Universidad de Salamanca
Grado en Matemáticas

REVISIÓN DE MÉTODOS MULTIVARIANTES SUPERVISADOS Y NO SUPERVISADOS

Trabajo Fin de Grado

VNiVERSiDAD D SALAMANCA

Alumno: Pedro Ángel Fraile Manzano

Tutoras: Ana Belén Nieto Librero y Nerea González García

Salamanca, Julio de 2023

A Loreto, que me inspiró a no abandonar
mis sueños aunque las tormentas fueran duras.
A Paula, la que me iluminó
en los momentos más oscuros.
A mi familia, que empujó mis aspiraciones
hasta que se convirtieron en realidades.
A Moloka'i, Titania y Pangea que me lo
dieron todo cuando el mundo se paró.

Índice general

Introducción

El Análisis Multivariante es, según *Martínez Arias, R.*[22], el conjunto de técnicas y métodos que busca describir y extraer información de las relaciones entre variables medidas en una o varias muestras u observaciones. Dentro de esta definición, entran todos los procedimientos que analizan de manera simultánea más de una variable.

La clasificación que se dará en esta memoria de los métodos multivariantes será la siguiente

- ***Métodos supervisados:*** Son aquellos que explora la relación estocástica entre varias variables, divididas en variables respuesta y variables predictoras, estas técnicas a su vez se pueden dividir según la naturaleza de las variables respuesta [21]:
  - *Regresión:* Cuando las variables de respuesta son continuas, dentro de esta tipología, entran métodos como las *Redes Neuronales[31], los Árboles de Regresión [32]* etc...
  - *Clasificación:* Cuando las variables de respuesta son discretas, como por ejemplo, el *Análisis discriminante [25], Regresión logística [33],*
- ***Métodos no supervisados:*** Los métodos no supervisados son aquellos que buscan analizar la estructura y las relaciones entre las distintas variables [9]. Las distintas técnicas se distinguen según la medida en la que se centren, por ejemplo la variabilidad común [29], la homogeneidad de grupos [30] etc...

Estas técnicas han visto un auge en los últimos años, debido a su utilidad en el análisis de grandes bases de datos y la creciente capacidad de recogida de datos que se tiene en la actualidad. Basta con ver que la mayoría de artículos recopilados en los que se aplican estas técnicas se han publicado en las dos primeras décadas de siglo [7, 24, 25, 26]. Además, el llevar a buen término el estudio anteriormente era complejo debido a la gran cantidad de cálculos que se necesitaban realizar. Actualmente esos cálculos son automatizados, en algunos casos, con una única línea de código es posible llevar a cabo ciertas técnicas de las descritas.

Muchas veces, el uso de estas técnicas se combinan con el aprendizaje automático como proponen *Hastie T., Tibshirani R. y Friedman J.* [9] o *James G, Witten D, Hastie T, y Tibshirani R.* [21]. De esta manera, los parámetros se estiman extrayendo la información de los propios datos, comparando distintos modelos etc... Utilizando esta aproximación se pueden crear modelos predictivos bastante precisos. Este aspecto no se desarrollará en profundidad aunque si se mencionará en ciertos métodos en los que sea importante.

La estructura de esta memoria se divide en tres partes, una dedicada a los métodos supervisados en la que se detallan los más importantes, como la regresión lineal, el análisis discriminante, los árboles de decisión y las redes neuronales. En la segunda parte se desarrollan los métodos no supervisados con técnicas como el análisis de componentes principales, el análisis factorial y el análisis de conglomerados. Por último, se realizarán aplicaciones en las que se desarrollarán las interpretaciones de las tecnicas descritas durante la memoria.

En resumen, el principal objetivo de este trabajo es describir las principales técnicas de análisis de multivariante y fundamentar su aplicación e interpretación sobre cualquier tipo de datos. Además se busca dar ejemplos de aplicación sobre datos, sacando conclusiones provechosas.

Capítulo 1

Métodos Supervisados

Sea un conjunto de variables aleatorias observables de manera simultánea en una población. Estas variables se pueden dividir en dos tipos [9]:

- ***Variables de entrada o predictoras***: Son las variables independientes que determinarán de manera aleatoria al segundo conjunto de variables. Al conjunto de variables aleatorias de entrada se la denotará como el vector aleatorio $\mathbf{x} = [X_1, \dots X_p]$.

  Definición 1.0.1. Tomando N observaciones de estas variables de manera simultánea se obtiene la *matriz de datos* $\mathbf{X}$. Esta matriz es de tamaño $N \times p$ y contiene como filas los vectores de longitud p que representan los datos de cada observación, se denotarán a lo largo de la memoria como $\mathbf{x}_i, i = 1 \dots N$. Por ejemplo, en el caso de que se recojan datos sobre distintos modelos de coches las observaciones serían los valores medidas de las distintas variables consideradas en cada uno de los coches.

- ***Variables de salida*** : Son las variables dependientes de las anteriores. A este conjunto de variables se les denota con el vector aleatorio $\mathbf{y} = [Y_1, \dots Y_K]$, en el caso de que se tenga una única variable respuesta se denotará como Y. Como en el caso de las variables de entrada, se tomarán N observaciones de las K variables, dando como resultado la matriz de respuestas $\mathbf{Y}$ de tamaño $N \times K$, donde cada observación es una fila y se denota como $\mathbf{y}_i, i = 1 \dots K$, en el caso de que $K > 1$ y como y_i cuando $K = 1$. Siguiendo el ejemplo anterior, se podrían tener variables respuesta como el tipo de etiqueta medioambiental siendo esta una variable discreta o el precio del vehículo que tiene naturaleza continua.

Por ende, se recogen observaciones simultáneas de las variables de entrada y de salida formando parejas $(\mathbf{y}_i, \mathbf{x}_i), i = 1 \dots N$, de manera que se obtiene un vector fila de tamaño $p + K$. Estas observaciones de las variables forman una muestra aleatoria de la población .

Definición 1.0.2. Se llama *muestra aleatoria* de una variable aleatoria con una cierta distribución de probabilidad F, a un conjunto de N variables aleatorias con la misma distribución.

Definición 1.0.3. Se llaman métodos supervisados [42] a aquellos métodos que buscan inferir una relación estocástica entre dos grupos de variables, predictoras y respuesta en un conjunto de datos contiene observaciones simultaneas de ambos conjuntos de datos.

Durante la memoria, se establecerán los fundamentos a nivel poblacional, solo teniendo en cuenta las propiedades de las variables aleatorias y luego en caso de que sea necesario se tomarán N observaciones o realizaciones del experimento para formar las matrices de datos correspondientes.

En conclusión, el objetivo de estos métodos supervisados es encontrar una relación estocástica que llamaremos predictor, de tal manera que para una nueva observación de las variables de entrada, que denotaremos con $\mathbf{x}_0$, se pueda hacer una predicción $\hat{\mathbf{y}}_0$ del valor real $\mathbf{y}_0$ de la variable respuesta siempre con un cierto error que será una variable aleatoria.

1.1. Métodos Lineales para regresión

El objetivo de la regresión es encontrar como un conjunto de variables respuesta se ven afectadas por otro conjunto de variables predictoras. En este caso, se estudia como una combinación lineal de las variables puede afectar a las variables respuesta, que en el caso de la regresión es una variable aleatoria continua. Esto se puede hallar con fines predictivos o con el fin de analizar cómo cada una de las variables predictivas afectan a las variables respuesta [35].

Para ello, se tendrá en cuenta que las observaciones de la variable respuesta siguen la siguiente relación estocástica [9, 35]:

$$Y = f(\mathbf{x}) + \varepsilon \tag{1.1.1}$$

Donde ε es una variable aleatoria con $\mathbb{E}(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2$, habitualmente $\varepsilon \sim N(0, \sigma^2)$. Además supóngase que el vector de variables de entrada sigue una distribución normal multivariante $N_p(\mu, \mathbf{\Sigma})$ donde $\mathbf{\Sigma}$ es una matriz semidefinido positiva y simétrica, [6]. A esta última suposición se le podría añadir que sea diagonal, de manera que las variables aleatorias predictoras no estén correladas,pero no es necesario.

En el caso que nos atañe actualmente, se supondrá que f es una función lineal de las variables de entrada del vector aleatorio $\mathbf{x} = [X_1 \dots X_p]$. De esta manera, se tiene que:

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{1.1.2}$$

Definición 1.1.1. Se llaman *parámetros de regresión* al vector columna $\beta = [\beta_0, \beta_1, \dots \beta_p]^T$ de tamaño $(p+1)$ con los coeficientes necesarios para la regresión.

Añadiendo al vector $\mathbf{x}$ una nueva variable aleatoria X_0 que sea constantemente 1, se puede dar la siguiente expresión matricial de la función f:

$$f(\mathbf{x}) = \mathbf{x}\beta \tag{1.1.3}$$

De esta manera, tomando la matriz de datos de entrada de tamaño $N \times (p+1)$ entonces se puede generar un vector de predicciones de tamaño N de la siguiente manera:

$$\hat{\mathbf{y}} = \mathbf{X}\beta \tag{1.1.4}$$

Si en particular se quiere hacer una predicción para una nueva observación $\mathbf{x}_0$ basta con calcular basta con calcular $f(\mathbf{x}_0)$.

Una vez establecida la notación y los supuestos que tomaremos de ahora en adelante, hay que estimar los parámetros de regresión, para ello utilizaremos distintos métodos como el método de los mínimos cuadrados o el de Máxima verosimilitud, a parte de esto, veremos las características inferenciales de los estimadores obtenidos y su interpretación geométrica.

1.1.1. Métodos de ajuste

Sea una matriz de datos $\mathbf{X}$ de tamaño $N \times (p+1)$ resultado de hacer N observaciones de p variables aleatorias y añadir a la primera componente de cada observación un 1. Sea también un vector de respuestas $\mathbf{y} = [y_1, \dots y_N]^T$ de tamaño N, resultado de observar la variable respuesta Y.

Sea el vector de parámetros de regresión β de tamaño $p+1$ definido como antes, entonces podemos definir el concepto de función de pérdida.

Definición 1.1.2. Se llama *función de pérdida* [9] a una función $L(\mathbf{y}, \hat{\mathbf{y}})$ monótona creciente de la diferencia entre la predicción y el valor real de una variable

Aunque el término sea más utilizado en el marco del aprendizaje automático [21], el concepto de error cuadrático o la función de pérdida cuadrática es el que se suele usar en el ajuste de los parámetros de regresión, Ya que esta pérdida al minimizarse, es equivalente al método de los mínimos cuadrados, muy utilizado en distintos campos. En particular, la pérdida cuadrática se define de la siguiente manera [2]:

Definición 1.1.3. Se llama *error de predicción cuadrático* entre una variable observada, y, y la predicción obtenida por un cierto modelo $\hat{y}$ a la expresión $(y - \hat{y})^2$

Tomando la suma de los errores cuadrados cometidos en todas las observaciones se obtiene la suma de residuos cuadrados, *RSS*, y minimizando se puede obtener una estimación del vector de parámetros β. En particular, si se utilizan las expresiones matriciales anteriores:

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}(y_i - \mathbf{x}_i\beta)^2 = (\mathbf{y} - \mathbf{X}\beta)(\mathbf{y} - \mathbf{X}\beta)^T \tag{1.1.5}$$

Para obtener el mínimo se debe hallar los puntos estacionarios, lo que implica calcular la derivada de la suma de residuos cuadrados respecto del vector β, y como esta es una forma cuadrática general respecto de β,*(Para el desarrollo de la expresión véase [18])*:

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)^T \tag{1.1.6}$$

La segunda derivada respecto del vector de parámetros es $2\mathbf{X}^T\mathbf{X}$, entonces es semidefinido positiva, ya que todos los valores propios son positivos o nulos. Entonces los β en los que $-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)^T = 0$ son mínimos.

Asumiendo que $\mathbf{X}$ es una matriz de rango máximo $p+1$, por tanto, que la matriz $\mathbf{X}^T\mathbf{X}$ es invertible. Lo que implica que la siguiente expresión tiene solución única:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{1.1.7}$$

Esa solución es :

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{1.1.8}$$

Para el cálculo de este vector de parámetros se ha utilizado el método de los mínimos cuadrados. Aún así, hay otros métodos de obtención del vector $\hat{\beta}$, como el geométrico o el de máxima verosimilitud.

Con la siguiente proposición se puede ver la equivalencia entre los mínimos cuadrados y el de máxima verosimilitud:

Proposición 1.1.1. Esta estimación de los parámetros β por mínimos cuadrados es equivalente a la estimación de estos mediante el método de máxima verosimilitud [9].

Demostración. Sea un conjunto de N observaciones, $y_i, i = 1, \dots N$ de una variable aleatoria Y, de manera que su función de probabilidad, $\mathbb{P}_\theta(y)$ depende de los parámetros θ. Entonces el método de máxima verosimilitud busca maximizar la siguiente función.

$$L(\theta) = \sum_{i=1}^{n} log(\mathbb{P}_\theta(y_i)) \tag{1.1.9}$$

Suponiendo que la variable respuesta cumple como antes que $Y = f_\theta(\mathbf{x}) + \varepsilon$, donde $\varepsilon \sim N(0, \sigma^2)$, provoca que si se suponen conocidos a priori el parámetro θ y el vector aleatorio $\mathbf{x}$ entonces :

$$\mathbb{P}(Y|\mathbf{x}, \theta) = N(f_\theta(\mathbf{x}), \sigma^2) \tag{1.1.10}$$

Teniendo esto en cuenta, la función $L(\theta)$ tiene la siguiente expresión:

$$L(\theta) = -\frac{n}{2}log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2 \tag{1.1.11}$$

Por tanto, teniendo en cuenta que el último término es $RSS(\theta)$, entonces maximizar $L(\theta)$ es equivalente a minimizar $RSS(\theta)$. □

Corolario 1.1.1. Las estimaciones para cualquier f_θ obtenidas por el método de los mínimos cuadrados son equivalentes a las del método de máxima verosimilitud.

Continuando con lo anterior, los valores predichos obtenidos de las observaciones recogidas en la matriz de datos $\hat{\mathbf{y}}$ se calculan de la siguiente manera:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{1.1.12}$$

De esta manera, se está calculando la predicción de la respuesta $\hat{y}_i$ para cada una de las observaciones $\mathbf{x}_i, i = 1, \dots N$.

Otra forma de considerar el problema del ajuste sería desde el punto de vista geométrico para el que hay que desarrollar las siguientes definiciones y proposiciones.

En lo sucesivo se considera a no ser que se diga lo contrario que $N > p + 1$ y que la matriz de datos $\mathbf{X}$ es de rango r.

Definición 1.1.4. Se llama $C_r(\mathbf{X})$ al subespacio lineal de $\mathbb{R}^N$ generado por las columnas linealmente independientes de la matriz $\mathbf{X}$ [8].

Proposición 1.1.2. El vector $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$ es ortogonal al subespacio $C_r(\mathbf{X})$.

Demostración. Premultiplicando por $\mathbf{X}^T$:

$$\mathbf{X^T}\hat{\varepsilon} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = \mathbf{X}^T\mathbf{y} - \mathbf{X^T}\mathbf{X}\beta = 0 \qquad \square$$

Abdi, H [2] desarrolla este tipo de ajuste de manera superficial, mientras que *Hastie et.al.* [9] detallan este marco geométrico utilizando el proceso de ortonormalización de Graham-Schmidt, para encontrar combinaciones ortonormales de las variables originales. Este proceso se basa como se puede ver, en el hecho de que el vector de residuo generado debe ser ortogonal al espacio $C_r(X)$. Aunque no se desarrolle el ajuste en su totalidad, la proposición anterior será utilizada para demostrar una propiedad de las distribuciones de los estimadores hallados.

1.1.2. Inferencias Estadísticas sobre $\hat{\beta}$

En esta parte se estudian las propiedades de los estimadores obtenidos mediante el método de los mínimos cuadrados. Tras ello, se formarán estadigrafos para realizar los contrastes de hipótesis necesarios.

Hay que tener en cuenta los siguientes supuestos el vector $\mathbf{x} \sim N_p(\mu, \mathbf{\Sigma})$ una distribución multivariante de dimensión p, del cual tomamos N observaciones independientes obteniendo la matriz de datos $\mathbf{X}$, que a no ser que se diga lo contrario, se supondrá conocida. Además, supóngase el vector de N observaciones de la variable respuesta sigue el siguiente modelo $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ donde $\varepsilon \sim N(0, \sigma^2\mathbf{I}_N)$,es un vector de longitud N en el que cada componente es una normal $N(0, \sigma^2)$ independientes entre sí. Esto nos permite dar la siguiente proposición [8]

Proposición 1.1.3. El vector aleatorio formado por N observaciones de la variable respuesta conocida la matriz de datos y el vector de parámetros β sigue una distribución $\mathbf{y} \sim N_N(\mathbf{X}\beta, \sigma^2\mathbf{I}_N)$, donde $\mathbf{I}_N$ es la matriz identidad de tamaño N.

Demostración. Basta con comprobar que:

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}(\mathbf{X}\beta) + \mathbb{E}(\varepsilon) = \mathbb{E}(\mathbf{X}\beta) = \mathbf{X}\beta \tag{1.1.13}$$

Y la varianza :

$$\begin{aligned} \mathbb{E}((\mathbf{X}\beta+\varepsilon)(\mathbf{X}\beta+\varepsilon)^T) &= \mathbf{X}\beta\beta^T\mathbf{X}^T + Var(\varepsilon) \\ Var(\mathbf{y}) = \mathbf{X}\beta\beta^T\mathbf{X}^T + Var(\varepsilon) - \mathbf{X}\beta\beta^T\mathbf{X}^T &= Var(\varepsilon) = \sigma^2\mathbf{I}_N \end{aligned} \tag{1.1.14}$$

Y se concluye que el vector $\mathbf{y} \sim N_N(\mathbf{X}\beta, \sigma^2\mathbf{I}_N)$, ya que el vector $\varepsilon \sim N_N(0, \sigma^2\mathbf{I}_N)$ por hipótesis. □

Añadamos a las suposiciones que la matriz $\mathbf{X}$ es de rango máximo y por tanto, $\mathbf{X}^T\mathbf{X}$ es semidefinido positiva. A continuación se detallarán las cualidades inferenciales del vector $\hat{\beta}$, como que es insesgado, su varianza y su distribución.

Proposición 1.1.4. El estimador $\hat{\beta}$ es un estimador insesgado [34]

Demostración. Teniendo en cuenta las hipótesis anteriores, se obtiene que: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta+\varepsilon) = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon$

Entonces, se tiene que conocida $\mathbf{X} \Rightarrow \mathbb{E}(\hat{\beta}|\mathbf{X}) = \mathbb{E}(\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon|\mathbf{X}) = \beta + \mathbb{E}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon|\mathbf{X}) = \beta$ debido a la distribución de $\varepsilon \sim N(0, \sigma^2)$.

Por tanto, como $\mathbb{E}(\mathbb{E}(\hat{\beta}|\mathbf{X})) = \mathbb{E}(\beta) = \beta$ queda demostrado. □

Proposición 1.1.5. La varianza del estimador $Var(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$[9],[34]

Demostración.

$$\begin{aligned} Var(\hat{\beta}) &= \mathbb{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}|\mathbf{X}) \\ &= \mathbb{E}(\varepsilon\varepsilon^T)(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned} \tag{1.1.15}$$

□

Por último, tenemos que el vector de estimaciones de los parámetros de estimación, $\hat{\beta} = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon$, es una combinación afín de ε una variable con distribución por tanto, tendremos que cada uno de los componentes del vector $\hat{\beta}, \hat{\beta}_j \sim N(\beta_j, \sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X}^{-1}_{jj}))$ donde $j = 1 \dots p$ y $(\mathbf{X}^\mathbf{T}\mathbf{X}^{-1}_{jj})$ es el j-ésimo elemento de la diagonal de la matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ [35].

Si el parámetro σ^2 fuera conocido, entonces se podrían llevar a cabo los contrastes de manera sencilla. En este caso, $z_j = \dfrac{\hat{\beta}_j}{\sigma\sqrt{v_j}} \sim N(0,1)$

Para poder hacer los estadígrafos de contrastes se debe construir primero un estimador de esta varianza, ya que habitualmente no es conocida. Para ello *Hastie et. al.*[9] y *Cuadras C.M* [8] proponen el siguiente estimador de la varianza, suponiendo que unicamente hay una variable respuesta:

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1.1.16}$$

Proposición 1.1.6. El estimador $\hat{\sigma}^2$ cumple lo siguiente :

- $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$ (Es un estimador insesgado de la varianza)
- $\hat{\sigma}^2 \sim \frac{\sigma^2}{N-p-1}\chi^2_{N-p-1}$

Demostración. El término $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = RSS(\hat{\beta})$, se puede expresar como un producto escalar, tomando el vector residuo $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \hat{\mathbf{y}}$, se puede ver que el $RSS(\hat{\beta}) = \hat{\varepsilon}^T\hat{\varepsilon}$.

Sabiendo que $\hat{\varepsilon} \in \mathbb{R}^N$, consideremos una nueva base del espacio, $\{t_1, \dots t_{p+1}, t_{p+2}, \dots, t_N\}$, de tal manera que los $p+1$ primeros son una base de $C_{p+1}(\mathbf{X})$ y que sean ortonormales entre si, entonces, se puede tomar la matriz de cambio de base $\mathbf{T}$, es una matriz ortogonal, de manera que $\mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T = \mathbf{I}$.

Entonces, el vector $\mathbf{T}\hat{\varepsilon} = (\overbrace{0 \dots 0}^{p+1}, e'_{p+2}, \dots e'_N)^T$ por la proposición 1.1.2, y se cumple que $\mathbb{E}(\hat{\varepsilon}) = \mathbb{E}(\mathbf{T}\hat{\varepsilon}) = 0$, por la distribución que tiene el vector ε además $\mathbf{E}((e'_i)^2) = \sigma^2, i = p+2 \dots N$ ya que para los anteriores es 0 y por último, $\mathbb{E}(\hat{\varepsilon}^T\hat{\varepsilon}) = \mathbb{E}(\hat{\varepsilon}^T\mathbf{T}^T\mathbf{T}\hat{\varepsilon}) = \sum_{p+2}^{N}(\mathbf{e'_i})^{\mathbf{2}} = (N-p-1)\sigma^2$

Por tanto:

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{N-p-1}\mathbb{E}(RSS(\hat{\beta})) = \sigma^2 \tag{1.1.17}$$

Se concluye entonces que el estimador $\hat{\sigma}^2$, es insesgado.

Y para terminar, se puede ver que el estimador es suma de $N-p-1$ normales estándar multiplicadas por σ^2 y divididas entre $N-p-1$, por tanto, tenemos que además $\hat{\sigma}^2 \sim \dfrac{\sigma^2}{N-p-1}\chi^2_{N-p-1}$. □

Observación: en el caso de que la matriz de datos $\mathbf{X}$ sea de rango $r < p+1$ se debe construir un estimador distinto, para el cual se puede seguir el mismo razonamiento:

$$\hat{\sigma}^2 = \frac{1}{N-r}\sum_{i=1} N(y_i - \hat{y}_i)^2 \tag{1.1.18}$$

Por tanto, ahora ya podemos establecer, conocidas las distribuciones de los estimadores construidos que [34, 9]:

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}, j = 1, \dots, p+1 \sim t_{N-p-1} \tag{1.1.19}$$

Donde v_j es como antes el $j-$ésimo elemento de la diagonal de $\mathbf{X}^T\mathbf{X}$

Conocer la distribución de los estimadores permite realizar contrastes sobre los distintos parámetros. De manera habitual, se plantea la hipótesis nula $H_0 : \beta_j = 0$. Esta hipótesis busca comprobar si la variable X_j es importante en el modelo lineal. En caso de que se acepte la hipótesis esa variable puede ser eliminada del modelo.

Teniendo en cuenta esto, hay ocasiones en las que se desea comprobar si un subconjunto de variables es más significativo estadísticamente que otro, de manera que

se pueda eliminar variables que no aporten información en pos de la sencillez del modelo y su posterior interpretación.

Sea $p_1 + 1$ el número del conjunto más grande de parámetros o de variables a considerar y RSS_1 su error cuadrático respectivo, sean igualmente RSS_0 y $p_0 + 1$ para el segundo conjunto o subconjunto menor, si además contamos que :

$$\frac{RSS(\hat{\beta})}{N-p-1} = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \sim \sigma^2 \chi_{N-p-1} \tag{1.1.20}$$

Se puede definir el estadígrafo:

$$F = \frac{\dfrac{(RSS_0 - RSS_1)}{p_1 - p_0}}{\dfrac{RSS_1}{N-p_1-1}} \tag{1.1.21}$$

Que cumple:

Proposición 1.1.7. El estadígrafo $F \sim F_{(p_1-p_0),(N-p_1-1)}$

Demostración. Teniendo en cuenta lo anterior, se puede comprobar que:

$$F \sim \frac{\chi^2_{p_1-p_0}}{\chi^2_{N-p_1-1}} = F_{(p_1-p_0),(N-p_1-1)} \tag{1.1.22}$$

□

Este estadígrafo se referencia en secciones sucesivas con el objetivo de reducir las variables involucradas en la regresión.

1.1.3. Regresión multivariante

Hasta ahora, se ha considerado una única variable aleatoria respuesta, sea $\mathbf{y}$ el vector aleatorio de variables respuesta $\mathbf{y} = [Y_1, \ldots, Y_K]$ y un vector aleatorio de variables de entrada $\mathbf{x} = [X_0, X_1, \ldots, X_p]$. Se puede establecer el siguiente modelo análogo, donde :

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \varepsilon_k = f_k(\mathbf{x}) + \varepsilon_k, \quad k = 1, \ldots K \tag{1.1.23}$$

Donde el término $\varepsilon_k \sim N(0, \sigma_k^2)$ es el error inevitable para la variable $Y_k, k = 1 \ldots K$. De esta manera, si se toman N observaciones se puede crear las matrices de datos de respuesta y de entrada y obtener la siguiente expresión:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1.1.24}$$

Dentro de esa expresión matricial se tiene que:

- $\mathbf{Y}$ es la matriz de tamaño $N \times K$ que contiene los valores observados de las variables respuesta

- $\mathbf{X}$ matriz de datos habitual de tamaño $N \times (p+1)$
- $\mathbf{B}$ matriz de tamaño $(p+1) \times K$ que contiene los parámetros de la regresión
- $\mathbf{E}$ es la matriz de tamaño $N \times K$ que contiene las errores cometidos en cada uno de las variables respuesta.

El proceso de ajuste, en el caso de que los errores $\varepsilon^T = [\varepsilon_1, \dots \varepsilon_K]$ no estén correlados, de la matriz de parámetros es análogo al de una sola variable respuesta de tal manera que minimizando el error cuadrático acumulado se obtiene $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. En caso de que los errores tengan una matriz de covarianzas conocida, $\mathbf{\Sigma}$, entonces es necesario hacer la siguiente modificación en el RSS:

$$RSS(\mathbf{B}, \mathbf{\Sigma}) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^T \mathbf{\Sigma}^{-1} (y_i - f(\mathbf{x}_i)) \tag{1.1.25}$$

En la última expresión se usa la *distancia de Mahalanobis*.

Definición 1.1.5. La *distancia de Mahalanobis* entre dos observaciones $\mathbf{x}_i, \mathbf{x}_j$ extraídas de una misma población con matriz de covarianzas $\mathbf{\Sigma}$ se define de la siguiente manera[8]:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j)} \tag{1.1.26}$$

En el caso anterior, $(y_i - f(\mathbf{x}_i)) = \varepsilon_i$ y se está haciendo la distancia de Mahalanobis respecto de su media, el 0.

Selección de subconjuntos y métodos penalizados

Anteriormente, se ha detallado un estadígrafo que permitía hacer un contraste sobre la cantidad de variables a considerar en la regresión. Esta reducción de características a considerar permite hacer mucho más interpretable el modelo obtenido durante todo el proceso de ajuste y en algunos casos incluso aumentar la precisión del modelo ya que puede ocurrir que se eliminen variables que introduzcan ruido en el modelo.

Se podría seguir un método exhaustivo que calcule cada uno de los subconjuntos posibles para cada número de variables que creamos necesarias y calcular el error cuadrático acumulado de cada uno de los subconjuntos posibles. Pero este método es de una complejidad computacional alta.

Otra posibilidad sería utilizar el estadígrafo de contraste F definido en la Ecuación (1.1.21) empezando con una sola variable e ir añadiendo cada variable que mejore el ajuste. También se puede hacer al revés, empezando con el modelo con todas las variables e ir reduciendo la cantidad de estas. A estos algoritmo se les llaman *algoritmos voraces* los cuales buscan una solución óptima en cada paso y no en global.

El problema de estos métodos de selección de variables es que es un proceso discreto, una variable es o no considerada para el modelo siguiente y esto puede generar sobreajuste o infraajuste, no habiendo un término medio.

Para ello, existen los métodos penalizados o de encogimiento, que añaden un término de penalización para que los parámetros β no sean muy grandes.

Definición 1.1.6. Llamamos *Errores Cuadrados Acumulados Penalizados, PRSS,* a la suma de los cuadrados de los errores cometidos con el modelo lineal que usa el vector de parámetros β, añadiendo un término regulador $\lambda > 0$:

$$PRSS(\beta) = \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{x}_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1.1.27}$$

El último término hace que parámetros β_j grandes sean considerados perjudiciales, y el parámetro λ es una forma de regular cuanta importancia tiene dicha penalización, cuanto mayor sea, este provocará un encogimiento mayor de los parámetros. De esta manera, se tiene una forma continua de considerar los pesos no eliminando en de manera total las variables o haciendo el β_j correspondiente cercano a 0.

1.2. Clasificación

Hasta ahora se ha considerado el caso en el que la variable respuesta, es una variable continua en la cual se puede definir una métrica usual [10], por ejemplo en el caso de valores respuesta continuos y reales se puede definir fácilmente métricas tomando el error cuadrático.

En el caso de la clasificación o discriminación, casos que se discutirán a continuación sus diferencias [35], se tienen variables respuestas que no tienen un orden en específico. Por tanto el modo de tratar los datos cambia de manera significativa.

El problema de estudiar la dependencia entre variables continuas como sería el vector aleatorio $\mathbf{x}$ y una variable que no es continua, tiene varias aproximaciones posibles dependendiendo de los objetivos a perseguir y el conocimiento previo que se tenga de los datos en sí [63].

- En el caso en el que no se conozca la distribución de las distintas poblaciones o categorías, se puede analizar con fines exploratorios el *Análisis canónico de poblaciones* [63], en el que se estudian las direcciones de mayor distancia entre las dos poblaciones, en particular, también se le llama *Análisis discriminante lineal de Fisher*[17]. En este caso, se quiere explorar las características independientes son importantes a la hora de separar las distintos valores de la variable
- En el caso de que se conozcan los parámetros de las distribuciones se pueden crear funciones discriminantes que permitan clasificar nuevas observaciones $\mathbf{x}_0$ en una de las poblaciones. A esto se le llama *clasificación*, uno de los métodos más habituales es la creación de funciones discriminantes por máxima verosimilitud o siguiendo un criterio geométrico[23].

Para simplificar los cálculos y los desarrollos en el caso de la primera parte en la que se desarrollan las funciones discriminante, se consideran dos poblaciones en las que en un inicio unicamente conocemos la distribución de probabilidad[35] y luego se asumirá que son normales para construir el discriminante lineal habitual en la que aparece la distancia de Mahalanobis [23] aunque *Morrison, D.F*[18] construye dicha función discriminante buscando una combinación lineal que maximice el estadístico T^2 el cual mide la diferencia de las medias de las dos poblaciones.

En la segunda parte de la sección se detallará el método del análisis canónico de poblaciones o *análisis discriminante de Fisher*[63] que toma la matriz de covarianzas dentro y fuera de las poblaciones permitiendo tanto hacer predicciones en subespacios de dimensión menor.

1.2.1. Funciones discriminantes

Supongase que se tienen dos valores de la variable respuesta y, entonces, se tienen dos poblaciones,π_1, π_2 de las cuales se conocen sus funciones de densidad de cada una de ellas, f_1, f_2 respectivamente, entonces si además se conoce la probabilidad de clasificación errónea $P(1|2), P(2|1)$. Se tiene que:

$$P(i|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|i)P(i)}{P(1)P(\mathbf{x}_0|1) + P(2)P(\mathbf{x}_0|2)} = \frac{f_i(\mathbf{x}_0)P(i)}{P(1)f_1(\mathbf{x}_0) + P(2)f_2(\mathbf{x}_0)} \qquad (1.2.1)$$

Donde $P(1), P(2)$ son las probabilidades de pertenecer a cada una de las poblaciones. Por tanto, se puede clasificar como una población u otra ya que se conoce la distribución de cada una de las poblaciones. Primero definamos lo que es una función discriminante[8]:

Definición 1.2.1. Se llama *función discriminante* aquella que:

$$D_i : \mathbf{\Omega} \longrightarrow \mathbb{R} \tag{1.2.2}$$

Donde Ω es el espacio de observaciones posibles. Definida de tal manera que si $D_i(\mathbf{x}_0) > 0 \Rightarrow \mathbf{x}_0 \in \pi_i$ y en caso contrario $\mathbf{x}_0 \notin P_i$.

En el caso anterior de dos poblaciones, $f_1(\mathbf{x}_0)P(1) - f_2(\mathbf{x}_0)P(2)$ sería la función discriminante para discernir si $\mathbf{x}_0$ pertenece a P_1 o no. Ya que representa la probabilidad de ser de dicha clase conociendo el valor de las variables predictoras.

Definición 1.2.2. Se llama *Error de clasificación* al coste de clasificar de manera errónea una observación y se denota como $c(i|j) =$ *"Error de clasificar como* π_i *una observación perteneciente a* π_j*"*

Se puede ponderar con respecto del coste de cada error de manera que una clasificación errónea con un coste mayor no se pueda dar, es decir, es penalizar a la que mayor coste tenga

$$\frac{f_1(\mathbf{x}_0)P(1)}{c(1|2)} - \frac{f_2(\mathbf{x}_0)P(2)}{c(2|1)} \tag{1.2.3}$$

Por consiguiente, a igualdad de los otros términos escogeremos el que menos coste, el que mayor verosimilitud o el que mayor probabilidad a priori tenga.

Johnson R.A. y Wichern D.W.[35] desarrollan otro enfoque, tomando el coste medio esperado ECM:

$$ECM = c(2|1)P(2|1)P(1) + c(1|2)P(1|2)P(2) \tag{1.2.4}$$

Además afirman que las regiones que minimizan el coste de clasificación son las que vienen dadas por la función discriminante anterior.

Supóngase ahora que tanto f_1, f_2 son densidades normales con medias distintas μ_1, μ_2 pero con una matriz de covarianzas común $\mathbf{\Sigma}$, Entonces tenemos la siguiente expresión [35, 8]:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left\{\frac{-1}{2}(\mathbf{x}-\mu_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)\right\} \tag{1.2.5}$$

De esta manera, la función discriminante se puede transformar sustituyendo las densidades por la expresión anterior y tomando logaritmos :

$$(\mathbf{x}-\mu_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1)+log\left(\frac{\pi_1}{c(1|2)}\right)-(\mathbf{x}-\mu_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2)-log\left(\frac{\pi_2}{c(2|1)}\right) \tag{1.2.6}$$

Hay que tener en cuenta que el término $D_i^2 = (\mathbf{x}-\mu_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)$ se puede interpretar como la *distancia de Mahalanobis* de la observación $\mathbf{x}$ a la $i-$ésima población.

Por otro lado, se tiene un término que relaciona probabilidad de pertenencia con el coste de clasificación errónea. En el caso de que tanto las probabilidades como el coste fueran iguales, entonces la función discriminante para la población 1 sería:

$$D_1^2 - D_2^2 \tag{1.2.7}$$

De esta manera, se obtiene un criterio geométrico [8] en el que una nueva observación se clasifica según a que media esté más cerca, es decir cuanto más se parezca la nueva observación a la media de la región más opciopnes tiene de ser clasificada de esa manera.

Wichern, D.W. y Johnson, R.A.[35] desarrollan el caso cuando no se conoce la matriz de covarianzas muestral y cuando ambas matrices de cada una de las poblaciones π_1, π_2 no son iguales.

1.2.2. Análisis canónico de poblaciones

Hasta ahora, se ha asumido que conocemos las distribuciones de las poblaciones que estamos estudiando. Utilizando estas probabilidades y los costos asociados, hemos desarrollado funciones discriminantes de manera simple. Sin embargo, la realidad rara vez se ajusta a estos supuestos. Por esta razón, hay métodos que afrontan el problema de manera exploratoria, por tanto, se busca analizar como y cuanto las variables influyen a la hora de esta discriminación.

Esto se puede realizar teniendo en cuenta como es la variabilidad entre grupos y como es dentro de los grupos, en este caso buscaremos cuales son las direcciones en la cual se maximiza la variabilidad entre grupos y se minimiza la varianza dentro de las poblaciones.

A continuación, se describirá el método para calcular una nueva base de observaciones que cumpla con el objetivo antes descrito. Este método se basa en el trabajo de *Lebart L., Morineau A. y Warwick K.M.*[17], aunque también se puede seguir el desarrollo de *Johnson R.A. y Wichern, D. W.* [35] en el que utilizan la matriz de varianzas entre grupos en vez de la matriz de covarianzas totales.

Sea $\mathbf{X}$ la matriz de datos de tamaño $N \times p$ donde las filas $\mathbf{x}_i$ son cada una de las observaciones de las p variables. Dichas observaciones están particionadas en general por L grupos, sea I_l el conjunto de observaciones pertenecientes al l-ésimo grupo, sea también N_l el número de observaciones que pertenecen al l-ésimo grupo.

Se definen las medias muestrales $\overline{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, j = 1 \dots p$ de la j-ésima variable en la población en total. También se define la media muestral dentro de cada grupo que es $\overline{x}_{jl} = \frac{1}{N_l}\sum_{i \in I_l} x_{ij}$

Por ende, podemos dar la distancia entre dos variables como:

$$Cov(X_j, X_{j'}) = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \overline{x}_j)(x_{ij'} - \overline{x}'_j) \tag{1.2.8}$$

Esto se puede particionar por grupos de la siguiente manera:

$$Cov(X_j, X_{j'}) = \frac{1}{N}\sum_{l=1}^{L}\sum_{i \in I_l}(x_{ij} - \overline{x}_j)(x_{ij'} - \overline{x}'_j) \tag{1.2.9}$$

y a su vez cada uno de los $(x_{ij} - \overline{x}_j)$ se pueden dividir en la parte intergrupos e intragrupos:

$$(x_{ij} - \overline{x}_j) = (x_{ij} - \overline{x}_{jl}) + (\overline{x}_{jl} - \overline{x}_j) \tag{1.2.10}$$

Sustituyendo y simplificando lo necesario:

$$Cov(X_j, X_{j'}) = \frac{1}{N}\sum_{l=1}^{L}\sum_{i\in I_l}(x_{ij}-\overline{x}_{jl})(x_{ij'}-\overline{x}_{j'l})+\sum_{l=1}^{L}\frac{N_l}{N}(\overline{x}_{jl}-\overline{x}_j)(\overline{x}_{j'l}-\overline{x}_{j'}) \tag{1.2.11}$$

Esto nos permite dar una descomposición de la matriz de covarianzas total de la siguiente forma :

$$\mathbf{T} = \mathbf{B} + \mathbf{W} \tag{1.2.12}$$

Donde:

- $\mathbf{T}$ es la matriz de tamaño $p \times p$ que expresa la covarianza total y sus coeficientes $t_{jj'} = Cov(X_j, X'_j)$
- $\mathbf{B}$ es la matriz que expresa la covarianza entre los grupos y sus coeficientes son $b_{jj'} = \sum_{k=1}^{q} \frac{n_k}{n}(\overline{x}_{jk} - \overline{x}_j)(\overline{x}_{j'k} - \overline{x}_{j'})$
- $\mathbf{W}$ es la matriz que expresa la covarianza dentro de los grupos y sus coeficientes son $w_{jj'} = \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in I_k}(x_{ij} - \overline{x}_{jk})(x_{ij'} - \overline{x}_{j'k})$

Para cualquier combinación lineal que se quiera hacer de las variables de entrada de la forma $\mathbf{a}^T\mathbf{x}$, donde el vector $\mathbf{a}$ es un vector de p constantes, entonces la varianza se transforma de la siguiente manera:

$$Var(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\Sigma\mathbf{a} \tag{1.2.13}$$

Entonces transformando por el vector $\mathbf{a}$ tenemos que la Ecuación (1.2.12) se transforma de la siguiente manera:

$$\mathbf{a}^T\mathbf{T}\mathbf{a} = \mathbf{a}^T\mathbf{B}\mathbf{a} + \mathbf{a}^T\mathbf{W}\mathbf{a} \tag{1.2.14}$$

Recopilando, el objetivo del análisis discriminante lineal es encontrar combinaciones lineales que maximicen la varianza entre grupos y minimicen la varianza dentro de los grupos. Eso es equivalente a encontrar el vector $\mathbf{a}$ tal que:

$$f(\mathbf{a}) = \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{T}\mathbf{a}} \tag{1.2.15}$$

Es máxima. Si además utilizamos la restricción $\mathbf{a}^T\mathbf{T}\mathbf{a} = 1$. En principio la función objetiva es homogénea, es decir, $f(\mu\mathbf{a}) = f(\mathbf{a})$ Utilizando el método de los multiplicadores de Lagrange derivamos respecto del vector $\mathbf{a}$ tendremos que:

$$L(\mathbf{a}) = \mathbf{a}^T\mathbf{B}\mathbf{a} - \lambda(\mathbf{a}^T\mathbf{T}\mathbf{a} - 1) \tag{1.2.16}$$

al derivarla respecto de $\mathbf{a}$ se obtiene que:

$$\frac{\partial L(\mathbf{a})}{\partial \mathbf{a}} = 2\mathbf{Ba} - 2\lambda\mathbf{Ta} \tag{1.2.17}$$

En consecuencia:

$$\mathbf{Ba} = \lambda\mathbf{Ta} \tag{1.2.18}$$

Si además $\mathbf{T}$ es no singular

$$\mathbf{T}^{-1}\mathbf{Ba} = \lambda\mathbf{a} \tag{1.2.19}$$

Es decir, el vector $\mathbf{a}$ es el vector de valor propio λ, tomando el valor propio máximo de la matriz $\mathbf{T}^{-1}\mathbf{B}$.

Definición 1.2.3. Al valor λ se le conoce como *potencia discriminante* de la combinación $\mathbf{a}$.

Proposición 1.2.1. El desarrollo antes detallado es análogo para $f(\mathbf{a}) = \dfrac{\mathbf{a}^T\mathbf{Ba}}{\mathbf{a}^T\mathbf{Wa}}$ [35]

En el caso de que tomamos el desarrollo para obtener los vectores propios de $\mathbf{W^{-1}B}$ es una matriz con rango $r = \text{mín}(p, q-1)$ entonces podemos utilizar la matriz de cambio de base $\mathbf{U_r}$ que tiene como columnas los r vectores propios con valores propios no nulos de la matriz $\mathbf{W^{-1}B}$. Sea $\mathbf{x}$ el vector a transformar, entonces tenemos que $\mathbf{z} = \mathbf{U_r^T x}$.

Aunque el fin de este método no sea este, si se desea determinar si una nueva observación $\mathbf{x}_0$ basta con calcular la distancia con las medias de los grupos transformadas, es decir, $||\mathbf{z_0} - \mathbf{U_r}\mu_\mathbf{i}||^2$ donde $\mathbf{z}_0$ es el vector en la nueva base y comprobar cual es la menor.

Para determinar cuantas variables canónicas se puede utilizar la varianza entre grupos que explica cada variable canónica.

Proposición 1.2.2. La varianza explicada por cada variable canónica es el valor propio de la matriz $\mathbf{W^{-1}B}$ de su vector propio asociado, su *potencia discriminante*.

Demostración. Sea el vector propio $\mathbf{a}_i$ con valor propio λ_i entonces tendremos que por construcción $\mathbf{a}_i^T\mathbf{Wa}_i = 1$, entonces los vectores son unitarios respecto a la métrica que induce la matriz de covarianzas intra-grupos, entonces la varianza explicada por cada variable es $VE(\mathbf{a}_i) = \mathbf{a}_i^T\mathbf{Ba}_i$, por ser valores propios de $\mathbf{W^{-1}B}$ se cumple que $\mathbf{Ba}_i = \lambda_i\mathbf{Wa}_i$, entonces

$$VE(\mathbf{a}_i)\mathbf{a}_i^T\mathbf{Ba}_i = \lambda_i\mathbf{a}_i^T\mathbf{Wa}_i = \lambda_i \tag{1.2.20}$$

□

Para elegir un número de variables canónicas a usar se puede fijar un umbral por el cual tomemos las m variables cuya varianza explicada acumulada sea superior. De esta manera, se puede llevar a cabo una reducción de la dimensionalidad de los datos de manera parecida a la que se desarrollará en el *Análisis de Componentes Principales.* La principal diferencia con este último es que mientras uno busca representar de una manera simple la variabilidad de los datos, esta busca las direcciones en las cuales las poblaciones se distinguen de manera más significativa.

1.3. Redes Neuronales

Las redes neuronales artificiales, son modelos predictivos basados en el funcionamiento de las propias neuronas del cerebro que reciben señales de entradas de las neuronas con las cuales están conectadas, las procesan y envían el resultado a las neuronas con las que estén conectadas.

La ventaja de este tipo de algoritmos es que en esencia, son un conjunto de parámetros y funciones de activación que pueden ser ajustados para cualquier tarea y cualquier tipo de función a aproximar solo hace falta la complejidad del modelo adecuada según *Hornik, K., Stinchcombe, M., y White, H.* [27].

Una neurona artificial es mucho más simple que una neurona, *López R., E. Balsa-Canto y E. Oñate* [16] definen una neurona en términos matemáticos, pero antes hay que definir los siguientes conceptos, para los cuales se han utilizado como base [62, 15]

Sea un vector aleatorio $\mathbf{x}$ de longitud p, llamamos datos de entrada a los números que se introducen en la neurona que pueden ser los propios valores observados de las

Definición 1.3.1. Llamaremos *pesos sinápticos* ω de una neurona, al vector de p constantes que regulan la importancia de cada entrada en la neurona. A este vector de pesos sinápticos se le puede añadir un término independiente que únicamente se sumará, se le llama *sesgo* y se denota como b.

Definición 1.3.2. Se llama *función de activación*, f de una neurona artificial a la función que transforma la suma ponderada de las entradas para obtener la salida.

Las funciones de activación más habituales son; la función *identidad (En este caso, es como si se hiciera una simple suma ponderada de los datos de entrada)*, la función *sigmoide*, la *tangente hiperbólica* o la función *lineal regularizada* para casos de regresión, es decir, en casos en los que la variable respuesta sea continua. En caso contrario, se pueden utilizar la *función softmax* o la *función de regresión logística* para casos discretos, ya que devuelven valores en el intervalo $[0, 1]$ y se puede asociar con la probabilidad de pertenecer a una clase u otra.

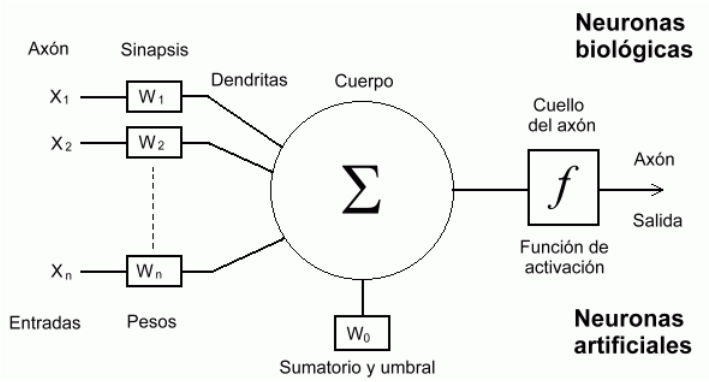
Definición 1.3.3. Una *neurona artificial* procesa una entrada $\mathbf{x}$ de acuerdo con unos *pesos sinápticos* (b, ω) que luego es transformada por una función de activación $f(\mathbf{x})$.

Una vez definidos los elementos que forman una neurona artificial estos se utilizan de manera que se utiliza la función:

$$\begin{aligned} g : \mathbb{R}^p &\longrightarrow \mathbb{R} \\ g(\mathbf{x}) &\longrightarrow g(\mathbf{x}; b, \omega) \end{aligned} \tag{1.3.1}$$

$$g(\mathbf{x}) = f\left(b + \sum_{i=1}^{p} \omega_i x_i\right) \tag{1.3.2}$$

El siguiente diagrama proporciona una forma sencilla de entender el funcionamiento de dicho modelo, incluyendo la analogía de las neuronas biológicas.



La principal ventaja de estos métodos es que las neuronas se pueden conectar entre ellas, es decir, estas se pueden organizar de manera que los datos de salida de un conjunto de neuronas sirvan como entrada del siguiente.

Definición 1.3.4. Se llama *capa de neuronas* al conjunto de neuronas artificiales que tienen el mismo conjunto de datos de entrada y cuyos datos de salida para la siguiente.

Se pueden establecer varios tipos de capas de neuronas [15]:

Definición 1.3.5. Se llama *capa de entrada* a la primera capa de neuronas que recibe los valores de las observaciones y las estandariza (*Se debe entrenar al modelo para ello*). Se establece una

Definición 1.3.6. Se llama *capa oculta* a cada una de las capas intermedias que se utilizan en las redes neuronales.

Definición 1.3.7. Se llama *capa de salida* a la última capa que tiene tantas neuronas como variables respuesta y sus datos de salida son las predicciones de

Para el proceso de ajuste se utiliza de manera habitual el método del gradiente con un conjunto de datos con N observaciones. En el capitulo 11 *Hastie et.al.* [9] se detallan en profundidad el ajuste mediante el método del gradiente. De esta manera, se tiene un proceso que se llama *back-propagation*.

La siguiente imagen es una representación de una red neuornal como un grafo, en el que cada nodo es una neurona, en particular, los nodos azules son neuronas comunes, formando 5 capas ocultas, mientras que las amarillas son capas de escalado, y las rojas de salida.

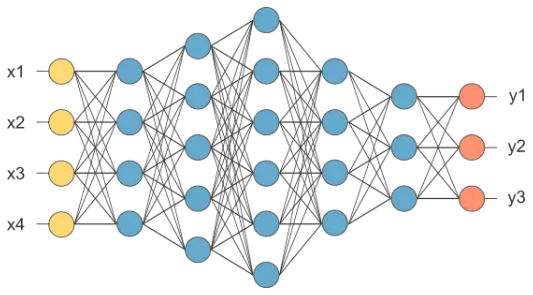


Figura 1.1: Imagen extraída directamente de www.neuraldesigner.com

Si se quisiera expresar el modelo como una única expresión, la expresión que se obtiene al hacer crecer un poco la red neuronal es bastante compleja de interpretar, por ejemplo en el caso anterior se tiene más de 100 parámetros y no es una red demasiado compleja, por tanto, la interpretación del modelo es compleja. Es por ello, que las redes neuronales se utilizan únicamente para fines predictivos [9, 21].

Las principales ventajas de las redes neuronales es que pueden ajustarse a cualquier estructura sin conocerla a priori. Por otro lado, debido a la gran cantidad de parámetros a ajustar pueden provocar *sobre-ajuste* pero para ello se pueden utilizar técnicas de validación cruzada *(Capítulo 5 de [21] o el capítulo 7 de [9])*

En esta memoria se han detallado los tipos más básicos de neuronas hay tareas específicas que este tipo de neuronas no pueden afrontar, por ejemplo, en el caso de datos que proceden de series temporales en las que estados previos influyen en los estados futuros como puede ser predicciones meteorológicas, bursátiles etc... se han desarrollado un tipo más complejo de neuronas llamadas LSTM *(Long-Short Term Memory)* de las que se puede ver su desarrollo y definición además de las propiedades que poseen en [58]. *(Para una descripción más esquemática véase [15])*

1.4. Árboles de Decisión y Bosques Aleatorios

Sea un vector aleatorio $\mathbf{x}$ de longitud p con las variables predictoras e Y la variable respuesta. Se toman N observaciones obteniéndose parejas $(\mathbf{x}_i, y_i)$. De esta manera, tenemos que se puede interpretar que $\mathbf{x}_i \in \mathbb{R}^p$.

Los árboles de decisión son métodos divisivos que dividen el espacio de observaciones $\mathbb{R}^p$ en varias regiones. En cada región, se ajusta un modelo más simple, como una constante.

La ventaja de este tipo de métodos es que son fácilmente interpretables, ya que pueden ser representados mediante un diagrama de tipo árbol. De esta manera, pueden ser utilizados según *Brown et.al.*[5] y *Song Y.Y y Ying, L* [49] describen los principales objetivos de los árboles de decisión, entre los que se encuentran la selección y evaluación de la importancia de las variables, además de que puede tener fines predictivos o incluso de manejo de los datos, valores perdidos etc...

Otra ventaja de este tipo de métodos es que permite trabajar con conjuntos de datos en los que se estudian más variables en comparación de las observaciones, es decir, $p > N$. Por ejemplo, el ejemplo que desarrollan *Díaz-Uriarte y De Andrés* [7], en el que trabajan con un microarray genético.

El problema que más se da en los árboles de decisión es su tendencia al sobre ajuste, ya que de la forma que están conceptualizados, provoca que el sesgo sea pequeño, pero con una gran varianza. Esto en ciertos algoritmos se soluciona con lo que se llama "poda´´

Para comprender mejor cómo se realizan las divisiones en los árboles de decisión es importante familiarizarse con la notación utilizada.

Notación

Definición 1.4.1. Se llama *separación, partición o división* de índice (j, s) [9] a la separación que particiona el espacio inicial dado en las siguientes regiones R_1, R_2:

$$R_1 = \{\mathbf{x}_i \in \mathbb{R}^p / x_{ij} > s\} \quad R_2 = \{\mathbf{x}_i \in \mathbb{R}^p / x_{ij} \leq s\} \tag{1.4.1}$$

Hay que tener en cuenta que esto sería el caso en el que la variable a separar X_j sea continua. Sí X_j es discreta, se pueden tomar dos opciones, hacerlo de la misma manera es decir con semihiperplanos o tomando la división $X_j = s, X_j \neq s$. Depende de como se desee particionar en el caso de variables discretas.

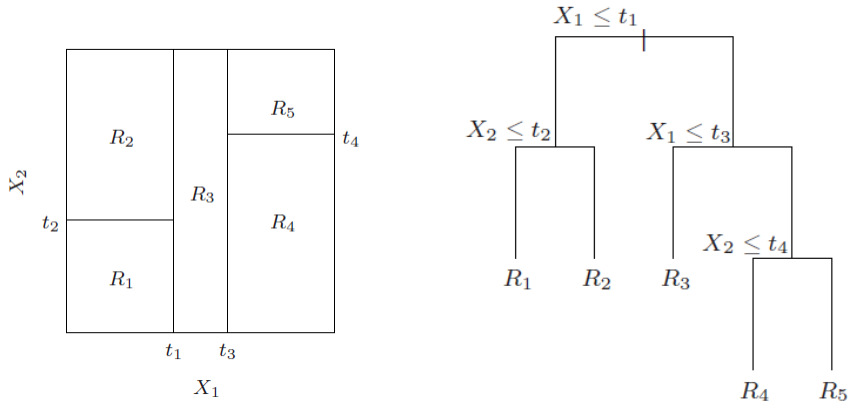
Definición 1.4.2. Se define un *nodo* en un árbol de decisión como cada una de las regiones resultantes después de aplicar una separación.

Definición 1.4.3. Se define un *nodo terminal o nodo hoja* [5] en un árbol de decisión como cada una de las regiones finales resultante de la partición del espacio de observaciones.

Definición 1.4.4. Llamaremos tamaño del árbol T, $|T|$ al número de nodos terminales.

Definición 1.4.5. Se llama profundidad del árbol al número máximo que de divisiones necesarias para llegar a un nodo terminal [9]. *En el caso del diagrama que se incluye después se dan 3 divisiones del espacio tanto para llegar a* R_4, R_5

Las anteriores imágenes procedentes de *Hastie et. al.*[9] muestran el diagrama resultante tras dividir el espacio de observaciones mediante un árbol.



(a) División de $\mathbb{R}^p$ (b) Diagrama resultante

Figura 1.2: Representación de la división de $\mathbb{R}^p$ y el diagrama de árbol resultante

Como se puede ver, cada uno de los nodos terminales representan cada una de las regiones en las que se ha separado el espacio de observaciones.

Dependiendo del tipo de variable respuesta, se utiliza un *Árbol de regresión* cuando se busca predecir una variable continua, y un *Árbol de clasificación* cuando se busca clasificar variables discretas. En nuestro caso, los algoritmos que se describirán pueden manejar tanto variables de entrada discretas como continuas. Se detallará el caso en el que las variables de entrada son continuas.

Sesgo y varianza de un modelo

Un aspecto en el que no se ha indagado en el trabajo por ahora es en la capacidad predictiva de los modelos. Es decir, a la capacidad de obtener dadas nuevas observaciones $\mathbf{x}_0$ de las variables predictoras, hallar el valor que tendría la variable respuesta Y. La razón de que se detalle ahora es que los árboles son modelos los cuales tienden al sobreajuste y algunos autores como *Breiman, L.*[52] o *Divakaran, S.*[20] proponen mecanismos para evitar dicha situación e incluso proponen u tilizar técnicas de *Bagging* para los *Random Forest*[55].

Supóngase variable respuesta sigue un modelo del tipo $Y = f(\mathbf{x}) + \varepsilon$, en el que la variable aleatoria $\varepsilon \sim N(0, \sigma^2)$ y la variable respuesta $Y \sim N(f(\mathbf{x}), \sigma^2)$ conociendo el vector de variables predictoras. Si la $f(\mathbf{x})$ fuera una función lineal y la variable respuesta Y fuera continua, se estaría ante un modelo de regresión lineal por ejemplo.

Consideremos que se toma una muestra con N observaciones, obteniéndose una matriz de datos $\mathbf{X}$, la matriz de respuesta $\mathbf{Y}$ y que se ajusta el modelo y sus

parámetros conocidas conociendo estas N observaciones. Entonces, para una nueva observación de las variables predictoras $\mathbf{x}_0$, se puede definir el siguiente concepto [9, 50]:

Definición 1.4.6. Se llama *error de predicción esperado* de la observación $\mathbf{x}_0$ a la siguiente expresión:

$$EPE(\mathbf{x}_0) = \mathbb{E}((Y - \hat{Y})^2 | \mathbf{x} = \mathbf{x}_0) = \mathbb{E}((Y - \hat{f}(\mathbf{x}_0))^2) \tag{1.4.2}$$

De esta manera, se tiene una forma de medir el rendimiento predictivo de un modelo. Es por ello, también que en aplicaciones de aprendizaje automático en las que se busca hacer predicciones lo más precisas posibles, se divide el conjunto de datos en el conjunto de entrenamiento y de validación ya que una vez ajustado el modelo, se evalúa su capacidad predictiva mediante un estimador del error de predicción esperado con una muestra aparte de los datos de entrenamiento o ajuste.

Este error de predicción se puede descomponer de manera sencilla

Proposición 1.4.1. El error de predicción esperado se puede dividir en un termino irreducible, el sesgo del modelo y la varianza [9]:

$$EPE(\mathbf{x}_0) = \sigma_\varepsilon^2 + Sesgo(\hat{f}(\mathbf{x}_0))^2 + Var(\hat{f}(\mathbf{x}_0)) \tag{1.4.3}$$

Donde el $Sesgo(\hat{f}(\mathbf{x}_0)) = \mathbb{E}(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0))$.

Demostración.

$$\begin{aligned}\mathbb{E}((Y - \hat{f}(\mathbf{x}_0))^2) &= \mathbb{E}((Y - f(\mathbf{x}_0) + f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0))^2) = \\ &= \mathbb{E}(\varepsilon^2) - 2\mathbb{E}(\varepsilon \cdot (f(\mathbf{x}_0) - \hat{f}(\mathbf{x})_0)) + \mathbb{E}((f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)^2) \\ &= \sigma^2 + \mathbb{E}((f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0))^2)\end{aligned}$$

El segundo término es el error cuadrático medio de un estimador, luego se obtiene que:

$$\mathbb{E}((Y - \hat{f}(\mathbf{x}_0))^2) = \sigma^2 + Sesgo(\hat{f}(\mathbf{x}_0))^2 + Var(\hat{f}(\mathbf{x}_0))$$

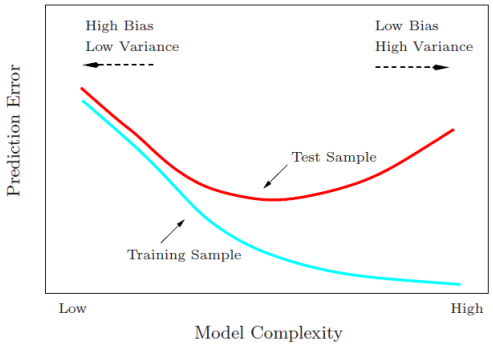
□

Estos dos parámetros están íntimamente relacionados con la complejidad del modelo, ya que cuanto más complejo sea el modelo, el sesgo se reduce de manera importante. Esto es debido a que los puntos del conjunto de entrenamiento están bastante cerca de las funciones aproximadas, en cambio la varianza se dispara, lo que implica que a la hora de hacer predicciones estas no sean lo mejor posible [15].

Para el caso en el que tengamos un modelo lineal que *Hastie et. al.* [9] desarrolla, se ve que la varianza del error esperado medio en observaciones del conjunto de entrenamiento depende de $\frac{p}{N}$, y el sesgo de $\frac{1}{N}$. Por tanto, a mayor numero de observaciones y menor de variables predictoras mejor. Esto se puede observar en el siguiente gráfico que incluyen *Hastie et.al.* [9] de manera que se puede ver a mismo número de observaciones que pasa si aumentamos las variables observadas. La línea

roja representa lo que ocurre con el error de predicción en observaciones fuera del conjunto de ajuste y la azul representa dentro del conjunto de ajuste.

Esta parte se va utilizar, tanto en los árboles como en los random forests, ya que son métodos que al hacer crecer el árbol de decisión el objetivo es el sobreajuste. Es decir, el objetivo en estos casos es bajar el sesgo del predictor, en particular, si se llega al punto en el que todas las observaciones en una región tengan la misma respuesta, entonces el sesgo del predictor es 0. En cambio la varianza de este caso, aumenta por que siguen siendo modelos lineales, y por tanto al reducir el número de observaciones, la varianza en cada región aumenta, lo que hace aumentar la varianza en general del modelo.



En la figura se puede bservar lo qué ocurre en el caso de que un modelo esté en una situación de infraajuste en la parte izquierda, no se ajusta el modelo ni para muestras dentro de los datos ni en el caso que cojamos una que no esté en los datos. En el caso contrario tenemos la parte derecha donde el modelo se ajusta demasiado a los datos recogidos y no tiene capacidad predictiva. Esto puede ocurrir por lo siguiente; que la complejidad del modelo en relación con la cantidad de muestras sea demasiada alta.

Como hemos dicho antes el crecimiento de los árboles tanto de regresión como de clasificación busca sobreajustarse. En la siguiente parte veremos como hay distintos algoritmos que buscan evitar dicho sobreajuste.

1.4.1. Árboles de regresión

A continuación, examinemos el proceso de crecimiento de un árbol de regresión. En este caso, se considera un vector aleatorio $\mathbf{x}$ con p variables predictoras y una variable respuesta continua Y. En particular, se desarrollan los conceptos y proceso llevado a cabo en el algoritmo *CART* de *Breiman, L.* [52].

Supongamos que hemos establecido un número máximo de particiones $M = |T|$. En consecuencia, el objetivo del árbol de decisión es dividir el espacio de observaciones en regiones R_m para $m = 1, \ldots, M$. Cada región R_m está asociada a su *función característica*, que denotaremos como $\mathbf{1}_m$.

En el caso de que se tomen N observaciones, podemos definir el estimador resultante $\hat{f}(\mathbf{x})$ de la siguiente manera:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} \hat{f}_m(\mathbf{x}) \cdot \mathbf{1}_m(\mathbf{x}) \tag{1.4.4}$$

Esto significa que en cada región R_m, se realiza una regresión utilizando los datos correspondientes a esa región. Generalmente, se busca una aproximación lo más sencilla posible en cada R_m, por ejemplo, utilizando una constante. [9, 3].

Si se aplica el método de los mínimos cuadrados, con la restricción requerida, se pueden definir las constantes $\hat{c}_m$ de la siguiente manera:

$$\hat{c}_m = \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} y_i \tag{1.4.5}$$

Donde N_m es el número de observaciones del total que hay en R_m. Es decir, $\hat{c}_m$ es la media muestral de las respuestas de las observaciones que entran en la región y esto provoca que $\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} \hat{c}_m \cdot \mathbf{1}_m(\mathbf{x})$. Es decir, se predice $\mathbf{x} \in R_m$ como la constante $\hat{c}_m$. [9]

Una vez se conoce como se va a dar la estimación final hay que saber cómo llegar a la mejor partición. A este proceso de elegir las particiones y elegir dichas particiones (j, s).

En regresión, hay que elegir j y s de tal manera que las regiones resultantes R_{m_1}, R_{m_2} son aquellas en las que se minimiza la siguiente expresión:

$$\sum_{i/\mathbf{x}_i \in R_{m_1}} (y_i - \hat{c}_{m_1})^2 + \sum_{i/\mathbf{x}_i \in R_{m_2}} (y_i - \hat{c}_{m_2})^2 \tag{1.4.6}$$

Es decir, el objetivo es encontrar las separaciones que minimicen la suma de los errores cuadráticos. A este criterio se le llama *CART*, aunque *Biau, G. y Scornet, E.*[3] lo plantea de manera distinta. Podemos definir el error cuadrático medio de cada región de la siguiente manera [9].

Definición 1.4.7. Se llama error cuadrático medio de una región R_m a $Q_m(T) = \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2$.

Por tanto, podemos definir una función de coste general

$$Q(T) = \sum_{m=1}^{M} \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2 \tag{1.4.7}$$

Ahora veamos como hacer crecer demasiado el árbol y por tanto, tener demasiadas

En el caso de que $\mathbf{x}_0 \in R_m$ y las mismas suposiciones que en la sección anterior, una observación que no está entre las recogidas en la matriz de datos, el error de predicción esperado cumple lo siguiente:

$$EPE(\mathbf{x}_0) = \sigma^2 + (Sesgo(\hat{f}_m(\mathbf{x}_0)))^2 + Var(\hat{f}(\mathbf{x}_0)) \tag{1.4.8}$$

Como $\mathbf{x}_0 \in R_m$ entonces, se tiene que:

$$EPE(\mathbf{x}_0) = \sigma^2 + (Sesgo(\hat{f}_m(\mathbf{x}_0)))^2 + Var(\hat{f}_m(\mathbf{x}_0)) \tag{1.4.9}$$

En particular, $\hat{f}_m(\mathbf{x}) = \hat{c}_m$, por tanto, es un estimador insesgado cuya varianza es $\frac{\sigma^2}{N_m}$ ya que procede de hacer una media muestral la variable $Y \sim N(0, \sigma^2)$, por tanto:

$$EPE(\mathbf{x}_0) = \sigma^2 + \frac{\sigma^2}{N_m} \tag{1.4.10}$$

Por tanto, hacer crecer un árbol, es decir hacer que tenga más nodos términales, va a provocar que N_m sea menor, por lo tanto, hacer crecer un árbol demasiado hace que estemos en un caso de sobre ajuste. Es por ello, que se crean métodos como la *poda*.

Para empezar hay que definir lo que significa la poda de un árbol.

Definición 1.4.8. Se llama *poda* al proceso en el que dado un árbol inicial de tamaño T_0 se revierten ciertas particiones terminales que no aportan en la relación coste-complejidad, es decir, aumentan demasiado la complejidad (*Aumentando la varianza*), sin reducir el coste en exceso.

Divakaran, S. [20], propone añadir un término al coste del árbol.

$$Q_\alpha(T) = \sum_{m=1}^{M} \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2 + \alpha|T| \tag{1.4.11}$$

Donde α es un término para controlar la complejidad del árbol, es decir, $\alpha = 0$ es el criterio habitual, mientras que cuanto mayor sea, más pequeños serán los árboles, ya que penaliza más el tamaño.

El proceso que sugiere *Divakaran S.*[20], basado en el que da *Breiman L.*[52] viene dado por los siguientes pasos:

- Primero prepara el conjunto de datos para poder realizar validación cruzada *(Véase el capítulo 7 de [9], en particular el método de K-folds)*
- Se hace una partición binaria hasta tener un árbol de gran tamaño T_0,parando con cualquiera de los criterios habituales como pueden ser un máximo de nodos terminales o un mínimo de observaciones por nodo.
- Se aplica la poda al árbol utilizando el coste que penaliza el tamaño del propio árbol, teniendo en cuenta el parámetro α.
- *Divakaran S.*[20] aplica el método de K-folds para elegir el parámetro α

Este método nos devuelve un subárbol T_α con el $Q_\alpha(T)$ menor posible. En caso de que $\alpha = 0$, $T_\alpha = T_0$.

En este caso, se está teniendo un intercambio entre varianza y sesgo, de manera que un árbol tiene menor sesgo, ya que se ajusta de mejor manera a los datos de

ajuste pero a la hora de hacer predicciones no son las mejores. De esta manera, se da a cambio de un poco más de sesgo, se reduce la varianza. (*Véase el capitulo 7 de [9] o el capitulo 5 de [21]*)

Otro algoritmo de selección de las particiones es el algoritmo *C4.5* de *Quinlan, J*[59] y aunque se desarrollará aplicado a árboles de clasificación, su implementación para árboles de regresión es análoga, cambiando únicamente el criterio, ya que el dado no es válido para variables respuesta continuas.

1.4.2. Árboles de Clasificación

Sea ahora el caso en el que las variables predictoras son como antes, dadas por un vector aleatorio $\mathbf{x}$ de longitud p , discretas o continuas de manera indiscretas, pero en el que la variable respuesta es una variable aleatoria con L posibles valores.

Tómense ahora N observaciones simultáneas del vector $\mathbf{x}$ y la variable respuesta Y para el cual se ha hecho crecer un árbol T de tamaño M, entonces, podemos denotar de la siguiente manera [20, 5]

$$\hat{p}_{lm} = \text{Proporción de observaciones en las que } y_i = l, \text{ en la región } R_m. \quad (1.4.12)$$

Teniendo en cuenta esto, se pueden definir los siguientes conceptos.

Definición 1.4.9. Se dice que un nodo correspondiente a la región R_m es puro, si $\exists l_0 / p_{l_0 m} = 1$ y $\hat{p}_{lm} = 0 \quad \forall l \neq l_0$ [20, 9, 21, 5].

Definición 1.4.10. Se define la *impureza de un nodo*, correspondiente a la región R_m como [5]:

$$1 - \max_{l \in L} \hat{p}_{lm} \quad (1.4.13)$$

Es decir, si se hablara de coste, estamos asumiendo que en esa región R_m, la l tal que $\hat{p}_{lm}$ es máxima es la correcta. Entonces, la impureza se puede interpretar como la "probabilidad"de error.

Otra forma de medir la impureza son el índice Gini y la entropía. *Hastie et. al.* y *Divakaran, S.* [9, 20], definen el índice Gini de la siguiente manera:

Definición 1.4.11. Se llama *índice Gini* de una región R_m a la siguiente expresión [9, 21]:

$$G = \sum_{l=1}^{L} \hat{p}_{lm}(1 - \hat{p}_{lm}) \quad (1.4.14)$$

Esta medida es la varianza que tiene un nodo.

Con esta medida se desarrolla el algoritmo de *CHAID*, en el que la variable X_j que se elige, se elige mediante un test de similitud y aquellas particiones de las planteadas que sean demasiado parecidas se fusionan. *(Véase [61] para leer el desarrollo inicialmente propuesto por Kass G.V.)*

Definición 1.4.12. Se llama *entropía* a la siguiente medida [5]:

$$H(\hat{p}_{lm}) = -\hat{p}_{lm}log_2(\hat{p}_{lm}) - (1-\hat{p}_{lm})log_2(1-\hat{p}_{lm}) \tag{1.4.15}$$

Esta medida toma valores en el intervalo $[0,1]$ siendo $H(\frac{1}{2}) = 1$ el máximo cuando hay tantas clasificaciones posibles correctas como incorrectas. Y tiene el mínimo en el 0 y el 1 donde toma el valor 0 ya que en ambos casos hay solo erróneas o correctas.

Entonces se puede definir la entropía total del nodo m como la siguiente expresión:

$$H(R_m) = -\sum_{l=1}^{L}(\hat{p}_{lm})log(\hat{p}_{lm}) \tag{1.4.16}$$

El algoritmo *C4.5* de *Quinlan J.R.*[59] hace uso de un criterio para hacer crecer el árbol de clasificación. Pero para ello hay que definir el concepto de ganancia de información y de ratio de ganancia.

Definición 1.4.13. Supóngase que se tiene un nodo padre que se divide en m nodos hijos, entonces la *ganancia de información* se define de la siguiente manera:

$$Ganancia = H(padre) - \sum_{i=1}^{M}\frac{N_m}{N_{padre}}H_i \tag{1.4.17}$$

Es decir, es la entropía del nodo padre menos la media ponderada de las entropías de cada uno de los nodos.

Definición 1.4.14. Llamamos ratio de ganancia a la siguiente cantidad que está entre 0 y 1:

$$\frac{Ganancia}{H_{padre}} \tag{1.4.18}$$

Por tanto, un valor cercano a 1 aporta un gran cambio y un valor cercano a 0 es que casi la ganancia es nula.

Teniendo esta medida se puede plantear el método de construcción de árboles *C4.5* [60] de la siguiente manera.

Para cada tipo de variable se da un tipo sistemático de particiones del espacio, en el caso de que $X_j, j = 1 \dots p$ sea una variable aleatoria discreta sin orden, que toma l valores distintos en el nodo que se va a particionar, entonces se hacen l particiones y en cada partición tomará cada uno de los valores.

En el caso de que la variable X_j sea continua se hace una partición binaria a partir de la media en el nodo padre. Es decir, tomamos la regla discriminante $X_j > c$, $X_j \geq c$, donde c es la media de la variable X_j en el nodo padre.

Una vez calculada las particiones para cada una de las variables, se calcula la ganancia de cada una de las particiones, y se toma la que mayor ratio de ganancia tenga.

De esta manera, se obtiene otro algoritmo voraz para el crecimiento del árbol. Al igual que en el resto, se puede hacer un proceso de poda para evitar el sobre ajuste.

1.4.3. Algoritmo Random Forest

El término bosque aleatorio se puede referir a dos conceptos distintos, uno el hecho de utilizar varios árboles sin importar como se obtienen y luego utilizar la media de los resultados o el voto por mayoría dependiendo del tipo de variable respuesta sea discreta o continua.

El término *random forest* se refiere al algoritmo de *Breiman, L.* [55]. Este tipo de algoritmo si hace hincapié en el método con el se construyen los árboles.

El proceso de construcción y de utilización de los *random forest* lo detallan *Biau, G. , Scornet, E. y Breiman, L.*[3, 4] :

- En cada árbol se utiliza una muestra aleatoria sin reemplazamiento del conjunto de datos iniciales. La razón es buscar una muestra *bootstrap* de los datos iniciales *(Véase [53] para más detalles de la técnica.)*
- Se establece un número j_{try} por el usuario. Para cada partición a realizar se escogen de manera aleatoria j_{try} variables y se elige la que más se ajusta al criterio dado. Y se hace crecer el árbol todo lo necesario o hasta un criterio dado por el usuario.
- Una vez se tienen crecidos los árboles, se toma como predicción de una nueva observación la media de las predicciones de los datos en el caso de que la variable respuesta sea continua. En el caso de que sea una variable discreta, se toma el voto por mayoría, es decir se toma el valor que más veces haya salido en todos los árboles.

Breiman, L. [4] detalla las propiedades básicas de un random forest. Entre estas propiedades se encuentran el sesgo, la varianza y sus acotaciones. Además *Biau G. y Scornet, E.* [3] destacan que no es habitual ver un estudio profundo de estos métodos debido

Breiman, L. [54] llama a este tipo de métodos los métodos de *bagging*. Este tipo de métodos lo que busca es reducir la varianza de métodos con sobreajuste utilizando la media y el teorema central del límite de manera que el predictor final es media de los predictores, lo que consigue reducir la varianza de manera importante. Además detalla que en los casos en los que los predictores no tengan esa varianza alta en comparación con el sesgo, el método de *bagging* no tiene un gran impacto e incluso puede resultar perjudicial debido al incremento de coste computacional que se puede dar

Capítulo 2

Métodos no supervisados

Los métodos no supervisados son aquellos que utilizan la matriz de datos en la cual no se tiene un conocimiento previo, es decir, no intenta ajustarse a una variable respuesta conocida. Esto provoca que no se tenga una *función de pérdida* en el sentido que se ha definido anteriormente [9].

El objetivo de la mayoría de métodos no supervisados es estudiar la estructura de los datos, con el objetivo de encontrar información subyacente a los datos. En este proceso, se pueden observar distintas características, como puede ser la variabilidad del conjunto de los datos, la homogeneidad de las observaciones etc...

Al estudiar la estructura de los datos y dar los puntos comunes a todos los datos, se puede llevar a cabo una reducción de la dimensionalidad, es decir, nuestro problema puede pasar a explicarse con menos variables de manera que se simplifica la compresión [1] y a veces los sucesivos análisis.

En particular, métodos como el Análisis factorial o el Análisis de componentes principales, permiten en casos en los que las ciertas variables estén muy correladas [36], crear nuevas variables que representen la misma información o gran parte de ella, en el caso de las componentes la variabilidad general y en el caso de los factores la variabilidad común a todas las variables. En cambio en el análisis de clústers centra el foco la variabilidad de las observaciones, teniendo en cuenta la homogeneidad dentro de los clusters y heterogeneidad entre clusters distintos, de esta manera, se pueden identificar las distintas observaciones como el cluster al que pertenecen.

Aunque no se tenga una función de pérdida como se daba en el caso de los métodos supervisados, se puede dar en estos casos también una medida de la bondad de ajuste que suele depender de la característica observada en cada uno de los casos. Esto puede ser en el caso de los factores la variabilidad común explicada [23], o la proporción de variabilidad explicada en cada componente [6].

Por falta de tiempo y de extensión del trabajo, no se desarrollarán técnicas como las reglas de asociación o el escalado multidimensional para los cuales *Hastie et.al.* aporta una visión introductoria a los mismos [9]. Por otro lado, *James et.al.*[21] desarrollan de manera similar con la particularidad que además lo implementan en el entorno `R` con ejemplos. *(Veánse también [36] y [35]).*

2.1. Análisis de Componentes Principales

El Análisis de componentes principales es una de las técnicas multivariantes más antiguas. En primera instancia, esta técnica fue desarrollada en paralelo por *Pearson, K.* [43] y *Hotelling* [44] con distintos enfoques.

Pearson, K.[43] en su aproximación, buscó la forma de ajustar de mejor manera puntos de un espacio de dimensión p a una recta o plano. Es decir, Pearson buscó un enfoque de optimización geométrica que llevó a las componentes principales.

Por el otro lado, *Hotelling, H.* [44], partiendo de un conjunto de datos provenientes de estudiar un conjunto de variables, buscaba un subconjunto de variables menor que pudiera determinar de igual manera o parecida los datos. Hotelling maximizó la contribución que aportaba cada componente a la varianza. Al utilizar para este fin el método de los multiplicadores de Lagrange, obtuvo el problema de valores propios que se desarrolla en esta sección.

Aún así, el trabajo de Hotelling tenia ciertas diferencias con el enfoque que actualmente se le da actualmente a las componentes principales [11], por ejemplo, consideraba que las variables originales debían ser combinaciones de las componentes y no al revés. Tampoco usaba la notación matricial ni la matriz de covarianzas, en su lugar utilizaba la de correlaciones. Añadir que aunque en esta memoria se trabajará con la matriz de covarianzas, *Chatfield, C. y Collins A.J.*[6] desarrollan ambos enfoques.

Según *Abdi, H* [1] esta técnica tiene varios objetivos, entre los cuales están el extraer la información más importante de los datos, reduciendo la dimensionalidad y simplificando la descripción en el proceso de los mismos. Y por último, analizar la estructura de los datos.

De esta manera, variables bastante correladas, se simplificarán en una que tenga toda la variabilidad de ambas [6, 36] ya que a fin de cuentas la información que expresan estaría repetida.

Para empezar, se detallan cómo se calculan las componentes principales solo teniendo en cuenta las propiedades del vector aleatorio $\mathbf{x}$ de longitud p. Tras esto, se tomarán N observaciones con las que poder construir estimadores de las componentes y se analizará como esta reducción de la dimensionalidad es la que permite una mejor reconstrucción de los datos mediante el Teorema de Eckart-Young [39].

Para desarrollar esta última parte en la que se habla de la reconstrucción de los datos se da una idea de una medida sobre las matrices, en particular, la norma de Frobenius que detallan *Eckart C. y Young, G* [39] sin llamarla de esa manera y demuestran que es invariante ante transformaciones ortogonales.

Y por último, se detalla un ejemplo para ilustrar la interpretación de los resultados obtenidos tras aplicar las técnicas detalladas y hallar ciertos parámetros de bondad de ajuste, además de mostrar aplicaciones reales de este método.

2.1.1. Definición y cálculo de las Componentes

Sea un vector aleatorio de longitud p, $\mathbf{x} = [X_1, \dots X_p]$ con una distribución normal p-multivariante $N_p(\mu, \mathbf{\Sigma})$

Definición 2.1.1. *Cuadras C.M.*,[8] define las componentes principales

$$Z_j = v_{1j}X_1 + \dots v_{pj}X_p = \mathbf{v}_j^T\mathbf{x}^T \quad j = 1 \dots p \tag{2.1.1}$$

Donde $\mathbf{v}_j$ es un vector columna con p escalares y la nueva variable aleatoria Z_j cumple lo siguiente:

- Si $j = 1$ $Var(\mathbf{z}_1)$ es máxima restringido a $\mathbf{v}_1^T\mathbf{v}_1 = 1$
- Si $j > 1$ debe cumplir:
  - $Cov(\mathbf{z}_j, \mathbf{z}_i) = 0 \quad \forall i \neq j$
  - $\mathbf{v}_j^T\mathbf{v}_j = 1$
  - $Var(Z_j)$ es máxima.

En resumen, lo que se busca es una nueva base que reúna las direcciones de máxima variación y que sean ortogonales respecto a la matriz de covarianzas, es decir, que sean no correladas.

Chatfield C. y Collins A.J [6], utilizan el método de los multiplicadores de Lagrange para resolver el problema de maximizar $Var(Z_1)$ sujeto a la restricción $\mathbf{v}_1^T\mathbf{v}_1 = 1$. Todo esto con el objetivo de calcular el vector $\mathbf{v}_1$, el cálculo de sucesivas componentes cambia en ciertos aspectos.

Aplicando el método de los multiplicadores, la función objetivo es la varianza de la combinación lineal, es decir, $f(\mathbf{x}) = \mathbf{x}^T\mathbf{\Sigma}\mathbf{x}$ y la restricción aplicada es $g(\mathbf{x}) = \mathbf{x}^T\mathbf{x} = 1$.

Tomando $\mathbf{x} = \mathbf{v}_1$ se puede establecer $L(\mathbf{v}_1) = \mathbf{v}_1^T\mathbf{\Sigma}\mathbf{v}_1 - \lambda[\mathbf{v}_1^T\mathbf{v}_1 - 1]$. Que al derivarla se obtiene:

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{v}_1} &= 2\mathbf{\Sigma}\mathbf{v}_1 - 2\lambda\mathbf{v}_1 \\ &= 2(\mathbf{\Sigma} - \lambda)\mathbf{v}_1\end{aligned}$$

Igualando a 0 tenemos la siguiente ecuación:

$$(\mathbf{\Sigma} - \lambda I)\mathbf{v}_1 = 0 \tag{2.1.2}$$

Para que $\mathbf{v}_1$ sea un vector no trivial, se elige λ de tal manera que $|\mathbf{\Sigma} - \lambda\mathbf{I}| = 0$, es decir, λ es un vector propio de la matriz de covarianzas, $\mathbf{\Sigma}$. Al ser ésta una matriz semidefinido positiva y simétrica, los valores propios son reales y positivos. Por tanto, $\mathbf{v}_1$ es un vector propio de la matriz de covarianza.

La función a maximizar es $Var(Z_1) = Var(\mathbf{v}_1^T\mathbf{x}^T) = \mathbf{v}_1^T\mathbf{\Sigma}\mathbf{v}_1 = \mathbf{v}_1^T\lambda\mathbf{v}_1 = \lambda$, y para maximizarla basta tomar $\lambda = \text{máx}\{\lambda_1 \dots \lambda_p\}$. Reordenando si es necesario, se tiene

que $\lambda = \lambda_1$, por tanto la primera componente es el vector propio Z_1 y además la varianza de la nueva variable cumple $Var(Z_1) = \lambda_1$

Una vez calculada la primera componente principal Z_1, la segunda componente se calcula de manera análoga, maximizando $Var(Z_2) = Var(\mathbf{v}_2^T \mathbf{x}^T)$ condicionada por $\mathbf{v}_2^T \mathbf{v}_2 = 1$. A esta restricción tenemos que añadir la restricción $Cov(Z_1, Z_2) = 0$

Proposición 2.1.1. La condición $Cov(Z_1, Z_2) = 0$ equivale a la condición $\mathbf{v}_2^T \mathbf{v}_1 = 0$.

Demostración. Utilizando que $Z_j = \mathbf{v}_j^T \mathbf{x}^T \quad \forall j = 1, \dots p$, se tiene entonces que :

$$\begin{aligned} Cov(Z_2, Z_1) &= Cov(\mathbf{v}_2^T \mathbf{x}^T, \mathbf{v}_1^T \mathbf{x}) \\ &= \mathbb{E}(\mathbf{v}_2^T (\mathbf{x}^T - \mu^T)(\mathbf{x}^T - \mu^T)^T \mathbf{v}_1) \\ &= \mathbf{v}_2^T \mathbb{E}((\mathbf{x}^T - \mu^T)(\mathbf{x}^T - \mu^T)^T) \mathbf{v}_1 \\ &= \mathbf{v}_2^T \mathbf{\Sigma} \mathbf{v}_1 \\ &= \mathbf{v}_2^T \lambda_1 \mathbf{v}_1 \end{aligned}$$

De manera que, si $\mathbf{v}_2^T \lambda_1 \mathbf{v}_1 = 0 \Rightarrow \mathbf{v}_2^T \mathbf{v}_1 = 0$, luego son vectores ortogonales entre sí. □

Observación: Esta proposición se puede extender de manera simple al caso de tener que calcular la j-ésima componente principal habiendo calculado las anteriores de las cuales se sepan los valores propios asociados.

Corolario 2.1.1. Las componentes principales son todas ortogonales entre sí.

Para $k = 2$, se dan dos restricciones, $\mathbf{v}_2^T \mathbf{v}_2 = 1$ y además $\mathbf{v}_1^T \mathbf{v}_2 = 0$. Para este caso existen λ, ϕ de manera que la función a maximizar es:

$$L(\mathbf{v}_2) = \mathbf{v}_2^T \mathbf{\Sigma} \mathbf{v}_2 - \lambda[\mathbf{v}_2^T \mathbf{v}_2 - 1] - \phi(\mathbf{v}_1^T \mathbf{v}_2) \tag{2.1.3}$$

Que al ser derivado respecto $\mathbf{v}_2$ obtenemos:

$$2\mathbf{\Sigma} \mathbf{v}_2 - 2\lambda \mathbf{v}_2 - \phi \mathbf{v}_1 = 0 \tag{2.1.4}$$

Que al multiplicar todo por $\mathbf{v}_1^T$ resulta que $\phi = 0$. De esta manera, se obtiene en la ecuación lo mismo que en el cálculo de la primera.

Por tanto, $\lambda = \lambda_2$ que es el segundo valor propio más grande, y $\mathbf{v}_2$ es el vector propio de valor propio λ_2.

El proceso para calcular el resto de componentes principales es análogo únicamente hay que tener en cuenta que se debe dar la ortogonalidad entre las distintas componentes respecto de la varianza.

Por ende, se obtiene que las componentes principales vienen dadas por los vectores propios de la matriz de covarianzas $\mathbf{\Sigma}$. Además sabemos que $Var(\mathbf{v}_j^T \mathbf{x}^T) = \lambda_j$, $j = 1, \dots p$, donde λ_j es el j-ésimo valor propio más grande.

En esencia, calcular las componentes principales es calcular una base ortonormal que cumple una ciertas condiciones. Por lo tanto, podemos definir lo siguiente:

Definición 2.1.2. Se llama matriz de cargas $\mathbf{V}$ a la matriz ortogonal que tiene por columnas a los $\mathbf{v}_j$. Dichas columnas forman una base ortonormal del espacio $\mathbb{R}^p$.

Observación: Se utilizará la condición de que al ser $\mathbf{V}$ una matriz ortogonal $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ y por tanto $\mathbf{V}^{-1} = \mathbf{V}^T$

Esa matriz nos permite calcular las componentes de la siguiente manera

$$\mathbf{z} = \mathbf{x}\mathbf{V} \tag{2.1.5}$$

Donde el vector $\mathbf{z} = Z_1, \ldots, Z_p$, y donde cada $Z_j = \mathbf{v}_j^T\mathbf{x}$

2.1.2. PCA muestral

Ahora, veamos como esto se puede aplicar a una matriz de datos con la matriz de cuasivarianzas muestral o con la matriz de covarianzas.

Sea $\mathbf{x}$ el vector aleatorio de longitud p definida de la misma manera. A continuación se toman N observaciones independientes de este vector y se obtiene la matriz de datos $\mathbf{X}$ de tamaño $N \times p$, en estos casos las componentes principales se definen de la misma manera.

Definición 2.1.3. Dado el vector aleatorio de longitud p del cual se han realizado N observaciones entonces se puede definir la j-ésima componente de la i-ésima observación $\mathbf{z}_{ij} = \mathbf{x}_i\mathbf{v}_j, i = 1 \ldots N, j = 1, \ldots p$.

Donde el vector $\mathbf{v}_j$ cumple las mismas condiciones que para las variables aleatorias.

Por tanto, el proceso que se detalla para un vector aleatorio $\mathbf{x}$ con matriz de covarianzas $\mathbf{\Sigma}$ se puede extender a este caso en el conocemos la matriz de covarianzas muestrales $\mathbf{S}$.

Con el objetivo de hacer las demostraciones más sencillas y compactas tomaremos la matriz $\mathbf{X}$ como la matriz centrada $\overline{\mathbf{X}}$, es decir:

$$\overline{x}_{ij} = x_{ij} - \overline{x}_j \tag{2.1.6}$$

Donde $\overline{x}_j$ es la media muestral de la j-ésima variable. Esto hace que $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$. Esto permite hablar de los valores y vectores propios de $\mathbf{S}$ y de $\mathbf{X}^T\mathbf{X}$ indistintamente, ya que los vectores propios son los mismos y los valores propios son proporcionales.

2.1.3. Reconstrucción de los datos

Lo más importante de esta técnica es que si la varianza se puede explicar por las m primeras componentes, entonces se puede afirmar que la dimensionalidad efectiva de los datos es $m < p$ [6].

Otra visión del problema es que se quiere buscar una proyección sobre una subvariedad de dimensión $m < p$ que contenga la máxima variabilidad posible de los datos y que brinde una mayor capacidad de interpretación de los datos, ya que en el caso de que $m = 2$ o $m = 3$ se podrán hacer representaciones gráficas de manera sencilla. En virtud de conseguir esto se deben definir los siguientes conceptos:

Definición 2.1.4. Dada una matriz $\mathbf{X} \in \mathbb{M}_{N\times p}(\mathbb{R})$ existe la descomposición en valores singulares *(SVD en inglés)*[1]:

$$\mathbf{X} = \mathbf{UDV}^T \tag{2.1.7}$$

Donde:

- $\mathbf{U}$ matriz ortogonal y de tamaño $N \times N$
- $\mathbf{D}$ matriz de tamaño $N \times p$ diagonal, cuyos elementos no nulos son los valores singulares $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$ que son los valores propios de la matriz $\mathbf{X}^T\mathbf{X}$ y $r = rg(\mathbf{X})$
- $\mathbf{V}$ matriz ortogonal y de tamaño $p \times p$.

Observación: En lo restante, se considerará que $N > p$ y que el rango de la matriz de datos es máximo, es decir, que el rango es p. En el caso contrario, se puede reducir las matriz $\mathbf{D}$ a una matriz de tamaño $r \times p$, con las columnas no nulas y por otro lado la matriz $\mathbf{U}$ de tamaño $N \times r$ a la que se le pueden quitar las columnas correspondientes a los vectores propios de valor propio nulo.

Proposición 2.1.2. La matriz $\mathbf{V}$ de tamaño $(p \times p)$ es la matriz que contiene los vectores para hacer la combinación lineal que definen las componentes principales. Es decir es la matriz con la base de diagonalización de $\mathbf{X}^T\mathbf{X}$ [40].

Demostración. La matriz $\mathbf{X}^T\mathbf{X}$ es la matriz de covarianzas $(p \times p)$ por la descomposición en valores singulares tenemos que:

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= (\mathbf{UDV}^T)^T(\mathbf{UDV}^T)\\ &= \mathbf{VD}^T\mathbf{U}^T\mathbf{UDV}^T\\ &= \mathbf{VD}^T\mathbf{DV}^T\end{aligned}$$

Donde la matriz $\mathbf{D}^T\mathbf{D}$ es una matriz diagonal de tamaño $p \times p$ cuyos elementos son los cuadrados de los valores singulares de $\mathbf{X}$, que son a su vez los valores propios de $\mathbf{X}^T\mathbf{X}$.

Añadiendo la condición de ortogonalidad de $\mathbf{V} \Rightarrow \mathbf{V}^{-1} = \mathbf{V}^T$ es fácil ver que la matriz $\mathbf{V}$ es la matriz cuyas columnas son los vectores propios de $\mathbf{X}^T\mathbf{X}$. □

Proposición 2.1.3. La matriz $\mathbf{U}^T$ es la matriz que contiene la base de diagonalización de la matriz $\mathbf{XX}^T$ [40].

Demostración. Se sigue un razonamiento análogo pero con la matriz $\mathbf{XX}^T$

$$\begin{aligned}\mathbf{XX}^T &= (\mathbf{UDV}^T)(\mathbf{UDV}^T)^T\\ &= \mathbf{UDV}^T\mathbf{VD}^T\mathbf{U}^T\\ &= \mathbf{UDD}^T\mathbf{U}^T\end{aligned}$$

Entonces, tenemos que como es ortonormal $\mathbf{U}^T = \mathbf{U}^{-1}$ de esta manera, se puede comprobar al ser $\mathbf{DD}^T$ una matriz diagonal $N \times N$ que efectivamente se cumple la proposición. □

Corolario 2.1.2. El cálculo de las componentes principales mediante los valores propios de la matriz de datos $\mathbf{X}^T\mathbf{X}$ es equivalente a calcular la descomposición en valores singulares de la misma.

El problema de la reconstrucción es demostrar que la matriz reducida, que se definirá más tarde, es la mejor aproximación de rango menor a la matriz de datos, para ello, tenemos que definir la siguiente norma sobre las matrices adecuada a la situación.

Definición 2.1.5. Sea $\mathbf{A} \in \mathbb{M}_{N\times p}(\mathbb{R})$ definimos la *norma de Frobenius* de la matriz $\mathbf{A}$ como [41]:

$$||\mathbf{A}||_F = (tr(\mathbf{A}^T \cdot \mathbf{A}))^{\frac{1}{2}} = \left(\sum_{i=1}^{N}\sum_{j=1}^{p} a_{ij}^2\right)^{\frac{1}{2}} \tag{2.1.8}$$

Proposición 2.1.4. La norma de Frobenius es invariante a transformaciones ortogonales

Demostración. Sea $\mathbf{U}$ una matriz ortogonal, que cumple $\mathbf{U}^T \cdot \mathbf{U} = \mathbf{U} \cdot \mathbf{U}^T = \mathbf{I}$, sea una matriz cualquiera $\mathbf{A}$, entonces:

$$\begin{aligned} ||\mathbf{U}\cdot\mathbf{A}||_F^2 &= tr((\mathbf{UA})^T \cdot (\mathbf{UA})) \\ &= tr((\mathbf{A}^T\mathbf{U}^T)\cdot\mathbf{UA})) \\ &= tr(\mathbf{A}^T\mathbf{A}) \\ &= ||\mathbf{A}||_F^2 \end{aligned}$$ □

Se ha elegido la norma de Frobenius en particular por la siguiente propiedad:

Proposición 2.1.5. Dada una matriz de datos $\mathbf{X}$ de tamaño $N \times p$ entonces

$$||\mathbf{X}||_F^2 = (N-1)\sum_{j=1}^{p} s_{jj}^2 \tag{2.1.9}$$

Donde las s_{ii}^2 son las varianzas muestrales.

Demostración. Debido a la centralidad impuesta a la matriz $\mathbf{X}$, sabemos que la matriz de covarianzas es $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$ por tanto,se tiene que utilizar la definición de la norma:

$$\begin{aligned} ||\mathbf{X}||_F^2 &= tr(\mathbf{X}^T\mathbf{X}) \\ &= (N-1)tr(\mathbf{S}) \\ &= (N-1)\sum_{i=1}^{p} s_{ii}^2 \end{aligned}$$ □

Por tanto, la norma de Frobenius da una imagen del tamaño de la matriz de datos en función de la varianza total de los datos, lo que concuerda con la idea de buscar una matriz que aproxime la matriz de datos con la mínima pérdida de variación de los datos.

Definición 2.1.6. Se llama matriz reducida de orden $m \leq p$ de $\mathbf{X}$ y se denota como $\mathbf{X}_m$, a la matriz $N \times m$ resultado de:

$$\mathbf{X}_m = \mathbf{U}_m \mathbf{D}_m \mathbf{V}_m^T \tag{2.1.10}$$

Donde:

- $\mathbf{U}_m$ matriz ortogonal de tamaño $N \times m$, resultado de tomar de $\mathbf{U}$ únicamente la matriz las m primeras columnas.
- $\mathbf{D}_m$ matriz cuadrada de tamaño m diagonal con los m primeros valores singulares.
- $\mathbf{V}_m$ matriz ortogonal de tamaño $p \times m$ obtenida al tomar las m primeras columnas de $\mathbf{V}$.

Es decir, es la matriz de rango m tras tomar las m primeras columnas y el resto hacerlas nulas o combinaciones lineales, en este caso.

Teorema 2.1.1 (De Eckart-Young)**.** Sea $\mathbf{A}$ una matriz de coeficientes reales de tamaño $N \times p$ y rango r entonces se cumple que [39, 41]:

$$||\mathbf{A} - \mathbf{B}||_F \leq ||\mathbf{A} - \mathbf{A}_m||_F \quad \forall \mathbf{B}/rg(\mathbf{B}) = m \leq r \tag{2.1.11}$$

Johnson R.M. [40] desarrolla la demostración del teorema.

Por tanto, la matriz reducida brinda la mejor aproximación de la matriz de datos teniendo un criterio de aproximación basado en la variación de los datos. En consecuencia se puede

Como conclusión, se puede definir un criterio para elegir el orden de la matriz reducida m. Se puede entonces definir la variación acumulada de la siguiente manera [6]:

$$t_m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_j} \tag{2.1.12}$$

Donde los λ_i son los valores propios de la matriz $\mathbf{S} = \frac{1}{N-1}\mathbf{X}^T\mathbf{X}$.

Ahora el número de componentes m que se van a seleccionar se puede hacer de varias maneras [23], [11], ya sea representando en un gráfico el número de componentes frente a la proporción de variabilidad explicada $t_j, j = 1, \ldots p$. Tomar hasta la primera componente con la cual se supere una cota de variabilidad explicada como el $0,8$ o $0,9$ [11] o desechar las componentes que no expliquen más de una cierta cota inferior. Sin embargo, estas reglas o condiciones suelen ser arbitrarias.

Todos estas condiciones se suelen resumir en desechar las componentes que no aporten a la representación.

Los valores de la matriz de cargas se pueden interpretar como las contribuciones de cada una de las variables a explicar la mayoría de la variabilidad de los datos.

Johnson, R. A., y Wichern, D. W [35] desarrollan un ejemplo sencillo en el cual se estudian 5 variables socioeconómicas de ciertas zonas de Estados Unidos, como

puede ser la población en miles, el porcentage de graduados universitarios, el porcentaje de empleo de mayores de 16 años, el porcentaje de funcionariado y la valor medio de la vivienda.

En este ejemplo, las dos primeras componentes explican el 93 % de la variabilidad y la primera es una diferencia ponderada de el empleo y el empleo público, mientras que la segunda una suma ponderada de las mismas. Por tanto, se puede entender que las variables que mayor importancia tienen en este caso son el porcentaje de funcionarios y de empleados en general.

Otro ejemplo destacable lo desarrolla *Ringnér, W.*[37] en el que se reduce un conjunto de datos con más de 8000 variables y 105 observaciones a uno de solo 63 variables para retener el 90 % y 105 para retener casi el 100 % de la variabilidad total de los datos. Es decir, no siempre se se busca la reducción para representar los datos, en este ejemplo únicamente se busca eliminar datos redundantes en el conjunto total.

Otro ejemplo de aplicación sería el que desarrollan *Gottumukkal, R., y Asari, V.* [38] que mediante una pequeña modificación del algoritmo habitual de las componentes principales en el campo del reconocimiento facial con el objetivo de adaptarse mejor a ciertas condiciones, en las que se tienen grandes variaciones de posición y expresiones. Esto lo consiguen dividiendo cada foto y aplicando de manera independiente a cada parte la técnica.

2.2. Análisis Factorial

El análisis factorial es un conjunto de técnicas que en primera instancia fue desarrollado en el campo de la psicología ya que estos buscaban encontrar las causas que provocaban la variabilidad de los datos que recogían [45].

En 1888 *Galton, F* escuchó la exposición de un articulo que relacionaba las sociedades que habían abandonado un modelo matriarcal con una estructura más compleja sociológicamente hablando. Ante esto, *Galton* no estuvo de acuerdo, ya que detectó un fallo en el razonamiento de su compañero y es por ello, que escribió el artículo [46] dando pie a los trabajos sobre el estudio de las correlaciones.

Es en 1904 cuando *Spearman, C.* [47] viendo el trabajo de *Galton, F.* aplicó esas ideas al estudio de la inteligencia, buscando una medida objetiva de la misma. En este desarrollo, dio las primeras nociones de qué era un factor y calculó una aproximación para un modelo factorial con 6 factores comunes. A pesar de que no especifica el modelo usado, se ha comprobado a posteriori que eran bastante precisos. A partir de este momento, fue cuando el análisis factorial empezó a ser importante y varios psicólogos comenzaron a usarlo de forma habitual. Desde hace años el Análisis factorial, al igual que otras técnicas multivariantes, ha experimentado un nuevo impulso en el desarrollo teórico y práctico debido al auge de la computación.

El Análisis factorial nació con un caracter confirmatorio, en el que se tenían unas ciertas ideas preconcebidas y se buscaba aceptar o rechazar dichas ideas. Actualmente, la mayoría de los investigadores lo utilizan de manera exploratoria para extraer información de los datos.[10]

El objetivo específico del Análisis factorial es, dado un conjunto de variables aleatorias, construir un modelo en el cual con un conjunto de nuevas variables llamadas *factores comunes*, que expliquen la mayor variabilidad posible de las variables iniciales. De esta manera, se descompondrá la varianza de cada una de las variables en una parte común llamada *comunalidad* y otra específica denominada *especificidad*.

La mayor diferencia con el análisis de componentes principales en el que se busca nuevas direcciones que maximicen la varianza explicada, es que en este caso, son de interés tanto las comunalidades como las especifidades ya que nos permiten analizar si una variable da información de manera común al resto o tiene algo diferencial.

Para empezar, empezaremos formalizando el modelo factorial, en particular el modelo multifactorial ortogonal [35] sin pérdida de generalidad. Tras la modelización, se desarrolla el método de estimación de las matrices de carga mediante el método del factor principal [23] y se mencionarán los fundamentos de otros métodos como el de máxima verosimilitud. Seguidamente, se darán algunos detalles sobre el problema de la no unicidad del modelo y soluciones ante ello. Y por último, se dará una introducción a la interpretación de los datos que arroja el modelo factorial.

2.2.1. Formalización

Supóngase un vector aleatorio $\mathbf{x}$ de longitud p con una distribución $N_p(0, \Sigma)$, centrada sin pérdida de generalidad, donde la matriz de covarianzas $\mathbf{\Sigma}$ es simétrica y definido positiva. El modelo factorial afirma lo siguiente [6]:

$$X_j = \lambda_{1j}F_1 + \ldots + \lambda_{mj}F_m + \psi_j U_j \quad j = 1 \ldots p \tag{2.2.1}$$

Donde:

- λ_{jk} son las *cargas del factor común* k-ésimo de la j-ésima variable.
- F_k es el k-ésimo factor común
- ψ_j es la *carga específica* de la variable X_j
- U_j es el *factor específico* para la variable X_j

En este caso haremos las siguientes suposiciones[8]:

- Los factores comunes F_k son variables aleatorias que siguen una distribución marginal $N(0, 1)$. Además se supondrá que $Cov(F_k, F_{k'}) = 0, k \neq k'$, de manera que el vector aleatorio $\mathbf{f}$ de longitud m con distribución $N_m(0, \mathbf{I})$. Y son completamente independientes de los factores específicos.
- Los factores específicos U_j son variables aleatorias con una distribución normal $N(0, 1)$ no correladas, que conjuntamente forman el vector aleatorio $\mathbf{u} \sim N_p(0, \mathbf{I})$ de longitud p. Son completamente independientes de los factores comunes.

En consecuencia, se pueden definir las siguientes matrices:

Definición 2.2.1. Llamaremos *matriz de cargas de los factores o matriz de cargas*, $\mathbf{\Lambda}$ de tamaño $p \times m$ a la matriz :

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pm} \end{pmatrix} \tag{2.2.2}$$

Definición 2.2.2. Llamaremos matriz específica a la matriz diagonal $\mathbf{\Psi}$ de tamaño $p \times p$ a aquella que tiene los términos ψ_j donde $j = 1, \ldots, p$:

$$\mathbf{\Psi} = \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{pmatrix} \tag{2.2.3}$$

Observación: En distintas fuentes no aparece la matriz $\mathbf{\Psi}$ ya que los factores específicos $U_j, j = 1 \ldots p$ en esos casos no se consideran variables estandarizadas no correladas, como sería el caso descrito. Por ejemplo, *Mardia, K.V.* [48] denotan como $\mathbf{\Psi}$ a la matriz de covarianzas del vector $\mathbf{u}$.

De esta manera, se puede dar una versión matricial de la expresión 2.2.1:

$$\mathbf{x}^T = \Lambda \mathbf{f}^T + \mathbf{\Psi} \mathbf{u}^T \tag{2.2.4}$$

Ya que se consideran los vectores aleatorios $\mathbf{x}, \mathbf{f}, \mathbf{u}$ vectores ordenados como filas.

Proposición 2.2.1. La varianza de la variable aleatoria X_j, σ_j^2 se puede descomponer de la siguiente manera:

$$\sigma_j^2 = \sum_{k=1}^{m} \lambda_{kj}^2 + \psi_j^2 \tag{2.2.5}$$

Demostración.

$$\mathbb{E}(X_j^2) = \sigma_j^2 = \sum_{k=1}^{m} \lambda_{kj}^2 \mathbb{E}(F_k^2) + \sum_{k=1}^{m} \sum_{n=1, n\neq k}^{m} \lambda_{kj}\lambda_{nj}\mathbb{E}(F_k \cdot F_n)$$
$$+ \sum_{k=1}^{m} \lambda_{kj}\psi_j \mathbb{E}(F_k \cdot U_j) + \psi_j^2 \mathbb{E}(U_j^2)$$

Utilizando las hipótesis del modelo:

$$\sigma_j^2 = \sum_{k=1}^{m} \lambda_{kj}^2 + \psi_j^2$$

□

Teniendo en cuenta esta descomposición se puede definir el siguiente concepto.

Definición 2.2.3. Se llama *comunalidad* de la variable X_j, $h_j^2 = \sum_{k=1}^{m} \lambda_{kj}^2$ [23]

Esto permite interpretar la varianza de la variable como la varianza explicada por los factores comunes por un lado y la variabilidad específica de la propia variable.

Proposición 2.2.2. La covarianza de dos variables $Cov(X_j, X_j')$, $\sigma_{jj'}$ cumple que [18, 6]:

$$\sigma_{jj'} = \sigma_{j'j} = \sum_{k=1}^{m} \lambda_{kj}\lambda_{kj'} \tag{2.2.6}$$

Demostración.

$$\sigma_{jj'} = \sum_{k=1}^{m} \lambda_{kj}\lambda_{kj'}\mathbb{E}(F_k^2) + \sum_{k=1}^{m} \sum_{n=1, n\neq k}^{m} \lambda_{kj}\lambda_{nj'}\mathbb{E}(F_k \cdot F_n) \tag{2.2.7}$$
$$+ \sum_{k=1}^{m} \lambda_{kj}\psi_{j'}\mathbb{E}(F_k \cdot U_{j'}) + \sum_{k=1}^{m} \lambda_{kj'}\psi_j \mathbb{E}(F_k \cdot U_j) + \psi_j\psi_{j'}\mathbb{E}(U_j \cdot U_{j'}) \tag{2.2.8}$$

De nuevo, teniendo en cuenta las hipótesis

$$\sigma_{jj'} = \sigma_{j'j} = \sum_{k=1}^{m} \lambda_{kj}\lambda_{kj'} \tag{2.2.9}$$

□

Y tomando estas dos proposiciones se puede tener que la matriz $\mathbf{\Sigma}$, cumple lo siguiente:

Corolario 2.2.1. La matriz $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$

Bajo esta idea en la que la varianza se puede descomponer en una parte específica y una común a todas las variables.

2.2.2. Estimación de la matriz de cargas

Supóngase el vector aleatorio x como antes del cual se recogen N observaciones para obtener la matriz de datos $\mathbf{X}$ de tamaño $N \times p$que se supondrá centrada sin pérdida de generalidad. Esto provoca que la matriz $\mathbf{S} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}}$ [23].

Entonces el objetivo es estimar la matriz de cargas $\mathbf{\Lambda}$, para ello se puede tomar la matriz $\mathbf{S} - \hat{\mathbf{\Psi}}$, si de esta matriz se obtiene la descomposición espectral:

$$\mathbf{S} - \hat{\mathbf{\Psi}} = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{2.2.10}$$

Donde la matriz $\mathbf{D}$ es una matriz diagonal de tamaño $r \times r$ con los $r = rg(\mathbf{X})$ valores propios no nulos, y la matriz $\mathbf{U}$ de tamaño $p \times r$ de manera las columnas son los vectores propios asociados a los valores propios no nulos.

De esta manera, *Cuadras C.M;* [8] toma como estimador de $\hat{\mathbf{\Lambda}} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$, ya que $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T = \mathbf{S} - \hat{\mathbf{\Psi}}$. El problema que surge aquí es que la estimación $\hat{\mathbf{\Psi}}$ no es conocido, pero se desarrollará en lo siguiente.

De esta manera, supóngase que es conocido el número de factores, m. Por tanto, como el objetivo es que los factores comunes es maximizar su varianza común explicada, entonces siguiendo el mismo razonamiento que en las componentes principales, se puede tomar como estimación de las matrices factoriales la matriz siguiente:

$$\hat{\mathbf{\Lambda}} = (\sqrt{\mu_1}e_1, \ldots, \sqrt{\mu_m}e_m) = \mathbf{U}_m\mathbf{D}_m^{\frac{1}{2}} \tag{2.2.11}$$

Donde la pareja (μ_i, e_i) es el vector propio e_i en formato columna y μ_i su valor propio asociado. Además la matriz $\mathbf{U}_m$ es la matriz que tiene como columnas los m vectores propios de la matriz $\mathbf{S} - \mathbf{\Psi}$ que tienen asociados los m valores propios más grandes.

El problema de que no se conoce $\hat{\mathbf{\Psi}}$, se puede resolver planteando el siguiente método iterativo, para ello se introduce la notación:

- $\mathbf{S}_i^* = \mathbf{S} - \hat{\mathbf{\Psi}}_i$ es la matriz reducida en la i-ésima iteración del método
- $\hat{\mathbf{\Lambda}}_i$ es la matriz de cargas obtenida en la i-ésima iteración del método.

El método se inicializa con $\hat{\mathbf{\Psi}}_0 = 0$ de tal manera que $\mathbf{S}_0^* = \mathbf{S}$ y consiste en cada iteración hacer las siguientes operaciones [35, 23, 8]:

1. Se calcula $\mathbf{S}_i^* = \mathbf{S} - \hat{\mathbf{\Psi}}_i$

2. Se calcula la descomposición $\hat{\mathbf{\Lambda}}_i$ como se ha detallado antes obteniendo $\mathbf{U}_i$ y $\mathbf{D}_i$ de tamaños $p \times m$ y $m \times m$.

3. Se obtiene la nueva estimación de la $\hat{\mathbf{\Psi}}_{i+1} = \mathbf{S} - \hat{\mathbf{\Lambda}}_i \hat{\mathbf{\Lambda}}_i^T$

El proceso se para cuando no haya cambios en las matrices $\hat{\mathbf{\Psi}}_i, \hat{\mathbf{\Psi}}_{i+1}$ o cuando $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \mathbf{\Psi}$ sea lo más parecido posible a la matriz de covarianzas $\mathbf{S}$.

Peña D [23] lo generaliza utilizando la función $F = tr(\mathbf{S} - \mathbf{\Lambda}\mathbf{\Lambda}^T - \Psi)^2$. Que en esencia es minimizar la distancia de Frobenius, implicando un razonamiento desde la optimización, ya que se busca minimizar una función de pérdida, la distancia entre ambas matrices en este caso.

Una vez estimadas las cargas de los factores, se pueden estimar las *puntuaciones de los factores*, que son los valores que toman los factores pero no se detallarán en esta memoria. *(Para más detalles de esta parte véase la sección 9.5 de [35] o la 12.7 de [23])*

2.2.3. Unicidad del modelo

Un detalle en el que no se ha profundizado, es que el modelo factorial, no es único, es decir, teniendo en cuenta las suposiciones hechas anteriormente, todas las rotaciones de los factores $\mathbf{T}\mathbf{f}^T$, donde $\mathbf{T}$ es una matriz ortogonal de tamaño $m \times m$, entonces al transformar la matriz $\mathbf{\Lambda}$ a la matriz $\mathbf{\Lambda}\mathbf{T}^T$ se cumple que [48]:

$$\mathbf{x}^T = (\mathbf{\Lambda}\mathbf{T}^T)(\mathbf{T}\mathbf{f}^T) + \Psi\mathbf{u}^T \tag{2.2.12}$$

Es equivalente al modelo 2.2.1, ya que se cumplen las mismas suposiciones.

Para evitar estas indeterminaciones se suelen tomar dos restricciones [48]:

$$\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda} \text{ ó } \mathbf{\Lambda}^T\mathbf{\Lambda} \text{ son diagonales} \tag{2.2.13}$$

En particular, la segunda restricción es la que se toma en el método del factor principal, ya que $\hat{\mathbf{\Lambda}} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$, donde $\mathbf{U}$, es una matriz ortogonal, por tanto, $\hat{\mathbf{\Lambda}}^{\mathbf{T}}\hat{\mathbf{\Lambda}} = \mathbf{D}^{\frac{1}{2}}\mathbf{U}^T\mathbf{U}\mathbf{D}^{\frac{1}{2}} = \mathbf{D}$.

Proposición 2.2.3. La segunda restricción hace que el modelo sea único [23].

Demostración. Sea la descomposición de la varianza *(Corolario 2.2.1)*:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$$
$$\mathbf{\Sigma} - \mathbf{\Psi} = \mathbf{\Lambda}\mathbf{\Lambda}^T$$

Si postmultiplicamos por $\mathbf{\Lambda}$ a ambos lados:

$$(\mathbf{\Sigma} - \mathbf{\Psi})\mathbf{\Lambda} = \mathbf{\Lambda}\mathbf{\Lambda}^T\mathbf{\Lambda}$$

Como se supone que $\mathbf{\Lambda}^T\mathbf{\Lambda} = \mathbf{D}$ se tiene que:

$$(\mathbf{\Sigma} - \mathbf{\Psi})\mathbf{\Lambda} = \mathbf{\Lambda}\mathbf{D}$$

Por tanto, la matriz $\mathbf{\Lambda}$ tiene por columnas los m vectores propios de la matriz $\mathbf{\Sigma} - \mathbf{\Psi}$, como se ha señalado anteriormente, esta propiedad ha sido usada anteriormente en el método del factor principal, es decir, la estimación por este método es única. □

2.2.4. Interpretación del modelo factorial

Para interpretar el modelo factorial hay que tener en cuenta los siguientes aspectos:

- *Proporción de varianza común explicada* Es decir, analizar que parte de la varianza se puede explicar por los factores comunes. A mayor sea, más relación entre las variables explican los factores,
- *Especificidad de cada una de las variables*: De esta manera, se puede analizar si una variable está alejada de las demás y no tiene tanto que ver con el resto.
- *Cargas de cada uno de los factores en cada variable* Esto también nos puede dar datos, por ejemplo si un grupo de variables tiene cargas mayores en uno de los factores F_i y en otro grupo es más prevalente otro factor, eso también es significativo.

Johnson y Wichern analizan en el capitulo 9 de [35] un ejemplo con un conjunto de datos de varias medidas de aves.

Como medidas observables tienen las medidas de 6 huesos y aplican el modelo factorial con 3 factores. Si nos centramos en los datos obtenidos aplicando el método del factor principal [23, 35], lo primero a destacar es que 3 factores consiguen acumular un 95 % de la varianza, por lo tanto, es posible afirmar que hay 3 variables latentes, a las que luego daremos interpretación. Por otro lado, la varianza específica es poca o nula en todas las variables, por lo tanto, nos reafirma en nuestra sospecha de que hay 3 variables latentes.

Por último, podemos ver que las cargas que tiene en todas el primer factor son cercanas a 1 por tanto tiene gran importancia en todas. Sin embargo el 3er factor en las ultimas 4 variables no tiene esa misma importancia. Por tanto el 3er factor casi no influye en las últimas medidas.

Johnson, R.A. y Wichern, D.W.[35] además utilizan el método de máxima verosimilitud y utilizan rotaciones para mejorar la interpretación.

2.3. Análisis de Clusters

Según *Jain,A.K. y Richard, C.D;* [13] el análisis de clusters tiene como objetivo conseguir una clasificación de las observaciones en subconjuntos que tengan sentido de acuerdo al contexto de la investigación. El clustering es un tipo de clasificación de los datos que tiene las siguientes características:

- *Exclusividad:* Una observación no puede pertenecer a más de un cluster a la vez.
- *Es intrínseco:* No hay ninguna etiqueta que permita clasificar las observaciones, son las características de las propias muestras las que servirán como diferenciadores. Lo que hace que sea un método *no supervisado*.

Una vez clarificado las características esenciales del clustering, este se puede dar de manera *jerárquica* o *particional*.

Definición 2.3.1. Se dice que un método de clustering es *jerárquico* si genera una secuencia de particiones del espacio de observaciones las cuales están anidadas

Definición 2.3.2. Un método de clustering *particional* genera una única partición del espacio.

Por otro lado, los algoritmos se pueden clasificar según sean *aglomerativos*, en los que se empieza teniendo cada una de las observaciones y se van formando los clusters hasta tener un solo cluster con todas las muestras. Por el contrario, los *algoritmos divisivos*, se empieza con un solo cluster y se van haciendo subconjuntos del mismo.

Otra clasificación puede venir dada por el modo en el que se consideran las variables, si es *Monotético*, en cada paso se centra en una variable, mientras que si es *Politético* se utilizan todas las variables a la vez.

También hay que distinguir métodos que usan como fundamento el álgebra matricial, centrándose en los datos que nos puede dar la matriz de distancias entre observaciones o en la Teoría de Grafos.

En primer lugar, es necesario definir los conceptos principales de esta sección que es un clustering y un cluster:

Definición 2.3.3. Se llama *clustering* a la partición que provoca la relación $\mathcal{R}$ y se definen los *clusters* como las clases de equivalencia de dicha relación, $\{c_i\}$.

Por tanto, el análisis de clusters es equivalente a establecer una relación de equivalencia entre las observaciones.

2.3.1. Algoritmos Jerárquicos

Los algoritmos jerárquicos son aquellos que utilizan como entrada no la matriz de los datos per sé, en su lugar, estas utilizan como entrada la matriz de distancias o similaridad.

Uno de los inconvenientes mayores de este tipo de algoritmos es como calcular la matriz de similaridades o distancias. En el caso de que las variables sean numéricas y continuas entonces se puede utilizar la distancia euclídea entre elementos de $\mathbb{R}^p$. Además hay que tener en cuenta si se va a estandarizar ya que en el caso de que si se estandarice se dará el mismo peso a todas las variables, dando igual su escala, en cambio si no se estandariza, la escala provoca que las características con mayores valores tengan mayor peso en las distancias. Por otro lado, si se tienen variables binarias o categóricas hay que tener en cuenta cómo medir sus distancias, todo esto se soluciona con las siguiente definición y consideraciones:

Definición 2.3.4. Llamaremos similaridad entre dos muestras i, h según la variable j a una función s_{jih} la cual cumpla que:

- La similaridad de una muestra consigo misma es igual a la unidad $s_{jii} = 1 \forall j, i$
- $0 \leq s_{jih} \leq 1 \quad \forall j, i, h$
- Simetría: $s_{jih} = s_{jhi}$

Una vez considerado esto, se puede crear un coeficiente de similaridad entre dos mediante el coeficiente de *Gower*:

$$s_{ih} = \frac{\sum_{j=1}^{p} w_{jih} s_{jih}}{\sum_{j=1}^{p} w_{jih}} \tag{2.3.1}$$

Donde la variable w_{jih} puede ser 0 ó 1, dependiendo de si se quiere tomar o no dicha variable o no en la comparación de dichas muestras.

Lo más habitual para formalizar las similaridades es tomar lo siguiente:

- Para continuas se puede tomar el valor $s_{jih} = 1 - \frac{|x_{ji} - x_{hi}|}{rango(X_j)}$
- En cambio para variables binarias en el caso de que coincida el atributo $x_{ji} = x_{jh}$ la similaridad es 1 y 0 en el caso contrario. También se puede decir que si hay presencia del atributo y coinciden es 1 y en caso contrario 0.

Una vez decidida la forma de decir de inicialmente dar la matriz de distancias o similaridades, se da el algoritmo aglomerativo de generación de la jerarquía:

- Se comienza con un *clustering* que tiene n *clusters* con una observación cada uno y se calcula la matriz de distancias
- Se toman los elementos que más cercanos están y se forman una nueva clase.
- Se sustituyen los dos elementos anteriores por la clase y se calcula la distancia entre la clase y el resto de observaciones de acuerdo con uno de los criterios preestablecidos.
- Repetir los dos pasos anteriores hasta obtener una única clase.

Dependiendo de la forma que se de unir los grupos se obtendrá un algoritmo u otro de clustering:

- Si $f(x, y) = min(x, y)$ es decir, se toma como distancia entre la clase y el resto de observaciones como el mínimo de las que se den, a este algoritmo se le llama *single linkage*, este tipo de enlace tiende a formar.
- Si $f(x, y) = max(x, y)$ es análogo al anterior, pero tomando la distancia máxima, a este método se le llama *complete linkage*.
- Si $f(x, y) = \frac{x + y}{2}$ se hace la media de las distancias de los dos grupos antes de unirlos, a este método se le llama *average linkage*.

Este tipo de métodos aporta una jerarquía en la agrupación de los datos en función de su similitud, de manera que a distintos grados de similitud se pueden establecer distintas reparticiones del espacio de observaciones. Además esto se puede representar fácilmente con un dendograma o diagrama de árboles otorgando una manera sencilla de establecer un criterio de similitud máxima.

2.3.2. Algoritmos Particionales

El clustering particional busca dado un conjunto de datos heterogéneos dividirlos en grupos homogéneos, como por ejemplo la segmentación de clientes o votantes. , buscaremos un criterio por el cual podamos afirmar que una partición del espacio es mejor que la anterior o no. En estos casos la variabilidad total de todas las observaciones se puede descomponer como la variabilidad intra-cluster y la inter-cluster como se detalla en la sección del análisis discriminante.

Definición 2.3.5. Se llama *centroide* de un cluster, C_k, $k = 1 \dots K$ al elemento que resulta de hacer la media de todos los elementos del cluster C_k:

$$\overline{x}_{jk} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij} \tag{2.3.2}$$

Definición 2.3.6. Se llama variabilidad o variación intra-cluster del cluster C_k $k = 1 \dots K$ a la siguiente expresión:

$$\mathbf{W}(C_k) = \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \overline{x}_{jk})^2 \tag{2.3.3}$$

Esta medida también se podría dar como la media de las diferencias cuadradas entre las observaciones del cluster, pero se seleccionará esta expresión.

Una vez dadas estas definiciones y criterios se puede construir un algoritmo iterativo como es el método de K-means que se puede resumir de la siguiente manera, sabiendo el número inicial de clusters, por ejemplo K:

1. Se inicializa la asignación de manera aleatoria, agrupando en los K clusters observaciones aleatoriamente.
2. Se itera lo siguiente hasta que no haya cambio en los clusters:
   a) Se calculan los centroides de cada uno de los K clusters

b) Se reasignan las observaciones de acuerdo a qué centroide es más cercano según la distancia euclidiana.

Proposición 2.3.1. El algoritmo de K-means sólo puede mejorar el criterio de la variación dentro de clusters.

Demostración. El paso 2a minimiza la suma de las desviaciones cuadradas y en el paso 2b la reasignación solo puede mejorar eso ya que se busca que la distancia euclídea al *centroide* sea la mínima. □

Esto provoca que el algoritmo no para hasta que no se da una mejora en el criterio. Sin embargo, este minimo que alcanza la función no tiene por qué ser global, ya que esta no es convexa y por tanto la consecución de este mínimo puede estar condicionada por la asignación inicial. Para solucionar estos problemas, se deben hacer varias ejecuciones del algoritmo para poder compararlas entre sí, por ejemplo *Scikit-Learn* implementa el método *k-means++* que no es más que realizar multiples ejecuciones con asignaciones iniciales distintas y luego las compara y otorga la mejor. Para poder compararlas se puede utilizar el parámetro inercia.

Definición 2.3.7. Se llama *inercia* de un clustering a la suma de las variaciones intra-clusters:

$$inercia = \sum_{k=1}^{K} \mathbf{W}(C_k) \tag{2.3.4}$$

La inercia representa como de compacta es nuestra partición a menor inercia más compacta es y por tanto más homogéneos son.

Otro de los problemas de este tipo de algoritmos es que se deben conocer a priori el número de clusters en los que se quieren dividir las observaciones. Para ello *Peña D.*[23] propone un contraste de hipótesis utilizando la inercia haciéndola depender del número de clusters

$$F = \frac{inercia(K) - inercia(K+1)}{\frac{inercia(K+1)}{n-K-1}} \tag{2.3.5}$$

Este cociente compara la disminución de la variación entre una partición con K y $K+1$ clústers, de esta manera, puede compararse con una F con $p, p(n-K-1)$ grados de libertad, pero que en la práctica no se puede asumir las hipótesis, por tanto, en la práctica si se da un valor mayor que 10 se asume que es conveniente añadir un clúster más.

Capítulo 3

Aplicación sobre datos

A lo largo de este capítulo se aplicarán algunas de las técnicas desarrolladas anteriormente a ejemplos prácticos con bases de datos reales obtenidas de distintos repositorios.

La principal herramienta usada ha sido el lenguaje de programación Python, utilizando librerías como Pandas para el manejo de las bases de datos, Scikit-Learn para la implementación de los modelos y otras librerias de visualización como Matplotlib para la inclusión de gráficos sencillos.

3.1. Tipo de alubias

Descripción del DataSet

Este conjunto de datos, [28] está formado por 12 variables predictoras que detallan *Koklu, M. y Ozkan, I.A.*[56]

Objetivos

Los objetivos son estudiar la estructura de la matriz de covarianzas, es decir se quiere estudiar cuales son las direcciones de mayor variabilidad y si las variables observadas guardan un factor latente. Es decir, se está utilizando una óptica exploratoria, ya que no queremos confirmar ninguna suposición previa.

Además, se busca definir un modelo predictor para poder crear una clasificador de las propias alubias.

Métodos a utilizar

Para estudiar la estructura de los datos, se utilizará el análisis de componentes principales. Por otro lado, para comprobar la existencia de factores latentes se aplicará el análisis de factores principales. Tras aplicar ambas técnicas,

Desarrollo y resultados

Conclusiones

Bibliografía

[1] **Abdi, H., and Williams, L. J. (2010).** Principal component analysis. Wiley *Interdisciplinary Reviews: Computational Statistics* 2(4), 433–459.

[2] **Abdi, H. (2007).** The method of least squares. *Encyclopedia of measurement and statistics*, 1, 530-532.

[3] **Biau, G., and Scornet, E. (2016)**. *A random forest guided tour*. Test, 25, 197-227.

[4] **Breiman, L. (2004).** Consistency for a simple model of random forests. *University of California at Berkeley. Technical Report,* 670.

[5] **Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., and Brown, S. D. (2004).** An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6), 275-285.

[6] **Chatfield,C and Collins A.J (1989)**. *Introduction to multivariate analysis*, Chapman and Hall.

[7] **Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006).** Gene selection and classification of microarray data using random forest. *BMC bioinformatics,* 7, 1-13.

[8] **Cuadras, C.M. (2014)**, *Nuevos métodos de Análisis Multivariante*, CMC Editions, Barcelona.

[9] **Hastie, T., Tibshirani, R. and Friedman J. (2001)**, *The Elements of Statistical Learning, Data Mining, Inference and Prediction* Springer

[10] **Hair, J. F., Black, W. C., Tatham, R. L., and Anderson, R. E. (1995).** *Multivariate data analysis with readings.* Prentice-Hall International, Englewood Cliffs, New Yersey.

[11] **Jollife I.T.(1986)**. *Principal Component Analysis*, Springer-Verlag.

[12] **Köppen, M. (2000)** The curse of dimensionality. *5th online world conference on soft computing in industrial applications (WSC5)* (Vol. 1, pp. 4-8).

[13] **Jain, A.K., Richard, C.D., (1988)** *Algorithms for clustering data.* Prentice-Hall, Inc.

[14] *Scikit Learn Documentation, Clustering Methods.* `https://scikit-learn.org/stable/modules/clustering.html#clustering`

[15] Neural Designer Blog https://www.neuraldesigner.com/

[16] **Lopez, R., Balsa-Canto, E. and Oñate, E. (2008)** Neural networks for variational problems in engineering *Int. J. Numer. Meth. Engng.*, 75 https://doi.org/10.1002/nme.2304 1341-1360.

[17] **Lebart, L., Morineau, A., Warwick, K. M. (1984)**.*Multivariate Descriptive Analysis: Correspondence analysis and related techniques for large matrices*. Wiley.

[18] **Morrison. D.(1976).** *Multivariate statistical methods (2nd ed)*. McGraw-Hill Kogakusha.

[19] **Batista Foguet, J.M. y Martínez Arias, R. (1989)***Análisis Multivariante: Análisis en Componentes Principales.* Editorial Hispano Europea.

[20] **Divakaran, S. (2022).** Data Science: Principles and Concepts in Modeling Decision Trees. *Data Science in Agriculture and Natural Resource Management*, 55-74.

[21] **James G, Witten D, Hastie T, Tibshirani R.** *An Introduction to Statistical Learning: With Applications in R.* New York, NY: Springer; 2021.

[22] **Martínez Arias R.** *El análisis multivariante en la investigación científica.* Madrid: La Muralla; 1999.

[23] **Peña D.** *Análisis de datos multivariantes.* Madrid: McGraw-Hill; 2002.

[24] **Andrade-Sánchez, A. I., Galindo-Villardón, M. P., y Cuevas Romo, J. (2015).** Análisis multivariante del perfil psicológico de los deportistas universitarios: Aplicación del CPRD en México. *Educación Física y Ciencia*, 17(2), 00-00.

[25] **Hernandis, S. P., Diez, J. P., y Rouma, A. C. (2002).** El consumo de inhalables y cannabis en la preadolescencia: Análisis multivariado de factores predisponentes. *Anales de Psicología/Annals of Psychology*, 18(1), 77-93.

[26] **Machado-Duque, M. E., Echeverri Chabur, J. E., y Machado-Alba, J. E. (2015).** Somnolencia diurna excesiva, mala calidad del sueño y bajo rendimiento académico en estudiantes de Medicina. *Revista colombiana de Psiquiatría,* 44(3), 137-142.

[27] **Hornik, K., Stinchcombe, M., and White, H.** Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359-366, 1989. https://doi.org/10.1016/0893-6080(89)90020-8

[28] **Dry Bean Dataset. (2020)**. *UCI Machine Learning Repository.* https://doi.org/10.24432/C50S4B.

[29] **Pagès, J. (2005).** Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food quality and preference*, 16(7), 642-649. https://doi.org/10.1016/j.foodqual.2005.01.006

[30] **Okazaki, S. (2006).** What do we know about mobile Internet adopters? A cluster analysis. *Information & Management,* 43(2), 127-141.

[31] **Mamidi, R. R. (2021)**. Application of Neural Networks for Prediction of Double Fed Induction Generator's Equivalent Circuit Parameters used in Wind Generators. *IRJCS:: International Research Journal of Computer Science*, Volume VIII, 23-31.

[32] **Nerini, D., & Ghattas, B. (2007).** Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis* 51(10), 4984-4993.

[33] **Ensum, J., Pollard, R., & Taylor, S. (2005).** Applications of logistic regression to shots at goal at association football. *In Science and football V: the proceedings of the Fifth World Congress on Science and Football* (pp. 214-221).

[34] **Greene, W.H.** (2008). *Econometric analysis (6th ed.)* Pearson Prentice Hall

[35] **Johnson, R. A., & Wichern, D. W. (2007)**. *Applied multivariate statistical analysis.(6th edition)* Pearson Prentice Hall.

[36] **Everitt, B., and Hothorn, T. (2011)**. *An introduction to applied multivariate analysis with R.* Springer Science & Business Media.

[37] **Ringnér, M. (2008).**What is principal component analysis?.*Nature biotechnology, 26(3), 303-304.*

[38] **Gottumukkal, R., & Asari, V. K. (2004).** An improved face recognition technique based on modular PCA approach. *Pattern Recognition Letters,* 25(4), 429-436.

[39] **Eckart, C., & Young, G. (1936).** The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211-218.

[40] **Johnson, R. M. (1963).** On a theorem stated by Eckart and Young. *Psychometrika*, 28(3), 259-263.

[41] **Golub, G. H., Hoffman, A., & Stewart, G. W. (1987)**. A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and its applications*, 88, 317-327.

[42] **Mahesh, B. (2020).** Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, 9, 381-386.

[43] **Pearson, K. (1901)**. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science,* 2(11), 559-572.

[44] **Hotelling H. (1933)** Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 25:417–441.

[45] **Vincent, D. F. (1953).** The origin and development of factor analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 2(2), 107-117.

[46] **Galton, F. (1889).** Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London,* 45(273-279), 135-145.

[47] **Spearman, C. (1904).** "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology,* 15(2), 201–292. `https://doi.org/10.2307/1412107`

[48] **Mardia, K.V., Kent, J.T. and Bibby J.M.; (1979)**. *Multivariate analysis.* Academic Press.

[49] **Song, Y. Y., and Ying, L. U. (2015).** Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry,* 27(2), 130.

[50] **Lawless, J. F., & Yuan, Y. (2010).** Estimation of prediction error for survival models. *Statistics in medicine,* 29(2), 262-274.

[51] **Loh, W. Y. (2011).** Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1), 14-23.

[52] **Breiman, L. (1984).** *Classification and regression trees.* Chapman & Hall.

[53] **Hesterberg, T. (2011).** Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics,* 3(6), 497-526.

[54] **Breiman, L. (1996).** Bagging predictors. *Machine learning*, 24, 123-140.

[55] **Breiman, L. (2001).** Random forests. *Machine learning*, 45, 5-32.

[56] **Koklu, M., & Ozkan, I. A. (2020).** Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture,* 174, 105507.

[57] **Livingstone, D. J., Manallack, D. T., & Tetko, I. V. (1997).** Data modelling with neural networks: Advantages and limitations. *Journal of computer-aided molecular design*, 11, 135-142.

[58] **Hochreiter, S., & Schmidhuber, J. (1997).** Long short-term memory. *Neural computation*, 9(8), 1735-1780.

[59] **Quinlan, J. R. (2014)**. *C4. 5: programs for machine learning.* Elsevier.

[60] **Loh, W. Y. (2014).** Fifty years of classification and regression trees. *International Statistical Review*, 82(3), 329-348.

[61] **Kass, G. V. (1980)**. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2), 119-127.

[62] **Grossi, E., & Buscema, M. (2007).** Introduction to artificial neural networks. *European journal of gastroenterology & hepatology,* 19(12), 1046-1054.

[63] **Villardón, J. L. V. (2006).** *Análisis discriminante: introducción.* Universidad de Salamanca: Departamento de estadistica.