FORMULACIÓN BIPLOT DE TÉCNICAS DE ANALISIS DE DATOS DE TRES MODOS

Mª. PURIFICACIÓN GALINDO VILLARDÓN

Dpto. de Estadística. Universidad de Salamanca. España.

pgalindo@usal.es

RESUMEN

El propósito de este trabajo es poner de manifiesto cómo los métodos Biplot enriquecen las técnicas estadísticas utilizadas en análisis de datos de tablas de tres modos, poniendo especial énfasis en las técnicas de la escuela francesa, Análisis Triádico, STATIS y Análisis Factorial Multiple, las cuales contemplan tres fases: 1) Análisis de la interestructura, que consiste en la comparación global de las matrices originales. 2) Búsqueda de la matriz consenso y 3) Análisis de la intraestructura que consiste en el estudio de individuos o variables de las matrices originales sobre el subespacio común. Se presenta la versión Biplot del Triádico, el Statis Canónico y el Biplot Múltiple y se compara, además, con el Meta- Biplot, el Biplot Conjunto y el Biplot interactivo.

Palabras clave: ***Biplot***, ***Statis, AFM, Triádico.***

1.- INTRODUCCIÓN: DATOS DE TRES MODOS

La forma tradicional de presentar la información en un Análisis Multivariante, es a partir de una matriz que contiene los valores de p variables observadas en n individuos. Para poder caracterizar los individuos en función de las variables observadas, es necesario reducir la dimensionalidad del problema; es decir, representar los individuos no en el p-hiperespacio de partida, sino en un subespacio de dimensión reducida, generalmente de dimensión 2. Esto está resuelto en la literatura desde principios del siglo pasando sin más que aplicar un Análisis de Componentes Principales (ACP). Si se pretende representar conjuntamente individuos y variables podemos utilizar BIPLOT (GABRIEL, 1971; GALINDO, 1985, 1986)

Los Métodos Clásicos del Análisis Multivariante trabajan con matrices de dos vías, esto es matrices de Individuos x Variables. Cuando las interrelaciones están medidas en distintos lugares o en distintos momentos en el tiempo, se tendrá en cada lugar o en cada momento, una tabla de dos vías y por tanto, la información se presenta en varias tablas de Individuos x Variables, es decir, cada observación es originada por *tres modos*: Individuos, Variables y Condiciones/Tiempos.

La información se concentra en las denominadas *Matrices Multivía*.

Existen varios tipos de datos de tres modos (KIERS (1988, 1991)):
Datos de Tres Vías (Three Way Data): se considera un solo conjunto de individuos, un solo conjunto de variables y un solo conjunto de condiciones. La información queda recogida en T matrices de orden IxJ; siendo I, J, T el número de categorías respectivas. Son observaciones para todos los individuos, sobre todas las variables, en todas las condiciones experimentales o en todas las ocasiones.
Datos de Conjuntos Múltiples (Multiway Set of Data): cuando uno de los modos está compuesto por varios conjuntos; podemos tener:

- *Varios conjuntos de individuos*, un *solo conjunto de variables* y un solo conjunto de condiciones, en tal caso en cada condición se miden las mismas variables en diferentes individuos. Tenemos por tanto T matrices de orden I_txJ, siendo I_t el número de individuos que se evalúan en la t-ésima ocasión . Cada tabla $\mathbf{X_t}$ recoge la información de J variables sobre I_t individuos. Existen T condiciones o situaciones experimentales diferentes.
- *Varios conjuntos de variables* y un *mismo conjunto de individuos* y condiciones; es decir, en cada condición se miden diferentes variables a los mismos individuos. En este caso tenemos T matrices de orden IxJ_t ; siendo J_t el número de variables medidas en la t-ésima ocasión. Cada tabla $\mathbf{X_t}$ recoge información de J_t variables sobre I individuos. Existen T condiciones o situaciones experimentales diferentes.

2.- MÉTODOS PARA EL TRATAMIENTO DE DATOS DE TRES MODOS

Los primeros trabajos de integración de matrices se dan con el Análisis Canónico (HOTELLING (1936)); es una técnica que consiste en buscar relaciones entre dos conjuntos de variables a partir de ejes canónicos. La extensión a más de dos conjuntos de variables se conoce más tarde como Análisis Canónico Generalizado (CARROLL (1968); KETTENRING (1971)).
Actualmente los trabajos de integración de matrices quedan recogidos en dos vertientes fundamentales: los métodos franceses y los métodos anglosajones.
En los métodos franceses el estudio se divide en tres etapas fundamentales:
1) Análisis de la interestructura, que consiste en la comparación global de las matrices originales.
2) Búsqueda de una matriz consenso o compromiso construida a partir de la concatenación de las matrices originales ponderadas; donde la elección de la ponderación es, en general lo que diferencia los distintos métodos.
3) Análisis de la intraestructura que consiste en un estudio más detallado de cada elemento (individuos y variables) de las

matrices originales sobre un subespacio común creado en la etapa anterior.
Los métodos más conocidos que tratan la integración de matrices desde esta perspectiva son por orden cronológico: el *Statis* (L'HERMIER DES PLANTES,1976), el Análisis *Triádico* (JAFFRENOU, 1978) y el Análisis *Factorial Múltiple* (ESCOFIER y PAGÈS, 1984).
Estos métodos, tratan a los modos en forma *asimétrica* y trabajan, a excepción del Triádico, con *datos derivados* ya que el consenso lo forman a partir de las configuraciones obtenidas de algún análisis factorial o a partir de matrices de productos cruzados (matrices de distancias, similaridades o covarianzas).

La escuela anglosajona se caracteriza por ajustar modelos que reproduzcan lo más fiable posible el dato original. En tal sentido podemos citar entre otros: el modelo de TUCKER (1966); el *PARAFAC = Para*llel *Fac*tor analysis (HARSHMAN, 1970; HARSHMAN y LUNDY, 1984) o *CANDECOMP = Can*onical *Decomp*osition (CARROLL y CHANG 1970), el *INDSCAL* (CARROLL y CHANG, 1970), y el *IDIOSCAL* (CARROLL y CHANG, 1972) para diferencias individuales. Los métodos TUCKALS: *Tuckals2 y Tuckals3* (KROONENBERG y de LEEUW, 1980), basados en los modelos: Tucker2 y Tucker3 (TUCKER, 1966, 1972); el modelo de *Escalamiento de Tres Modos* (Three Mode Model) (KROONENBERG, 1983) que es un caso particular del método Tuckals3, donde dos de los tres modos coinciden.
El problema de la integración de matrices se recoge además en los métodos de Análisis Procrustes (GOWER, 1975; GOWER y HAND, 1996); Meta-Componentes Principales (KRZANOWSKI, 1979; 1982) y Análisis de Componentes Principales Comunes (FLURY, 1984; 1988); los cuales se basan en la búsqueda de una configuración consenso ¨óptima¨, en el sentido de aproximar lo máximo posible las distintas configuraciones asociadas a cada matriz.

En este trabajo nos centraremos en los métodos de la escuela francesa los cuales se fundamentan en el ACP de dos vías.

Existen varios e interesantes trabajos en los que se realizan estudios comparativos de diferentes técnicas entre los que cabe destacar el trabajo de KIERS, (1988) en el que se comparan las técnicas de la escuela francesa y la anglosajona, con especial énfasis en las relaciones entre el STATIS y el CANDECOMP/PARAFAC y entre el STATIS y el TUCKALS 3; BOVE y Di CIACCIO (1994) pusieron de manifiesto las relaciones entre el STATIS y el modelo de TUCKER 2 (TUCKALS 2); PAGES, (1996) compara el AFM y el STATIS y el trabajo de MURES y col, (2006) en el que se comparan las técnicas desde un punto de vista empírico. En ninguno de estos trabajos se contempla su expresión BIPLOT.

3.- EL PAPEL DE LOS MÉTODOS BIPLOT EN EL TRATAMIENTO DE DATOS DE TRES MODOS

Un BIPLOT (GABRIEL 1971) es una representación gráfica de datos multivariantes. La característica fundamental que lo hace diferenciar de las distintas representaciones gráficas asociadas a los métodos clásicos de reducción de dimensionalidad; es que en este caso es posible una representación conjunta de filas y columnas de la matriz de datos. Desde el punto de vista algebraico, el BIPLOT se basa en el mismo principio sobre el que se sustentan la mayoría de las técnicas factoriales de reducción de dimensionalidad, es decir, hace uso de la descomposición en valores y vectores singulares de la matriz. La diferencia fundamental es que en este caso se trata de reproducir el dato (Biplots de Gabriel) y se incorpora una representación conjunta de filas y columnas.
La interpretación del BIPLOT se basa en conceptos geométricos muy sencillos, así por ejemplo:

- La similitud entre individuos (filas) es una función inversa de la distancia entre los mismos.
- Las longitudes y los ángulos de los vectores que representan a las variables, se interpretan en términos de variabilidad y covariabilidad respectivamente.
- Las relaciones entre filas y columnas se interpretan en términos de producto escalar, es decir, en términos de las proyecciones de los puntos “fila” sobre los vectores “columna”.

Las dos factorizaciones BIPLOT más importantes propuestas por GABRIEL (1971) fueron denominadas: GH-Biplot y JK-Biplot. El GH-Biplot consigue una alta calidad en la representación de las columnas (variables) y no tan alta para las filas (individuos); mientras que el JK-Biplot consigue una alta calidad de representación para las filas, y no tan alta para las columnas.
GALINDO (1986) demuestra que con una conveniente elección de los marcadores es posible representar las filas y las columnas simultáneamente sobre un mismo sistema de coordenadas, con una alta calidad de representación tanto para las filas como para las columnas. GALINDO denomina a este tipo de BIPLOT, HJ-Biplot.

El BIPLOT, *ha dado lugar a nuevos métodos de análisis multivariante de datos*, al ser combinado con otras técnicas clásicas; en este sentido podemos citar los modelos AMMI (GOLLOB (1968)), (VARELA, 2002), el cual inicialmente combina las técnicas de Análisis de Varianza y Análisis de Componentes Principales, y posteriormente incorpora el BIPLOT en lugar del ACP (GABRIEL (1978); KEMPTON (1984); GAUCH (1988)). Consiste en hacer un BIPLOT a la matriz de residuales del modelo.

GABRIEL (1972) combina el Biplot con el MANOVA; introduce algunas características del MANOVA-BIPLOT de una vía; consiste en representar mediante un BIPLOT los resultados del

MANOVA. Más tarde, AMARO (2001) y AMARO y col, (2004) lo generalizan al caso de dos factores de variación; lo cual facilita el estudio de los efectos principales e interacciones para cada una de las variables analizadas.

En la línea de búsqueda de una configuración consenso, MARTÍN-RODRIGUEZ (1996) y MARTÍN-RODRIGUEZ y col (2002), hacen una generalización para el caso en el que las configuraciones son el resultado de aplicar un análisis biplot a cada matriz inicial de datos.

Así mismo, VALLEJO, (2004) y VALLEJO y col (2006), proponen un método, que denominan STATIS CANONICO, basado en la metodología STATIS, que permite recoger la estructura de grupos de los datos y la evolución de la misma en diferentes ocasiones, el cual se complementa mediante una representación simultánea de grupos, variables y ocasiones (Biplot) que amplia las capacidades de los métodos clásicos. El Statis Canónico supera las limitaciones del MANOVA y/o el Análisis Canónico que aunque recogen la estructura de grupos no capturan la evolución temporal y las limitaciones del Statis que aunque es sensible a la evolución temporal, confunde la variabilidad “entre” y “dentro” de los grupos.

También el BIPLOT ha permitido enriquecer el Análisis Triádico. BASSO, (2006) formuló las representaciones Biplot del Análisis Triádico, proyectando en el sistema de referencia común que proporcionan las variables o los individuos consenso. La formulación Biplot del Análisis Triádico le da una nueva visión que amplia su aplicabilidad ya que permite, para datos de tres vías, la definición correcta de un conjunto de trayectorias en el espacio consenso y la posibilidad de definir índices que permiten evaluar la calidad de las representaciones

El Biplot ha permitido, también, la formulación de una alternativa al Análisis Factorial Múltiple BACCALÁ, (2004) que nos permite representar los individuos y los distintos grupos de variables en el mismo espacio, y obtener un patrón de comparación común para todos los grupos donde es posible calcular medidas de la calidad de representación para cada uno de los elementos representados. Las proyecciones de los marcadores filas sobre los marcadores columnas reproducen aproximadamente los elementos de las columnas en la matriz original yuxtapuesta, permitiendo una ordenación aproximada de los individuos respecto a cada una de las variables. En la factorización HJ- Biplot Múltiple, los marcadores filas y los marcadores columnas se pueden representar en el mismo sistema de referencia, con óptima calidad de representación. En la factorización JK Biplot Múltiple, los marcadores para las columnas constituyen las componentes "consenso o compromiso"; por tanto los marcadores fila serán las coordenadas de los individuos sobre el espacio de las componentes consenso. En esta factorización es posible

calcular la bondad del ajuste de la matriz ponderada y de los datos originales.

Para finalizar, señalar que el Biplot ha sido generalizado al caso de varias matrices de datos, en el contexto de la escuela anglosajona. En tal sentido podemos citar el Biplot Conjunto y el Biplot Interactivo (CARLIER Y KROONENBERG, 1996); los cuales operan con tres matrices de marcadores.

4.- REFERENCIAS

AMARO, R. I. (2001). *Manova-Biplot para diseños con varios factores basado en modelos lineales generales multivariantes*. Tesis Doctoral. Universidad de Salamanca.

AMARO, R. I. ; **VICENTE-VILLARDÓN, J. L. y GALINDO, M. P** (2004). Manova Biplot para arreglos de tratamientos con dos factores basados en Modelos Lineales Generales Multivariantes. *Interciencia*. *29 (1). 26-32.*

BACCALÁ, N. (2004). *Contribuciones al Análisis de datos Multivía: Tipología de las variables*. Tesis Doctoral. Universidad de Salamanca. España.

BASSO, L.C. (2006). *Análisis Conjunto de varias matrices de Datos: Contribuciones a la tipología de los individuos.* Tesis Doctoral. Universidad de Salamanca. España.

BOVE, G.; Di CIACCIO, A. (1994). A user-oriented overview of multiway methods and software. *Computacional Statistics and Data Analysis. 18. 15-37.*

CARLIER, A. & KROONENBERG, P. M. (1996). Decompositions and Biplots in Three-way Correspondence Analysis. *Psychometrika, 61(2). 355-373.*

CARROLL, J. D. (1968): 'A generalization of canonical correlation analysis to three or more sets of variables'. *Proceedings of 76th annual convention of the American Psychological Associations, 227-228*.

CARROLL, J. D & CHANG, J. J. (1970): Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35. 283-320.*

CARROLL, J. D. y CHANG, J. J. (1972): IDIOSCAL (Individual Differences In Orientation Scaling): A Generalization of INDSCAL allowing idiosyncratic reference systems as well as an analytic approximation to INDSCAL. *Trabajo presentado en la Psychometric Society*, Princeton, NJ.

ESCOUFIER, Y. (1973). Le traitement des variables vectorielles. *Biometrics. 29. 750-760.*

FLURY, B.D. (1984). Common Principle Components in K groups. *Journal of the American Statistical Associations*. *79. 892-898.*

FLURY, B. (1988). *Common Principal Components and Related Multivariate Models*. Wiley. New York.

ESCOFIER, B. et PAGÈS, J. (1984). Analyse factorielle multiple. *Cahiers du BURO*, *2*. ISUP. Paris.

GABRIEL, K. R. (1971): The Biplot graphic display of matrices with applications to principal components análisis. *Biometrika*, *58 (3). 453-467.*

GABRIEL, K.R. (1972). 'Analysis of meteorological data by means of canonical decomposition and Biplots'. *Journal of Applied Meteorology*, *11: 1071-1077.*

GABRIEL, K.R. (1978). 'Least Squares Approximation of Matrices by Additive and Multiplicative Models'. *Journal of the Royal Statistical Society*, Series B. *40. 186-196*.

GALINDO, M.P. (1985). *Contribuciones a la representación simultánea de datos multidimensionales*. Tesis Doctoral. Universidad de Salamanca.

GALINDO, M.P. (1986): Una alternativa de representación simultánea: HJ-Biplot. *Qüestiio*, *10(1): 13-23.*

GOLLOB, H.F. (1968). 'A statistical model which combines features of factor analytic and analyses of variance techniques'. *Psychometrika*, *33. 73-115.*

GOWER, J. C. (1975): 'Generalized Procrustes analysis'. *Psychometrika*, *40. 33-51.*

GOWER, J. C. & HAND, D. J. (1996): *Biplots*. Chapman and Hall. London.

GAUCH, H.G. (1988). 'Model Selection and Validation for Yield Trials with Interaction'. *Biometrics, 44. 705-715*.

HARSHMAN, R. A. (1970). 'Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-mode factor analysis'. UCLA *Working Papers in Phonetics*, *16. 1-84.*

HARSHMAN, R. A. & LUNDY, M. E. (1984). 'The PARAFAC model for three-way factor analysis and multidimensional scaling'. In: H. G. Law, C. W. Snyder Jr., J. A. Hattie & R. P. Mc Donald (eds.). *Research Methods for Multimode Data Analysis*, *122-215*. Praeger. New York.

HOTELLING, H. (1936): 'Relations between two sets of variates'. *Biometrika, 28. 321-377*.

KEMPTON, R.A. (1984). 'The use of Biplots in interpreting variety by environment interactions'. *Journal of Agricultural Science*. Cambridge *103: 123-135.*

KETTENRING, J. R. (1971): 'Canonical analysis of several sets of variables'. *Biometrika*, **58**: 433-460.

KIERS, H. A. L. (1988). 'Comparison of "Anglo-Saxon" and "French" Three-Mode Methods'. *Statistique et Analyse des Données*, *13. 14-32.*

KIERS, H.A.L. (1991): Hierarchical relations among three-way methods. *Psychometrika*, *56. 449-470.*

KROONENBERG, P.M. and DE LEEUW, J. (1980): Principal Component Analysis of Three-Mode Data by means of Alternating Least Squares Algorithms. *Psychometrika*, *45. 69-97.*

KROONENBERG, P. M. (1983): *Three-Mode Principal Components Analysis. Theory and Applications*. DSWO-Press. Leiden. The Netherlands.

KRZANOWSKI, W.J. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Association, 74 (367). 703-707*. Correccion en *76. 1022*.

KRZANOSWSKI, W. J. (1982): Between-group comparison of principal components. Some sampling results. *Journal of Statistical Computation and Simulation*, *15. 141-154*.

L´HERMIER DES PLANTES. (1976). *Structuration des tableaux à trois indices de la statistique. Theory et applications d'une méthode d'analysie conjointe.* Thèse de 3º cycle, USTL, Montpellier.

MARTÍN-RODRIGUEZ, J. (1996): *Contribuciones a la Integración de Subespacios desde una Perspectiva Biplot*. Tesis Doctoral. Departamento de Estadística y Matemática Aplicadas. Universidad de Salamanca.

MARTÍN-RODRIGUEZ, J.; GALINDO, M. P. & VICENTE-VILLARDÓN, J. L. (2002): 'Comparison and integration of subspaces from a biplot perspective'. *Journal of Statistical Planning and Inference, 102 (2). 1-13*.

MURES, M.J.; VALLEJO, M.E.; GARCIA, A. (2006). Comparación Empírica de técnicas estadísticas para tablas de tres entradas: La construcción en Castilla y León en el periodo 2002-2004. *Pecvnia 3. 95-140.*

PAGÈS, J. (1996): Eléments de comparaison entre l´Analyse Factorielle Multiple et la Méthode Statis. *Revue de Statistique Appliquée*, *44 (4). 81-95.*

TUCKER, L. R. (1966): 'Some mathematical notes on three-mode factor analysis'. *Psychometrika, 31. 279-311*.

TUCKER, L. R. (1972): 'Relation between multidimensional scaling and three-mode factor analysis'. *Psychometrika, 37. 3-27*.

VARELA, M. (2002): *Los Métodos Biplot Como Herramienta de Análisis de Interacción de Orden Superior en un Modelo Lineal/Bilineal*. Tesis Doctoral. Departamento de Estadística y Matemática Aplicadas. Universidad de Salamanca.

VALLEJO, A. (2004). *Análisis de datos Multivía con estructura de grupos*. Tesis Doctoral. Universidad de Salamanca. España.

VALLEJO, A.; VICENTE-VILLARDÓN, J.L. y GALINDO, M.P. (2006). Canonical STATIS: Biplot Análisis of multi-table group structured data based on STATIS-ACT methodology. *Computacional Statistics and Data Analysis 51. 4193- 4205.*