FACULTAD DE CIENCIAS ECONÓMICAS

FAMAF
Facultad de Matemática, Astronomía, Física y Computación

UNC
Universidad Nacional de Córdoba

DIPLOMATURA

CIENCIA DE DATOS, INTELIGENCIA ARTIFICIAL Y SUS APLICACIONES EN ECONOMÍA Y NEGOCIOS

Análisis Multivariado

Dra. María Inés Stimolo

Mgter. Mariana Gonzalez

Dra. Ana Georgina Flesia

- **¿Cómo vamos a trabajar?**

  Presentación conceptual de los métodos multivariados básicos con aplicaciones

- **Cronograma de clases**
- **Bibliografía** (para profundizar temas)

En aula virtual

- **Comunicación**

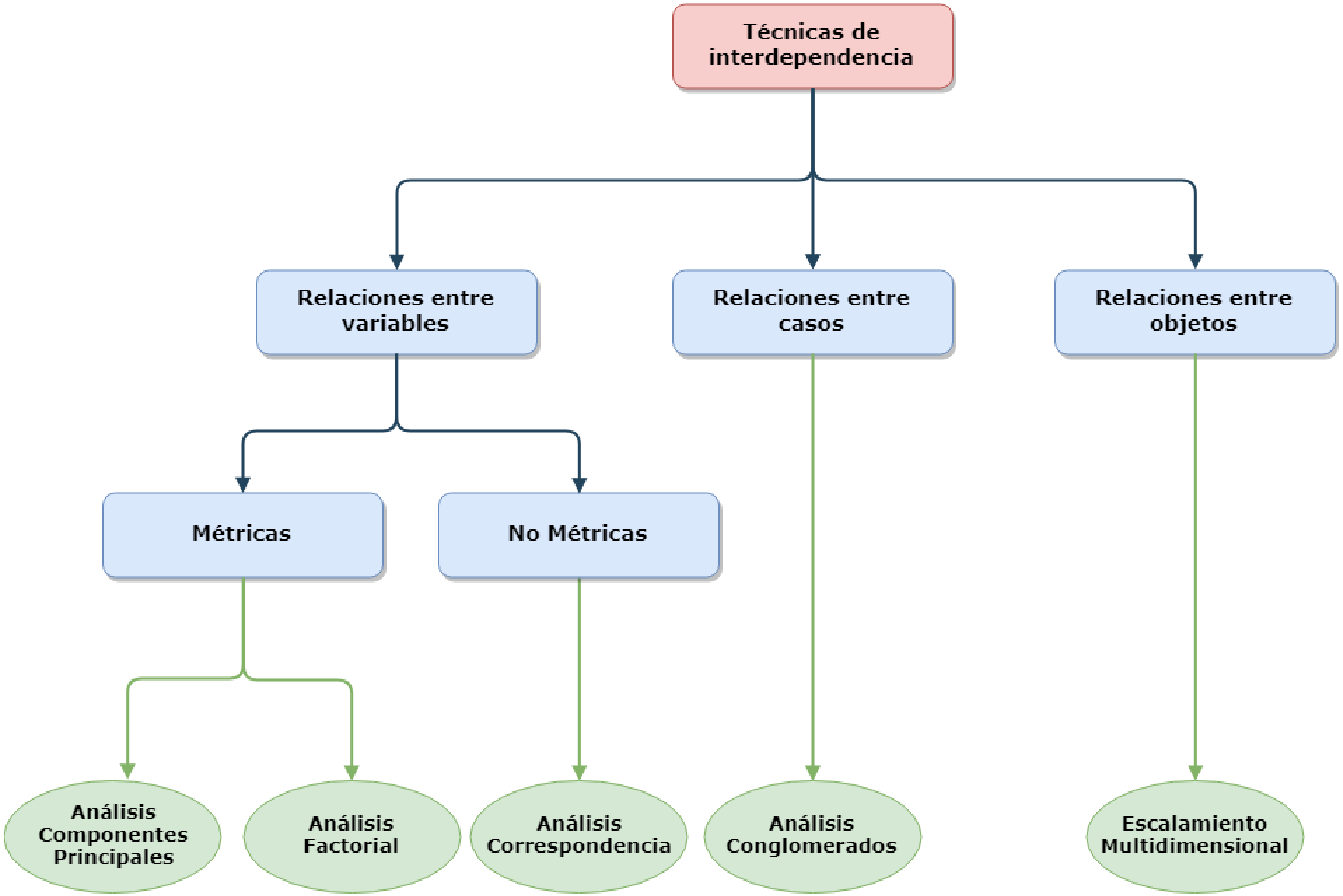
  Foros de intercambio aula virtual

  Cta slack

TÉCNICAS MULTIVARIADAS

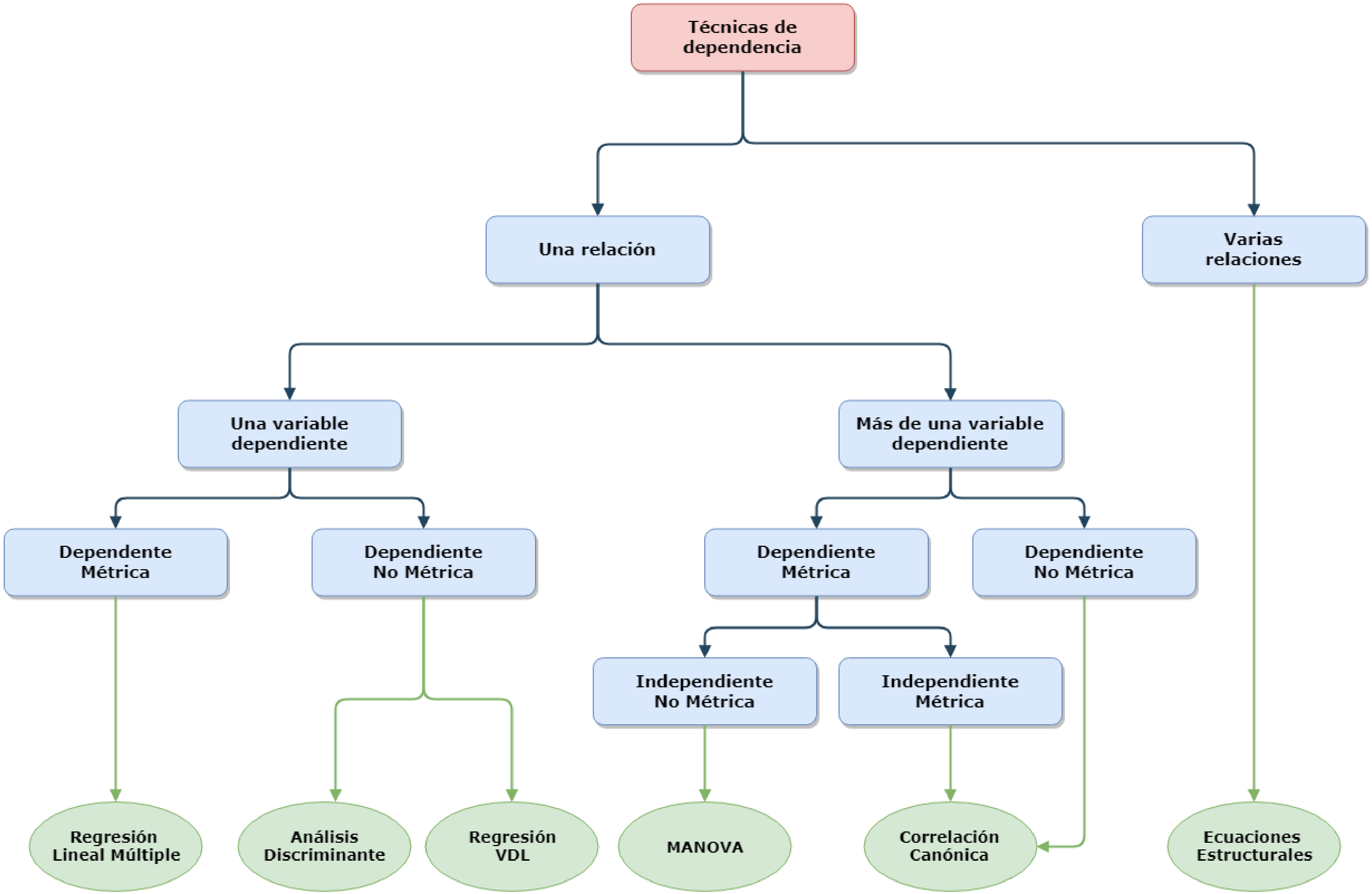
Clasificación de las técnicas multivariadas

Uriel Jiménez y Manzano 2017, adaptado de Hair *et al.* 2014



Clasificación de las técnicas multivariadas

Uriel Jiménez y Manzano 2017, adaptado de Hair *et al.* 2014



MEDIDAS DESCRIPTIVAS MULTIVARIADAS

Matriz de datos

$$\mathbf{X}_{n\times p} = \begin{array}{c|cccc}
\text{Obs} & \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \\
\hline
1 & x_{11} & x_{12} & \cdots & x_{1p} \\
\hline
2 & x_{21} & x_{22} & \cdots & x_{2p} \\
\hline
\cdots & \cdots & \cdots & \ddots & \cdots \\
\hline
\text{n} & x_{n1} & x_{n2} & \cdots & x_{np}
\end{array} \quad \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \cdot \\ \mathbf{x}'_n \end{bmatrix}$$

$$(\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots\ldots \quad \mathbf{x}_p)$$

Cada elemento $\boldsymbol{x}'_i$ es un vector fila *px1* que representa el valor de las *p* variables de la observación *i*

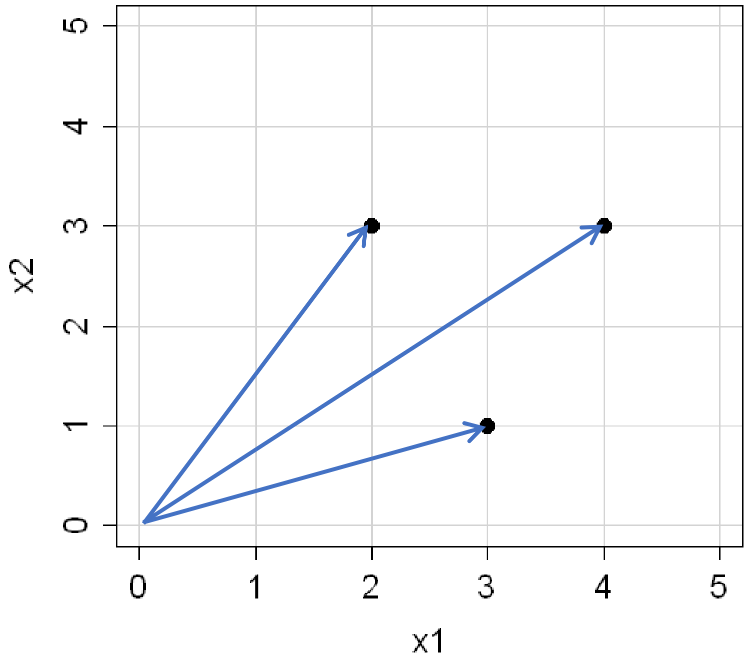
R^p Se representan las observaciones

Cada elemento $\boldsymbol{x}_j$ es un vector columna *nx1* que representa la variable x_j medida en los *n* elementos de la población.

R^n Se representan las variables

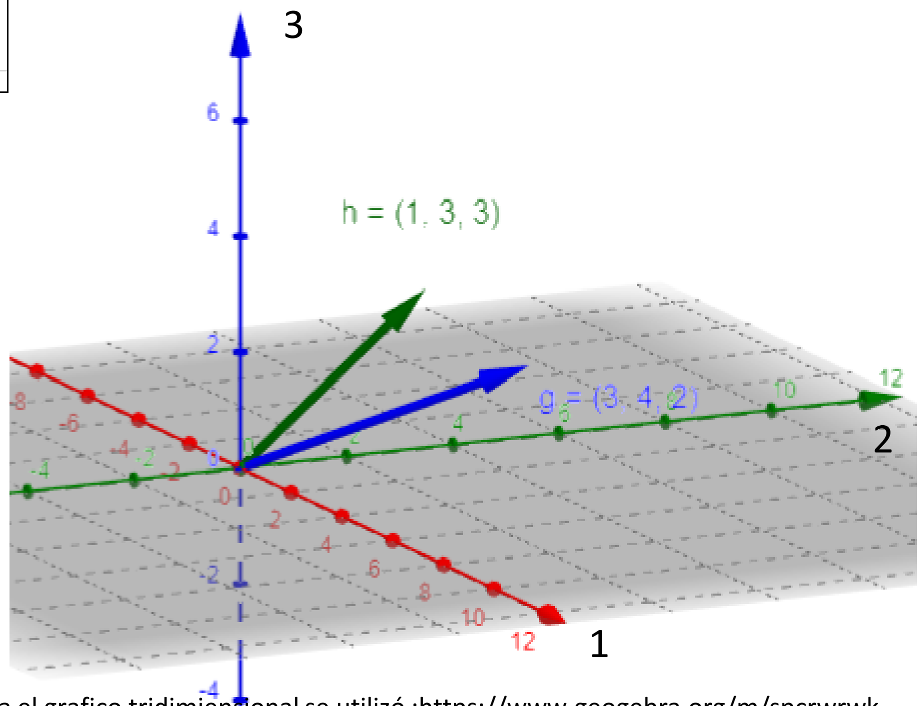
Representación de espacio de variables y observaciones

Observaciones en R^2



$$X = \begin{bmatrix} 3 & 1 \\ 4 & 3 \\ 2 & 3 \end{bmatrix}$$

Variables en R^3

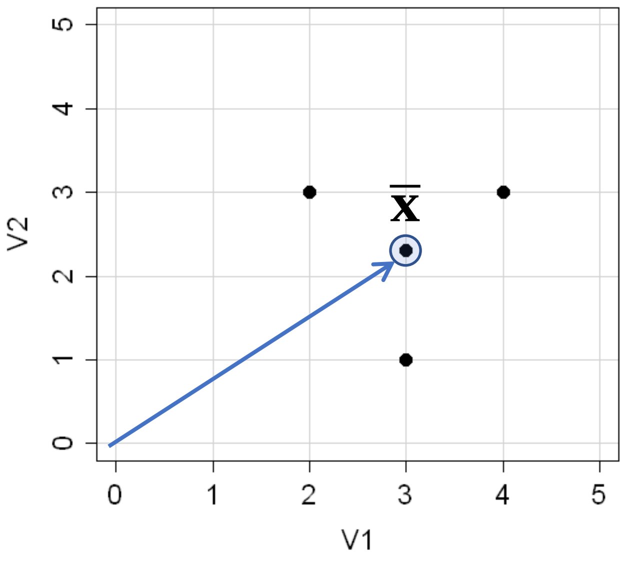


Para el grafico tridimiensional se utilizó :https://www.geogebra.org/m/spcrwrwk

Análisis descriptivo Multivariado

Media $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1}{n}\mathbf{x}_j^{'}\mathbf{1} \qquad j=1,\ 2,\ldots,\ p$

Vector de Medias



$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{n} = \frac{1}{n}\mathbf{X}'\mathbf{1}_n$$

$$\bar{\mathbf{x}} = \begin{bmatrix} 3 \\ 2{,}3 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 3 & 4 & 2 \\ 1 & 3 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

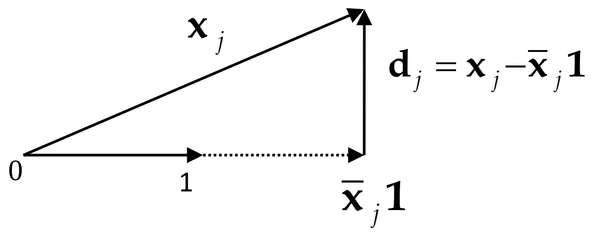
Interpretación geométrica del vector de medias

La proyección del vector de datos de la j-ésima variable sobre el vector constante es igual:

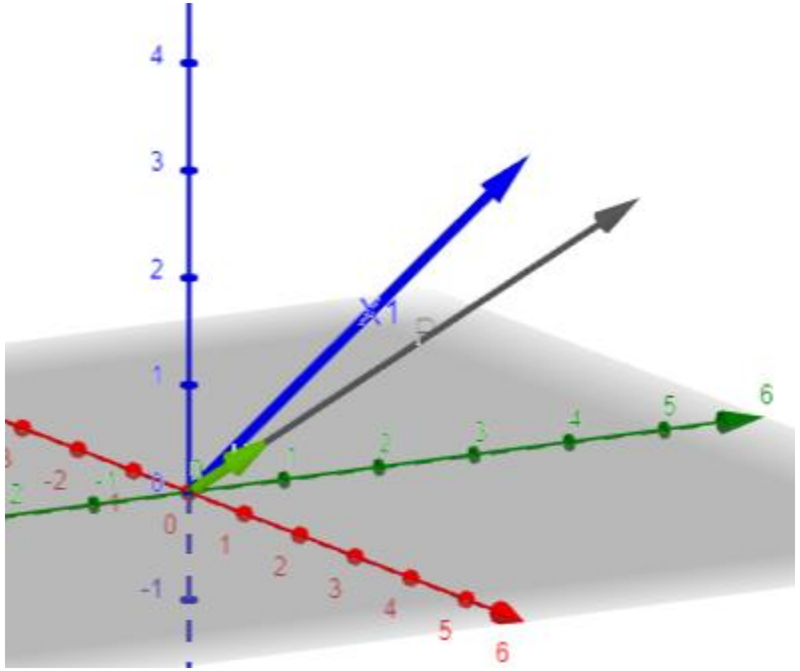
$\frac{1}{\sqrt{n}}\mathbf{1}_n$ Vector constante de norma uno en R_n

$$P_{\mathbf{x}_j} = k\frac{1}{\sqrt{n}}\mathbf{1}_n \qquad k = \frac{1}{\sqrt{n}}\mathbf{1}_n'\mathbf{x}_j = \bar{x}_j\sqrt{n}$$

$$= \bar{x}_j\mathbf{1}_n$$



Para el ejemplo proyección de la variable X1



Medidas de variabilidad univariadas

Vector diferencia $\mathbf{d}_j = \mathbf{x}_j - \overline{x}_j \mathbf{1}_n$

$$\mathbf{d}_j = \begin{bmatrix} x_{1j} - \overline{\mathbf{x}}_j \\ x_{2j} - \overline{\mathbf{x}}_j \\ \cdot \\ x_{nj} - \overline{\mathbf{x}}_j \end{bmatrix} \qquad \mathbf{d}_1 = \begin{bmatrix} 3 - \overline{\mathbf{x}}_1 \\ 4 - \overline{\mathbf{x}}_1 \\ 2 - \overline{\mathbf{x}}_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \qquad \mathbf{d}_2 = \begin{bmatrix} 1 - \overline{\mathbf{x}}_1 \\ 3 - \overline{\mathbf{x}}_1 \\ 3 - \overline{\mathbf{x}}_1 \end{bmatrix} = \begin{bmatrix} -1,33 \\ 0,67 \\ 0,67 \end{bmatrix}$$

Varianza

$$s_j^2 = s_{jj} = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1} = \frac{\sum_{i=1}^{n} d_{ij}^2}{n-1} = \frac{\mathbf{d}_j^{'}\mathbf{d}_j}{n-1}$$

$$s_{x_1}^2 = s_{11} = \frac{\sum_{i=1}^{3} (x_{i1} - \overline{x}_1)^2}{3-1} = \frac{\mathbf{d}_1'\mathbf{d}_1}{n-1}$$

$$s_{11} = \frac{1}{2}\begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = \frac{1}{2}(0 + 1 + 1)$$

$$= 1$$

$$s_{x_2}^2 = s_{22} = 1,33$$

Medidas de variabilidad bivariadas

Covarianza

$$s_{jk} = s_{kj} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n-1} = \frac{\mathbf{d_j^{'} d_k}}{n-1} \qquad j,\ k = 1,\ 2,..,\ p$$

$$s_{12} = s_{12} = \frac{\sum_{i=1}^{3}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{3-1} = \frac{\mathbf{d}_1^{'}\mathbf{d}_2}{3-1}$$

$$s_{12} = s_{12} = \frac{1}{2}\begin{bmatrix} 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} -1,33 \\ 0,67 \\ 0,67 \end{bmatrix} = 0$$

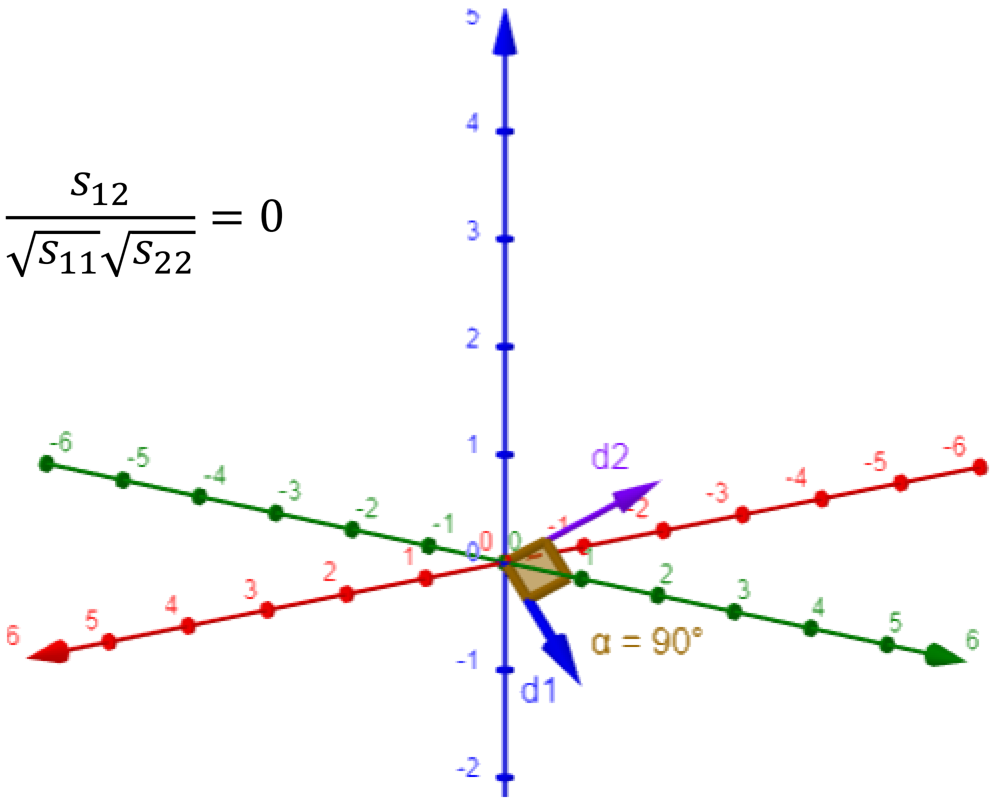
Medidas de variabilidad bivariadas

Coeficiente de correlación

$$r_{jk} = r_{kj} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \cos\theta$$

θ es el ángulo formado por los vectores $\mathbf{d}_j$ y $\mathbf{d}_k$

$$r_{12} = r_{21} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = 0$$



Matriz de datos centrados

Matriz de datos centrados

$$\tilde{\mathbf{X}}' = \mathbf{X}' - \bar{\mathbf{x}}\mathbf{1}_n' = \mathbf{X}'(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')$$

En la expresión anterior se reemplazó el vector de medias por su cálculo matricial $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{1}_n^t\,\mathbf{X}$

y resolviendo algebraicamente se obtiene la matriz P

$$\mathbf{P} = (\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n')$$

Matriz simétrica, idempotente y de rango igual a n-1 (es ortogonal al espacio definido por $\mathbf{1}_n$

$$\tilde{X}' = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 3 & 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2.3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

$$\tilde{X}' = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 3 & 3 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \right)$$

$$\tilde{X}' = \begin{bmatrix} 0 & 1 & -1 \\ -1{,}33 & 0{,}67 & 0{,}67 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

Matriz de varianzas y covarianzas

Matriz de varianzas-covarianzas muestral (S)

$$\mathbf{S} = \frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{n-1} = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'}{n-1}$$

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \cdots & \cdots & \ddots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1{,}33 \end{bmatrix}$$

Sumas de cuadrados y productos cruzados

$$\mathbf{T} = \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$$

$$\mathbf{T} = \tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \mathbf{X}'(\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}$$

$$\mathbf{T} = \tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \mathbf{X}'\mathbf{P}\mathbf{X}$$

$$\mathbf{P} = (\mathbf{I}_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n')$$

Matriz simétrica, idempotente y de rango igual a n-1 (es ortogonal al espacio definido por $\mathbf{1_n}$

Si los datos están expresados en “desvíos”con respecto a la media de cada variable, efectuando el producto **X’X** los elementos de la matriz resultante son los numeradores de varianzas y covarianzas

Matriz de correlación

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \cdots & \cdots & \ddots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

$$\mathbf{D} = diag(s_{11}\ , s_{22}\ \ldots\ldots s_{pp})$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{1}{\sqrt{1,33}} \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 1,33 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{1}{\sqrt{1,33}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

MÉTODOS FACTORIALES COMPONENTES PRINCIPALES

Componentes principales (ACP)

Se aplica cuando tenemos un número elevado de variables **cuantitativas** correlacionadas entre sí

Objetivos

- Permitir analizar la interdependencia entre las variables originales

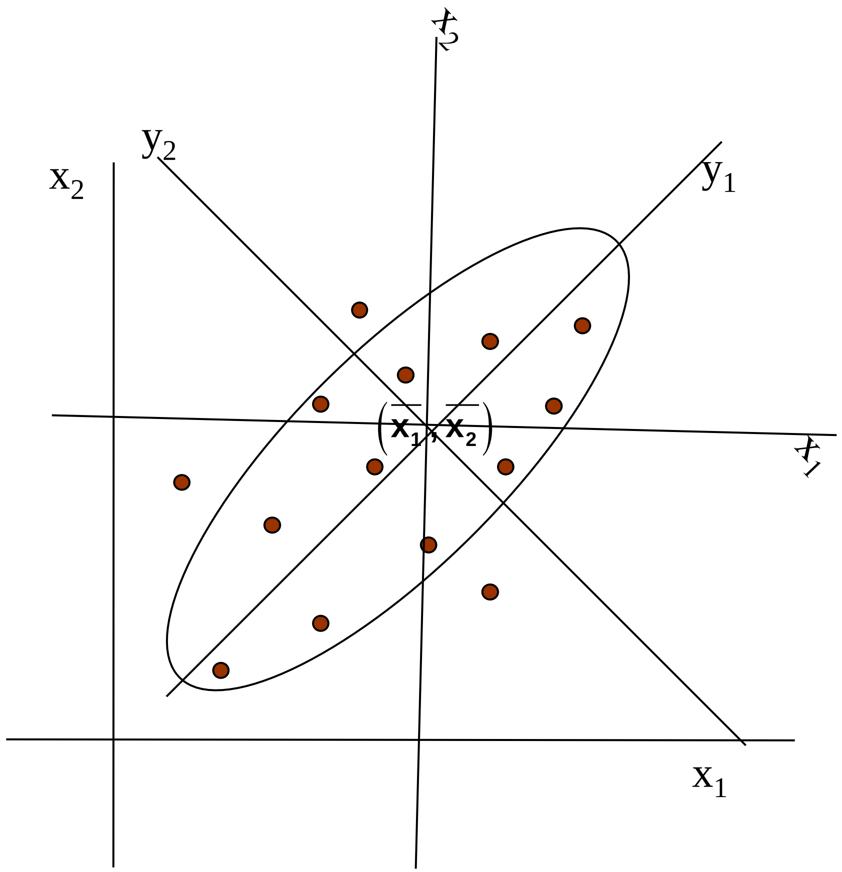
- Obtener nuevas variables llamadas **componentes principales**, que se calculan como combinación lineal de las p variables originales.

$$\mathbf{Y_i} = \mathbf{u}'\widetilde{\mathbf{X}}' \qquad \text{para i } = 1,2,3 \ldots p$$

- Utilizar pocas componentes en la aplicación de otras técnicas, tales como Cluster y regresión Múltiple.

Componentes principales

Geométricamente: Los ejes originales son transformados efectuando primero una traslación del origen al centroide, y luego una rotación que determina los nuevos ejes.



Componentes principales

$$\left.\begin{aligned} Y_1 &= u_{11} X_1 + u_{12} X_2 \\ Y_2 &= u_{21} X_1 + u_{22} X_2 \end{aligned}\right\} \quad \mathbf{Y} = \mathbf{u}\,\mathbf{X}$$

Es posible calcular tantas combinaciones lineales como variables; la primera componente principal es aquella que explica la mayor parte de la varianza de la muestra, la segunda es la que sigue en magnitud de explicación y es independiente de la primera, y así sucesivamente.

$$Var(X_1) \quad Var(X_2)$$

$$\text{Varianza total} = \sum_{i=1}^{p} Var(X_i)$$

$$Var(Y_1) = \lambda_1 \geq Var(Y_2) = \lambda_2$$

$$\text{Varianza total} = \sum_{i=1}^{p} Var(Y_i) = \sum_{j=1}^{p} \lambda_j$$

Componentes principales

- Es posible calcular tantas combinaciones lineales como variables

- El primer eje minimiza las distancias entre los puntos originales y sus proyecciones.

- La primera componente principal es aquella que explica la mayor parte de la varianza de la muestra

- La segunda es la que sigue en magnitud de explicación y es independiente de la primera, y así sucesivamente.

- Los datos transformados se proyectan sobre un nuevo conjunto de ejes ortogonales, de manera tal que las varianzas de los puntos con respecto a las nuevas direcciones estén en orden decreciente en magnitud.

$$\mathrm{var}(\mathbf{Y_1}) \geq \mathrm{var}(\mathbf{Y_2}) \geq \ldots\ldots \geq \mathrm{var}(\mathbf{Y_p})$$

Componentes principales: cálculo de los coeficientes

$$\mathbf{Y_1} = \mathbf{u_1}'\widetilde{\mathbf{X}}' \qquad \mathrm{var}(\mathbf{Y_1}) = \mathbf{u_1}'\boldsymbol{\Sigma}\mathbf{u_1}$$

Función a maximizar

$$\frac{\partial L}{\partial \mathbf{u}} = 2\boldsymbol{\Sigma}\mathbf{u} - 2\lambda\mathbf{u} = \mathbf{0} \qquad \mathbf{u_1}'\mathbf{u_1} = 1$$

$$L = \mathbf{u_1}'\boldsymbol{\Sigma}\mathbf{u_1} - \lambda(1 - \mathbf{u_1}'\mathbf{u_1})$$

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0}$$

Ecuación característica de la matriz $\boldsymbol{\Sigma}$

Los coeficientes que multiplican las variables son obtenidos a través de los vectores propios de la matriz de covarianzas o de la matriz de correlación.

$\mathrm{u}_1, \mathrm{u}_2, \dots, \mathrm{u}_p$ conjunto ortonormal de vectores $\mathrm{u}_j'\mathrm{u}_j = 1 \quad \mathrm{u}_j'\mathrm{u}_k = 0$

$\lambda_1, \lambda_2, \dots, \lambda_p$ donde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0 \quad \mathrm{var}(\mathbf{Y}_{\boldsymbol{i}}) = \lambda_i$ para $i = 1,2, \dots, p$

Componentes principales : parámetros

Los $\mathbf{u}_j$ forman la matriz $\mathbf{U}$, la que diagonaliza a Σ

$$\mathbf{U}'\Sigma\mathbf{U} = \Lambda = diag(\lambda_1, \lambda_2, \ldots\ldots\lambda_p) \qquad \mathbf{Y} = \mathbf{U}'(\mathbf{X} - \overline{\mathbf{x}})$$

$$\mathrm{Var}(Y_j) = \mathbf{u}_j'\Sigma\mathbf{u}_j = \lambda_j$$

$$\mathrm{Cov}(Y_j Y_k) = \mathbf{u}_j'\Sigma\mathbf{u}_k = 0$$

$$\begin{aligned} \mathrm{VarCov}\,\mathbf{Y} &= \mathbf{U}'\Sigma\mathbf{U} \\ &= \Lambda \\ &= diag(\lambda_1, \lambda_2, \ldots\ldots\lambda_p) \end{aligned}$$

Propiedades de las componentes

✓**Conservan la variabilidad inicial:**

$$\mathrm{tr}(\boldsymbol{\Sigma}) = \sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \lambda_j$$

✓**La proporción de varianza explicada por un componente es el cociente entre su varianza (el valor propio que corresponda) y la suma de todos los valores propios de la matriz. Para la componente h:**

$$\frac{\lambda_h}{\sum_{j=1}^{p} \lambda_j}$$

✓**Covarianza entre componentes principales y variables originales**

$$Cov(X,\ Y) = \frac{1}{n} X'Y$$

$$Cov(X,\ Y) = \frac{1}{n} X'XU$$

$$Cov(X,\ Y) = \Sigma U \quad \text{lo que es igual a} \quad Cov(X,\ Y) = DU$$

$$\text{donde} \quad D = diag(\lambda\ , \lambda_2, \ldots\ldots, \lambda_n)$$

Propiedades de las componentes

✓**Correlación entre componentes principales y variables originales**

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Entre la componente principal *l* y la variable *k*.

$$Corr(X_l, Y_k) = \frac{u_{lk}\sqrt{\lambda_k}}{\sqrt{s_{ll}}} \qquad \mathbf{u}_k = \begin{bmatrix} u_{k1} \\ u_{k2} \\ \vdots \\ u_{kp} \end{bmatrix}$$

✓**En el análisis de las componentes principales normado a partir de la matriz de correlaciones.**

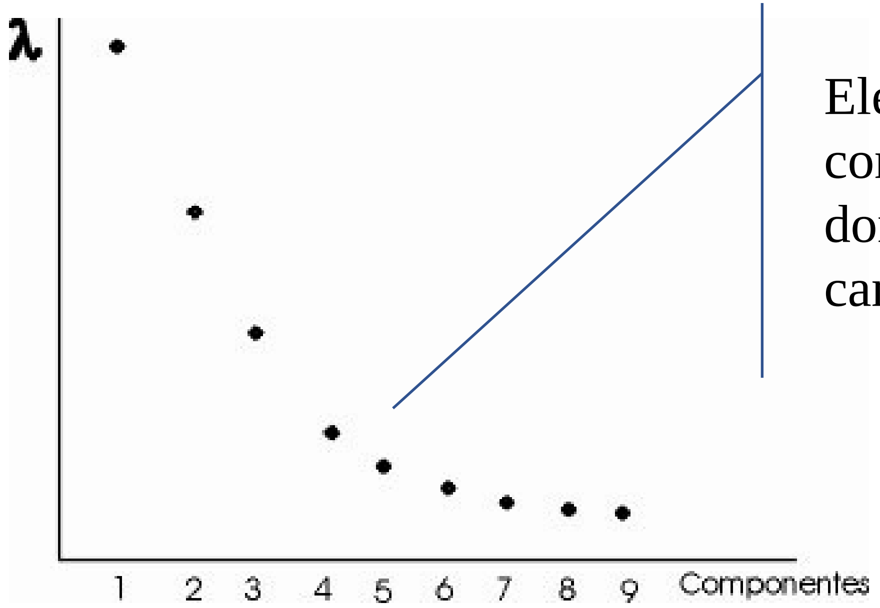
$$\mathrm{tr}(\mathrm{R}) = \sum_{j=1}^{p} \lambda_j^R = p$$

La proporción de varianza explicada por un componente h $\frac{\lambda_h^R}{p}$

La correlación con cada componente $Corr(X_l, Y_k^R) = u_{lk}\sqrt{\lambda_k}$

Selección del número de componentes

- ✓Gráfica



Elegir las componentes hasta donde empieza a cambiar la dirección

- ✓Seleccionar las componentes hasta cubrir un porcentaje determinado de la varianza P

$$\sum_{h=1}^{k} \frac{\lambda_h}{\sum_{j=1}^{p} \lambda_j} \leq P$$

- ✓ Establecer una cota, por ejemplo la varianza media $\dfrac{\sum_{j=1}^{p} \lambda_j}{p}$ con la matriz de correlación esto lleva a seleccionar los $\lambda > 1$

Gráfico de variables y observaciones en el espacio de las componentes: Biplot