Análisis Descriptivo de Datos Multivariados

José A. Perusquía Cortés

Análisis Multivariado Semestre 2024-1

¿Qué es el análisis multivariado?

&

¿Qué tipo de datos nos interesan?

- El estudio de "muchas" variables correlacionadas.

- Para n variables $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n} \in S^p$, i.e., $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ se tiene la notación

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \qquad \mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)})$$

- Problemas de interés
  - Graficar/describir la estructura de los datos
  - Selección de variables
  - Aprendizaje supervisado, semi-supervisado y no supervisado
  - Analizar correlación entre variables
  - Etc…

- Retos
  - $n >> 1$, $p >> 1$
  - $p > n$

- En R generalmente representados a través de data frames

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

- ¿Cómo podemos visualizar los datos?

Diagramas de Dispersión y de 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

Diagrama de Dispersión

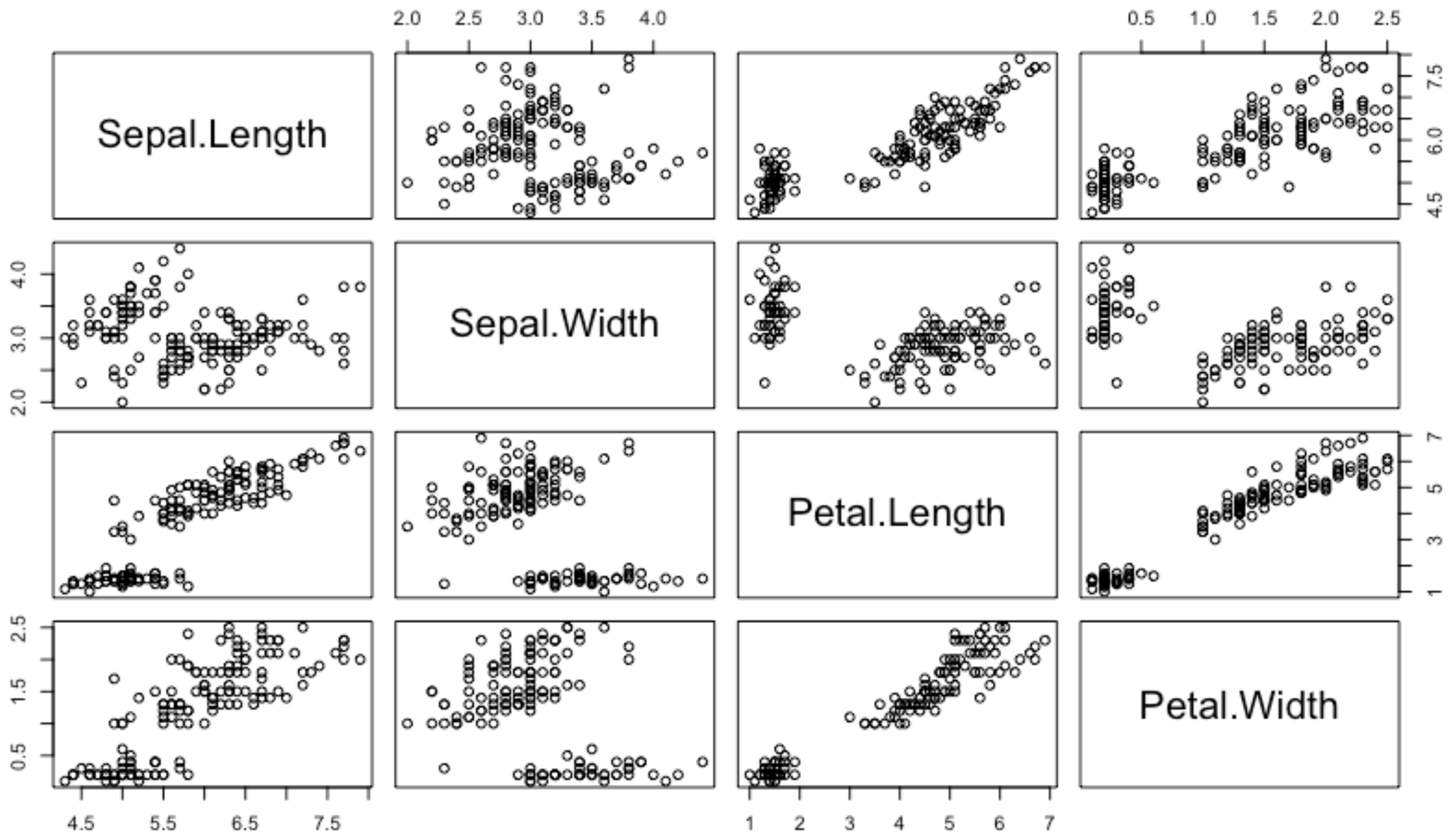


Diagrama de Correlación

- 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

Diagrama de Correlación

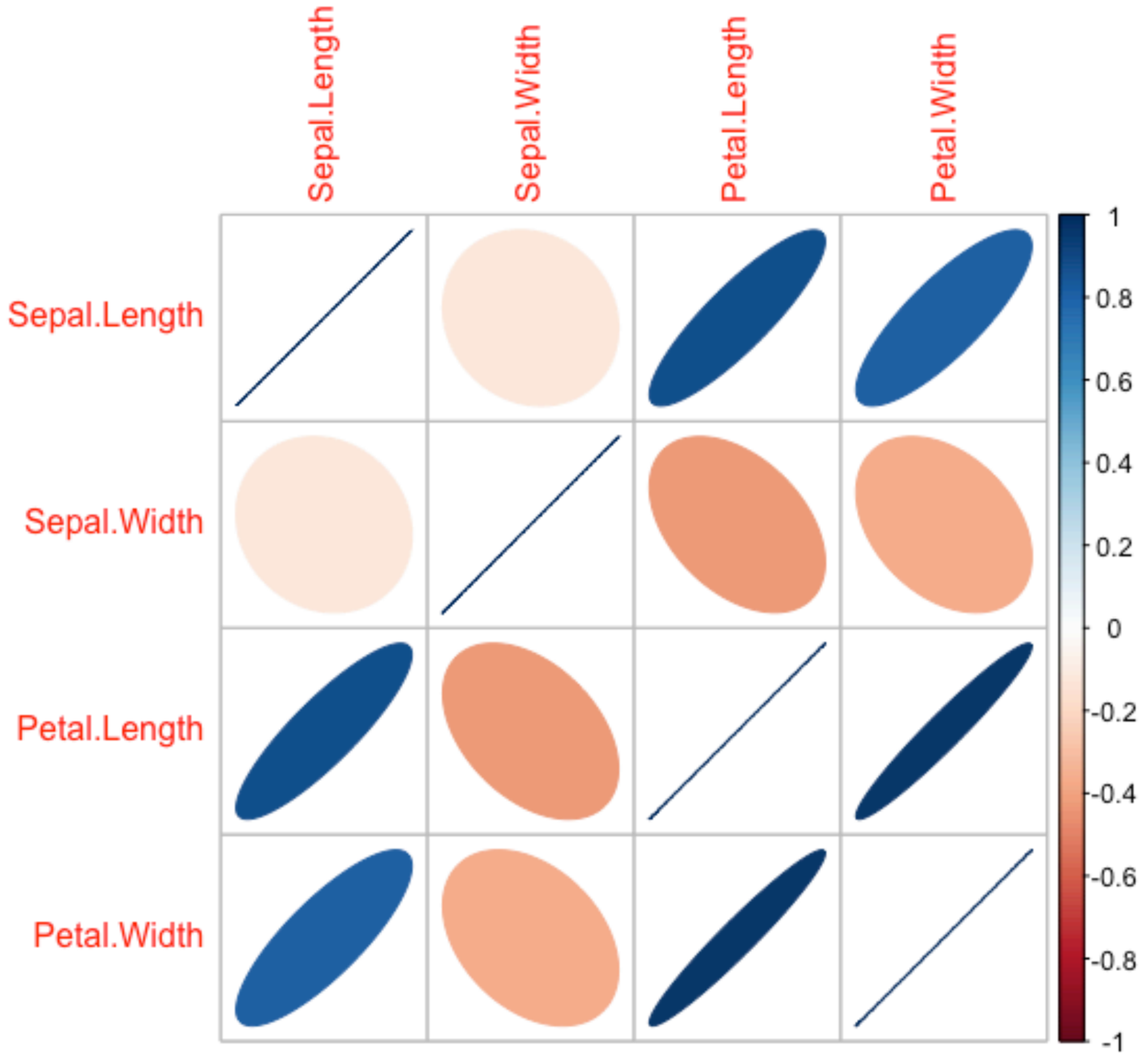


Diagrama de Dispersión II

- ¿Qué sucede si se utiliza la información de la especie?

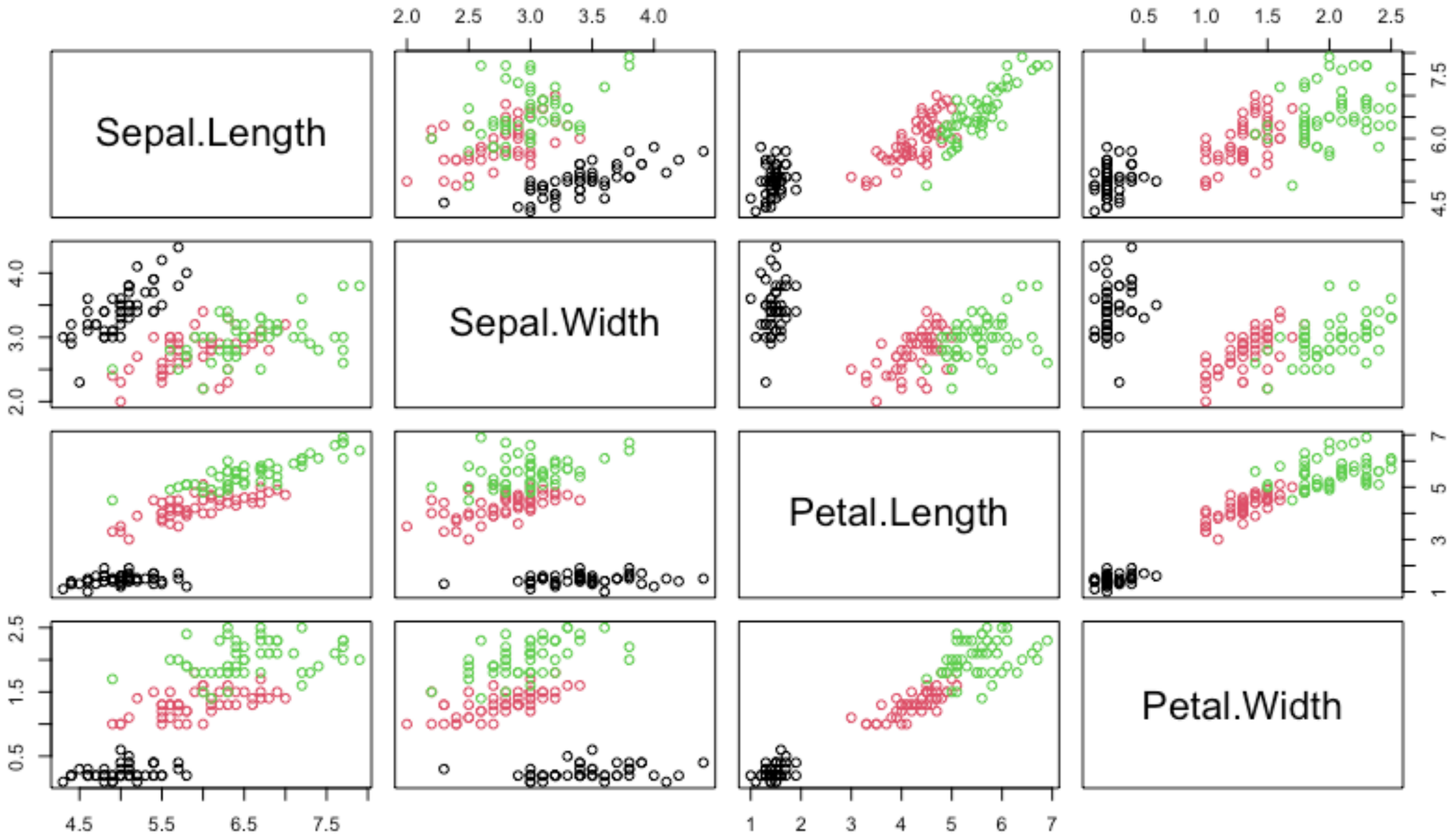
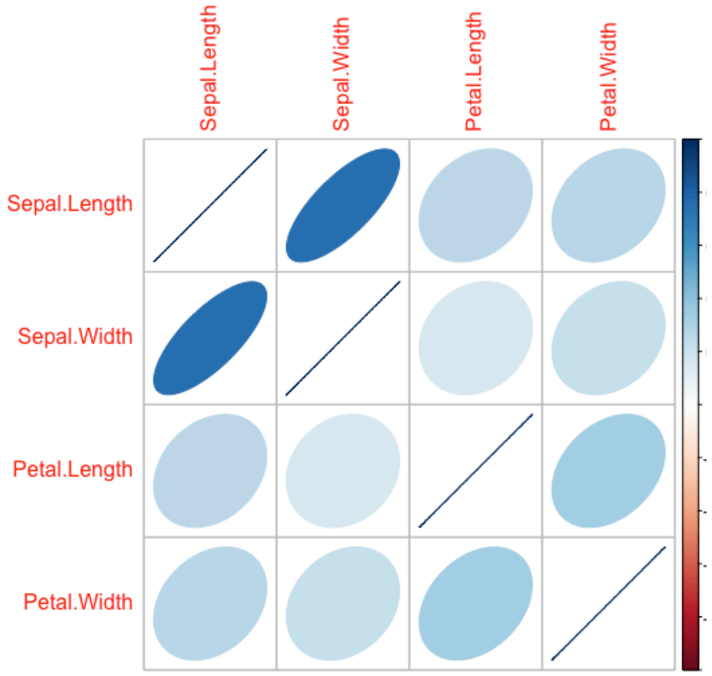
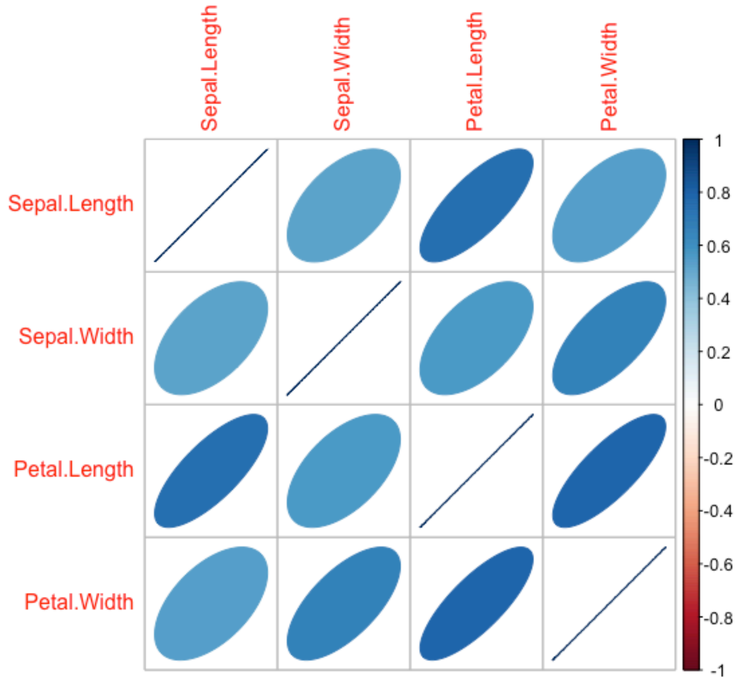


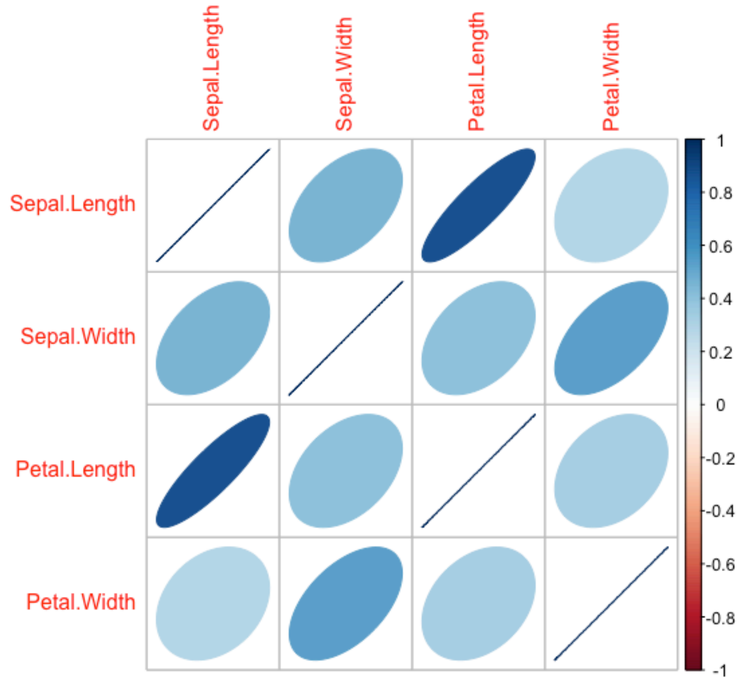
Diagrama de Correlación



Setosa



Versicolor



Virginia

Diagrama de Dispersión III

- Gráficas de R no son muy estéticas

- Podemos explotar las bondades de ggplot2 para crear gráficas más estéticas e ilustrativas

- Para diagramas de dispersión y correlación:
  - Librería: GGally

Diagrama de Dispersión III

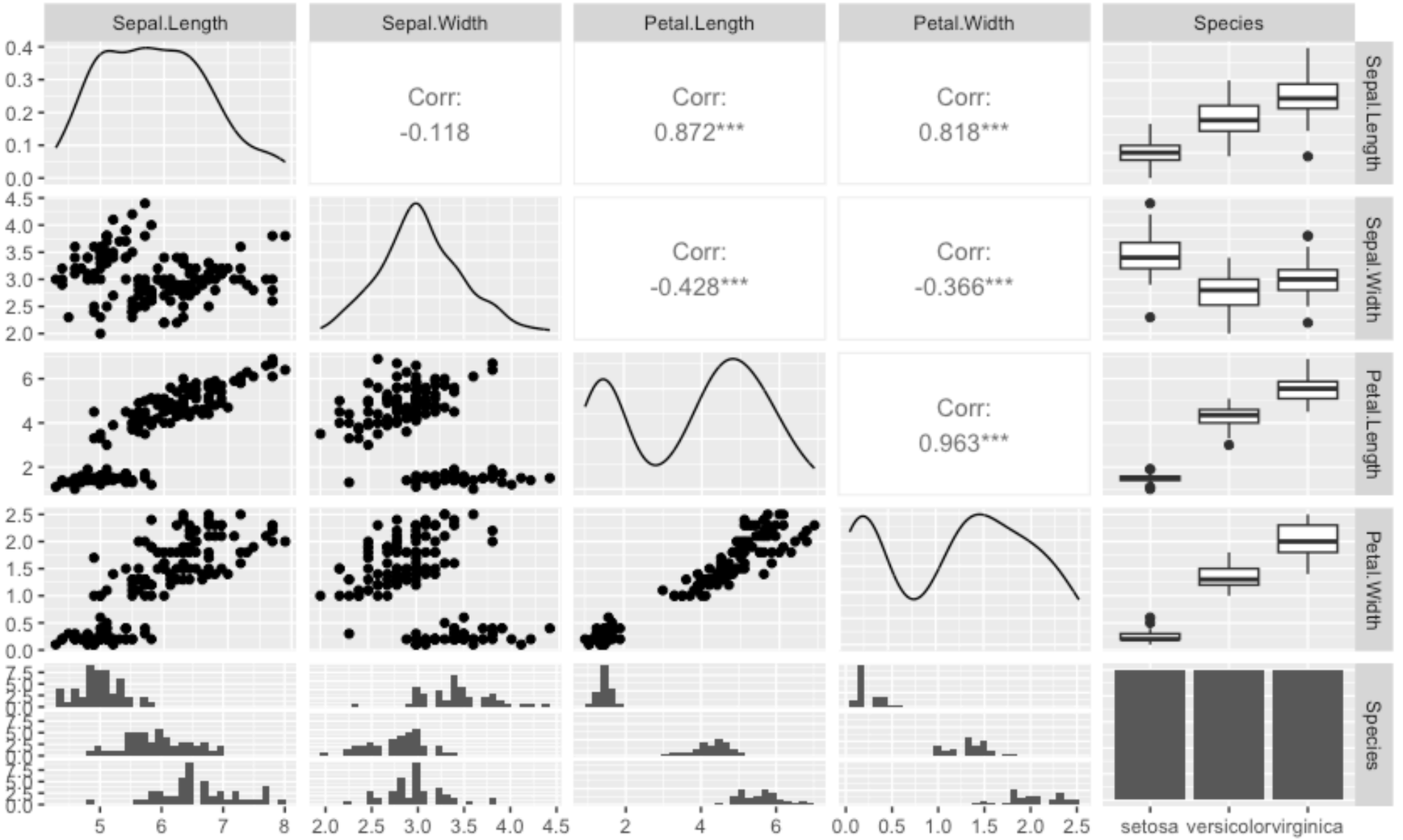


Diagrama de Dispersión III

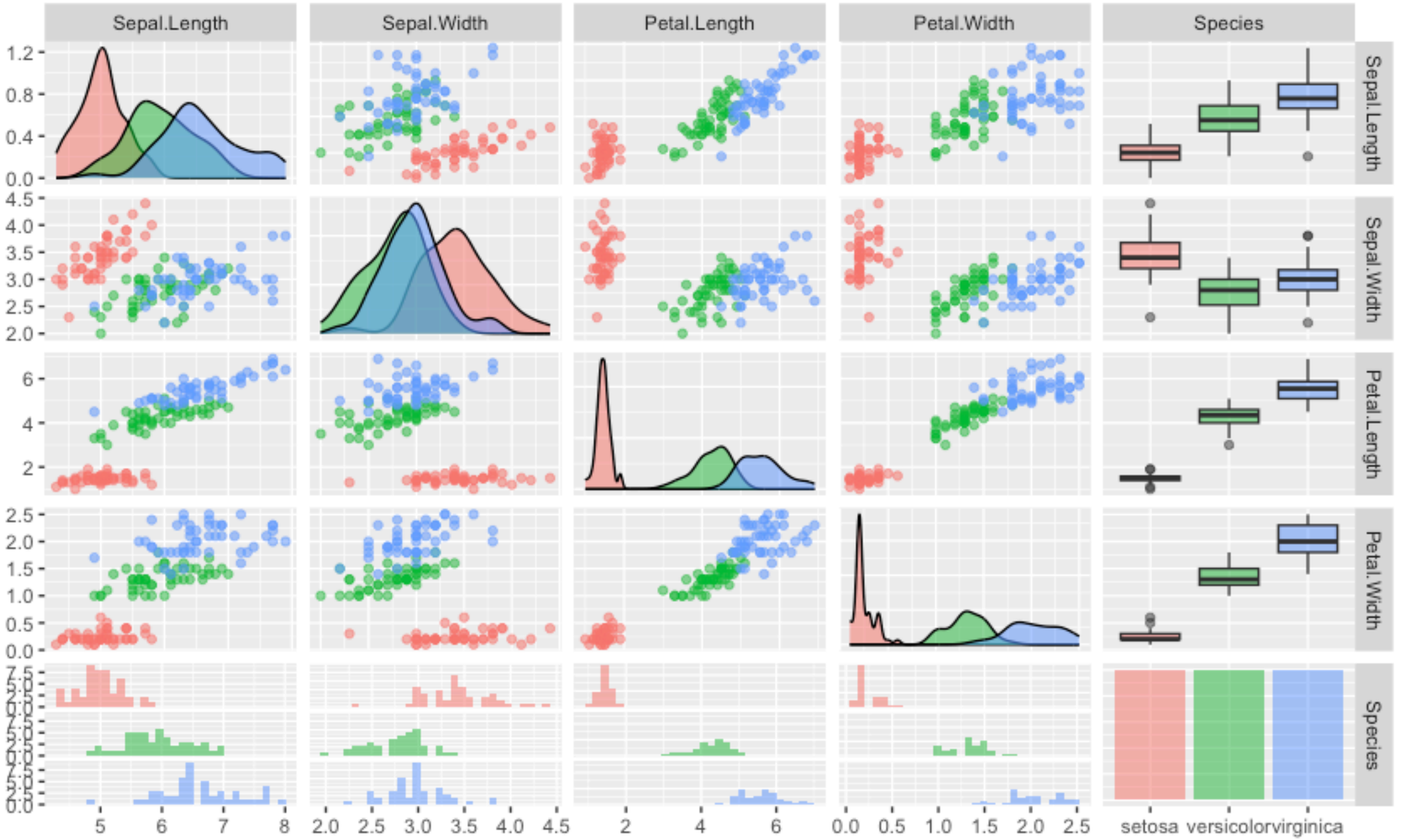
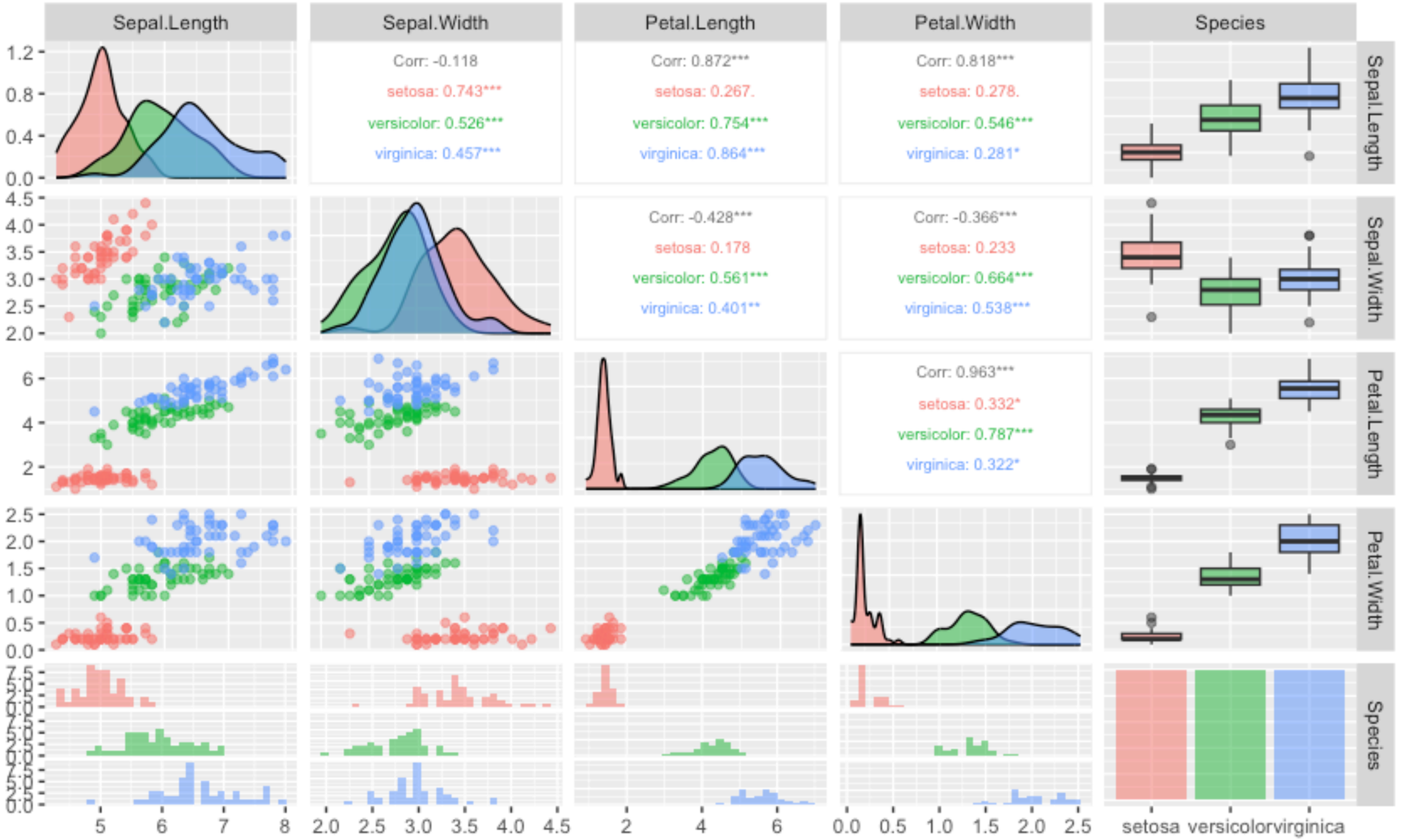


Diagrama de Dispersión III



Gráfica de Estrellas

Gráfica de 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

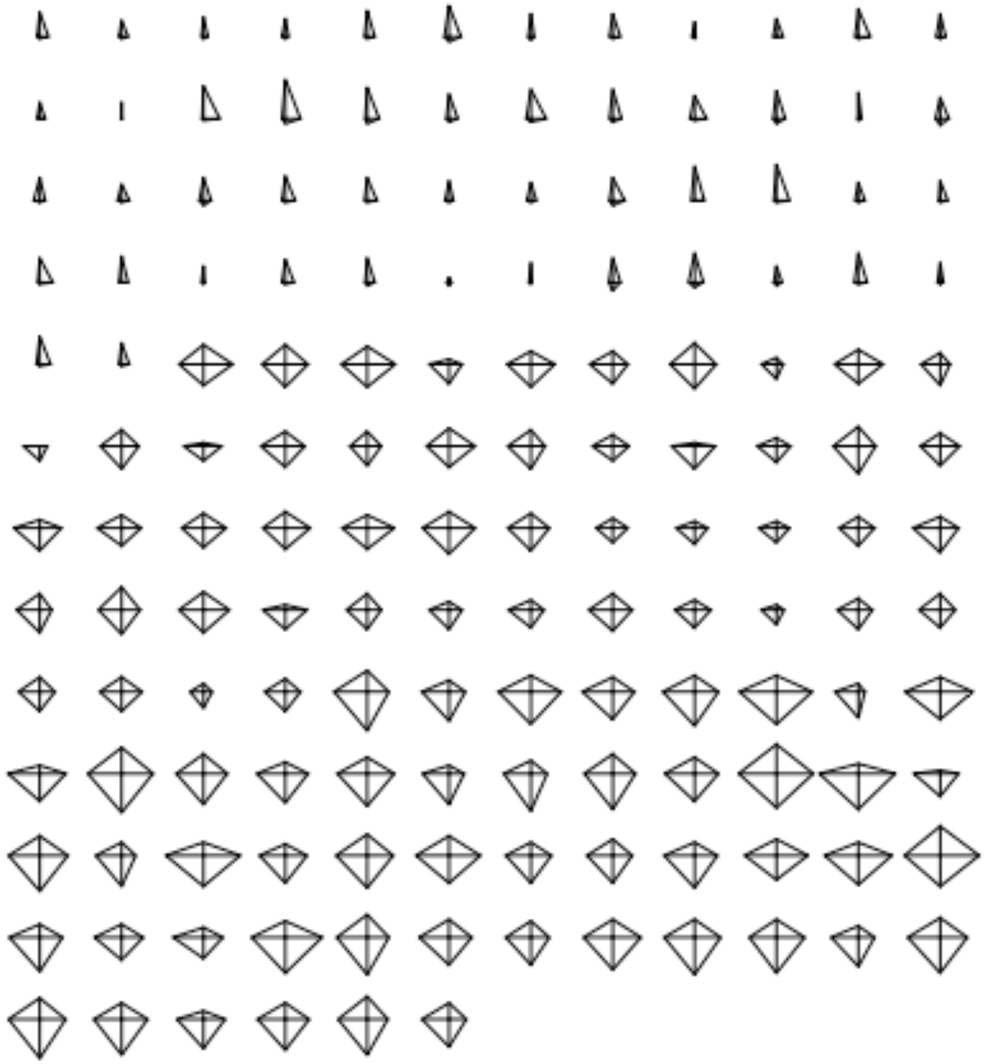
- Útil para:

  Identificar clusters, outliers y variables "importantes"

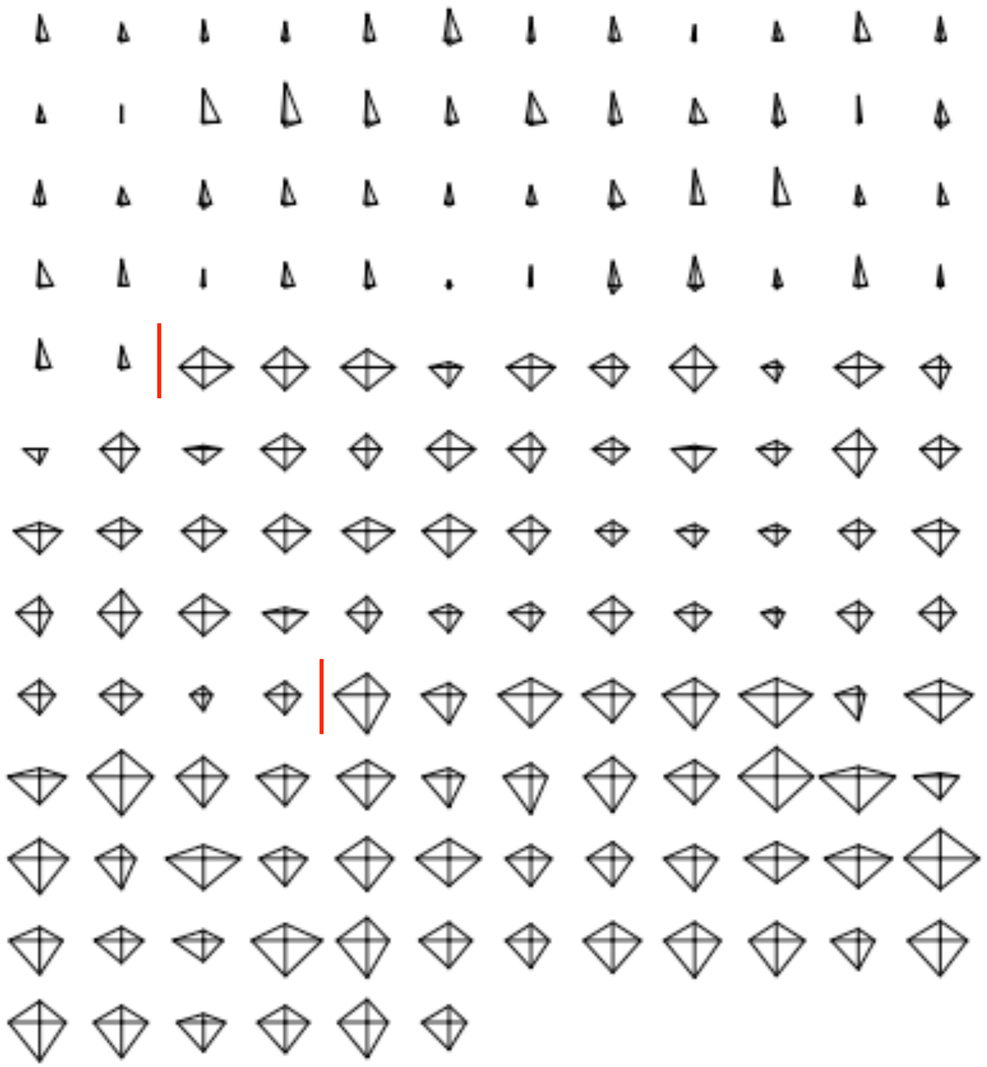
- Desventajas:

  Complicado de analizar si hay muchas observaciones y/o muchas variables

Gráfica de Estrellas



Gráfica de Estrellas



Caras de Chernoff

Caras de Chernoff

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

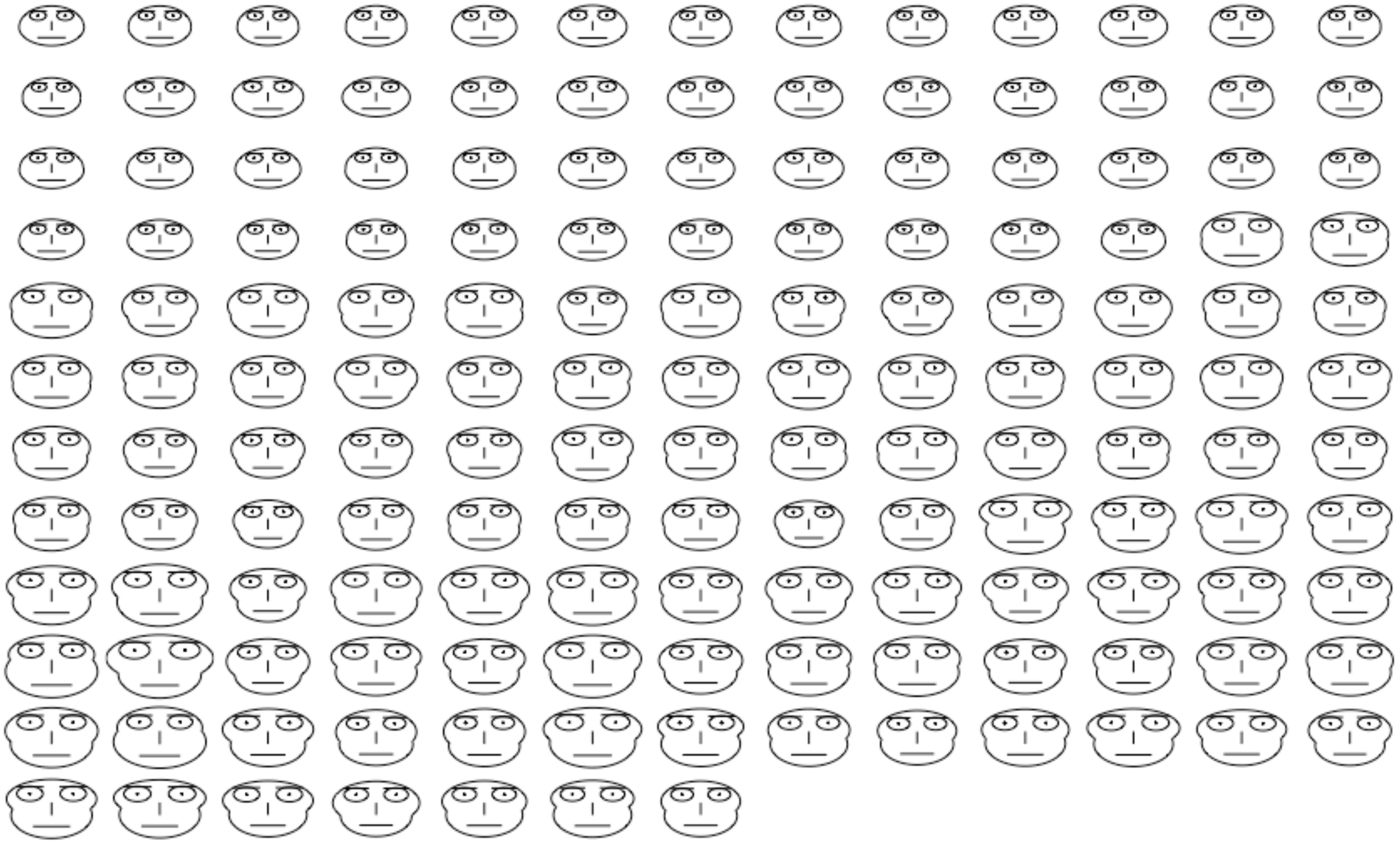
- Desarrollado por Chernoff, Herman (1973). **The use of Faces to Represent Points in K-Dimensional Space Graphically**

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

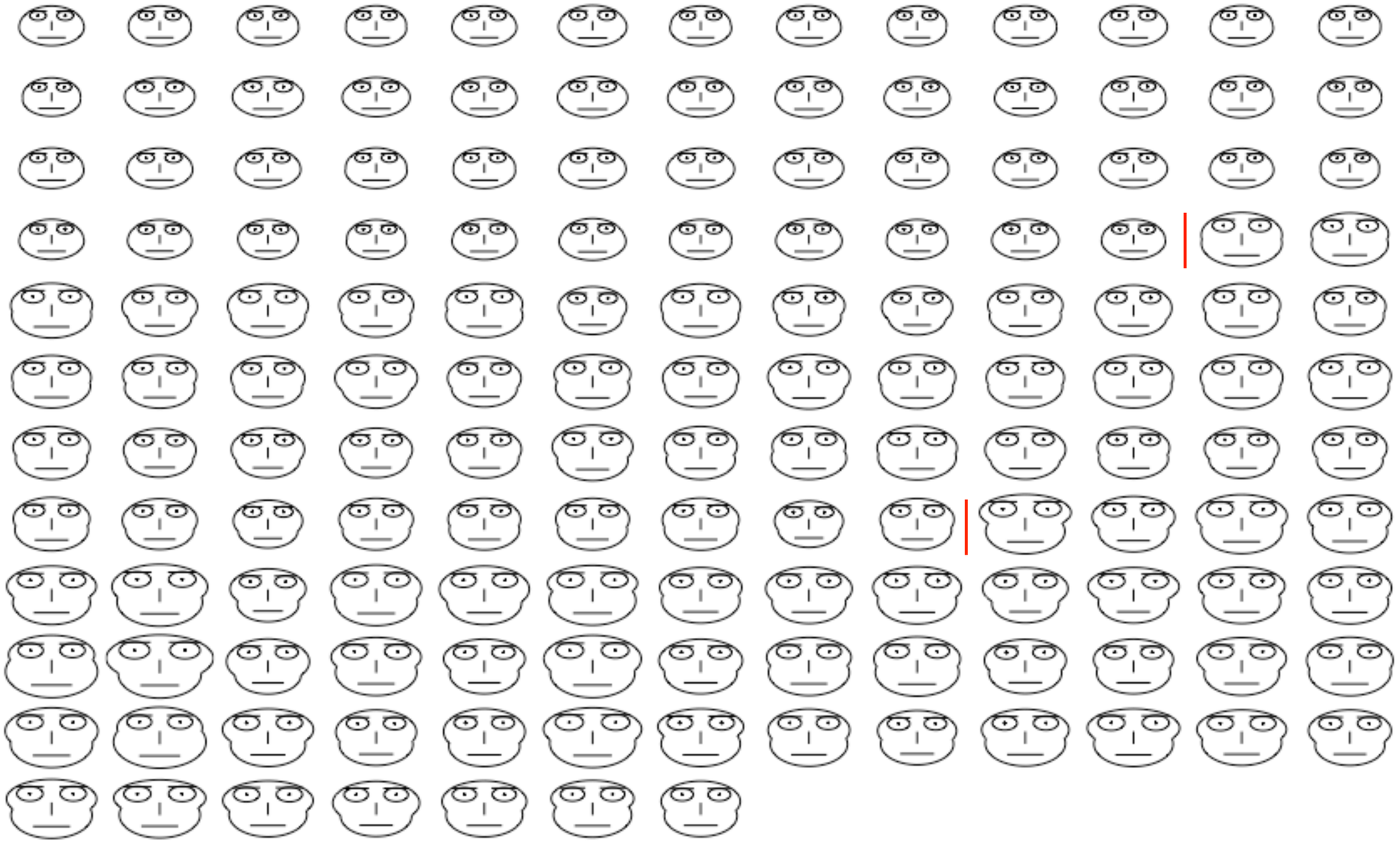
- Desventajas:
  Limitado a $p \leq 18$
  El orden de las variables importa

- En R: Librería TeachingDemos

Caras de Chernoff



Caras de Chernoff



Caras de Chernoff

- El orden importa

Curvas de Andrews

Curvas de Andrews

- Transformación ad hoc para graficar datos multivariados en el plano cartesiano o en 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:

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

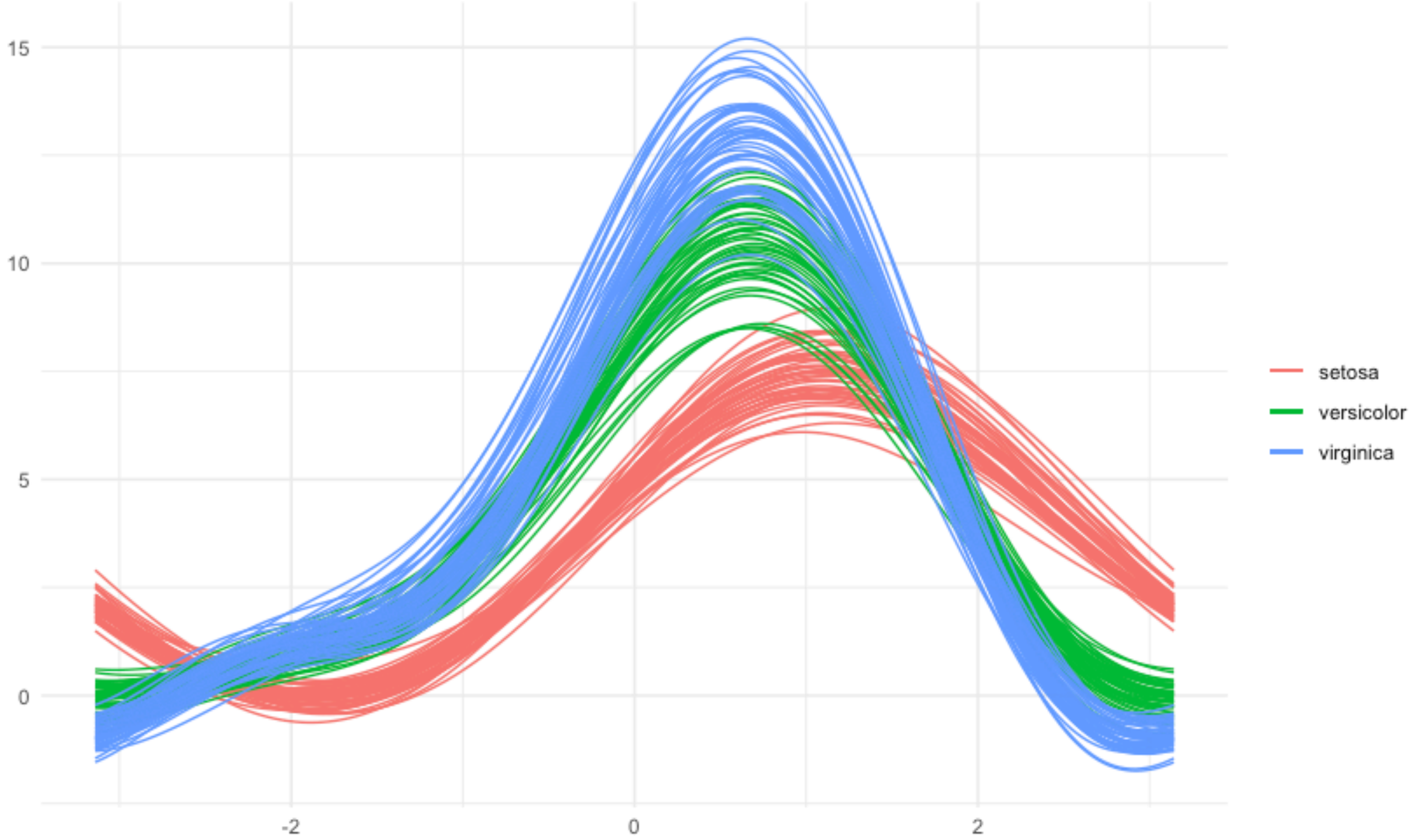
- Ventajas

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

- Desventajas

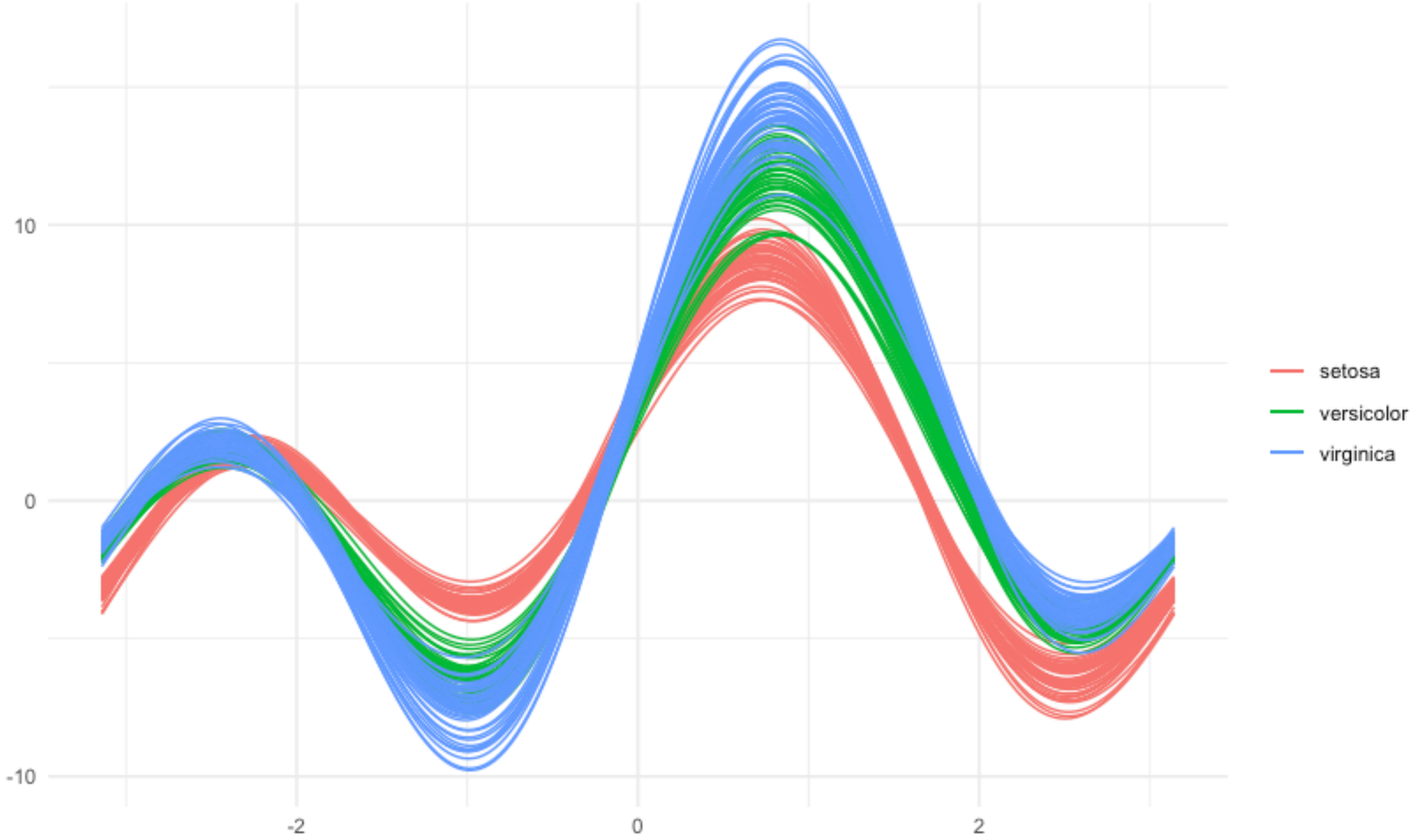
  - El orden de las variables importa
  - Mayor peso a las primeras variables.

Curvas de Andrews



Curvas de Andrews

- El orden importa



Curvas de Andrews

- Otros posibles mapeos

  - Andrews, 1972

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

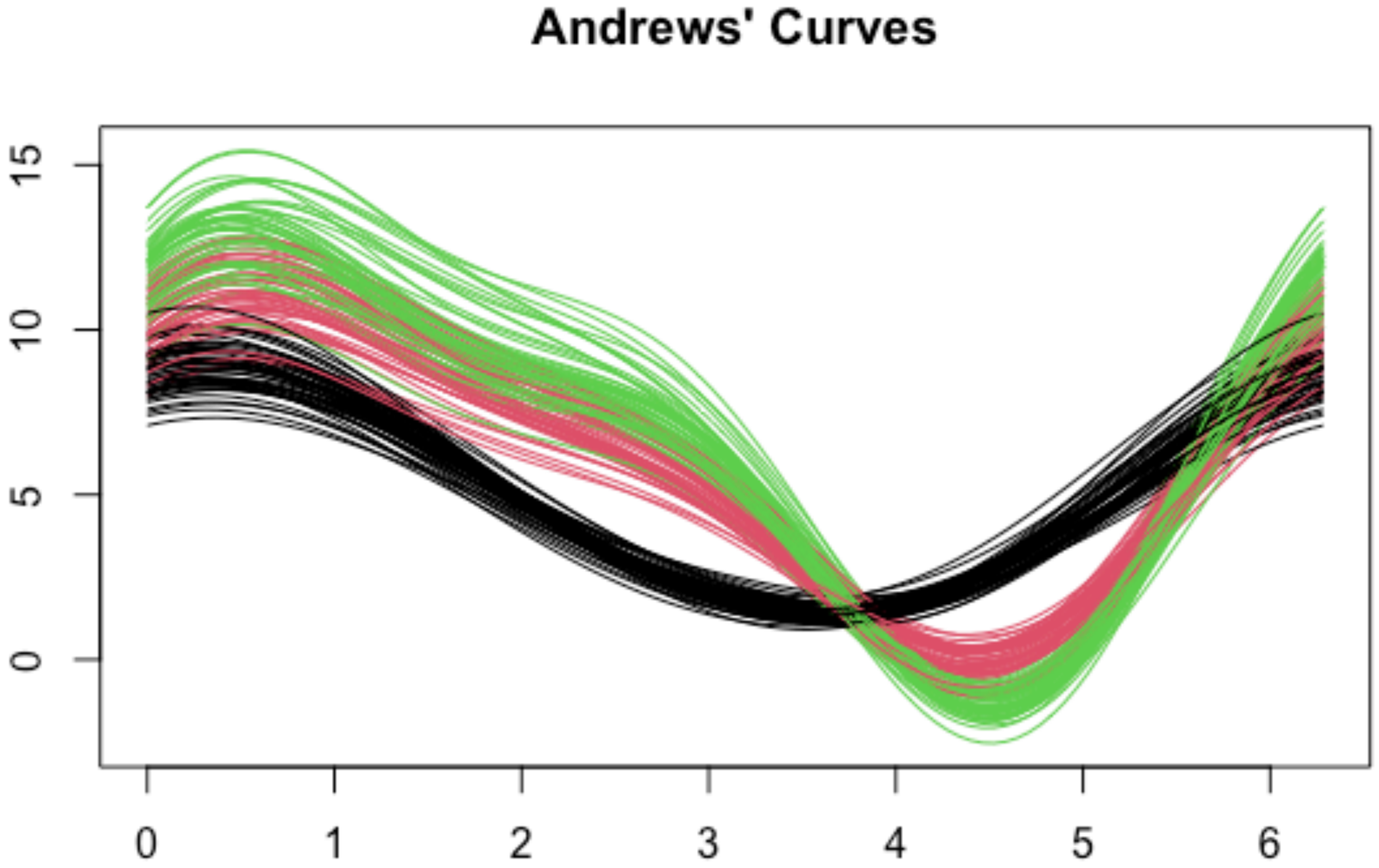
$$f_{\mathbf{x}}(t) = x_1 \sin(2t) + x_2 \cos(2t) + x_3 \sin(4t) + x_4 \cos(4t) + \cdots, \qquad 0 \leq t \leq \pi$$

  - Khattree, R. & Naik, D. (2002). **Andrews plots for multivariate data: some new suggestions and applications.**

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

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

Curvas de Andrews



Estadísticas Descriptivas

Media Muestral

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

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

- Para obtener el vector de medias

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

- Formalmente, definimos al vector de medias como

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

Media Muestral

- Proposición

La media muestral de una matriz de datos $\mathbf{X}$ está dada por

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

donde

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

- Observaciones

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

Media Muestral

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

Varianza y covarianza muestral

- Varianza de $\mathbf{x}^{(j)}$

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

- Covarianza 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 varianza y covarianza

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

Varianza y covarianza muestral

- Formalmente definimos a 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 $\mathbf{W}$

- **Observación**

$$
\begin{aligned}
\mathbf{W} &= \mathbf{X} - \begin{pmatrix} \bar{\mathbf{x}}^T \\ \bar{\mathbf{x}}^T \\ \vdots \\ \bar{\mathbf{x}}^T \end{pmatrix} \\
&= \mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}^T \\
&= \mathbf{X} - \mathbf{1}_n \left[\frac{1}{n} \mathbf{X}^T \mathbf{1}_n \right]^T \\
&= \mathbf{X} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \mathbf{X} \\
&= \left(\mathbf{I}_{n \times n} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \mathbf{X} \\
&= \mathbf{H}_n \mathbf{X}
\end{aligned}
$$

Varianza y covarianza muestral

- **Definición**

A la matriz $\mathbf{H}_n$ se le conoce como matriz de centrado

- **Proposición**

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

- Proposición

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

i. $\mathbf{S}$ es simétrica

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

- Proposición

La matriz de varianza y covarianza muestral es semidefina positiva

Varianza y covarianza muestral

- En **R**:
  - var ()
  - cov()
  - sweep( ) - para construir la matriz **W**
  - by( ) - para la varianza/covarianza muestral por grupos

Correlación muestral

- Finalmente, la correlación entre $\mathbf{x}^{(j)}$ y $\mathbf{x}^{(k)}$

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

- La matriz de correlación dada por

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

- Otra representación útil

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$$

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

- **Proposición**

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

i. $\mathbf{R}$ es simétrica.

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

Correlación muestral

- En **R**:
  - cor ()
  - by( ) - para la correlación muestral por grupos