

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

1	Métodos Supervisados	3
1.1	Métodos Lineales para regresión	4
1.1.1	Métodos de ajuste	5
1.1.2	Inferencias estadísticas sobre $\hat{\beta}$	7
1.1.3	Regresión multivariante	11
1.2	Clasificación	14
1.2.1	Funciones discriminantes	15
1.2.2	Análisis canónico de poblaciones	16
1.3	Redes Neuronales	20
1.4	Árboles de decisión y bosques aleatorios	23
1.4.1	Conceptos básicos	23
1.4.2	Árboles de regresión	25
1.4.3	Árboles de clasificación	27
1.4.4	Sesgo y varianza de un modelo	29
1.4.5	Algoritmo Random Forest	31
2	Métodos no supervisados	33
2.1	Análisis de componentes principales	34
2.1.1	Definición y cálculo de las componentes principales	35
2.1.2	Componentes principales por la descomposición en valores singulares	38
2.2	Análisis Factorial	42
2.2.1	Formalización	43
2.2.2	Obtención de la matriz de saturaciones	45
2.2.3	Unicidad del modelo	46
2.2.4	Interpretación del modelo factorial	47
2.3	Análisis de cluster	48
2.3.1	Algoritmos jerárquicos	49
2.3.2	Algoritmos particionales	50

<i>ÍNDICE GENERAL</i>	III
3 Aplicación sobre datos	52
3.1 Cemento, análisis estructural y capacidades predictivas.	52
4 Conclusiones	56
Bibliografía	57

Introducción

El análisis multivariante es, según Martínez, 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 [52]. 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 exploran 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 [36]:
 - Regresión: Cuando las variables de respuesta son continuas. Dentro de esta tipología, entran métodos como las redes neuronales [50] o los árboles de regresión [54].
 - Clasificación: Cuando las variables de respuesta son cualitativas o categóricas, como por ejemplo, el análisis discriminante [13] o la regresión logística [18].
- Métodos no supervisados: Los métodos no supervisados son aquellos que buscan analizar la estructura y las relaciones entre las distintas variables [30]. Las distintas técnicas se distinguen según la medida en la que se centren, por ejemplo, la variabilidad común [57], la homogeneidad de grupos [56] 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 del siglo XXI [12, 3, 13, 16]. 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, implementados para que el usuario sólo tenga que usar una línea de código.

Muchas veces, el uso de estas técnicas se combinan con el aprendizaje automático como proponen Hastie, Tibshirani y Friedman [30] o James, Witten, Hastie, y Tibshirani [36]. De esta manera, los parámetros se estiman extrayendo la información de los propios datos, comparando distintos modelos, entre otros. 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.

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

Teniendo en cuenta el objetivo, 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, la clasificación, 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 *cluster*. Por último, se realizará una aplicación en la que se desarrollarán las interpretaciones de las técnicas descritas anteriormente.

Capítulo 1

Métodos Supervisados

En esta parte del trabajo se realizará una descripción de los métodos supervisados [49].

Definición 1.0.1. Se llaman métodos supervisados a aquellos métodos que buscan inferir una relación estocástica entre dos grupos de variables, predictoras y respuestas en un conjunto de datos que contiene observaciones simultáneas de ambos conjuntos de datos.

Sea un conjunto de variables aleatorias observables de manera simultánea en una población. Estas variables se pueden dividir en dos tipos [30]:

- 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 $\mathbf{x} = [X_1, \dots, X_p]$.
- Variables de salida o respuesta: 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 .

Estas $K + p$ variables son observables en una población, pero no siempre es posible estudiar los valores de estas variables en el conjunto total de la población. Es por ello, que normalmente se utiliza una muestra.

Definición 1.0.2. Se llama muestra a un subconjunto de la población. Para que nuestro estudio sea correcto, esta debe ser representativa.

De manera habitual se trabajará con muestras aleatorias. De esta forma se podrán hacer inferencias sobre las estructuras de medidas de la población.

Definición 1.0.3. 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.

Sea una muestra representativa con N observaciones de las $K + p$ variables. Entonces, se pueden definir los siguientes conceptos [10]:

Definición 1.0.4. Se define matriz de datos \mathbf{X} como la matriz de tamaño $N \times p$ que contiene como filas los vectores de longitud p que representan los datos de cada observación. Estas se denotarán a lo largo de la memoria como \mathbf{x}_i , donde $i = 1, \dots, N$. Por ejemplo, en el caso de que se recojan datos sobre distintos modelos de coches las observaciones serían los valores medidos de las distintas variables consideradas en cada uno de los coches.

Definición 1.0.5. La matriz de respuestas \mathbf{Y} es una matriz de tamaño $N \times K$, donde cada observación de las variables respuesta es una fila. En este trabajo cada observación del vector respuesta se denota como \mathbf{y}_i donde $i = 1, \dots, N$, 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 categórica 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 .

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 lineal es estudiar como un conjunto de variables respuesta continuas están relacionadas con otro conjunto de variables predictoras. En el caso de la regresión lineal, se estudia como una combinación lineal de las variables predictoras puede relacionarse con las variables respuesta. 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 [37].

De manera concreta, sea $\mathbf{x} = [X_1, \dots, X_p]$ el vector de variables predictoras. Además supóngase hasta que se diga lo contrario que únicamente hay una variable respuesta Y continua.

Teniendo en cuenta lo anterior, se define el modelo de regresión lineal como la siguiente relación estocástica [30, 37]:

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

donde se supone que ε es una variable aleatoria con $\mathbb{E}(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2$. Habitualmente se asume a mayores que $\varepsilon \sim N(0, \sigma^2)$. Además, supóngase que el vector de variables de entrada sigue una distribución normal multivariante $N_p(\mu, \Sigma)$ donde Σ es una matriz semidefinida positiva y simétrica [10].

En el caso de la regresión lineal, 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:

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

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 :

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

De esta manera, tomando una matriz de datos de N realizaciones, entonces se puede generar un vector de predicciones de tamaño N de la siguiente manera:

$$\hat{\mathbf{y}} = \mathbf{X}\beta \quad (1.1.4)$$

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

Una vez establecida la notación y los supuestos que se tomarán de ahora en adelante, hay que estimar los parámetros de regresión, β . Para ello, se utilizan distintos métodos como el método de los mínimos cuadrados o el de máxima verosimilitud. A parte de esto, se detallarán las características inferenciales de los estimadores obtenidos y su interpretación geométrica.

1.1.1.1. 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 con anterioridad. Para poder hallar una estimación del mismo se debe definir el concepto de función de pérdida [30].

Definición 1.1.2. Se llama función de pérdida $L(Y, \hat{Y})$ a aquella que mide las diferencia entre el valor real de una variable y su valor predicho por un cierto modelo.

Aunque el término sea más utilizado en el marco del aprendizaje automático [36], el concepto de error cuadrático o la función de pérdida cuadrática es utilizado en el ajuste de los parámetros de regresión, ya que esta pérdida, al minimizarse, es equivalente al método de los mínimos cuadrados. En particular, el error cuadrático se define de la siguiente manera [1]:

Definición 1.1.3. Se llama error de predicción cuadrático entre una variable observable 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:

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

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 ésta es una forma cuadrática general respecto de β :

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

La segunda derivada respecto del vector de parámetros es $2\mathbf{X}^T\mathbf{X}$, entonces es semi-definida positiva, ya que todos los valores propios son positivos o nulos. Entonces los β en los que $-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$ son mínimos (para el desarrollo de los anteriores cálculos véase [53]).

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

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

Esa solución es :

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

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

Con la siguiente proposición se puede ver la equivalencia entre el método de los mínimos cuadrados y de máxima verosimilitud [30]:

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.

Demostración. Sea un conjunto de N observaciones independientes y_i , donde $i = 1, \dots, N$ de una variable aleatoria Y . Sea su función de probabilidad $\mathbb{P}_\theta(y)$ que depende de unos ciertos parámetros θ que no tienen por qué ser los de regresión. Entonces el método de máxima verosimilitud busca maximizar la siguiente función:

$$L(\theta) = \sum_{i=1}^N \log(\mathbb{P}_\theta(y_i)), \quad (1.1.9)$$

suponiendo que la variable respuesta cumple como antes que $Y = f_\theta(\mathbf{x}) + \varepsilon$, en el caso de la regresión lineal tendríamos que $f_\theta(\mathbf{x}) = \mathbf{x}\theta$, es decir, θ sería el vector

de parámetros de regresión y donde $\varepsilon \sim N(0, \sigma^2)$. En consecuencia, si se suponen conocidos a priori el vector de parámetros θ y el vector aleatorio \mathbf{x} entonces :

$$Y \sim N(f_\theta(\mathbf{x}), \sigma^2) \quad (1.1.10)$$

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

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

En particular la derivada de la función de la verosimilitud logística queda:

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{1}{2\sigma^2} \frac{\partial RSS(\theta)}{\partial \theta} \quad (1.1.12)$$

Por tanto, es equivalente minimizar el error cuadrático y maximizar la verosimilitud respecto al vector de parámetros. En consecuencia, ambos métodos de estimación obtendrán los mismos resultados. \square

Continuando con las estimaciones calculadas por el método de los cuadrados, los valores predichos obtenidos de las observaciones recogidas en la matriz de datos $\hat{\mathbf{y}}$ se calculan de la siguiente manera:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (1.1.13)$$

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

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 .

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 estadígrafos para realizar los contrastes de hipótesis necesarios.

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.

Hay que tener en cuenta los siguientes supuestos el vector $\mathbf{x} \sim N_p(\mu, \Sigma)$, 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 [11]

Proposición 1.1.2. 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 $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_N)$, donde \mathbf{I}_N es la matriz identidad de tamaño $N \times N$.

Demostración. Basta con comprobar que:

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}(\mathbf{X}\beta) + \mathbb{E}(\varepsilon) = \mathbb{E}(\mathbf{X}\beta) = \mathbf{X}\beta \quad (1.1.14)$$

Además :

$$\mathbb{E}((\mathbf{X}\beta + \varepsilon)(\mathbf{X}\beta + \varepsilon)^T) = \mathbf{X}\beta\beta^T\mathbf{X}^T + \text{Var}(\varepsilon) \quad (1.1.15)$$

Entonces, la varianza

$$\text{Var}(\mathbf{y}) = \mathbf{X}\beta\beta^T\mathbf{X}^T + \text{Var}(\varepsilon) - \mathbf{X}\beta\beta^T\mathbf{X}^T = \text{Var}(\varepsilon) = \sigma^2 \mathbf{I}_N \quad (1.1.16)$$

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

Añadamos a las suposiciones que la matriz \mathbf{X} es de rango máximo y por tanto, $\mathbf{X}^T\mathbf{X}$ es semidefinida positiva. A continuación se detallarán las cualidades inferenciales del vector $\hat{\beta}$, como que es insesgado [26], su varianza [26, 30, 37] y su distribución [30].

Proposición 1.1.3. El estimador $\hat{\beta}$ es un estimador insesgado.

Demostración.

Teniendo en cuenta las hipótesis anteriores se obtiene que:

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{y}) \quad (1.1.17)$$

En particular, se sabe que $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$. Entonces se obtiene lo siguiente:

$$\mathbb{E}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta \quad (1.1.18)$$

\square

Proposición 1.1.4. La varianza del estimador:

$$\text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \quad (1.1.19)$$

Demostración.

$$\begin{aligned}
 \text{Var}(\hat{\beta}) &= \mathbb{E}(\hat{\beta}^T \hat{\beta}) \\
 &= \mathbb{E}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}^T \mathbf{y} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} | \mathbf{X}) \\
 &= \mathbb{E}(\mathbf{y}^T \mathbf{y}) (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{X})^{-1} \\
 &= \text{Var}(\mathbf{y}) (\mathbf{X}^T \mathbf{X})^{-1} \\
 &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}
 \end{aligned} \tag{1.1.20}$$

Por tanto, el vector de estimaciones de los parámetros, denotado como $\hat{\beta}$, se calcula como una combinación lineal del vector de errores ε . Cada componente del vector $\hat{\beta}$, denotado como $\hat{\beta}_j$, $j = 1, \dots, p$, sigue una distribución normal. Específicamente, $\hat{\beta}_j$ sigue una distribución normal con media β_j y varianza $\sigma^2 (\mathbf{X}^T \mathbf{X}_{jj}^{-1})$. Aquí, β_j representa el valor verdadero del parámetro, σ^2 es la varianza y $(\mathbf{X}^T \mathbf{X}_{jj}^{-1})$ es el elemento de la diagonal correspondiente a la posición j en la matriz $(\mathbf{X}^T \mathbf{X})^{-1}$. \square

Si el parámetro σ^2 fuera conocido, entonces se puede construir un estadígrafo de contraste para cada parámetro de la siguiente manera:

$$z_j = \frac{\hat{\beta}_j}{\sigma \sqrt{(\mathbf{X}^T \mathbf{X}_{jj}^{-1})}} \sim N(0, 1) \quad \forall j = 1, \dots, p. \tag{1.1.21}$$

donde $(\mathbf{X}^T \mathbf{X}_{jj}^{-1})$ se define como antes.

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. y Cuadras proponen el siguiente estimador de la varianza, suponiendo que únicamente hay una variable respuesta [11, 30]:

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \tag{1.1.22}$$

Pero antes de detallar las propiedades del estimador definido hay que explicar el siguiente concepto geométrico [11].

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} .

En el caso de la regresión lineal que estamos describiendo se cumple una importante propiedad.

Proposición 1.1.5. El vector $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$ es ortogonal al subespacio $C_r(\mathbf{X})$.

Una vez definido lo anterior, se pueden establecer las propiedades del estimador descrito:

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

1. $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$ (Es un estimador insesgado de la varianza)

$$2. \hat{\sigma}^2 \sim \frac{\sigma^2}{N-p-1} \chi_{N-p-1}^2$$

Demostración.

1. El término $RSS(\hat{\beta} = \sum_{i=1}^N (y_i - \hat{y}_i)^2)$, 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 $\mathbf{T}\hat{\varepsilon} = (\overbrace{0 \dots 0}^{p+1}, \dots, \mathbf{e}'_N)^T$ por la proposición 1.1.5, y se cumple que $\mathbb{E}(\hat{\varepsilon}) = \mathbb{E}(\mathbf{T}\hat{\varepsilon}) = 0$, por la distribución que tiene el vector ε además $\mathbf{E}((\mathbf{e}'_i)^2) = \sigma^2, i = p+2 \dots N$, es decir, cada $\mathbf{e}_i \sim N(0, \sigma^2)$ ya que para los anteriores es 0 y por último:

$$\mathbb{E}(\hat{\varepsilon}^T \hat{\varepsilon}) = \mathbb{E}(\hat{\varepsilon}^T \mathbf{T}^T \mathbf{T} \hat{\varepsilon}) = \sum_{p+2}^N (\mathbf{e}'_i)^2 = (N-p-1)\sigma^2 \quad (1.1.23)$$

Por tanto:

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{N-p-1} \mathbb{E}(RSS(\hat{\beta})) = \sigma^2 \quad (1.1.24)$$

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

2. Y para terminar, se puede ver en la ecuación 1.1.23 que el estimador es suma de $N-p-1$ variables que siguen una distribución normales estándar multiplicadas por σ^2 y divididas entre $N-p-1$, por tanto, tenemos que además.

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{N-p-1} \chi_{N-p-1}^2 \quad (1.1.25)$$

□

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:

$$\hat{\sigma}^2 = \frac{1}{N-r} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (1.1.26)$$

Por tanto, ahora ya podemos establecer, conocidas las distribuciones de los estimadores construidos que [26, 30]:

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}}, j = 1, \dots, p+1 \sim t_{N-p-1} \quad (1.1.27)$$

Este estadígrafo permite hacer contrastes en los que se comprobará si una variable es estadísticamente significativa para el modelo. Basta con plantear la hipótesis nula $H_0 : \beta_j = 0$.

Ahora, 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 que sea subconjunto del anterior, si además contamos que :

$$\frac{RSS(\hat{\beta})}{N - p - 1} = \frac{1}{N - p - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi_{N-p-1}^2 \quad (1.1.28)$$

Se puede definir el estadígrafo:

$$F = \frac{\frac{(RSS_0 - RSS_1)}{p_1 - p_0}}{\frac{RSS_1}{N - p_1 - 1}} \quad (1.1.29)$$

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:

$$F \sim \frac{\chi_{p_1-p_0}^2}{\chi_{N-p_1-1}^2} = F_{(p_1-p_0), (N-p_1-1)} \quad (1.1.30)$$

Basta tener en cuenta la ecuación 1.1.28 entonces $RSS_0 - RSS_1$ es una suma de $p_1 - p_0$ cuadrados de variables que siguen la normal estándar multiplicada por σ^2 . \square

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

En resumen, podemos hacer contrastes sobre la importancia o de una única variable o de un conjunto de variables. Así como sobre la varianza del modelo en general.

1.1.3. Regresión multivariante

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

$$Y_k = \beta_{0k} + \sum_{j=1}^p X_j \beta_{jk} + \varepsilon_k = f_k(\mathbf{x}) + \varepsilon_k, \quad k = 1, \dots, K \quad (1.1.31)$$

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

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad (1.1.32)$$

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 de las variables explicativas o independientes 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 los 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, Σ , entonces es necesario hacer la siguiente modificación en el RSS :

$$RSS(\mathbf{B}, \Sigma) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^T \Sigma^{-1} (y_i - f(\mathbf{x}_i)) \quad (1.1.33)$$

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 Σ se define de la siguiente manera[11]:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)} \quad (1.1.34)$$

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 (1.1.29) 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 algoritmos 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 $\lambda > 0$:

$$PRSS(\beta) = \sum_{i=1}^n (\mathbf{y}_i - \mathbf{x}_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (1.1.35)$$

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 sin eliminar de manera total las variables.

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 [28].

En los casos de la clasificación o discriminación, se tienen variables respuestas que son cualitativas o categóricas. Por tanto, el modo de tratar los datos cambia de manera significativa en un primer momento.

En este caso, la clasificación y la discriminación se diferencian, según Johnson, y Wichern, en que la discriminación busca describir, ya sea gráficamente (análisis canónico de poblaciones) o algebraicamente (análisis discriminante lineal, logístico o cuadrático) las características de las distintas poblaciones [37]. En el otro lado, la clasificación busca asociar nuevas observaciones a las distintas clases o poblaciones. Aún así, las distintas técnicas descritas se pueden usar para ambas tareas, ya que al describir las características que posee una cierta población se pueden realizar nuevas asignaciones.

Las distintas técnicas de estudio de la dependencia entre variables continuas y no continuas asumen hipótesis y buscan objetivos distintos:

- 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 [58]. Es en este caso en el que se difuminan los límites entre discriminación y clasificación, en particular, se hace un trabajo de discriminación para luego realizar nuevas clasificaciones.
- Cuando no se conozca las distribuciones de probabilidad de las distintas poblaciones, se puede utilizar con fines exploratorios el análisis canónico de poblaciones en el que se estudian las direcciones de mayor distancia entre las dos poblaciones, también se le llama análisis discriminante lineal de Fisher [45].

Para simplificar los cálculos y los desarrollos, en el caso de la primera parte en la que se desarrollan las funciones discriminantes, se consideran dos poblaciones en las que en un inicio se conocen las distribuciones de probabilidad de cada una de las poblaciones [37] y luego se asumirá que son normales para construir el discriminante lineal habitual en la que aparece la distancia de Mahalanobis [58] aunque Morrison [53] construye dicha función discriminante buscando una combinación lineal que maximice el estadístico T^2 de Hotelling el cual mide la diferencia de las medias entre varias poblaciones [37, 51]. Añadir que para esto último, se utiliza la matriz de covarianzas muestral, mientras que el desarrollo dado durante la memoria se hará suponiendo conocida la matriz de covarianzas poblacional.

En la segunda parte de la sección se detallará el método del análisis canónico de poblaciones o análisis discriminante de Fisher que toma la matriz de covarianzas dentro y entre de las poblaciones permitiendo reducir la dimensión de la matriz en un espacio de dimensión menor [45].

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, f_1, f_2 respectivamente. Entonces, si además se conoce la probabilidad de clasificación errónea $P(1|2), P(2|1)$. Se tiene que:

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

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 [11]:

Definición 1.2.1. Se llama función discriminante, f_d aquella que:

$$f_d : \Omega \longrightarrow \mathbb{R} \quad (1.2.2)$$

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

En el caso anterior de dos poblaciones

$$f_d(\mathbf{x}_0) = P(1|\mathbf{x}_0) - P(2|\mathbf{x}_0) = f_1(\mathbf{x}_0)P(1) - f_2(\mathbf{x}_0)P(2) \quad (1.2.3)$$

sería la función discriminante más simple para discernir si \mathbf{x}_0 pertenece a π_1 o no.

Definición 1.2.2. Se llama error de clasificación al coste de clasificar en una población erróneamente una observación y se denota como $c(i|j) = \text{"Error de clasificar en } \pi_i \text{ una observación perteneciente a } \pi_j \text{"}$.

En general, la función de pérdida de un problema de clasificación se da en forma de matriz de pérdida \mathbf{L} de tamaño $L \times L$ con la diagonal nula, que puede ser simétrica o no y donde $L_{ij} = c(i|j)$ [30].

Podemos penalizar las probabilidades de cada una de las poblaciones utilizando el coste. De esta manera, aunque se cometa error se potencia clasificar en la población que menor coste tenga:

$$f_d(\mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)P(1)}{c(1|2)} - \frac{f_2(\mathbf{x}_0)P(2)}{c(2|1)} \quad (1.2.4)$$

En el caso de que la probabilidad de pertenecer a cada una de las dos poblaciones sea 0,5 y en cambio, el coste de clasificar mal en la primera población algo de la segunda sea menor, entonces se clasificará como un elemento de la primera población.

En resumen, cuando los demás términos sean iguales, elegiremos la opción con el menor costo, la mayor verosimilitud o la mayor probabilidad a priori [58].

Johnson y Wichern desarrollan otro enfoque, tomando el coste medio esperado (*ECM* en inglés) [37]:

$$ECM = c(2|1)P(2|1)P(1) + c(1|2)P(1|2)P(2) \quad (1.2.5)$$

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 las funciones de densidad de distribuciones normales con medias distintas μ_1, μ_2 pero con una matriz de covarianzas común Σ . Entonces tenemos la siguiente expresión [11, 37]:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i) \right)} \quad (1.2.6)$$

De esta manera, la función discriminante se puede transformar sustituyendo las funciones de densidad f_1, f_2 por la expresión anterior y tomando logaritmos :

$$(\mathbf{x} - \mu_1)^T \Sigma^{-1} (\mathbf{x} - \mu_1) + \log \left(\frac{P(1)}{c(1|2)} \right) - (\mathbf{x} - \mu_2)^T \Sigma^{-1} (\mathbf{x} - \mu_2) - \log \left(\frac{P(2)}{c(2|1)} \right) \quad (1.2.7)$$

Hay que tener en cuenta que el término $D_i^2 = (\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i)$ es la *distancia de Mahalanobis* de la observación \mathbf{x} a la media de i -ésima población. En cambio, los términos logarítmicos son términos que ponderan el coste y la probabilidad a priori, de manera que si la probabilidad es menor que el coste, penalizarán esa clasificación y si no la potencian. En el caso de que tanto las probabilidades como el coste fueran iguales, entonces la función discriminante para la población 1 sería:

$$D_1^2 - D_2^2 \quad (1.2.8)$$

De esta manera, se obtiene un criterio geométrico 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 opciones tiene de ser clasificada de esa manera [11]. En particular, Peña desarrolla el enfoque geométrico desde la función discriminante antes obtenida dando, en el caso de 2 variables independientes, una visualización de las dos poblaciones [58].

Wichern, D.W. y Johnson, desarrollan el caso cuando no se conoce la matriz de covarianzas poblacional y cuando ambas matrices de cada una de las poblaciones π_1, π_2 no son iguales, obteniendo la regla discriminante cuadrática cuyo razonamiento es similar, sustituyendo las funciones de densidad adecuadas [37].

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 costes asociados, se han desarrollado funciones discriminantes. Sin embargo, hay métodos que afrontan el problema de manera exploratoria, por tanto, se busca analizar cómo y cuánto se estructura la diferenciación de las poblaciones.

Este estudio de las diferencias entre poblaciones se puede realizar teniendo en cuenta cómo es la variabilidad entre grupos y cómo es dentro de las poblaciones. En este caso, se buscarán cuales son las direcciones en las cuales se maximiza la variabilidad entre grupos con respecto a la varianza dentro de las poblaciones.

A continuación, se describirá el método para calcular una nueva base del espacio de las variables que cumpla con el objetivo antes descrito [37, 45].

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 por L poblaciones. Sea I_l el conjunto de observaciones pertenecientes al l -ésimo grupo $\forall l = 1, \dots, L$, sea también N_l el número de observaciones que pertenecen al l -ésimo grupo $\forall l = 1, \dots, L$.

Definición 1.2.3. Se define la media muestral de la j -ésima variable como:

$$\bar{x}_j = \frac{1}{N} \sum_{i=1}^N x_{ij}, \forall j = 1 \dots p \quad (1.2.9)$$

Para un conjunto de N observaciones realizadas en p variables.

Definición 1.2.4. Se define la media muestral de la j -ésima variable en la l -ésima población :

$$\bar{x}_{jl} = \frac{1}{N_l} \sum_{i \in I_l} x_{ij} \quad (1.2.10)$$

Por ende, podemos evaluar la relación entre dos variables $X_j, X_{j'}$ como:

$$Cov(X_j, X_{j'}) = \frac{1}{N} \sum_{i=1}^N (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \quad (1.2.11)$$

Esto se puede particionar por grupos de la siguiente manera:

$$Cov(X_j, X_{j'}) = \frac{1}{N} \sum_{l=1}^L \sum_{i \in I_l} (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \quad (1.2.12)$$

A su vez cada uno de los $(x_{ij} - \bar{x}_j)$ se pueden dividir en la parte intergrupos e intragrupos.

Definición 1.2.5. La covarianza intragrupo de un par de variables $X_j, X_{j'}$ en la l -ésima población se define como:

$$W_l(X_j, X_{j'}) = \frac{1}{N_l} \sum_{i \in I_l} (x_{ij} - \bar{x}_{jl})(x_{ij'} - \bar{x}_{j'l}) \quad (1.2.13)$$

Definición 1.2.6. La covarianza intergrupos de un par de variables $X_j, X_{j'}$ se define como:

$$B(X_j, X_{j'}) = \sum_{l=1}^L \frac{N_l}{N} (\bar{x}_{jl} - \bar{x}_j)(\bar{x}_{j'l} - \bar{x}_{j'}) \quad (1.2.14)$$

Es decir, se puede simplificar la covarianza entre dos variables de la siguiente manera.

$$(x_{ij} - \bar{x}_j) = (x_{ij} - \bar{x}_{jl}) + (\bar{x}_{jl} - \bar{x}_j) \quad (1.2.15)$$

Sustituyendo y simplificando lo necesario:

$$Cov(X_j, X_{j'}) = \frac{1}{N} \sum_{l=1}^L \sum_{i \in I_l} (x_{ij} - \bar{x}_{jl})(x_{ij'} - \bar{x}_{j'l}) + \sum_{l=1}^L \frac{N_l}{N} (\bar{x}_{jl} - \bar{x}_j)(\bar{x}_{j'l} - \bar{x}_{j'}) \quad (1.2.16)$$

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

$$\mathbf{T} = \mathbf{W} + \mathbf{B} \quad (1.2.17)$$

Donde:

- \mathbf{T} es la matriz de tamaño $p \times p$ que expresa la covarianza total y sus coeficientes $t_{jj'} = Cov(X_j, X_{j'})$, $\forall j, j' = 1, \dots, p$
- \mathbf{W} es la matriz que expresa la covarianza dentro de los grupos y sus coeficientes son $w_{jj'} = \frac{1}{N} \sum_{l=1}^L \sum_{i \in I_l} (x_{ij} - \bar{x}_{jl})(x_{ij'} - \bar{x}_{j'l})$, $\forall j, j' = 1, \dots, p$
- \mathbf{B} es la matriz que expresa la covarianza entre los grupos y sus coeficientes son $b_{jj'} = \sum_{l=1}^L \frac{N_l}{N} (\bar{x}_{jl} - \bar{x}_j)(\bar{x}_{j'l} - \bar{x}_{j'})$, $\forall j, j' = 1, \dots, p$

Para cualquier combinación lineal que se quiera hacer de las variables de entrada X_1, \dots, X_p de la forma $\mathbf{x}\mathbf{a}$, donde el vector \mathbf{a} es un vector columnas de p constantes, entonces la varianza se transforma de la siguiente manera:

$$Var(\mathbf{x}\mathbf{a}) = \mathbf{a}^T \Sigma \mathbf{a} \quad (1.2.18)$$

Entonces transformando por el vector \mathbf{a} tenemos que la ecuación (1.2.17) se escribe de la siguiente manera:

$$\mathbf{a}^T \mathbf{T} \mathbf{a} = \mathbf{a}^T \mathbf{B} \mathbf{a} + \mathbf{a}^T \mathbf{W} \mathbf{a} \quad (1.2.19)$$

Recopilando, el objetivo del análisis discriminante lineal es encontrar combinaciones lineales que maximicen la varianza entre grupos con respecto a la varianza total. Eso es equivalente a encontrar el vector \mathbf{a} tal que [45]:

$$f(\mathbf{a}) = \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{T} \mathbf{a}} \quad (1.2.20)$$

sea máxima. utilizando la restricción $\mathbf{a}^T \mathbf{T} \mathbf{a} = 1$. Utilizando el método de los multiplicadores de Lagrange derivamos respecto del vector \mathbf{a} tendremos que:

$$L(\mathbf{a}) = \mathbf{a}^T \mathbf{B} \mathbf{a} - \lambda(\mathbf{a}^T \mathbf{T} \mathbf{a} - 1) \quad (1.2.21)$$

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

$$\frac{\partial L(\mathbf{a})}{\partial \mathbf{a}} = 2\mathbf{B}\mathbf{a} - 2\lambda\mathbf{T}\mathbf{a} \quad (1.2.22)$$

En consecuencia:

$$\mathbf{B}\mathbf{a} = \lambda\mathbf{T}\mathbf{a} \quad (1.2.23)$$

Si además \mathbf{T} es no singular

$$\mathbf{T}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a} \quad (1.2.24)$$

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

Definición 1.2.7. Al valor λ se le conoce como *potencia discriminante* los pesos de la combinación \mathbf{a} .

Johnson y Wichern realizan un desarrollo análogo maximizando $\frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}$ [37].

Las variables canónicas aunque no sea su objetivo principal también pueden usarse para construir funciones discriminantes en pos de clasificar nuevas observaciones. Sea \mathbf{x}_0 la nueva observación y \mathbf{x}'_0 su expresión en la nueva base de las variables canónicas, sea además \bar{x}_l la media muestral de cada una de las $l = 1, \dots, L$ poblaciones y \bar{x}'_l su expresión en la nueva base $\forall l = 1, \dots, L$. Entonces se puede clasificar la nueva observación en la l -ésima población como aquella que tenga la mínima distancia a la media de la población l -ésima.

Por otro lado, para determinar cuantas variables canónicas se puede utilizar la varianza entre grupos que explica cada variable canónica.

Proposición 1.2.1. La varianza entre grupos explicada por cada variable canónica es el valor propio de la matriz $\mathbf{T}^{-1}\mathbf{B}$ 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 $\mathbf{a}_i^T\mathbf{T}\mathbf{a}_i = 1$, entonces los vectores son unitarios respecto a la métrica que induce la matriz de covarianzas, entonces, la varianza explicada por cada variable es $B(\mathbf{a}_i) = \mathbf{a}_i^T\mathbf{B}\mathbf{a}_i$, por ser valores propios de $\mathbf{T}^{-1}\mathbf{B}$ se cumple que $\mathbf{B}\mathbf{a}_i = \lambda_i\mathbf{T}\mathbf{a}_i$, entonces

$$B(\mathbf{a}_i) = \mathbf{a}_i^T\mathbf{B}\mathbf{a}_i = \lambda_i\mathbf{a}_i^T\mathbf{T}\mathbf{a}_i = \lambda_i \quad (1.2.25)$$

□

Para reducir la dimensionalidad, podemos seleccionar un número de variables canónicas estableciendo un umbral. Elegiremos las $m < \min(p, L - 1)$ primeras variables canónicas cuya varianza explicada acumulada sea mayor que dicho umbral [37]. Esto nos permite reducir la dimensionalidad de los datos de manera similar al Análisis de componentes principales. Sin embargo, hay una diferencia clave: mientras el análisis de componentes principales busca representar la variabilidad de los datos explicando la máxima variabilidad total, las variables canónicas buscan identificar 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, Stinchcombe y White [33].

Una neurona artificial es mucho más simple que una neurona, López, Balsa-Canto y Oñate definen una neurona en términos matemáticos [44]. pero antes hay que definir los siguientes conceptos, para los cuales se han utilizado como base [27, 55].

Sea un vector aleatorio \mathbf{x} de longitud p , llamamos datos de entrada al vector \mathbf{x}_0 que representa cada una de las observaciones de las p variables medidas. Por otro lado, se llamará datos de salida al vector \mathbf{y}_0 obtenido tras haber introducido en la red neuronal el vector \mathbf{x}_0 .

Ahora se detallan todos los elementos de una neurona.

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 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 rectificada* 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 se puede definir la siguiente 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 1.3.1 proporciona una forma sencilla de entender el funcionamiento de dicho modelo, incluyendo la analogía de las neuronas biológicas.

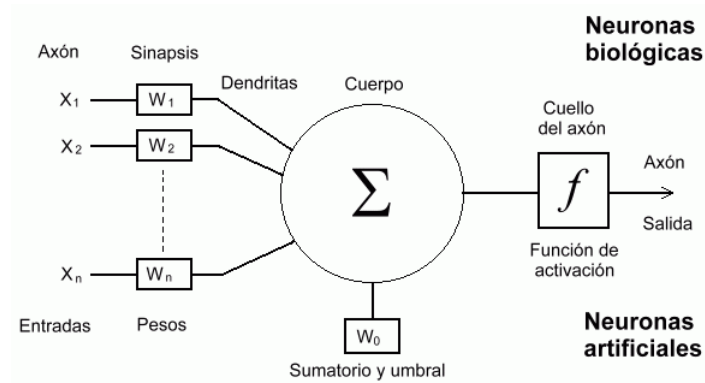


Figura 1.3.1: Representación de una neurona en la que se comparan los elementos biológicos y artificiales. Obtenida de [62]

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 son la entrada del siguiente.

Se pueden establecer varios tipos de capas de neuronas [55]:

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).

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 que hace la red neuronal de las variables de salida.

Para el proceso de ajuste se utiliza de manera habitual el método del gradiente con un conjunto de datos con N observaciones. En el capítulo 11 Hastie et.al. detallan en profundidad el ajuste [30]. A este proceso se le llama *back-propagation*, ya que una vez aplicado el método del gradiente se van actualizando los parámetros anteriores.

Si queremos expresar el modelo de manera concisa, la complejidad de interpretación aumenta significativamente al expandir la red neuronal. Incluso en casos simples como en la figura 1.3.2 la interpretación se vuelve complicada. Por lo tanto, las redes neuronales se utilizan principalmente con propósitos predictivos, en lugar de brindar una explicación clara de los parámetros y relaciones involucrados en el modelo [30, 36].

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 sobreajuste.

La siguiente imagen 1.3.2 es una representación de una red neuronal como un grafo, en el que cada nodo es una neurona, en particular, los capas azules son capas ocultas, formando 5 capas ocultas, mientras que las amarillas son capas de entrada, y las rojas de salida.

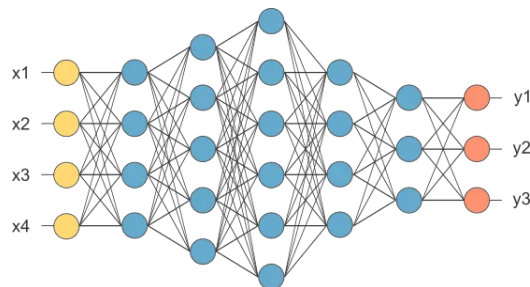


Figura 1.3.2: Imagen extraída directamente de www.neuraldesigner.com

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 [32, 55].

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.

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, entre los principales objetivos de los árboles de decisión, se encuentran la selección y evaluación de la importancia de las variables[9, 67]. Además de que puede tener fines predictivos o incluso de manejo de los datos, valores perdidos etc...[54].

Otra ventaja de este tipo de métodos es que permite trabajar con conjuntos de datos en los que se estudian un mayor número de variables que de observaciones, es decir, $p > N$ [12].

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 se considera en los algoritmos desarrollados durante la sección, normalmente se eliminan particiones que no aporten demasiado. A este proceso se le llama “poda”. Es por esta problemática que se desarrolla el análisis de modelos mediante la descomposición en sesgo y varianza de los mismos.

Se detallan los principales conceptos y notación a continuación.

1.4.1. Conceptos básicos

Para empezar hay que definir el concepto inicial de un árbol de decisión [30].

Definición 1.4.1. Un árbol de decisión es un conjunto de reglas discriminantes que fraccionan el espacio de observaciones de acuerdo a un cierto criterio. Este conjunto genera un grafo de tipo árbol con un único nodo raíz y donde los nodos hoja representan cada una de las regiones en las que se ha fraccionado el espacio.

Como se ha dicho un árbol de decisión genera particiones del espacio de observaciones que se definen de la siguiente manera [30].

Definición 1.4.2. Se llama separación, partición o división de índice (j, s) a la fragmentación del espacio inicial dado en las siguientes regiones R_1, R_2 :

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

Hay que tener en cuenta que esto sería en el caso en el que la variable a separar X_j sea continua.

En el caso de que la variable a particionar sea categórica o cualitativa Breiman propone la siguiente solución [5]. Supongamos que X_j toma L valores distintos en un nodo en específico. Entonces podemos tomar un conjunto de valores $L_1 \subset L$ de tal manera que las regiones que se obtienen son:

$$R_1 = \{\mathbf{x}_i \in \mathbb{R}^p / x_{ij} = l / l \in L_1\} \quad R_2 = \{\mathbf{x}_i \in \mathbb{R}^p / x_{ij} = l / l \notin L_1\} \quad (1.4.2)$$

Definición 1.4.3. Se define como nodo de un árbol de decisión, a cada una de las regiones resultantes después de aplicar una separación.

En particular se tienen los nodos terminales o hoja y los nodos padre y nodos hijos [9].

Definición 1.4.4. Se define un nodo terminal o nodo hoja en un árbol de decisión como cada una de las regiones finales resultante de la partición del espacio de observaciones.

Definición 1.4.5. Se denomina nodo padre al nodo previo a una separación y se definen los nodos hijos a los resultantes de una determinada partición.

Además podemos dar varias medidas de la complejidad del modelo [30].

Definición 1.4.6. Se denomina tamaño del árbol T , $|T|$ al número de nodos terminales.

Definición 1.4.7. Se llama profundidad del árbol al número máximo que de divisiones necesarias para llegar a un nodo terminal.

Las siguientes imágenes 1.4.1b, 1.4.1a muestran el diagrama resultante tras dividir el espacio de observaciones mediante un árbol.

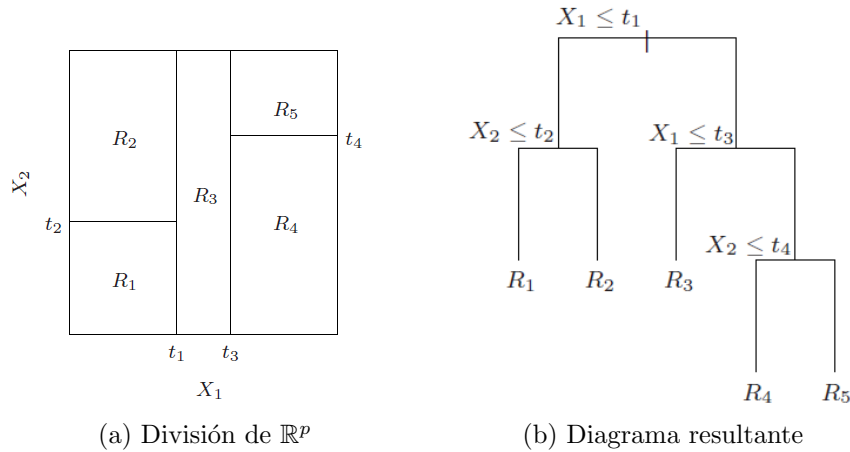


Figura 1.4.1: Representación de la división de \mathbb{R}^p y el diagrama de árbol resultante. Ambas procedentes de [30]

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 estudiar la relación de un conjunto de variables con otra variable continua, y un árbol de clasificación cuando la variable respuesta es categórica o cualitativa. 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.

Observación: La razón de usar particiones binarias es que son fáciles de entender y manejar. Esto hace que los árboles de decisión sean más fáciles de usar y nos ayuda a comprender las áreas o regiones que hemos creado al particionar.

1.4.2. Árboles de regresión

A continuación, examinemos el proceso de crecimiento de un árbol de regresión. Se considera un vector aleatorio \mathbf{x} con p variables predictoras y una variable respuesta continua Y . Aunque hay varios algoritmos para hacer crecer los árboles de regresión se detallará el algoritmo CART (árboles de clasificación y regresión) de Breiman [5]. Se elige este ya que es el más importante y fundamental.

Supongamos que hemos establecido un número máximo de particiones $M = |T|$. El objetivo del árbol de regresión es dividir el espacio de observaciones en regiones R_m para $m = 1, \dots, M$. 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:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^M \hat{f}_m(\mathbf{x}) \cdot \mathbf{1}_m(\mathbf{x}) \quad (1.4.3)$$

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. [30, 4].

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 (Para ver demostración completa véase la sección de 9.1 [5]) :

$$\hat{c}_m = \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} y_i \quad (1.4.4)$$

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 pertenecen a 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, la predicción \hat{y} para un $\mathbf{x} \in R_m$ otorgada por el modelo es \hat{c}_m . [30, 5]

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 [5]:

$$\sum_{i/\mathbf{x}_i \in R_{m_1}} (y_i - \hat{c}_{m_1})^2 + \sum_{i/\mathbf{x}_i \in R_{m_2}} (y_i - \hat{c}_{m_2})^2 \quad (1.4.5)$$

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 y Scornet lo plantean de manera distinta [4, 5]. Podemos definir el error cuadrático medio de cada región de la siguiente manera [30].

Definición 1.4.8. 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

$$Q(T) = \sum_{m=1}^M \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2 \quad (1.4.6)$$

Ya que el error cometido medio en el espacio de observaciones es la media de todos los errores cometidos en cada una de las regiones en las que se particiona el espacio [5].

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:

$$EPE(\mathbf{x}_0) = \sigma^2 + (Sesgo(\hat{f}_m(\mathbf{x}_0)))^2 + Var(\hat{f}(\mathbf{x}_0)) \quad (1.4.7)$$

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

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

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:

$$EPE(\mathbf{x}_0) = \sigma^2 + \frac{\sigma^2}{N_m} \quad (1.4.9)$$

Por tanto, hacer crecer un árbol, es decir, hacer que tenga más nodos terminales, 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 1.4.9. 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 y Breiman proponen añadir un término al coste del árbol [5, 14].

$$Q_\alpha(T) = \sum_{m=1}^M \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2 + \alpha|T| \quad (1.4.10)$$

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 sugieren Divakaran, basado en el que da Breiman viene dado por los siguientes pasos [5, 14]:

- Primero prepara el conjunto de datos para poder realizar validación cruzada (Véase el capítulo 7 de [30], 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 aplica el método de *K-folds* para elegir el parámetro α [14].

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 capítulo 7 de [30] o el capítulo 5 de [36]).

Otro algoritmo de selección de las particiones es el algoritmo *C4.5* de Quinlan 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 [60].

1.4.3. 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 , categóricas o cualitativas, pero en el que la variable respuesta es una variable aleatoria con L posibles modalidades o clases.

Supóngase que se han tomado N observaciones de las variables predictoras y respuesta. Supóngase que la variable respuesta Y , en este caso, categórica tiene L categorías posibles las cuales se denotan como $l, l = 1, \dots, L$. Sea también un árbol T de tamaño M que genera una partición del espacio con las regiones $R_m, m = 1, \dots, M$. Entonces se puede denotar de la siguiente manera:

$$\hat{p}_{lm} = \text{Proporción de observaciones en las que } y_i = l, \text{ en la región } R_m. \quad (1.4.11)$$

Teniendo en cuenta esto, se pueden definir los siguientes conceptos [14, 30, 36, 9].

Definición 1.4.10. Se dice que un nodo correspondiente a la región R_m es puro, si

$$\exists l_0 \in \{1, \dots, L\} / p_{l_0 m} = 1 \text{ y } \hat{p}_{lm} = 0 \quad \forall l \neq l_0 \quad (1.4.12)$$

Es decir, en ese nodo la variable respuesta pertenece a una única categoría.

Definición 1.4.11. Se define la impureza de un nodo, correspondiente a la región R_m como :

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

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 [30, 14, 36]:

Definición 1.4.12. Se llama índice Gini de una región R_m a la siguiente expresión :

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

Esta medida es la impureza que tiene un nodo.

Definición 1.4.13. Se llama entropía a la siguiente medida :

$$H(\hat{p}_{lm}) = -\hat{p}_{lm} \log(\hat{p}_{lm}) - (1 - \hat{p}_{lm}) \log(1 - \hat{p}_{lm}) \quad (1.4.15)$$

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 sólo clasificaciones erróneas o correctas.

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

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

Esta medida nos da una medida de la impureza del nodo, ya que si tuviéramos algún l que cumpliera $\hat{p}_{lm} = 1$ entonces $H(R_m) = 0$. Esto se da ya que $\log(\hat{p}_{lm}) = 0$ y $\hat{p}_{l'm} = 0 \quad \forall l' \neq l = 1, \dots, L$.

El algoritmo *C4.5* de Quinlan hace uso de otro criterio para hacer crecer el árbol de clasificación [60]. Hay que definir el concepto de ganancia de información y de ratio de ganancia [9].

Definición 1.4.14. Supóngase que se tiene un nodo padre que se divide en k nodos hijos, entonces la *ganancia de información* se define de la siguiente manera:

$$Ganancia = H(padre) - \sum_{i=1}^k \frac{N_i}{N_{padre}} H_i \quad (1.4.17)$$

donde $H(padre)$ es la entropía del nodo padre antes de la separación, H_i es la entropía de cada una de las k regiones creadas en la separación. N_i, N_{padre} es el número de observaciones en los nodos hijos y el nodo padre respectivamente.

Definición 1.4.15. Llamamos ratio de ganancia a la siguiente cantidad que está entre 0 y 1 :

$$\frac{Ganancia}{H_{padre}} \quad (1.4.18)$$

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* de la siguiente manera [47].

En el caso de que $X_j, j = 1 \dots p$ sea una variable categórica o cualitativa, 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.

En particular, tomando este criterio de separación se puede hacer crecer un árbol de manera análoga al crecimiento de un árbol de regresión. Es decir, se van aplicando las particiones más beneficiosas en cada punto siguiendo el criterio dado hasta llegar a un número de particiones máximo o un número de observaciones mínimo en cada nodo hoja.

Otro tipo de árboles sería el algoritmo *CHAID* (*Chi-Square Automatic Interaction*) que en cada partición a realizar se realiza análisis Chi cuadrado para discernir que variable predictora tiene una mayor relación con la variable respuesta, de esta manera obtiene los grupos más diferenciados [40].

En general, los distintos algoritmos de crecimiento se diferencian en el criterio escogido para crear las particiones.

1.4.4. 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, la capacidad de obtener, dadas nuevas observaciones \mathbf{x}_0 de las variables predictoras, el valor que tendría la variable respuesta Y . La razón de que se detalle ahora es que los árboles son modelos que tienden al sobreajuste y algunos autores como Breiman o Divakaran proponen mecanismos para evitar dicha situación e incluso proponen utilizar técnicas alternativas basadas en los árboles de decisión como los bosques aleatorios [5, 7, 14].

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. En particular en un árbol de regresión

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 [30, 43].

Definición 1.4.16. Se llama *error de predicción esperado* de la observación \mathbf{x}_0 a la siguiente expresión:

$$EPE(\mathbf{x}_0) = \mathbb{E}((Y - \hat{Y})^2 | \mathbf{x} = \mathbf{x}_0) = \mathbb{E}((Y - \hat{f}(\mathbf{x}_0))^2) \quad (1.4.19)$$

De esta manera, se tiene una forma de medir el rendimiento predictivo de un modelo. En aplicaciones de aprendizaje automático, donde el principal el objetivo es la predicción, dividimos los datos en conjuntos de entrenamiento y validación. Después de ajustar el modelo, evaluamos qué tan bien puede predecir utilizando una muestra separada de los datos de entrenamiento. Esto nos ayuda a estimar el error esperado en las predicciones.

Además, 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 [30]:

$$EPE(\mathbf{x}_0) = \sigma_\epsilon^2 + Sesgo(\hat{f}(\mathbf{x}_0))^2 + Var(\hat{f}(\mathbf{x}_0)) \quad (1.4.20)$$

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

Demostración.

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

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

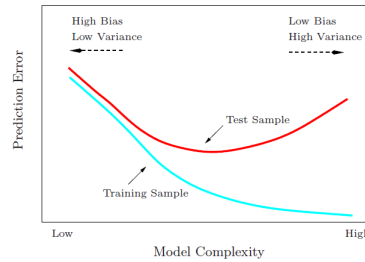
$$\mathbb{E}((Y - \hat{f}(\mathbf{x}_0))^2) = \sigma^2 + Sesgo(\hat{f}(\mathbf{x}_0))^2 + Var(\hat{f}(\mathbf{x}_0))$$

□

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 [55].

Esto se puede observar en el siguiente gráfico que incluyen *Hastie et.al.* de manera que se puede ver a mismo número de observaciones que pasa si aumentamos las variables observadas [30]. 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.

En la figura se puede observar 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.

1.4.5. 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.* [7]. 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.* [4, 8] :

- 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 [31] 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. [8] 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.* [4] destacan que no es habitual ver un estudio profundo de estos métodos debido

Breiman, L. [6] 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 buscan dentro de los datos, patrones, estructuras latentes, etc... Es decir, se puede llegar a estudiar el comportamiento de los datos observando las relaciones entre las distintas variables u observaciones. En estos casos, no hay una o varias variables respuesta que dependan de un conjunto de variables predictoras de las cuales se estudie su dependencia.

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 mismos, 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 [2] 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 [21], obtener nuevas variables que representen la misma información o gran parte de ello. 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, en el sentido de una medida de la diferencia entre el valor predicho y el valor real de una variable respuesta, 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 [58], o la proporción de variabilidad explicada en cada componente [10].

Por falta de tiempo y de extensión del trabajo, no se desarrollarán técnicas como las reglas de asociación o el escalamiento multidimensional para los cuales Hastie et.al. aportan una visión introductoria a los mismos [30]. Por otro lado, James et.al. desarrollan de manera similar con la particularidad que además lo implementan en el entorno R con ejemplos [36] . (Veáanse también [21] y [37]).

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 y Hotelling con distintos enfoques [59, 34].

Pearson, en su aproximación, buscó la forma de ajustar de mejor manera puntos de un espacio de dimensión p a una recta o plano [59]. Es decir, Pearson buscó un enfoque de optimización geométrica que llevó a las componentes principales.

Por el otro lado, Hotelling, 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 [34]. 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 y vectores propios que se desarrolla en esta sección.

Aún así, el trabajo de Hotelling tenía ciertas diferencias con el enfoque que actualmente se le da actualmente a las componentes principales [39]. 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 y Collins desarrollan ambos enfoques [10].

Según Abdi [2] 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 ya que a fin de cuentas la información que expresan estaría repetida [10, 21].

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 [17].

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 y Young y demuestran que es invariante ante transformaciones ortogonales [17].

Como ejemplo de las aplicaciones que puede tener esta técnica Johnson y Wichern desarrollan un ejemplo sencillo en el cual se estudian 5 variables socioeconómicas de ciertas zonas de Estados Unidos [37].

En este ejemplo, las dos primeras componentes explican el 93 % de la variabilidad y la primera es una diferencia ponderada de el empleo privado 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 [63]. Estudio 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 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 y Asari 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 [25].

2.1.1. Definición y cálculo de las componentes principales

Sea un vector aleatorio de longitud p , $\mathbf{x} = [X_1, \dots, X_p]$, con una distribución $N(\mu, \Sigma)$, donde μ es el vector de medias y Σ la matriz de covarianzas con $r = \text{rg}(\Sigma)$, entonces, se pueden definir las componentes principales [11].

Definición 2.1.1. Se definen las componentes principales

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

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

- Si $j = 1$ $\text{Var}(Z_1)$ es máxima restringido a $\mathbf{v}_1^T \mathbf{v}_1 = 1$
- Si $j > 1$ debe cumplir:
 - $\text{Cov}(Z_j, Z_i) = 0 \quad \forall i \neq j$
 - $\mathbf{v}_j^T \mathbf{v}_j = 1$
 - $\text{Var}(Z_j)$ 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. Añadir que como máximo se pueden obtener r componentes principales.

Chatfield y *Collins* utilizan el método de los multiplicadores de Lagrange para resolver el problema de maximizar la varianza de la primera componente sujeto a la restricción $\mathbf{v}_1^T \mathbf{v}_1 = 1$ [10]. Todo esto con el objetivo de calcular el vector \mathbf{v}_1 . El cálculo de sucesivas componentes cambiará en ciertos aspectos.

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

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

$$\frac{\partial L}{\partial \mathbf{v}_1} = 2\Sigma \mathbf{v}_1 - 2\lambda \mathbf{v}_1 = 2(\Sigma - \lambda \mathbf{I})\mathbf{v}_1$$

Igualando a 0 tenemos la siguiente ecuación:

$$\Sigma \mathbf{v}_1 = \lambda \mathbf{v}_1 \quad (2.1.2)$$

Para que \mathbf{v}_1 sea un vector no trivial, se elige λ de tal manera que $|\Sigma - \lambda \mathbf{I}| = 0$, es decir, λ es un vector propio de la matriz de covarianzas, Σ . Al ser ésta una matriz semidefinida 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 $Var(Z_1) = Var(\mathbf{v}_1^T \mathbf{x}^T) = \mathbf{v}_1^T \Sigma \mathbf{v}_1 = \mathbf{v}_1^T \lambda \mathbf{v}_1 = \lambda$, y para maximizarla basta tomar $\lambda = \max\{\lambda_1, \dots, \lambda_r\}$. El resto de valores propios que se puedan calcular serán nulos ya que Σ es de rango r y semidefinida positiva. Reordenando si es necesario, se tiene que $\lambda = \lambda_1$. Por tanto, la primera componente viene dada por el vector propio \mathbf{v}_1 de valor propio λ_1 de la matriz de covarianzas Σ .

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 $\mathbf{v}_2^T \mathbf{v}_1 = 0$. 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 $Z_j = \mathbf{v}_j^T \mathbf{x}^T \quad \forall j = 1, \dots, r$, se tiene entonces que :

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

donde μ es el vector de medias del vector aleatorio \mathbf{x} .

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. \square

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 $\mathbf{v}_2^T \mathbf{v}_1 = \mathbf{v}_1^T \mathbf{v}_2 = 0$. Para este caso existen λ, ϕ de manera que la función a maximizar es:

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

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

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

Al multiplicar todo por \mathbf{v}_1^T , se obtiene que:

$$2\mathbf{v}_1^T \Sigma \mathbf{v}_2 - 2\lambda \mathbf{v}_1^T \mathbf{v}_2 - \phi \mathbf{v}_1^T \mathbf{v}_1 = 0 \quad (2.1.5)$$

Por tanto, teniendo en cuenta las condiciones impuestas al problema, $\phi = 0$. De esta manera, se obtiene en la ecuación lo mismo que en el cálculo de la primera.

Y de manera análoga al caso de la primera componente, $\lambda = \lambda_2$, el segundo valor propio más grande, y \mathbf{v}_2 es el vector propio de valor propio λ_2 de la matriz de covarianzas Σ .

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.

Por ende, se obtiene que las componentes principales vienen dadas por los vectores propios de la matriz de covarianzas Σ . Además sabemos que $Var(\mathbf{v}_j^T \mathbf{x}^T) = \lambda_j$, $\forall j = 1, \dots, r$, siendo donde λ_j es el j -ésimo valor propio más grande no nulo.

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 $\forall j = 1, \dots, r$. Estos a su vez son un conjunto de vectores ortonormales en el 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

$$\mathbf{z} = \mathbf{x}\mathbf{V} \quad (2.1.6)$$

Donde el vector $\mathbf{z} = Z_1, \dots, Z_r$, y donde cada $Z_j = \mathbf{v}_j^T \mathbf{x}$, $\forall j = 1, \dots, r$

Ahora, veamos como esto se puede aplicar a una matriz de datos observados con la matriz de varianzas muestral.

Sea \mathbf{x} el vector aleatorio de longitud p definida de la misma manera. Entonces se puede definir lo siguiente :

Definición 2.1.3. Dados los datos de N observaciones recogidos sobre p variables X_1, \dots, X_p , entonces $\mathbf{z}_{ij} = \mathbf{x}_i \mathbf{v}_j$, $\forall i = 1, \dots, N, \forall j = 1, \dots, p$ es el valor de la j -ésima componente para la observación \mathbf{x}_i .

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 Σ se puede extender al 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 $\bar{\mathbf{X}}$, es decir:

$$\bar{x}_{ij} = x_{ij} - \bar{x}_j \quad \forall i = 1, \dots, N, \forall j = 1, \dots, p \quad (2.1.7)$$

donde \bar{x}_j es la media muestral de la j -ésima variable. Esto hace que $\mathbf{S} = \frac{1}{N-1} \mathbf{X}^T \mathbf{X}$ [58]. Esto permite hablar indistintamente de los vectores propios y valores propios de \mathbf{S} y $\mathbf{X}^T \mathbf{X}$. Ya que si se tiene un endomorfismo T , del cual se tiene un vector propio \mathbf{v} de valor propio λ y tomamos un escalar μ . Entonces, el endomorfismo μT cumple que $\mu T(\mathbf{v}) = \mu \lambda \mathbf{v}$. Por tanto, los vectores propios son los mismos y los valores propios son proporcionales.

2.1.2. Componentes principales por la descomposición en valores singulares

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. Es decir, que los datos pueden representarse con una dimensionalidad menor con la mínima pérdida mediante la matriz reducida.

Otra visión del problema es que se quiere buscar una proyección sobre un subespacio 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. Antes de desarrollar los cálculos debemos revisar las siguientes conceptos.

Hay que recordar que toda matriz de tamaño arbitrario se puede descomponer en tres matrices en lo que se llama la descomposición en valores singulares (*SVD* en inglés) [2].

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

$$\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T \quad (2.1.8)$$

Donde:

- \mathbf{U} matriz ortogonal y de tamaño $N \times r$
- \mathbf{D} matriz de tamaño $r \times r$ diagonal, cuyos elementos no nulos son los valores singulares $\sigma_1 \geq \dots \geq \sigma_r \geq 0$, es decir

$$\mathbf{D} = \begin{pmatrix} \sigma_1 & 0 & \dots & \dots \\ 0 & \sigma_2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & 0 & \sigma_r \end{pmatrix} \quad (2.1.9)$$

- \mathbf{V} matriz ortogonal y de tamaño $p \times r$.

Cada una de las matrices tiene un origen distinto que hay que detallar [38].

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

Demostración. Por la descomposición en valores singulares tenemos que:

$$\begin{aligned}\mathbf{X}^T \mathbf{X} &= (\mathbf{U} \mathbf{D} \mathbf{V}^T)^T (\mathbf{U} \mathbf{D} \mathbf{V}^T) \\ &= \mathbf{V} \mathbf{D}^T \mathbf{U}^T \mathbf{U} \mathbf{D} \mathbf{V}^T \\ &= \mathbf{V} \mathbf{D}^T \mathbf{D} \mathbf{V}^T\end{aligned}$$

donde la matriz $\mathbf{D}^T \mathbf{D}$ es una matriz diagonal de tamaño $r \times r$ 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 r vectores propios de $\mathbf{X}^T \mathbf{X}$ con los valores propios $\sigma_1, \dots, \sigma_r$ respectivamente. \square

Proposición 2.1.3. La matriz \mathbf{U}^T es la matriz de tamaño $N \times r$ que contiene como columnas los r vectores propios con valores propios no nulos de la matriz $\mathbf{X} \mathbf{X}^T$

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

$$\begin{aligned}\mathbf{X} \mathbf{X}^T &= (\mathbf{U} \mathbf{D} \mathbf{V}^T) (\mathbf{U} \mathbf{D} \mathbf{V}^T)^T \\ &= \mathbf{U} \mathbf{D} \mathbf{V}^T \mathbf{V} \mathbf{D}^T \mathbf{U}^T \\ &= \mathbf{U} \mathbf{D} \mathbf{D}^T \mathbf{U}^T\end{aligned}$$

Entonces, tenemos que como es ortonormal $\mathbf{U}^T = \mathbf{U}^{-1}$ de esta manera, se puede comprobar al ser $\mathbf{D} \mathbf{D}^T$ una matriz diagonal $r \times r$. Por tanto, la matriz \mathbf{U}^T es la matriz de diagonalización de la matriz $\mathbf{X} \mathbf{X}^T$ \square

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

En el problema de reconstrucción de datos hay que dar una métrica en el espacio de las matrices, en particular se da la norma de Frobenius ya que cumple ciertas propiedades interesantes [24].

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:

$$\|\mathbf{A}\|_F = (\text{tr}(\mathbf{A}^T \mathbf{A}))^{\frac{1}{2}} = \left(\sum_{i=1}^N \sum_{j=1}^p a_{ij}^2 \right)^{\frac{1}{2}} \quad (2.1.10)$$

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

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

$$\begin{aligned}\|\mathbf{U} \mathbf{A}\|_F^2 &= \text{tr}((\mathbf{U} \mathbf{A})^T (\mathbf{U} \mathbf{A})) \\ &= \text{tr}(\mathbf{A}^T \mathbf{U}^T \mathbf{U} \mathbf{A}) \\ &= \text{tr}(\mathbf{A}^T \mathbf{A}) \\ &= \|\mathbf{A}\|_F^2\end{aligned} \quad \square$$

Además teniendo en cuenta la definición se puede obtener la siguiente propiedad [10].

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

$$\|\mathbf{X}\|_F^2 = (N-1) \sum_{j=1}^p s_{jj}^2 \quad (2.1.11)$$

Donde las s_{jj}^2 son las varianzas muestrales pertenecientes a cada una de las variables X_1, \dots, X_p observadas .

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:

$$\begin{aligned} \|\mathbf{X}\|_F^2 &= \text{tr}(\mathbf{X}^T \mathbf{X}) \\ &= (N-1) \text{tr}(\mathbf{S}) \\ &= (N-1) \sum_{j=1}^p s_{jj}^2 \end{aligned} \quad \square$$

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 $m \leq r$ de \mathbf{X} y se denota como \mathbf{X}_m , a la matriz $N \times m$ resultado de:

$$\mathbf{X}_m = \mathbf{U}_m \mathbf{D}_m \mathbf{V}_m^T \quad (2.1.12)$$

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.

$$\mathbf{D}_m = \begin{pmatrix} \sigma_1 & 0 & \dots & \dots \\ 0 & \sigma_2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & 0 & \sigma_m \end{pmatrix} \quad (2.1.13)$$

- \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.

Una vez definidos los anteriores conceptos, Eckart y Young desarrollan el siguiente teorema que permite afirmar que la matriz reducida es la más próxima a la matriz de datos [17, 24].

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 :

$$\|\mathbf{A} - \mathbf{B}\|_F \leq \|\mathbf{A} - \mathbf{A}_m\|_F \quad \forall \mathbf{B} / \text{rg}(\mathbf{B}) = m \leq r \quad (2.1.14)$$

Johnson desarrolla la demostración del teorema [38].

Por otro lado, 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 [10]:

$$t_m = \frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^r \lambda_j} \quad (2.1.15)$$

donde los λ_j son los valores propios de la matriz \mathbf{S} .

Ahora el número de componentes m a considerar se puede elegir de varias maneras, ya sea representando en un gráfico el número de componentes frente a la proporción de variabilidad explicada $t_j, j = 1, \dots, p$ [39, 58]. O se puede establecer una cota de variabilidad explicada acumulada mínima que deben representar.

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 afirmar que la matriz reducida es la que mejor reconstrucción de los datos da.

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 [69].

En 1888, Galton escuchó la exposición de un artículo 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 un artículo dando pie a los trabajos sobre el estudio de las correlaciones [23].

Es en 1904 cuando Spearman viendo el trabajo de Galton, aplicó esas ideas al estudio de la inteligencia, buscando una medida objetiva de la misma [68]. En este desarrollo, dio las primeras nociones de qué era un factor y calculó una aproximación para un modelo factorial. 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 se puede usar tanto con fines exploratorios como confirmatorios. Es decir, se puede usar para encontrar nuevas informaciones no anticipadas o para confirmar una hipótesis planteada, por ejemplo, Pearson buscaba confirmar que había una estructura latente en la inteligencia [28].

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 comunales como las especificidades 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 [37] sin pérdida de generalidad. Tras la modelización, se desarrolla el método de estimación de la matriz de covarianzas mediante el método del factor principal [58]. Seguidamente, se darán algunos detalles sobre el problema de la no unicidad del modelo y soluciones ante ello. Y por último, se hará una introducción a la interpretación de los datos que arroja el modelo factorial.

Si se quieren observar un estudio completo aplicando el análisis factorial léase [3], estudio en el que se estudia los factores de riesgo psicológico en deportistas universitarios. También es interesante leer [13] en el que se estudia los factores que pueden llevar al consumo de drogas en la preadolescencia. Otro ejemplo sería [57], en este caso para analizar las características de ciertas denominaciones enológicas en el Valle del Loira, en particular, se utiliza el análisis factorial con el fin de reducir la dimensionalidad.

2.2.1. Formalización

Supóngase un vector aleatorio $\mathbf{x} = [X_1, \dots, X_p]$ de longitud p con una distribución $N_p(0, \Sigma)$, centrada sin pérdida de generalidad, donde la matriz de covarianzas Σ es simétrica y definido positiva. El modelo factorial se define [10]:

$$X_j = \lambda_{1j}F_1 + \dots + \lambda_{mj}F_m + \psi_j U_j \quad j = 1 \dots p \quad (2.2.1)$$

Donde:

- λ_{kj} es la saturación de la j -ésima variable en el factor común k -ésimo.
- F_k es el k -ésimo factor común
- ψ_j es la saturación específica de la variable X_j en el factor único.
- U_j es el factor específico para la variable X_j .

El análisis factorial se apoya en las siguientes hipótesis [11]:

- Los factores comunes F_k son variables aleatorias que siguen una distribución marginal $N(0, 1)$. Además se supondrá que $Cov(F_k, F_{k'}) = 0, k \neq k', \forall k, k' = 1, \dots, m$. Además se suponen completamente independientes de los factores específicos.
- Los factores específicos U_j son variables aleatorias con una distribución normal $N(0, 1)$ no correladas. Se suponen completamente independientes de los factores comunes.

En consecuencia, se pueden definir las siguientes matrices:

Definición 2.2.1. Llamaremos matriz de saturaciones de los factores o matriz de saturaciones, Λ de tamaño $p \times m$ a la matriz :

$$\Lambda = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{p1} & \dots & \lambda_{pm} \end{pmatrix} \quad (2.2.2)$$

Definición 2.2.2. Llamaremos matriz específica a la matriz diagonal Ψ de tamaño $p \times p$ a aquella que contiene los términos ψ_j donde $j = 1, \dots, p$:

$$\Psi = \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{pmatrix} \quad (2.2.3)$$

Observación: En distintas fuentes no aparece la matriz Ψ ya que los factores específicos $U_j, (j = 1, \dots, p)$ no se consideran variables estandarizadas no correladas, como sería el caso descrito. Por ejemplo, Mardia et.al. denotan como Ψ a la matriz de covarianzas del vector \mathbf{u} [51].

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

$$\mathbf{x}^T = \Lambda \mathbf{f}^T + \Psi \mathbf{u}^T \quad (2.2.4)$$

Ya que se consideran los vectores aleatorios $\mathbf{x} = [X_1, \dots, X_p]$, $\mathbf{f} = [F_1, \dots, F_m]$, $\mathbf{u} = [U_1, \dots, U_p]$ vectores ordenados en filas.

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

$$\sigma_j^2 = \sum_{k=1}^m \lambda_{kj}^2 + \psi_j^2 \quad \forall j = 1, \dots, p \quad (2.2.5)$$

Demostración.

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

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 [58].

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

Además de la comunalidad se puede definir el concepto de especificidad o unicidad [11].

Definición 2.2.4. Se llama especificidad o unicidad de la variable X_j a ψ_j^2 , $\forall j = 1, \dots, p$

Es decir, la parte de la varianza de una variable X_j , $\forall j = 1, \dots, p$ explicada por los factores comunes es la comunalidad. 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. Pero también se puede interpretar las covarianzas en función de las saturaciones de los factores comunes [53, 10].

Proposición 2.2.2. La covarianza de dos variables $Cov(X_j, X_{j'})$, $\sigma_{jj'}$ cumple que

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

Demostración.

$$\sigma_{jj'} = \sum_{k=1}^m \lambda_{kj} \lambda_{kj'} \mathbb{E}(F_k^2) + \sum_{k=1}^m \sum_{n=1, n \neq k}^m \lambda_{kj} \lambda_{nj'} \mathbb{E}(F_k \cdot F_n) \quad (2.2.7)$$

$$+ \sum_{k=1}^m \lambda_{kj} \psi_{j'} \mathbb{E}(F_k \cdot U_{j'}) + \sum_{k=1}^m \lambda_{kj'} \psi_j \mathbb{E}(F_k \cdot U_j) + \psi_j \psi_{j'} \mathbb{E}(U_j \cdot U_{j'}) \quad (2.2.8)$$

De nuevo, teniendo en cuenta las hipótesis

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

□

Y tomando las proposiciones anteriores se puede tener que la matriz Σ , cumple lo siguiente:

Teorema 2.2.1. La matriz $\Sigma = \Lambda \Lambda^T + \Psi^2$

Este teorema es el más importante del análisis factorial, ya que prueba que el modelo factorial nos permite descomponer la variabilidad de un grupo de variables $X_j, j = 1, \dots, p$ en una parte común y una parte específica a cada una de las variables.

2.2.2. Obtención de la matriz de saturaciones

Supóngase el vector aleatorio \mathbf{x} como antes del cual se recogen N observaciones en la matriz de datos \mathbf{X} de tamaño $N \times p$ que se supondrá centrada sin pérdida de generalidad. Esto provoca que la matriz $\mathbf{S} = \Lambda \Lambda^T + \Psi^2$ [58].

Entonces el objetivo inicial es estimar la matriz de cargas Λ . Para ello se puede tomar la matriz $\mathbf{S} - \Psi^2$. Si de esta matriz se obtiene la descomposición espectral:

$$\mathbf{S} - \Psi^2 = \mathbf{U} \mathbf{D} \mathbf{U}^T \quad (2.2.10)$$

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 que sus columnas son los vectores propios asociados a los valores propios no nulos.

De esta manera, Cuadras toma como estimador de $\hat{\Lambda} = \mathbf{U} \mathbf{D}^{\frac{1}{2}}$, ya que $\hat{\Lambda} \hat{\Lambda}^T = \mathbf{U} \mathbf{D} \mathbf{U}^T = \mathbf{S} - \Psi^2$ [11]. El problema que surge aquí es que la estimación $\hat{\Psi}$ no es conocida, pero se desarrollará en lo siguiente.

De esta manera, supóngase que es conocido el número de factores, m . Entonces se tiene que obtener una matriz reducida de rango $m < rg(\mathbf{X})$ de la siguiente manera.

$$\hat{\Lambda} = (\sqrt{\theta_1} e_1, \dots, \sqrt{\theta_m} e_m) = \mathbf{U}_m \mathbf{D}_m^{\frac{1}{2}} \quad (2.2.11)$$

Donde la pareja $(\theta_k, e_k), k = 1, \dots, m$ es el vector propio e_k en formato columna y θ_k (Se utiliza la letra θ para no confundirla con la saturación, λ_{kj}), su valor

propio asociado. Además la matriz \mathbf{U}_m es la matriz que tiene como columnas los m vectores propios de la matriz $\mathbf{S} - \mathbf{\Psi}^2$ que tienen asociados los m valores propios más grandes.

En el caso de considerar que $\mathbf{\Psi} = 0$, estaríamos ante el caso de la descomposición en valores y vectores propios que se utilizaba en el análisis de componentes principales. Por tanto, la matriz $\hat{\mathbf{\Lambda}}$ tendría por columnas vectores que son proporcionales a los vectores propios de \mathbf{S} , lo que provoca que entonces sean proporcionales a las componentes principales [61].

El mayor problema en este punto es que no se conoce la matriz específica. $\hat{\mathbf{\Psi}}$. Se puede resolver planteando el siguiente método iterativo. En primer lugar, se introduce la siguiente notación :

- $\mathbf{S}_i^* = \mathbf{S} - \hat{\mathbf{\Psi}}_i^2$ 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 inicia con $\hat{\mathbf{\Psi}}_0 = 0$ de tal manera que $\mathbf{S}_0^* = \mathbf{S}$ y consiste en cada iteración hacer las siguientes operaciones [11, 37, 58]:

1. Se calcula $\mathbf{S}_i^* = \mathbf{S} - \hat{\mathbf{\Psi}}_i^2$
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 $\hat{\mathbf{\Psi}}_{i+1}^2 = \mathbf{S} - \hat{\mathbf{\Lambda}}_i \hat{\mathbf{\Lambda}}_i^T$

El proceso se para cuando no haya cambios en las matrices $\hat{\mathbf{\Psi}}_i, \hat{\mathbf{\Psi}}_{i+1}$ o cuando $\hat{\mathbf{\Lambda}} \hat{\mathbf{\Lambda}}^T + \mathbf{\Psi}^2$ sea lo más parecido posible a la matriz de covarianzas \mathbf{S} .

Peña D [58] lo generaliza utilizando la función $F = \text{tr}(\mathbf{S} - \mathbf{\Lambda} \mathbf{\Lambda}^T - \mathbf{\Psi})^2$. 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 saturaciones de los factores, se pueden estimar las puntuaciones en los factores, que son los valores de los factores en cada una las observaciones. Es decir, las puntuaciones de las observaciones en el espacio de los factores comunes. (Para más detalles de esta parte véase la sección 9.5 de [37] o la 12.7 de [58])

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{Tf}^T , donde \mathbf{T} es una matriz ortogonal de tamaño $m \times m$, cumplen las hipótesis del modelo planteado. Entonces al transformar la matriz $\mathbf{\Lambda}$ a la matriz $\mathbf{\Lambda T}^T$ se cumple que [51]:

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

Sigue manteniendo las características y suposiciones hechas en el modelo 2.2.1.

Para evitar estas indeterminaciones se suelen tomar dos restricciones [51]:

$$\mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda} \text{ ó } \mathbf{\Lambda}^T \mathbf{\Lambda} \text{ son diagonales con los valores en orden decreciente} \quad (2.2.13)$$

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, $\hat{\mathbf{\Lambda}}^T \hat{\mathbf{\Lambda}} = \mathbf{D}^{\frac{1}{2}} \mathbf{U}^T \mathbf{U} \mathbf{D}^{\frac{1}{2}} = \mathbf{D}$.

Proposición 2.2.3. La segunda restricción hace que el modelo sea único [58].

Demostración. Sea la descomposición de la varianza (Teorema 2.2.1):

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Psi} \\ \mathbf{\Sigma} - \mathbf{\Psi} &= \mathbf{\Lambda} \mathbf{\Lambda}^T \end{aligned}$$

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 $\mathbf{\Sigma} - \mathbf{\Psi}^2$, 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. \square

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: Analizar qué parte de la variabilidad explica por cada uno de los factores, cuanto de la variabilidad total de los datos recogen los factores comunes etc. . . . Esto nos puede ayudar a la hora de determinar si el modelo obtenido es o no útil para nuestros objetivos.
- Especificidad de cada una de las variables: Una alta especificidad implica que poca de la información de la variable es explicada por los factores comunes, es decir, que estos no recogan la información.
- Saturaciones de la variable en los factores: Puede ocurrir que haya variables que tengan en el mismo factor menor saturación que el resto, esto es significativo ya que en este caso, la variabilidad de esa variable no es recogida por ese factor en específico.
- Significado de los factores: analizar una vez obtenidas las cargas, si estas representan una medida subyacente notable o no. Detallar el origen del significado que se le otorga en el caso de tenerlo.

Una vez interpretados los resultados , se puede llegar a los objetivos del análisis factorial que es estudiar la estructura latente de los datos en términos de variabilidades comunes.

2.3. Análisis de cluster

Según Jain y Dubes el análisis de *cluster* es un conjunto de técnicas exploratorias que toman un conjunto de datos resultantes de recoger N observaciones sobre p medidas diferentes y buscan agrupar las observaciones. También se puede utilizar agrupaciones de variables pero no se desarrollará.

Para empezar, debemos definir que es un *cluster*. Como en nuestro caso no buscamos agrupar variables, sino observaciones o mediciones utilizaremos la definición dada en [21].

Definición 2.3.1. Diremos que un *cluster* o conglomerado es un subconjunto de las observaciones que son similares entre sí.

Es fácil observar que los clusters definen la siguiente relación de equivalencia \mathcal{R} en la que dos observaciones, $\mathbf{x}_i, \mathbf{x}_{i'}, i, i' = 1, \dots, N$ están relacionadas si pertenecen al mismo clúster [11]. Por tanto, cada *cluster* genera una clase de equivalencia $[c_k], k = 1 \dots K$ donde K es el número de clusters

El conjunto de *clusters* generan un *clustering* o partición, es decir:

Definición 2.3.2. Se llama *clustering* a la partición que provoca la relación \mathcal{R} del espacio de observaciones.

Para llegar a dichas particiones podemos usar distintos tipos de algoritmos según el número de particiones distintas que se generen a lo largo del proceso [35]:

- Particional Son aquellos procesos de *clustering* en los que de manera previa se conocen el número de *clusters* y por tanto, se genera una partición que se va modificando.
- Jerárquico, si el proceso genera una secuencia anidada de particiones, es decir, que dependiendo de cómo se desarrolle puede ser de los siguientes tipos:
 - Aglomerativo Cuando se inicia con una partición que tiene N *clusters* con una observación en cada uno. Este tipo de algoritmos va tomando los *clusters* y junta aquellos según un criterio establecido hasta llegar a un criterio de parada.
 - Divisivo Cuando empieza con un único *cluster* que contiene a las N observaciones. En cada paso se va dividiendo los cluster siguiendo algún criterio hasta llegar al criterio de parada.

En resumen, el análisis de cluster es un conjunto de técnicas que permiten la simplificación estructural de las observaciones recogidas de un vector aleatorio mediante la agrupación de las mismas en conjuntos llamados clusters [21].

El análisis de clusters ha sido aplicado en infinidad de áreas como el marketing con el objetivo de segmentar los anuncios [56], la psicología para detectar si se da un sólo trastorno en una población [19] o en el caso de la medicina para intentar estudiar la supervivencia de pacientes con carcinoma renal [64]. Estos son ejemplos de unas pocas aplicaciones, pero también nos pueden ayudar en la predicción del riesgo crediticio, segmentación electoral etc...

2.3.1. Algoritmos jerárquicos

Los algoritmos jerárquicos calculan en cada paso la similaridad o disimilaridad entre *clusters* para generar nuevas particiones del espacio uniendo o separando los distintos *clusters*.

Un algoritmo jerárquico genera una secuencia de particiones anidada que está definida de la siguiente manera [65] :

Definición 2.3.3. Una partición del espacio $\Pi^{(r)}$ está anidada en otra partición $\Pi^{(k)}$ y se denota $\Pi^{(r)} \sqsubset \Pi^{(k)}$ si cumple lo siguiente:

- El número de clusters de $\Pi^{(k)}$ es menor que el de $\Pi^{(r)}$.
- Cada cluster de la partición $\Pi^{(r)}$ es un subconjunto de algún cluster de $\Pi^{(k)}$.

En el caso de un algoritmo aglomerativo tendremos que $k > r$, es decir en el paso k -ésimo el número de *clusters*.

La pregunta que hay que hacer antes de aplicar este tipo de algoritmos, es cómo calcular las distancias o similaridad entre dos observaciones. Hay que distinguir dos casos, el cálculo inicial de las distancias entre observaciones del espacio y la distancia entre *clusters*. Dependiendo de como se elijan sobre todo las últimas se tendrá un algoritmo de *clustering* u otro [58].

En particular los métodos jerárquicos se pueden aplicar a distancias y similaridades, (véase [51]) Estas se pueden definir de distintas formas pero la más habitual es la distancia de Mahalanobis como se detallaba en la definición 1.1.5. Aunque esta sólo se detalla para variables continuas.

Dependiendo de como se defina dicha distancia va a provocar que tengamos uno u otro tipo de algoritmo jerárquico.[21, 37, 58]

- El vecino más proximo: $d(C, C') = \min_{\mathbf{x}_i \in C, \mathbf{x}_{i'} \in C'} (\mathbf{x}_i, \mathbf{x}_{i'})$, es decir, se toma como la distancia entre los *clusters* y una observación, $\mathbf{x}_{i'}$, el mínimo de las distancias entre las observaciones del propio cluster y la observación considerada.
- El vecino más lejano: $d(C, C') = \max_{\mathbf{x}_i \in C, \mathbf{x}_{i'} \in C'} (\mathbf{x}_i, \mathbf{x}_{i'})$, es decir, se toma como la distancia máxima entre todas las posibles.
- Enlace medio

$$d(C, C') = \frac{1}{N_C} \frac{1}{N_{C'}} \sum_{\mathbf{x}_i \in C} \sum_{\mathbf{x}_{i'} \in C'} d(\mathbf{x}_i, \mathbf{x}_{i'}) \quad (2.3.1)$$

donde $N_C, N_{C'}$ son el número de observaciones que hay en cada cluster. Es decir, es la media de todas las distancias.

- Ward:

$$d(C, C') = \frac{1}{N_C} \frac{1}{N_{C'}} \sum_{\mathbf{x}_i \in C} \sum_{\mathbf{x}_{i'} \in C'} (d(\mathbf{x}_i, \mathbf{x}_{i'}))^2 \quad (2.3.2)$$

donde $N_C, N_{C'}$ son igual que antes. En particular, el algoritmo de Ward calcula las variación en los *clusters*

En general, un algoritmo jerárquico aglomerativo funciona de la siguiente manera:

- Se inicia con una partición que tiene N *clusters* con una única observación cada uno.
- Se calcula la matriz de distancias entre los *clusters*.
- Se unen los dos clusters que menor distancia tengan entre ellos. De esta manera, se reduce el número de clusters en 1.
- Se repiten los dos pasos anteriores hasta tener un único cluster.

El resultado de estos algoritmos es un dendograma [51]

Definición 2.3.4. Un dendograma es un diagrama de árbol en el que en el eje horizontal se sitúan las observaciones mientras que en el eje vertical se representan las distancias. Cada nodo representa una unión de los clusters.

Los siguientes dendogramas 2.3.1a 2.3.1b son resultado de aplicar un algoritmo jerárquico al conjunto *IRIS* de Fisher [22].

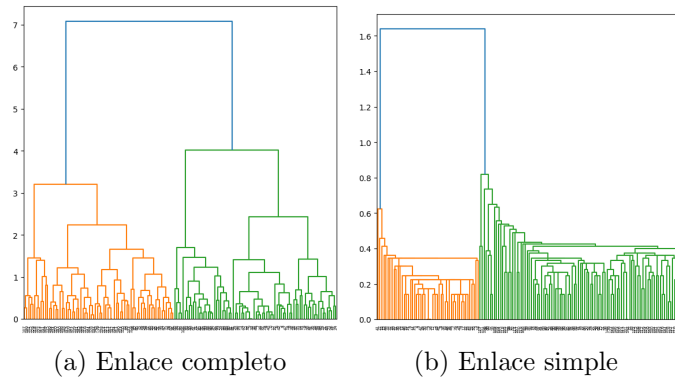


Figura 2.3.1: Diagramas obtenidos utilizando la biblioteca de Python Scikit-Learn

2.3.2. Algoritmos particionales

Dada una matriz de datos \mathbf{X} de tamaño $N \times p$ resultado de realizar N observaciones de p variables aleatorias. El *clustering* particional busca hacer una partición que divida las observaciones en K grupos homogéneos

Definición 2.3.5. Se llama *centroide* de un cluster C_k , $k = 1, \dots, K$ al vector de tamaño p cuyas componentes son las medias de cada una de las variables de todas las observaciones del *cluster* C_k :

$$\bar{x}_{jk} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij} \quad \forall j = 1, \dots, p \quad (2.3.3)$$

El algoritmo de K -medias es el más importante de este tipo [37]. Se desarrolla de la siguiente manera :

1. Se asigna a cada una de las observaciones en cada uno de los C *clusters*. Se calculan los centroides de dichos *clusters*.
2. Se comprueba la distancia euclidiana de cada una de las observaciones a los centroides de los *clusters* y se reasignan las observaciones de acuerdo al *cluster* más cercano. Tras esto se recalculan los centroides.
3. Se repiten los dos pasos anteriores hasta que las asignaciones no cambien.

Resaltar que este algoritmo tiene un problema y es que los resultados finales

Hay otras alternativas que utilizan criterios de variabilidad intra grupos o entre grupos para realizar las asignaciones óptimas fijados el número de clusters, para ello léase [21, 58, 29]

Capítulo 3

Aplicación sobre datos

3.1. Cemento, análisis estructural y capacidades predictivas.

Descripción del dataset

El dataset [70] es un conjunto de datos recogidos de distintos tipos de cementos con distintas densidades, en particular tenemos las siguientes variables:

Nombre	Tipo de dato	Medida	Descripción
Cemento (variable 1)	continuo	kg en una mezcla m ³	Variable predictora
Escoria de alto horno (variable 2)	continuo	kg en una mezcla m ³	Variable predictora
Ceniza volante (variable 3)	continuo	kg en una mezcla m ³	Variable predictora
Agua (variable 4)	continuo	kg en una mezcla m ³	Variable predictora
Superplastificante (variable 5)	continuo	kg en una mezcla m ³	Variable predictora
Agregado grueso (variable 6)	continuo	kg en una mezcla m ³	Variable predictora
Agregado fino (variable 7)	continuo	kg en una mezcla m ³	Variable predictora
Edad	continuo	Día (1-365)	Variable predictora
Resistencia a la compresión	continuo	MPa	Variable respuesta

Cuadro 3.1: Tabla resumen de las variables estudiadas

Se recogen de manera esquemática en el cuadro 3.1.

El tamaño de la muestra es de 1030 observaciones, de las cuales un 70 % se separan para realizar el ajuste de los distintos modelos y el 30 % restante se utilizará para calcular la capacidad predictiva de los modelos.

Objetivos

Este estudio tiene dos objetivos principales:

- Detallar y analizar patrones de comportamiento en los cementos en base a las características descritas.
- Crear y analizar distintos modelos que permitan la predicción de la resistencia a la compresión.

Metodología

Para conseguir el primer objetivo se utilizará el análisis de componentes principales. Con ello, se podrá estudiar la estructura de los datos de manera que se vea que variables afectan en mayor medida a la varianza.

Para el otro objetivo se ajustarán tres modelos distintos. Un modelo de regresión lineal, un árbol de regresión y una red neuronal. Se describirán las cualidades que aporta cada uno de los modelos y se compararán. Es por ello que se separará el conjunto de datos en dos partes una de ajuste y otra de comprobación de las capacidades predictivas.

En particular, se utilizará el error cuadrático medio, MSE , en cada uno de los modelos, de manera que en el conjunto de entrenamiento servirán como medida de bondad de ajuste y en el conjunto de comprobación para probar sus capacidades predictivas.

A nivel técnico, para esta parte se utilizará el lenguaje de programación Python. En particular, se usan las siguientes librerías:

- *Pandas*: para el tratamiento previo de los datos.
- *Scikit-Learn*: usado en el ajuste de los distintos modelos [66].
- *Numpy*: utilizado en el manejo de las matrices y los vectores
- *Matplotlib* empleado para la generación de los distintos gráficos <https://matplotlib.org/>.

Desarrollo y resultados

Para empezar, se hace un análisis descriptivo básico de cada una de las variables:

	Cement (component 1) (kg in a m ³ mixture)	Blast Furnace Slag (component 2) (kg in a m ³ mixture)	Fly Ash (component 3) (kg in a m ³ mixture)	Water (component 4) (kg in a m ³ mixture)	Superplasticizer (component 5)(kg in a m ³ mixture)	Coarse Aggregate (component 6) (kg in a m ³ mixture)	Fine Aggregate (component 7) (kg in a m ³ mixture)	Age (day)	Concrete compressive strength(MPa. megapascals)
count	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000
mean	281.167864	73.895825	54.188350	181.567282	6.204660	972.918932	773.580485	45.662136	35.817961
std	104.506364	86.279342	63.997004	21.354219	5.973841	77.753954	80.175980	63.169912	16.705742
min	102.000000	0.000000	0.000000	121.800000	0.000000	801.000000	594.000000	1.000000	2.330000
25%	192.375000	0.000000	0.000000	164.900000	0.000000	932.000000	730.950000	7.000000	23.710000
50%	272.900000	22.000000	0.000000	185.000000	6.400000	968.000000	779.500000	28.000000	34.445000
75%	350.000000	142.950000	118.300000	192.000000	10.200000	1029.400000	824.000000	56.000000	46.135000
max	540.000000	359.400000	200.100000	247.000000	32.200000	1145.000000	992.600000	365.000000	82.600000

Figura 3.1.1: Análisis descriptivo de los datos

En este primer análisis resumido en la figura 3.1.1 podemos ver que la media de la fuerza de compresión o resistencia a la compresión tiene una media muestral de 35,82 MPa, con una desviación estándar de 16,70 MPa lo cual nos dice que hay una alta dispersión de los datos en este caso. Hay que tener esto en cuenta ya que puede afectar a los errores de los distintos modelos.

Hay otras medidas a las que le ocurre lo mismo, es decir, que tienen una alta dispersión como pueden ser las cantidad de superplastificante, la ceniza volante o la escoria de horno.

También hay que destacar que al menos un 50 % de las observaciones no tiene ceniza volante.

Ahora se calculan las componentes principales utilizando la matriz estandarizada. Después de aplicar este proceso, obtenemos la proporción acumulada de la varianza explicada en cada una de las componentes:

$$[0,25, 0,47, 0,63, 0,74, 0,85, 0,95, 0,98, 1., 1.]$$

En particular, cuando tomamos las 3 primeras componentes principales estas representan un 62,5 % de la variabilidad total, en particular esas componentes representan una variabilidad del 25 %, 22 %, 16 % respectivamente. Podríamos llegar a un 85 % si cogiésemos las 5 primeras. En nuestro caso, nos quedaremos con 3 para poder representar los datos en un espacio de dimensión reducida.

Los coeficientes de las componentes principales obtenidos son los siguientes:

$$Z_1 \quad (3.1.1)$$

$$Z_2 \quad (3.1.2)$$

$$Z_3 \quad (3.1.3)$$

Al haber trabajado con la matriz estandarizada, estas cargas también representan la correlación de cada una de las variables con la componente en cuestión.

Aquellas variables que tienen una carga en las componentes principales mayor que 0,5 en valor absoluto tienen una relación importante con las componentes. En cambio, todas aquellas cercanas al 0 no tienen una relación menor con la componente.

Se ajustan 3 modelos distintos, un modelo lineal, un árbol de regresión, y una red neuronal.

Las características de los modelos y de los procesos de ajuste son los siguientes:

El modelo lineal es ajustado obteniéndose un mínimo de error cuadrático de 106,14 en el conjunto de entrenamiento. Mientras que en el conjunto de comprobación se obtiene un error cuadrático medio de 108,21. Además aunque no se pueda usar en el resto de modelos se obtiene en este caso un coeficiente de determinación R^2 de 0,62 en el conjunto de ajuste. Por tanto, no tiene un buen nivel predictivo.

El árbol de decisión utiliza el criterio *CART* de Breiman, ya que es el implementado por defecto en la librería *Scikit-Learn* de Python [5]. Este modelo proporciona un error cuadrático medio de 108,22 sobre el conjunto de ajuste mientras que sobre el conjunto de testeo proporciona un error cuadrático de 57,55, lo que implica que tiene una mejor capacidad predictiva que el modelo lineal pero aún así deja que desear.

En la red neuronal se ha elegido un modelo simple con dos capas, la primera con 13 neuronas con una función de activación lineal rectificada, en cambio, la otra capa contiene únicamente una neurona con una función lineal. En este caso el ajuste de

la red neuronal obtiene un error cuadrático medio de $MSE = 46,56$ en el ajuste y un $MSE = 47,54$ en el conjunto de prueba, en principio, tenemos un modelo que comete menos error en el conjunto de prueba.

Hay que destacar que ninguno de los modelos ha incurrido en problemas de sobreajuste, es decir, no se ha tenido casos en los que el valor del error medio en el conjunto de ajuste sea mucho menor que en el conjunto de prueba.

Una vez ajustados los modelos, se pueden tomar los valores predichos para los datos de prueba y graficarlos frente a los valores reales, en cada caso tendremos los puntos en azul y la recta roja representaría los puntos los cuales su predicción es igual al valor real.

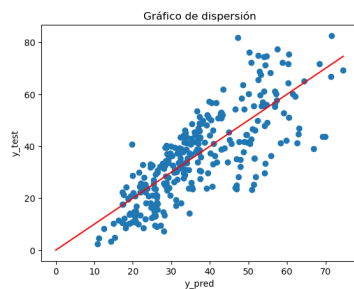


Figura 3.1.2: Regresión Lineal

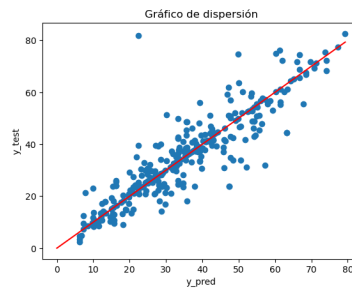


Figura 3.1.3: Árbol de Regresión

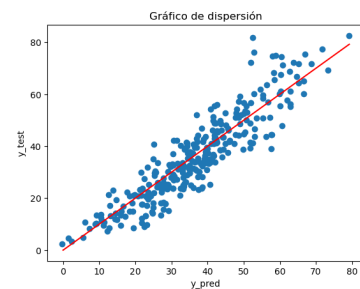


Figura 3.1.4: Red Neuronal

Se puede ver que en el caso del modelo lineal 3.1.2 los puntos son más lejanos de la recta. Por tanto, tenemos que el primer caso en principio tendrá peores predicciones.

En resumen, obtenemos los siguientes resultados:

	Modelo Lineal	Árbol de Regresión	Red Neuronal
Conjunto Ajuste	106,14	108,22	46,56
Conjunto de Prueba	108,21	57,55	47,54

Cuadro 3.2: Errores cuadráticos medios en cada conjunto

En conclusión, los modelos en este caso no han conseguido el objetivo principal que era diseñar un modelo con buenas capacidades predictivas. Concretamente, ha sido el modelo de la red neuronal el que mejores resultados ha obtenido, pero su alta complejidad lastra su interpretabilidad. En cambio, el árbol de regresión obtiene en el proceso de prueba resultados similares y además con una fácil interpretación debido a que se puede representar como un diagrama de árbol.

Capítulo 4

Conclusiones

Los métodos multivariantes son un tema amplio e interesante con miles de aplicaciones posibles.

Bibliografía

- [1] **Abdi, H. (2007).** The method of least squares. *Encyclopedia of measurement and statistics*, 1, 530-532.
- [2] **Abdi, H., and Williams, L. J. (2010).** Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(4), 433–459.
- [3] **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.
- [4] **Biau, G., and Scornet, E. (2016).** *A random forest guided tour*. *Test*, 25, 197-227.
- [5] **Breiman, L. (1984).** *Classification and regression trees*. Chapman & Hall.
- [6] **Breiman, L. (1996).** Bagging predictors. *Machine learning*, 24, 123-140.
- [7] **Breiman, L. (2001).** Random forests. *Machine learning*, 45, 5-32.
- [8] **Breiman, L. (2004).** Consistency for a simple model of random forests. *University of California at Berkeley. Technical Report*, 670.
- [9] **Brown, S. D., Myles, A. J., Feudale, R. N., Liu, Y., and Woody, N. A., (2004).** An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6), 275-285.
- [10] **Chatfield, C. and Collins A.J (1989).** *Introduction to multivariate analysis*, Chapman and Hall.
- [11] **Cuadras, C.M. (2014),** *Nuevos métodos de Análisis Multivariante*, CMC Editions, Barcelona.
- [12] **Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006).** Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7, 1-13.
- [13] **Diez, J. P., Hernandis, S. 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.

- [14] **Divakaran, S. (2022).** Data Science: Principles and Concepts in Modeling Decision Trees. *Data Science in Agriculture and Natural Resource Management*, 55-74.
- [15] **Dry Bean Dataset. (2020).** *UCI Machine Learning Repository*. <https://doi.org/10.24432/C50S4B>.
- [16] **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.
- [17] **Eckart, C., & Young, G. (1936).** The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211-218.
- [18] **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).
- [19] **Everitt, B., Ismail, K., David, A. S., & Wessely, S. (2002).** Searching for a Gulf War syndrome using cluster analysis. *Psychological medicine* 32(8), 1371-1378.
- [20] **Everitt, B., and Hothorn, T. (2011).** *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.
- [21] **Everitt, B. S., Leese, M., Stahl, D., & Landau, S. (2011).** *Cluster analysis*. John Wiley & Sons.
- [22] **Fisher, R. A.. (1988).** Iris. *UCI Machine Learning Repository*. <https://doi.org/10.24432/C56C76>.
- [23] **Galton, F. (1889).** Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279), 135-145.
- [24] **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.
- [25] **Gottumukkal, R., & Asari, V. K. (2004).** An improved face recognition technique based on modular PCA approach. *Pattern Recognition Letters*, 25(4), 429-436.
- [26] **Greene, W.H. (2008).** *Econometric analysis (6th ed.)* Pearson Prentice Hall.
- [27] **Grossi, E., & Buscema, M. (2007).** Introduction to artificial neural networks. *European journal of gastroenterology & hepatology*, 19(12), 1046-1054.
- [28] **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 Jersey.

- [29] **Hartigan, J. A. (1975).** *Clustering algorithms*. John Wiley & Sons, Inc.
- [30] **Hastie, T., Tibshirani, R. and Friedman J. (2001),** *The Elements of Statistical Learning, Data Mining, Inference and Prediction* Springer.
- [31] **Hesterberg, T. (2011).** Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6), 497-526.
- [32] **Hochreiter, S., & Schmidhuber, J. (1997).** Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- [33] **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](https://doi.org/10.1016/0893-6080(89)90020-8)
- [34] **Hotelling H. (1933)** Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 25:417-441.
- [35] **Jain, A.K., Richard, C.D., (1988)** *Algorithms for clustering data*. Prentice-Hall, Inc.
- [36] **James G, Witten D, Hastie T, Tibshirani R.** *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer; 2021.
- [37] **Johnson, R. A., & Wichern, D. W. (2007).** *Applied multivariate statistical analysis.*(6th edition) Pearson Prentice Hall.
- [38] **Johnson, R. M. (1963).** On a theorem stated by Eckart and Young. *Psychometrika*, 28(3), 259-263.
- [39] **Jolliffe I.T.(1986).** *Principal Component Analysis*, Springer-Verlag.
- [40] **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.
- [41] **Köppen, M. (2000)** The curse of dimensionality. *5th online world conference on soft computing in industrial applications (WSC5)* (Vol. 1, pp. 4-8).
- [42] **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.
- [43] **Lawless, J. F., & Yuan, Y. (2010).** Estimation of prediction error for survival models. *Statistics in medicine*, 29(2), 262-274.
- [44] **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.
- [45] **Lebart, L., Morineau, A., Warwick, K. M. (1984).** *Multivariate Descriptive Analysis: Correspondence analysis and related techniques for large matrices*. Wiley.

- [46] **Loh, W. Y. (2011).** Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1), 14-23.
- [47] **Loh, W. Y. (2014).** Fifty years of classification and regression trees. *International Statistical Review*, 82(3), 329-348.
- [48] **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.
- [49] **Mahesh, B. (2020).** Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, 9, 381-386.
- [50] **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.
- [51] **Mardia, K.V., Kent, J.T. and Bibby J.M.; (1979).** *Multivariate analysis*. Academic Press.
- [52] **Martínez Arias R.** *El análisis multivariante en la investigación científica*. Madrid: La Muralla; 1999.
- [53] **Morrison. D.(1976).** *Multivariate statistical methods (2nd ed)*. McGraw-Hill Kogakusha.
- [54] **Nerini, D., & Ghattas, B. (2007).** Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis* 51(10), 4984-4993.
- [55] Neural Designer Blog <https://www.neuraldesigner.com/>
- [56] **Okazaki, S. (2006).** What do we know about mobile Internet adopters? A cluster analysis. *Information & Management*, 43(2), 127-141.
- [57] **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>
- [58] **Peña D.** *Análisis de datos multivariantes*. Madrid: McGraw-Hill; 2002.
- [59] **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.
- [60] **Quinlan, J. R. (2014).** *C4. 5: programs for machine learning*. Elsevier.
- [61] **Rencher, A. C. (2002).** *Methods of multivariate analysis*. New York: A John Wiley & Sons. Inc. Publication.

- [62] **A. Requena, R. Quintanilla, J.M. Bolarín, A. Vázquez, A. Bastida, J. Zúñiga y L.M. Tomás.** Nuevas Tecnologías y Contaminación de Atmósferas, para PYMEs. VI-3-4, Pag. 3 *Universidad de Murcia. Spain*. <https://www.um.es/LEQ/Atmosferas/Ch-VI-3/F63s4p3.htm>.
- [63] **Ringnér, M. (2008).** What is principal component analysis?. *Nature biotechnology*, 26(3), 303-304.
- [64] **Witten, D. M., & Tibshirani, R. (2010).** Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1), 29-51.
- [65] **Scitovski, R., Sabo, K., Martínez-Álvarez, F., & Ungar, Š. (2021).** *Cluster analysis and Applications* Springer.
- [66] *Scikit Learn Documentation* <https://scikit-learn.org/stable>
- [67] **Song, Y. Y., and Ying, L. U. (2015).** Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130.
- [68] **Spearman, C. (1904).** "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201-292. <https://doi.org/10.2307/1412107>
- [69] **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.
- [70] **Yeh, I-Cheng. (2007).** Concrete Compressive Strength. *UCI Machine Learning Repository*. <https://doi.org/10.24432/C5PK67>.