Este análisis lo vamos a utilizar con dos fines:

- Reducir la dimensionalidad de los datos

  Se pretende pasar de p variables a k comp. principales que son comb. lineales de todas las variables.

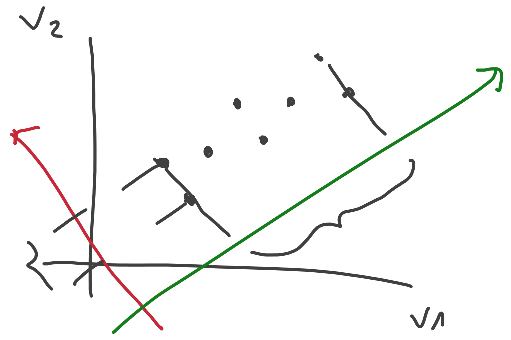
- Estudiar la relación entre las variables.

- Aplicar un método que <u>NO</u> requiere que NO haya correlación entre sus variables (p.e: Regresión lineal)

<u>Interpretación geométrica</u>

Cada comp. va a definir una dirección $a_i = (a_{i1}, \dots, a_{ip})$. Se quiere encontrar aquellas direcciones que expliquen la mayor variabilidad posible de los datos.

Ej: 2 variables



La 1a comp. principal max. la varianza de las proy. de las variables.

La 2a comp. principal será ortogonal a la 1a comp. y max. la varianza restante.

Observaciones:

1) Si hay p variables, se necesitan p CP para capturar el 100% de la variabilidad.

2) En general, se puede capturar gran parte de la variabilidad con $k \ll p$.

3) El dataset se queda como $n \times k$

Un poco de teoría

Consideremos el vector aleatorio

$$\mathbb{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \quad \text{con matriz de cov. } \Sigma$$

Queremos construir las comp. principales:

$$Y_1 = a_1' \mathbb{X} = a_{11}X_1 + \dots + a_{1p}X_p \quad (1 \text{ CP})$$
$$\vdots$$
$$Y_p = a_p' \mathbb{X} = a_{p1}X_1 + \dots + a_{pp}X_p \quad (p \text{ CP})$$

Tenemos que:

$$Var(Y_i) = a_i' \Sigma a_i$$
$$i = 1, \dots, p$$
$$Cov(Y_i, Y_k) = a_i' \Sigma a_k = 0 \quad i \neq k$$

↳ Las CP son comb. lineales de $\mathbb{X}$ NO correlacionadas entre sí.

Por tanto, buscamos las a_i's tal que:

- La 1a comp. principal es $a_1'\mathbb{X}$ tal que a_1 max. $Var(a_1'\mathbb{X})$
- La 2a " " " $a_2'\mathbb{X}$ " " a_2 " $Var(a_2'\mathbb{X})$,

la $cov(a_1'\mathbb{X}, a_2'\mathbb{X}) = 0$ y a_2 es ortogonal a a_1.

⋮

- La i-ésima comp. principal es $a_i'\mathbb{X}$ tal que a_i max $Var(a_i'\mathbb{X})$,

la $cov(a_i'\mathbb{X}, a_k'\mathbb{X}) = 0$ y a_i es ortogonal a las anteriores.

¿Qué valen a_i's?

Teorema: Sea Σ la matriz de cov. de $\mathbb{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$. Suponga que Σ tiene pares de valores y vectores propios:

$$(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p) \quad \text{con} \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

La i-ésima comp. principal se define como:

$Y_i = e_i' \mathbb{X}$ (es decir, $a_i = e_i$ vector propio)
"scores"

Además:

$Var(Y_i) = e_i' \Sigma e_i = \lambda_i$
$\quad i = 1, \ldots, p$
$Cov(Y_i, Y_k) = 0 \quad i \neq k$

Observaciones:

1) La varianza total se preserva.

2) La prop. de varianza explicada por la k-ésima CP es:

$$\frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p}$$

3) En muchas ocasiones para las 1as comp. principales esta proporción es 0.8 - 0.9

4) Si el vector principal i es $e_i = \begin{pmatrix} e_{i1} \\ \vdots \\ e_{ip} \end{pmatrix}$, entonces $|e_{ik}|$ mide la "importancia" de la variable k en la i-ésima CP.

Más aún, $e_{ik} \sim \rho_{Y_i, X_k}$ (corr. entre CP i y variable k)

Concretamente:

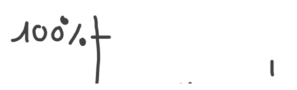
$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, \ldots, p$$
"loadings"

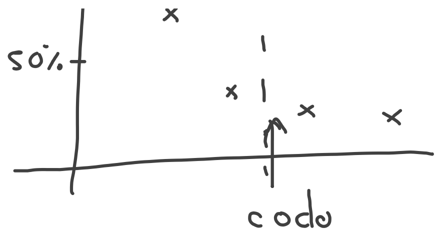
Nota: Si los datos están estandarizados utilizamos ρ en vez de Σ

<u>Número apropiado de comp. principales</u>

i) Mirado la prop. de varianza explicada acumulada y decidir cuanta info. es suficiente (80-85%)

ii) Regla del codo (Elbow)





Cogeríamos 2 CP.

Ejemplo numérico.

Example 8.1 (Calculating the population principal components)

Suppose the random variables X_1, X_2 and X_3 have the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

It may be verified that the eigenvalue–eigenvector pairs are

$$\left.\begin{array}{ll} \lambda_1 = 5.83, & \mathbf{e}_1' = [.383, -.924, 0] \\ \lambda_2 = 2.00, & \mathbf{e}_2' = [0, 0, 1] \\ \lambda_3 = 0.17, & \mathbf{e}_3' = [.924, .383, 0] \end{array}\right\} \text{ dirección CP}$$

Therefore, the principal components become

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} = .383X_1 - .924X_2 \\ Y_2 &= \mathbf{e}_2'\mathbf{X} = X_3 \\ Y_3 &= \mathbf{e}_3'\mathbf{X} = .924X_1 + .383X_2 \end{aligned}$$

The variable X_3 is one of the principal components, because it is uncorrelated with the other two variables.

Equation (8-5) can be demonstrated from first principles. For example,

$$\left.\begin{aligned} \operatorname{Var}(Y_1) &= \operatorname{Var}(.383X_1 - .924X_2) \\ &= (.383)^2 \operatorname{Var}(X_1) + (-.924)^2 \operatorname{Var}(X_2) \\ &\quad + 2(.383)(-.924) \operatorname{Cov}(X_1, X_2) \\ &= .147(1) + .854(5) - .708(-2) \\ &= 5.83 = \lambda_1 \\ \operatorname{Cov}(Y_1, Y_2) &= \operatorname{Cov}(.383X_1 - .924X_2, X_3) \\ &= .383 \operatorname{Cov}(X_1, X_3) - .924 \operatorname{Cov}(X_2, X_3) \\ &= .383(0) - .924(0) = 0 \end{aligned}\right\} \text{ Comprobación}$$

It is also readily apparent that

$$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2.00 + .17$$

validating Equation (8-6) for this example. The proportion of total variance accounted for by the first principal component is $\lambda_1/(\lambda_1 + \lambda_2 + \lambda_3) = 5.83/8 = .73$. Further, the first two components account for a proportion $(5.83 + 2)/8 = .98$ of the population variance. In this case, the components Y_1 and Y_2 could replace the original three variables with little loss of information.

Next, using (8-8), we obtain

$$\rho_{Y_1,X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{.383\sqrt{5.83}}{\sqrt{1}} = .925$$

$$\rho_{Y_1,X_2} = \frac{e_{21}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-.924\sqrt{5.83}}{\sqrt{5}} = -.998$$

Notice here that the variable X_2, with coefficient $-.924$, receives the greatest weight in the component Y_1. It also has the largest correlation (in absolute value) with Y_1. The correlation of X_1, with Y_1, .925, is almost as large as that for X_2, indicating that the variables are about equally important to the first principal component. The relative sizes of the coefficients of X_1 and X_2 suggest, however, that X_2 contributes more to the determination of Y_1 than does X_1. Since, in this case, both coefficients are reasonably large and they have opposite signs, we would argue that both variables aid in the interpretation of Y_1.

Finally,

$$\rho_{Y_2,X_1} = \rho_{Y_2,X_2} = 0 \quad \text{and} \quad \rho_{Y_2,X_3} = \frac{\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{\sqrt{2}}{\sqrt{2}} = 1 \quad \text{(as it should)}$$

The remaining correlations can be neglected, since the third component is unimportant. ■

Ejemplo gráfico (R)