COLABORACIONES

Taxonomía de las técnicas clásicas de Análisis Multivariante

1. INTRODUCCIÓN

El término *análisis multivariante* engloba una serie de técnicas estadísticas dispares que tienen en común el hecho de que todas se aplican a conjuntos de datos o medidas realizadas sobre un cierto número de individuos u objetos (items, en notación aséptica). Por lo general, los individuos observados constituyen una muestra aleatoria extraída de una población que incluye el conjunto de todos ellos. De esta forma, dentro del análisis multivariante surgen, como en otras áreas de la estadística, problemas de inferencia. Estos últimos intentan, con carácter general, establecer relaciones sobre toda la población a partir de las relaciones detectadas en la muestra. Las observaciones realizadas sobre el item *i* se presentan mediante un vector $x_i=(x_{i1},...,x_{ip})^t$ de dimensión idéntica al número de observaciones registradas, el cual identifica un punto de $\boldsymbol{R}^p$. El calificativo *multivariante* es sinónimo a *multidimensional*, reflejando que el número *p* de observaciones registradas sobre cada item puede ser múltiple, a diferencia del caso univariante donde necesariamente $p=1$. Muy habitualmente, un aspecto relevante a considerar es la dependencia existente entre las distintas variables observadas. La dependencia entre dos cualesquiera de ellas se sintetiza a través de su covarianza, bien global o bien normalizada en unidades del producto de las desviaciones típicas de las dos variables en cuestión, la cual determina el *coeficiente de correlación simple*. Frente a observaciones múltiples, los estadísticos sumariales más relevantes son el vector de medias y la matriz de varianzas/covarianzas, los cuales sintetizan la centralización y la dispersión (junto a las dependencias binarias entre variables). Adicionalmente, la medida y el análisis de la dependencia entre una variable y un conjunto de éstas o entre dos conjuntos de variables constituyen dos elementos fundamentales en el análisis de los datos. En este sentido, el coeficiente de *correlación múltiple* extiende el *simple* entre dos variables, para medir la dependencia entre una variable y un conjunto de éstas; las *correlaciones canónicas*, por su parte, hacen lo propio entre dos conjuntos de variables.

Con carácter general, las técnicas clásicas de análisis multivariente son útiles para: (a) Descubrir constancias o regularidades en el comportamiento de las variables observadas; y (b) Contrastar modelos de asociación bien entre distintas variables, o bien entre (grupos de) items u objetos, incluyendo, en este último caso, cómo y en qué medida dos o más grupos difieren en sus respectivos perfiles multivariantes. El primer punto es típico de las llamadas investigaciones *exploratorias*. El segundo, por su parte, presenta un marcado cariz *confirmatorio*. En conjunto, las citadas ***técnicas de análisis multivariante constituyen un conjunto de procedimientos de análisis de datos que esencialmente busca relaciones o asociaciones entre variables o entre items u objetos.*** En este sentido, la aplicabilidad de tales herramientas o procedimientos no está confinada a una disciplina específica sino, por el contrario, a una gran diversidad de campos o áreas. Tradicionalmente, estos último incluyen las ciencias de conducta (más concretamente, la psicología experimental), la biología, la medicina y, en general, las ciencias de la salud. Hoy día, sin embargo, su aplicación se ha extendido a prácticamente todas las áreas del conocimiento, en especial a la economía, la gestión, la demografía, el marketing o el análisis de mercados, la sociología, etc.

Adicionalmente a su gran difusión de uso, el conjunto de técnicas de análisis multivariante se ha ido incrementando, tanto en variedad como en sofisticación. Si en el pasado el interés estuvo muy sesgado hacia la *regresión* y, en menor medida, hacia el *análisis factorial*, hoy día las restantes técnicas (e.g., *correlaciones canónicas*, *análisis discriminante*, *modelos de respuesta discreta*, *análisis de cluster* o *de conglomerados*, *multidimensional scaling*, etc.) se encuentran a un nivel muy similar de interés. Por otra parte, tanto los paquetes informáticos integrados de análisis de datos como las rutinas existentes en bibliotecas de programas se han ido extendiendo enormemente, al tiempo que se han incorporado versiones rápidas con objeto de minorar, en la medida de lo posible, los tiempos de respuesta ante ficheros masivos de datos. Finalmente, algunas de las técnicas que originalmente fueron desarrolladas para el tratamiento de datos continuos se han extendido y/o modificado para admitir el tratamiento de rangos y/o datos categóricos.

2. CLASIFICACIÓN DE LAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE

El conjunto de técnicas de análisis multivariante es muy extenso. Por este motivo, resulta conveniente emplear un cierto tiempo en caracterizar su taxonomía. El punto clave a tener en cuenta en la clasificación de las citadas técnicas es la propia matriz de datos $X=(x_{ij})$, cuyo sentido se refleja en la Tabla 1.

Por lo general, el conjunto de las *p* variables observadas sobre cada item u objeto suele dividirse en dos categorías o subconjuntos. La primera categoría está formada por las llamadas variables *dependientes* o *criterio*, las cuales se intentan explicar, relacionar o asociar con las contenidas en la segunda categoría,

Tabla 1. *Matriz de datos*

Objetos o items	*Variables*				
	1	...	*j*	...	*p*
1	x_{11}	...	x_{1j}	...	x_{1p}
.	.	...	.	...	.
i	x_{i1}	...	x_{ij}	...	x_{ip}
.	.	...	.	...	.
n	x_n	...	x_{nj}	...	x_n

llamadas por este motivo variables *explicativas, predictoras* o *independientes*. Cuando se analiza una matriz de datos, muy habitualmente se intenta explicar, en algún sentido, la variación de un grupo de variables en términos de su *covariación* con otro grupo. Tal explicación se lleva a cabo de acuerdo con alguno(s) de los criterios siguientes: (a) ***Determinar la naturaleza y/o el grado de asociación entre un conjunto de variables criterio o dependientes con un conjunto de variables independientes o predictoras***; esto ocurre, por ejemplo, con las técnicas de *correlaciones canónicas*, *componentes principales* y *análisis discriminante*; (b) ***Encontrar una función o fórmula con la cual se puedan estimar los valores de la(s) variable(s) criterio a partir de las variables predictoras;*** aunque esto es típico de los métodos de regresión, aparece también en otros contextos (e.g., *discriminación*, *modelos de respuesta discreta* tales como *probit*, *logit*, etc.); y (c) ***Valorar la confianza esta-dística en los resultados obtenidos con respecto a los dos puntos anteriores***, mediante contrastes de hipótesis o intervalos de confianza de los parámetros estimados. Por último, las técnicas de análisis multivariante pueden ser encuadradas en términos de las dos siguientes preguntas básicas: *¿El interés esencial del estudio está centrado en los objetos o items de observación o, por el contrario, se centra sobre las variables de la matriz de datos?, ¿las variables de la matriz de datos han sido divididas previamente por el experimentador entre variables criterio o dependientes y explicativas o independientes?*

Bajo el esquema dual previamente citado, es posible clasificar el conjunto de técnicas clásicas de análisis multivariante en cuatro subdivisiones esenciales de interés:

1. **Asociación entre una variable independiente única y múltiples variables predictoras**. Incluye las técnicas de *regresión*, de *análisis de la varianza* y *de la covarianza* y, finalmente, de *análisis discriminante* y *de respuesta discreta* (binario, en ambos casos).
2. **Asociación entre variables criterio múltiples y variables independientes igualmente múltiples.** Este apartado adscribe las técnicas de *correlaciones canónicas*, *análisis multivariante de la varianza* y *de la covarianza* y *técnicas de discriminación* en grupos múltiples.
3. **Análisis de interdependencia entre variables y/o de reducción de dimensión**, el cual incluye *análisis factorial*, *componentes principales*, *multidimensional scaling* y otros métodos alternativos de reducción de la dimensión.
4. **Análisis de similaridad inter-objetos o inter-items.** Esencialmente, este apartado engloba las técnicas de *análisis cluster*, así como otros tipos de procedimientos de agrupación de items u objetos.

Los dos primeros puntos involucran estructuras de dependencia en donde la matriz de datos se encuentra particionada en sendos subconjuntos de variables criterio o dependientes y variables explicativas o independientes; adicionalmente, en ambos puntos el interés se centra sobre las variables, constituyendo los objetos o items simples replicaciones de aquéllas. Por el contrario, los puntos 3 y 4 se refieren a *interdependencias* mutuas (en lugar, de dependencias de las variables criterio de las explicativas), las cuales afectan bien a variables o bien a objetos. En los cuatro apartados citados, existen técnicas específicas cuya diferenciación estriba, fundamentalmente, en la existencia o no de variables categóricas o cualitativas. Tradicionalmente, los métodos multivariantes han considerado dos tipos esenciales de variables. Por un lado, las variables numéricas, bien continuas (e.g., longitud, peso, altura, etc.), cuyas medidas varían sobre un cierto intervalo, o bien discretas 2(e.g., edad en años, número de hijos, etc.), donde se vulnera la propiedad de que los valores posibles llenan un cierto intervalo. Por otro lado, existen las variables cualitativas (sin sentido numérico, salvo código); entre estas últimas, las más sencillas son las variables binarias (o *dummy*, en notación sajona), que toman exclusivamente los valores cero y uno (e.g., sexo –hembra≡0, varón≡1–; ¿fumador? –sí≡1, no≡0–; etc.). A través de conjuntos de variables binarias se pueden representar todas las restantes variables cualitativas, las cuales pueden tener un número de categorías (exhaustivas y excluyentes) mayor que dos. Por poner un ejemplo, la ocupación laboral de una persona activa, clasificada en cinco categorías, podría codificarse mediante las cuatro variables *dummy* que figuran en la Tabla 2. Es evidente que la anterior

Tabla 2. *Codificación de la ocupación laboral*

	Variable dummy o binaria			
Ocupación	1ª	2ª	3ª	4ª
Profesional	1	0	0	0
Trabajador especializado	0	1	0	0
Trabajador no especializado	0	0	1	0
Religioso	0	0	0	1
Otros	0	0	0	0

variable *ocupación laboral* podría codificarse mediante una única variable cualitativa tomando cinco valores, por ejemplo, 0, 1, 2, 3 y 4 o, alternativamente, 1, 2, 3, 4 y 5. Pese a ello, la utilización de variables dummy resulta muy conveniente en muchas de las técnicas multivariantes, puesto que normaliza las codificaciones empleadas en términos de rangos y evita, además, la componente subjetiva que subyace en la elección de una opción concreta entre las distintas posibles codificaciones que pudieran emplearse.

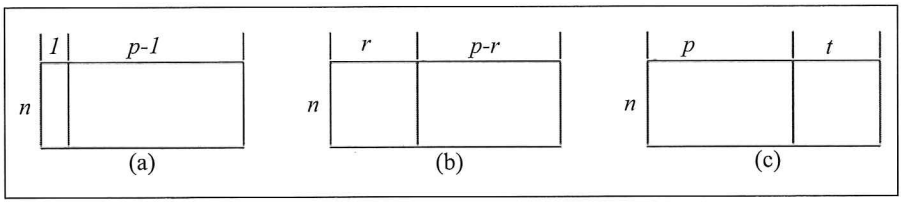


Figura 1. *Distintas particiones de la matriz de datos.*

A continuación se comentarán someramente cada una de las cuatro subdivisiones de las técnicas multivariantes citadas. Puesto que algunas de ellas visualizan la matriz de datos en forma particionada, la Figura 1 será de utilidad posterior.

2.1 Asociación entre una variable dependiente única y múltiples variables predictoras

En este caso, la matriz de datos $X=(x_{ij})$ se encuentra particionada en la forma que refleja la Figura 1(a). La primera columna incluye las medidas de la única variable *criterio* o *dependiente*, mientras que en las restantes columnas se incorporan los valores de las variables *predictoras* o *independientes*. Por ejemplo, la variable dependiente podría ser el consumo semanal de cerveza de cada individuo observado *i*, mientras que las variables dependientes podrían ser el salario, la edad (en años), el nivel de educación, el sexo, etc. En este caso, la variable dependiente es numérica y continua. Por su parte, las variables dependientes pueden ser bien numéricas o bien categóricas (convenientemente codificadas mediante distintos conjuntos de variables *binarias* o *dummy*). El caso citado, podría constituir un caso típico de *regresión lineal múltiple* (en el número de variables *predictoras*), donde el consumo de cerveza se intenta determinar a través de los valores de las restantes variables registradas.

Alternativamente a la situación antes descrita, la variable *criterio* única pudiera ser cualitativa (necesariamente binaria, al codificarse en términos *dummy*). Por ejemplo, se podría haber dividido la población bajo estudio en dos segmentos, incluyendo, respectivamente, los bebedores ligeros de cerveza (por debajo de un determinado nivel de consumo) y los bebedores consumados (por encima del citado nivel). Bajo este contexto nuevo, se podría intentar clasificar un individuo bien como bebedor ligero o bien como bebedor consumado, a partir de los valores observados de las restantes variables independientes. Para abordar este reto, podría emplearse un *análisis discriminante* (con dos únicos grupos) o bien desarrollar un *modelo de respuesta discreta* (por ejemplo, *probit* o *logit*).

Otra posibilidad podría ser que, como en el primer caso, la variable *criterio* fuera el consumo semanal de cerveza (variable numérica), mientras que las *independientes* fueran todas binarias, resultado de haber codificado de manera normalizada (mediante un conjunto de variables dummy, como se sabe) una determinada variable categórica con distintos valores (por ejemplo, la anteriormente citada *ocupación laboral del consumidor*). En este caso, se podría emplear la técnica de *análisis de la varianza* para analizar si existen diferencias significativas en los consumos de cerveza entre las distintas ocupaciones laborales.

Finalmente, si el conjunto de variables *predictoras* incluyera, además de la variable cualitativa *ocupación laboral*, la cuantitativa *salario anual* del consumidor, podría preguntarse si el consumo semanal de cerveza difiere entre las distintas ocupaciones laborales una vez que se ha controlado el efecto derivado del salario anual. En este caso, lo conveniente sería emplear un *análisis de la covarianza.*

2.2 Asociación entre variables dependientes e independientes múltiples

Aquí se asume que la matriz de datos $X = (x_{ij})$ se puede particionar como se refleja en la Figura 1(b). El número *r* de variables *dependientes* es mayor que la unidad. Por ejemplo, si $r = 3$, las variables *criterio* podrían ser, en una primera opción, los gastos mensuales en cine, teatro y espectáculos musicales (todas cuantitativas), mientras que las restantes variables *explicativas* pudieran ser determinadas características económicas y/o demográficas. En estas condiciones, uno podría estar interesado en encontrar una medida de asociación lineal entre las dos baterías de variables. Se podría emplear para ello la técnica multivariante conocida como *correlaciones canónicas.*

Alternativamente, supóngase que cada individuo ha sido previamente clasificado en uno de los cuatro grupos siguientes: (a) espectadores de cine, pero no asistentes a representaciones teatrales ni espectáculos musicales; (b) asistentes a representaciones teatrales o espectáculos musicales, pero no a salas cinematográficas; (c) asistentes a los tres tipos de espéctaculos; y (d) no asistentes a ninguno de ellos. Como antes, la variable cualitativa que refleja los cuatro grupos anteriores podría codificarse mediante tres variables dummy ($r = 3$). En estas circunstancias, se podría intentar asignar, de la mejor forma posible (en algún sentido previamente fijado), cada individuo a uno de los anteriores grupos (a), (b), (c) o (d), a partir de sus valores observados de las restantes variables económicas y/o demográficas. Este ejemplo refleja una situación típica que puede abordarse mediante el *análisis discriminante* (o los *modelos de respuesta discreta*) multigrupo.

Otra posibilidad podría ser haber registrado (en continuo) los gastos semanales en cine, teatro y espectáculos músicales, mientras que las variables *predictoras* fueran el conjunto antes indicado de variables *dummy* que permite codificar la variable cualitativa *ocupación laboral* del individuo. En este caso, el *análisis multivariante de la varianza* podría ser el procedimiento adecuado de análisis de los datos registrados. Si además de la ocupación laboral se dispusiera del *salario anual* de

cada individuo y uno se preguntara si, controlado el efecto del salario, los gastos citados difieren de una ocupación laboral a otra, se podría utilizar un *análisis multivariante de la covarianza* como herramienta de análisis.

2.3 Métodos de reducción de dimensión y/o factores latentes

Supóngase que se añade a la matriz de datos un conjunto de t ($\leq n$) variables adicionales, siendo cada una de ellas una combinación lineal de las p variables originales observadas. Las nuevas t variables citadas podrían ser no observables (o latentes). La representación conjunta de todas ellas se refleja en la Figura 1(c). Si se intenta que la información asociativa/disociativa suministrada por las p variables originales sea suministrada en la mayor medida posible por las nuevas t variables (menores en número) se podrían emplear bien las técnicas de *análisis factorial* y/o *componentes factoriales*, bien el *multidimensional scaling*, o bien otros métodos de reducción de dimensión. Así, se puede representar el conjunto original de variables correladas como combinaciones lineales (o no lineales) de las componentes del conjunto de $t \leq n$ variables subyacentes o latentes que, como se ha indicado, mantienen la mayor parte posible de la información original disponible. Puede exigirse, además, que las componentes citadas se elijan de modo que se cumplan ciertas condiciones adicionales (por ejemplo, que sean mutuamente incorreladas, lo sean o no las p variables originales).

2.4 Similaridad inter-items

Hasta el momento presente, la atención se ha centrado sobre las columnas de la matriz de datos X, representando los individuos distintas realizaciones de las citadas variables. Supongamos ahora que las p columnas de la matriz de datos X representan los gastos medios de consumo familiares (*per cápita*) en distintos bienes de consumo (e.g., alimentación, vestido, transporte, vivienda, educación, ocio,...). Los perfiles (*per cápita*) de los consumos familiares podrían ser comparados entre sí y se podría intentar desarrollar una medida de similaridad inter-familias con respecto a los patrones de consumo existentes. Hecho esto, se podría además proceder a agrupar las distintas familias en grupos de hábitos similares. Obsérvese que, en este caso, las relaciones entre variables carecen de interés, ya que se intentan obtener las citadas medidas de similaridad inter-objetos. Por ejemplo, medidas de *similaridad global* o de *interagrupabilidad* entre objetos. Este enfoque es propio del *análisis cluster*, muy utilizado en campos económicos, biológicos, demográficos...

Como resumen de todo lo anteriormente citado, queda puesto de manifiesto que los distintos procedimientos englobados dentro del análisis multivariante responden a objetivos dispares, cuya taxonomía se justifica mejor por medio de los datos registrados que a través de su propia unidad conceptual (de facto, muy liviana).

3. TRES ESTUDIOS DE CASOS CONCRETOS

No pudiendo ser exhaustivos, se muestran a continuación tres estudios de casos donde se hace patente el interés (y disparidad) de algunas de las técnicas citadas anteriormente. Imagínese que una cierta empresa pretende analizar el absentismo de sus empleados. Sus registros de personal incluyen los siguientes datos referidos al año anterior al actual.

- $Y \equiv$ *Número de días faltados al trabajo.*
- $X_1 \equiv$ *Calificación del empleado en un test de aptitud (con valoraciones de 1 a 13, con 1 ≡ aptitud muy favorable y 13 ≡ aptitud extremadamente desfavorable).*
- $X_2 \equiv$ *Número de años que el trabajador lleva en la compañía.*

Los valores de las variables citadas se muestran en la Tabla 3, junto con su medias y desviaciones típicas. Como es lógico, el pequeño tamaño muestral (de únicamente 12 empleados) refleja el interés pedagógico de la situación presentada. Se han añadido a cada variable dos normalizaciones habitualmente usadas, reconocibles mediante los subíndices d y s. Son, respectivamente, las *desviaciones a la media* (subíndice d) y las *desviaciones estandarizadas* en unidades de desviación típica (subíndice s). Por ejemplo para la variable Y, los valores de Y_d e Y_s para el individuo i coinciden con

$$Y_{di} = Y_i - \overline{Y}, \qquad Y_{si} = \frac{Y_{di}}{s_Y},$$

siendo

$$\overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i \quad \text{y} \quad s_Y^2 = n^{-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

Tabla 3. *Datos registrados*

	Nº de días ausente			Nota de aptitud			Años en la empresa		
Empleado	Y	Y_d	Y_s	X_1	X_{d1}	X_{s1}	X_2	X_{d2}	X_{s2}
1	1	-5.25	-0.97	1	-5.25	-1.39	1	-3.92	-1.31
2	0	-6.25	-1.15	2	-4.25	-1.13	1	-3.92	-1.31
3	1	-5.25	-0.97	2	-4.25	-1.13	2	-2.92	-0.98
4	4	-2.25	-0.41	3	-3.25	-0.86	2	-2.92	-0.98
5	3	-3.25	-0.60	5	-1.25	-0.33	4	-0.92	-0.31
6	2	-4.25	-0.78	5	-1.25	-0.33	6	1.08	0.36
7	5	-1.25	-0.23	6	-0.25	-0.07	5	0.08	0.03
8	6	-0.25	-0.05	7	0.75	0.20	4	-0.92	-0.31
9	9	2.75	0.51	10	3.75	0.99	8	3.08	1.03
10	13	6.75	1.24	11	4.75	1.26	7	2.08	0.70
11	15	8.75	1.61	11	4.75	1.26	9	4.08	1.37
12	16	9.75	1.80	12	5.75	1.53	10	5.08	1.71
Media	6.25			6.25			4.92		
Desviación típica	5.43			3.77			2.98		

la media y la varianza de Y. Si se comparasen los respectivos gráficos de dispersión de los puntos (X_{1i}, Y_i), (X_{d1i}, Y_{di}) y (X_{s1i}, Y_{si}), podría comprobarse fácilmente que las desviaciones a la media simplemente modifican el origen del gráfico, manteniéndose la forma. Por el contrario, la *estandarización* puede llegar a cambiar la forma, puesto que supone un cambio de escala eventualmente distinto en cada eje.

A la vista de los datos contenidos en la Tabla 3, se pueden plantear las siguientes baterías de preguntas, las cuales responden a tres posibilidades distintas de análisis de los datos citados.

1. ¿Cómo responde la variable Y ante los cambios producidos sobre X_1 y X_2? ¿Puede encontrarse una función lineal que nos permita predecir los valores de Y a partir de los de X_1 y X_2? ¿Es muy elevada la dependencia de Y de las otras dos variables? Por último, ¿es significativa la dependencia parcial de Y de cada una de las variables X_1 y X_2?
2. ¿La variabilidad conjunta de las variables X_1 y X_2 puede explicarse por medio de una sola variable (combinación lineal de ambas)? En otros términos, ¿reflejan X_1 y X_2 que existe un único factor subyacente a ambas variables? ¿Qué parte de la variación total entre X_1 y X_2 se explica por medio de ese factor? Al respecto, obsérvese que si la citada variación fuera elevada, dicho factor único podría sustituir las dos variables originales, produciéndose una reducción en la dimensión de los datos.
3. Si la variable Y nos ha servido para clasificar el conjunto de empleados en tres grupos o niveles de absentismo (e.g.: bajo, medio y alto), ¿cómo se puede clasificar un individuo dentro de uno de dichos grupos conociendo los valores de X_1 y X_2? ¿Cúal es el nivel de aciertos en la asignación?

Como se verá, una buena parte de las técnicas de análisis multivariante (aunque no todas) se basan en la búsqueda de ciertas transformaciones lineales de las variables observadas cumpliendo alguna condición de optimalidad. Esta última se fija de acuerdo con el contexto de interés del problema a resolver. En los tres subapartados siguientes se muestran ciertas posibilidades de contestación de cada una de las baterías de preguntas antes citadas. En cada caso, el título del epígrafe referencia la técnica clásica de análisis multivariante empleada.

3.1 Regresión múltiple

Para responder al primer grupo de preguntas, se puede buscar una transformación lineal

$$\hat{Y} = a_0 + a_1 X_1 + a_2 X_2,$$

donde $\hat{Y}$ denota las predicciones de Y. Los errores cometidos en la predicción se corresponden con las diferencias $e=Y-\hat{Y}$. Así pues, parece natural buscar aquellos parámetros que minimicen la suma de los errores individuales al cuadrado (para que no se produzca una eventual compensación debido al signo). Esta condición de optimalidad se conoce como *principio de mínimos cuadrados*, originalmente propuesto por Legendre y Gauss a comienzos del siglo XIX. En resumen, los parámetros $a = (a_0, a_1, a_2)^t$ buscados son

$$a = \underset{b \varepsilon R^3}{arg\,min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - b_2 X_{2i})^2$$

Respecto a la pregunta sobre en qué medida se manifiesta el grado de dependencia (afín) que existe entre Y y las dos restantes variables X_1 y X_2, la respuesta se establece mediante el cómputo del coeficiente R^2, de correlación múltiple al cuadrado. R^2 puede expresarse simplemente mediante:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} \quad .$$

Es claro que si todos los errores individuales son nulos ($R^2=1$), $\hat{Y}$ predice exactamente la variable original Y, para cada item i. Por el contrario, si el uso de X_1 y X_2 no mejora el que cada Y_i sea estimado mediante el valor medio de Y, entonces $R^2=0$, lo cual significa que la dependencia lineal de Y con respecto a las restantes variables es nula.

Finalmente, la última pregunta dentro de la primer batería de ellas puede responderse asumiendo el modelo $Y_i = a_{0i} + a_1 X_{1i} + a_2 X_{2i} + e_i$, donde los términos de error son independiente e idénticamente distribuidas, siguiendo una distribución normal $N(0,\sigma)$. La citada última pregunta involucra un contraste de hipótesis bien para contrastar que $R_p = 0$, o bien para hacer lo propio con respecto a la hipótesis $\alpha_1 = \alpha_2 = 0$, donde R_p, α_1 y α_2 denotan respectivamente el coeficiente de correlación múltiple poblacional y los coeficientes de regresión poblacionales correspondientes a las dos variables explicativas.

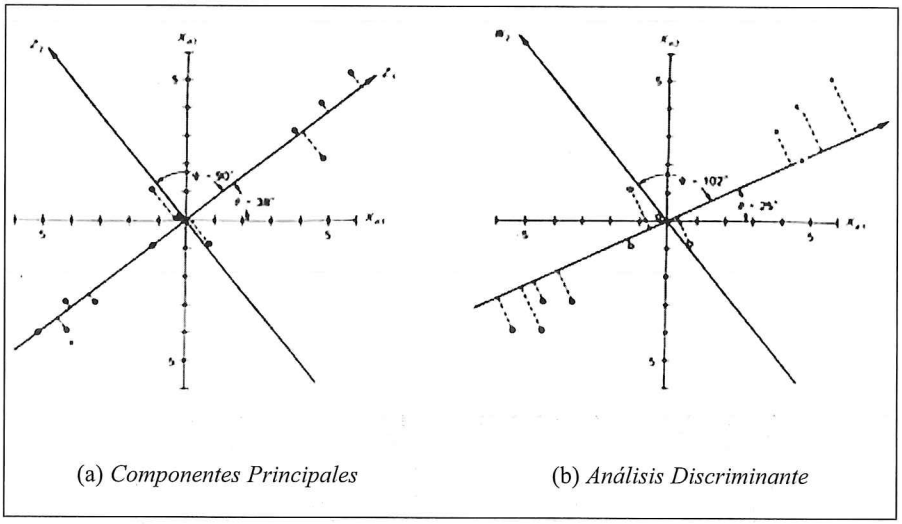
3.2 Análisis factorial

La segunda batería de preguntas está muy relacionada con las técnicas de análisis factorial y componentes principales, cuyas primeras referencias históricas se deben, respectivamente, a Spearman (1904) y Hotelling (1935, 1936). En dichas técnicas el interés se centra sobre la interdependencia entre conjuntos de variables y la posibilidad de representar las observaciones de los distintos items mediante un conjunto de variables de dimensión inferior al de las variables originales. Por poner un ejemplo, supongamos que se han graficado los valores de las variables X_{d1} y X_{d2} (véase la Figura 2(a)). En esta misma figura se representa un nuevo eje Z_1, que forma un ángulo de 38° con el eje horizontal. Supongamos ahora que se proyectan (perpendicularmente) los puntos observados (X_{d1i}, X_{d2i}) sobre el eje Z_1. Finalmente, manteniendo el mismo punto origen, representemos mediante z_{1i} la distancia al origen de la citada proyección del item i. La varianza del conjunto de estos últimos valores es

$$s_{z1}^2 = n^{-1} \sum_{i=1}^{n} \left(z_{1i} - \bar{z}_1\right)^2,$$

siendo $\bar{z}$ la media de la variable Z_1. La idea que subyace en este procedimiento es típica del análisis de componentes principales, representando un tipo de análisis factorial. De hecho, el citado eje Z_1 que forma un ángulo de 38° es aquél que tiene varianza máxima entre todos los ejes posibles. Dicho eje Z_1 óptimo (en el sentido mencionado previamente) puede representarse, como todos los ejes similares, mediante una determinada combinación lineal de las variables originales X_{d1} y X_{d2}, digamos $Z_1=\gamma_1 X_{d1}+\gamma_2 X_{d2}$.

Uno podría preguntarse en qué medida la variabilidad de las observaciones originales (X_{d1},X_{d2}) se mantendría si ambas variables se sustituyeran por la unidimensional Z_1. La contestación a esta pregunta es que, con los datos mostrados en la Tabla 3, Z_1 mantiene el 98% de la variabilidad de los datos originales. En resumen, la conclusión que se deriva de este hecho es que la disminución de la dimensión (de dos a uno) no redunda en una pérdida sustancial (de hecho, ésta es mínima) en la variabilidad de los datos originales. Para llegar a intuir el interés de la anterior reducción en la complejidad de los datos (en términos de dimensión), basta recordar que la citada variabilidad es el elemento esencial a la hora de encontrar similitudes y/o diferencias entre los individuos o items de observación, cuando éstos se identifican mediante sus respectivos valores observados. La reducción que supone, en los datos que se están considerando, pasar de las dos variables iniciales a una sola no es muy sustancial. Sin embargo, lo que merece la pena resaltar es que, cuando el número de variables originales es muy elevado (e.g., muy por encima de la capacidad computacional disponible), la reducción, en términos bien prácticos o bien de parsimonia, puede llegar a ser relevante. En este caso, el primer eje obtenido (que conserva la máxima variabilidad de las observaciones originales) se puede complementar con otros ejes adicionales. Se pueden exigir a estos últimos, condiciones de optimalidad similares a la del primer eje y, adicionalmente, otras más. Entre



(a) *Componentes Principales* (b) *Análisis Discriminante*

Figura 2. *Ejes relevantes en las obtenidos a partir de los datos de la Tabla 3.*

estas últimas, lo habitual es exigir condiciones de perpendicularidad con todos los ejes anteriores (lo cual, estadísticamente, se concreta en condiciones de independencia). En la Figura 2(a) está representado un segundo eje Z_2 (perpendicular al primero) con el cual se podría ampliar la variabilidadad de Z_1 bajo condiciones de independencia entre las nuevas variables. Si se considerara este segundo eje, la reducción en la dimensión no se haría efectiva con respecto a los datos bidimensionales originales, puesto que solo se manifestaría si se partiera de dimensiones más elevadas. Sin embargo, sí que se obtendría la independencia, otra propiedad que en muchas ocasiones resulta ser muy conveniente. Es de resaltar finalmente que, bajo condiciones de normalidad, los anteriores ejes o componentes principales Z_1 y Z_2 tienen una interpretación geométrica muy sencilla: coinciden con los ejes principales de los elipsoides que identifican las curvas de nivel de la función de densidad bivariante correspondiente a las variables originales (X_{d1},X_{d2}). Aunque esta propiedad se mantiene en dimensiones superiores a dos, no se incidirá en este punto con mayor detalle.

3.3 Análisis discriminante

Finalmente, la tercera batería de preguntas apunta directamente a distintos aspectos del *Análisis Discriminante*, técnica inicialmente propuesta por Fisher (1936). Aunque existen distintos procedimientos de discriminación (incluyendo los métodos *bayesianos*, los basados en *vecindades*, e incluso los *modelos de respuesta discreta –logit, probit,...–*), se centrará aquí la atención en la llamada *discriminación de Fisher*. Ésta involucra, como ocurre en los dos apartados anteriores, una transformación lineal de los datos originales, cumpliendo una determinada condición de optimalidad establecida en concordancia con la naturaleza del problema que se trata de resolver. Supongamos que, tras haber dividido en rangos los días de absentismo laboral contemplados en la Tabla 3, se han clasificado los distintos empleados en los tres grupos, clases o subpoblaciones siguientes: subpoblación Π_1 de *bajo* absentismo formada por los trabajadores {1,2,3,6}; y los grupos Π_2 = {4,5,7,8} y Π_3 = {9,10,11,12}, de absentismo *intermedio* y *alto*, respectivamente. El problema que se trata de abordar es cómo encontrar vías de clasificación de un individuo en uno de los tres anteriores grupos a partir de las restantes variables observadas. En lo que sigue, se supondrá que estas últimas son X_{d1} y X_{d2}, como en el apartado anterior. El planteamiento de Fisher ante este problema consiste en buscar, en primer lugar, aquella combinación lineal de las variables (X_{d1},X_{d2}), digamos $W_1=\omega_1 X_{d1}+\omega_2 X_{d2}$,

de modo que cuando se proyecten los puntos individuales (X_{d1i}, X_{d2i}) sobre el citado eje W_1 se maximice, tras la proyección, la varianza *entre-grupos* (es decir, entre las medias de los grupos proyectados) en unidades de varianza *dentro de los grupos* (es decir, la media ponderada de las varianzas de los grupos proyectados, con pesos de ponderación iguales a la importancia de cada grupo). Si solo existieran dos grupos, la idea de Fisher es clara: se intenta conseguir que las medias de los grupos (tras la proyección citada) se separen lo más posible (en unidades de las varianzas *dentro de los grupos*). De esta forma, se llega a determinar una primera regla de clasificación: *un individuo se asigna a un grupo determinado si, tras la proyección citada, la distancia entre el individuo y la media del grupo es la mínima entre todas las posibles*. La Figura 2(b) muestra que el citado primer eje óptimo W_1 forma un ángulo de 25° con el eje cartesiano horizontal X_{d1} (a diferencia de la primera componente principal del apartado anterior, para el cual, como se indicó, el ángulo era de 38°). Al igual que en el análisis factorial, pueden buscarse ejes adicionales de discriminación, si bien ahora la condición de perperdicularidad de las distintas transformadas no tiene sentido, puesto que el interés se centra en los grupos. En la Figura 2(b) se ha incluido el segundo eje W_2 de discriminación asociado a los datos considerados en este apartado, donde se observa la citada falta de ortogonalidad con el anterior eje W_1.

4. CONCLUSIONES

Las técnicas clásicas de *Análisis Multivariantes* engloban toda una serie de técnicas de análisis de datos con objetivos muy dispares. Puesto que la naturaleza del problema que se pretende abordar en cada caso ha motivado el desarrollo de las distintas técnicas citadas, se ha presentado en el apartado 2 una taxonomía donde se pivota a partir de la propia matriz de datos. Aunque no siempre, una buena parte de dichas técnicas involucran (en parsimonia) transformaciones lineales de los datos originales, admitiendo por ello una sencilla interpretación geométrica. Los ejemplos concretos comentados someramente en el apartado anterior corroboran estos dos últimos puntos, al tiempo que constituyen una muestra no exhaustiva, como es lógico, de los mencionados procedimientos clásicos de análisis de datos.

REFERENCIAS

- Anderson, T.W. *Introduction to Multivariate Statistical Analysis.* Wiley (1958).
- Arnold, S.F. *The Theory of Linear Models and Multivariate Analysis.* Wiley (1981).
- Bock, R.D. *Multivariate Statistical Methods in Behavioral Research.* McGraw-Hill (1975).
- Carroll, J.D. y Green, P.E. *Mathematical Tools for Applied Multivariate Analysis*. Academic Press (1997).
- Eaton, M.L. *Multivariate Statistics.* Wiley (1983).
- Fisher, R.A. The use of multiple measurement in taxonomic problems. *Annals of Eugenics*, **7**, 179-188 (1936).
- Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417-441, 498-520 (1933).
- Hotelling, H. The most predictible criterion. *Journal of Educational Psychology*, **26**, 139-142 (1935).
- Hotelling, H. Relations between two sets of variates. *Biometrika*, **28**, 321-377 (1936).
- Morrison, D.F. *Multivariate Statistical Methods.* McGraw-Hill (1976).
- Spearman, C. "General-intelligence objectively determined and measured". *American Journal of Psychology*, **15**, 201-293 (1904).
- Srivastava, M.S. y Khatri, C.G. *An Introduction to Multivariate Statistics.* North Holland (1979).

Teófilo Valdés
Dpto. de Estadística
Facultad de Matemáticas
Universidad Complutense de Madrid

Beowulf Superconductores caseros

INTRODUCCIÓN

To Beowulf now the glory was given, and Grendel thence death-sick his den in the dark moor sought, noisome abode: he knew too well that here was the last of life, and end of his days on earth.

Beowulf, Anónimo

Beowulf, guerrero escandinavo del siglo VI, es famoso por ser el único héroe capaz de derrotar al aterrador monstruo Grendel. Su historia se relata en un manuscrito, del que sólo existe un ejemplar, que data de algo antes del siglo X. Este manuscrito constituye la pieza épica en Inglés Antiguo más antigua que se conserva [1, 2].

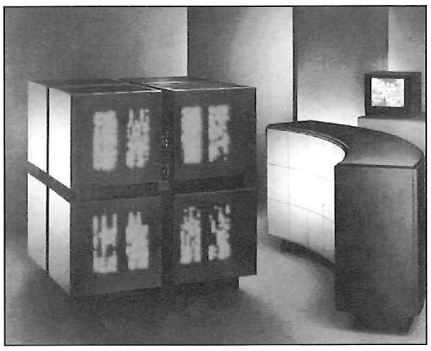


Figura 1. *Connection Machine.*

En el siglo XXI, los modernos guerreros de la computación desafían problemas que manejan ingentes cantidades de datos y que realizan cantidades astronómicas de cálculos. Para ello utilizan un computador de alto rendimiento construido con medios caseros y que recibe su nombre de este héroe épico: Beowulf.

La clave de Beowulf es el sistema operativo Linux [3], una versión libre de Unix que funciona en sistemas basados en procesadores Intel-compatibles, DEC Alpha o Power PC, entre otros. El código fuente de Linux se distribuye libremente, por lo que cualquiera puede mejorarlo, extenderlo o incluso contribuir a su desarrollo.