

Análisis de Datos para la Ingeniería Informática^a

Jorge Mateu

Departamento de Matemáticas, Universitat Jaume I

1 Introducción

La asignatura de *Análisis de Datos* es optativa en el plan de estudios de Ingeniería Informática de la Universitat Jaume I. Por su potencial interés, también se oferta como asignatura de libre configuración. Se impartirá durante el primer semestre y se recomienda a los alumnos que la cursen en cuarto.

Los 5 créditos de esta asignatura se reparten por igual en 2.5 créditos teóricos y otros 2.5 créditos de laboratorio.

Respecto al número de matriculados en la asignatura, el hecho de que sea una asignatura optativa de segundo ciclo hace que no tenga una matrícula demasiado elevada. La media del número de estudiantes durante estos años es de alrededor de 35.

Por tratarse de una asignatura de último curso, es claramente de especialización y debe servir como aplicación de unos métodos de análisis estadísticos avanzados proporcionando al alumno una formación especializada en un campo muy concreto de trabajo. Las técnicas de análisis de datos suponen una disciplina en continua expansión siendo el sustento científico de otras muchas (por sus enormes aplicaciones potenciales, ha atraído la atención de muchos profesionales de otros campos). De hecho, no todos los alumnos proceden de la titulación de Ingeniería Informática, sino que por esta asignatura han pasado alumnos procedentes de las licenciaturas en Químicas, Administración y Dirección de Empresas, etc. En cualquier caso la mayoría de los alumnos son futuros ingenieros informáticos.

Por todo ello, esta asignatura se orientará más a la aplicación, sin descuidar el aspecto científico. Hay que tener en cuenta que los estudiantes que acceden a esta asignatura, por lo general, tienen una buena base teórica, porque han cursado (y obligatoriamente aprobado) 50 créditos de matemáticas.

Para una asignatura que recoge algunas técnicas de análisis de datos, hay que hacer un esfuerzo de ponderación entre varios factores: interés y aplicabilidad de la técnica, dificultad matemática subyacente (teniendo en cuenta a quien va dirigida), adaptación al poco tiempo existente para su explicación y su puesta en práctica con algún software. En este contexto, hemos optado por elegir una colección de métodos estadísticos muy útiles en la práctica diaria, de fácil implementación con un software y sin demasiados tecnicismos.

2 Las técnicas del análisis de datos en la ciencia y la ingeniería

Las técnicas de análisis estadísticos que actualmente se utilizan en investigaciones de carácter empírico han avanzado de forma asombrosa en los últimos años, gracias a la utilización de los ordenadores. El concepto "palingenesia" (literalmente 'volver a nacer') fue acuñado por Arnold Toynbee para definir un conjunto de cambios sociales y de personalidad que se producen en determinados momentos históricos. Este término se ha aplicado a la situación

de los últimos tiempos con motivo de la generalización de la informática. El concepto más reciente de la palingenesia informática probablemente sea el de la necesaria alfabetización informática ("computer literacy") o más concretamente de la estadística informática (análisis estadísticos por ordenador).

Con estos preliminares, no es de extrañar que en muchas ramas científicas, y en particular en las ingenierías, los procedimientos y modelos estadísticos se hayan convertido en una necesidad evidente. Valga como ejemplo la rama de la informática médica, una especialidad de la ingeniería informática que día a día demanda este tipo de ingenieros con plenos conocimientos estadísticos. Otro ejemplo sería la necesidad de las técnicas de análisis de datos en ciencias experimentales, en todas sus ramificaciones.

Es claro, pues, que un profesional capacitado y adaptado a las necesidades de hoy en día necesita de un conocimiento, por lo menos básico, de una colección de las técnicas de estadística más usuales. Estas técnicas tienen como elemento de trabajo los fenómenos no deterministas (aleatorios), fenómenos que forman parte de los datos con los que se tiene que enfrentar un ingeniero.

La modelización de estos fenómenos, o al menos el intentar extraer la mayor cantidad de información posible a nuestros datos se lleva a cabo mediante la utilización adecuada de unos métodos estadísticos como los que mostraremos en esta asignatura. Métodos que van desde el análisis de dependencias lineales entre variables, análisis de diferencias en medias de una variable medida bajo diferentes circunstancias, técnicas de búsqueda de grupos de individuos y/o variables, métodos de reducción de la dimensionalidad o análisis de la componente temporal de nuestros datos.

3 Objetivos de la asignatura

Al formular los objetivos para la asignatura de Análisis de Datos se ha de tener en cuenta que, a diferencia de la asignatura Estadística, se trata de una asignatura específica.

Las asignaturas específicas en una carrera tecnológica tienen como objetivo general orientar al alumno hacia el campo específico de la materia tecnológica de que se trate y poner el mecanismo optativo de aprendizaje con el fin de facilitar la formación adicional de los alumnos con un mayor interés. También se plantean dotar al alumno de suficientes conocimientos prácticos sobre la materia específica, y proporcionar un contacto del alumno con el entorno profesional correspondiente.

En concreto, un curso sobre análisis de datos estadísticos debería cumplir los siguientes objetivos:

- Dar una panorámica genérica de los métodos de tratamiento estadístico de datos desde los niveles más bajos (tratamiento descriptivo y modelos simples) hasta niveles medios altos (modelos múltiples y multivariantes). En cualquier caso, se pretende que los temas tratados durante el curso sean lo más "estables" posible y dejar los aspectos más novedosos y en desarrollo para los cursos de doctorado.
- Hacer que el alumno conozca las principales técnicas estadísticas concretando posibles restricciones de utilización de cada método y generalizaciones de los mismos.
- Mostrar las muchas aplicaciones del Análisis de Datos y hacer que el alumno sea capaz de identificar un problema como abordable por estos medios, así como escoger el hardware y software necesario para su resolución.
- El alumno debe de familiarizarse con uno o varios softwares suficientemente potentes (pero al mismo tiempo relativamente sencillos de usar) para abordar los problemas estudiados en clase, y en general problemas reales.

- Ya que el nombre de esta asignatura incluye el de datos, se pretende que el alumno sepa buscar bases de datos reales susceptibles de ser analizadas con las técnicas aprendidas. En este sentido (ver capítulo 1) se le darán a conocer varias direcciones Web para navegar entre páginas y páginas de datos reales.

desarrollo de las prácticas será más dinámico, en el sentido de que el tiempo dedicado a cada tema dependerá de la dificultad que muestren los alumnos en su comprensión.

4 Relación con otras asignaturas

La asignatura de *Análisis de Datos* tiene restricción de matrícula con respecto a las asignaturas de Estadística (curso 2º) e Investigación Operativa (curso 2º). Por otra parte existen relaciones de dependencia temática con otras asignaturas:

Asignaturas previas

- Cálculo (curso 1º) y Ampliación de Matemáticas (curso 2º). Aunque estas asignaturas son requisito inexcusable de muchas otras, aquí incidimos en que estas dos asignaturas proporcionarán las bases matemáticas de los modelos que estudiemos en Análisis de Datos.
- Estadística (curso 2º). Ya comentamos anteriormente la importancia de la Estadística en la modelización concreta de nuestros datos.
- Robótica (curso 3º). Inteligencia Artificial e Ingeniería del Conocimiento (curso 4º). Algunas de las técnicas que se explican durante estos cursos se basan en métodos de análisis discriminante y cluster.

5 Prácticas

En esta asignatura las prácticas de laboratorio ocupan 2.5 créditos del total de 5, es decir, son tan importantes como las clases de teoría. Y así tiene que ser por el enfoque eminentemente práctico que se le da a esta signatura (teniendo en cuenta a quién va dirigida).

Como ya ocurría en la asignatura Estadística, tampoco es el objetivo de estas prácticas el incrementar la habilidad del alumno en un determinado lenguaje de programación. Por el contrario, se considera interesante que el alumno desarrolle cierta soltura en el manejo de paquetes informáticos específicos (en nuestro caso el StatGraphics). Puesto que, probablemente, en un hipotético futuro trabajo la situación que se encontrará estará más próxima a desarrollar aplicaciones a partir de un soporte previamente elaborado que a programarlos desde la base. Además, desde el punto de vista de desarrollo de las prácticas, la opción de la programación de los métodos podría ser muy costosa temporalmente.

Como tanto en la asignatura de *Estadística* en la Ingeniería Técnica en Informática de Gestión, como en la misma asignatura en la Ingeniería Informática, se ha utilizado el paquete estadístico StatGraphics, por razones obvias de continuidad también se ha escogido este software para la realización de las prácticas en la asignatura Análisis de Datos.

Cierto es que en general el alumno que llega a esta asignatura conoce el StatGraphics por las prácticas realizadas en las asignaturas anteriores, pero lo ha utilizado únicamente en el contexto de dichas asignaturas, es decir, desconoce las herramientas de este software para llevar a cabo las técnicas nuevas que va a estudiar. Es por ello, que en esta nueva asignatura se les va a pasar un manual de prácticas (ver el capítulo correspondiente en esta memoria) que recoge todas las nuevas instrucciones necesarias.

La idea general es plantear prácticas que cubran todos y cada uno de los temas que se desarrollan en teoría ajustándose lo máximo a posible al temario. Nuestra experiencia es que hay que dedicar al menos 2 sesiones de 2.5 horas por tema de teoría. Sin embargo, el

6 Programa de la Asignatura

De acuerdo con todas las consideraciones anteriores proponemos la siguiente relación de contenidos para la asignatura de Análisis de Datos:

Capítulo I: INTRODUCCIÓN GENERAL DE LA ASIGNATURA

- Lección 1. Introducción.
- Lección 2. Conceptos básicos en inferencia estadística univariante.
- Lección 3. Conceptos básicos en análisis multivariante. Distribución Normal multivariante.
- Lección 4. Direcciones Web donde encontrar bancos de datos estadísticos.

Capítulo II: ANALISIS DE LA VARIANZA

- Lección 1. Introducción.
- Lección 2. Análisis de la varianza de una vía.
- Lección 3. Análisis de la varianza en un diseño de bloques aleatorizados.
- Lección 4. Análisis de la varianza de dos vías.

Capítulo III: ANALISIS DE REGRESION

- Lección 1. Introducción.
- Lección 2. Modelo de regresión lineal simple.
- Lección 3. Modelo general de regresión.

Capítulo IV: ANALISIS DISCRIMINANTE

- Lección 1. Introducción.
- Lección 2. Notación y estructura de los datos.
- Lección 3. Método basado en la distancia de Mahalanobis.
- Lección 4. Método de variedades canónicas.

Capítulo V: ANALISIS CLUSTER

- Lección 1. Introducción.
- Lección 2. Tipos de datos para el análisis cluster.
- Lección 3. Formulación geométrica: medidas de similitud.
- Lección 4. Cluster jerárquico.
- Lección 5. Cluster no jerárquico.

Capítulo VI: ANALISIS DE COMPONENTES PRINCIPALES

- Lección 1. Introducción.
- Lección 2. Propiedades y significado geométrico.
- Lección 3. Método de obtención de las componentes.

Capítulo VII: ANALISIS FACTORIAL

- Lección 1. Introducción.
- Lección 2. El modelo de análisis factorial.
- Lección 3. Métodos de extracción de factores.
- Lección 4. Contrastes en el modelo factorial.
- Lección 5. Rotaciones en el análisis factorial.

Capítulo VIII: SERIES TEMPORALES

- Lección 1. Introducción.
- Lección 2. Series temporales y procesos estocásticos.
- Lección 3. Procesos ARIMA.
- Lección 4. Ajuste del modelo y predicciones.

7 Capítulo I: Introducción general a la asignatura

7.1 Lección 1: Introducción

La investigación aplicada frecuentemente se encuentra con fenómenos complejos que requieren para su análisis de una considerable cantidad de variables. En este curso se pretende dar una serie de técnicas estadísticas de análisis multivariante desde un enfoque práctico que nos permitan abordar el problema de tal análisis. Podemos considerar el análisis multivariable o multivariante como el conjunto de técnicas estadísticas que analizan simultáneamente más de dos variables en una muestra de observaciones.

Los métodos que estudiaremos serán los siguientes:

Análisis de la Varianza Conjunto de técnicas estadísticas que permiten estudiar la influencia de una o más variables categóricas sobre una cuantitativa.

Análisis de Regresión Conjunto de técnicas estadísticas que permiten estudiar la influencia de una o más variables cuantitativas sobre otra cuantitativa.

Análisis Discriminante Es una técnica de clasificación y asignación de un individuo a un grupo, conocidas sus características. En el análisis discriminante se dispone de una serie de grupos definidos 'a priori', con una serie de observaciones para cada individuo referidas a un conjunto de variables relevantes. En base a esta información se llega a calcular una función discriminante que se puede utilizar para hacer predicciones futuras.

Análisis Cluster Consiste en la construcción de una clasificación jerárquica de los individuos mediante la obtención de sucesivas particiones ('clusterings'), organizadas en diferentes niveles jerárquicos, estando cada partición formada por clases disjuntas ('clusters').

Análisis de Componentes Principales Este método tiene por objeto transformar un conjunto de variables, a las que denominaremos *originales*, en un nuevo conjunto de variables denominadas *componentes principales*. Estas últimas se caracterizan por estar incorreladas entre sí. Se busca explicar la mayor parte de la variabilidad total de un conjunto de variables con el menor número de componentes posibles.

Análisis Factorial Al igual que la técnica anterior examina las interdependencias entre las variables pero en este caso el objetivo es seleccionar factores para explicar las relaciones entre las variables.

Series Temporales Conjunto de métodos para modelizar las dependencias temporales intra y entre variables. Las observaciones están medidas en el tiempo y son dependientes unas de las otras. Averiguar el grado de dependencia es uno de los objetivos de estas técnicas.

Pero para tratar todos estos temas necesitamos 'refrescar' algunos conceptos básicos de estadística univariante e introducir otros de estadística multivariante.

7.2 Lección 2: Conceptos básicos en inferencia estadística univariante

En esta lección se recuerda al alumno los conceptos de estadística univariante que aprendió en la asignatura de Estadística, sea en la Ingeniería Técnica de Gestión o en la Ingeniería Informática.

En primer lugar se recuerda el concepto de *variables aleatorias y su distribución*. Una variable aleatoria es una función que asigna un número real a cada resultado del espacio muestral de un experimento aleatorio.

Podemos distinguir varios tipos de variables según su estructura:

- Llamaremos variable aleatoria *discreta* a aquella cuyo soporte es un número finito de números reales o bien un número infinito contable (i.e. puede tomarse como una sucesión de valores distintos) y variable *continua* a aquella variable cuyo soporte es un conjunto infinito no numerable (generalmente un intervalo).
- A nivel de aplicaciones es más interesante distinguir entre variables *medibles* que expresan una magnitud numérica y las llamadas *categóricas* que expresan categorías (cualitativas o discretas con rango pequeño).

La distribución de probabilidad de una variable aleatoria viene caracterizada por una función $f(x)$ que en el caso discreto recibe el nombre de *función de probabilidad* y en el continuo de *función de densidad* de probabilidad.

- Si X es discreta, $P(X \in A) = \sum_{x \in A} f(x)$ (el rango de valores es numerable y podemos 'sumar').
- Si X es continua $P(X \in A) = \int_A f(x) dx$.

A continuación describimos algunas medidas que describen características de la distribución de probabilidad de una variable aleatoria. Distinguiremos dos tipos de medidas importantes, las de tendencia central y las de dispersión.

La medida de centralización más utilizada es la *media o esperanza matemática*. En el caso discreto, $\mu = E(X) = \sum x f(x)$, y en el caso continuo, $\mu = E(X) = \int x f(x) dx$.

La medida de dispersión asociada a la media es la desviación típica o su cuadrado, *la varianza*. Para el caso discreto, $\sigma^2 = \text{var}(X) = \sum (x - \mu)^2 f(x)$. En el continuo, $\sigma^2 = \text{var}(X) = \int (x - \mu)^2 f(x) dx$.

Otras medidas interesantes sobre todo en la parte de la inferencia estadística son los *percentiles*. El *percentil* de orden p , donde $0 \leq p \leq 1$, es aquel valor x_p tal que $P(X \leq x_p) = p$. Es decir, la probabilidad de que la variable tome valores menores o iguales que él es p .

Posteriormente describimos algunos modelos de probabilidad interesantes.

Un *modelo o familia de distribuciones de probabilidad* es un conjunto de variables aleatorias que tienen sus funciones de densidad o de probabilidad con la misma estructura funcional matemática. Esta estructura matemática suele depender de uno o más parámetros, que según los valores que tomen darán lugar a una u otra distribución concreta. Repasaremos los modelos de distribución de probabilidad discretos: *Bernoulli*, *Binomial*, *Hipergeométrica* y *Poisson*. En el caso continuo estudiamos las distribuciones *Uniforme*, *Exponencial* y *Normal*.

La distribución Normal es una de las distribuciones de probabilidad más importantes. Debe su importancia principalmente a:

- Un gran número de fenómenos reales se pueden modelizar mediante esta distribución.
- Muchas distribuciones de uso frecuente tienden a aproximarse a la distribución normal bajo ciertas condiciones.
- Tiene propiedades matemáticas muy útiles.

La distribución normal es una distribución de variable continua que queda especificada por dos parámetros de los que depende su función de densidad y que resultan ser la media y la varianza. Llamamos distribución *normal tipificada* a la distribución normal con media cero y desviación típica 1. La función de distribución de la distribución normal tipificada está tabulada (la denotaremos por $\Phi(x)$) y se utiliza para calcular las probabilidades en el caso general. Si $X \sim N(\mu, \sigma^2)$, $P(X \leq x) = \Phi(\frac{x-\mu}{\sigma})$.

Existen otras tres distribuciones relacionadas con la distribución normal:

- Distribución Ji-cuadrado. $X \sim \chi_n^2$ si $X = Z_1^2 + \dots + Z_n^2$ con $Z_i \sim N(0,1)$.
- Distribución t de Student. $T \sim t_n$ si $T = \frac{Z}{\sqrt{X/n}}$ con $Z \sim N(0,1)$ y $X \sim \chi_n^2$.
- Distribución F de Snedecor. $F \sim F_{n,m}$ si $F = \frac{X/n}{Y/m}$ con $X \sim \chi_n^2$ y $Y \sim \chi_m^2$.

Sus percentiles se encuentran recogidos en tablas.

Acabaremos esta introducción con los conceptos básicos de *inferencia estadística y contraste de hipótesis*.

Una *muestra aleatoria de tamaño n* de X es un conjunto de variables aleatorias X_1, \dots, X_n con la misma distribución de probabilidad, la que describe el comportamiento de X , e independientes entre sí.

La inferencia estadística es la parte de la estadística que se encarga de extraer conclusiones sobre alguna característica desconocida de la variable de interés a partir de la muestra.

Llamaremos *parámetro* a cualquier característica desconocida de la variable de interés, que denotaremos en general por θ , sobre la cuál queremos hacer alguna inferencia.

Llamaremos *estadístico* a cualquier función $T(X_1, \dots, X_n)$ de la muestra. Algunos ejemplos de estadísticos son: a) Media muestral, $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$; b) Varianza muestral, $s_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}$

Un estadístico es una variable aleatoria con una distribución de probabilidad asociada que estará directamente asociada con la distribución de la variable (ya que es función de ella). Llamaremos *distribución en el muestreo* a la distribución de probabilidad del estadístico. Es importante conocerla porque nos basaremos en ella para dar una fiabilidad a las inferencias que hagamos sobre el parámetro.

La inferencia estadística aborda dos tipos de problemas:

Estimación Determinar el valor desconocido de una característica de la población.

- *Estimación puntual*: basándose en la muestra observada se da como aproximación del parámetro un valor concreto. Para ver si ese valor es adecuado o no, se analizan sus propiedades (insegadez, consistencia...)
- *Estimación por intervalos*: basándose en la muestra se da un intervalo que contenga al parámetro con una cierta 'confianza'.

Contraste de hipótesis Decidir entre dos hipótesis sobre las características de la distribución de la variable.

Cada uno de estos conceptos serán brevemente recordados, insistiendo en que volverán a salir en el desarrollo de la asignatura.

7.3 Lección 3: Conceptos básicos en análisis multivariante. Distribución Normal multivariante

En muchos experimentos es necesario considerar las propiedades de dos o más variables aleatorias simultáneamente. Un vector aleatorio n-dimensional es una aplicación:

$$\begin{aligned} X: \Omega &\rightarrow \mathbb{R}^p \\ \omega &\mapsto X(\omega)' = (X_1(\omega), \dots, X_p(\omega)) \end{aligned}$$

donde cada componente del vector aleatorio es variable aleatoria.

Por ejemplo, se puede considerar que además de medir el tiempo que se tarda en grabarse un fichero (X_1), estudiamos el tiempo de vida (X_2) y el tipo de ordenador (X_3). Tendríamos así un vector aleatorio de dimensión 3, $X' = (X_1, X_2, X_3)$.

La distribución de probabilidad del vector aleatorio viene caracterizada por una función $f(x_1, \dots, x_n)$ que en el caso discreto recibe el nombre de *función de probabilidad conjunta* y en el continuo de *función de densidad conjunta*. Además,

- Si X es discreto, $P(X \in A) = \sum \dots \sum_{x \in A} f(x)$.
- Si X es continuo, $P(X \in A) = \int \dots \int_A f(x) dx_1 \dots dx_n$.

Dado un vector aleatorio p-dimensional, a la distribución de cada una de las variables componentes le llamamos *distribución marginal*. Para denotar a las distribuciones marginales usaremos la siguiente nomenclatura: $f_i(x_i)$ será la función de densidad o de probabilidad, según corresponda, de la variable componente X_i con $i = 1, \dots, p$ (se obtiene a partir de la distribución conjunta sumando o integrando sobre el soporte del resto de variables).

Por ejemplo, para $p = 2$, las distribuciones marginales serán:

Vector Discreto $f_i(x_i) = \sum_{x_j} f(x_i, x_j)$

Vector Continuo $f_i(x_i) = \int_{-\infty}^{\infty} f(x_i, x_j) dx_j$

Diremos que n variables aleatorias X_1, \dots, X_n son independientes si para n conjuntos A_1, \dots, A_n cualesquiera de números reales, se cumple:

$$P(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^n P(X_i \in A_i)$$

o, equivalentemente:

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n$$

siendo f , y f_i funciones de densidad o probabilidad según el caso de estudio

Algunas características de los vectores aleatorios serán:

Vector de medias Viene dado por $\mu' = (E(X_1), \dots, E(X_p))$ con

$$E(X_i) = \int \dots \int x_i f(x_1, \dots, x_n) dx_1 \dots dx_n$$

En el caso discreto la integral se transforma en sumatorio.

Covarianza La covarianza entre dos variables X_i y X_j se define como:

$$Cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))] = \sigma_{ij}$$

Y para $i = j$, obtenemos la varianza,

$$Var(X_i) = Cov(X_i, X_i) = E[(X_i - E(X_i))(X_i - E(X_i))] = \sigma_i^2$$

Las diferentes varianzas y covarianzas se representan en forma matricial a través de una matriz $p \times p$ simétrica y definida positiva:

$$Cov(X, X') = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p^2 & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

Correlación Se define como $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$. Si X_i y X_j son independientes se cumple que $\sigma_{ij} = \rho_{ij} = 0$. El recíproco es cierto sólo en el caso Gaussiano. Las correlaciones también se representan de forma matricial:

$$Corr(X, X') = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix}$$

Finalmente nos centraremos en la *distribución normal multivariante*.

Se dice que las variables aleatorias X_1, \dots, X_p siguen una distribución normal multivariante, y lo denotaremos por $X \sim N_p(\mu, \Sigma)$, si su función de densidad conjunta es de la forma:

$$f(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)\Sigma^{-1}(x - \mu)\right)$$

Para $p = 2$ la podemos representar gráficamente. Tendremos la llamada superficie de Gauss que tiene un máximo en el punto, $\mu = (\mu_1, \mu_2)$, la relación $(x - \mu)\Sigma^{-1}(x - \mu) = c^2$ define elipses concéntricas y la función de densidad es constante en tales elipses, constituyendo las llamadas curvas de equiprobabilidad.

Propiedad

Si $X \sim N_p(\mu, \Sigma)$, A es una matriz $r \times p$ y b un vector r -dimensional, se cumple que

$$Y = AX + b \sim N_r(A\mu + b, A\Sigma A^t)$$

Una muestra aleatoria de una distribución multivariante será de la forma X_1, \dots, X_n observaciones, con $X'_i = (X_{i1}, \dots, X_{ip})$. Normalmente lo expresaremos en forma matricial:

$$[X'_1, \dots, X'_n] = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix}$$

donde cada vector X_i supondremos que sigue la misma distribución que X , una distribución normal multivariante. Si suponemos que repetimos n veces el experimento de forma independiente, para una muestra de este tipo podemos calcular el *vector media muestral* y la matriz de *varianzas-covarianzas muestral*:

Vector de medias

$$\begin{aligned} \bar{X}' &= (\bar{X}_1, \dots, \bar{X}_p) \\ \bar{X}_j &= \sum_{i=1}^n \frac{X_{ij}}{n} \end{aligned}$$

Matriz de varianzas-covarianzas muestral

$$S = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ s_p^2 & s_{p2} & \dots & s_p^2 \end{pmatrix}$$

con:

$$s_{ij} = \sum_{k=1}^n \frac{(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{n}$$

7.4 Lección 4: Direcciones Web donde encontrar bancos de datos estadísticos

Uno de los objetivos de esta asignatura es la aplicación directa de las técnicas aprendidas a bancos de datos reales. Es por ello que al alumno, en estas primeras sesiones de la asignatura, se le proporcionan una colección de direcciones web en las que podrá encontrar bancos de datos de muy diversas fuentes. El objetivo es que elija algunos bancos de datos que le puedan interesar para que durante el resto del curso vaya aplicando aquellas técnicas susceptibles de poder ser aplicadas en el contexto de esos datos. De hecho este trabajo es complementario a cualquier otra actividad de la asignatura y forma parte de la evaluación final (ver capítulo de evaluación de la asignatura).

Algunas de estas direcciones web son:

- <http://www.statistics.com/>.
- <http://www.nilesonline.com/data/>.
- <http://www.worldbank.org/data/>.

8 Capítulo II: Análisis de la varianza

8.1 Lección 1: Introducción

En muchos problemas científicos el interés se centra en explicar la variabilidad de un fenómeno en función de un conjunto de otras variables, determinando si entre estas variables existe una relación, y en tal caso, cual es la función que las liga.

En este capítulo estudiaremos cómo construir un modelo para representar la dependencia lineal de una variable respuesta, Y , respecto a otras variables explicativas, X_1, \dots, X_n . El modelo lineal permite cómodas deducciones matemáticas, y puede ser utilizado en muchos problemas mediante adecuadas transformaciones de los datos observados.

La utilidad de encontrar un modelo que ligue a las dos o más variables, radica en la posibilidad de predecir el valor de una de ellas a partir del conocimiento de la otra. Sin embargo conviene advertir a los alumnos que el hecho de que entre ambas variables exista una buena correlación no significa que exista una relación de causalidad. La alta correlación puede deberse a que una tercera variable influya sobre ambas de forma regular.

Los modelos lineales pueden subdividirse a su vez en dos grandes grupos: *modelos de regresión* (que estudiaremos en el capítulo siguiente), en los cuales las variables explicativas son generalmente continuas y pueden ser no controlables, y *modelos del análisis de la varianza*, en los cuales las variables explicativas son generalmente cualitativas y controlables por el investigador. En este contexto se les suele denominar *factores*.

Ambos modelos pueden tratarse matemáticamente de una manera unificada, pero los problemas principales en su aplicación práctica son muy distintos. Por ejemplo, un aspecto central en los primeros es el diseño de la muestra, mientras que en los segundos esta cuestión se supone resuelta. Es por ello que los estudiaremos en temas diferentes.

En este capítulo nos centramos en tres tipos de modelos de ANOVA: con un único factor, bloques aleatorizados y diseño de dos factores con interacción. Creemos que la presentación en clase de teoría de únicamente estos tres modelos es suficiente para que el estudiante comprenda las ideas teóricas en que se basan estos modelos y pueda usar igualmente diseños más complicados con cualquier software estadístico.

Antes de empezar a modelizar, creemos importante hacer algunos comentarios sobre tipos de factores y de diseños.

Los factores pueden ser de *efectos fijos*, si los niveles (tratamientos dentro del factor) son fijos y todos los que nos interesan y de *efectos aleatorios*, si los niveles son una muestra aleatoria del conjunto de posibles niveles.

Por otra parte, según la forma de asignación de la muestra a cada nivel del factor (diseño del experimento) tendremos:

- *Anova con diseño completamente aleatorizado*, si la muestra es asignada al azar en cada nivel del factor.
- *Anova con diseño aleatorizado en bloques*, si agrupamos los individuos en bloques según una característica que pueda influir en la variable respuesta y dentro de cada bloque asignamos los individuos al azar en los niveles.
- *Anova con medidas repetidas*, si asignamos todos los individuos a todos los niveles del factor.

8.2 Lección 2: Análisis de la varianza de una vía

Iniciamos el estudio del análisis de la varianza considerando el modelo de un factor, diseño completamente aleatorizado con efectos fijos. En este modelo se supone que se han obtenido

k muestras de la variable respuesta, cada una de ellas observada en uno de los valores o niveles del factor.

Las hipótesis básicas de este modelo son:

$$Y_{ij} = \mu_j + \epsilon_{ij} \quad i = 1, \dots, n_j \quad j = 1, \dots, k,$$

donde haremos las siguientes suposiciones para las variables aleatorias $\{\epsilon_{ij}\}$.

- Tienen esperanza nula.
- Su varianza es siempre constante, σ^2 .
- Tienen una distribución normal.
- Son independientes entre sí.

Una formulación alternativa de estas hipótesis es la siguiente:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

con μ la media general, $\alpha_j = \mu_j - \mu$ el efecto incremental en la media debido al nivel j del factor.

Los estimadores puntuales de los parámetros del modelo son los habituales:

$$\hat{\mu} = \bar{Y} = \frac{\sum_i \sum_j Y_{ij}}{\sum_i \sum_j n_j}$$
$$\hat{\mu}_j = \bar{Y}_{.j} = \frac{\sum_i Y_{ij}}{n_j}$$
$$\hat{\alpha}_j = \hat{\mu}_j - \hat{\mu}, \text{ donde } n = \sum_j n_j$$

La hipótesis de interés en este tipo de problemas es la de que *no hay diferencias significativas entre los niveles del factor*, que queda formalmente expresada por $H_0 : \mu_1 = \dots = \mu_k$ o equivalentemente $H_0 : \alpha_j = 0 \forall j$.

La idea para deducir el estadístico de contraste se basa en descomponer la *variabilidad total* de los datos en dos términos: la *variabilidad entre* las medias de cada muestra y la media general, y la *variabilidad dentro* de cada grupo o *residual*.

$$\sum_i \sum_j (Y_{ij} - \bar{Y})^2 \equiv \sum_j n_j (\bar{Y}_{.j} - \bar{Y})^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_{.j})^2$$

SCE SCE SCR

Cuando haya diferencias reales entre las medias en cada nivel, la variabilidad entre será grande comparada con la variabilidad residual. Juzgar su tamaño relativo requiere conocer su distribución en el muestreo.

Se demuestra que cuando H_0 es cierta SCE/σ^2 y SCR/σ^2 tienen una distribución ji-cuadrado con $k - 1$ y $n - 1$ grados de libertad respectivamente.

El estadístico de contraste será pues:

$$F = \frac{SCE/(k-1)}{SCR/(n-k)}$$

que tendrá una distribución F con $k - 1$ y $n - k$ grados de libertad.

Los términos de la descomposición en que se basa este contraste suelen disponerse en una tabla conocida como *tabla ANOVA*. En la tabla 1 podemos ver la tabla ANOVA para este primer modelo.

Cuando el efecto del factor resulta significativo por el test F , interesa a menudo establecer nuevas hipótesis entre los niveles del factor. En particular consideramos contrastes del tipo:

Fuente variación	Suma de cuadrados	G. L.	Varianzas	Estadístico
Entre niveles	$\sum_j n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$k - 1$	$VE = SCE/(k - 1)$	$\frac{VE}{VR}$
Residual	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{.j})^2$	$n - k$	$VR = SCR/(n - k)$	
Total	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$	$n - 1$	$SCT/(n - 1)$	

Table 1: Tabla ANOVA de un factor.

$H_0 : \sum_j c_j \mu_j = 0$ con c_1, \dots, c_k constantes conocidas con $\sum_j c_j = 0$. Un posible test para contrastar este tipo de hipótesis es el test de Scheffé basado en el estadístico:

$$S = \frac{(\sum_j c_j \bar{Y}_{.j})^2}{(k - 1) \frac{SCR}{(n - k)} \sum_j \frac{c_j^2}{n_j}},$$

el cual se demuestra que tiene, bajo H_0 , una distribución F con $(k - 1)$ y $(n - k)$ grados de libertad.

La segunda suposición de este modelo se basa en la igualdad de varianzas entre niveles. Un contraste de la hipótesis $H_0 : \sigma_1^2 = \dots = \sigma_k^2$ viene dado por el test de Bartlett cuyo estadístico se distribuye como una ji-cuadrado con $k - 1$ grados de libertad.

En el caso de efectos aleatorios, el modelo toma la forma

$$Y_{ij} = \mu + A_j + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2), A_j \sim N(0, \sigma_A^2)$$

En este caso la hipótesis de interés será $H_0 : \sigma_A^2 = 0$.

8.3 Lección 3: Análisis de la varianza en un diseño de bloques aleatorizados

El siguiente modelo que presentamos es el modelo ANOVA de un factor con un diseño en bloques aleatorizados o equivalentemente un diseño de dos factores sin interacción o un diseño de medidas repetidas (donde el bloque sería el individuo). En el anterior modelo los factores no controlados por el experimentador y que podían influir en los resultados se asignaban al azar a las observaciones. En este modelo, las unidades experimentales han sido agrupadas según otra causa de variabilidad que puede influir en los resultados o *variable de bloqueo*.

Las hipótesis básicas del modelo para efectos fijos son:

$$Y_{ij} = \mu + \alpha_j + \beta_i + \epsilon_{ij} \quad i = 1, \dots, n \quad j = 1, \dots, k,$$

donde los ϵ_{ij} son variables $N(0, \sigma^2)$ independientes.

El modelo descompone la respuesta en:

- Una media global μ .
- El efecto incremental en la media debida al nivel del factor, α_j ($\sum_j \alpha_j = 0$).
- El efecto incremental en la media debida a la unidad experimental o al bloque, β_i ($\sum_i \beta_i = 0$).
- El error experimental, ϵ_{ij} , que recoge el efecto de todas la restantes causas posibles de variabilidad del experimento.

F. V.	Suma de cuadrados	G. L.	Varianzas	Estadístico
Entre niveles	$\sum_j n (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$k - 1$	$VE = \frac{SCE}{k - 1}$	$F_N = \frac{VE}{VR}$
Entre bloques	$\sum_j n_j (\bar{Y}_{.i} - \bar{Y}_{..})^2$	$b - 1$	$VB = \frac{SCB}{b - 1}$	$F_B = \frac{VB}{VR}$
Residual	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{.j} - \bar{Y}_{.i} + \bar{Y}_{..})^2$	$(b - 1)(k - 1)$	$VR = \frac{SCR}{(b - 1)(k - 1)}$	
Total	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$	$kb - 1$	$\frac{SCT}{kb - 1}$	

Table 2: Tabla ANOVA de un factor diseño en bloques aleatorizados.

Notar que este modelo supone que los efectos del factor y de la variable de bloqueo son aditivos, es decir no existe interacción entre ambos.

La hipótesis de interés será $H_0 : \alpha_j = 0 \forall j$ pero en algunas ocasiones también puede ser de interés contrastar $H_0 : \beta_i = 0 \forall i$.

La deducción de los estadísticos de contraste y sus distribuciones muestrales se obtiene de una manera completamente análoga al anterior. En la tabla 2 podemos ver la tabla de ANOVA para este modelo. En la misma, F_N se distribuye como $F_{k-1, (k-1)(b-1)}$ y F_B se distribuye como $F_{b-1, (k-1)(b-1)}$. Además, $n = kb$.

Nuevamente, podemos realizar comparaciones múltiples mediante aplicaciones del test de Scheffé en cada factor.

En el caso de efectos aleatorios, el modelo toma la forma:

$$Y_{ij} = \mu + A_j + B_i + \epsilon_{ij} \quad i = 1, \dots, n \quad j = 1, \dots, k,$$

donde $\epsilon_{ij} \sim N(0, \sigma^2)$, $A_j \sim N(0, \sigma_A^2)$ y $B_i \sim N(0, \sigma_B^2)$ son variables independientes. En este caso las hipótesis a contrastar serán: $H_0 : \sigma_A^2 = 0$ y $H_0 : \sigma_B^2 = 0$.

8.4 Lección 4: Análisis de la varianza de dos vías

El último modelo de ANOVA que se estudiara es el de dos factores. Suponemos ahora que la observación de la variable Y está influida por dos factores y tomamos una muestra de tamaño n en cada combinación de ambos niveles.

Las hipótesis básicas del modelo con efectos fijos son:

$$Y_{ijl} = \mu + \alpha_j + \beta_i + (\alpha\beta)_{ij} + \epsilon_{ijl} \quad i = 1, \dots, b \quad j = 1, \dots, k \quad l = 1, \dots, r,$$

donde b y k son el número de niveles de cada factor y r es el número de observaciones para cada combinación de niveles. Además, las variables ϵ_{ijl} son $N(0, \sigma^2)$ e independientes. El modelo descompone la respuesta en:

- Una media global μ .
- El efecto incremental en la media debida al nivel del factor, α_j ($\sum_j \alpha_j = 0$).
- El efecto incremental en la media debido al bloque, β_i ($\sum_i \beta_i = 0$).
- $(\alpha\beta)_{ij}$ representa la interacción entre ambos factores.
- El error experimental, ϵ_{ijl} , que recoge el efecto de todas la restantes causas posibles de variabilidad del experimento.

F. V.	Suma de cuadrados	G. L.	Varianzas	Estadístico
Factor A	$\sum_j br (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$k - 1$	$VA = \frac{SCA}{k-1}$	$F_A = \frac{VA}{VR}$
Factor B	$\sum_j kr (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$b - 1$	$VB = \frac{SCB}{b-1}$	$F_B = \frac{VB}{VR}$
Interacción	$\sum_i \sum_j r (\bar{Y}_{ij.} - \bar{Y}_{.j.} - \bar{Y}_{i..} + \bar{Y}_{..})^2$	$(k-1)(b-1)$	$VAB = \frac{SCAB}{(k-1)(b-1)}$	$F_{INT} = \frac{VAB}{VR}$
Residual	$\sum_i \sum_j \sum_l (Y_{ijl} - \bar{Y}_{ij.})^2$	$kb(r-1)$	$VR = \frac{SCR}{kb(r-1)}$	
Total	$\sum_i \sum_j \sum_l (Y_{ijl} - \bar{Y}_{..})^2$	$kbr - 1$	$\frac{SCT}{kbr-1}$	

Table 3: Tabla ANOVA de dos factores con interacción.

Tres son las hipótesis de interés: $H_0 : \alpha_j = 0 \forall j$, $H_0 : \beta_i = 0 \forall i$ y $H_0 : (\alpha\beta)_{ij} = 0 \forall i, j$.

En la tabla 3 podemos ver la tabla de ANOVA para este modelo.

Con los estadísticos F_A, F_B contrastamos las igualdades de medias en cada factor. Con el estadístico F_{INT} contrastamos la significatividad del efecto interacción. Podemos de nuevo utilizar la versión del test de Scheffé para las comparaciones múltiples.

Finalmente, podemos considerar modelos con efectos aleatorios en los dos factores o en uno sólo.

9 Capítulo III: Análisis de regresión

9.1 Lección 1: Introducción

Ya hemos comentado en la introducción al ANOVA que el *análisis de regresión* es un ejemplo de *modelo lineal* en el que tanto la variable(s) explicativa(s) como la respuesta son generalmente continuas y pueden ser no controlables.

La idea de la regresión es determinar un modelo estocástico funcional para representar la dependencia de una variable cuantitativa respecto a una colección de variables explicativas (aleatorias o no). Normalmente, expresaremos por Y la variable respuesta o dependiente y por X_1, \dots, X_n el resto de variables explicativas. El modelo propuesto será de la forma $Y = f(X_1, \dots, X_n) + \epsilon$, donde aquí nos centraremos en expresiones lineales de $f(\cdot)$. Si en la expresión anterior, $n = 1$, hablaremos de *regresión lineal simple*, y si $n > 1$ hablamos de *regresión lineal múltiple*. Aunque no lo consideraremos en este curso, si disponemos en el modelo anterior de varias variables dependientes, hablaremos de *modelo de regresión multivariante*.

La segunda lección está dedicada a los modelos de regresión simple. Empezamos poniendo especial atención en las hipótesis de partida, cuyo olvido es una inagotable causa de errores en el trabajo aplicado. La regresión múltiple se presentará como una generalización de la anterior y para ello se utilizará extensamente el álgebra matricial. Es conveniente dejar la resolución de los problemas de regresión múltiple para las clases de prácticas, debido a su gran complejidad de cálculo.

9.2 Lección 2: Modelo de regresión lineal simple

Empezamos la lección estudiando las hipótesis básicas del *modelo de regresión simple* que son:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n$$

donde las letras minúsculas y_i, x_i denotan las observaciones de las variables Y, X . Haremos las siguientes suposiciones para las variables aleatorias $\{\epsilon_i\}$

- Tienen esperanza nula.
- Su varianza es siempre constante, σ^2 , y no depende de x .
- Tienen una distribución normal.
- Son independientes entre sí.

El primer objetivo consiste en estimar los parámetros β_0, β_1 y σ^2 , para ello usamos el método de los mínimos cuadrados (aquí recordaremos brevemente en que consiste este método que los alumnos dieron en la asignatura de Estadística) para obtener la *recta de regresión*:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

que estima el valor medio de Y , \hat{y}_i , para cada valor de X .

Llamaremos *vector de residuos* a (e_1, \dots, e_n) con $e_i = y_i - \hat{y}_i$, los residuos serán muy útiles (como veremos más adelante) para estudiar la validez de las hipótesis del modelo.

El estimador de σ^2 se denomina *varianza residual* y viene dado por

$$s_r^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}.$$

Parámetro	Estimador	Media	Varianza	Distribución
β_0	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	β_0	$\frac{\sigma^2}{n} (1 + (\frac{\bar{x}}{s_x})^2)$	Normal
β_1	$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$	β_1	$\frac{\sigma^2}{ns_x^2}$	Normal
σ^2	$\hat{\sigma}^2 = s_r^2$	σ^2	$\frac{2\sigma^4}{n-2}$	$\frac{(n-2)s_r^2}{\sigma^2} \sim \chi_{n-2}^2$

Table 4: Distribuciones muestrales de los estimadores de los parámetros en el modelo de regresión simple.

Parametro	Intervalo
β_0	$(\hat{\beta}_0 \pm_{1-\alpha/2} t_{n-2} \frac{s_r}{\sqrt{n}} \sqrt{1 + \bar{x}^2/s_x^2})$
β_1	$(\hat{\beta}_1 \pm_{1-\alpha/2} t_{n-2} \frac{s_r}{s_x \sqrt{n}})$
σ^2	$(\frac{(n-2)s_r^2}{1-\alpha/2 \chi_{n-2}^2}, \frac{(n-2)s_r^2}{\alpha/2 \chi_{n-2}^2})$

Table 5: Intervalos de confianza para los estimadores de los parámetros en el modelo de regresión simple.

Una vez estimado el modelo, para poder contrastar hipótesis o hacer inferencias respecto a los parámetros, obtenemos la distribución en el muestreo de los estimadores anteriores. La tabla 4 resume estas distribuciones.

La distribución de probabilidad de los estimadores se utiliza, en primer lugar para obtener intervalos de confianza sobre los parámetros del modelo de regresión. La tabla 5 contiene estos intervalos.

Los estadísticos utilizados para construir intervalos de confianza permiten también realizar cualquier contraste sobre los valores de los parámetros. En particular, es de especial interés el contraste $\beta_1 = 0$, que implica la falta de relación lineal entre las variables.

Este mismo contraste denominado *contraste de regresión* se presenta a continuación mostrando su relación con el análisis de la varianza. Así pues, se descompone la variación total como:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

esta expresión descompone la variabilidad de Y en dos términos independientes: el primero refleja la variabilidad no explicada por la regresión, que es debida al carácter aleatorio de la relación; el segundo contiene la variabilidad explicada por la regresión, y puede interpretarse como la parte determinista de la variabilidad de la respuesta. Se verifica que:

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n-2)}$$

sigue una distribución F con 1 y $n-2$ grados de libertad bajo la hipótesis nula $H_0 : \beta_1 = 0$. De hecho, esto lo podemos expresar en formato de tabla ANOVA como sigue. Denotamos por $SCT = \sum_{i=1}^n (y_i - \bar{y})^2$, $SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ y por $SCReg = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, entonces tenemos como resultado la tabla 6.

Estudiamos a continuación medidas de la bondad de ajuste. Una medida de bondad del ajuste puede ser la varianza residual, sin embargo, esta medida no es útil para comparar rectas de regresión porque depende de las unidades de medida. Una medida más adecuada del ajuste es el *coeficiente de determinación* R^2 definido como la proporción de variabilidad explicada. Se demuestra fácilmente que el coeficiente de determinación es el cuadrado del coeficiente de correlación lineal.

Fuente variación	Suma de cuadrados	G. L.	Varianzas	Estadístico
Explicada	SCReg	1	$SCReg$	$\frac{SCReg}{s_r^2}$
Residual	SCR	$n-2$	$s_r^2 = SCR/(n-2)$	
Total	SCT	$n-1$	$SCT/(n-1)$	

Table 6: Tabla ANOVA para el contraste de regresión.

Empezábamos la lección presentando las hipótesis básicas del modelo de regresión simple. Estas hipótesis son fundamentales para las inferencias sobre los parámetros que hemos presentado. Nos planteamos ahora la comprobación de estas hipótesis del modelo. Si dispusiéramos de varias observaciones de Y para cada valor de X , sería posible utilizar contrastes estadísticos para comprobar las hipótesis de normalidad, de igualdad de varianzas y de independencia. Pero habitualmente se tiene únicamente un valor de Y para cada valor de X por lo que la contrastación de las hipótesis básicas se efectúa, a posteriori, mediante el análisis de los residuos.

Analizaremos los residuos para comprobar:

- si su distribución es aproximadamente normal;
- si su variabilidad es constante, y no depende de X ;
- si presentan evidencia de una relación no lineal entre las variables;
- si existen observaciones atípicas o heterogéneas respecto a la variable X , a la Y , o a ambas.

Para contrastar la normalidad se utilizan tests no paramétricos, aunque los residuos no son independientes, pero si n es grande este efecto es despreciable. La heterocedasticidad y la no linealidad puede detectarse representando gráficamente los residuos.

Es importante señalar la importancia de la transformación de los datos en muchos casos para lograr la adaptabilidad de éstos al modelo lineal.

Una vez ajustado el modelo y comprobado la validez de sus hipótesis, lo lógico es usarlo para estimar el valor de la variable respuesta. Un modelo de regresión permite:

- Estimar las medias de las distribuciones de Y para cada valor de X .
- Predecir futuros valores de la variable respuesta, conociendo el valor de X .

Aunque en ambos casos los valores numéricos son idénticos, la precisión de estas estimaciones es distinta. Con estas estimaciones terminamos con el modelo de regresión simple.

9.3 Lección 3: Modelo general de regresión

El *modelo de regresión lineal múltiple* es una extensión para k variables explicativas del modelo simple para una. Para la observación i -ésima, el modelo queda de la forma

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \epsilon_i \quad i = 1, \dots, n$$

con $n > k + 1$ y con las siguientes suposiciones para las variables aleatorias $\{\epsilon_i\}$:

- Tienen esperanza nula.
- Su varianza es siempre constante, σ^2 .

- Tienen una distribución normal.
- Son independientes entre sí.

El modelo puede expresarse en forma matricial:

$$Y = X\beta + \epsilon$$

donde:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Como vemos, las hipótesis básicas del modelo de regresión múltiple son análogas a las de regresión simple. La metodología es también similar: estimación de los parámetros, contrastes de simplificación, diagnosis y validación. Su expresión en forma matricial simplifica enormemente los cálculos a la hora de obtener los estimadores de los parámetros así como las propiedades de los mismos.

La estimación por el método de los mínimos cuadrados (y por el de máxima verosimilitud) del vector de parámetros es:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

que sigue una distribución Normal de parámetros, $\hat{\beta} \sim N_{k+1}(\beta, \sigma^2(X'X)^{-1})$.

El estimador de la varianza σ^2 vendrá dado por la varianza residual s_r^2 , que toma la forma

$$s_r^2 = \frac{Y'Y - \hat{\beta}X'Y}{n - (k + 1)}$$

Cabe destacar algunas diferencias respecto a lo visto en el contexto de regresión simple:

- El contraste de regresión es ahora $H_0 : \beta_1 = \dots = \beta_k = 0$, que se puede resolver mediante una tabla ANOVA definiendo las correspondientes sumas de cuadrados, cuadrados medios y estadístico F.
- El coeficiente de determinación que introducíamos en regresión simple, y que ahora toma la forma $R^2 = \frac{SCR_{Reg}}{SCT}$, aumenta con el número de variables explicativas. Para evitar esto se define el *coeficiente de determinación corregido por grados de libertad*. Este coeficiente se define como

$$R_{adj}^2 = 1 - \frac{s_r^2}{SCT/(n-1)} = 1 - \frac{n-1}{n-k-1} \frac{SCR}{SCT}$$

- Un problema que puede surgir al construir un modelo de regresión múltiple es el de la *multicolínealidad*, que se presenta cuando las variables explicativas son muy dependientes entre sí, lo que hace que los estimadores sean inestables y con gran varianza.

- Un problema muy importante en el análisis de regresión múltiple es determinar cuáles de las variables explicativas en la lista inicial deberán incluirse en el modelo de regresión. Cuando k es grande, puede no ser práctico determinar y evaluar todas las posibles ecuaciones de regresión. Para estos casos se han desarrollado *técnicas para selección de las variables*. La técnica más usual de selección de variables emplea un procedimiento de regresión por pasos para obtener la mejor ecuación de regresión. Existen dos versiones principales de esta técnica: la selección hacia adelante (método forward) y la eliminación hacia atrás (método backward).

10 Capítulo IV: Análisis Discriminante

10.1 Lección 1: Introducción

El análisis discriminante es una técnica de clasificación y asignación de un individuo a un grupo conocidas sus características. En el análisis discriminante se dispone de una serie de grupos definidos a priori, con un conjunto de observaciones para cada individuo referidas a una colección de variables relevantes. En base a esta información se calcula una función discriminante que se puede utilizar para hacer predicciones futuras.

El objetivo del análisis discriminante consiste en lo siguiente. En primer lugar se determina si en función de las variables originales disponibles, los grupos quedan suficientemente discriminados. Esto podría ser una explicación al fenómeno de las diferencias entre grupos. Se trata de analizar cuáles son las variables que contribuyen más a discriminar entre los grupos que se han formado. Para ello lo que se hace es "reducir" las variables que mejor discriminan a unas pocas nuevas variables, que se denominan *variedades o variables canónicas*. Generalmente, una sola variable canónica es la que aporta la mayor explicación. Estas variables canónicas son combinación lineal de las variables originales y vienen expresadas por una *función discriminante*.

La función discriminante es una ecuación lineal con una variable dependiente que representa la pertenencia a un grupo. Combinaciones lineales de variables independientes (predictores) sirven de base para clasificar a los individuos entre los grupos. De hecho, cuando se trabaja con dos grupos, la función discriminante no es más que una ecuación de regresión múltiple en la cual la variable dependiente es una variable nominal con valores 0-1, que representan la pertenencia al grupo respectivo.

En general cuando se dispone de k grupos, se pueden calcular hasta $k-1$ funciones discriminantes, todas ellas incorreladas. Para que una función discriminante sea óptima debe proporcionar una regla de clasificación que minimice la probabilidad de cometer errores.

Aunque tenemos varios métodos de análisis discriminante (métodos geométricos y estimativos), teniendo en cuenta a quién va dirigida esta asignatura, nos centraremos en los métodos geométricos basados en la distancia de Mahalanobis y transformaciones a variedades canónicas.

10.2 Lección 2: Notación y estructura de los datos

En el contexto del análisis discriminante, disponemos de p variables clasificadoras X_1, \dots, X_p , de k grupos o poblaciones sobre los que hay que discriminar y la matriz usual de datos tiene la siguiente estructura:

	X_1	X_2	\dots	X_p
Grupo 1	$x_{11}^{(1)}$	$x_{12}^{(1)}$	\dots	$x_{1p}^{(1)}$
	$x_{21}^{(1)}$	$x_{22}^{(1)}$	\dots	$x_{2p}^{(1)}$
	\vdots	\vdots	\dots	\vdots
	$x_{n_1}^{(1)}$	$x_{n_1 2}^{(1)}$	\dots	$x_{n_1 p}^{(1)}$
Grupo 2	$x_{11}^{(2)}$	$x_{12}^{(2)}$	\dots	$x_{1p}^{(2)}$
	$x_{21}^{(2)}$	$x_{22}^{(2)}$	\dots	$x_{2p}^{(2)}$
	\vdots	\vdots	\dots	\vdots
	$x_{n_2}^{(2)}$	$x_{n_2 2}^{(2)}$	\dots	$x_{n_2 p}^{(2)}$
\vdots	\vdots	\vdots	\dots	\vdots
Grupo k	$x_{11}^{(k)}$	$x_{12}^{(k)}$	\dots	$x_{1p}^{(k)}$
	$x_{21}^{(k)}$	$x_{22}^{(k)}$	\dots	$x_{2p}^{(k)}$
	\vdots	\vdots	\dots	\vdots
	$x_{n_k}^{(k)}$	$x_{n_k 2}^{(k)}$	\dots	$x_{n_k p}^{(k)}$

donde $x_{ij}^{(l)}$ es el valor observado de la variable j -ésima en el individuo i -ésimo de la población l -ésima con $j = 1, \dots, p$, $i = 1, \dots, n_l$ y $l = 1, \dots, k$. Por tanto, n_l representa el tamaño de la muestra en el grupo l -ésimo y $n = n_1 + \dots + n_k$ es el número total de individuos.

A partir de aquí podemos definir toda una colección de matrices y vectores que utilizaremos en los métodos de discriminación.

La matriz de observaciones en el grupo l -ésimo viene definida por

$$X^{(l)} = \begin{pmatrix} x_{11}^{(l)} & \dots & x_{1p}^{(l)} \\ \vdots & \dots & \vdots \\ x_{n_l 1}^{(l)} & \dots & x_{n_l p}^{(l)} \end{pmatrix}$$

El vector de observaciones del individuo i -ésimo en el grupo l -ésimo es

$$X_i^{(l)} = \begin{pmatrix} x_{i1}^{(l)} \\ \vdots \\ x_{ip}^{(l)} \end{pmatrix}$$

De donde es claro que $X^{(l)} = (X_1^{(l)}, \dots, X_{n_l}^{(l)})$ y el vector media muestral general será $\bar{X} = (\bar{X}_1, \dots, \bar{X}_p)$, siendo $\bar{X}_j = \sum_i \sum_l x_{ij}^{(l)} / n$.

El vector media muestral dentro del grupo l -ésimo es $\bar{X}^{(l)} = (\bar{X}_1^{(l)}, \dots, \bar{X}_p^{(l)})$, siendo $\bar{X}_j^{(l)} = \sum_i x_{ij}^{(l)} / n_l$.

Finalmente, si denotamos por $S^{(l)}$ la matriz de varianzas-covarianzas dentro del grupo l -ésimo, ésta vendrá dada por la expresión

$$S^{(l)} = \sum_{i=1}^{n_l} \frac{(X_i^{(l)} - \bar{X}^{(l)})(X_i^{(l)} - \bar{X}^{(l)})'}{n_l - 2}.$$

10.3 Lección 3: Método basado en la distancia de Mahalanobis

La distancia de Mahalanobis fue propuesta por este autor en 1936 y es una generalización de la distancia euclídea, que tiene en cuenta la matriz de covarianzas intragrupos para corregir por la dispersión y relaciones entre variables.

Dado un vector de observaciones X , definimos la distancia de Mahalanobis de X a la población l -ésima con vector media $\mu^{(l)}$ y matriz de varianzas-covarianzas $\Sigma^{(l)}$ como

$$D_l = (X - \mu^{(l)})' (\Sigma^{(l)})^{-1} (X - \mu^{(l)})$$

Sin embargo, con datos experimentales los parámetros teóricos son desconocidos y debemos estimarlos mediante los muestrales. Así, $\hat{\mu}^{(l)} = \bar{X}^{(l)}$ y $\hat{\Sigma}^{(l)} = \frac{\sum_{i=1}^k n_i S^{(l)}}{n-k}$. Finalmente, la distancia de Mahalanobis muestral será

$$D_l = (X - \bar{X}^{(l)})' (\hat{\Sigma}^{(l)})^{-1} (X - \bar{X}^{(l)})$$

y clasificaremos en vector de observaciones X en el grupo l -ésimo sobre el que se minimice D_l .

10.4 Lección 4: Método de variedades canónicas

La idea subyacente a las variedades canónicas es realizar una transformación lineal a las p variables predictoras que consiga:

1. La media total sea cero.
2. Separar lo más posible los grupos.
3. La matriz de varianzas-covarianzas intragrupo sea la identidad (es decir, variables incorreladas).

Además con esta transformación conseguiremos reducir el número de variables y nos permitirá ver qué variables influyen más en la discriminación. A esta transformación lineal buscada se le llamará variedad canónica.

A partir de X_1, \dots, X_p , buscamos nuevas variables Z_1, \dots, Z_p (después veremos que no hace falta llegar a p) tales que

$$Z_j = u_{0j} + u_{1j}X_1 + \dots + u_{pj}X_p, \quad j = 1, \dots, p$$

cumpliéndose las condiciones 1-3 anteriores. Sin embargo, para llegar aquí empezamos buscando una transformación que sólo cumpla 2-3. Para ello, buscamos nuevas variables Y_j cumpliendo que

$$Y_j = t_{1j}X_1 + \dots + t_{pj}X_p = t_j'X, \quad j = 1, \dots, p$$

donde $t_j' = (t_{1j}, \dots, t_{pj})$. El criterio para la obtención de los vectores coeficientes de esta transformación se basa en el problema de maximización de

$$\text{Max}_t \frac{t'Bt}{t'Wt} \text{ sujeto a } t'Wt = n - k$$

siendo B la matriz que define la variabilidad entre grupos,

$$B = \sum_{l=1}^k n_l (\bar{X}^{(l)} - \bar{X}) (\bar{X}^{(l)} - \bar{X})'$$

y W la matriz que define la variabilidad intra grupos,

$$W = \sum_{l=1}^k \sum_{i=1}^{n_l} (\bar{X}_i^{(l)} - \bar{X}^{(l)}) (\bar{X}_i^{(l)} - \bar{X}^{(l)})' = \sum_{l=1}^k n_l S^{(l)}.$$

Los máximos buscados se alcanzan en relación a los valores propios de $W^{-1}B$ de tal forma que si $\lambda_1 > \dots > \lambda_s$ son los $s = \min(p, k-1)$ mayores valores propios y r_1, \dots, r_s son los vectores propios asociados, entonces los coeficientes $t_j = \sqrt{n - kr_j}$.

Finalmente, para conseguir media cero, restamos el vector media general para obtener las variedades canónicas de la forma $Z_j = \sqrt{n - kr_j}(X - \bar{X}) = \sqrt{n - kr_j}X - \sqrt{n - kr_j}\bar{X}$.

El método para discriminar será el siguiente:

1. A cada $X_i^{(l)}$ le aplicamos la transformación $Z_i^{(l)}$.
2. Calculamos los vectores centroides de cada grupo, es decir, $\bar{Z}^{(l)} = \sum_{i=1}^{n_l} Z_i^{(l)} / n_l$.
3. A cada nuevo vector de observaciones $Z = (Z_1, \dots, Z_p)'$ le aplicamos la transformación para dar $Z = (Z_1, \dots, Z_p)'$.
4. Calculamos la distancia al grupo l -ésimo dada por $D_l = (Z - \bar{Z}^{(l)})' (Z - \bar{Z}^{(l)})$ y clasificamos el nuevo vector en aquel grupo en el que D_l sea menor.

Finalmente, acabaremos la lección comentando algunos contrastes de significación. Para contrastar la hipótesis nula de que $H_0 : \lambda_i = 0$ (que implicará que el resto de valores propios $\lambda_j = 0$ con $j = i + 1, \dots, s$) utilizamos el estadístico V de Bartlett dado por $V = -\left\{n - 1 - \frac{p+k}{2}\right\} \ln \Lambda$, donde $\Lambda = \prod_{j=i}^s \frac{1}{1 + \lambda_j}$. Este estadístico se distribuye como una $\chi_{p(k-1)}^2$. Existe una versión secuencial de este contraste.

11 Capítulo V: Análisis Cluster

11.1 Lección 1: Introducción

El análisis cluster es el arte de buscar grupos en datos. Se trata de una técnica utilizada para combinar observaciones en grupos o clusters de forma que:

1. Cada grupo sea homogéneo con respecto a ciertas características. Es decir, las observaciones en cada grupo son similares unas a otras.
2. Cada grupo debería ser diferente de los otros grupos con respecto a las mismas características; es decir, las observaciones de un grupo deben ser diferentes de las de otros grupos.

La definición de *similaridad* u *homogeneidad* varía de análisis en análisis y depende de los objetivos del estudio. La clasificación de objetos similares en grupos es una importante actividad humana pues forma parte del proceso diario de aprendizaje. Es por ello que el análisis cluster es considerado como una parte del *reconocimiento de patrones* y de la *inteligencia artificial*.

Hay que hacer notar que el análisis cluster es bastante diferente al análisis discriminante en que el primero define y busca los grupos, mientras que el segundo asigna objetos o individuos a grupos previamente establecidos.

A lo largo del tiempo, el análisis cluster ha sido renombrado de muy diversas formas tales como por ejemplo taxonomía numérica, clasificación automática o análisis tipológico.

Geoméricamente, el concepto de análisis cluster es bastante sencillo. En el caso de un análisis cluster basado en dos variables, cada observación puede ser representada como un punto en un espacio bidimensional. Y, en general, cada observación se puede representar como un punto en un espacio p-dimensional, donde p es el número de características o variables que se usan para describir los individuos.

Como hemos comentado, el análisis cluster agrupa las observaciones más semejantes entre sí y diferentes con respecto al resto en función de una colección de variables de agrupación. Pero también es posible agrupar variables de tal forma que las variables en cada grupo sean similares con respecto a las observaciones de agrupación. Si disponemos de n observaciones, geoméricamente esto es equivalente a representar los datos en un espacio n-dimensional e identificar clusters de variables. Este segundo objetivo del análisis cluster parece similar al del análisis factorial. Hay que recordar que el análisis factorial intenta identificar clusters de variables, de forma que las variables en cada cluster tengan algo en común, es decir, que de alguna forma midan al factor latente en común. Es por tanto posible utilizar análisis factorial para buscar grupos de observaciones y el análisis cluster para agrupar variables. La técnica del análisis factorial utilizada para agrupar observaciones se conoce como análisis *Q-factorial*. Sin embargo esta técnica no es aconsejable en este contexto. En general se recomienda que:

1. Si se está interesado en identificar factores latentes y sus indicadores, se debe utilizar el análisis factorial pues se trata de una técnica desarrollada específicamente para este objetivo.
2. Si el interés está puesto en agrupar observaciones, entonces se debe utilizar el análisis cluster.

Los pasos en un análisis cluster son:

1. Seleccionar una medida de similaridad.

2. Selección de la técnica de clustering: método jerárquico o no jerárquico.
3. Selección del método de agrupación adecuado a la técnica escogida en el paso anterior.
4. Selección del número de clusters.
5. Interpretación de la solución.

11.2 Lección 2: Tipos de datos para el análisis cluster

Para un análisis cluster podemos disponer de dos tipos de estructuras de datos:

1. En formato de matriz individuo \times variable. En este caso tendremos los siguientes datos: p variables y n observaciones con el formato matricial

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

donde x_{ij} denota la observación i-ésima de la variable j-ésima. Denotaremos por $X_i = (x_{i1}, \dots, x_{ip})'$ el vector asociado al individuo i-ésimo.

2. En formato de matriz simétrica con datos de proximidades entre cada par de individuos. En este caso tendremos una matriz simétrica $n \times n$. Las proximidades pueden definirse como similaridades (miden la cercanía entre pares de individuos) o disimilaridades (miden lo distantes que están entre sí cada par de individuos).

A la hora de llevar a cabo un análisis cluster, podemos encontrarnos con diferentes tipos de datos cuya estructura (y posibles transformaciones) hay que tenerlas en cuenta a la hora de utilizar unas medidas u otras.

Por ejemplo, supongamos que las variables vienen medidas en escala de intervalo, es decir, variables normalmente continuas. Un simple cambio en las unidades de medida puede provocar un cambio en la solución final de los grupos encontrados. Esto nos hace plantearnos, para evitar la dependencia en las unidades de medida, una estandarización de los datos. Para ello, para la variable j-ésima calculamos la media $m_j = \sum_{i=1}^n x_{ij}/n$ y la desviación estándar $s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - m_j)^2}$. Sin embargo, esta medida está afectada por la posible presencia de outliers. En este caso podemos utilizar la medida de desviación absoluta con respecto a la media definida por $dam_j = \frac{1}{n} \sum_{i=1}^n |x_{ij} - m_j|$. Con estas medidas podemos definir las medidas estandarizadas como $z_{ij} = \frac{x_{ij} - m_j}{s_j}$ o bien $z_{ij} = \frac{x_{ij} - m_j}{dam_j}$. Finalmente, en lugar de trabajar con la matriz de datos original, se aplicarán los procedimientos de cluster sobre la matriz $n \times p$ formada por las puntuaciones z.

Sin embargo, la estandarización no siempre es beneficiosa. Se trata de una opción que según el caso puede ser útil o no. Por ejemplo, puede ocurrir que algunas variables sean intrínsecamente más importantes que otras en una cierta aplicación y una estandarización nos podría hacer perder esta información.

11.3 Lección 3: Formulación geométrica: medidas de similaridad

Cualquier algoritmo de clustering requiere de algún tipo de medida para estudiar la similitud o disimilitud entre un par de observaciones o clusters. Las medidas de similaridad pueden clasificarse en tres tipos:

1. Medidas de distancia

2. Coeficientes de asociación

3. Coeficientes de correlación

11.3.1 Medidas de distancia

La distancia más popular es la *distancia Euclidea*, la cual se define entre un par de puntos u observaciones de p dimensiones como

$$d_{ij} = d(X_i, X_j) = \left(\sum_{k=1}^p (X_{ik} - X_{jk})^2 \right)^{1/2}$$

Esta distancia es un caso particular de la *distancia de Minkowski* (métrica L_q) dada por

$$d_{ij} = d(X_i, X_j) = \left(\sum_{k=1}^p (|X_{ik} - X_{jk}|)^q \right)^{1/q}$$

con $n = 2$. Si consideramos $n = 1$, tenemos una nueva definición de distancia dada por

$$d_{ij} = d(X_i, X_j) = \sum_{k=1}^p |X_{ik} - X_{jk}|$$

que se llama *distancia city-block o de Manhattan*.

La versión *estandarizada* de la distancia Euclidea viene dada por

$$d_{ij} = d(X_i, X_j) = \left(\sum_{k=1}^p \left(\frac{X_{ik} - X_{jk}}{s_k} \right)^2 \right)^{1/2}$$

donde s_k representa la desviación estandar de la variable k-ésima. s_k puede ser sustituido por dam_k .

Por otra parte, esta versión estandarizada es un caso particular de una *distancia Euclidea ponderada*, la cual viene definida por

$$d_{ij} = d(X_i, X_j) = \left(\sum_{k=1}^p w_k (X_{ik} - X_{jk})^2 \right)^{1/2}$$

donde cada variable recibe un peso de acuerdo con su importancia.

Otro ejemplo de distancia estadística es la distancia de *Mahalanobis* (utilizada en el análisis discriminante). Se trata de una generalización de la distancia Euclidea para tener en cuenta la correlación entre las variables y es invariante frente cambios de escala

$$d_{ij} = d(X_i, X_j) = (X_i - X_j)' S^{-1} (X_i - X_j)$$

donde S es la matriz de varianzas-covarianzas.

En general cualquier medida de distancia de similitud está basada en el concepto de una métrica cuyas propiedades son:

1. $d(X_i, X_j) \geq 0$
2. $d(X_i, X_i) = 0$
3. $d(X_i, X_j) = d(X_j, X_i)$
4. $d(X_i, X_j) \leq d(X_i, X_h) + d(X_h, X_j)$

11.3.2 Coeficientes de asociación

En el caso de disponer de variables binarias, sobre una tabla 2×2 de la forma

	$X_j = 1$	$X_j = 0$	
$X_i = 1$	a	b	a+b
$X_i = 0$	c	d	c+d
	a+c	b+d	p

un coeficiente de similitud entre los dos individuos viene dado por

$$s(X_i, X_j) = \frac{a + d}{a + b + c + d}$$

Lógicamente, el correspondiente coeficiente de disimilitud vendrá dado por

$$dis(X_i, X_j) = \frac{c + b}{a + b + c + d}$$

En el caso de disponer de variables nominales, la forma más común de medir similitudes y/o disimilitudes entre un par de individuos es mediante la técnica de "simple matching" consistente en

$$s(X_i, X_j) = \frac{u}{p}; \quad dis(X_i, X_j) = \frac{p - u}{p} \quad (1)$$

donde u es el número de coincidencias entre los individuos i-ésimo y j-ésimo.

En el caso de variables en escala ordinal, la mayoría de los autores tratan los rangos como si fueran variables en escala de intervalo y por tanto aplican las técnicas antes comentadas. Pero previo a esto, hay que transformar los rangos a puntuaciones entre 0 y 1 de la siguiente forma

$$z_{ij} = \frac{r_{ij} - 1}{M_j - 1}$$

donde r_{ij} es el rango de la i-ésima observación en la variable j-ésima y M_j representa el mayor rango para la variable j-ésima.

Finalmente, supongamos que disponemos de p variables con diferentes estructuras. Entonces, podemos definir una medida de disimilitud $dis(X_i, X_j)$ (conocida como disimilitud de Gower) entre los individuos i y j como

$$dis(X_i, X_j) = \frac{\sum_{k=1}^p \delta_{ij}^k dis_{ij}^k}{\sum_{k=1}^p \delta_{ij}^k} \quad (2)$$

δ_{ij}^k es una función indicatriz que toma el valor 1 si ambas medidas x_{ik} y x_{jk} para la variable k-ésima no son faltantes, y 0 en otro caso. Además, $\delta_{ij}^k = 0$ si la variable k-ésima es un atributo binario no simétrico y los individuos i y j forman un emparejamiento 0-0. La expresión (2) no se puede calcular si todos los δ_{ij}^k son cero. En este caso, o bien eliminamos los individuos i o j, o bien damos a $dis(X_i, X_j)$ un valor cualquiera.

Por otra parte, dis_{ij}^k representa la contribución de la variable k-ésima a la disimilitud entre los individuos i y j. Si la variable k-ésima es *binaria o nominal*, $dis_{ij}^k = 1$ si $x_{ik} \neq x_{jk}$ y $dis_{ij}^k = 0$ si $x_{ik} = x_{jk}$. Si todas las variables fueran nominales, (2) es el número de coincidencias sobre el total de pares posibles, es decir, coincide con (1). Si la variable k-ésima es de *escala intervalo*, dis_{ij}^k viene dada por

$$dis_{ij}^k = \frac{|x_{ik} - x_{jk}|}{\max_h x_{hk} - \min_h x_{hk}} \quad (3)$$

Para variables *ordinales*, podemos utilizar (3) tras sustituir los x_{ik} por sus rangos. Si todas las variables fueran continuas (de escala intervalo), (2) pasa a ser la distancia de Manhattan. Si, finalmente queremos calcular un índice de similitud, basta con hacer $s(X_i, X_j) = 1 - dis(X_i, X_j)$. Hay que advertir que en general los coeficientes de asociación no satisfacen las propiedades de una verdadera métrica.

11.3.3 Coeficientes de correlación

Otras medidas para aplicar análisis cluster vienen definidas por medio de coeficientes de correlación. Por ejemplo, podemos calcular el estadístico paramétrico de correlación de Pearson dado por

$$r(X_k, X_j) = \frac{\sum_{i=1}^n (x_{ik} - m_k)(x_{ij} - m_j)}{\sqrt{\sum_{i=1}^n (x_{ik} - m_k)^2} \sqrt{\sum_{i=1}^n (x_{ij} - m_j)^2}}$$

o bien la versión no paramétrica dada por el coeficiente de correlación de Spearman. Ambos coeficientes pueden ser convertidos a medidas de disimilitud mediante $dis(X_k, X_j) = (1 - r(X_k, X_j))/2$ o $dis(X_k, X_j) = 1 - |r(X_k, X_j)|$. También pueden ser transformadas a medidas de similitud mediante $dis(X_k, X_j) = (1 + r(X_k, X_j))/2$ o $dis(X_k, X_j) = |r(X_k, X_j)|$.

Una versión del índice de correlación, es el índice de *correlación cofenética* en el que utilizamos las variables d_{ij} (distancia entre individuos i y j) y d_{ij}^* (número de pasos para juntar los individuos i y j). Se trata de una medida de fiabilidad.

11.4 Lección 4: Cluster jerárquico

Dada una matriz de similaridad entre individuos, hemos de definir técnicas para agrupar ya no individuos, sino grupos de individuos, clusters. Existen dos tipos de métodos jerárquicos: *aglomerativos* y *divisivos*. Los métodos jerárquicos aglomerativos se caracterizan por el hecho de empezar con una matriz de similitud de $n \times n$ e ir reduciéndola en cada paso hasta llegar a la dimensión $g \times g$, siendo g el número de clusters que había que encontrar. Es decir, en el paso 0 empiezan con n clusters y en cada paso se van juntando dos clusters. Los métodos jerárquicos divisivos empiezan con todos los clusters juntos y en cada paso se van separando dos de ellos. Las distintas etapas suelen representarse mediante un *dendrograma*. La diferencia entre los distintos tipos de métodos jerárquicos viene dada por el tipo de distancias utilizadas para unir clusters. Podemos distinguir entre los siguientes métodos:

1. *Vecino más proximo (nearest-neighbour, single linkage)*. Dados los clusters R y Q , este método calcula la disimilitud (similitud o distancia) entre ambos clusters $d(R, Q)$ como la mínima disimilitud (similitud, distancia) entre todos los posibles pares de individuos en los dos clusters, es decir, $d(R, Q) = \min d(i, j), i \in R, j \in Q$.
2. *Vecino más lejano (furthest-neighbour, complete linkage)*. Dados los clusters R y Q , este método calcula la disimilitud (similitud o distancia) entre ambos clusters $d(R, Q)$ como la máxima disimilitud (similitud, distancia) entre todos los posibles pares de individuos en los dos clusters, es decir, $d(R, Q) = \max d(i, j), i \in R, j \in Q$.
3. *Distancia media*. Dados los clusters R y Q , este método calcula la disimilitud (similitud o distancia) entre ambos clusters $d(R, Q)$ como la media de las disimilitudes (similitudes, distancias) entre cualquier par de individuos en los dos clusters. Así, $d(R, Q) = \text{mediad}(i, j), i \in R, j \in Q$.

4. *Distancia mediana*. Dados los clusters R y Q , este método calcula la disimilitud (similitud o distancia) entre ambos clusters $d(R, Q)$ como la mediana de las disimilitudes (similitudes, distancias) entre cualquier par de individuos en los dos clusters. Así, $d(R, Q) = \text{medianad}(i, j), i \in R, j \in Q$.
5. *Centroide*. Para cada cluster, se calculan los centroides, y la disimilitud (similitud, distancia) entre los clusters será la disimilitud (similitud, distancia) entre los respectivos centroides.
6. *Ward*. El método de Ward no calcula disimilitudes (similitudes, distancias) entre clusters. Lo que hace más bien es formar clusters maximizando la homogeneidad dentro de cada cluster. Como medida de homogeneidad utiliza las sumas de cuadrados dentro de cada cluster. Por tanto este método intenta minimizar las sumas de cuadrados totales dentro de cada grupo (también llamadas sumas de cuadrado del error).

Cada uno de estos métodos jerárquicos aglomerativos se utiliza en determinados contextos. Por ejemplo, para detectar clusters "con formas redondeadas" hay que escoger el método de la media. Pero si que quiere detectar clusters compactos pero no muy separados, el método adecuado será el del vecino más alejado.

11.5 Lección 5: Cluster no jerárquico

En el caso de análisis cluster no jerárquico, los datos se agrupan en k particiones o grupos con k conocido a priori. Las técnicas basadas en este método siguen los siguientes pasos:

1. Selección de k semillas (centroides de los clusters) iniciales, siendo k el número de clusters deseado.
2. Asignación de cada observación al cluster más próximo.
3. Reasignación de cada observación a uno de los k clusters de acuerdo a una cierta regla de parada.
4. El algoritmo para si no hay más realocación de los datos o si la reasignación satisface el criterio de parada. En otro caso, volver al paso 2.

La mayoría de los algoritmos de cluster no jerárquico difieren con respecto a: (1) el método utilizado para obtener las semillas iniciales; (2) la regla utilizada para la reasignación de las observaciones.

De forma general, algunos de los métodos utilizados para obtener las semillas iniciales son:

1. Seleccionar las primeras k observaciones no faltantes como centroides o semillas.
2. Seleccionar la primera observación no faltante como la semilla del primer cluster. La semilla para el segundo se selecciona como aquella que su distancia con respecto a la primera semilla es mayor que un cierto límite impuesto. La tercera semilla se seleccionará como aquella cuya distancia con respecto a las dos primeras semillas es mayor que el límite, y así sucesivamente.
3. Seleccionar aleatoriamente k observaciones no faltantes como semillas.
4. Modificar las semillas seleccionadas mediante ciertas reglas como por ejemplo que estén lo más separadas posible.

5. Utilizar semillas asignadas por el investigador.

Una vez asignadas las semillas, se forman los clusters iniciales asignando cada una de las restantes $n - k$ observaciones a la semilla más cercana.

Con respecto a los métodos utilizados para la reasignación de las variables, algunos más destacados son:

1. Calcular el centroide de cada cluster y reasignar a los individuos a aquel cluster para el que el centroide sea el más cercano. Si el cambio en los centroides de los clusters es mayor que un cierto criterio de convergencia, se procede a una nueva reasignación.
2. Calcular el centroide de cada cluster y reasignar los individuos a aquel cluster para el que el centroide sea el más cercano. Para la asignación de cada observación, se recalcula el centroide del cluster al que la observación es asignada y el del cluster de donde viene la observación. Este proceso continúa hasta que el cambio en los centroides sea menor que un cierto criterio de convergencia.
3. Reasignar las observaciones según se minimice un cierto criterio estadístico. Algunas de las funciones objetivas a minimizar son:
 - (a) Traza de la matriz SSCP (sum of squares and cross-product) dentro de cada grupo.
 - (b) Determinante de la matriz SSCP dentro de cada grupo.
 - (c) Traza de $W^{-1}B$, donde W es la matriz SSCP dentro de cada grupo y B la correspondiente entre grupos.
 - (d) Mayor valor propio de la matriz $W^{-1}B$.

Un algoritmo conocido que cumple con alguno de los criterios comentados es el *K-means*. Este algoritmo consiste en seleccionar las primeras k observaciones como centros de clusters. Las observaciones se reasignan mediante el cálculo de la distancia de cada observación al centroide.

12 Capítulo VI: Análisis de componentes principales

12.1 Lección 1: Introducción

El método de *componentes principales* tiene por objeto transformar un conjunto de variables, a las que se denomina como *variables originales*, en un nuevo conjunto de variables denominadas *componentes principales*. Estas últimas se caracterizan por estar incorreladas entre sí.

En muchas ocasiones el investigador se enfrenta a situaciones en las que, para analizar un fenómeno, dispone de información de muchas variables que están correladas entre sí en mayor o menor grado. Estas correlaciones son como un velo que impiden evaluar adecuadamente el papel que juega cada variable en el fenómeno estudiado. El análisis de componentes principales permite pasar a un nuevo conjunto de variables que gozan de la ventaja de estar incorreladas entre sí y que, además, pueden ordenarse de acuerdo con la información que llevan incorporada. De hecho, si las variables originales estuvieran completamente incorreladas, el análisis de componentes principales carecería de interés, ya que éstas coincidirían con las propias componentes.

Como medida de la cantidad de información incorporada en una componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza, mayor es la información que lleva incorporada la componente. Es por ello que se selecciona como primera componente aquella que tenga mayor varianza. En general, la extracción de componentes principales se efectúa sobre variables tipificadas para evitar problemas derivados de escala, aunque también se puede aplicar sobre variables expresadas en desviaciones respecto la media.

Desde el punto de vista de su aplicación, el método de componentes principales es considerado un *método de reducción*, es decir, un método que permite reducir la dimensión del número de variables inicialmente consideradas.

Esta metodología es similar a otros procedimientos como el análisis discriminante en el hecho de que ambos métodos se basan en la construcción de combinaciones lineales de variables correladas cuyos pesos en dichas combinaciones se obtienen maximizando alguna propiedad estadística.

También las componentes principales se usan en análisis de regresión para determinar los posibles problemas de multicolinealidad. Por tanto se enfatiza al alumno que se trata de una metodología que comparte y establece lazos de unión entre otros métodos estudiados en esta asignatura.

12.2 Lección 2: Propiedades y significado geométrico

Conviene señalar que la obtención de componentes principales es un caso típico de cálculo de vectores y valores propios en matrices simétricas. Por tanto se le recordará al alumno algunos conceptos y operaciones básicas con matrices, como de forma muy concreta el polinomio y sus raíces características.

Algunas propiedades que hay que destacar de este tipo de análisis son:

1. Las componentes principales son combinaciones lineales de las variables originales.
2. Los coeficientes de las combinaciones lineales son los elementos de los vectores propios asociados a la matriz de covarianzas de las variables originales.
3. La primera componente principal está asociada a la mayor raíz característica (valor propio) de la matriz de covarianzas de las variables originales.
4. La varianza de cada componente es igual al valor propio al que va asociado.

- En el caso de que las variables estén tipificadas, la proporción de la variabilidad total de las variables originales captada por una componente es igual al valor propio correspondiente dividido por el número de variables originales.
- La correlación entre una componente y una variable original se determina con el valor propio correspondiente.

Por otra parte, un vector de p variables se puede considerar como un punto en un espacio p -dimensional donde una observación particular representa los valores de los ejes de coordenadas. En este caso, el análisis de componentes principales encuentra un espacio de dimensión menor (subespacio) que proporciona el mejor ajuste para los datos en el espacio p -dimensional. Esto es así de forma que una observación en lugar de ser representada en el espacio p -dimensional, sólo se representará en el subespacio.

Las propiedades geométricas derivadas de las componentes principales serán estudiadas mediante representaciones gráficas en 2-3 dimensiones.

12.3 Lección 3: Método de obtención de las componentes

Supongamos que disponemos de una muestra aleatoria de tamaño n a partir de X_1, \dots, X_p , y que las observaciones están expresadas bien en desviaciones respecto la media o bien como variables tipificadas.

La primera componente, igual que el resto de ellas, se expresa como una combinación lineal de las variables originales

$$Z_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + \dots + u_{1p}X_{pi}, \quad i = 1, \dots, n$$

Para el conjunto de n observaciones muestrales, esta ecuación se puede expresar matricialmente de la forma

$$\begin{pmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1n} \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \dots & X_{p1} \\ X_{12} & X_{22} & \dots & X_{p2} \\ \vdots & \vdots & \dots & \vdots \\ X_{1n} & X_{2n} & \dots & X_{pn} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{pmatrix}$$

que de forma abreviada es

$$Z_1 = X u_1$$

Esta primera componente se obtiene de forma que su varianza es máxima, sujeta a la restricción de que la suma de los pesos (coeficientes u_{1j}) al cuadrado sea igual a la unidad.

Es claro que la primera componente tiene esperanza cero y su varianza viene dada por

$$Var(Z_1) = u_1' \left[\frac{1}{n} X'X \right] u_1$$

Si las variables vienen expresadas en desviaciones respecto la media, $V = (1/n)X'X$ es la matriz de covarianzas muestral; para variables tipificadas $R = (1/n)X'X$ es la matriz de correlaciones. Por generalidad supongamos que vamos a trabajar con V .

Por tanto la varianza a maximizar es $Var(Z_1) = u_1' V u_1$ con la restricción $\sum_{j=1}^p u_{1j}^2 = u_1' u_1 = 1$. La resolución pasa por calcular $(V - \lambda I)u_1 = 0$, de donde se deduce que u_1 es el vector propio asociado al mayor valor propio de la matriz V .

El resto de componentes se calculan de forma similar. La forma matricial de la componente genérica h -ésima es de la forma $Z_h = X u_h$ con la restricción de que $u_h' u_h = 1$ y una

serie de restricciones adicionales dadas por $u_h' u_1 = u_h' u_2 = \dots = u_h' u_{h-1} = 0$, que definen al vector propio asociado a la componente h -ésima como ortogonal al resto de vectores propios calculados anteriormente. Finalmente, el vector u_h estará asociado al valor propio h -ésimo, ordenados los valores propios de mayor a menor.

A continuación se le proponen al alumno una serie de expresiones que caracterizan diferentes propiedades de las componentes principales:

- La varianza de cada componente es igual al valor propio asociado de la matriz V , $Var(Z_h) = u_h' V u_h = \lambda_h$.
- La proporción de la variabilidad total recogida por la componente h -ésima es $\lambda_h / (\sum_{h=1}^p \lambda_h)$, que en el caso de variables tipificadas será λ_h / p .
- La correlación entre la variable X_j y la componente Z_h viene dada por

$$r_{jh} = \frac{Cov(X_j, Z_h)}{\sqrt{Var(X_j)} \sqrt{Var(Z_h)}} = \frac{\lambda_h u_{hj}}{\sqrt{Var(X_j)} \sqrt{\lambda_h}}$$

que en el caso de que X_j esté tipificada, es de la forma $r_{jh} = u_{hj} \sqrt{\lambda_h}$.

- Una vez calculados los coeficientes u_{hj} , se pueden obtener las puntuaciones Z_{hi} , es decir, los valores de las componentes correspondientes a cada observación, a partir de la relación

$$Z_{hi} = u_{h1}X_{1i} + u_{h2}X_{2i} + \dots + u_{hp}X_{pi}, \quad i = 1, \dots, n; \quad h = 1, \dots, p$$

- Podemos trabajar también con componentes tipificadas, es decir, con $Y_h = Z_h / \sqrt{\lambda_h}$. En este caso,

$$Z_{hi} = c_{h1}X_{1i} + c_{h2}X_{2i} + \dots + c_{hp}X_{pi}, \quad i = 1, \dots, n; \quad h = 1, \dots, p$$

con $c_{hj} = u_{hj} / \sqrt{\lambda_h}$.

Finalmente, si el objetivo de esta metodología es reducir las p variables iniciales a $m < p$ componentes principales, hemos de saber cuantas componentes retener. Por tanto acabaremos la lección comentando algunos contrastes de significación de los valores propios. Básicamente disponemos de dos criterios:

- Criterio de la media aritmética*, se retienen aquellas componentes cuyo valor propio excede la media de los valores propios. En el caso de variables tipificadas, esto equivale a retener si el valor es mayor que 1.
- Contraste de los valores propios no relevantes*, nos planteamos la hipótesis nula de $H_0 : \lambda_{m+1} = \lambda_{m+2} = \dots = \lambda_p = 0$ y obtenemos el estadístico

$$Q = \left\{ n - \frac{2p+11}{6} \right\} \left\{ (p-m) \ln \bar{\lambda}_{p-m} - \sum_{j=m+1}^p \ln \lambda_j \right\}$$

que bajo el supuesto de normalidad multivariante de las variables originales se distribuye, bajo H_0 , como una χ_q^2 , siendo $q = (p-m+2)(p-m+1)/2$.

13 Capítulo VII: Análisis factorial

13.1 Lección 1: Introducción

El análisis factorial es un método de análisis multivariante que intenta explicar, según un modelo lineal, un conjunto extenso de variables observables mediante un número reducido de variables hipotéticas llamadas *factores*. Es un aspecto esencial del análisis factorial el que los factores no sean directamente observables, obedeciendo a conceptos de naturaleza más abstracta que las variables originales.

El análisis de componentes principales y el análisis factorial tienen en común que son técnicas para examinar la interdependencia de variables. Difieren en su objetivo, sus características y su grado de formalización.

Mientras que el objetivo del análisis de componentes principales es explicar la mayor parte de la variabilidad total de un conjunto de variables con el menor número de componentes posibles, en el análisis factorial, los factores son seleccionados para explicar las interrelaciones entre variables.

En componentes principales se determinan los pesos o ponderaciones que tienen cada una de las variables en cada componente; es decir, las componentes principales se explican en función de las variables observables. Sin embargo, en el análisis factorial las variables originales juegan el papel de variables dependientes que se explican por factores comunes y únicos, que no son observables.

Por otra parte, el análisis de componentes principales es una técnica estadística de reducción de datos que puede situarse en el dominio de la estadística descriptiva, mientras que el análisis factorial implica la elaboración de un modelo que requiere la formulación de hipótesis estadísticas y la aplicación de métodos de inferencia.

El hecho de que las componentes principales se utilicen como uno de los procedimientos para la extracción de factores, ha llevado a hacer pensar a algunos que ambos métodos son equivalentes, lo cual no es cierto.

El análisis factorial puede ser exploratorio o confirmatorio. El *análisis exploratorio* se caracteriza porque no se conoce a priori el número de factores, y es en la aplicación empírica donde se determina este número. Por el contrario, en el *análisis de tipo confirmatorio* los factores están fijados a priori, utilizándose contrastaciones empíricas para su corroboración.

En este curso nos centraremos en el primer tipo de análisis factorial.

13.2 Lección 2: El modelo de análisis factorial

Por simplicidad en los cálculos, consideraremos que las variables observables X_1, \dots, X_p son variables tipificadas (variables con media 0 y varianza 1). El modelo de análisis factorial se define de la siguiente forma:

$$\begin{aligned} X_1 &= l_{11}F_1 + l_{12}F_2 + \dots + l_{1m}F_m + e_1 \\ X_2 &= l_{21}F_1 + l_{22}F_2 + \dots + l_{2m}F_m + e_2 \\ &\dots \\ X_p &= l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pm}F_m + e_p \end{aligned}$$

donde F_1, \dots, F_m son factores comunes, e_1, \dots, e_p son factores únicos o específicos y l_{jh} es el peso del factor h en la variable j . A los coeficientes de este tipo se les denomina cargas factoriales o saturación de la variable j en el factor h .

En el modelo anterior, cada una de las p variables observables es una combinación lineal de m factores comunes ($m < p$) y de un factor único. Así pues, todas las variables originales

vienen influidas por todos los factores comunes, mientras que existe un único factor que es específico de cada variable. Debe tenerse en cuenta que tanto los factores únicos como los comunes no son observables. Este modelo se puede expresar matricialmente como

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \vdots & \vdots & \dots & \vdots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{pmatrix}$$

o bien $X = LF + e$.

Para poder realizar inferencias sobre el modelo factorial, es preciso formular algunas hipótesis estadísticas sobre los factores comunes y sobre los factores únicos. Las hipótesis sobre los comunes son las siguientes:

1. La esperanza de cada uno de los factores comunes es nula, es decir, $E(F) = 0$.
2. La matriz de covarianzas de los factores comunes es $E(FF') = I$. Esto implica que la varianza de cada uno de los factores es 1 y éstos están incorrelados entre sí. Por tanto, los factores comunes son variables tipificadas.

Las hipótesis de los factores únicos son las siguientes:

1. La esperanza de cada uno de los factores únicos es nula, es decir, $E(e) = 0$.
2. La matriz de covarianzas de los factores únicos es $E(ee') = \Omega$, una matriz diagonal. Es decir, las varianzas de los factores únicos pueden ser distintas y estos factores están incorrelados entre sí.
3. Finalmente, la matriz de covarianzas entre los factores comunes y los únicos es $E(Fe') = 0$.

A partir de estas propiedades se puede demostrar que la matriz de correlación poblacional para las variables X_1, \dots, X_p se puede descomponer en $R_p = LL' + \Omega$ (Teorema de Thurstone) de donde es claro que la varianza de la variable tipificada X_j se podrá descomponer como

$$1 = l_{j1}^2 + l_{j2}^2 + \dots + l_{jm}^2 + w_j^2$$

Si denotamos por $h_j^2 = l_{j1}^2 + l_{j2}^2 + \dots + l_{jm}^2$ (*comunalidad*) y llamamos a w_j^2 *especificidad*, entonces tenemos que $1 = h_j^2 + w_j^2$.

13.3 Lección 3: Métodos de extracción de factores

En la práctica hemos de sustituir los elementos poblacionales por los muestrales. En este sentido la descomposición de Thurstone empírica será de la forma $\hat{R}_p = \hat{L}\hat{L}' + \hat{\Omega}$, siendo \hat{R}_p la matriz de correlaciones muestrales. La cuestión que se plantea es la obtención de las matrices estimadas a partir del conocimiento de \hat{R}_p . En este sentido surgen dos problemas: grados de libertad y la no unicidad de la solución. Respecto al primero, para que el proceso de estimación pueda efectuarse se requiere que el número de ecuaciones sea mayor o igual que el número de parámetros a estimar. Respecto a la unicidad de la solución cabe decir que si L es una solución, cualquier transformación ortogonal de L también será solución.

En el análisis factorial se parte del supuesto de que las variables originales están correladas entre sí debido a que comparten unos mismos factores comunes.

Existen varios métodos para la obtención de los factores que serán comentados al alumno pero nos centraremos, por conexión con el tema anterior, en el método de componentes principales.

Las p componentes principales se pueden expresar de la siguiente forma

$$\begin{aligned} Z_1 &= u_{11}X_1 + u_{12}X_2 + \dots + u_{1p}X_p \\ Z_2 &= u_{21}X_1 + u_{22}X_2 + \dots + u_{2p}X_p \\ &\dots \\ Z_p &= u_{p1}X_1 + u_{p2}X_2 + \dots + u_{pp}X_p \end{aligned}$$

Este conjunto de ecuaciones es reversible, pudiéndose demostrar que es posible expresar las variables en función de las componentes de la siguiente forma

$$\begin{aligned} X_1 &= u_{11}Z_1 + u_{21}Z_2 + \dots + u_{p1}Z_p \\ X_2 &= u_{12}Z_1 + u_{22}Z_2 + \dots + u_{p2}Z_p \\ &\dots \\ X_p &= u_{1p}Z_1 + u_{2p}Z_2 + \dots + u_{pp}Z_p \end{aligned} \quad (4)$$

Un problema que se plantea en (4), para utilizarlo como base para la estimación de los factores, es que las componentes Z no están tipificadas, mientras que los factores teóricos F los hemos definido con varianza 1. Para obviar este problema podemos utilizar componentes tipificados definidos de la siguiente forma $Y_h = Z_h/\sqrt{\lambda_h}$, con $h = 1, \dots, p$ y siendo λ_h la varianza de Z_h . Por tanto, $Z_h = \sqrt{\lambda_h}Y_h$ y sustituyendo en (4), la ecuación j -ésima puede expresarse como

$$X_j = u_{1j}\sqrt{\lambda_1}Y_1 + u_{2j}\sqrt{\lambda_2}Y_2 + \dots + u_{pj}\sqrt{\lambda_p}Y_p \quad (5)$$

Teniendo en cuenta que $u_{hj}\sqrt{\lambda_h}$ es precisamente el coeficiente de correlación entre la variable j -ésima y la componente h -ésima, (5) puede expresarse de esta forma

$$X_j = r_{1j}Y_1 + r_{2j}Y_2 + \dots + r_{pj}Y_p$$

y añadiendo los $(p-m)$ últimos términos obtenemos

$$X_j = r_{1j}Y_1 + r_{2j}Y_2 + \dots + r_{mj}Y_m + (r_{m+1j}Y_{m+1} + \dots + r_{pj}Y_p) \quad (6)$$

Ahora bien, por otra parte, tenemos que la ecuación j -ésima para X_j es la siguiente

$$X_j = l_{j1}F_1 + l_{j2}F_2 + \dots + l_{jm}F_m + e_j \quad (7)$$

De la comparación entre (6) y (7) obtenemos que $\hat{l}_{j1} = r_{1j}, \dots, \hat{l}_{jm} = r_{mj}$.

Una vez estimados los coeficientes anteriores, se puede estimar la comunalidad de la variable X_j de la siguiente forma $\hat{h}_j^2 = \hat{l}_{j1}^2 + \hat{l}_{j2}^2 + \dots + \hat{l}_{jm}^2$. La especificidad se puede estimar directamente mediante la expresión $\hat{w}_j^2 = 1 - \hat{h}_j^2$.

13.4 Lección 4: Contrastes en el modelo factorial

En el análisis factorial, se pueden realizar dos tipos de contrastes según se apliquen previamente al análisis o a posteriori. Con los primeros, se trata de analizar la adecuación de realizar un análisis factorial sobre un determinado conjunto de variables. Con los contrastes a posteriori se pretende evaluar el modelo factorial estimado.

Entre los contrastes a priori tenemos:

1. *Contraste de esfericidad de Bartlett.* La cuestión esencial previa a la realización de un análisis factorial es si están correlacionadas entre sí las variables originales. Si no lo estuvieran, no existirían factores comunes y, por tanto, no tendría sentido aplicar el análisis factorial. Bartlett propuso el siguiente estadístico $B = -[n - 1 - (2p + 5)\frac{1}{6}] \ln |\hat{R}_p|$ que se distribuye como una $\chi_{p^2-p}^2$ para contrastar que todos los coeficientes de correlación teóricos entre cada par de variables son nulos.

2. Los estadísticos Kaiser, Meyer y Olkin propusieron una medida de adecuación de la muestra al análisis factorial conocida por las iniciales KMO y definida de la siguiente forma

$$KMO = \frac{\sum \sum_{h \neq j} r_{jh}^2}{\sum \sum_{h \neq j} r_{jh}^2 + \sum \sum_{h \neq j} a_{jh}^2}$$

siendo r_{jh}^2 los coeficientes de correlación observados entre las variables originales y a_{jh}^2 los coeficientes de correlación parcial entre las variables originales.

En el caso de que exista adecuación de los datos a un modelo de análisis factorial, el término del denominador que recoge los coeficientes será pequeño y, consecuentemente, la medida KMO estará próxima a 1.

Entre los contrastes a posteriori tenemos:

1. *Correlación observada y reproducida.* Si el modelo factorial estimado es adecuado, las diferencias entre los coeficientes de correlación observados y reproducidos deben ser pequeñas. Por tanto se evalúan estas diferencias y si existe un porcentaje elevado de diferencias superiores a una cantidad pequeña prefijada (por ejemplo 0.05) esto será indicativo de que el modelo factorial estimado no se adecúa a los datos.

2. *Medida de la bondad del ajuste.* Se pueden realizar contrastes estadísticos formales de la bondad del ajuste en el caso de que el método de estimación aplicado haya sido el de máxima verosimilitud o el de mínimos cuadrados generalizados, bajo el supuesto de que los datos muestrales proceden de una población normal multivariante. Así, bajo estos supuestos, se puede construir un contraste que tiene una distribución Chi-cuadrado para contrastar la hipótesis nula H_0 : Existen m factores comunes.

13.5 Lección 5: Rotaciones en el análisis factorial

Los procedimientos de rotación de factores se han ideado para obtener, a partir de la solución inicial, unos factores que sean fácilmente interpretables. Con los factores rotados se trata de que cada una de las variables originales tenga una correlación lo más próxima a 1 que sea posible con uno de los factores y correlaciones próximas a 0 con el resto de factores. Existen dos formas básicas de realizar la rotación de factores:

1. *Rotación Ortogonal.* Los ejes se rotan de forma que quede preservada la incorrelación entre los factores, es decir, los nuevos ejes son perpendiculares de igual forma que lo son los factores sin rotar. Entre los diversos procedimientos de rotación ortogonal, el denominado *método Varimax* es el más conocido y aplicado.

2. *Rotación Oblicua.* Con la denominación de oblicua se indica que los ejes no son ortogonales, es decir, no son perpendiculares. Al realizarse una rotación oblicua, los factores ya no estarán incorrelados pero puede conseguirse una asociación más nítida de cada una de las variables con el factor correspondiente. El método de rotación oblicua más conocido es el denominado *Oblimin*.

14 Capítulo VIII: Series temporales

14.1 Lección 1: Introducción

La mayoría de fenómenos aleatorios que se presentan en el mundo de las computadoras varían aleatoriamente sobre el tiempo, por ejemplo si consideramos la variable número de tareas esperando a ser procesadas. Es pues lógico el pensar que la distribución de esta variable no será la misma si se considera a las 8 de la mañana o a las 12. En este capítulo estudiaremos cómo construir un modelo para explicar la estructura y prever la evolución a lo largo del tiempo de este tipo de variables. El marco teórico para el estudio de variables dinámicas es la teoría de procesos estocásticos, es por ello que empezaremos con una introducción general a los procesos estocásticos, se presentarán a continuación los procesos de ARIMA y por último se desarrollará la metodología estadística para ajustar un modelo a una serie observada. También es aconsejable en este caso dejar la resolución de problemas para las clases de prácticas, debido a la gran cantidad de cálculos necesarios.

14.2 Lección 2: Series temporales y procesos estocásticos

El modelo matemático para una serie temporal es el concepto de *proceso estocástico*, es por ello que empezamos el tema introduciendo conceptos básicos del mismo.

Un *proceso estocástico* es una familia de variables aleatorias $\{X(t), t \in T\}$. Así para cada $t \in T$, donde T es el *conjunto de índices* del proceso, $X(t)$ es una variable aleatoria. A un elemento de T se le suele llamar *parámetro tiempo*, porque éste es su significado habitual. Si T es discreto, decimos que el proceso es de *tiempo discreto* y, si T es continuo, diremos que es de *tiempo continuo*. A partir de ahora consideraremos $T = \{\pm 1, \pm 2, \dots\}$ discreto y denotaremos $X_t = X(t)$. El *espacio de estados* del proceso es el conjunto de valores posibles para X_t .

A cualquier conjunto particular de observaciones del proceso le llamaremos *realización* del proceso. Una *serie temporal de observaciones* $\{X_1, \dots, X_T\}$ es una parte de la realización de un proceso.

Diremos que conocemos la estructura probabilística de un proceso estocástico si conocemos la distribución conjunta de cualquier conjunto de n variables $\{X_{t_1}, \dots, X_{t_n}\}$.

Llamaremos *función de medias* μ_t del proceso a la función del tiempo que proporciona las medias de las distribuciones marginales X_t para cada t y *función de varianzas* σ_t^2 a la que proporciona las varianzas. Por último llamaremos *función de covarianzas* $cov(t, t+j)$ a la función que describe las covarianzas en dos instantes y *función de autocorrelación* $\rho(t, t+j)$ a la estandarización de la función de covarianzas, esto es, al coeficiente de correlación entre X_t y X_{t+j} .

Para poder estudiar estas características es necesario suponer que son estables a lo largo del tiempo. Esto conduce a la definición de estacionariedad. Se dice que un proceso es *estacionario en sentido estricto* si la distribución conjunta de cualquier conjunto de variables del proceso no cambia al someter a todas las variables a un mismo desplazamiento en el tiempo. Decimos que es *estacionario en sentido débil* cuando las funciones de medias y varianzas del proceso no depende de t y la función de autocovarianza entre dos variables del proceso sólo depende del tiempo transcurrido entre los dos correspondientes periodos de tiempo.

Si el proceso es estacionario, la función de autocorrelación, como función del periodo de separación k , ρ_k entre las dos variables, recibe el nombre de *función de autocorrelación simple o correlograma*.

Llamaremos *coeficiente de autocorrelación parcial de orden k* a la relación lineal entre observaciones separadas k periodos con independencia de los valores intermedios. Llamaremos *función de autocorrelación parcial* a la representación de los coeficientes de autocorrelación parcial en función del retardo.

Para llevar a cabo el tratamiento de series temporales nos apoyaremos en procesos estocásticos estacionarios en sentido débil. La mayoría de series temporales que encontramos en el mundo de la economía o de la computación no son estacionarias, esto requiere aplicar una serie de transformaciones que conducen a la estacionariedad de la serie. En cuanto a la media, la serie original se somete a d diferencias sucesivas. En cuanto a la varianza las transformaciones más utilizadas son las *transformaciones de Box-Cox*.

Un proceso estacionario muy útil en el estudio de series temporales es el *proceso de ruido*, que es aquel proceso estacionario con media cero y con función de autocovarianza constante e igual a cero. El proceso es de *ruido blanco* si la distribución de cada variable es normal.

Por último definiremos el *operador retardo* B como $BX_t = X_{t-1}$ y denotaremos por $\phi_p(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ a un polinomio de grado p en el retardo cuyo primer término es la unidad.

14.3 Lección 3: Procesos ARIMA

Tras estos conceptos introductorios se introducen los modelos básicos para el estudio de series temporales. A partir de ahora $\{a_t\}$ será un proceso de ruido blanco y X_t, X_{t+1}, \dots se toman como desviaciones respecto a la media.

- *Procesos autorregresivos*, $AR(p)$: $\phi_p(B)X_t = a_t$. Llamaremos *ecuación característica* del proceso a $\phi_p(B) = 0$ considerada como función de B . El proceso es estacionario si todas las raíces de la ecuación característica son menores que la unidad.

La función de autocorrelación simple se obtiene con la ecuación en diferencias finitas de orden p : $\phi_p(B)\rho_k = 0 \quad k > 0$.

Se demuestra que la función de autocorrelación es una mezcla de exponenciales y sinusoidales con una estructura muy compleja y que se amortigua al avanzar el retardo.

Particularizando la anterior ecuación para $k = 1, \dots, p$, se obtiene un sistema de ecuaciones que relacionan las p primeras autocorrelaciones con los parámetros del proceso y que reciben el nombre de ecuaciones de Yule-Walker.

La función de autocorrelación parcial de un proceso autorregresivo tendrá solamente los p primeros coeficientes de autocorrelación distintos de cero.

- *Procesos de medias móviles*, $MA(q)$: $X_t = \theta_q(B)a_t$. Estos procesos son siempre estacionarios.

Se demuestra que cualquier proceso $AR(p)$ estacionario es equivalente a un $MA(\infty)$.

Diremos que un proceso es *invertible* si el efecto del pasado decrece con el tiempo. El proceso es invertible si las raíces de $\theta(B) = 0$ son en módulo, mayores que la unidad. Cualquier $MA(q)$ invertible equivale a un $AR(\infty)$.

La función de autocorrelación simple de un proceso $AR(q)$ cumple que todas las autocorrelaciones superiores a q son nulas. De forma dual al modelo autoregresivo, la función de autocorrelación parcial será una sucesión infinita, mezcla de exponenciales y sinusoidales y decreciendo al aumentar el retardo.

- *Procesos Autorregresivos de Medias Mviles*, ARMA(p,q): $\phi_p(B)X_t = \theta_q(B)a_t$. El proceso será estacionario si las raíces de $\phi_p(B) = 0$ están dentro del círculo unidad y será invertible si las raíces de $\theta_q(B) = 0$ están fuera.

Dadas las características de las funciones de autocorrelación parcial y simple en procesos AR y MA, resulta fácil encontrar las correspondientes a un proceso ARMA. Dado que éste es la suma de los procesos AR y MA.

Respecto a la función de autocorrelación simple, a partir del q decrecerá como una mezcla de exponenciales y sinusoidales, determinada por la parte autorregresiva. En la función de autocorrelación parcial los efectos se acumularán hasta superar el orden del proceso autorregresivo, y a partir de p sólo se mantendrá el efecto del proceso MA y así pues decrecerá como mezcla de exponenciales y sinusoidales.

- Los *Procesos* ARIMA(p, d, q) se presentan cuando d raíces de la parte AR de un proceso ARMA(p,q) son la unidad, dando lugar a procesos no estacionarios. Si definimos el operador diferencia $\Delta = 1 - B$, tenemos que

$$\phi_p(B)\Delta^d X_t = \theta_q(B)a_t$$

. Casos particulares son el camino aleatorio (ARIMA(0,1,0)) y el alisado exponencial (ARIMA(0,1,1)).

La función de autocorrelación simple de un proceso ARIMA se caracteriza por tener coeficientes positivos que se amortiguarán muy lentamente.

Para finalizar el tema, una vez que se ha modelizado la serie temporal ésta puede usarse para *predecir* futuras observaciones. Si se ha observado (X_1, \dots, X_t) , la predicción del valor X_{t+k} se basa en el hecho de que la predicción óptima en el sentido de minimizar el error cuadrático medio de predicción coincide con la esperanza de la distribución condicionada: $E[X_{t+k}|X_1, \dots, X_t]$.

14.4 Lección 4: Ajuste del modelo y predicciones

Hasta ahora hemos estudiado las propiedades teóricas de los procesos ARIMA. Veamos a continuación como ajustar estos modelos a una serie real. Para ello haremos uso de la metodología propuesta por Box-Jenkins que consta de tres etapas:

Identificación del modelo Utilizar los datos recogidos para sugerir un conjunto reducido de posibles modelos. Consiste en:

- Decidir qué transformaciones se aplicarán para convertir el proceso subyacente en estacionario.
- Determinar un modelo para el proceso estacionario, es decir, los órdenes p y q calculando las funciones de autocorrelación simple y parcial del proceso.

Estimación de los parámetros del modelo Esta operación se puede llevar a cabo mediante la minimización de la suma de cuadrados de los errores.

Diagnóstico del modelo Requiere comprobar:

- Los residuos del modelo estimado se aproximan al comportamiento de un ruido blanco.
- El modelo es estacionario e invertible.
- Los coeficientes son estadísticamente significativos.
- Los coeficientes del modelo están poco relacionados entre sí.
- El grado de ajuste es elevado en comparación al de otros modelos alternativos.

15 Bibliografía

La bibliografía sobre los temas de esta asignatura es bastante amplia y diversa en muchos contextos aplicados. Sin embargo si buscamos referencias que se ajusten de forma exacta al temario propuesto para esta asignatura reducimos muchísimo estas aportaciones. En este sentido podemos distinguir entre bibliografía básica y aconsejada para el alumno.

En muchas ocasiones, presentaremos libros que se pueden recomendar a los estudiantes que estén interesados en profundizar en alguno de los capítulos de la asignatura, pero sin estar especialmente dirigidos a la ingeniería.

15.1 Bibliografía básica

Los siguientes textos de forma parcial o completa podrían ser utilizados como libros de texto en esta asignatura, aunque muchos otros podrían ser también recomendados:

- Abascal y Grande (1989) [1]. Contiene ejemplos ilustrativos de la teoría aplicados a la investigación comercial.
- Aczel (1989) [2]. Es un libro autocontenido y de nivel asequible aunque enfocado a los métodos estadísticos en la economía de la empresa. El capítulo dedicado al análisis de la varianza es muy completo.
- Afifi y Clark (1990) [3]. Se hace una aproximación de carácter intuitivo y claro de distintos métodos de análisis multivariante, con referencias a programas de ordenador.
- Bisquerra (1989) [6]. Libro que cubre parte del temario con muchos ejemplos prácticos resueltos con diferentes softwares.
- Cooper y Weekes (1981) [8]. La exposición de los métodos de análisis multivariante está realizada de forma intuitiva.
- Cuadras (1981) [9]. Contiene información abundante sobre los algoritmos que se utilizan en el análisis multivariante. Es quizás muy teórico.
- Green (1978) [12]. Es un libro clásico de análisis multivariante en que se desarrollan todos los cálculos que hay que realizar en los distintos métodos.
- Jobson (1992) [16]. Es un manual actual y completo de métodos de análisis multivariante, aunque con cierto grado de dificultad.
- Peña (1991) [21]. Excelente libro que se ha convertido en un clásico en la estadística aplicada. Consta de dos volúmenes que cubren desde los conceptos básicos de estadística hasta los modelos lineales.
- Sharma (1996) [25]. Libro muy ameno de leer, al tiempo que presenta metodológicamente casi todos los métodos del análisis de datos multivariantes. Contiene bastantes ejercicios resueltos.
- Uriel (1995) [26]. Es un libro bastante completo, que bien podría ser considerado como libro de texto de la asignatura. Muchos ejemplos a nivel econométrico.

15.2 Bibliografía para consultar

A continuación presentamos una serie de libros que pueden ser recomendados a los estudiantes para ampliar determinados temas.

- Para ampliar el capítulo de *Modelos Lineales*, análisis de la varianza y de regresión, se pueden consultar los siguientes textos:
 - Searle (1971) [24]. Libro clásico sobre modelos lineales.
 - Berry y Feldman (1985) [5]. Buena exposición en forma de libro de bolsillo de los métodos de regresión múltiple.
 - Girder (1987) [10]. Métodos de anova con medidas repetidas en formato de libro de bolsillo.
- Para ampliar los capítulos de *Análisis Discriminante y Cluster*, se pueden consultar los siguientes textos:
 - Kaufman y Rousseeuw (1989) [17]. Libro muy didáctico pero monotemático, dedicándose exclusivamente al análisis cluster.
 - Anderberg (1973) [4]. Exposición intuitiva de los métodos del análisis cluster en vistas a la aplicación.
 - Romesburg (1984) [22]. Buen libro de consulta para cualquier investigador. Tiene ejemplos en muy diversos contextos.
 - Goldstein y Dillon (1978) [11]. Buen libro para profundizar en una versión particular del análisis discriminante con datos discretos.
 - Hartigan (1975) [14]. Buena exposición de los distintos métodos de clustering.
- Para ampliar los capítulos de *Análisis de Componentes Principales y Factorial*, se pueden consultar los siguientes textos:
 - Harman (1976) [13]. Recopilación de las técnicas de análisis factorial.
 - Jackson (1991) [15]. Buen libro, desde la práctica, para el uso de las técnicas de componentes principales.
 - McDonald (1985) [19]. Libro muy didáctico, particularmente en el intento de relacionar diferentes metodologías con el análisis factorial.
- Para ampliar el capítulo de *Series Temporales*, se pueden consultar los siguientes textos:
 - Ostrom (1978) [20]. Buen libro sobre series temporales incluyendo métodos de regresión.
 - Box y Jenkins (1976) [7]. Libro básico y clásico de consulta sobre series temporales.

15.3 Relación de referencias bibliográficas

[1] Abascal, E. y Grande, I. (1989) *Métodos multivariantes para investigación comercial. Teoría, aplicaciones y programación BASIC*. Ariel Economía, Barcelona.

[2] Aczel, A.D. (1989) *Complete business statistics*. Irwin, Boston.

- [3] Affi, A.A. y Clark, V. (1990) *Computer-aided multivariate analysis*. Segunda edición, VNR. Nueva York.
- [4] Anderberg, M.R. (1973) *Cluster analysis for applications*. Academic, New York.
- [5] Berry, W.D. y Feldman, S. (1985) *Multiple regression in practice*. Quantitative Applications in the Social Sciences, Sage University Paper.
- [6] Bisquerra, R. (1989) *Introducción conceptual al análisis multivariable*. Volúmenes 1 y 2. Editorial PPU, Barcelona.
- [7] Box, G.E.P. y Jenkins, G.M. (1976) *Time series: data analysis, forecasting and control*. Holden Day.
- [8] Cooper, R.A. y Weekes, A.J. (1983) *Data, models and statistical analysis*. Philip Allan, Oxford.
- [9] Cuadras, C.M. (1981) *Métodos de análisis multivariable*. Eumibar, Barcelona.
- [10] Girden (1987) *ANOVA: Repeated measures*. Quantitative Applications in the Social Sciences, Sage University Paper.
- [11] Goldstein, M. y Dillon, W.R. (1978) *Discrete discriminant analysis*. Wiley, New York.
- [12] Green, P.E. (1978) *Analyzing multivariate data*. The Dryden Press, Illinois.
- [13] Haman, H.H. (1976) *Modern factor analysis*. Chicago.
- [14] Hartigan, J. (1975) *Clustering algorithms*. Wiley, New York.
- [15] Jackson, J.E. (1991) *A User's guide to principal components*. Wiley, New York.
- [16] Jobson, J.D. (1992) *Applied multivariate data analysis*. Springer-Verlag.
- [17] Kaufman, L. y Rousseeuw, P.J. (1989) *Finding groups in data. An introduction to cluster analysis*. Wiley Series.
- [18] Lebart, L., Morineau, A. y Fenelon, J.P. (1987) *Tratamiento de los datos estadísticos. Métodos y programas*. Bordas, Paris.
- [19] McDonald, R. (1985) *Factor analysis and related techniques*. Lawrence Erlbaum, Hillsdale, N.J.
- [20] Ostrom, C.J. (1978) *Time series analysis: regression techniques*. Beverly Hills, CA. Sage.
- [21] Peña, D. (1991) *Estadística. Modelos y métodos*. Volúmenes 1 y 2. Alianza Editorial, segunda edición.
- [22] Romesburg, H.C. (1984) *Cluster analysis for researchers*. Lifetime Learning Publications, California.
- [23] Scheffe, H. (1959) *The Analysis of variance*. John Wiley and Sons.
- [24] Searle, S.R. (1971) *Linear Models*. Wiley.
- [25] Sharma, S. (1996) *Applied multivariate techniques*. John Wiley & Sons.
- [26] Uriel, E. (1995) *Análisis de datos. Series temporales y análisis multivariante*. Editorial AC, Madrid.