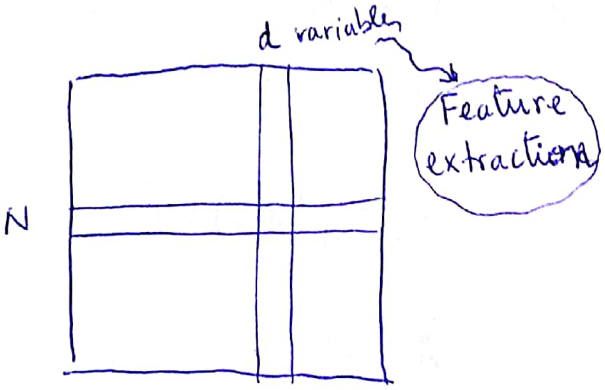
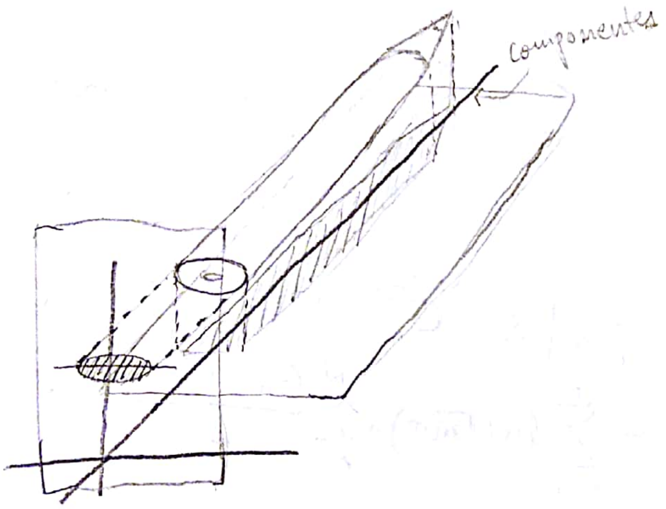
(2) Reducción de dimensión



- La probabilidad de sobreajuste se incrementa con la dimensión
- Visualización de datos.
- Preservar la información relevante en los datos

Métodos

(1) Analisis de componentes principales (ACP/PCA)



→ Para PCA: "información relevante" es dispersión

- Tenemos una muestra de datos $\{x_1, x_2, \ldots, x_N\}$, $x_n \in \mathbb{R}^d$ (→ son vectors, punts d-dimensionals); que provienen de un vector aleatorio $X = (X_1, \ldots, X_d)^T$, con media μ y matriz de covarianza $\Sigma \longrightarrow \Sigma = (\sigma_{ij}^2)$, $Cov(X_i, X_j) = \sigma_{ij}^2$, $Var(X_i) = \sigma_{ii}^2 =: \sigma_i^2$
- Consideramos el problema de encontrar un nuevo sistema de coordenadas (vector de variables). $Y = (Y_1, \ldots, Y_d)^T$

Condiciones:

1) $Cov(Y_i, Y_j) = 0$, $i \neq j$
2) $Var(Y_1) > Var(Y_2) > \cdots > Var(Y_d)$
3) $\sum_{i=1}^{d} Var(X_i) = \sum_{i=1}^{d} Var(Y_i)$

La solución más simple es que las Y_j sean c.l. de las X_i
↳ Combinación lineal

$$Y_j = a_j^T X$$

Imponemos la normalización $\|a_j\| = 1 \Leftrightarrow \|a_j\|^2 = 1$ → lo mismo (para simplificar)

<u>Objetivo</u>: a_1 ha de maximizar la varianza de Y_1 sujeto a $\|a_1\|^2 = 1$

- $Var(Y_1) = Var(a_1^T X) = a_1^T Var(X) a_1 = a_1^T \cdot \Sigma \cdot a_1$

vector aleatorio

vector de números

Σ

vector matriz

<u>Solución</u> a maximizar $a_1^T \Sigma a_1$ sujeto a $\|a_1\|^2 = 1 \rightsquigarrow \|a_1\|^2 - 1 = 0$

$$L(a_1) = a_1^T \Sigma a_1 - \lambda(\|a_1\|^2 - 1)$$ → Función lagrangiana.

$$\frac{\partial L}{\partial a_1} = 2\Sigma a_1 - 2\lambda a_1 = 0 \Rightarrow \Sigma \cdot a_1 = \lambda a_1$$

$x \cdot y = (\,)(\,)$

$x^T y = (—)(|)$

→ encontrar un vector propio a_1 de valores propios λ. → mirar!!

¡Ya está! $\lambda_{(1)} > \lambda_{(2)} > \cdots > \lambda_{(d)} \geq 0$ → propiedad matriz covarianza Σ

→ maxim. $Var(Y_1) = a_1^T \Sigma a_1 = a_1^T(\lambda a_1) = \lambda(a_1^T a_1) = \lambda \underbrace{\|a_1\|^2}_{1} = \boxed{\lambda} \Rightarrow$

$\Rightarrow$ elijo como $\lambda = \lambda_{(1)} \Rightarrow a_1$ es su vector propio correspondiente.

- Resulta que a_2 se encuentra igual, con la condición adicional $\boxed{a_2^T a_1 = 0}$ → perpendiculares (como las componentes)

→ a_2 es el vector propio de Σ con valor propio $\lambda_{(2)}$
... y así $a_3, a_4, \dots$

- <u>Resultado</u>

$$A = \begin{pmatrix} a_1 ; & a_2 ; & \cdots ; & a_d \\ \downarrow & \downarrow & & \end{pmatrix} \qquad Y = A^T X$$

m

hay que coger m componentes principales

Si nos quedamos con $m (\leq d)$ de los C.P. (componentes principals)

$$\frac{\sum_{i=1}^{m} \lambda_{(i)}}{\sum_{i=1}^{d} \lambda_{(i)}} \times 100\% \rightarrow \text{el \% de varianza retenida.}$$

¿Qué porcentaje nos interesa mantener?

⊛ Si el objetivo es reducir la dimensión (eliminando ruido), un criterio es del 90% de la varianza retenida.

⊛ Si el objetivo es visualizar, $m = 2$ ó 3

Propiedades

1) La matriz de covarianza de las Y es:

$$\Delta = \mathrm{diag}(\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(d)})$$

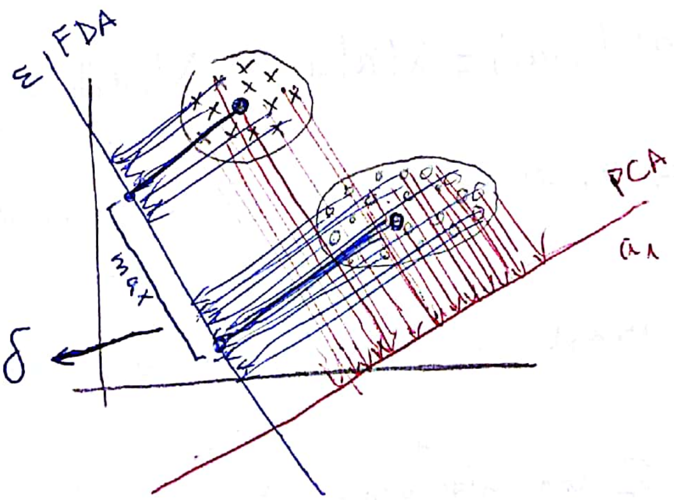
2) $\mathrm{Var}(Y_1) > \mathrm{var}(Y_2) > \ldots > \mathrm{Var}(Y_d)$

$$\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_d$$

traza ⇒ suma dels elements de la diagonal

3) $$\sum_{i=1}^{d} \mathrm{Var}(X_i) = \sum_{i=1}^{d} \lambda_{(i)} = \mathrm{Tr}(\Sigma) = \sum_{i=1}^{d} \mathrm{Var}(X_i)$$

traza matriz = suma valores propios



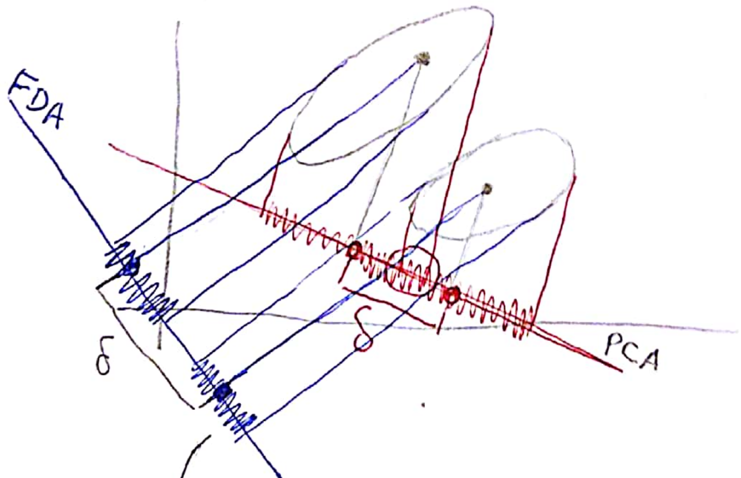
(2) Analisis discriminante de Fischer (ADF / FDA)

$$D = \{(x_1, t_1), \dots, (x_N, t_N)\}, \quad x_n \in \mathbb{R}^d, \quad t_n \in \{1, 2\}$$

N_1, N_2 observaciones de cada clase, $N_1 + N_2 = N$

$$m_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n \; ; \quad m_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n$$

$$\delta = |\mu_1 - \mu_2| = |w^T m_1 - w^T m_2| = |w^T (m_1 - m_2)|$$



minimizar δ y dispersión