Análisis Descriptivo Multivariado de Iris

Análisis Multivariado: Semestre 2025-II

José A. Perusquía Cortés

1 Iris

La base de datos *iris*, cotiene las medidas en centímetros de la longitud y el ancho del sépalo y del petalo de 50 flores, de las tres especies de iris, i.e. *setosa*, *versicolor* y *virginica*. Los primeros cinco registros son:

Tabla 1: Primeros cinco registros de la base de datos iris

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa

En este archivo trabajaremos con estos datos para explorar algunas posibles representaciones gráficas multivariadas, así como la correcta presentación de resúmenes y estadísticas utilizando RMarkdown. Antes de proceder al análisis descriptivo, es importante notar que para cada una de las 150 observaciones se conoce la especie de iris a la que pertenece. Por lo que esta información puede ser utilizada para condicionar y crear gráficas por grupo. Es claro que en la práctica, no siempre tendremos acceso a este tipo de información; sin embargo, las representaciones gráficas que se muestran a continuación, pueden ser fácilmente adaptadas para esta situación, por lo que siguen siendo representaciones válidas (siempre y cuando representen de forma clara y concisa lo que se busca).

1.1 Gráficas de dispersión y correlación

Las gráficas de dispersión y correlación nos permiten estudiar las relaciones por pares entre las variables de interés. Es claro que a partir de esto podemos darnos una idea de la estructura de los datos; sin embargo, este entendimiento será parcial, ya que no tendremos la imagen completa. Otra potencial desventaja de este tipo de representaciones, es que no son adecuadas para cuando se tienen muchas variables. Por lo que en estas situaciones, se tendrá que recurrir a otro tipo de gráficas. Para hacer gráficas de dispersión en **R**, se utilizará la librería **GGally**, la cual explota las bondades de *ggplot* para generar representaciones gráficas de gran calidad. Como ejemplo de esto, se tiene la Figura 1, donde se puede apreciar las gráficas de dispersión por pares, así como histogramas y diagramas de caja y bigote por grupo de cada una de las variables. A partir de esta figura, es claro que la información de la longitud y ancho del pétalo deberían sernos útil para discriminar el grupo de *setosa* de las otras dos especies de iris.

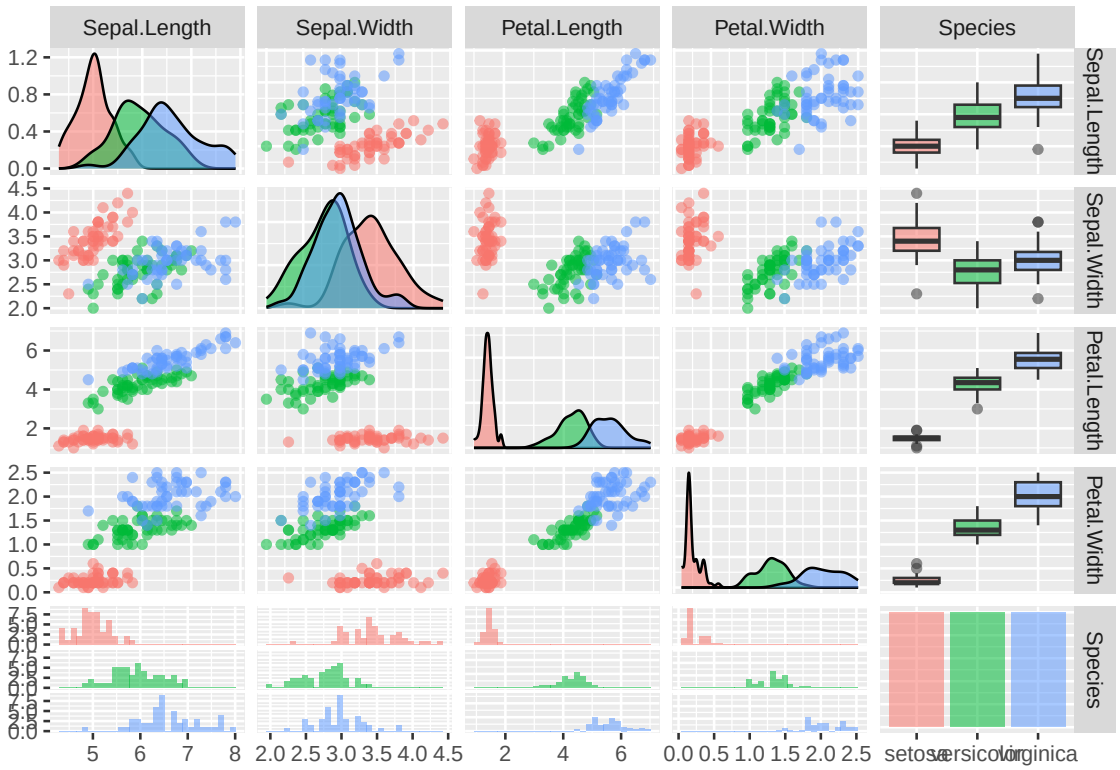

Figura 1: Diagrama de dispersión.

Ahora bien, para entender las posibles relaciones lineales de las variables en cada uno de los grupos, podemos hacer uso de la librería **corrplot**, la cual nos permite generar la Figura **??**. Así hay varios comentarios por hacer, en primera instancia, podemos apreciar que en todos los grupos se observa una correlación positiva para todos los pares de variables. Sin embargo, la correlación entre la longitud y el ancho del sepalo es mucho más fuerte en setosa que versicolor y virginica. En esta última, además se tiene una correlación muy fuerte entre la longitud del sepalo y del petalo. Finalmente, para versicolor se observa una correlación muy alta para todos los pares de variables.

```
{r cor1, fig.cap='Gráficas de correlación por grupo.', fig.subcap=c('Setosa', 'Versicolor',
'Virginica'),fig.ncol = 3,  , fig.align = "center",echo=F} corrplot(cor(iris[which(iris$Species=="setosa
corrplot(cor(iris[which(iris$Species=="versicolor"),-5]),method="ellipse") corrplot(cor(iris[which(iris
```

Diagrama de estrellas Otra representación gráfica de gran utilidad son las estrellas, que nos permiten sobre todo identificar clusters, outliers y variables importantes. En esta representación, cada observación tiene asociada una estrella de p picos, que a su vez representan las p variables. Es importante notar que para poder graficar las estrellas, los datos deben estar escalados en el intervalo $[0, 1]$. Ahora bien, este método también cuenta con algunas desventajas, como el hecho de que las estrellas son claras y concisas cuando no se tienen muchas variables u observaciones, que impidan apreciar la estructura de los datos. Para el caso de iris, las estrellas se pueden ver en la Figura 2.

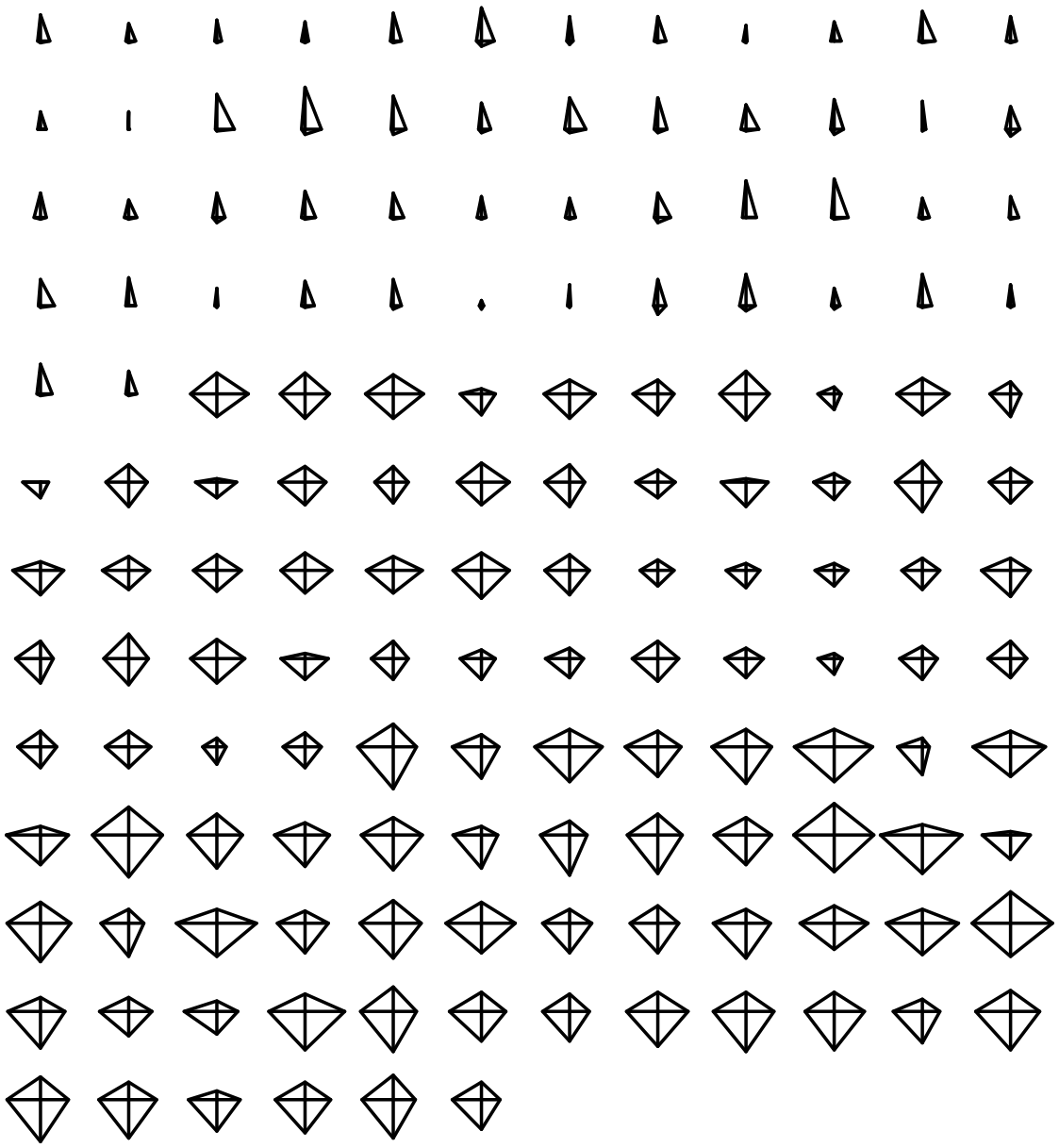

Figura 2: Diagrama de estrellas.

En este caso es posible distinguir los tres grupos que ya sabíamos que existían de antemano. Sin embargo, en el caso de no tener datos etiquetados *a priori*, esta técnica sigue siento útil y válida para intentar encontrar clusters, así como posibles outliers, cuyas estrellas diferirán del resto de forma significativa.

1.2 Caras de Chernoff

Una posible alternativa a las estrellas son las caras de Chernoff, donde cada osbervación tendrá una cara asociada y las variables representarán diferentes características faciales. Al igual que las estrellas, las caras de Chernoff nos permiten identificar clusters y outliers, cuyas caras diferirán del resto de forma clara y significativa. De igual forma, esta representación será adecuada siempre que no se tengan demasiadas observaciones y/o variables. Para el caso de iris, las caras de Chernoff se pueden apreciar en la Figura 3.

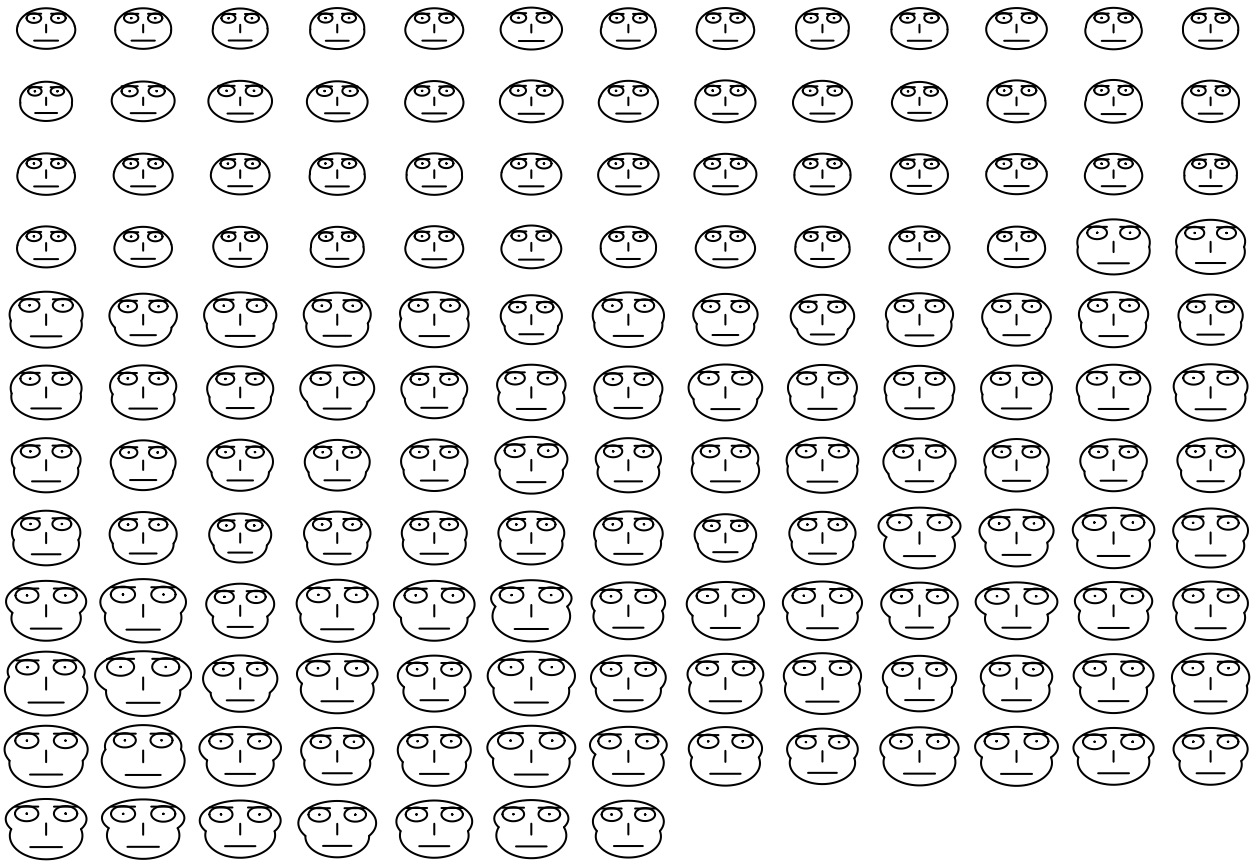

Figura 3: Caras de Chernoff.

Es valioso mencionar que el orden de las variables importa, ya que si se asigna una característica facial a otra variable esta generará un conjunto de caras completamente diferentes.

1.3 Curvas de Andrews

Finalmente, la última representación que se considerará son las curvas de Andrews. Estas curvas tienen la ventaja de que *a priori*, no tienen limitaciones en el número de observaciones ni de variables. Para un conjunto de n observaciones, se tendrán n curvas diferentes, que permiten identificar clusters y outliers, ya que el mapeo f se construye de tal forma que preserva medias y distancias. De esta forma, observaciones cercanas tendrán una curva cercana y observaciones alejadas de la media y del resto de observaciones podrán ser identificadas. También es valioso notar, que al igual que las caras de Chernoff, el orden de las variables importa, ya que el mapeo f tenderá a darle una mayor importancia a las primeras variables y al cambiar el order, también se generarán un conjunto de curvas diferentes. Para el caso de iris, las curvas de Andrews se pueden apreciar en la Figura 4.

Es claro, que al tener las etiquetas esto hace que las curvas de Andrews permitan ver las diferencias entre los grupos. Sin embargo, para cuando no se tienen los grupos *a priori*, identificar clusters será más complicado.

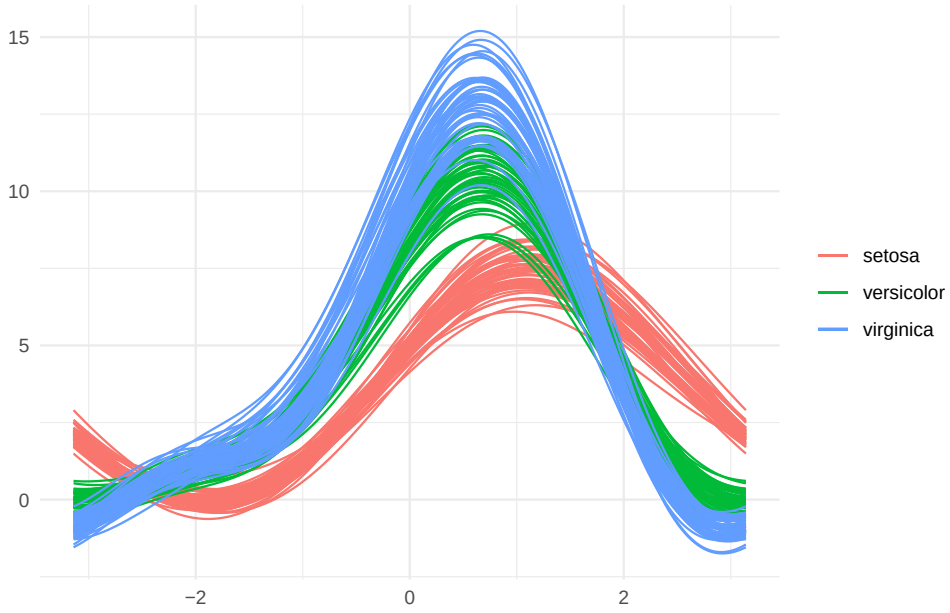

Figura 4: Curvas de Andrews.