Fourth LACCEI International Latin American and Caribbean Conference for Engineering and Technology (LACCET'2006)
"Breaking Frontiers and Barriers in Engineering: Educationl, Research and Practice"
21 – 23 June 2006, Mayagüez, Puerto Rico

.

ANÁLISIS MULTIVARIANTE: CLASIFICACIÓN, ORGANIZACIÓN Y VALIDACIÓN DE RESULTADOS.

Miriam M. Álvarez Suárez, Ph. D.
Profesor Titular del Centro de Investigaciones Avanzadas en Ingeniería Industrial,
Universidad Autónoma del Estado de Hidalgo, Hidalgo, México,
miriamsu@uaeh.reduaeh.mx

Amaury Caballero, Ph. D., P.E.
Assistant Profesor, Department of Construction Management, Florida International University, Florida, USA
caballer@fiu.edu

Gilberto Pérez Lechuga, Ph. D.
Profesor Titular del Centro de Investigaciones Avanzadas en Ingeniería Industrial,
Universidad Autónoma del Estado de Hidalgo, Hidalgo, México,
glechuga2004@hotmail.com

Resumen

Las técnicas estadísticas multivariantes son cada día más utilizadas en diferentes ramas de la ciencia. La ingeniería y la administración de empresas no es una excepción a esto. Los métodos exploratorios y los confirmatorios, que en la mayoría de las ocasiones se utilizan de forma combinada, requieren de un conocimiento previo del problema a estudiar y de la información con que se cuenta. El presente trabajo tiene como objetivo relacionar una serie de aspectos importantes para la aplicación de modelos multivariantes a diferentes problemas de investigación. Aunque el análisis multivariante tiene sus raíces en la estadística univariante y bivariante, la extensión al dominio multivariante introduce conceptos y cuestiones adicionales, que van desde el "valor teórico" hasta las escalas de medida utilizadas, los errores de medición, los resultados estadísticos de las pruebas de significación y los intervalos de confianza. La utilización de un modelo multivariante conlleva la elaboración de un plan de investigación bien definido que incluye los objetivos analíticos en términos conceptuales, la selección de la técnica, la evaluación de los supuestos básicos de dicha técnica, la estimación del modelo y su interpretación, para finalizar con la aplicación de las técnicas de validación para determinar la estabilidad de los resultados obtenidos.

Palabras clave: análisis multivariante, análisis previo, técnicas de validación

1. Introducción

Las tareas implícitas en el examen previo de los datos pueden parecer insignificantes y sin consecuencias a primera vista; no obstante, son una parte esencial del análisis multivariante. Si bien estas técnicas suponen un tremendo poder analítico en manos de cualquier investigador, y además, hay que asegurarse de que se mantengan las bases teóricas y estadísticas sobre las que éstas se sustentan (Hair et al., 1999).

En primer lugar, el investigador obtiene un conocimiento básico de los datos y las relaciones entre las variables. Las técnicas multivariantes plantean grandes demandas al analista en cuanto a la comprensión, interpretación y articulación de resultados basados en relaciones cuya complejidad puede llegar a ser muy grande. El conocimiento de algunas interrelaciones importantes o evidentes puede ayudar en la especificación y refinamiento del modelo multivariante a utilizar, así como proporcionar una perspectiva razonable para la interpretación de los resultados.

La naturaleza y distribución de las variables incluídas en el estudio, las representaciones de perfiles multivariantes para una observación, el examen de los datos ausentes y casos atípicos, y la solución a tomar en cada caso, así como la verificación de los supuestos de normalidad, linealidad y homocedasticidad incluyendo el estudio de las posibles transformaciones a realizar para resolver los problemas encontrados, son, de forma general, los aspectos a tener en cuenta antes de emprender la tarea de realizar un análisis multivariante .

Muchos autores han clasificado los métodos multivariantes (Lebart et al.,1981; Dagnelie, 1981; Hair et al.,1999) pero todos coinciden en que los tres aspectos más importantes a tener en cuenta sin orden de prioridad, son: la dependencia o no entre las variables, las escalas de medición utilizas para cada una de las ellas y el objetivo que se persigue en el estudio.

El análisis y la interpretación de cualquier técnica multivariante no conduce a una única respuesta, aunque puede ayudarse por un conjunto general de directrices, no exhaustivas, pero que representan una filosofía del análisis multivariante. Entre ellas podemos citar, la significación estadística y la significación práctica; la discusión de la relación de la potencia estadística con el tamaño muestral y con la significación estadística, procurar la parsimonia del modelo, el análisis de los errores de predicción no como una medida del error, sino como un punto de partida para diagnosticar la validez de los resultados obtenidos y como una indicación de las relaciones que quedan sin explicar, así como la validación de los resultados.

Al discutir las numerosas técnicas multivariantes a disposición del investigador y la gran cantidad de supuestos que implica su aplicación, se hace evidente que finalizar con éxito un análisis multivariante implica algo más que la selección del modelo correcto. Deben resolverse problemas que van desde la definición del problema hasta el diagnóstico crítico de los resultados. Sin intentar proporcionar un conjunto rígido de procedimientos a seguir, utilizaremos una aproximación al análisis multivariante en seis pasos, donde los tres primeros se refieren al análisis previo de los datos, el cuarto se refiere al análisis propiamente dicho y los dos últimos se refieren a la interpretación y posible generalización de los resultados obtenidos (Hair et al., 1999). Ellos son:

1- Definición del problema de investigación, objetivos y técnica multivariante conveniente,
2- Desarrollo del plan de análisis (tamaños de muestra mínimos, tipos de variables permitidas y métodos de estimación),
3- Evaluación de los supuestos básicos de la técnica propuesta,
4- Estimación del modelo multivariante y valoración del ajuste del modelo,
5- Interpretación del valor teórico, y

6- Validación del modelo multivariante.

El presente trabajo se realizó con el objetivo de organizar el trabajo a seguir para la aplicación de una técnica multivariante; contar con una clasificación previa de las técnicas y algunas consideraciones sobre las dos etapas olvidadas de la aplicación de una técnica multivariante: el análisis previo de los datos y la validación de los resultados para conocer su posible generalización.

2. Análisis previo de los datos

El análisis cuidadoso de los datos conduce a una mejor predicción y a una evaluación más precisa de la dimensionalidad. Para ello existen técnicas analíticas y técnicas gráficas que ofrecen al investigador un conjunto de formas simples de examinar, tanto las variables individuales, como las relaciones entre ellas. Más concretamente, se trata de la evaluación de datos faltantes, la identificación de casos atípicos, y la comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes.

Para ello hay que pasar por 4 fases del examen previo de los datos. Éstas incluyen:

- un examen gráfico de la naturaleza de las variables a analizar y sus distribuciones así como de las relaciones que forman las bases del análisis multivariante,
- un proceso de evaluación para entender el impacto que pueden tener los datos ausentes sobre el análisis,
- las técnicas que mejor se ajustan para la identificación de casos atípicos, y
- los métodos analíticos necesarios para evaluar adecuadamente la capacidad de los datos para cumplir los supuestos estadísticos específicos de muchas técnicas multivariantes.

2.1. Examen gráfico de los datos.

Como paso previo se hace necesario realizar un examen de cada una de las variables individualmente. Las técnicas de la Estadística Clásica; distribuciones de frecuencia, histogramas, diagramas de tallo y hojas, diagramas de caja y bigotes y el cálculo de estadígrafos e intervalos de confianza, así como gráficos de dispersión entre variables (para dos y tres dimensiones), nos permitirán tener una idea más clara y simple de los datos, sus distribuciones y sus relaciones.

Para el caso de más de tres variables se utilicen las representaciones gráficas multivariantes (Johnson, 2000) entre las que se encuentran: los *perfiles multivariantes*, que representa un diagrama de barras de todas las variables para cada observación; *los gráficos de rayos o estrellas*, que representan la distancia a la que se encuentra cada variable de cero sobre rayos o ejes que irradian de un punto central generándose un rayo para cada variable; las *representaciones icónicas*, siendo las más utilizada las caras (Chernof, 1973) y en la cual cada cara corresponde a un individuo y cada rasgo de la cara se corresponde con una variable; y por último, *las curvas de Andrews*, (Andrews, 1972) donde los diferentes parámetros de las curvas son las variables y que conlleva una transformación matemática de los datos originales en una relación que puede ser representada gráficamente. Aunque estas comparaciones para un valor único son más difíciles, esta forma de representación gráfica nos presenta en un solo gráfico una comparación generalizada y la agrupación de observaciones.

2.2. Datos ausentes.

Antes de que se pueda instrumentar cualquier solución para la ausencia de datos, el investigador debe diagnosticar los procesos de ausencia de datos que subyacen en este fenómeno. Algunas veces estos procesos se encuentran bajo el control del investigador y pueden ser identificados

explícitamente. En tales casos, la ausencia de casos se denomina "*prescindible*", lo que significa que no se necesitan soluciones específicas para la ausencia de datos dado que los límites de la ausencia de dichos datos son inherentes a la técnica usada.

Un ejemplo de datos ausentes prescindibles es aquel o aquellas observaciones de una población que no están incluidas en la muestra. La muestra probabilística permite al investigador especificar que los procesos de datos ausentes causantes de las observaciones omitidas son aleatorios y que dichos datos ausentes pueden explicarse como un error muestral en los procedimientos estadísticos. Otro caso de datos ausentes prescindible tiene lugar cuando los datos están censurados. Estos datos son observaciones incompletas como consecuencia de su etapa en el proceso de toma de datos. Un ejemplo típico es un análisis de las causas de fallecimiento.

La ausencia de datos puede ocurrir por otras muchas razones y en muchas situaciones. Estos datos ausentes pueden ser causados por errores en la introducción de datos, o problemas de su recolección, o también una no respuesta por parte del encuestado. Los primeros a veces se pueden resolver, pero los últimos no son tan sencillos. Entonces se hace necesario saber si estos datos ausentes están distribuidos aleatoriamente entre las observaciones o se pueden identificar algunas pautas, además de saber en qué medida son relevantes.

El impacto de los datos ausentes es perjudicial no sólo por sus sesgos potenciales sino también por su efecto en el tamaño de la muestra disponible para el análisis. Luego, para decidir si se puede aplicar una solución para dichos datos, el investigador debe averiguar el grado de aleatoriedad presente en ellos, ya que una consideración errónea sobre este aspecto introduciría un sesgo en los resultados. Para dicho diagnóstico existen tres métodos:

- valoración de los datos ausentes mediante una única variable **Y** formando dos grupos (uno con valores ausentes y otro con valores válidos de **Y**) y comparando ambos grupos; si la diferencia es significativa implica que existe un proceso de pérdida de datos no aleatorio,
- utilizando las correlaciones dicotomizadas para evaluar la correlación de los datos ausentes en cualquier par de valores. Para cada variable, se representa por 1 los valores válidos y por 0 los valores faltantes. Las correlaciones indican el grado de asociación entre los valores perdidos sobre cada par de variables. Bajas correlaciones implican aleatoriedad en el par de variables estudiada, y
- se puede hacer un test conjunto de aleatoriedad que determine si los datos ausentes pueden ser clasificados como "completamente aleatorios", analizando el patrón de datos ausentes sobre todas las variables y comparándolas con el patrón esperado para un proceso de datos ausentes aleatorio. Si las diferencias son no significativas, los datos ausentes pueden ser clasificados como "completamente aleatorios" y si son significativas, se debe utilizar alguno de las soluciones anteriores para identificar los procesos específicos de datos ausentes que no son aleatorios.

Las aproximaciones o soluciones que tratan con los datos ausentes están basadas en la aleatoriedad antes descrita. Si se encuentran procesos de datos ausentes "aleatorios" o no aleatorios, el investigador debe aplicar sólo el método diseñado específicamente para este proceso, ya que la aplicación de cualquier otro método, introduciría sesgos en los resultados (Little y Roderick,1987). Las soluciones que veremos a continuación sólo pueden utilizarse si el investigador determina que el proceso de ausencia de datos puede clasificarse como "completamente aleatorio". Estas son:
:

- Utilizar, si es posible, aquellas observaciones con datos completos.
- Suprimir el caso y/o la variable que peor se comporten con respecto a los datos ausentes,..
- Estimación de valores ausentes basado en valores válidos de otras variables y/o casos de la muestra..

- Utilizar otras técnicas de imputación (pairwise en el SPSS)
- Sustitución de caso (media, valor constante, por regresión u otro).

2.3. Casos atípicos.

Los casos atípicos pueden identificarse desde una perspectiva univariante, bivariante o multivariante. El investigador debe utilizar cuantas perspectivas sean posibles, para buscar una consistencia entre los métodos de identificación de casos atípicos.

No obstante, el detectar los casos atípicos no implica su eliminación inmediata. Una vez identificados y especificados, no se deben eliminar a menos que exista una prueba demostrable de que son verdaderas aberraciones y no son representativos de las observaciones de la población. Pero si representan a un segmento de la población, deben retenerse para asegurar su generalidad al conjunto de la población. Si se eliminan los casos atípicos, el investigador corre el riesgo de mejorar el análisis pero limitar su generalidad. Si los casos atípicos son problemáticos en una técnica particular, muchas veces pueden ser manejados de una forma tal que se ajusten al análisis sin que lo distorsionen significativamente.

2.3.1. Detección univariante.

Esta perspectiva se basa en el examen de la distribución de las observaciones, seleccionando como casos atípicos aquellos que caigan fuera de los rangos de la distribución utilizando para ello un diagrama de caja (boxplot) y el cálculo de la variable "z" (ó z-score). La cuestión principal consiste en establecer el umbral para la designación como caso atípico. El enfoque típico convierte los valores de los datos en valores estandarizados, con media cero y desviación estándar igual a uno. Para menos de 80 muestras, las pautas sugeridas identifican como casos atípicos aquellos con valores estándar mayores o iguales a 2.5 y cuando las muestras son mayores, el valor umbral del estandarizado se sitúa entre 3 y 4.

2.3.2. Detección bivariante.

Además de la evaluación univariante, pueden evaluarse conjuntamente pares de variables mediante un gráfico de dispersión. Casos que caigan fuera del rango del resto de las observaciones, pueden identificarse como puntos aislados en el gráfico de dispersión. Para ayudar a identificar el rango esperado de las observaciones, se puede superponer sobre el gráfico de dispersión, una elipse que represente un intervalo de confianza especificado (entre el 50 y 90% de la distribución) para una distribución normal bivariante. Esto proporciona una representación gráfica de los límites de confianza y facilita la identificación de casos atípicos.

2.3.3. Detección multivariante.

La medida D^2 de Mahalanobis puede usarse para tener una forma objetiva de medición de la posición multidimensional de cada observación relativa a un punto común. Es decir, proporciona una medida común de centralidad multidimensional y además tiene propiedades estadísticas que tienen en cuenta las pruebas de significación. Dada la naturaleza de las pruebas estadísticas, se sugiere un nivel muy conservador (0.001) como valor umbral para la designación como caso atípico.

2.4. Verificación de los supuestos del análisis multivariante.

La complejidad de las relaciones en el análisis multivariante aumenta la necesidad de comprobar los supuestos estadísticos, ya que la gran cantidad de variables hace que las distorsiones y los sesgos potenciales sean más potentes cuando se incumplen éstos. Los supuestos fundamentales que hay que corroborar son los siguientes: normalidad, homocedasticidad, linealidad y ausencia de errores correlacionados.

2.4.1. Normalidad.

El test más simple para diagnosticar la normalidad es una comprobación visual del histograma que compare los valores de los datos observados con una distribución aproximada a la distribución normal. Además de examinar el gráfico, se pueden examinar los valores de la simetría y la curtosis y los tests estadísticos específicos como el Shapiro-Wilks y el de Kolmogorov-Smirnov, que aparecen en muchos programas computacionales. La forma de corregir la normalidad es transformando las variables.

2.4.2. Homocedasticidad.

La homocedasticidad se refiere al supuesto de que las variables dependientes tengan iguales varianzas a lo largo del rango del predictor de las variables. La prueba de igualdad de varianzas entre dos variables métricas se puede realizar gráficamente y estadísticamente.

La aplicación más común de la evaluación gráfica se realiza a partir de un análisis de regresión múltiple. Dado que el eje del análisis de regresión es el valor teórico, el gráfico de residuos se usa para revelar la presencia de homocedasticidad. Los tests estadísticos de igualdad de varianzas se refieren a la varianza en grupos formados por variables métricas. El test más común es el de Levene, que se utiliza para evaluar si las varianzas de una única variable métrica son iguales a lo largo de cualquier cantidad de grupos. Si se compara más de una variable métrica, implicando la igualdad de las matrices de varianzas y covarianzas, se aplica el test M de Box.

La forma de corregir esta situación es a través de la transformación de datos, similares a las usadas para conseguir la normalidad, ya que en muchos casos, la heterocedasticidad es el resultado de la no normalidad de una de las variables y la corrección de la normalidad, resuelve igualmente la dispersión de la varianza.

2.4.3. Linealidad.

La linealidad es un supuesto implícito de todas las técnicas multivariantes basadas en medidas de correlación, incluyendo la regresión múltiple, la logística, el análisis factorial y los modelos de ecuaciones estructurales. La forma más común de evaluar la linealidad es examinar los gráficos de dispersión de las variables e identificar cualquier pauta no lineal en los datos. Otra forma es realizar el análisis de regresión múltiple y realizar el análisis de los residuos. La corrección más directa de la no linealidad, es la transformación de una o ambas variables para conseguir la linealidad.

2.4.4. Ausencia de errores correlacionados.

Debemos asegurarnos que cualquiera de los errores de predicción no está correlacionado con el resto. Por ejemplo, si encontráramos un indicio que sugiera que los errores son positivos y negativos alternativamente, debemos entender que hay alguna relación sistemática no explicada de la variable dependiente. Si existe tal situación, no podemos estar seguros de que nuestros errores de predicción sean independientes de los niveles que estamos intentando predecir. Existe otro factor que está afectando los resultados, pero que no está incluido en el análisis. Este error se debe, en

muchos casos, a la recogida de datos. Si estos se hacen por grupos, por personas diferentes, etc. Pueden haber errores sistemáticos, y hay que analizar las diferencias entre esos grupos; si eso existe, hay que incluir el factor "grupos" en el análisis. Es decir; este error puede ser corregido incluyendo el factor causante omitido en el análisis.

3. Clasificación de los Análisis Multivariados

Los métodos estadísticos multivariados se pueden seleccionar teniendo en cuenta varios aspectos pero todos ellos deben incluir: a) la *estructura* de la *matriz de datos*, b) el *objetivo* perseguido, y c) la *naturaleza* de esos *datos* (Dagnelie, 1981).

a) Según la estructura de la matriz de datos, los métodos pueden clasificarse según sean las variables o los individuos de la matriz de datos de base. La estructura se refiere a si las variables o los individuos son diferentes, o si pertenecen a un grupo o a más grupos de variables o de individuos:

- *sin* ninguna *estructura* en particular, (análisis de componentes principales y análisis factorial; conglomerados)
- una *estructura entre variables*, (métodos de regresión múltiple ó análisis de correlación canónica)
- una *estructura entre individuos*, (análisis discriminante)
- *ambas estructuras* (análisis de correspondencias múltiples)

b) Según el objetivo perseguido, los métodos son muy difíciles de clasificar, pues puede haber muchos y muy diferentes, pero los agruparemos en dos grandes grupos: los *descriptivos*, y los *inferenciales*. También pueden clasificarse de la siguiente forma:

- *Reducción de datos o simplificación estructural*. El problema de estudio se debe representar tan simplemente como se pueda sin sacrificar información valiosa, y esto hará la interpretación más sencilla.
- *Selección y agrupamiento***.** Se crean grupos de individuos o variables "similares" basándose en las características que se midieron. En este caso se requieren reglas para clasificar los individuos en grupos bien definidos.
- *Investigación de la dependencia entre variables***.** Estamos interesados en las relaciones entre variables. No sabemos si todas las variables son mutuamente independientes, o una ó más variables dependen de otras. Si ocurre esto, se desea conocer cómo se relacionan.
- *Predicción*. Las relaciones entre variables deben ser halladas con el propósito de predecir los valores de una ó más variables sobre la base de las observaciones de otras variables.
- *Construcción de hipótesis y prueba de ellas*. Se desean probar algunas hipótesis estadísticas específicas, formuladas en función de los parámetros de poblaciones multivariadas. Esto debe realizarse para validar las suposiciones o para reforzar algunas convicciones previas).

c) Según la naturaleza de los datos

- En el caso de los *métodos descriptivos:*

- si las "p" variables son cuantitativas (Análisis Factorial Clásico (Análisis de Componentes Principales y Análisis Factorial común)
- si las "p" variables son cualitativas y/o cuantitativas (Métodos de Conglomerados (clusters) y Análisis de Correspondencias (Simple y Múltiple)

- En el caso de los *métodos inferenciales:*

En este caso, siempre hay dos grupos de variables y casi siempre se reconocen como variables independientes y variables dependientes (Tabla 1). Por esto, tenemos que tener en cuenta la naturaleza y la cantidad de variables de cada uno de los grupos:

Tabla 1: Métodos multivariantes inferenciales según el número y naturaleza de las variables.

Variables dependientes	Variables independientes	Método
1 variable cuantitativa	1 ó n variables cuantitativas	Regresión múltiple
1 variable cualitativa	n variables cuantitativas	Análisis Discriminante
p variables cuantitativas	p variables cuantitativas	Correlación canónica
p variables cuantitativas	1 ó n variables cualitativas	MANOVA

4. Validación de los Resultados

Entre los métodos que permiten conocer la estabilidad de los ejes, de las formas o de las clases se encuentran: los métodos de validación empíricos, los métodos de validación por remuestreo, el análisis de las zonas de confianza que se pueden trazar alrededor de los puntos en los espacios de visualización y el caso de la clasificación así como el número y la significación de las clases.

4.1. Métodos de validación empíricos.

Los cálculos de estabilidad y de sensibilidad son probablemente los procedimientos de validación más probatorios. Lo esencial de las operaciones consiste en una verificación de la estabilidad de las configuraciones después de realizadas diversas perturbaciones a la tabla inicial de datos. Desde el punto de vista teórico, la estabilidad de los factores en el análisis de componentes principales y en el análisis de correspondencias se debe acometer, estudiando las variaciones máximas de los factores y de los valores propios cuando se realizan modificaciones bien precisas a los datos de base; entre los que se encuentran: añadir o eliminar elementos a la tabla de datos, reagrupar varios elementos, modificar valores de la tabla, cambiar la métrica y la ponderación (Escofier y Leroux, 1972; Escofier, 1979).

Existen tres elementos de “estabilidad interna” que pueden condicionar la calidad y la estabilidad de los resultados en un análisis factorial: la elección y el peso de las variables, la codificación de las variables y los errores de medición (Greenacre, 1984). Hay un cuarto elemento (Lebart et al., 1995) referido a los pesos de los individuos conjuntamente con las fluctuaciones de muestreo que responde sobre todo a las demandas de “estabilidad externa”.

Las cuatro fuentes de perturbación dan lugar a modificaciones de la tabla inicial y permiten verificar la permanencia de la configuración inicial. Además, pueden ser implícitamente estudiadas en la medida en que no se necesite un solo análisis, sino una serie de análisis por etapas, y en cada una de ellas, la tabla de datos es modificada por la incorporación de nuevos individuos o por la selección de nuevas variables, por correcciones de algunos errores eventuales, o por recodificación de algunos datos. Esta aproximación de la “estructura en escalada”, (Mallows y Tukey, 1982), permite un conocimiento progresivo del fenómeno y constituye en sí, un procedimiento de validación de los resultados. Un ejemplo de inestabilidad es el del valor atípico que tiene demasiada influencia sobre el plano principal, y por lo tanto, quitarlo, cambia sustancialmente la orientación de dicho plano (Holmès, 1985).

4.2. Métodos de validación por remuestreo.

Estos son los métodos de cálculos intensivos que se basan en las técnicas de simulaciones de muestras a partir de una sola muestra y son los únicos procedimientos posibles cuando la complejidad analítica del problema no permite el uso de la inferencia estadística clásica. En general, consisten en la repetición de los análisis para las diferentes muestras simuladas para estudiar las fluctuaciones de los resultados obtenidos (valores propios, factores o cualquier otro parámetro

estadístico a estimar). Por esto, se evalúa la variabilidad real de un parámetro mediante la división de su variabilidad para el conjunto de dichas series de datos. Existen varios métodos de validación que permiten obtener, de manera diferente, las muestras artificiales. Los más conocidos son: jackknife (Quenouille, 1949; Tukey, 1958; Miller, 1974), bootstrap (Efron, 1979; Efron y Tibshirani, 1993) y la validación cruzada (Lachenbruch y Mickey, 1968).

4.3. Zonas de confianza y número de ejes.

Los resultados a los que se llega en un análisis factorial no son afirmaciones, sino representaciones; es decir, objetos complejos a los que se aplican mal las diferentes técnicas de medición de información usuales en estadística. Una forma observada en un plano factorial se puede validar mediante:

- procedimientos externos: conocimiento a priori, o posicionamiento de variables suplementarias,
- cálculos de estabilidad adaptados (exploración de una vecindad de los datos construida a partir de los errores de medición o de respuesta),
- cálculo de las zonas de confianza por las posiciones de los puntos-filas y de los puntos-columnas. Estos cálculos pueden ser analíticos, basados en hipótesis probabilísticas, o por el contrario, basados en las técnicas de remuestreo expuestas anteriormente.

4.3.1. Zonas de confianza establecidas por bootstrap.

La técnica bootstrap es idónea para estudiar la estabilidad de las formas, y por tanto, después de su aplicación, podemos contar con muchas réplicas del análisis deseado. Como el trabajo de superposición de las estructuras puede ser laborioso, se pueden tener en cuenta las siguientes posibilidades:

- analizar las yuxtaposiciones de las tablas de contingencia por filas (para estudiar la variabilidad de las filas) y en columnas (para posicionar las columnas simuladas), o
- proyectar como elementos suplementarios, las filas (y las columnas simuladas) en los planos factoriales salidos del análisis de la tabla de contingencia inicial, o
- calcular una tabla de contingencia promedio y proyectar las filas o las columnas como en la posibilidad anterior.

Los tres procedimientos dan resultados parecidos para los casos donde efectivamente existe una estructura estable.

4.4. Número de clases y validación de las clasificaciones.

Existen varios trabajos realizados con vistas a la existencia y la determinación del número de clases. Entre ellos, podemos referirnos al análisis del marco inferencial general donde se puede probar la hipótesis de ausencia de estructura, pero que por ser muy severa, casi siempre es rechazada; también los procedimientos empíricos son ampliamente utilizados, y los cálculos de estabilidad utilizando los métodos de simulación o de remuestreo, permitirán probar la calidad de los resultados y lograr una apreciación de la realidad de las clases producidas por los algoritmos. No se puede descartar el rol importante que juegan, también los criterios externos; sin embargo, los métodos para conocer cuántas clases se deben retener, son procedimientos empíricos en los cuales no interviene ninguna información externa.

5. Referencias

Andrews, D.F. (1972). Plots of High Dimensional Data, Biometrics, 28, p. 125 – 136.

Chernof, H. (1973). Using Faces to Represent Points in K – Dimensional Space Graphically, Journal of the American Statistical Association, 68, N° 342, p. 361 – 368.
Dagnelie, P. (1981). Principes d'expérimentation. Les Presse Agronomique de Gembloux, Gembloux, Bélgica.
Efron, B. (1979). Bootstraps methods: another look at the Jackknife. Ann. Statist. 7, p. 1 – 26.
Efron, B. y Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman and Hall.
Escofier, B. y Leroux, B. (1972). Etude de trois problèmes de stabilité en analyse factorielle. Publication de l'Ínstitut Statistique de l' Université de Paris, 11, p. 1 – 48.
Escofier, B. (1979). Stabilité et approximation en analyse factorielle. Thèse d'Etat, Université Pierre et Marie Curie, Paris, Francia.
Greenacre, M. (1984). Theory and applications of correspondence analysis. Academic Press, London.
Hair, J.F., Anderson, R.E., Tatham, R.L. y Black, W.C. (1999). Análisis Multivariante, 5° ed. Prentice Hall, IBERIA, Madrid, España.
Holmès, S. (1985). Outils informatiques pour l'évaluation de la pertinence d'un résultat en analyse des données. Thèse USTL, Montpellier, Francia.
Johnson D.E. (2000). Métodos Multivariados aplicados al análisis de datos. International Thomson Editores, S.A.deC.V., México.
Lachenbruch, P.A. y Mickey, M.R. (1968). Estimation of error rate in discriminant analysis. Technometrics, 10, p. 1 – 11.
Little, Roderick, J.A. y Rubin, D.B. (1987). Statistical Analysis with Missing Data. John Wiley and Sons, New York.
Lebart, L., Morineau, A., Fénelon, J.P. (1981). Traitement des données statistiques. DUNOD, París, Francia.
Lebart, L. Morineau, A. y Piron, Marie (1995). Statistique exploratoire multidimensionnelle. DUNOD, París, Francia.
Mallows, C.L. y Tukey, J. W. (1982). An overviews of technique of data analysis emphasizing its exploratory aspects. In: Some recent advances in Statistics. (J. Tiago de Oliveira, ed.), Academic Press, p. 11 – 172.
Miller, R.G. (1974). The Jackknife – a review. Biometrika, 61, p. 1 – 15.
Quenouille, M. (1949). Approximate tests of correlation in time series. J. Royal Statist. Soc., B, 11, p. 18 – 44.
Toussaint, G.T. (1974). Bibliography on estimation of misclassification. IEEE, Trans. Inform. Theory, IT – 20, p. 472 – 479.
Tukey, J. W. (1958). Bias and confidence in not quite large samples. Ann. Math. Statist., (Abstract), 29, p. 614.

Autorización

Los autores autorizan a LACCEI la publicación de este artículo en las memorias de la conferencia. Ni LACCEI ni los editores son responsables del contenido y de la implicaciones que se expresan en este artículo.