La distribución normal multivariante

Simulación de vectores normales

Para generar una muestra de datos con distribución normal multivariante se usa el comando **mvrnorm** del paquete **MASS**. Los tres principales argumentos de este comando son:

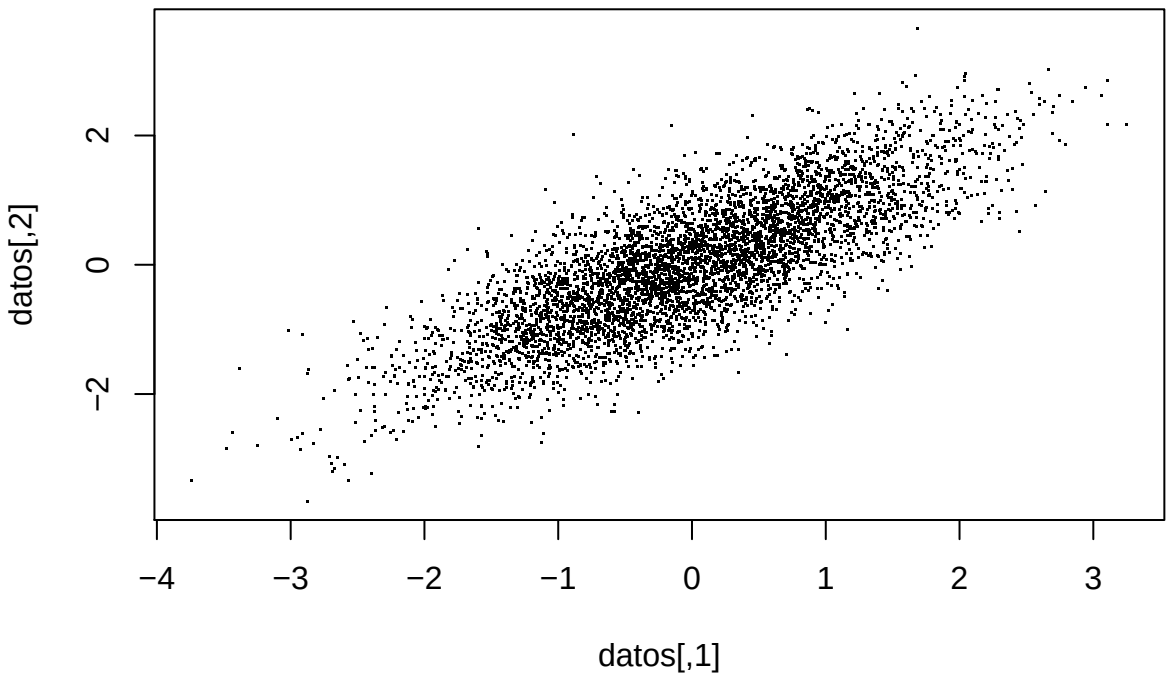
- `n`, el número de puntos que queremos generar,
- `mu`, el vector de medias de la distribución,
- `Sigma`, la matriz de covarianzas de la distribución.

Por ejemplo, para generar (y representar) 5000 puntos con distribución normal bidimensional con vector de medias el origen y matriz de covarianzas

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

podemos ejecutar los comandos siguientes:

```
library(MASS)
n <- 5000
mu <- c(0,0)
sigma <- matrix(c(1,0.8,0.8,1),2)
datos <- mvrnorm(n,mu,sigma)
plot(datos, pch='.')
```



Estimación del vector de medias y de la matriz de covarianzas

Si queremos estimar a partir de datos que hemos generado los parámetros del modelo mediante sus cantidades análogas muestrales podemos utilizar el siguiente código:

```
muest <- colMeans(datos)
muest
```

```
## [1] -0.009227642  0.002816826
```

```
sigmaest <- cov(datos)
sigmaest
```

```
##           [,1]      [,2]
## [1,] 0.9837414 0.7927420
## [2,] 0.7927420 0.9991218
```

Compara los estimadores obtenidos con los valores poblacionales. Dado que el tamaño muestral es muy grande, deben parecerse bastante.

Distancia de Mahalanobis

Se calcula con el comando **mahalanobis** cuyos principales argumentos son:

- `x`, la matriz de datos
- `center`, el centro respecto al que se calcula la distancia
- `cov`, la matriz de covarianzas respecto a la que se calcula la distancia (normalmente, la matriz de covarianzas muestral)

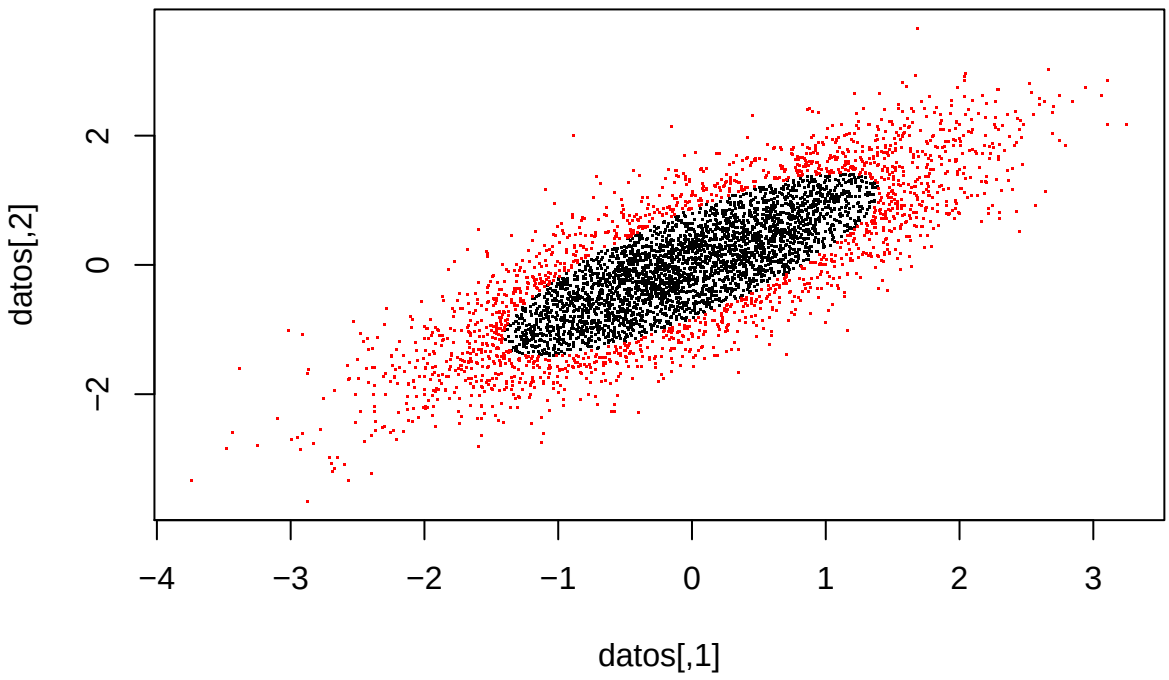
Para los datos y estimadores calculados anteriormente:

```
distancias <- mahalanobis(datos, muest, sigmaest)
```

Ejercicio: Calcula las principales medidas numéricas (media y desviación típica) y representa los principales gráficos (histograma y diagrama de cajas) para describir la distribución del vector de distancias. Compara los resultados con lo esperado bajo el modelo normal.

Con los siguientes comandos se representan los puntos en distinto color según su distancia esté por debajo o por encima del valor medio teórico (que en este caso es $p = 2$):

```
plot(datos, pch='.')
points(datos[distancias>2,], pch='.', col='red')
```



Ejercicios

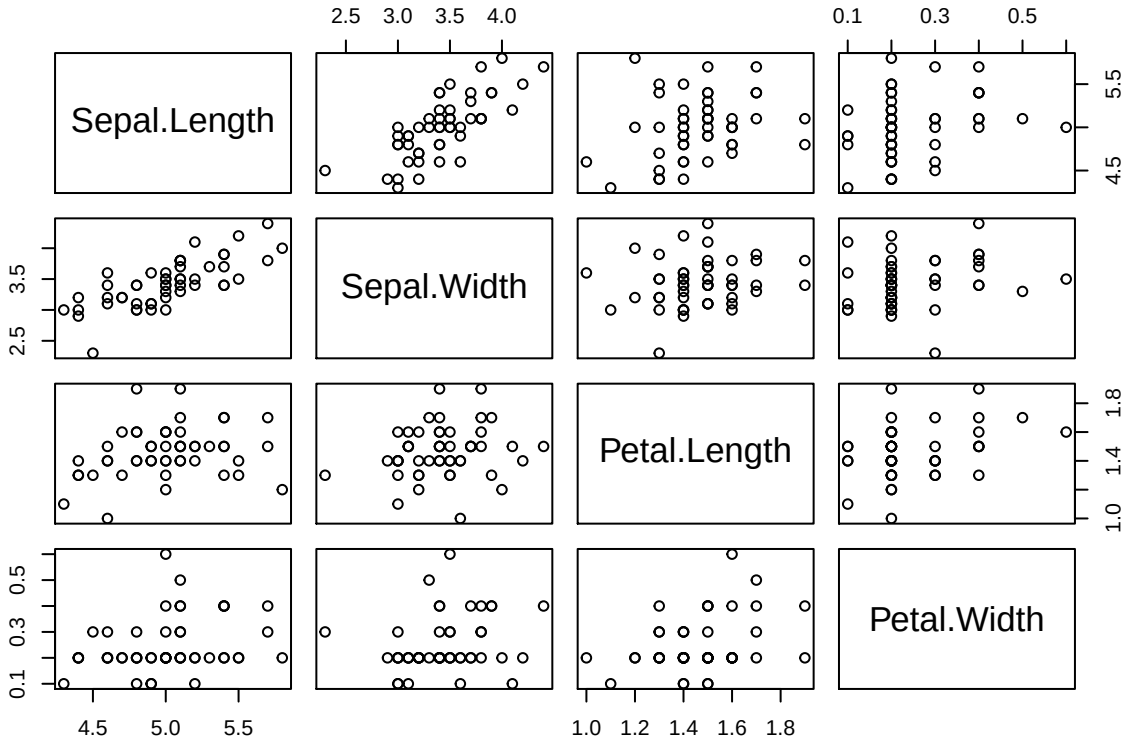
Los primeros ejercicios se refieren a los datos de lirios de Fisher. El fichero **iris**, ya incluido en la distribución básica de **R**, contiene los datos de los lirios de Fisher. Primero leemos las primeras filas del fichero para confirmar su estructura:

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Nos quedamos solo con las 50 observaciones correspondientes a la primera de las tres especies, *setosa*.

```
datos <- iris[1:50,-5] # eliminamos la quinta columna que ya no es necesaria
pairs(datos) # matriz de diagramas de dispersion
```



Para estos datos se proponen los siguientes ejercicios:

1. Calcula el vector de medias muestral y las matrices de covarianzas y de correlaciones (**cor**) muestrales. ¿Entre qué par de variables es más alta la correlación? ¿Qué variable tiene la mayor varianza?
2. Calcula las distancias de Mahalanobis entre cada uno de los lirios y el vector de medias. Representa los datos, usando el color rojo para el 25 % de los lirios más lejanos al vector de medias.
3. Representa un histograma de las distancias y compáralo con la funcion de densidad de una variable χ^2 con 4 grados de libertad.

El resto de ejercicios ya no están relacionados con los datos de los lirios:

4. Genera 100 observaciones con distribución normal bidimensional con vector de medias el origen y matriz de covarianzas:
$$\Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}$$
Representa la nube de puntos generados, su vector de medias y su matriz de covarianzas.
5. Para la misma distribución del apartado anterior, calcula el valor esperado teórico de la segunda coordenada condicionada a la primera. Si no lo conocieras y solo dispusieras de los datos generados, ¿cómo lo estimarías? Calcula el valor resultante para el estimador que has propuesto.