Revista Venezolana de Análisis de Coyuntura, 2001, Vol. VII, No. 2 (jul-dic), pp. 333-347
recibido 25-07-01 / arbitrado 28-11-01

ESTIMACIÓN MÁXIMO VEROSÍMIL EN APROXIMACIONES BIPLOTS

Olesia Cárdenas
ESCUELA DE ECONOMÍA, UCV

María P. Galindo
José Vicente-Villardón
UNIVERSIDAD DE SALAMANCA

Resumen:

Los métodos biplots clásicos de Gabriel (1971) se utilizan generalmente con propósitos descriptivos, sin hacer supuestos sobre distribuciones poblacionales, sin embargo, el biplot de una matriz de datos puede interpretarse también como un modelo bilineal multiplicativo (Gollob, 1968). Desde esta ultima perpectiva, en esta investigación se analiza su geometría y se formaliza matemáticamente un método de estimación alternativo a los existentes. El método propuesto puede ser de mucha utilidad en la práctica, ya que permite su generalización para introducir información externa que ayude en la interpretación, y en la obtención de variables latentes continuas en Ciencias Sociales.

Palabras claves: Biplots, regresión, modelos bilineales generalizados, estimación máximo verosímil, geometría.

1. INTRODUCCIÓN

La posibilidad de interpretar el biplot (Gabriel, 1971) de una matriz de datos Y de orden (nxp), como un modelo bilineal multiplicativo (Gollob, 1968), permite su utilización no solo desde una perspectiva descriptiva, sino también para la descripción de aspectos resaltantes en tablas de dos vías, tal como la interacción entre los dos factores de clasificación en los que se agrupa una variable (Denis, 1991; Falguerolles, 1995; Van Eeuwijk, 1995; Choulakian, 1996), o para visualizar el modelo subyacente en los datos (Bradu & Gabriel, 1978; Gabriel, Galindo y Vicente-Villardón, 1998).

Para la aproximación de los biplots, los autores citados se fundamentan en generalizaciones heurísticas de los métodos de estimación utilizados en los modelos bilineales, considerándolos como extensiones de los modelos lineales generalizados (Nelder & Wedderburn, 1972), cuando la variable respuesta tiene cualquiera de las distribuciones de la familia exponencial (normal, binomial, poisson, multinomial, etc).

Por su parte, Gower (1992), Gower & Harding (1988) y Gower & Hand (1996), aunque utilizan los biplots para la descripción de una matriz dan otro enfoque diferente al clásico, el cual se puede relacionar con la forma factorial clásica de la escuela francesa de análisis de datos y con los métodos de ordenación de la escuela biométrica. Ellos describen a priori la geometría de los biplots en términos de proyecciones de subespacios, en contraposición a la geometría a posteriori utilizada en la diagnosis de modelos, tal como lo hacen Vicente-Villardón & Galindo (1998) para el caso de variables con respuestas no lineales de tipo sigmoidal.

En otro contexto, la forma factorial para variables con distribuciones de la familia exponencial se puede comparar a la obtención de variables latentes continuas en las Ciencias Sociales, tal como sucede por ejemplo, en la Teoría de Respuesta al Ítem (Baker, 1992).

La finalidad de este trabajo de investigación tiene el propósito de describir una matriz de datos Y (individuos por variables), conformada por variables con distribuciones pertenecientes a la familia exponencial. Se utilizan en la aproximación de los biplots los modelos bilineales generalizados multiplicativos, analizando su geometría y proponiendo formalmente un método alternativo de estimación (Cárdenas, 2000).

2. APROXIMACIONES EN BIPLOTS CLÁSICOS

La fundamentación teórica de los biplots clásicos (Gabriel, 1971) se basa en la aproximación de una cierta matriz de datos Y de orden (nxp) y de rango r, por una de bajo rango (q < r), a través de la descomposición en valores singulares, para luego hacer una factorización en matrices de marcadores filas A de orden (nxq), y de marcadores columnas B de orden (qxp), tal que:

$$\mathbf{Y}_{(r)} \cong \mathbf{Y}_{(q)} = \mathbf{U}_{(q)}\mathbf{D}_{(q)}\mathbf{V}'_{(q)} = \mathbf{A}_{(q)}\mathbf{B}'_{(q)} \qquad (1)$$

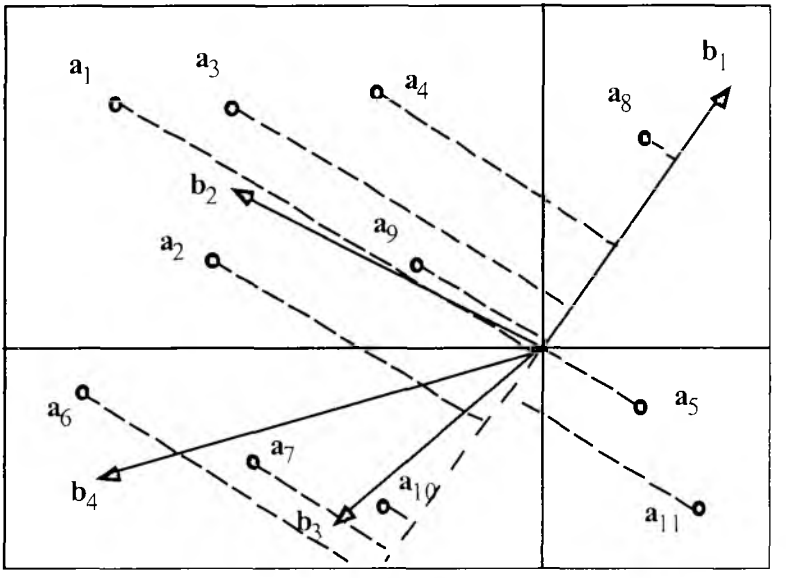
La notación utilizada es la usual en la terminología estadístico-matemática, siendo: q el rango de la matriz Y (q < min(n,p)), U y V matrices cuyos vectores columna ortonormales son los vectores singulares de (YY') e (Y'Y) respectivamente y D una matriz diagonal (D = [diag (α_k)]; α_k = valores singulares).

Los distintos tipos de biplots: el GH (column metric preserving), el JK (row metric preserving) y el SQRT (Gabriel, 1971), así como también el HJ (row column metric preserving) (Galindo, 1985), dependen de las métricas introducidas

en el espacio de las filas o de las columnas, razón por la cual los marcadores tienen distintas propiedades de acuerdo a la factorización Biplot elegida en (1).

El producto interno de la fila i de $A_{(q)}$ y la columna j de $B'_{(q)}$ constituye una *forma bilineal* (Gollob, 1968), admitiendo por ende una representación gráfica a través de la proyección ortogonal de $a_{i(q)}$ sobre $b_{j(q)}$ y viceversa, tal como se muestra en la figura 1.

Figura 1: Proyección ortogonal de los marcadores fila a_i sobre el marcador columna b_1, para la representación biplot de una matriz de datos Y de orden (4x11)



Basándose en las propiedades geométricas del producto escalar entre marcadores fila y columna, a partir de esa representación se puede aproximar: la similitud global y el orden de los individuos en relación con una variable particular, la variabilidad y correlación de las variables, el orden de las medias de las variables, los efectos filas y columna y los datos originales.

3. APROXIMACIONES BIPLOTS A TRAVÉS DEL AJUSTE DE MODELOS BILINEALES

Esta forma de aproximación surge como consecuencia de la estrecha relación existente entre la teoria de aproximación mínimo cuadrática de matrices

(Eckart & Young, 1936) y el álgebra de la descomposición de una matriz en sus valores y vectores singulares. Householder & Young (1938) demuestran que una aproximación mínimo cuadrática para una matriz dada también puede hallarse a través de la descomposición en valores singulares.

Así pues, las matrices de marcadores A y B en la factorización biplot clásica (1), son equivalentes a matrices de parámetros desconocidos en el siguiente modelo bilineal generalizado, el cual se considera como una extensión de los modelos lineales generalizados (Nelder y Wedderburn, 1972):

$$\eta = g(\mu) = A B' \Rightarrow g(\mu_{ij}) = a'_i b_j \qquad (2)$$

donde la distribución de las p variables contenidas en la matriz Y pertenece a la familia exponencial, y sus valores esperados denotados por $\mu_j = E(y_j)$, se encuentran relacionados con predictores lineales η_j a través de "funciones link" (g) como la identidad, la logit, la probit, etc.; de donde $g(\mu_{ij})$ resulta en una forma linearizada de la "función link".

A los biplots ajustados a través de modelos de ese tipo cuando la "función link" utilizada es diferente de la identidad ($g \neq I$), los denominamos *Biplots de Regresión no Lineal*, mientras que si la "función link" es la identidad ($g = I$), los denominamos *Biplots de Regresión Lineal*, siguiendo la terminología utilizada por Gower & Hand (1996).

Una vez que se han estimado las matrices de parámetros A o B en el modelo (2), se pueden considerar a posteriori las restricciones de ortonormalidad realizando la descomposición en valores singulares de la solución final obtenida, para luego recalcular en forma definitiva dichas matrices, pudiéndose representar cualquier tipo de biplot y conservar las propiedades clásicas los marcadores (Vásquez, 1995).

4. GEOMETRÍA DE LOS BIPLOTS DE REGRESIÓN

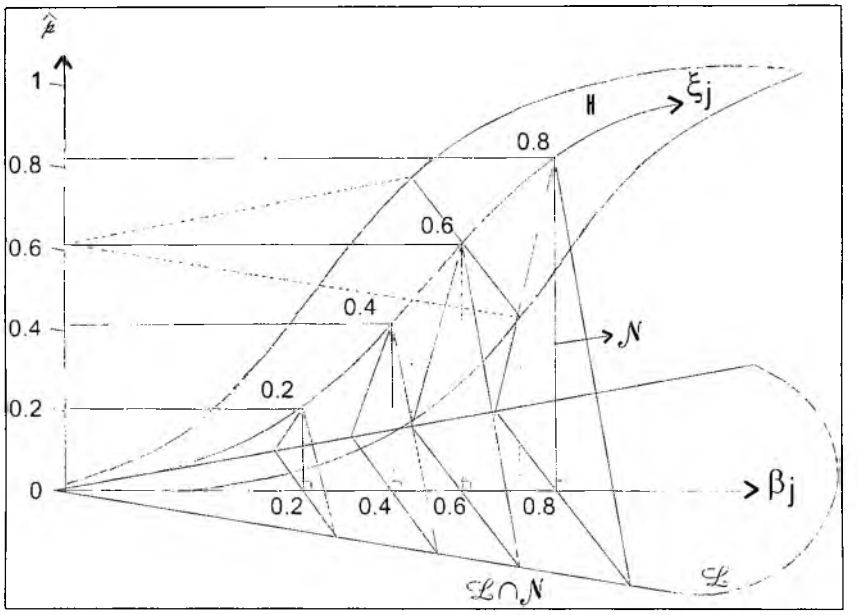
La geometría en el ajuste de los *Biplots de Regresión Lineal* es similar a la geometría en un ajuste de regresión lineal (Gower & Hand, 1996), pero considerando la proyección ortogonal de cada columna o fila de la matriz Y sobre el subespacio óptimo generado por las columnas de A o B (suponiendo conocidas respectivamente las matrices A y B).

De allí que, en los *Biplots de Regresión Lineal*, ajustar un plano de regresión a cada columna de la matriz A conlleva a la obtención de una superficie de res-

puesta lineal, cuya proyección sobre un subespacio de baja dimensión es también lineal.

En el caso del ajuste de los *Biplots de Regresión no Lineal*, Vicente-Villardón y Galindo (1998), analizan la geometría de los *Biplots Logísticos* (cuando la "función link" utilizada en el ajuste es la logit), demostrando que ajustar un plano de regresión a cada columna de la matriz A conlleva a la obtención de una superficie de respuesta sigmoidal H (tal como se muestra en la figura 2), siendo sin embargo la proyección de la curva sigmoidal ξ_j sobre el subespacio L una línea recta, que coincide con el eje Biplot $ß_j$ de predicción lineal, cuya escala (a diferencia de los *Biplots de Regresión Lineal*) no se encuentra igualmente espaciada.

Figura 2: Geometría de los biplots logísticos



En el caso de "funciones link", como la probit y log-log, dado que ellas generan curvas de respuesta sigmoidales muy similares a la logit, entonces su geometría será similar a la de los *Biplots Logísticos*.

Dado que la proyección de una curva de respuesta no lineal sobre un subespacio en baja dimensión es siempre lineal, el resultado anterior se puede generalizar a cualquier *Biplot de Regresión no Lineal* considerando otras "funciones link" como la logarítmica, raíz cuadrada e inversa.

5. ESTIMACIÓN MÁXIMO VEROSÍMIL EN BIPLOTS DE REGRESIÓN

La estimación en el ajuste de los *Biplots de Regresión Lineal* puede hacerse a través del método de Mínimos Cuadrados Alternados (Blázquez, 1998), mientras que en los *Biplots de Regresión no Lineal* se puede utilizar el método de Regresiones Generalizadas Alternadas de una forma linearizada de la "función link", como una generalización del método de máxima verosimilitud utilizado en los Modelos Lineales Generalizados (Falguerolles, 1995; Van Eeuwijk, 1995; Choulakian, 1996; Gabriel, 1998; Vicente-Villardón & Galindo, 1998).

A continuación se propone un método alternativo a los ya citados, el cual permite realizar la estimación en forma simultánea o en dos etapas (Cárdenas, 2000).

5.1. Estimación simultánea

Para el ajuste de los *Biplots de Regresión* se propone un método de estimación simultánea, a través del cual se obtienen en forma conjunta los estimadores máximo-verosímiles para las matrices de parámetros contempladas en el modelo (2). A tal efecto, se generaliza el método de estimación utilizado en Teoría de Respuesta al Ítem (Baker, 1992) adecuándolo a este contexto (Vicente-Villardón, Galindo y Cárdenas, 2000).

Contemplando la posibilidad de inclusión de un término independiente (para el caso por ejemplo de datos dicotómicos con distribución Binomial), se propone el siguiente modelo general:

$$g(\mu) = A^*B^{*\prime} = a'_0 b_0 + a'_1 b_1 + \ldots\ldots + a'_q b_q = A_0 B'_0 + AB' \qquad (3)$$

Concatenando las filas de las matrices A* y B*, se pueden reescribir las mismas como vectores columna, designándolas <A*´> y <B*´> respectivamente. Aplicando el método de Newton-Raphson, se obtiene el siguiente sistema de ecuaciones:

$$\begin{bmatrix} \langle \hat{\mathbf{B}}^{*\prime} \rangle \\ \langle \hat{\mathbf{A}}^{*\prime} \rangle \end{bmatrix}_{(t+1)} = \begin{bmatrix} \langle \hat{\mathbf{B}}^{*\prime} \rangle \\ \langle \hat{\mathbf{A}}^{*\prime} \rangle \end{bmatrix}_{(t)} - \left[\Im(\mathbf{B}^*, \mathbf{A}^*) \right]^{-1}_{(t)} \left[\mathbf{U}(\mathbf{B}^*, \mathbf{A}^*) \right]_{(t)} \qquad (4)$$

La matriz de información $\mathfrak{I}(B^*, A^*)$ está constituida por las siguientes cuatro submatrices:

$$\mathfrak{I}(\mathbf{B}^*, \mathbf{A}^*) = \begin{bmatrix} \mathfrak{I}(\mathbf{B}^*) = \left[\partial^2 L / \partial b^2\right] & \mathfrak{I}(\mathbf{B}^* \mathbf{A}^*) = \left[\partial^2 L / \partial b \partial a\right] \\ \mathfrak{I}(\mathbf{A}^* \mathbf{B}^*) = \left[\partial^2 L / \partial a \partial b\right] & \mathfrak{I}(\mathbf{A}^*) = \left[\partial^2 L / \partial a^2\right] \end{bmatrix} \quad (5)$$

El vector U(B*, A*) a su vez, está constituido por los siguientes 2 subvectores:

$$\mathbf{U}(\mathbf{B}^*, \mathbf{A}^*) = \begin{bmatrix} \mathbf{U}\langle \mathbf{B}^{*\prime} \rangle \\ \mathbf{U}\langle \mathbf{A}^{*\prime} \rangle \end{bmatrix} = \begin{bmatrix} \partial L / \partial b_{jk} \\ \partial L / \partial a_{ik} \end{bmatrix} \quad (6)$$

Para simplificar la labor computacional, en la inversión de la matriz de información $\mathfrak{I}(B^*,A^*)$, se puede transformar la misma en una matriz diagonal, bajo los mismos supuestos de Baker (1992) en Teoría de Respuesta al Item, o sea:

- Los "n" individuos se seleccionan en forma aleatoria, de donde las filas de la matriz A son independientes, siendo por lo tanto los productos cruzados entre pares de individuos nulos ya que están incorrelacionados (para $i \neq i'$), pudiendo existir sin embargo, covariación entre los parámetros de un mismo individuo (para $k \neq k'$) y por consiguiente:

$$\mathfrak{I}(\mathbf{A}) = diag\begin{bmatrix} \left[\partial^2 L / \partial a_{ik}^2\right] & \left[\partial^2 L / \partial a_{ik} a_{ik'}\right] \\ \left[\partial^2 / \partial a_{ik'} a_{ik}\right] & \left[\partial^2 L / \partial a_{ik'}^2\right] \end{bmatrix} \quad (7)$$

- Los parámetros de los individuos a_{ik} y de las variables b_{jk} son independientes, esto es:

$$\left[\mathfrak{I}(\mathbf{BA})\right] = \left[\mathfrak{I}(\mathbf{AB})\right] = \left[\mathbf{0}\right] \quad (8)$$

- Los parámetros b_{jk} de cada una de las variables son independientes, por lo que los productos cruzados entre pares de variables serán nulos (para $j \neq j'$), pudiendo existir sin embargo, covariación entre los parámetros de una misma variable (para $k \neq k'$) , o sea:

$$\left[\mathfrak{I}(\mathbf{B})\right] = diag\begin{bmatrix} \left[\partial^2 L / \partial b_{jk}^2\right] & \left[\partial^2 L / \partial b_{jk} \partial b_{jk'}\right] \\ \left[\partial^2 L / \partial b_{jk'} \partial b_{jk}\right] & \left[\partial^2 L / \partial b_{jk'}^2\right] \end{bmatrix} \quad (9)$$

Bajo esos tres supuestos, el sistema de ecuaciones (4) se transforma en:

$$\langle \hat{\mathbf{B}}^{*\prime} \rangle_{(t+1)} = \langle \hat{\mathbf{B}}^{*\prime} \rangle_t - \left[\Im(\mathbf{B}^*)\right]_t^{-1} \left[\mathbf{U}(\mathbf{B}^*)\right]_t$$
$$\langle \hat{\mathbf{A}}^{*\prime} \rangle_{(t+1)} = \langle \hat{\mathbf{A}}^{*\prime} \rangle_t - \left[\Im(\mathbf{A}^*)\right]_t^{-1} \left[\mathbf{U}(\mathbf{A}^*)\right]_t \qquad (10)$$

La resolución del mismo requiere la sustitución de las derivadas respectivas de acuerdo al tipo de *Biplot de Regresión* (lineal o no lineal) a ajustar.

Una de las ventajas de utilizar este procedimiento de estimación es que se pueden obtener fácilmente las varianzas asintóticas de los estimadores, de requerirse, así como también los estadísticos Deviance y χ^2 para verificar la bondad del ajuste.

El procedimiento de estimación descrito se puede particularizar al caso en que el modelo (3) no considere el término independiente.

5.2. Estimación por etapas

Como una alternativa al método de estimación simultánea, se pueden estimar en dos etapas las matrices de parámetros, tal como en el método de regresiones bilineales segmentadas (Gabriel, 1998), del cual se demuestra que éste es un caso particular.

A tal efecto, el modelo (3) se puede reescribir (siendo $A_o = 1$) como:

$$g(\mu) = \mathbf{A}^* \mathbf{B}^{*\prime} = \mathbf{I}_n \mathbf{A}_o \mathbf{B}'_o \mathbf{I}_p + \mathbf{I}_n \mathbf{A}\mathbf{B}' \mathbf{I}_p = \mathbf{I}_n \mathbf{1}_n \mathbf{B}'_o \mathbf{I}_p + \mathbf{I}_n \mathbf{A}\mathbf{B}' \mathbf{I}_p \quad (11)$$

y realizar las estimaciones a través de las siguientes etapas:

Etapa 1: Se estiman los parámetros de las variables b_{jk} (suponiendo conocidas a priori las coordenadas de las filas a_{ik}), en forma equivalente a la realización de la siguiente regresión generalizada:

$$g(\mu) = \mathbf{A}^* \mathbf{B}^{*\prime} = \mathbf{I}_n \mathbf{1}_n \mathbf{B}'_o \mathbf{I}_p + \mathbf{I}_n \mathbf{A}\mathbf{B}' \mathbf{I}_p \qquad (12)$$

Modelo que se puede reescribir de la siguiente manera, concatenando las columnas de la matriz Y, y considerando al mismo tiempo el producto Kronecker

entre matrices ($A \otimes B = a_{ij} B$) y la concatenación de columnas de matrices producto ($<A B' I> = (I \otimes A) <B'> = [diag (A)] <B'>$):

$$g < \mu >= [(\mathbf{I}_p \otimes \mathbf{1}_n); (\mathbf{I}_p \otimes (\mathbf{I}_n \mathbf{A})] < \mathbf{B}^{*\prime} > \qquad (13)$$

Por lo que la resolución de esa regresión, para las columnas de la matriz Y, es equivalente a la resolución de la primera ecuación del sistema de ecuaciones (10) utilizado en la estimación simultánea, o sea:

$$\langle \hat{\mathbf{B}}^{*\prime} \rangle_{(t+1)} = \langle \hat{\mathbf{B}}^{*\prime} \rangle_t - \left[\Im(\mathbf{B}^*) \right]_t^{-1} \left[\mathbf{U}(\mathbf{B}^*) \right]_t \qquad (14)$$

Etapa 2: Se estiman las coordenadas de los individuos a_{ik}, suponiendo conocidos los parámetros de las variables b_{jk}, mediante a un procedimiento análogo al de la Etapa 1.

Concluyéndose que esta etapa es equivalente a la resolución de la segunda ecuación del sistema de ecuaciones (10), o sea :

$$\langle \hat{\mathbf{A}}^{*\prime} \rangle_{(t+1)} = \langle \hat{\mathbf{A}}^{*\prime} \rangle_t - \left[\Im(\mathbf{A}^*) \right]_t^{-1} \left[\mathbf{U}(\mathbf{A}^*) \right]_t \qquad (15)$$

En el caso del modelo sin término independiente, el procedimiento de estimación es un caso particular del antes descrito, cuando $B_o = 0$.

5.3. Una aplicación

Para evaluar la metodología propuesta, se utiliza el ejemplo de Gower & Hand (1996, 74), comparando los resultados por ellos obtenidos mediante la aplicación de un análisis de correspondencias múltiples, con los aquí obtenidos a través de estimación simultánea.

El ejemplo original consiste de una matriz de datos Y de orden (20x4), que se transforma en una matriz de variables indicatrices de orden (20x16). Las observaciones se realizan sobre 20 granjas y las 16 variables categóricas corresponden a: nivel de humedad (mínimo nivel: H1; H2; H3; máximo nivel: H4), tipo de granja (granja normal: GN; granja biológica: GB; granja hobby: GH; granja conservacionista: GC), uso de la granja (producción: U1; uso intermedio: U2; pasto: U3) y nivel de abono (mínimo nivel: A0; A1; A2; A3; máximo nivel: A4).

Se ajusta a los datos un *Biplot de Regresión no Lineal,* de tipo logístico, para cada una de las categorías, usando el modelo (3) con término independiente y la función link logit, de donde: logit (p) = log [p / (1-p)] = $b_{i0} + a_{i1} b_{j1} + a_{i2} b_{i2}$, siendo $p = (e^{\eta} / 1 + e^{\eta})$.

A tal efecto, se elaboró un programa en MATLAB que utiliza en la estimación el sistema de ecuaciones (10), en el cual se sustituyen las derivadas considerando que las variables tienen distribución binomial (1,p). Los valores iniciales $<A^{0*'}>$ se obtienen mediante la ordenación de las filas de la matriz a través de un Análisis de Coordenadas Principales, aplicado a la matriz de similaridades obtenida del coeficiente de Jaccard (véase Cuadras, 1996, 297) . Con estos valores se calculan los iniciales para $<B^{0*'}>$ y luego se inicia el proceso iterativo de Newton-Raphson hasta que converja. Los resultados se muestran en la tabla 1.

Para medir la bondad del ajuste se usa la Deviance al igual que en los modelos lineales generalizados (sus valores relativos aproximan la calidad de representación de las variables) y el porcentaje de bien clasificados.

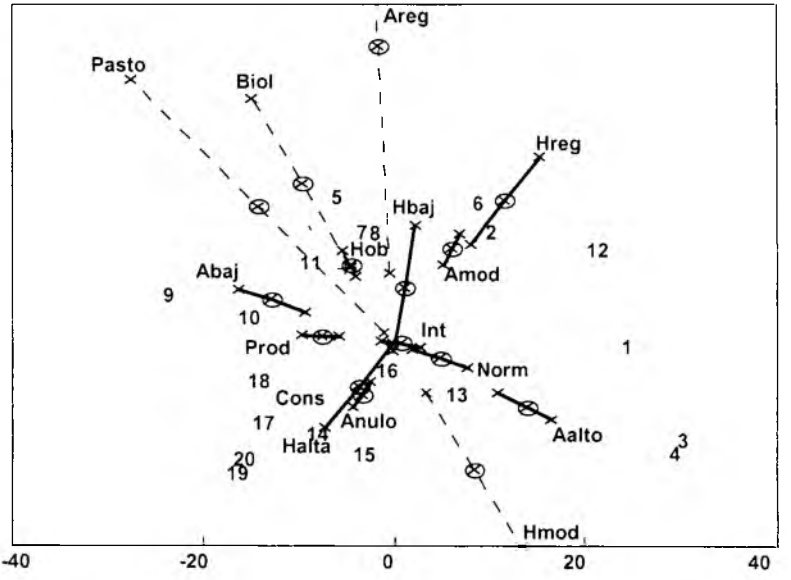
Se observa en la tabla 1 anexa que la Deviance no es significativa (p valor > 0.10), para las variables GB, U1, U2, U3, A3, y A4. Sin embargo, los porcentajes de granjas Bien Clasificados son altos para todas las variables, por lo que la predicción de presencia-ausencia de cada variable, para cada una de las granjas, es bastante aproximada.

Las variables mal representadas son las que coinciden en la figura 3, con los segmentos (pespunteados) de mayor tamaño, como A3, U3, H2 y GB, mientras que las variables bien representadas aparecen con segmentos de menor longitud, como GC, U1, U2, A0, A2 y A4. Esto contrasta con la interpretación habitual de los biplots clásicos, en los que las variables bien representadas tienen segmentos grandes, mientras que las mal representadas aparecen en torno al origen (ello debido al tipo de modelo utilizado en el ajuste).

Tabla 1: Parámetros estimados para las matrices A* y B*, Deviance, p valor, bien clasificados (%B.C), calidad de representación (C.Rx 1000), escalas 1 y 2 de predicción para el percentil 50 (p= 0.5)

Granja	a_{i0}	a_{i1}	a_{i2}	*Variable*	b_{j0}	b_{j1}	b_{j2}	*Deviance*	*P valor*	*%B.C*	*C:R*	*Escala 1*	*Escala 2*
1	1	24.613	-0.372	H1	-0.953	0.015	0.121	8.053	0.0045	80	315	0.96	7.76
2	1	9.479	15.229	H2	-1.809	0.038	-0.082	2.658	0.103	90	197	8.42	-18.16
3	1	30.695	-14.968	H3	-3.564	0.077	0.133	3.404	0.065	90	318	11.62	20.07
4	1	30.695	-14.968	H4	-1.142	-0.081	-0.136	6.87	0.0088	75	422	-3.69	-6.20
5	1	-6.288	20.85	GN	-1.749	0.303	-0.146	3.731	0.0534	95	743	4.68	-2.26
6	1	8.628	19.924	GB	-2.109	-0.034	0.079	2.239	0.1346	85	148	-9.69	22.52
7	1	-2.804	14.822	GH	-6.285	-0.202	0.477	17.552	0.0000	100	825	-4.73	11.17
8	1	-1.832	15.001	GC	-4.382	-0.229	-0.485	17.046	0.0000 0.0000	100	947	-3.49	-7.39
9	1	-23.689	6.767	U1	-4.426	-0.571	0.069	1.532	0.2158	100	868	-7.64	0.92
10	1	-14.749	3.637	U2	-0.378	0.515	-0.118	2.352	0.1251	100	889	0.70	-0.16
11	1	-5.12	10.462	U3	-1.186	-0.029	0.04	0.966	0.3257	75	65	-14.09	19.43
12	1	21.625	12.473	A0	-4.382	-0.229	-0.485	17.046	0.0000	100	947	-3.49	-7.39
13	1	7.162	-7.074	A1	-4.077	-0.259	0.122	3.668	0.0555	90	524	-12.88	6.07
14	1	-7167	-12.841	A2	-6.754	0.189	0.423	15.855	0.0001	100	851	5.95	13.31
15	1	-1.933	-15.647	A3	-1.434	-0.001	0.034	0.649	0.4206	80	32	-0.001	42.14
16	1	-0.645	-4.128	A4	-5.477	0.277	-0.181	0.2156	0.2156	100	935	13.86	-9.05
17	1	-11.666	-11.127										
18	1	-12.955	-6.025										
19	1	-15.558	-16.954										
20	1	-15.558	-16.954										

Figura 3: Representación del biplot de regresión logístico



La dirección de los ejes biplot sobre la representación gráfica, está determinada por los parámetros b_{jk} estimados. La asociación entre las distintas variables, puede aproximarse a través del ángulo que forman entre sí. El producto escalar entre marcadores fila a_i y marcadores columna b_j, aproxima salvo un factor de escala la probabilidad (p_{ij}) de cada categoría para cada una de las 20 granjas. Procedimiento que se puede abreviar introduciendo en los ejes biplot escalas de predicción, que permitan visualmente predecir la presencia o ausencia de determinado carácter o categoría.

A tal efecto y considerando la geometría de los *Biplots Logísticos*, se obtienen las escalas de predicción sobre cada eje biplot (escalas 1 y 2 en la tabla 1) y se hacen marcas en los percentiles 25, 50 y 75. Para ello se parte de un punto cualquiera (y_1, y_2) sobre el eje biplot, representado a través de la recta que pasa por los puntos (0,0) y (b_{j1}, b_{j2}), o sea $y_2 = (b_{j1}/b_{j2})y_1$, para luego sustituir esa expresión en el modelo utilizado (logit (p) = $b_{j0} + b_{j1}\, y_1 + b_{j2}\, y_2$), de donde finalmente despejando se obtienen las escalas sobre la representación (para p = 0.25, 0.5, 0.75). Si la proyección de un punto fila a_{ik} cae por debajo de la marca central (percentil 50), la predicción será ausencia de esa característica específica, y en caso contrario, la predicción será presencia.

Las conclusiones obtenidas de la figura 3 son similares a las obtenidas por Gower & Hand (1996) y por Vicente-Villardón y Galindo (1998), reflejando fielmente la estructura de la matriz de datos original, o sea:

> Las granjas normales GN aparecen ubicadas en la región con niveles de humedad moderados H2, alto uso de abono A4 y uso de la granja U2. Las granjas biológicas GB y las granjas para hobby GH aparecen en la región de bajo uso de abono A1, humedad baja H1 y uso de la granja para la producción U1 y para pasto U3. Las granjas conservacionistas GC están en la región no abonada A0 y alta humedad H4.

Las predicciones de ausencia obtenidas para las variables A3, U3, H2 y GB, se deben utilizar con precaución por encontrarse mal representadas.

6. DISCUSIÓN

Se demuestra que es posible realizar el ajuste de un biplot a través de modelos bilineales generalizados de tipo multiplicativo, siendo su geometría similar a la de un ajuste de regresión lineal, sin embargo, dependiendo su interpretación de la función link utilizada en el ajuste, la cual contrasta con la de los biplots clásicos.

En la estimación se puede generalizar el método iterativo utilizado en los modelos bilineales generalizados, el cual se puede simplificar haciendo supuestos en la misma forma que en Teoría de Respuesta al Item, demostrando así su aplicabilidad en ese campo de las Ciencias Sociales, pudiéndose describir del gráfico biplot algunas variables latentes.

El método de estimación propuesto también puede ser de mucha utilidad práctica, para ordenar los individuos de acuerdo a ciertas variables externas, en el sentido del Análisis Canónico de Correspondencias (Ter Braak, 1986), ya que su generalización permite restringir los ejes en el ajuste para que sean combinaciones lineales de esas variables externas (Cárdenas, 2000, 123-174).

REFERENCIAS BIBLIOGRÁFICAS

Blázquez, A. (1998), *Análisis biplot basado en modelos lineales generalizados*, Tesis Doctoral, Universidad de Salamanca, España.

Baker, Frank B. (1992), *Item Response Theory*, Marcel Dekker, Inc. New York.

Bradu, D. & Gabriel, K. R. (1978), "The Biplot as a Diagnostic Tool for Models of Two-Way

Tables", *Technometrics* 20, (1), 47-68.

Cárdenas, O. C. (2000), *Biplot con información externa basado en modelos lineales generalizados*, Tesis Doctoral, Universidad de Salamanca, España.

Cuadras, C. (1996), *Métodos de análisis multivariante,* EUB, S.L., Barcelona.

Choulakian, V. (1996), "Generalized Bilinear Models", *Psychometrika* 61, (2), 271- 283.

Denis, J. B (1991), "Ajustements de Modelles Lineaires et Bilineaires sous Constraintes Lineaires avec Donnes Manquantes", *Statistique Applique*, XXXIX (2), 5-24.

Eckart, C. & Young, G. (1936), "The approximation of one matrix by another of lower rank", *Psychometrika*, 1, 211-18.

Falguerolles, A. (1995), "Generalized Bilinear Models and Generalized Biplots: Some Examples", *Publications du Laboratoire de Statistique et Probabilites*. Université Paul Sabatier, Toulouse.

Gabriel, K. R (1971), "The Biplot-graphic display of matrices with applications to principal component analysis", *Biometrika* 58, 453-467.

—(1998), "Generalised Bilinear Regression", *Biometrika*, 85, 3, 689-700.

—, Galindo, M. P. y Vicente-Villardón, J. L. (1998), "Use of Biplots to diagnose Independence Models in Three-Way Contingency Tables", In, M. Greenacre and J. Blasius (eds.), *Visualization of Categorical Data,* Academic Press, London

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

Gollob, H. (1968), "A statistical model wich combines features of factor analytic and analysis of variance techniques", *Psichometrika,* 33: 73-115.

Gower, J. C. (1992), "Generalized Biplots", *Biometrika* 79, 475-493.

— & Harding, S. (1988), "Nonlinear Biplots", *Biometrika,* 75, 445-455.

Gower, J. C. & Hand, D. J. (1996), *Biplots,* Chapman & Hall, London.

Householder, A. S. & Young, G. (1938), "Matrix Approximation and Latent Roots", *American Mathematics Monthly,* 45: 165-171.

Nelder, J. A., & Wedderburn, R. W. (1972), "Generalized Linear Models", *Journal of the Royal Statistical Society* A, 135, 370-384.

Ter Braak, C.J.F. (1986), "Canonical Correspondence Analysis: a new eigenvector technique for Multivariate Direct Gradient Analysis", *Ecology* 67 (5), 1167-1179.

Van Eeuwijk, F. (1995), "Multiplicative Interaction in Generalized Linear Models", *Biometrics*, 51, 1017-32

Vásquez, M. (1995), *Aportaciones al análisis biplot: Un enfoque algebraico,* Tesis Doctoral, Universidad de Salamanca, España.

Vicente-Villardón, J. L. & Galindo, M. P. (1998), *Biplot externo para datos presencia-ausencia basado en superficies logísticas de respuesta*, Departamento de Estadística, Universidad de Salamanca, España.

— y Cárdenas, O. (2000), "Biplot para Datos Binarios basado en Modelos Logísticos de Respuesta", *XXV Congreso Nacional de Estadística e Investigación Operativa,* Servicio de Publicaciones, Universidad de Vigo, 269-270.