Análisis descriptivo de datos multivariados

Análisis multivariado

- El estudio de “muchas” variables **correlacionadas**

- Se considera que se tiene un vector aleatorio $\mathbf{x} = (x_1, \ldots, x_p)$ y se registran n realizaciones

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- Otras notaciones

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \qquad \mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)})$$

Análisis multivariado

- (Algunos) problemas de interés
  - Graficar/describir la estructura de los datos
  - Selección de variables
  - Aprendizaje supervisado, semi-supervisado y no supervisado
  - Analizar correlación entre variables

- Retos
  - Muchas observaciones y muchas variables $(n \gg 1, p \gg 1)$
  - Más variables que observaciones $(p > n)$

Ánalisis descriptivo multivariado

Análisis descriptivo

- Medidas numéricas
  - Media muestral
  - Varianza/covarianza muestral
  - Curtosis y coeficiente de asimetría
- Gráficas
  - Diagramas de dispersión/correlación
  - Gráfica de estrellas
  - Caras de Chernoff
  - Curvas de Andrews

Estadísticas descriptivas

Media muestral

- Para la matriz $\mathbf{X}$ podemos obtener la **media muestral** para cada variable $\mathbf{x}^{(j)}$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

- Así el **vector de medias muestrales** queda definido como

$$\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)$$

- Formalmente, se define al **vector de medias** como

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

Media muestral

Proposición 1

Sea $\mathbf{X}$ una matriz de datos entonces la media muestral se puede calcular como

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}^T\mathbf{1}_n,$$

donde $\mathbf{1}_n \equiv (1,1,\ldots,1)^T$.

Observación 1

- $\mathbf{1}_n^T\mathbf{1}_n = n$
- $\mathbf{1}_n\mathbf{1}_p^T = \mathbf{J}_{n\times p}$

Media muestral

En **R** existen muchas formas de obtener el vector de medias como:

- summary( )
- apply( )
- colMeans( )
- by( ) : para la media muestral por grupos

Ejemplo: Iris

Base de datos con 150 observaciones y 5 variables:

- Sepal length (variable numérica)
- Sepal width (variable numérica)
- Petal length (variable numérica)
- Petal width (variable numérica)
- Species (variable categórica de 3 niveles : setosa, versicolor y virginica).

Ejemplo: Iris

Media muestral

Sepal length	Sepal width	Petal length	Petal width
5.843	3.057	3.758	1.199

Media muestral por grupo

Species	Sepal length	Sepal width	Petal length	Petal width
Setosa	5.006	3.428	1.462	0.246
Versicolor	5.936	2.77	4.26	1.326
Virginica	6.588	2.974	5.552	2.026

- Parece que la media para sepal length, petal length y petal width si cambia significativamente por grupos

Varianza y covarianza muestral

- **Varianza muestral** de cada variable $\mathbf{x}^{(j)}$

$$s_j^2 = s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

- **Covarianza muestral** entre $\mathbf{x}^{(j)}$ y $\mathbf{x}^{(k)}$

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

- Y así, la **matriz de covarianzas muestral**

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

Varianza y covarianza muestral

- Formalmente, se define a la matriz $\mathbf{S}$ como

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

- Considerando $\mathbf{w}_i = \mathbf{x}_i - \bar{\mathbf{x}}$

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{w}_i\mathbf{w}_i^T$$

-Podemos pensar a $\mathbf{w}_i$ como observaciones de una “nueva” matriz de datos $\mathbf{W}$

Varianza y covarianza muestral

Proposición 2

Sea $\mathbf{W}$ la matriz definida como $\mathbf{w}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ entonces

$$\mathbf{W} = \mathbf{H}_n \mathbf{X}$$

donde a la matriz $\mathbf{H}_n = \mathbf{I}_{n\times n} - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T$, se le conoce como **matriz de centrado**

Varianza y covarianza muestral

Proposición 3 (tarea)

i. $\mathbf{H}_n$ es simétrica

ii. $\mathbf{H}_n$ es idempotente

iii. $\mathbf{W} = \mathbf{H}_n\mathbf{X}$ tiene como media muestral al vector de ceros

iv. $\mathbf{S} = \dfrac{1}{n-1}\mathbf{X}^T\mathbf{H}_n\mathbf{X}$

Varianza y covarianza muestral

Proposición 4 (tarea)

Sea $\mathbf{B}$ una matriz cuadrada tal que $\mathbf{B} = \mathbf{A}^T\mathbf{A}$, donde $\mathbf{A}_{n\times p}$ entonces

i. $\mathbf{B}$ es simétrica

ii. $\mathbf{B}$ es semidefinida positiva, i.e., $\forall \alpha \in \mathbb{R}^p$ se cumple $\alpha^T\mathbf{B}\alpha \geq 0$

Proposición 5 (tarea)

La matriz de covarianza muestral $\mathbf{S}$ es semidefina positiva

Varianza y covarianza muestral

En **R** existen varias formas de encontrar la matriz de covarianzas muestral

- var ()

- cov()

- sweep( ) : para construir la matriz **W**

- by( ) : para la matriz de covarianza muestral por grupos

Ejemplo: Iris

Matriz de covarianza muestral

	Sepal length	Sepal width	Petal length	Petal width
Sepal length	0.6856935	**-0.0424340**	**1.2743154**	0.5162707
Sepal width	**-0.0424340**	0.1899794	**-0.3296564**	**-0.1216394**
Petal length	**1.2743154**	**-0.3296564**	**3.1162779**	**1.2956094**
Petal width	0.5162707	**-0.1216394**	**1.2956094**	0.5810063

- Varianza ‘grande’ en el petal length

- Covarianza positivas ‘grandes’ para sepal length con petal length y para petal length con petal width

- Presencia de covarianzas negativas

Ejemplo: Iris

Matriz de covarianza muestral por especie

Setosa	Sepal length	Sepal width	Petal length	Petal width
Sepal length	0.12424898	0.099216327	0.016355102	0.010330612
Sepal width	0.09921633	0.143689796	0.011697959	0.009297959
Petal length	0.01635510	0.011697959	0.030159184	0.006069388
Petal width	0.01033061	0.009297959	0.006069388	0.011106122

- Varianzas pequeñas

- Todas las covarianzas son positivas y pequeñas

Ejemplo: Iris

Matriz de covarianza muestral por especie

Versicolor	**Sepal length**	**Sepal width**	**Petal length**	**Petal width**
Sepal length	**0.26643265**	0.08518367	**0.18289796**	0.05577959
Sepal width	0.08518367	0.09846939	0.08265306	0.04120408
Petal length	**0.18289796**	0.08265306	**0.22081633**	0.07310204
Petal width	0.05577959	0.04120408	0.07310204	0.03910612

- Varianza de sepal length y petal length más grandes que la de sepal width y petal width

- Todas las covarianzas son positivas

- La covarianza de sepal length y petal length es la única 'grande'

Ejemplo: Iris

Matriz de covarianza muestral por especie

Virginica	**Sepal length**	**Sepal width**	**Petal length**	**Petal width**
Sepal length	**0.40434286**	0.09376327	**0.30328980**	0.04909388
Sepal width	0.09376327	0.10400408	0.07137959	0.04762857
Petal length	**0.30328980**	0.07137959	**0.30458776**	0.04882449
Petal width	0.04909388	0.04762857	0.04882449	0.07543265

- Varianza de sepal length y petal length más grandes que la de sepal width y petal width

- Todas las covarianzas son positivas

- La covarianza de sepal length y petal length es la única 'grande'

Correlación muestral

- La **correlación** entre $\mathbf{x}^{(j)}$ y $\mathbf{x}^{(k)}$

$$r_{jk} = \frac{s_{jk}}{s_j s_k}, \quad s_j = \sqrt{s_{jj}}$$

- La **matriz de correlación** está dada por

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

Correlación muestral

- Otra representación útil está dada por $\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$, donde

$$\mathbf{D} = \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_p \end{pmatrix}$$

Proposición 6 (tarea)

Sea $\mathbf{R}$ la matriz de correlación muestral entonces

i. $\mathbf{R}$ es simétrica.

ii. $\mathbf{R}$ es semidefinida positiva.

Correlación muestral

En **R** se puede calcular como

- cor ()

- by( ) - para la correlación muestral por grupos

Ejemplo: Iris

Matriz de correlación muestral

	Sepal length	Sepal width	Petal length	Petal width
Sepal length	1	-0.1175698	**0.8717538**	**0.8179411**
Sepal width	-0.1175698	1	**-0.4284401**	**-0.3661259**
Petal length	**0.8717538**	**-0.4284401**	1	**0.9628654**
Petal width	**0.8179411**	**-0.3661259**	**0.9628654**	1

- Correlación positiva fuerte entre petal length con petal width, sepal length con petal length y sepal length con petal width

- Presencia de correlaciones negativas posiblemente significativas

Ejemplo: Iris

Matriz de correlación muestral por especie

Setosa	Sepal length	Sepal width	Petal length	Petal width
Sepal length	1	**0.7425467**	0.2671758	0.2780984
Sepal width	**0.7425467**	1	0.1777000	0.2327520
Petal length	0.2671758	0.1777000	1	0.3316300
Petal width	0.2780984	0.3316300	0.006069388	1

- Todas las correlaciones son positivas

- La única correlación fuerte se da entre sepal width y sepal length

Ejemplo: Iris

Matriz de correlación muestral por especie

Versicolor	**Sepal length**	**Sepal width**	**Petal length**	**Petal width**
Sepal length	1	0.5259107	**0.7540490**	0.5464611
Sepal width	0.5259107	1	0.5605221	0.6639987
Petal length	**0.7540490**	0.5605221	1	**0.7866681**
Petal width	0.5464611	0.6639987	**0.7866681**	1

- Todas las correlaciones son positivas y posiblemente significativas

- Correlación de petal length con sepal length y petal length con petal width las más fuertes

Ejemplo: Iris

Matriz de correlación muestral por especie

Virginica	**Sepal length**	**Sepal width**	**Petal length**	**Petal width**
Sepal length	1	0.4572278	**0.8642247**	0.2811077
Sepal width	0.4572278	1	0.4010446	0.5377280
Petal length	**0.8642247**	0.4010446	1	0.3221082
Petal width	0.2811077	0.5377280	0.3221082	1

- Todas las correlaciones son positivas

- Correlación de petal length con sepal length es la más significativa

Propiedades de los estimadores

Proposición 7

Suponer que se tienen n observaciones independientes de un vector aleatorio $\mathbf{x}$ tal que $\mathbb{E}(\mathbf{x}) = \mu$ y $\mathrm{Var}(\mathbf{x}) = \Sigma$ entonces

i. $\mathbb{E}(\bar{\mathbf{x}}) = \mu$

ii. $\mathrm{Var}(\bar{\mathbf{x}}) = \dfrac{1}{n}\Sigma$

iii. $\bar{\mathbf{x}}$ es consistente, esto es, $\mathbb{P}(\,|\,|\,\bar{\mathbf{x}} - \mu\,|\,| < \epsilon) \to 1$ para todo $\epsilon > 0$

iv. $\mathbb{E}(\mathbf{S}) = \Sigma$

Gráficas

Gráficas de dispersión y correlación

Diagrama de dispersión

- Graficar todas las variables contra todas las variables

- Útil para:

  ‣ Observar la relación por pares entre las variables

  ‣ Identificar el tipo de correlación por pares entre ellas

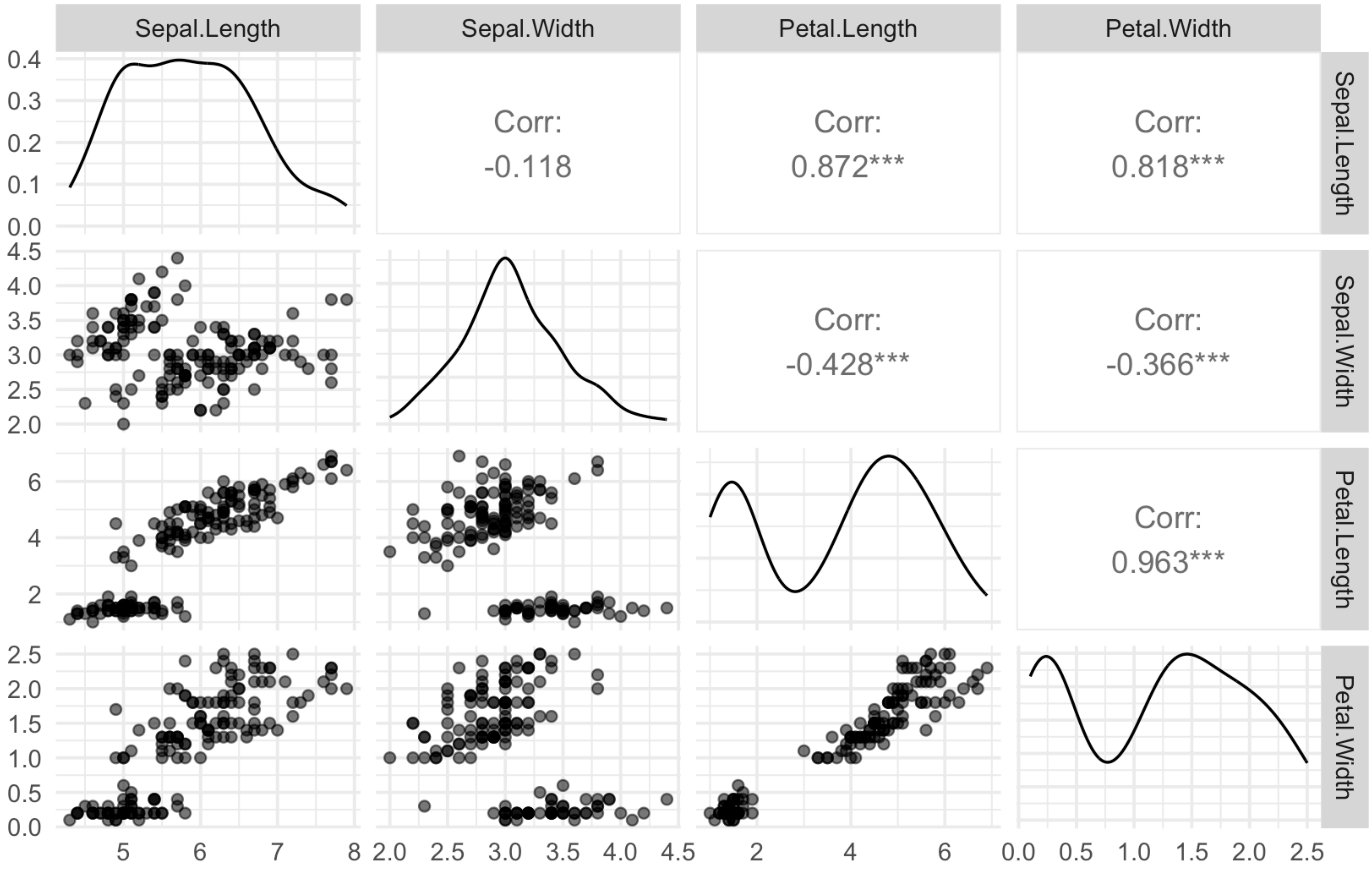
- Desventajas:

  ‣ Solo se puede analizar a las variables por pares

  ‣ Muy difícil de graficar/analizar si se tienen muchas variables

- En R: librería GGally (ggplot2)

Ejemplo: Iris (datos no agrupados)



Ejemplo: Iris (datos agrupados)

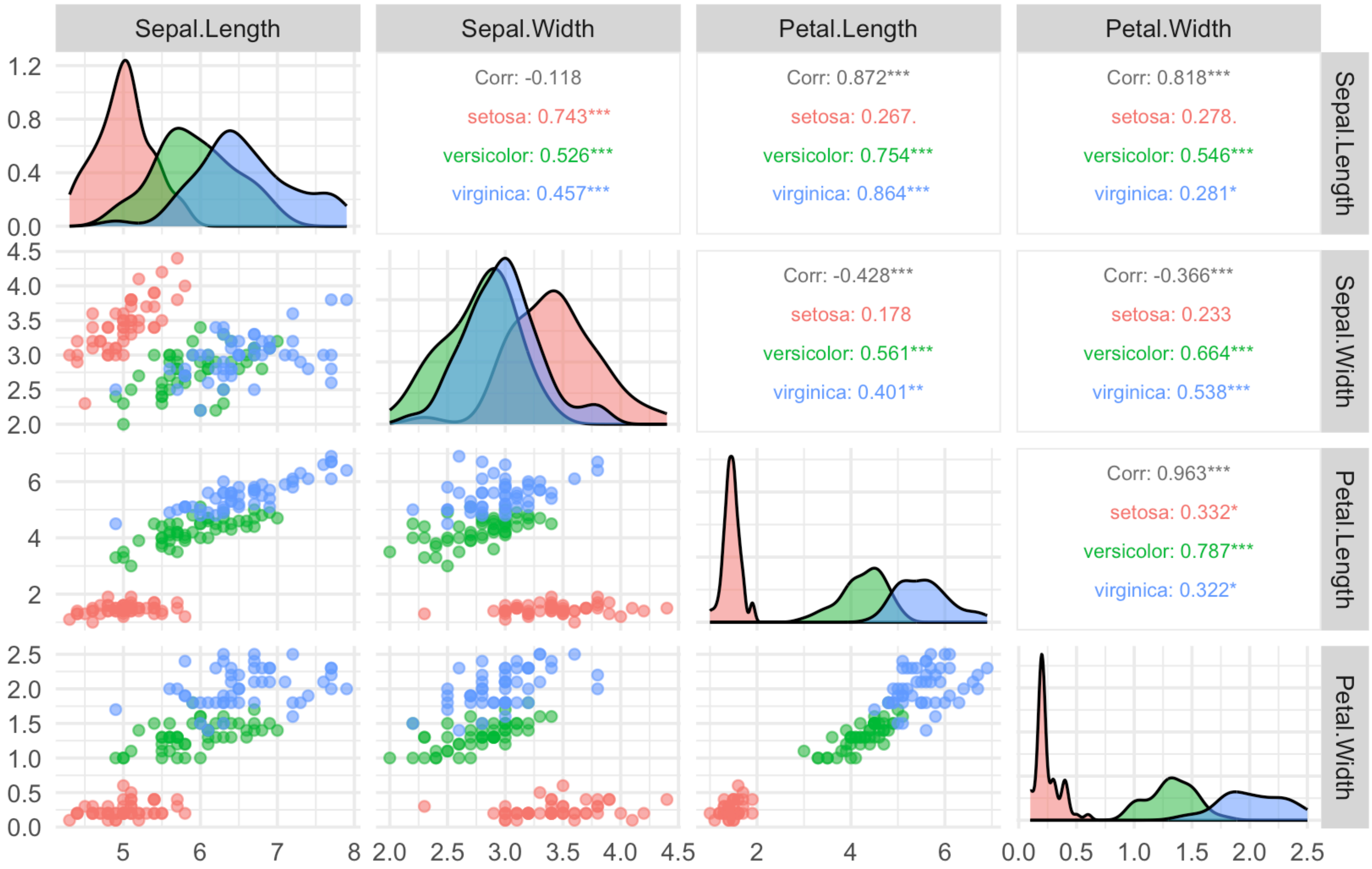


Diagrama de correlación

- Graficar la correlación por pares de las variables

- Útil para:

    - Identificar el tipo y el grado de correlación por pares entre ellas

- Desventajas:

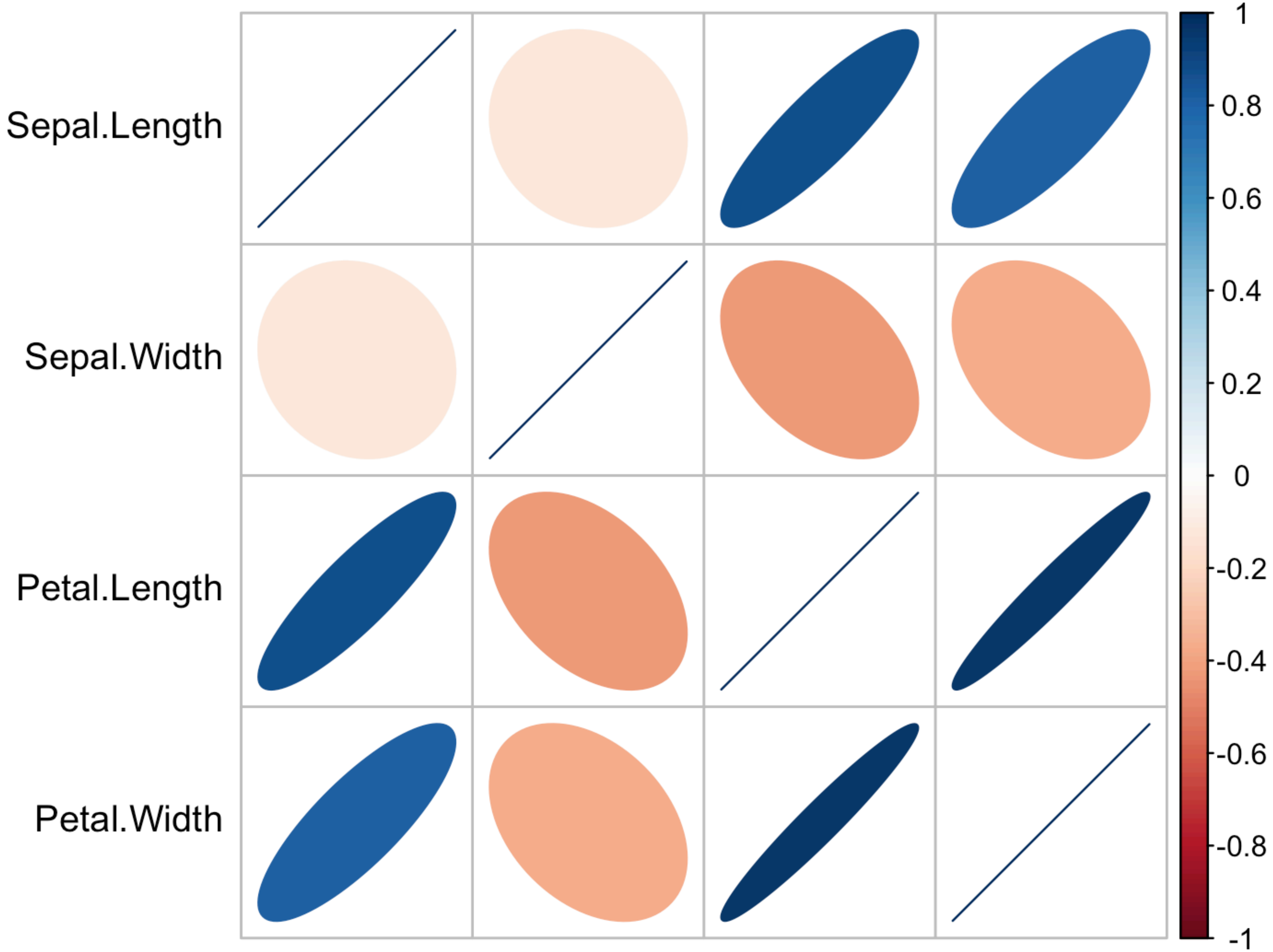
    - Solo se puede analizar a las variables por pares

    - Muy difícil de graficar/analizar si se tienen muchas variables

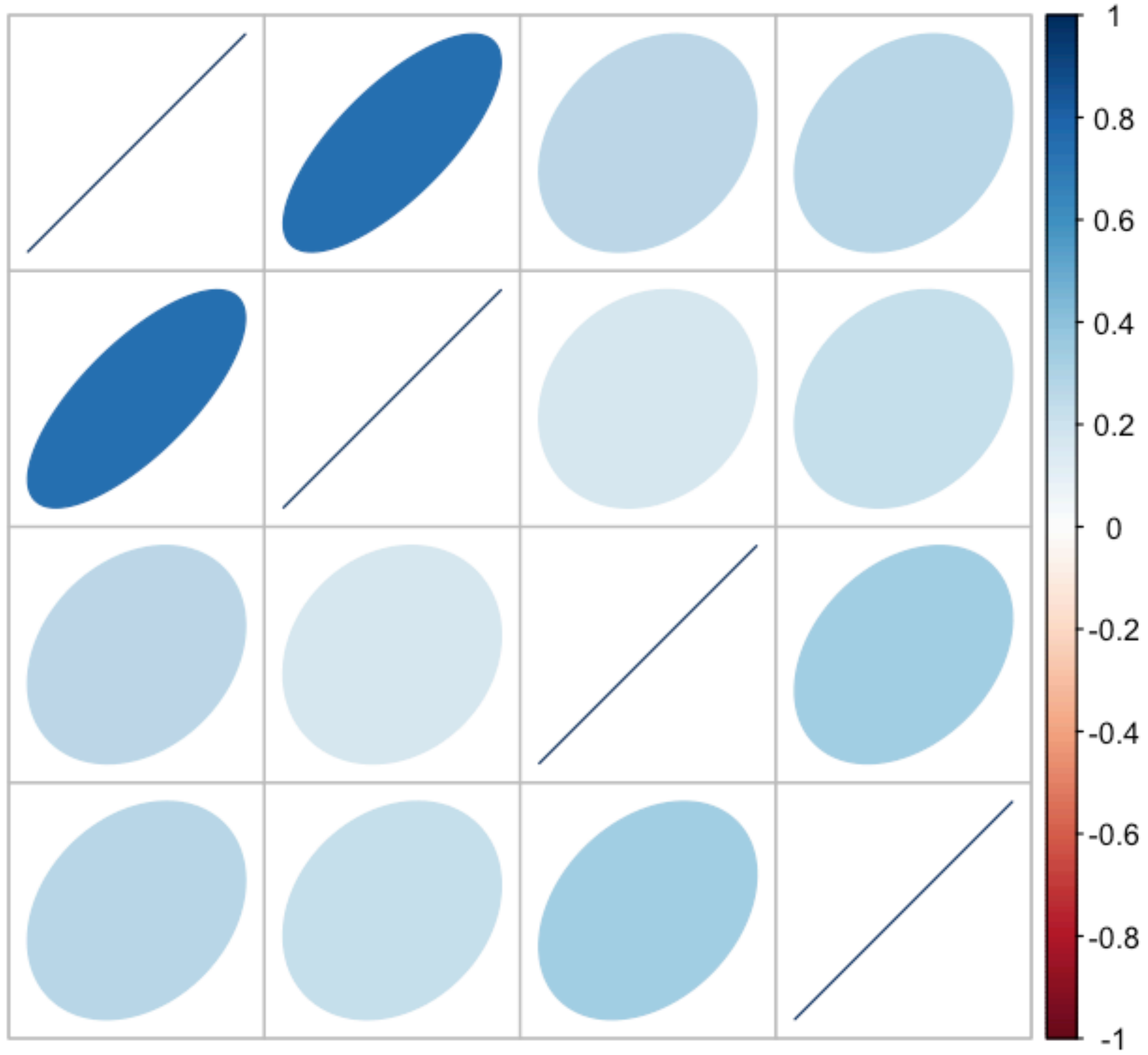
-En R:

    - Librería: corrplot

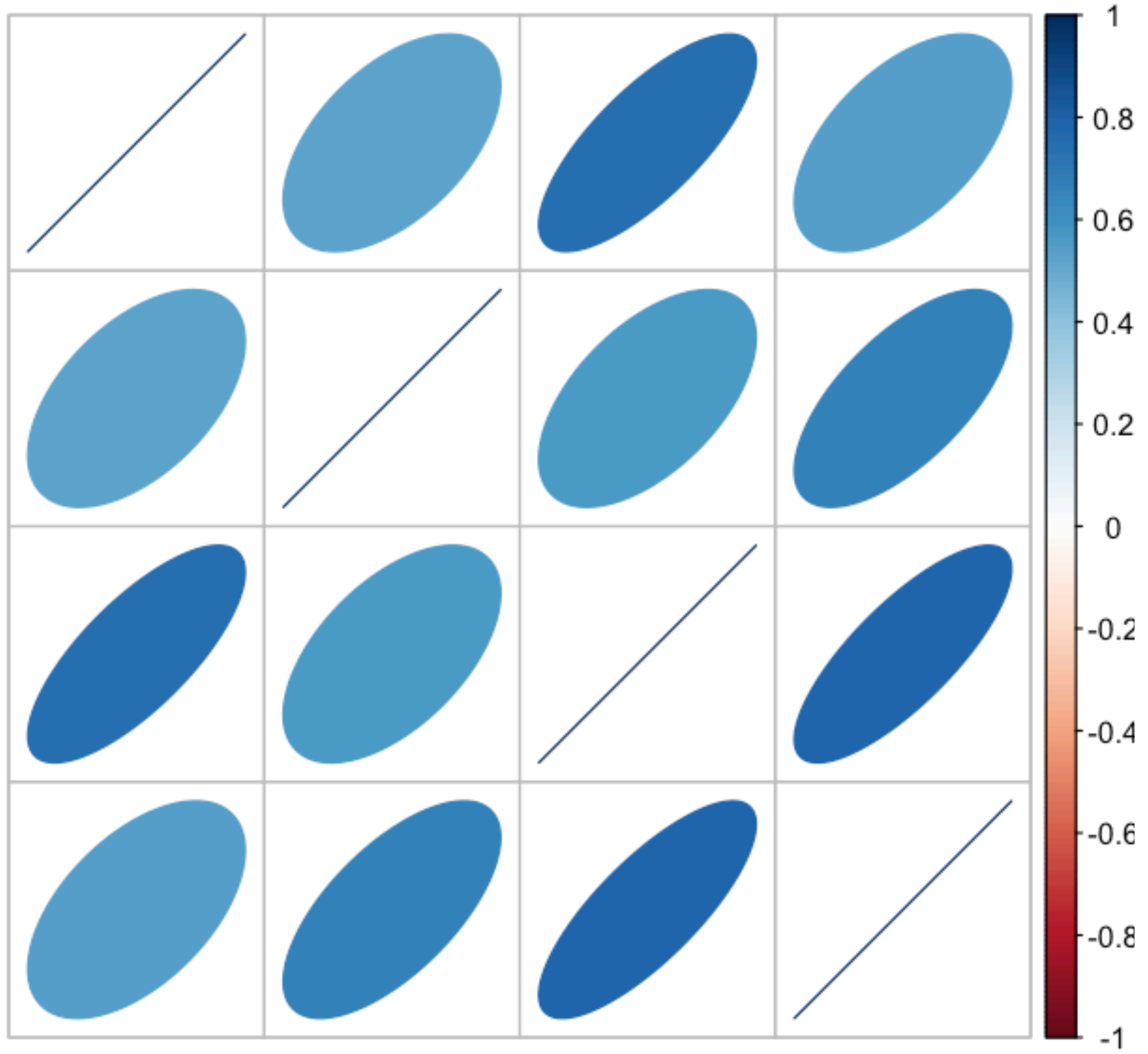
Ejemplo: Iris (datos no agrupados)



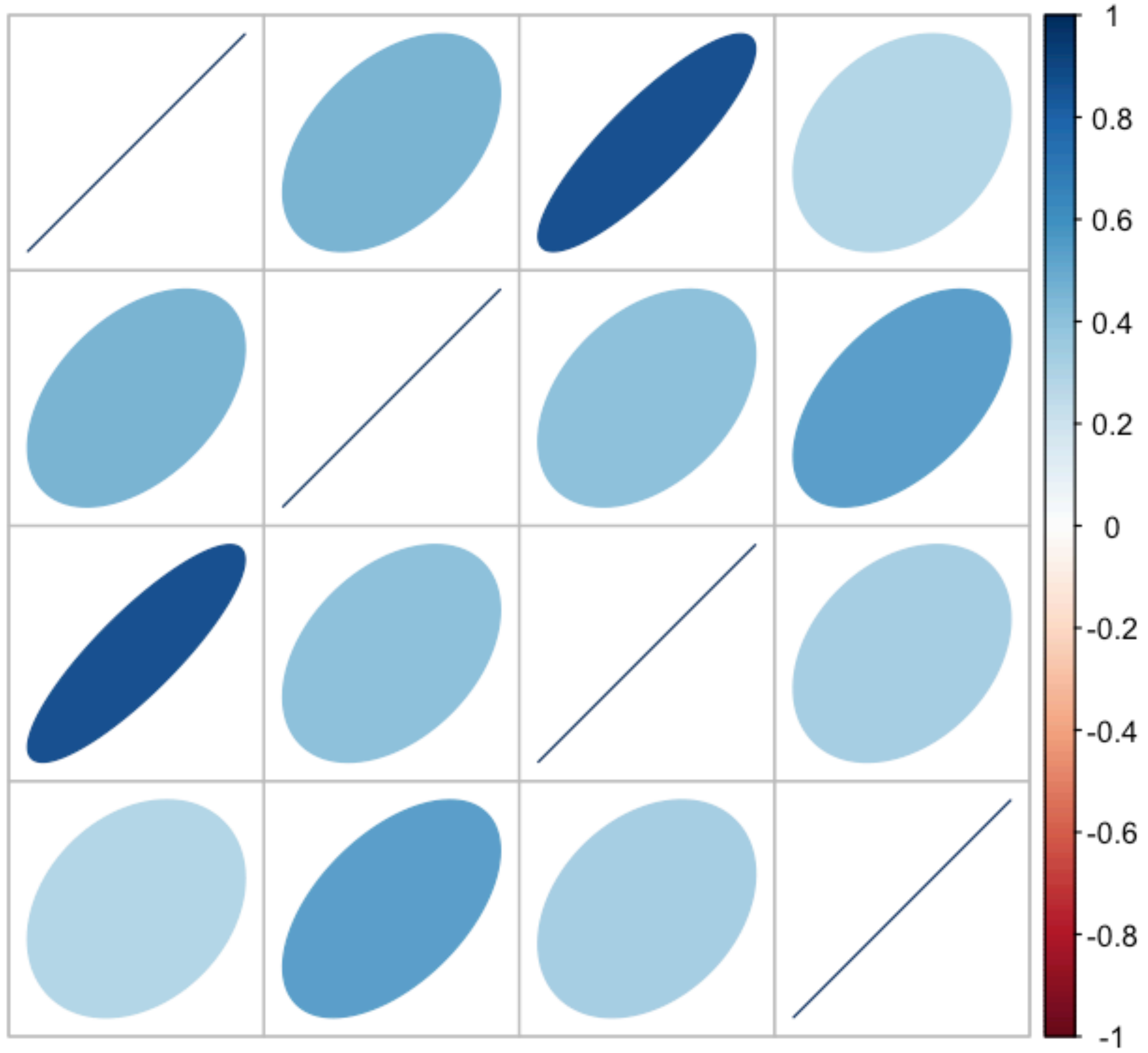
Ejemplo: Iris (datos agrupados)



Setosa



Versicolor



Virginica

Estrellas

Estrellas

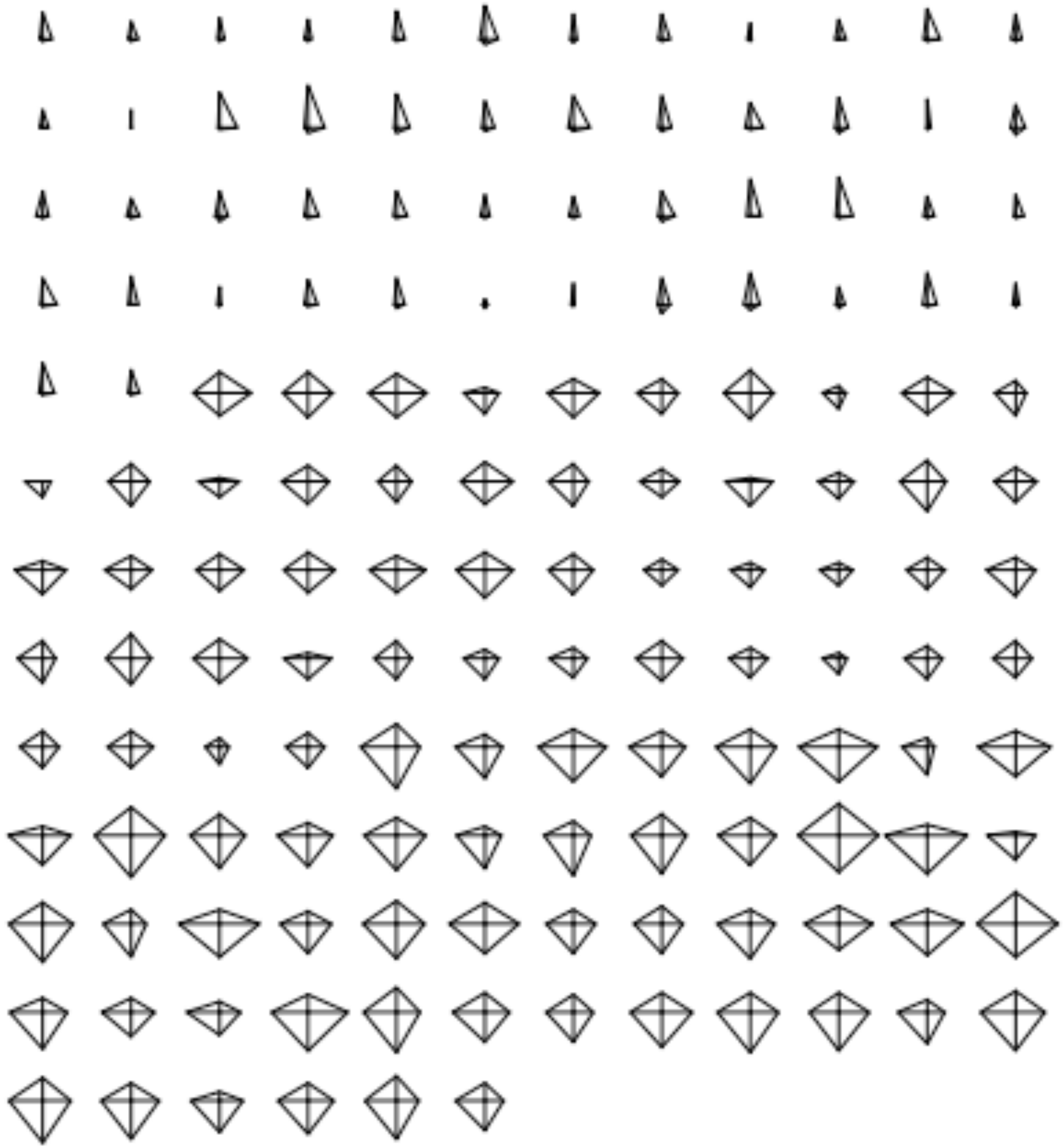
- Técnica para graficar datos multivariados en 2D (**escalados a** $[0,1]$)

- Se forma una "estrella" con p picos por cada una de las n observaciones

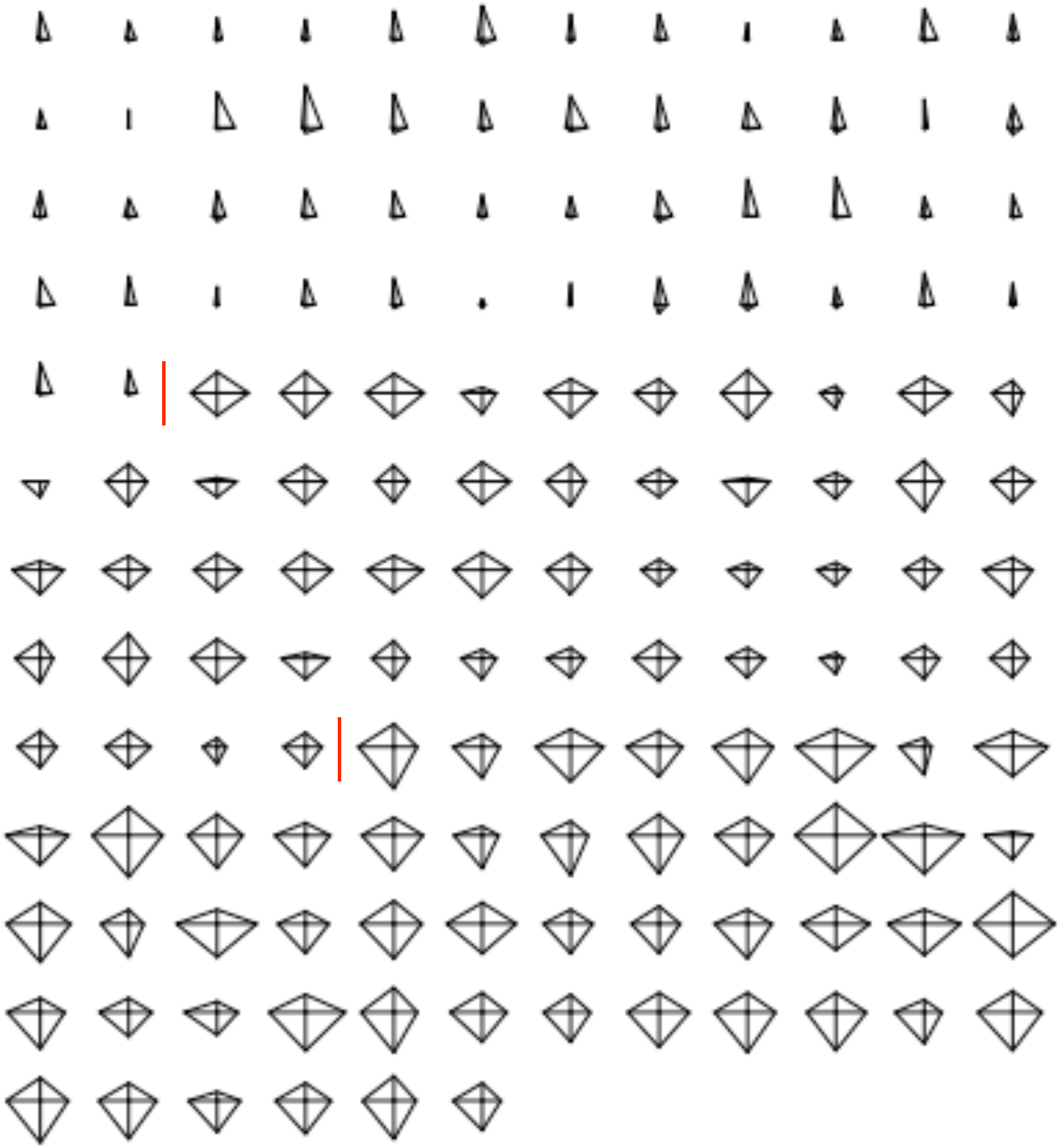
- Útil para:
  - Identificar clusters, outliers y variables "importantes"

- Desventajas:
  - Complicado de analizar si hay muchas observaciones y/o muchas variables

Estrellas



Estrellas



Caras de Chernoff

Caras de Chernoff

- Técnica similar a las estrellas para graficar datos multivariados (**escalados a** $[0,1]$)

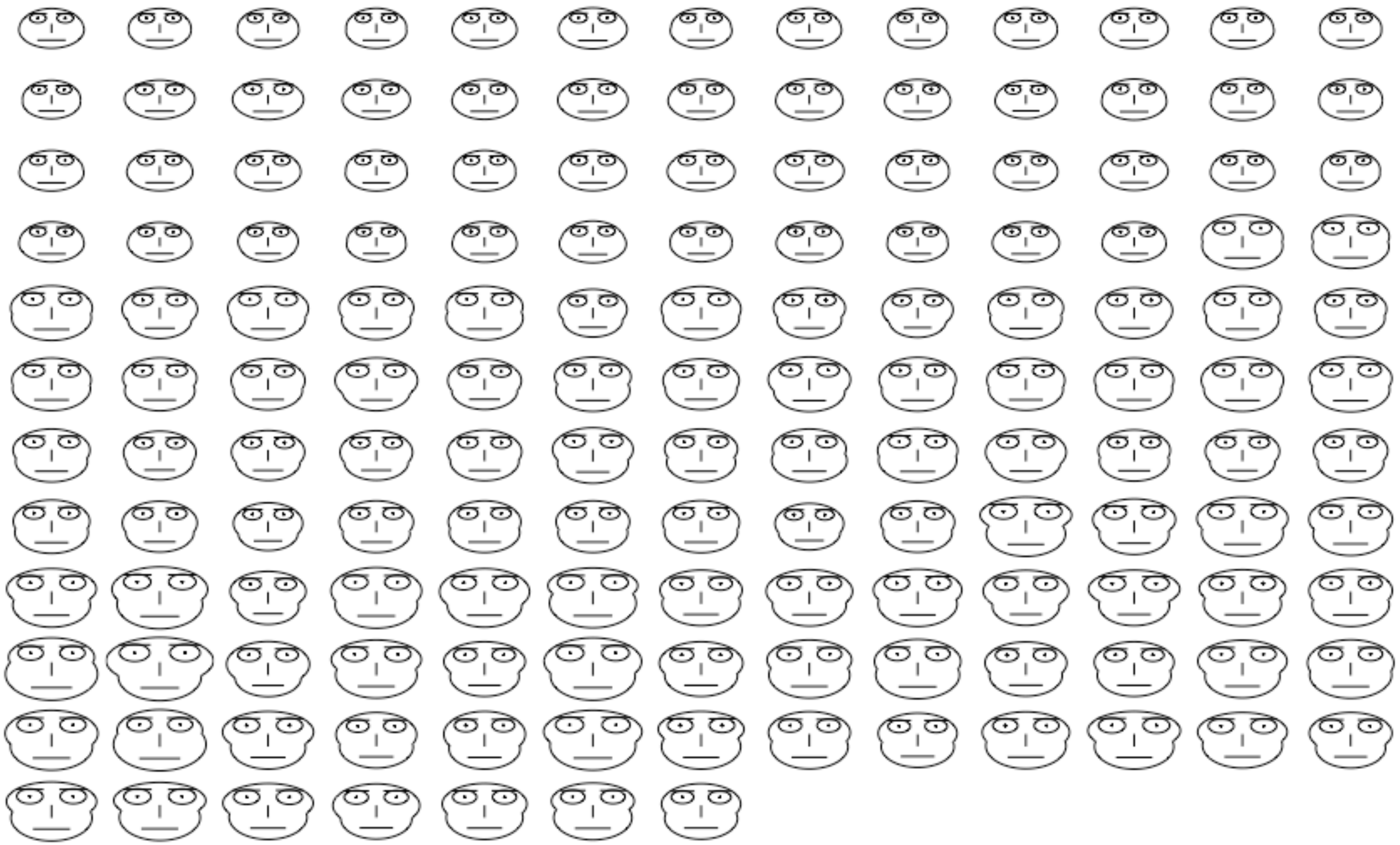
- Desarrollado por Chernoff, Herman (1973). **The use of Faces to Represent Points in K-Dimensional Space Graphically**

- Útil para:
    - Identificar rápidamente clusters, outliers y variables importantes

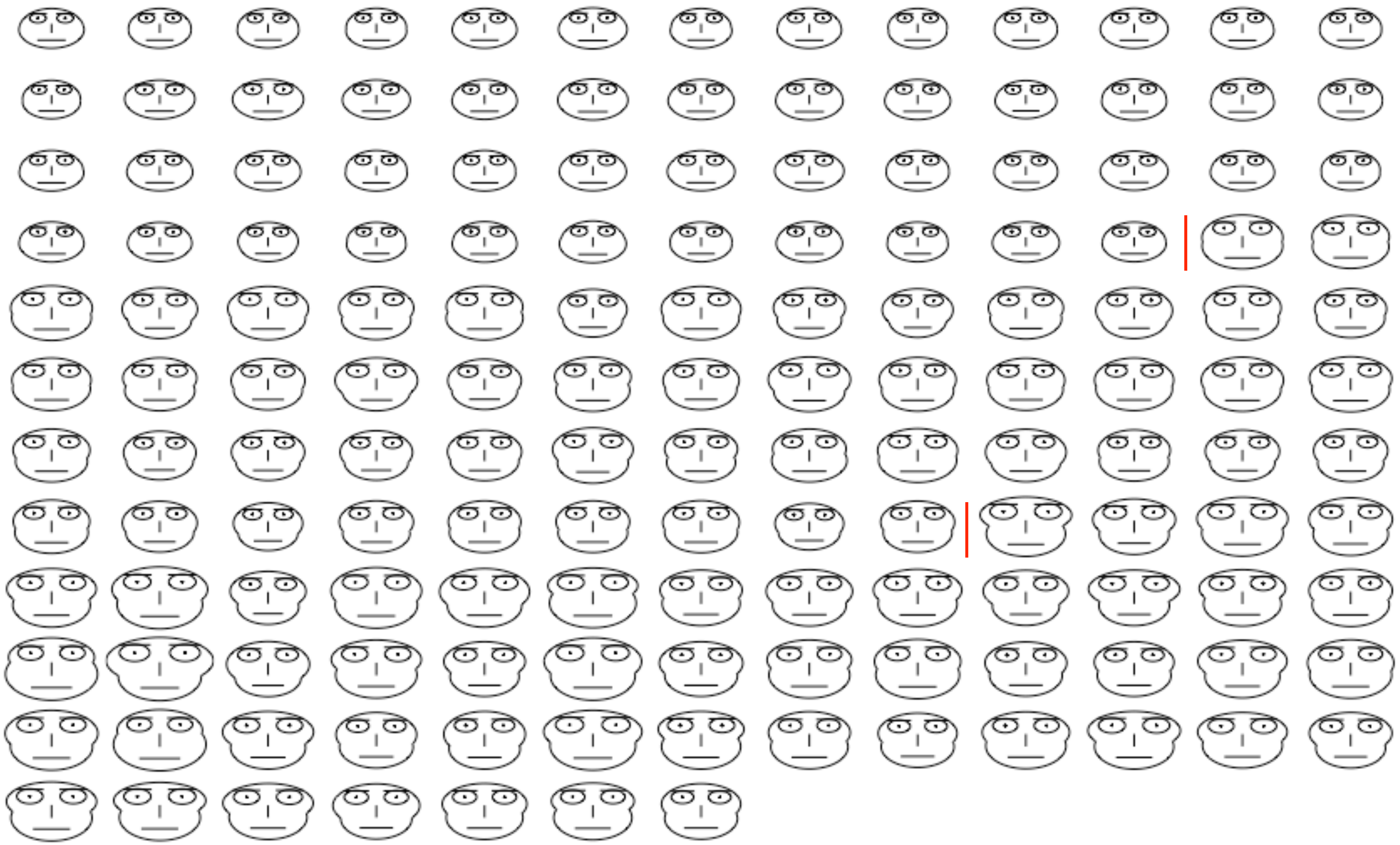
- Desventajas:
    - Limitado a $p \leq 18$
    - El orden de las variables importa

- En R: Librería TeachingDemos

Caras de Chernoff



Caras de Chernoff



Caras de Chernoff

- El orden de las variables es importante

Curvas de Andrews

Curvas de Andrews

- Transformación para graficar datos multivariados en el plano cartesiano (o coordenadas polares)

- Desarrollado por Andrews, D.F. (1972). **Plots of High-Dimensional Data.**

- Cada punto $\mathbf{x} = (x_1, \ldots, x_p)$ es mapeado a

$$f_{\mathbf{x}}(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \cdots, \qquad -\pi < t < \pi$$

- (Algunas) Propiedades útiles (tarea):

Preserva medias, i.e., $$f_{\bar{x}}(t) = \frac{1}{n} \sum_{i=1}^{n} f_{x_i}(t)$$

Preserva distancias, i.e., $$||f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)||_{L_2} = \int_{-\pi}^{\pi} [f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)]^2 dt = \pi ||\mathbf{x} - \mathbf{y}||^2$$

Curvas de Andrews

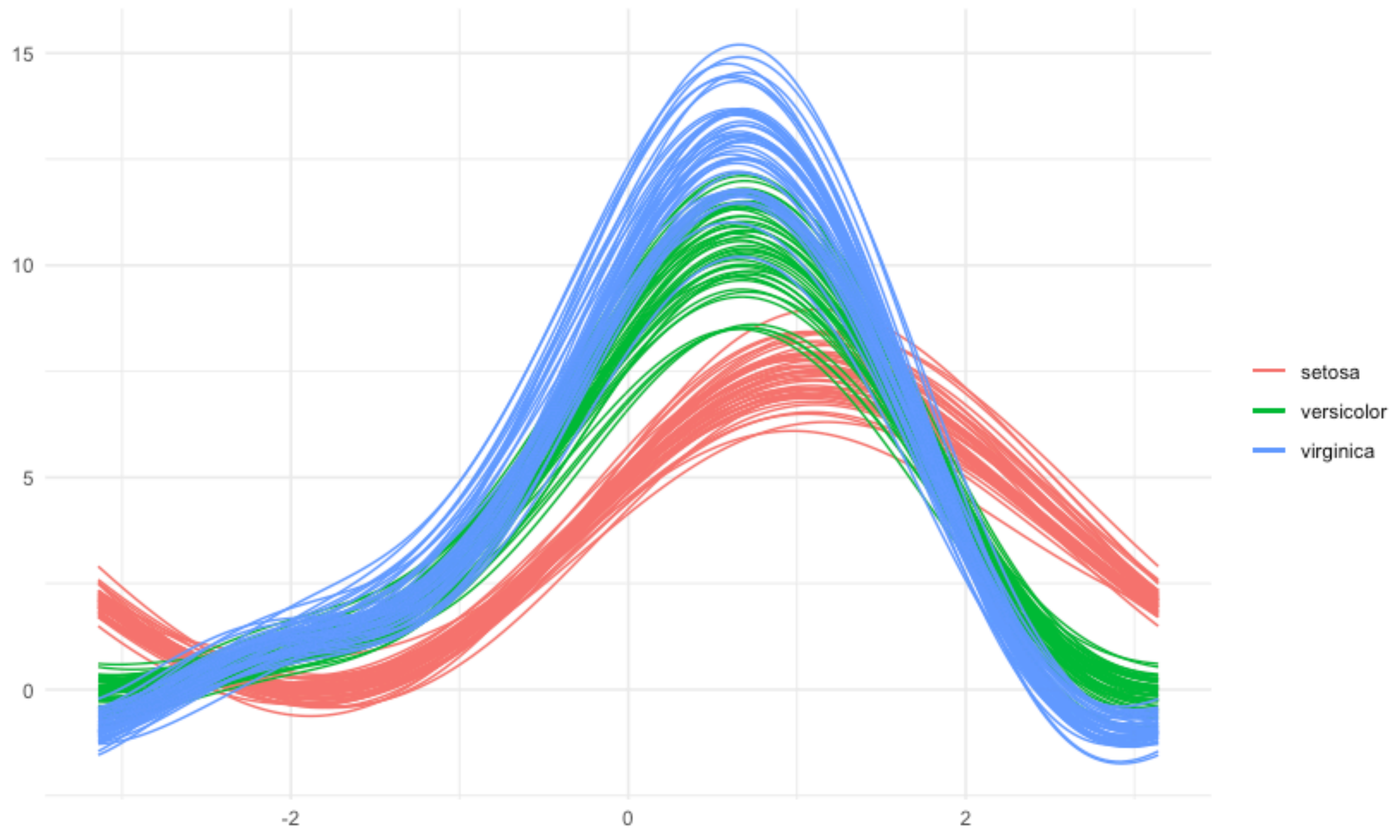
-Ventajas

- No hay restricciones en el número de variables ni de observaciones.
- Detección de outliers y clusters
- No requiere datos escalados

-Desventajas

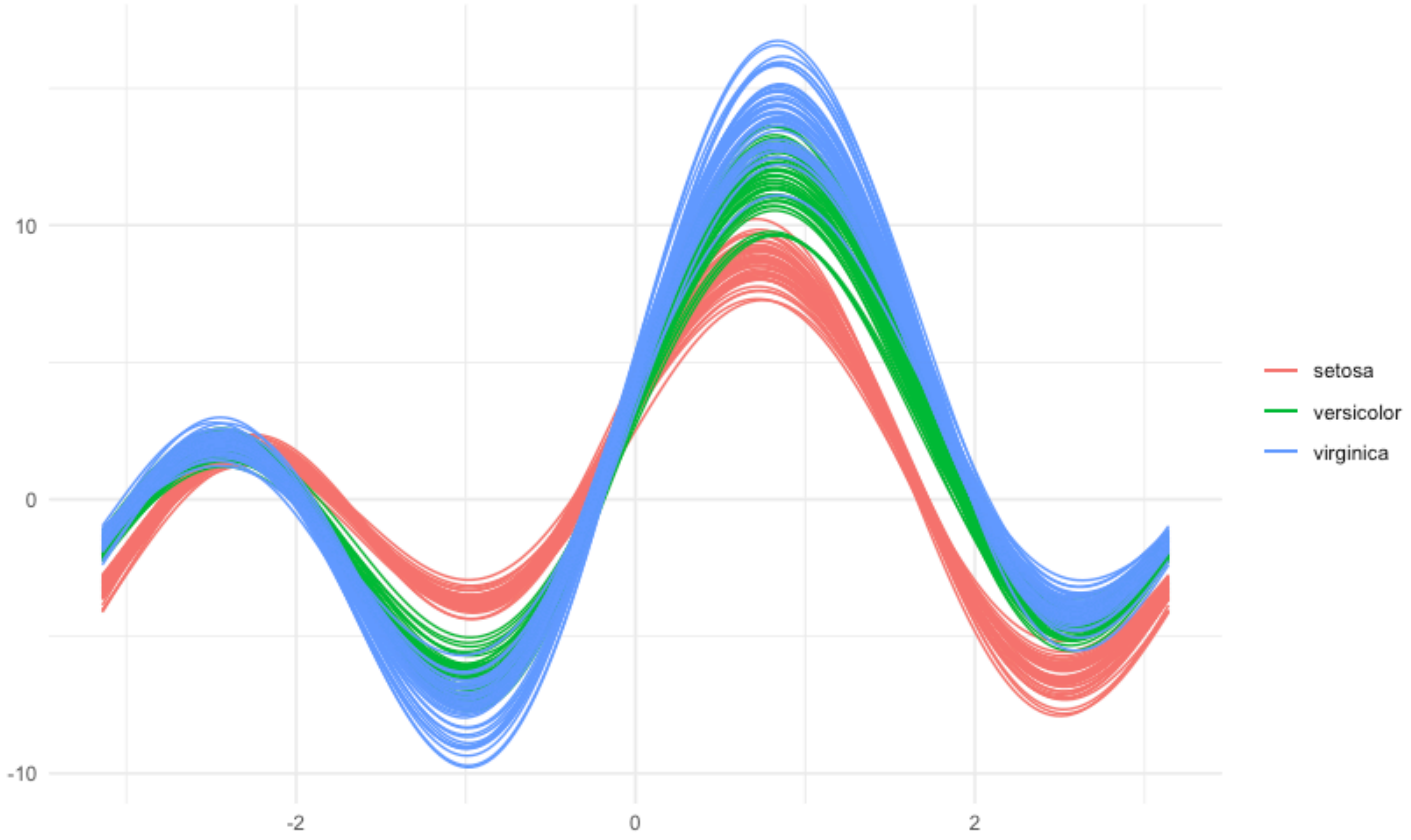
- El orden de las variables importa
- Mayor peso a las primeras variables.

Curvas de Andrews



Curvas de Andrews

- El orden de las variables es importante



Curvas de Andrews

- Otros posibles mapeos

  - Andrews, 1972

$$f_{\mathbf{x}}(t) = x_1 \sin(n_1 t) + x_2 \cos(n_1 t) + x_3 \sin(n_2 t) + x_4 \cos(n_2 t) + \cdots, \quad n_i \in \mathbb{N}, \quad -\pi \leq t \leq \pi$$

$$f_{\mathbf{x}}(t) = x_1 \sin(2t) + x_2 \cos(2t) + x_3 \sin(4t) + x_4 \cos(4t) + \cdots, \qquad 0 \leq t \leq \pi$$

  - Khattree, R. & Naik, D. (2002). **Andrews plots for multivariate data: some new suggestions and applications.** Para $-\pi \leq t \leq \pi$

$$f_{\mathbf{x}}(t) = \frac{1}{\sqrt{2}} \left[x_1 + x_2(\sin(t) + \cos(t)) + x_3(\sin(t) - \cos(t)) + x_4(\sin(2t) + cos(2t)) + \ldots\right]$$

- En R: Librería pracma implementa la función definida por Khattree pero con $0 \leq t \leq 2\pi$

Curvas de Andrews (librería pracma)