ANALISIS DE COMPONENTES PRINCIPALES (PCA)

Como estudiar datos $x_1, \ldots, x_n$, cuya dimensión puede ser grande, no es una tarea inmediata, nos gustaría una forma de representar $x_1, \ldots, x_n$ en dimensiones menores. En estadística hemos aprendido que un "buen" resumen de los datos es un promedio, pero una forma de hacerlo más "flexible" sería darle un peso a cada componente de x_i, la cual midiera una "importancia" en el promedio.

$x_1, \ldots, x_n$ realizaciones del vect. aleatorio $\mathbb{X}$

$$\alpha' \mathbb{X} = \sum_{i=1}^{p} \alpha_i X_i, \quad \text{donde} \quad \sum_{i=1}^{p} \alpha_i^2 = 1 .$$

La forma en que $\mathbb{X}$ varía es una característica de interés, por lo cual nos planteamos el problema de encontrar $\alpha \in \mathbb{R}^p$, tal que $\sum_{j=1}^{p} \alpha_j^2 = 1$ y $\alpha' \mathbb{X}$ tiene varianza máxima. En otras palabras hay que encontrar $\alpha \in \mathbb{R}^p$ tal que $\alpha' \mathbb{X}$ alcance el máximo

$$\max \left\{ VAR(\alpha' \mathbb{X}) : \alpha \in \mathbb{R}^p \text{ y } \|\alpha\| = 1 \right\}$$

$$= \max \left\{ \alpha' \, VAR(\mathbb{X}) \, \alpha : \alpha \in \mathbb{R}^p \text{ y } \|\alpha\| = 1 \right\}$$

Si regresamos al Teorema C, específicamente al caso en que $\mathbb{B} = \mathbb{I}_p$, es claro que una **solución** al problema de optimización anterior, es tomar $\alpha = \gamma_1$, donde γ_1 es el vector propio **correspondiente** al más grande valor propio λ_1 de la matriz de covarianzas $\Sigma \equiv VAR(\mathbb{X})$.

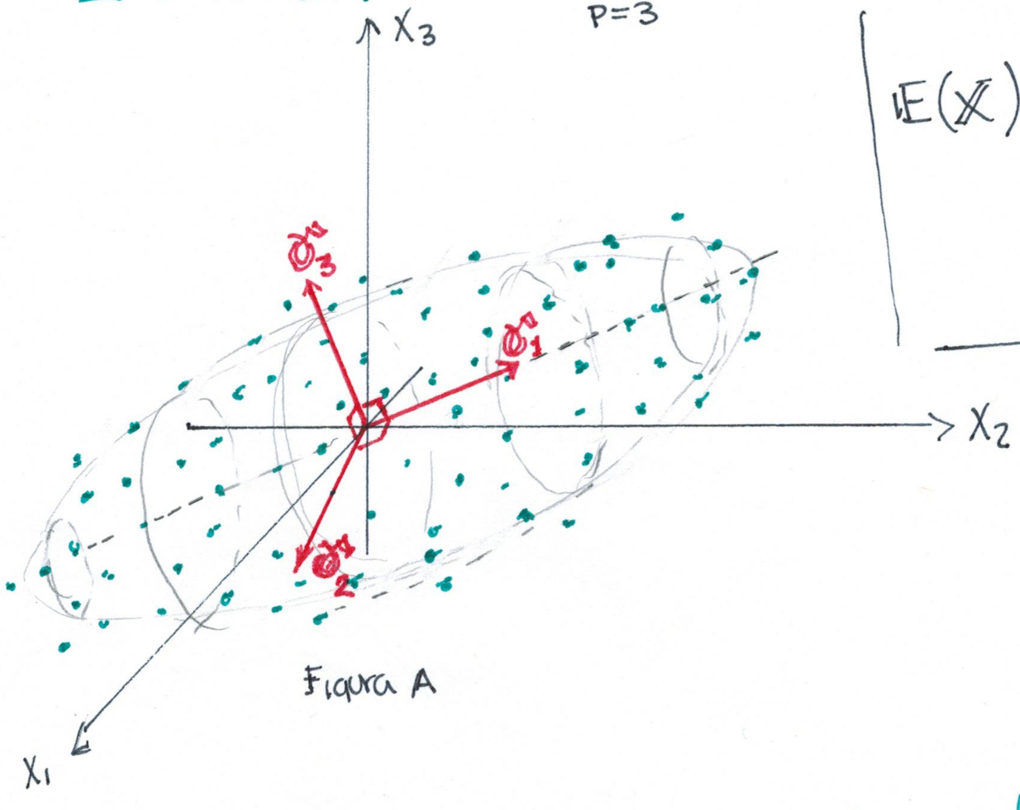

Figura A

$$\mathbb{E}(\mathbb{X}) = \begin{pmatrix}\mathbb{E}(X_1)\\ \mathbb{E}(X_2)\\ \mathbb{E}(X_3)\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix}$$

Si los datos tuvieran un comportamiento "ideal"[(1)] como en el dibujo, la dirección en $\mathbb{R}^3$ en la que existe mayor variabilidad **está**

(1) ver página 15

(1)

Asumiendo $\mathbb{x}_1, \dots, \mathbb{x}_n$ (los datos) son una muestra aleatoria del vector aleatorio $\mathbb{X}$ tal que $\mathbb{X} \sim F$, donde la función de distribución multivariada F es tal que:

$$\mathbb{E}(\mathbb{X}) = \mathbb{O} \quad \text{y} \quad VAR(\mathbb{X}) = \Sigma > 0.$$

Además $F: \mathbb{R}^3 \to [0,1]$ con soporte como en el dibujo de la Figura A.

$$\text{Soporte}(F) = \{\mathbb{x} \in \mathbb{R}^3 : \; f(\mathbb{x}) > 0\}$$

f = densidad de F.

dada por γ_1^* Pregunta: ¿Porqué es así? (*)

Por el Teorema C la elección $\alpha = \gamma_1^*$ garantiza que $VAR(\gamma_1^{*\prime} \mathbb{X}) = \gamma_1^{*\prime} VAR(\mathbb{X}) \gamma_1^*$ alcanza el máximo, <u>el</u> <u>cual</u> <u>vale</u> λ_1. $\left(VAR(\gamma_i^{*\prime} \mathbb{X}) = \lambda_1 \right)$.

(PC1) .... $\mathbb{Y}_1 \equiv \gamma_1^{*\prime} \mathbb{X}$ ← La primera componente principal

Es una combinación lineal "estandarizada" $\Big($es decir convexa: $\sum_{j=1}^{p} \gamma_{j1}^2 = 1$; $\gamma_i^{*\prime} = (\gamma_{i1}, \dots, \gamma_{ip})\Big)$ de las componentes de $\mathbb{X}$.

Está construída para tener varianza máxima.

(PC2) ... $\mathbb{Y}_2 \equiv \gamma_2^{*\prime} \mathbb{X}$ ← La segunda componente principal

donde γ_2^* es el vector propio de Σ asociado a $\lambda_2 (\leq \lambda_1)$, entonces $\Sigma \gamma_2^* = \lambda_2 \gamma_2^*$ por lo que $\gamma_2^{*\prime} \Sigma \gamma_2^* = \lambda_2$ $\left(VAR(\gamma_2^{*\prime} \mathbb{X}) = \lambda_2 \leq \lambda_1 \right)$.

(PC3) ... $\mathbb{Y}_3 \equiv \gamma_3^{*\prime} \mathbb{X}$ ← La tercera componente principal

etc ...