Análisis de Datos

Maura Vásquez y Guillermo Ramirez [*]

2012

[*]Escuela de Estadística y Ciencias Actuariales de la Universidad Central de Venezuela

Capítulo 1

Introducción al Análisis de Datos

1.1. Introducción

El presente curso tiene como objetivo principal introducir al estudiante en los conceptos esenciales del análisis multivariante de datos, específicamente en las técnicas de análisis de componentes principales (ACP) y de correspondencias binarias (ACB). Tendrá un enfoque que hace énfasis en los aspectos matemáticos de la teoría, y al mismo tiempo pretenderá que se obtenga una comprensión intuitiva de las diferentes técnicas, orientada fundamentalmente a resaltar aspectos de la interpretación de resultados

El análisis multivariante de datos está conformado por un conjunto de métodos y técnicas utilizadas en el estudio del comportamiento simultáneo de varias variables, que permiten obtener una visión de conjunto de fenómenos de la realidad cuya complejidad exige que sean estudiados con técnicas de mayor alcance que las de la estadística univariante o bivariante. Su objetivo fundamental es resumir y sintetizar la información contenida en grandes conjuntos de datos, con el fin de lograr una mayor comprensión del fenómeno en estudio.

Suele utilizarse el término "multivariante"(del inglés multivariate) para destacar el hecho de que se analizan simultáneamente varias variables, y se recurre al término como sinónimo de multivariable y multivariado. Es

conveniente aclarar sin embargo, que lo que fundamentalmente caracteriza al análisis multivariante es el estudio del comportamiento conjunto de las variables y de sus interrelaciones, y no la multiplicidad de ellas.

Hasta época relativamente reciente, los métodos multivariantes habían permanecido en el campo meramente teórico. Actualmente, con el uso de los potentes equipos de computación, estos métodos son utilizados en la mayora de las investigaciones cientficas, habiéndose comprobado ampliamente su eficacia en el tratamiento de grandes masas de datos. Precisamente, la expresión *análisis de datos* surge en la década de los 60 con la intención de distinguirlo del *análisis multivariante clásico*, basado en modelos y supuestos teóricos, y de enfatizar la idea de su utilidad para la descripción y análisis de grandes masas de datos.

Hoy por hoy se reconoce y aprecia la importancia de los métodos estadísticos en todas las esferas del quehacer científico, hasta el punto de que es utilizada incluso en disciplinas tales como historia, literatura y linguística, en las cuales la idea de realizar estudios cuantitativos era inconcebible hasta hace unos pocos años. La potencialidad de los métodos multivariantes puede apreciarse con mayor claridad si se presenta un inventario, obviamente breve e incompleto, de algunas de sus aplicaciones.

El análisis de componentes principales ha sido utilizado exitosamente en:

i La antropología física, para construir índices que capturen diferentes aspectos de la variabilidad biológica en los patrones de distribución de la grasa corporal en el ser humano, con el objeto de evaluar factores de riesgo de enfermedades metabólicas y endocrinas (Mueller y otros, 1986), (Vasquez y Perez, 1991).

ii La educación, para determinar los principales factores que permitan explicar la variabilidad que se observa en el desempeño académico de los estudiantes, mediante la construcción de medidas de habilidad académica que capten diferentes aspectos de esa variabilidad (Bartholomew, 2002).

iii La industria, para tratar de controlar las variables que limitan la optimización de los procesos, identificando las variables responsables de las

principales diferencias entre los productos obtenidos en un proceso de producción (Martens y Martens, 2001).

El análisis de correspondencias binarias ha sido utilizado exitosamente en:

i Las ciencias sociales, para obtener mediciones aproximadas de la pobreza. A estos efectos, las estadísticas de calidad de vida proporcionan indicadores que describen aspectos parciales de esta problemática, los cuales están directa y fuertemente asociados entre sí, y cuya escala de medición es ordinal. Con el ACB se puede construir una única medida en escala de intervalo o de razón, que resuma los principales características de los indicadores parciales (Camardiel y otros, 2000).

ii La medicina, para analizar la asociación entre índices clinimétricos que ordenan pacientes de acuerdo con el grado de severidad de la afección renal, y otros indicadores de presencia/ausencia de anticuerpos antinucleares. Este análisis permite apoyar la toma de decisiones clínicas para diagnóstico, pronóstico, selección de un tratamiento, y/o evaluación de los beneficios de una terapia (Tapanes y otros, 2000).

iii El análisis de mercados, para construir mapas perceptuales que describan el perfil de preferencias de los consumidores por diferentes marcas de productos en asociación con posibles atributos que los identifican (Levy y Varela, 2005).

1.2. Reseña Histórica

El análisis multivariante se apoya fundamentalmente en conceptos matemáticos originalmente formulados en el siglo XIX por el matemático italiano Eugenio Beltrami (1835-1900), y en forma independiente por el francés Camille Jordan (1838-1922), quienes desarrollaron la base algebraica de la factorización de una matriz en sus valores y vectores singulares (DVS).

Los primeros estudios estadísticos formales que podrían calificarse como multivariantes, se refieren a las generalizaciones de los análisis de correlación y regresión realizados a principios de siglo por Francis Galton (1822-1911), Karl Pearson (1857-1936) y Charles Spearman (1863-1945), científicos ingleses del área de la psicología y la biometría.

A Galton se debe la introducción del método de regresión y de las primeras ideas sobre correlación, conceptos que surgen en sus investigaciones sobre la tendencia de la talla de los seres humanos hacia la estatura promedio de la población a la cual pertenecen. Por su parte Pearson, eminente estadístico y director del Laboratorio de Biometría de la Universidad de Londres, hace uno de sus más importantes aportes a la Estadística al proponer el contraste chi-cuadrado de independencia y la expresión analítica del coeficiente de correlación muestral, diseñado para estudiar las relaciones lineales entre variables, dos a dos. Posteriormente presenta la primera medida de distancia multivariante conocida como coeficiente de parecido racial. Consciente de la necesidad de superar esta perspectiva limitada de análisis de relaciones entre variables, el mismo autor desarrolla en 1901 los fundamentos geométricos del análisis de componentes principales. Fue Spearman quien, para estimar la inteligencia en niños británicos, desarrolló el primer modelo de análisis de factores (AF), en el cual se postula que los resultados de cualquier test sicométrico se pueden expresar como una combinación lineal de un factor común a todas las pruebas que incluye el test y de un factor específico para cada prueba.

Los trabajos desarrollados posteriormente por Ronald Fisher (1890-1962) incorporan formalmente el lenguaje algebraico y el punto de vista geométrico a algunas distribuciones probabilísticas, al análisis de la varianza, al diseño de experimentos y al análisis discriminante. En particular la ley de distribución normal, que surgió anteriormente con los trabajos de Abraham De Moivre, Pierre-Simon Laplace y Carl Friedrich Gauss en el siglo XVIII, adquiere forma bivariante a finales del siglo XIX con los trabajos de Galton y Pearson, deviniendo en multivariante con las investigaciones de Fisher. Estos tres grandes maestros de la Estadística hacen importantes aplicaciones en las áreas de Antropometría, Genética, Agricultura y Biometría.

La década del 30 se caracterizó por ser un período de grandes contribuciones a la estadística multivariante. En Estados Unidos destacan por sus importantes contribuciones Harold Hotelling (1885-1973), Samuel Wilks y M. Bartlett. En particular Hotelling plantea el problema del ACP como un procedimiento de reducción de variables, estableciendo que es posible construir un conjunto de nuevas y pocas variables incorrelacionadas, denominadas componentes, que logran resumir la información contenida en las variables originales. Simultáneamente se desarrolla en la India un movimiento que

hace aportes fundamentales a los métodos multivariantes, iniciado por Prasanta Mahalanobis y Samarendra Roy, y posteriormente profundizado por Calyampudi Rao y Paruchuri Krisnaiah. En 1939 Hotelling discute una interpretación geométrica del ACP en términos de elipsoides de concentración de una distribución normal multivariante. Por esa misma época, sus aportes son complementados con 4 artículos fundamentales de M. Girschik, Fisher, P. Hsu y Roy sobre la distribución probabilística de los valores propios de la matriz de varianzas y covarianzas de una muestra procedente de una población normal multivariante. Estas ideas son ampliamente desarrolladas en los textos clásicos de Theodore Anderson (1958) y Maurice Kendall y Alan Stuart (1969).

Louis Thurstone en 1930 reformula el AF proponiendo un modelo con varios factores comunes e imprimiéndole un sentido geométrico al mismo. Además de los desarrollos del AF y del ACP en la década del 30, surge el análisis discriminante introducido por Fisher. La función lineal discriminante de Fisher se relaciona con la T^2 de Hotelling introducida por este autor en 1931, así como con la distancia D^2 de Mahalanobis. El análisis canónico, que constituye una generalización de la correlación múltiple a dos conjuntos de variables, es propuesto por Hotelling en 1935. Más tarde Paul Horst, James Carroll y Jon Kettenring extienden este enfoque a varios conjuntos de variables, surgiendo así lo que se conoce como análisis multicanónico.

En 1936, Carl Eckart y Gale Young publican un trabajo que resulta de fundamental importancia en el desarrollo de las técnicas multivariantes. En este artículo se presenta la teoría de aproximación de matrices, basada en la descomposición de una matriz en sus valores y vectores singulares, cuya álgebra y geometría constituyen el soporte matemático de la mayoría de las técnicas de análisis de datos.

El análisis de correspondencias binarias tiene su origen en el método de "reciprocal averaging"desarrollado inicialmente por Hans Hirchsfeld en 1935. Este método define un procedimiento de optimización para asignar puntuaciones a las modalidades de dos variables categóricas, que relaciona los vectores directores de los espacios de representación óptima de las dos variables mediante las relaciones de doble transición definidas por la DVS. De esta manera la puntuación asignada a la j-ésima modalidad de una de las variables es, salvo un coeficiente, una media ponderada de las puntuacio-

nes de la otra variable. Jean-Paul Benzecri presenta en 1969 el ACB desde una óptica geométrica y multidimensional, cercana a la que Pearson le imprimió al ACP. Este autor inicia los fructíferos trabajos de la denominada escuela francesa de análisis de datos, que posteriormente han sido continuados entre otros, por Ludovic Lebart, Alain Morineau y Jean-Pierre Fenelon.

En 1971 Karl Ruben Gabriel desarrolla los principios del biplot, técnica factorial que se diferencia de las anteriores en que garantiza la representación simultánea de los objetos de estudio y de sus atributos. En esta década se inicia la escuela sueca de análisis de datos, promovida fundamentalmente por Karl Jöreskog y Dag Sörbom.

A partir de los años 80 surge la escuela holandesa con los trabajos de John Van de Geer, Pieter Kroonenberg, Jan De Leeuw y el Grupo GIFI de la Universidad de Leiden, cuyas investigaciones se han centrado en el estudio y desarrollo de técnicas multivariantes aplicadas a datos categóricos.

1.3. Una clasificacion de los metodos multivariantes

Los procedimientos para abordar el conocimiento de los fenómenos reales son muy similares en todas las ramas del quehacer científico:

- Reconocer y formular con precisión el problema en cuestión.
- Recolectar y organizar datos relevantes producidos a través de experimentos, cuasiexperimentos, estudios observacionales o encuestas por muestreo.
- Efectuar un acercamiento sistemático a los datos, bien para su exploración o para la confirmación de hipótesis establecidas a priori.

Cuando hay tres o más variables involucradas en el problema, los métodos estadísticos multivariantes permiten analizar simultáneamente las interrelaciones que se producen entre ellas, aún cuando éstas se irán haciendo tanto más complejas cuanto mayor sea el número de variables a analizar. Si el interés del investigador consiste en estudiar la asociación entre dos conjuntos de variables, donde uno de ellos (variables independientes, explicativas o

predictoras) ayuda a predecir o a explicar el comportamiento del otro (variables dependientes, explicadas o respuestas), las técnicas apropiadas para el tratamiento de los datos corresponden a los métodos denominados de dependencia. En el caso en que el interés se centre en el estudio de las interrelaciones entre las variables, pero sin distinguir entre sus roles, se utilizan métodos de interdependencia, algunos de los cuales se conocen como métodos de reducción de la dimensión y otros como métodos de clasificación y escalamiento.

En el cuadro 1 se presenta una clasificación de los métodos multivariantes basada en el criterio de la existencia o no de variables dependientes, su número y naturaleza, y que además toma en cuenta la naturaleza de las variables independientes.

Para abordar el estudio del comportamiento de las variables y de sus interrelaciones, los métodos multivariantes consideran como elemento fundamental de análisis la variabilidad existente en los datos, buscando explicarla a través de las fuentes que la originan.

En el caso de los métodos de dependencia la variabilidad de la(s) variable(s) dependiente(s) es explicada por las independientes, que son variables observables. Usualmente esta explicación no es completa, y por ello se agrega un término de error que capta aquella parte de la variabilidad no recogida por las primeras. Por su parte, en los métodos de interdependencia, conocidos como de reducción, se asume que las interrelaciones entre un conjunto de variables observables pueden ser explicadas en términos de otro conjunto de variables no observables (componentes o factores). En este contexto las fuentes de variabilidad en los datos se atribuyen a estas últimas.

En relación con el análisis de conglomerados la información que constituye el input básico, está constituida por medidas de distancia o de semejanza entre los objetos de estudio, que son utilizadas para detectar patrones de agrupación conducentes a la formación de clases homogéneas de objetos, dando lugar a una partición de la variabilidad total de los datos en dos términos: el primero asociado con la variabilidad interna de los grupos y el segundo con la variabilidad entre ellos.

Algunos Métodos de Dependencia

Método	Dependiente	Independientes	Objetivo
Regresion Multiple	Cuantitativa (a explicar)	Cuantitativas (explicativas)	Describir el tipo, direccion y fuerza de la relacion entre una variable a explicar y un conjunto de variables explicativas.
Analisis de la Varianza	Cuantitativa (respuesta)	Cualitativas (factores)	Describir la relacion entre una variable respuesta y un conjunto de factores.
Analisis de la Covarianza	Cuantitativa (respuesta)	Cualit/Cuantit (factores y covariables)	Igual que el ANOVA, pero controlando por el efecto de una o varias covariables.
Regresion Logistica	Dicotomica (a explicar)	Cualit/Cuantit (explicativas)	Describir la relacion entre un conjunto de variables explicativas y la probabilidad de ocurrencia de un evento.

Algunos Métodos de Interdependencia

Método	Variables	Objetivo General
Analisis de Componentes Principales	Cuantitativas	Construir combinaciones lineales de las variables, con propiedades especiales de varianza y correlación, que retienen lo esencial de la informacion.
Analisis de Correspondencias	Cualiitativas	Explicar la asociación entre dos variables
Analisis de la Covarianza	Cuantitativa (respuesta)	Igual que el ANOVA, pero controlando por el efecto de una o varias covariables.
Analisis de Conglomerados	Cuantit./Cualit. (a explicar)	Agrupar objetos de acuerdo con las caracteristicas que poseen