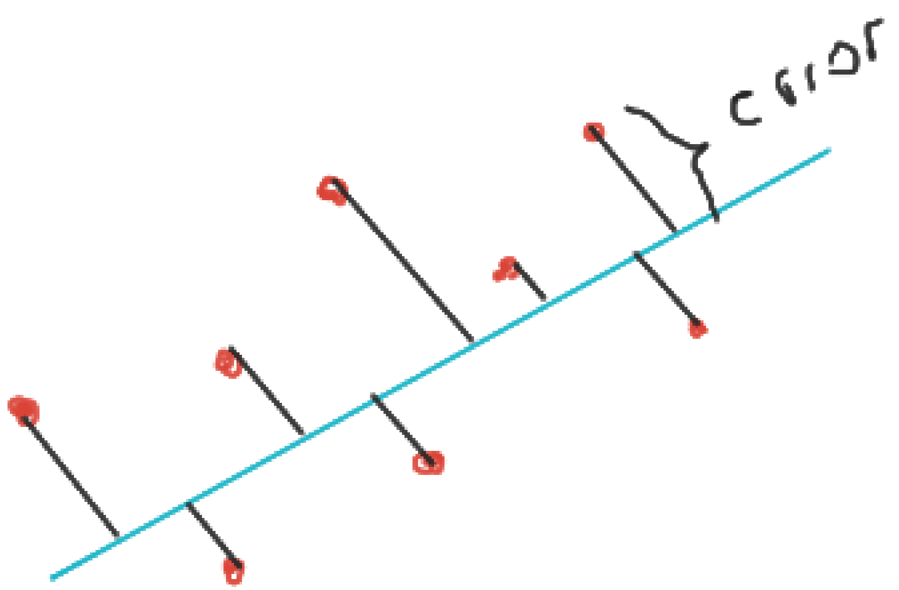
$$\frac{\partial \mathcal{L}}{\partial b_1} = 0 \Rightarrow \underbrace{S b_1 = \lambda_1 b_1}_{\text{Problema de valores y vectores propios}} \quad (1)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = 0 \Rightarrow b_1^T b_1 = 1$$

(1) $S \underset{\text{vector propio de } S}{b_1} = \underset{\text{autovalor correspondiente}}{\lambda_1} b_1$ $\qquad \sqrt{\lambda_1}$ loading factor : desvío de los datos sobre dirección de b_1

seleccionamos los autovectores asociados con los m autovalores más grandes de la matriz de covarianza $\Longrightarrow$

- Varianza contemplada $V_M = \sum_{m=1}^{M} \lambda_m$
- Varianza perdida $J_M = \sum_{j=m+1}^{D} \lambda_j$



Minimizar el error de reconstrucción

Para una base $(b_1, \dots, b_d)$ de $\mathbb{R}^D$, cualquier $x \in \mathbb{R}^D$ se puede escribir como combinación lineal de las bases

$$x = \sum_{d=1}^{D} \xi_d b_d = \sum_{m=1}^{M} \xi_m b_m + \sum_{j=M+1}^{D} \xi_j b_j$$

Queremos encontrar $\tilde{x} = \sum_{m=1}^{M} z_m b_m \in U \subseteq \mathbb{R}^D$ lo más similar posible a x

$$\underbrace{\boxed{\text{Encontrar } z \text{ y } [b_1 \dots b_m] \text{ que minimice } \|x - \tilde{x}\|}}$$

$$\tilde{x}_n = \sum_{m=1}^{M} z_{mn} b_m = B z_n \in \mathbb{R}^D, \quad z_n = [z_{1n}, \dots, z_{Mn}] \in \mathbb{R}^M$$

Minimizar el error cuadrático medio

$$\boxed{J_M = \frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2}$$

(1) optimizar z_n para una base dada

(2) encontrar la base óptima

- Asumimos una base ortonormal $(b_1 \ldots b_M)$ de $U \subseteq \mathbb{R}^D$

$$\frac{\partial J_M}{\partial z_{in}} = \frac{\partial J_M}{\partial \tilde{x}_n}\frac{\partial \tilde{x}_n}{\partial z_{in}} , \quad \frac{\partial J_M}{\partial \tilde{x}_n} = \frac{-2}{N}(x_n - \tilde{x}_n)^T \in \mathbb{R}^{1\times D}$$

$$\frac{\partial \tilde{x}_n}{\partial z_{in}} = \frac{\partial}{\partial z_{in}}\left(\sum_{m=1}^{M} z_{mn} b_m\right) = b_i \rightarrow \frac{\partial J_M}{\partial z_{in}} = \frac{-2}{N}(x_n - \tilde{x}_n)^T b_i = \frac{-2}{N}\left(x_n - \sum_{m=1}^{M} z_{mn} b_m\right)^T b_i$$

$$= \frac{-2}{N}\left(x_n^T b_i - z_{in}\underbrace{b_i^T b_i}_{=1}\right) = \frac{-2}{N}\left(x_n^T b_i - z_{in}\right)$$

minimizamos $\frac{\partial J_M}{\partial z_{in}} \Rightarrow \frac{-2}{N}(x_n^T b_i - z_{in}) = 0 \Rightarrow \underbrace{z_{in} = x_n^T b_i = b_i^T x_n}$

las coordenadas óptimas z_{in} dada una base b_i, son las proyecciones ortogonales de x_n en b_i

(2) Buscamos encontrar la base ortonormal óptima

$$\tilde{x}_n = \sum_{m=1}^{M} z_{mn} b_m = \sum_{m=1}^{M} (x_n^T b_m) b_m$$

$$\tilde{x}_n = \left(\sum_{m=1}^{M} b_m b_m^T\right) x_n \quad (a)$$

$$x_n = \sum_{d=1}^{D} z_{dn} b_d = \sum_{d=1}^{D} (x_n^T b_d) b_d = \left(\sum_{d=1}^{D} b_d b_d^T\right) x_n$$

$$x_n = \left(\sum_{m=1}^{M} b_m b_m^T\right) x_n + \left(\sum_{j=M+1}^{D} b_j b_j^T\right) x_n \quad (b)$$

De (a) y (b) $\rightarrow x_n - \tilde{x}_n = \sum_{j=M+1}^{D} (x_n^T b_j) b_j$

$$J_M = \frac{1}{N} \sum_{n=1}^{N} \left\| \sum_{j=M+1}^{D} (b_j^T x_n b_j) \right\|^2$$

$$J_M = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=M+1}^{D} b_j^T x_n x_n^T b_j = \sum_{j=M+1}^{D} b_j^T \underbrace{\left(\frac{1}{N} \sum_{n=1}^{N} x_n x_n^T\right)}_{S} b_j$$

$$J_M = \sum_{j=M+1}^{D} tr(b_j^T S b_j) = tr\left(\left(\underbrace{\sum_{j=M+1}^{D} b_j b_j^T}_{\text{MATRIZ DE PROYECCIÓN}}\right) S\right)$$

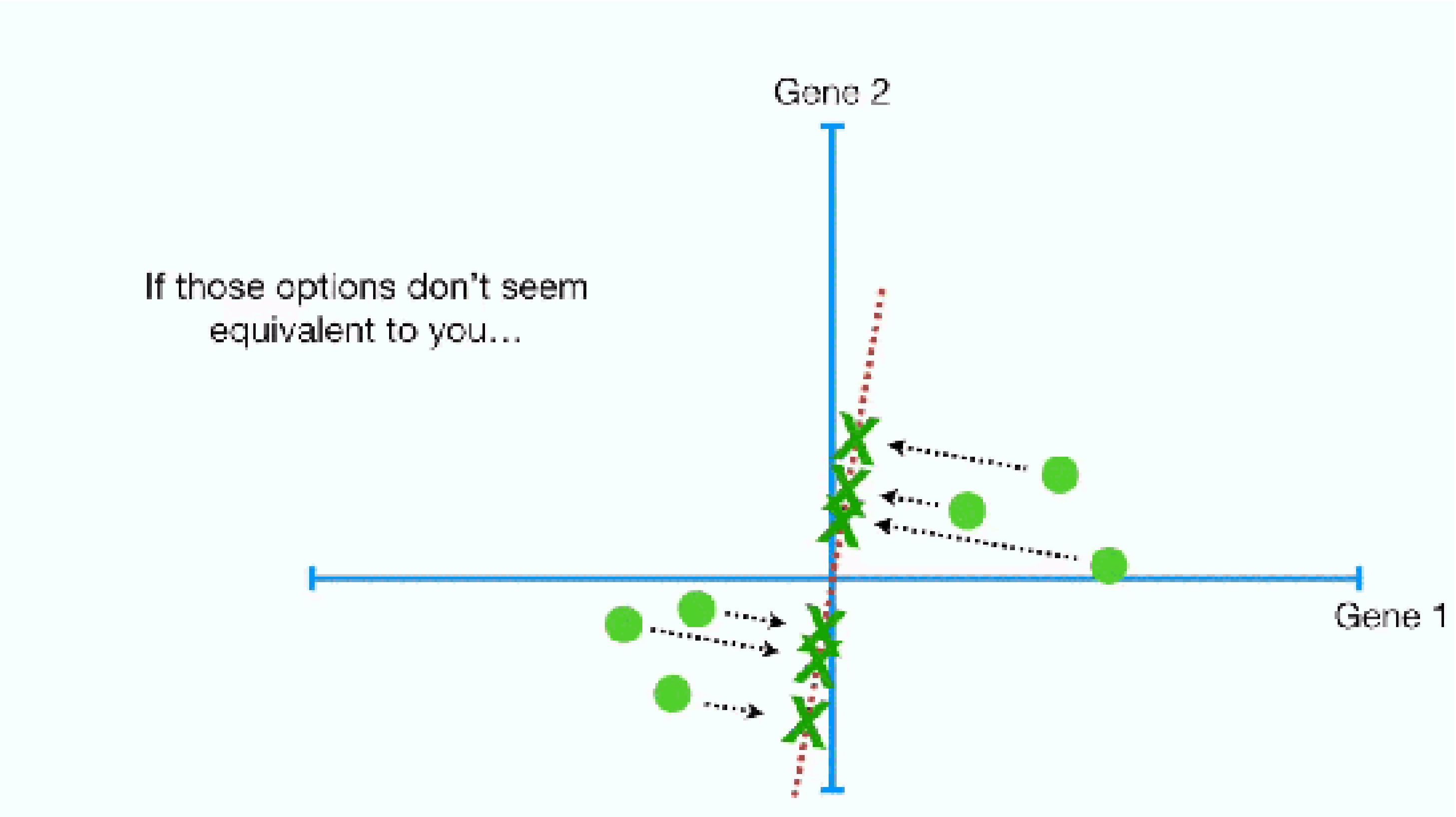
Minimizar el error de reconstrucción es equivalente a maximizar la varianza en las direcciones (componentes) escogidas.

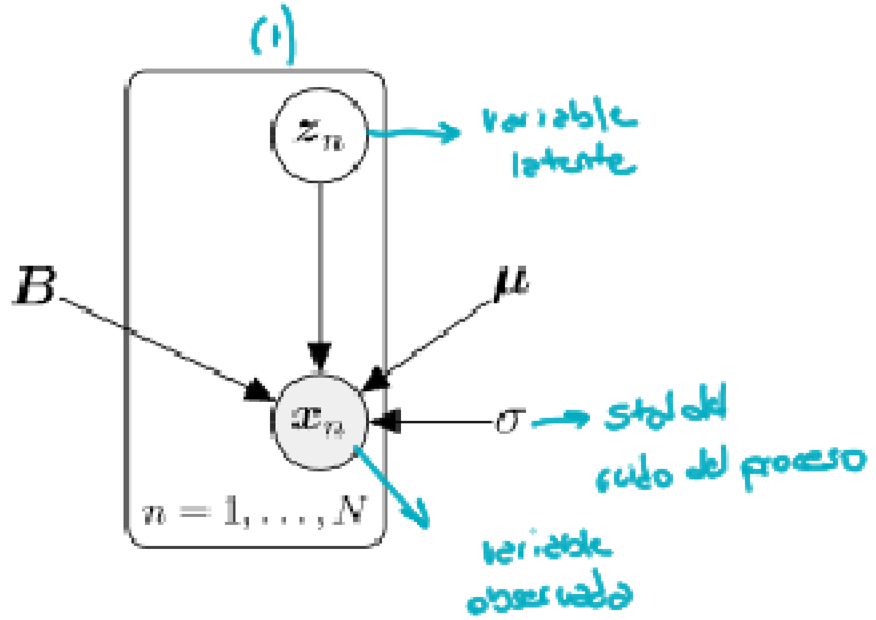
El error de reconstrucción se puede pensar como la matriz de covarianza proyectada sobre el complemento ortogonal de U

$\Downarrow$

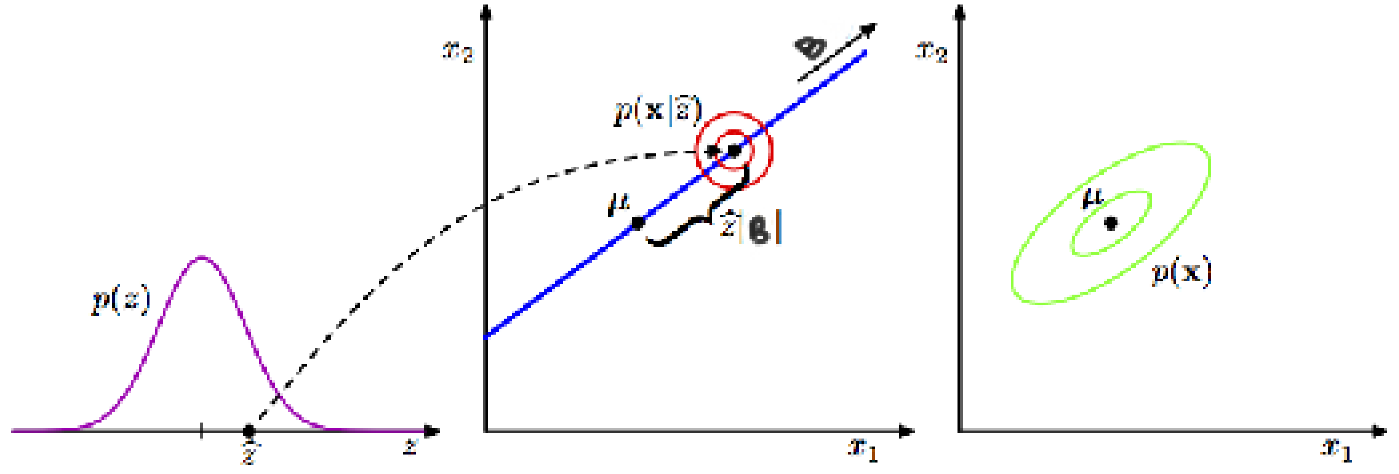
Equivalente a minimizar la varianza de los datos proyectados sobre el subespacio ignorado (ortogonal a U)

$$J_M = \sum_{j=M+1}^{D} \lambda_j$$





Enfoque por Variables Latentes



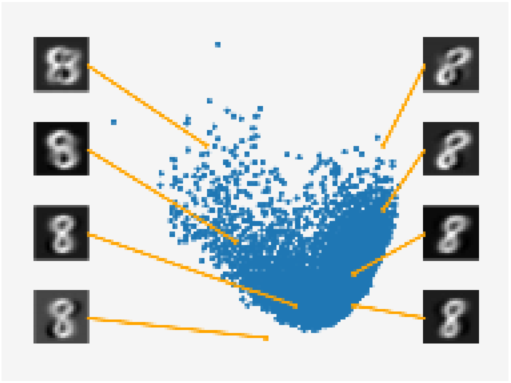
- Asumimos una variable latente $z \in \mathbb{R}^M$:
  - $p(z) = \mathcal{N}(0, I) \longrightarrow$ distribución a priori
  - $\underset{\sim}{x} = Bz + \mu + \epsilon \in \mathbb{R}^D$, $\underbrace{\epsilon \sim \mathcal{N}(0, \sigma^2 I)}_{\text{Ruido Gausiano}}$

    variable observada

El vínculo entre las variables latentes y observables es:

$$\boxed{p(x/z, B, \mu, \sigma^2) = \mathcal{N}(x/Bz+\mu, \sigma^2 I)}$$

Proceso generativo



$$\begin{cases} z_n \sim \mathcal{N}(z/0, I) \\ x_n/z_n \sim \mathcal{N}(x/Bz_n+\mu, \sigma^2 I) \end{cases}$$

Muestreamos z_n de $p(z)$ y después x_n

Distribución conjunta

$$p(x, z / B, \mu, \sigma^2) = p(x/z, B, \mu, \sigma^2)\, p(z)$$

Derivaciones

(1) Si $\sigma = 0$, PCA $\longrightarrow$ PPCA

(2) Si para cad d, $\sigma_d \longrightarrow$ FA

(3) $p(z)$ no Gaussiana $\longrightarrow$ ICA

Probabilistic PCA

$$\boxed{¿\, p(z/x)\, ?}$$

$$\underbrace{p(x) = \int p(x/z)\, p(z)\, dz}_{\text{marginal de } x}$$

↓

MÁXIMA VEROSIMILITUD