BIPROB: UN MÉTODO PARA OBTENER UN BIPLOT ROBUSTO

Sergio Hernández González [(1)] y Ma. Purificación Galindo Villardón [(2)]
(1) Facultad de Estadística e Informática, Universidad Veracruzana, Xalapa, Veracruz, México
(2) Departamento de Estadística, Universidad de Salamanca, Salamanca, España.

RESUMEN
El Biplot clásico está basado en la descomposición en valores singulares (DVS) de la matriz de datos $\mathbf{X}$. Convencionalmente se construye a partir de marcadores obtenidos de un Análisis de Componentes Principales (ACP), utilizando $\mathbf{X}^T\mathbf{X}$. Esta solución es susceptible ante la presencia de outliers en $\mathbf{X}$. En este trabajo se presenta un método para obtener un Biplot Robusto (Hernández, 2005), mediante regresiones alternadas, con lo cual se resuelve la potencial influencia de valores outliers en $\mathbf{X}$.

ABSTRACT
Biplot classic is based on the singular values decomposition (SVD) of the matrix of data $\mathbf{X}$. Conventionally it is constructed from obtained markers of an Principal Component Analysis (PCA) using $\mathbf{X}^T\mathbf{X}$. This solution is susceptible to the presence of outliers in $\mathbf{X}$. In this work a method to obtain a Robust Biplot (Hernández, 2005), by means of alternate regressions, is presented, with which the potential influence of outliers in $\mathbf{X}$ is solved.

Key words: singular values decomposition, principal component analysis, outliers, regressions.

MSC: 62H25

`1. INTRODUCCIÓN

El Biplot clásico (Gabriel, 1971) es un método utilizado para visualizar de manera conjunta las filas y las columnas de una matriz de datos $\mathbf{X}$. Las dos factorizaciones Biplot propuestas por Gabriel fueron denominadas: GH-Biplot y JK-Biplot. La primera consigue una alta calidad en la representación de las columnas (variables) y no tan alta para las filas (individuos); mientras que la segunda consigue una alta calidad de representación para las filas, y no tan alta para las columnas. Galindo (1985) propone una elección de los marcadores para representar las filas y las columnas simultáneamente sobre un mismo sistema de coordenadas, obteniendo una alta calidad de representación tanto para las filas como para las columnas, al cual denomina HJ-Biplot.

Como en la mayoría de las técnicas de análisis de datos, partimos de una matriz de datos $\mathbf{X}$ de n filas y p columnas, las cuales por lo general representan a n individuos a los que se les observan p variables. El objetivo es representar las filas y columnas de la matriz de datos en un espacio de dimensión reducida, con la pérdida mínima de información.

La base teórica sobre la que se sustenta el Biplot clásico es la DVS de la matriz $\mathbf{X}$, para obtener así su mejor aproximación mínimo cuadrática de rango menor. La matriz de datos $\mathbf{X}$ de orden nxp y de rango r, puede descomponerse según la DVS y siguiendo a Eckart y Young (1936), en la forma:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} = \sum_{\alpha=1}^{r} \sqrt{\lambda_{\alpha}}\,\mathbf{u}_{\alpha}\mathbf{v}_{\alpha}^{\mathrm{T}} \ ,$$

donde $\mathbf{U}$ es una matriz, cuyos vectores columna $\mathbf{u}$ son ortonormales y vectores propios de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$; $\mathbf{D}$ es una matriz diagonal de los valores singulares de $\mathbf{X}$, que son las raíces cuadradas no negativas de los valores propios λ_{α} de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$; y $\mathbf{V}$ es una matriz ortogonal cuyos vectores columna $\mathbf{v}$ son vectores propios de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Si se consideran los d primeros sumandos de esta descomposición, se obtiene una aproximación de la matriz $\mathbf{X}$ que coincide con su mejor aproximación mínima cuadrática en rango d (Eckart y Young, 1936); esto es:

$$\mathbf{X} \cong \mathbf{X}_{(d)} = \mathbf{U}_{(d)}\mathbf{D}_{(d)}\mathbf{V}_{(d)}^{\mathrm{T}} = \sum_{\alpha=1}^{d} \sqrt{\lambda_{\alpha}}\,\mathbf{u}_{\alpha}\mathbf{v}_{\alpha}^{\mathrm{T}} \ .$$

Por lo tanto, la mejor aproximación mínima cuadrática de rango d para un elemento cualquiera de la matriz $\mathbf{X}$ será de la forma:

$$x_{ij(d)} = \sum_{\alpha=1}^{d} \sqrt{\lambda_{\alpha}}\,\mathbf{u}_{i,\alpha}\mathbf{v}_{j,\alpha}^{\mathrm{T}}$$

La técnica del Biplot plantea aproximar la matriz $\mathbf{X}$ mediante $\mathbf{X}_{(d)}$, pero de tal forma que quede factorizada de la siguiente manera:

$$\mathbf{X}_{(d)} = \mathbf{A}_{(d)}\mathbf{B}_{(d)}^{\mathrm{T}} \ ,$$

donde $\mathbf{A}_{(d)}$ y $\mathbf{B}_{(d)}$ son matrices de rango completo por columnas, definidas como: $\mathbf{A}_{(d)} = \mathbf{U}_{(d)}\mathbf{D}_{(d)}^{\rho}$ y $\mathbf{B}_{(d)} = \mathbf{V}_{(d)}\mathbf{D}_{(d)}^{1-\rho}$, para distintos valores de la constante ρ, $0 \le \rho \le 1$.

Teniendo en cuenta estas ideas se replantea el problema partiendo de un análisis Biplot entendido como un modelo bilineal. Supongamos que disponemos de una matriz de datos $\mathbf{X}$, cuyos valores x_{ij} son observaciones de una variable aleatoria $\mathbf{X}$, clasificada con respecto a dos factores de variación. Esta matriz $\mathbf{X}$ contiene I filas correspondientes a los niveles del primer factor estudiado y J columnas que se corresponderán con los niveles del segundo factor considerado, siendo cada elemento x_{ij} la respuesta para la combinación de los niveles i-ésimo del factor en filas y j-ésimo del factor en columnas. En la literatura clásica los modelos bilineales quedan formulados genéricamente de la siguiente forma:

$$x_{ij} = \mu + \alpha_i + \beta_j + \sum_{l=1|}^{k} \boldsymbol{\lambda}_{jl}\mathbf{f}_{il} \ ,$$

donde $\boldsymbol{\lambda}_{jl}$ es el vector de cargas y $\mathbf{f}_{il}$ es el vector de marcadores.

Según esta formulación, tenemos un efecto global común a todas las observaciones μ; los efectos del factor fila α_i y del factor columna β_j; la interacción viene formulada mediante k sumandos distintos, cada uno de los cuales formula la interacción multiplicativamente a través de I+J parámetros, I para las filas y J para las columnas, correspondientes a los niveles de ambos factores.

Varios modelos pueden ser postulados para estimar x_{ij}. Básicamente, se pueden distinguir dos situaciones, según que el modelo considere como única fuente de variación los efectos principales y cualquier otra fuente de variación pase a formar parte del error experimental, o bien, que se postule algún tipo de interacción entre los factores. En el primer caso, en el cual no se postula ningún tipo de interacción entre los factores, el modelo quedaría de la siguiente manera:

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \ ;$$

es decir, los efectos pueden considerarse aditivos. A este modelo se le conoce como *Modelo Aditivo*. Bradu y Gabriel (1974) demuestran que si los datos se ajustan a un modelo aditivo, los marcadores que representan las filas están sobre una línea recta, los marcadores que representan las columnas están sobre otra recta y, además, estas rectas son perpendiculares. En algún sentido el efecto interacción puede ser considerado como una medida de no aditividad entre los factores; es decir, entre los efectos principales. En la mayor parte de las aplicaciones prácticas, no es posible saber a priori si un experimento debe considerar un término de interacción o no.

Presentaremos algunos modelos que postulan de manera sencilla algún tipo de interacción y pueden, por tanto, ser representados en rango dos. Es decir, sus geometrías necesitan únicamente de dos direcciones en el espacio que nos muestran la información original de la matriz de datos.

Tukey (1949) desarrolló un test, conocido como test con un grado de libertad para no aditividad, para diferenciar de algún modo entre el modelo aditivo anteriormente descrito y aquel que de forma sencilla postula la existencia de un término de interacción con un solo grado de libertad. La formulación clásica de este modelo, viene expresada de la siguiente forma:

$$x_{ij} = \mu + \alpha_i + \beta_j + \lambda\alpha_i\beta_j + e_{ij} \ .$$

Por lo tanto, hacer hipótesis sobre si el valor de λ es cero o no, es equivalente a decir si $\mathbf{X}$ sigue un modelo aditivo. A este modelo se le conoce como *Modelo Concurrente* de Tukey. Bradu y Gabriel (1974) demuestran que si los datos se ajustan a un modelo concurrente, los marcadores que representan las filas están sobre una línea recta, los marcadores que representan las columnas están sobre otra recta y, además, estas rectas no son perpendiculares.

Mandel (1961) propuso dos modelos para no aditividad conocidos como *Modelos de Regresión para Filas* y *Modelo de Regresión para Columnas*, los cuales pueden ser considerados como extensiones del modelo concurrente de Tukey; el primero está dado por la siguiente expresión:

$$x_{ij} = \mu + \alpha_i + \beta_j + \delta_i\beta_j \ ,$$

y el segundo por:

$$x_{ij} = \mu + \alpha_i + \beta_j + \alpha_i\gamma_j \ .$$

Bradu y Gabriel (1974) demuestran que si los datos se ajustan a un modelo de regresión para las filas, los marcadores que representan las columnas están sobre una línea recta. Si los

datos se ajustan a un modelo de regresión para las columnas, los marcadores que representan las filas están sobre una línea recta.

Gollob (1968), propone el modelo conocido como Modelo FANOVA, el cual combina aspectos del análisis de la varianza y del análisis factorial; este modelo tiene la siguiente expresión:

$$x_{ij} = \mu + \alpha_i + \beta_j + \sum_{i=1}^{k} \lambda_{jl} \mathrm{f}_{il} \ ,$$

donde μ es un efecto global común a todas las observaciones; α_i el efecto fila (individuos); β_j el efecto columna (variables) y $\sum_{l=1}^{k} \lambda_{\mathrm{jl}} \mathrm{f}_{\mathrm{il}}$ representa la interacción entre las filas y las columnas, y es factorizada como el producto escalar entre el vector de cargas $\boldsymbol{\lambda}_{\mathrm{j}\circ} = \left(\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jk}\right)^{\mathrm{T}}$ y un vector de marcadores $\mathbf{f}_{\mathrm{i}\circ} = \left(\mathrm{f}_{\mathrm{i}1}, \mathrm{f}_{\mathrm{i}2}, \ldots, \mathrm{f}_{\mathrm{ik}}\right)^{\mathrm{T}}$.

Denotando por $\boldsymbol{\theta} = \left(\mathbf{f}_{1\circ}^{\mathrm{T}}, \ldots, \mathbf{f}_{n\circ}^{\mathrm{T}}, \boldsymbol{\lambda}_{1\circ}^{\mathrm{T}}, \ldots, \boldsymbol{\lambda}_{p\circ}^{\mathrm{T}}\right)$ al vector de todos los marcadores y cargas, entonces, podemos decir que:

$$\hat{x}_{ij}\left(\boldsymbol{\theta}\right) = \sum_{l=1}^{k} \mathbf{f}_{\mathrm{il}} \, \boldsymbol{\lambda}_{\mathrm{lj}} = \mathbf{f}_{i\circ}^{\mathrm{T}} \, \boldsymbol{\lambda}_{\mathrm{j}\circ} = \boldsymbol{\lambda}_{\mathrm{j}\circ}^{\mathrm{T}} \mathbf{f}_{\mathrm{i}\circ}$$

será el valor aproximado de x_{ij} de acuerdo a la versión muestral de un modelo factorial.

Tomando $k = 2$ y representando en la misma gráfica bidimensional las filas $\left(\hat{\mathbf{f}}_{\mathrm{i}1}, \hat{\mathbf{f}}_{\mathrm{i}2}\right)$ y las columnas $\left(\hat{\boldsymbol{\lambda}}_{\mathrm{j}1}, \hat{\boldsymbol{\lambda}}_{\mathrm{j}2}\right)$; el resultado será un Biplot.

Ahora bien, consideremos la descomposición Biplot, como si fuera un Análisis Factorial, en la forma siguiente:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{E} \ ,$$

donde $\mathbf{A} = \mathbf{U}\mathbf{D}$ (matriz de marcadores), $\mathbf{B} = \mathbf{V}^{\mathrm{T}}$ (matriz de cargas) y $\mathbf{E}$ es una matriz de residuos. Una estimación de $\mathbf{X}$, la cual se puede obtener mediante la DVS de $\mathbf{X}$, sería:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} = \sum_{\alpha=1}^{r} \sqrt{\lambda_{\alpha}} \, \mathbf{u}_{\alpha} \mathbf{v}_{\alpha}^{\mathrm{T}} \ .$$

Si se consideran los k primeros sumandos de esta descomposición, se obtiene una aproximación de la matriz $\mathbf{X}$ que coincide con su mejor aproximación mínimo cuadrática en rango k (Eckart y Young, 1936). Esto es:

$$\hat{\mathbf{X}} \cong \mathbf{X}_{(k)} = \mathbf{U}_{(\mathrm{k})}\mathbf{D}_{(\mathrm{k})}\mathbf{V}_{(k)}^{\mathrm{T}} = \sum_{\boldsymbol{\alpha}=\mathbf{1}}^{k} \sqrt{\lambda_{\alpha}} \, \mathbf{u}_{\alpha} \mathbf{v}_{\alpha}^{\mathrm{T}} \ .$$

Convencionalmente la DVS es calculada por medio de un ACP de $\mathbf{X}'\mathbf{X}$, pero tanto la DVS, como el ACP, son muy susceptibles a la presencia de outliers, por lo que se deberá de contar con otras alternativas, las cuales sean robustas ante la presencia de los outliers.

En este trabajo se presenta una alternativa, para obtener un Biplot Robusto (Hernández, 2005), la cual hace uso de un procedimiento iterativo conocido como regresiones alternadas; éste fue originalmente propuesto por Wold (1966a y 1966b), también llamado regresiones criss-cross (Gabriel y Zamir, 1979). A este tipo de algoritmo se le conoce en la literatura como Algoritmo L1 (Martens y Naes, 1989). Utilizando el Algoritmo L1 se obtendrán estimadores, tanto para la matriz de marcadores $\mathbf{A}$, como para la matriz de cargas $\mathbf{B}$, en forma robusta. Se realiza un Biplot, tomando como marcadores para las filas a la matriz $\mathbf{A}$, y como marcadores para las columnas a la matriz $\mathbf{B}$. El resultado será, lo que se ha denominado Biplot Robusto (Hernández, 2005), el cual es resistente ante la presencia de outliers.

2 BIPLOT CLÁSICO

Para simplificar la explicación a partir de esta sección se obviara el subíndice (d). Las filas de la matriz $\mathbf{A}$ y las columnas de la matriz $\mathbf{B}$ se pueden interpretar como coordenadas de puntos en un espacio Euclídeo, referidas a los mismos ejes ortogonales. El producto interno entre filas de $\mathbf{A}$ y las de $\mathbf{B}$ proporcionan los valores x_{ij} por medio de marcadores $\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n$ para las filas y $\mathbf{b}_1,\mathbf{b}_2,\ldots,\mathbf{b}_p$ para las columnas, de tal forma que el producto interno $\mathbf{a}_i^{\mathrm{T}}\mathbf{b}_j$ representa al elemento x_{ij} según la expresión: $x_{ij}=\mathbf{a}_i^{\mathrm{T}}\mathbf{b}_j=\|\mathbf{a}_i\|\,\|\mathbf{b}_j\|\cos\Theta$, donde Θ es el ángulo entre $\mathbf{a}_i$ y $\mathbf{b}_j$.

Dependiendo del valor seleccionado de ρ se obtienen el Biplot clásico (Gabriel, 1971): GH-Biplot ($\rho=0$), JK-Biplot ($\rho=1$) y SQRT-Biplot ($\rho=1/2$). En el primero, el GH-Biplot, la elección de los marcadores es de la forma $\mathbf{A}=\mathbf{U}$ para las filas y $\mathbf{B}=\mathbf{VD}$ para las columnas, por lo tanto: $\mathbf{X}\cong\mathbf{AB}^{\mathrm{T}}=\mathbf{UDV}^{\mathrm{T}}$. Las columnas de $\mathbf{A}$ son ortonormales y las de $\mathbf{B}$ son ortogonales, imponiéndose la métrica $\mathbf{B}^{\mathrm{T}}\mathbf{B}=\mathbf{I}$, para que la representación sea única, excepto por rotaciones. En este tipo de Biplot, se preserva la distancia Euclídea entre las columnas, por lo que esta representación es usada cuando se estudia el comportamiento de las variables y el papel de los individuos es secundario. En el segundo, el JK-Biplot, los marcadores son: $\mathbf{A}=\mathbf{UD}$ y $\mathbf{B}=\mathbf{V}$. Se impone la métrica $\mathbf{B}^{\mathrm{T}}\mathbf{B}=\mathbf{I}$, por lo que la representación, también es única, excepto por rotaciones. Para adecuarse al nombre de este tipo de Biplot, se adopta la siguiente terminología para las matrices de marcadores fila $\mathbf{J}$ y marcadores columna $\mathbf{K}$, de tal forma que: $\mathbf{J}=\mathbf{UD}$ y $\mathbf{K}=\mathbf{V}$. En este tipo de Biplot solamente las filas presentan una calidad de representación aceptable, con lo que se consigue que la distancia entre individuos aproxime la similitud o disimilitud entre los mismos. Este tipo de representación es útil cuando se estudia el comportamiento de los individuos y el papel de las variables es secundario. En el tercero, el SQRT-Biplot, la factorización en este caso es de la forma: $\mathbf{A}=\mathbf{UD}^{1/2}$ y $\mathbf{B}=\mathbf{VD}^{1/2}$. Este tipo de representación se utiliza cuando el objetivo fundamental es la aproximación de los elementos de la matriz, y no el análisis de las características de las filas y las columnas por separado.

El HJ-Biplot (Galindo, 1985), posee importantes propiedades, como por ejemplo, que la calidad de representación para las filas y para las columnas de la matriz de datos es la misma. La elección de los marcadores es la siguiente: $\mathbf{A}=\mathbf{UD}$ y $\mathbf{B}=\mathbf{VD}$. Para adecuarse al nombre de este tipo de Biplot, se adopta la siguiente terminología para las matrices de marcadores fila y marcadores columna: $\mathbf{J}=\mathbf{UD}$ y $\mathbf{H}=\mathbf{VD}$. En este tipo de representación ya no se aproximan los elementos de la matriz original, sino que se interpretan las relaciones fila-columna a través de los ejes factoriales. Debido a que tanto las filas como las columnas tienen la misma calidad de representación, se pueden interpretar las posiciones de las filas, de las columnas y las relaciones fila-columna a través de las contribuciones relativas del factor al elemento y del elemento al factor (Galindo y Cuadras, 1986).

3 BIPLOT ROBUSTO (BIPROB)

Esta alternativa está basada en la siguiente idea. Supongamos, que las coordenadas para las filas de $\mathbf{A}$ están fijadas de antemano, tenemos entonces que las coordenadas para las columnas pueden calcularse como la matriz $\mathbf{B}$ que hace mínima la suma de cuadrados de los residuos dada por la siguiente expresión:

$$\mathbf{R} = \left\| \mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}} \right\| = \text{traza}\left[\left(\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}} \right)^{\mathrm{T}} \left(\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}} \right) \right],$$

cuya solución viene dada por la matriz:

$$\mathbf{B}^{\mathrm{T}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} \right)^{-1} \mathbf{A}^{\mathrm{T}}\mathbf{X} \ ;$$

es decir, las filas de $\mathbf{B}$ son los coeficientes de regresión obtenidos en la regresión de cada columna de la matriz original sobre las columnas de $\mathbf{A}$.

De la misma manera, si escribimos:

$$\mathbf{X}^{\mathrm{T}} = \mathbf{B}\mathbf{A}^{\mathrm{T}} + \mathbf{E}^{\mathrm{T}},$$

y fijamos ahora los valores de $\mathbf{B}$, podemos obtener los valores para $\mathbf{A}$ que hacen mínima la suma de cuadrados de los residuales dada por la siguiente expresión:

$$\mathbf{R} = \left\| \mathbf{X}^{\mathrm{T}} - \mathbf{B}\mathbf{A}^{\mathrm{T}} \right\| = \text{traza}\left[\left(\mathbf{X}^{\mathrm{T}} - \mathbf{B}\mathbf{A}^{\mathrm{T}} \right)^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}} - \mathbf{B}\mathbf{A}^{\mathrm{T}} \right) \right],$$

cuya solución viene dada por la matriz siguiente:

$$\mathbf{A}^{\mathrm{T}} = \left(\mathbf{B}^{\mathrm{T}}\mathbf{B} \right)^{-1} \mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} \ .$$

Partiendo de valores iniciales arbitrarios para $\mathbf{A}$ (ó $\mathbf{B}$) se puede construir un algoritmo con el que se obtienen los mismos valores que con la DVS. Para nuestro caso, se toma como inicio de $\mathbf{A}$ (matriz de marcadores) a la primera columna de $\mathbf{X}$ y se realiza una regresión para estimar la primera columna de $\mathbf{B}$ (matriz de cargas). Es decir:

$$\hat{\mathbf{B}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A} \ ;$$

se normaliza a $\hat{\mathbf{B}}$, esto es: $\hat{\mathbf{B}} = \dfrac{\hat{\mathbf{B}}}{\text{norma}(\hat{\mathbf{B}})}$; se hace una regresión para estimar la primera columna de $\mathbf{A}$, mediante la estimación obtenida de $\mathbf{B}$, con lo que:

$$\hat{\mathbf{A}} = (\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}})^{-1}\mathbf{X}\hat{\mathbf{B}} \ .$$

Ahora, se evalúa la suma de las desviaciones absolutas de los residuales, y se verifica si el proceso converge; es decir, si dicha suma de las desviaciones absolutas de los residuos ya es mínima (el nivel de tolerancia que se utiliza en este trabajo es de $1 * \mathrm{e}^{-12}$). Esto es, se verifica si:

$$\hat{\boldsymbol{\theta}}_{DVSR} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j=1}^{p} \left| x_{ij} - \hat{x}_{ij}(\boldsymbol{\theta}) \right| < 1 * \mathrm{e}^{-12} \ .$$

Cuando se haya cumplido la condición, entonces, ya se tiene la estimación de la primera columna de $\mathbf{A}$ (matriz de marcadores) y de $\mathbf{B}$ (matriz de cargas). Se estima $\mathbf{X}$ tomando los valores estimados de $\mathbf{A}$ y de $\mathbf{B}$ de la siguiente manera:

$$\hat{\mathbf{X}} = \mathbf{X} - \hat{\mathbf{A}}\hat{\mathbf{B}}^{\mathrm{T}};$$

Con lo que, se procede a estimar la segunda columna de $\mathbf{A}$ y de $\mathbf{B}$, tomando como inicio de $\mathbf{A}$ a la segunda columna de $\hat{\mathbf{X}}$ y se realiza una regresión para estimar la segunda columna de $\mathbf{B}$, tal y como se hizo para encontrar la primera columna. El proceso se continua hasta encontrar las k columnas, tanto de $\mathbf{A}$, como de $\mathbf{B}$. El valor de k es el número de componentes principales. Para seleccionar el número de componentes principales se realiza un gráfico de los valores propios λ_i frente a i. La idea es buscar un "codo" en el gráfico, es decir, un punto a partir del cual los valores propios son aproximadamente iguales. El criterio es quedarse con el número de componentes que excluya los asociados a valores pequeños y aproximadamente del mismo tamaño. A este tipo de procedimiento de factorizar a la matriz de datos se le conoce en la literatura como Algoritmo L1. Este algoritmo L1 plantea encontrar todas las columnas de $\mathbf{A}$, así como todas las columnas de $\mathbf{B}$, es estable y fácil de programar. Entonces, se realiza un Biplot, tomando como marcadores para las filas a la matriz $\mathbf{A}$, y como marcadores para las columnas a la matriz $\mathbf{B}$. El resultado será, lo que hemos denominado BIPROB (Hernández, 2005), el cual es resistente ante la presencia de outliers.

4 APLICACIÓN PRÁCTICA

Para mostrar las ventajas que presenta el Biplot Robusto, con respecto al clásico, se presenta una aplicación con datos reales. Se utiliza el Índice Metropolitano de la Calidad del Aire (IMECA) de la Ciudad de México, que se calcula de acuerdo a cinco contaminantes criterio del aire: i) Monóxido de Carbono (CO), ii) Bióxido de Azufre (SO2), iii) Ozono (O3), iv) Bióxido de Nitrógeno (NO2) y v) Partículas Suspendidas Fracción Aire (PM10). El Sistema de Monitoreo Atmosférico (SIMA) de la Zona Metropolitana de la Ciudad de México reporta el IMECA en 5 subzonas: Noroeste (NO), Noreste (NE), Centro (CE), Suroeste (SO) y Sureste (SE).

Se tomaron los datos (día/hora) del IMECA del mes de mayo del año 2003 para la subzona Noroeste (RAMA, 2003) de la siguiente dirección electrónica http://148.243.232.103/imecaweb/base_datos-htm. La matriz que se tenía para analizar estaba compuesta por 744 filas (individuos - día/hora) y 5 columnas (variables - contaminantes). Como el primer día del mes de mayo de ese año fue jueves, se decidió tomar los primeros 28 días para así tener cuatro semanas completas. Así la matriz quedó conformada de 672 filas y 5 columnas.

Se desarrolló un programa en Matlab Ver. 6.5, denominado BIPROB-2004 (Hernández, 2005), el cual realiza los cálculos para obtener, tanto las representaciones para el Biplot clásico, como la representación para el Biplot Robusto propuesto en este trabajo. Asimismo, dicho programa contempla la opción de encontrar outliers multivariantes, mediante tres formas: 1) Forma Clásica (utilizando la distancia de Mahalanobis -Mahalanobis 1939-), 2) Mediante el procedimiento propuesto por Rousseeuw y van Driessen (1999) que denominamos MCD, y 3) Mediante el procedimiento propuesto por Peña y Prieto (2001) que llamaremos KUR.

La Figura 1 presenta la gráfica de la serie cronológica de los 5 contaminantes. En ella podemos darnos cuenta que solamente el Ozono y las Partículas Suspendidas rebasan el nivel de IMECA satisfactorio (100 puntos) y que ninguna rebasa los 200 puntos (nivel malo).

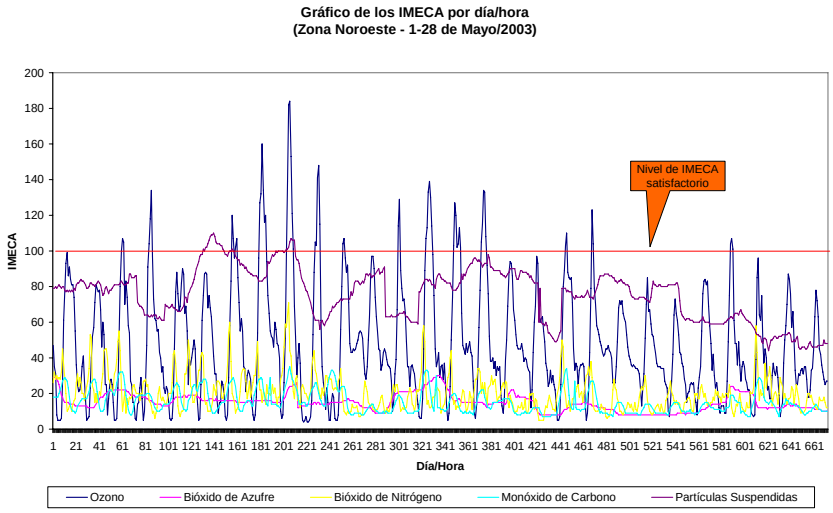

Figura 1.- Gráfica de los IMECA por día/hora de los contaminantes

4.1 Identificación de outliers multivariantes

En primer lugar detectamos los posibles outliers. La Figura 2 muestra los posibles día/hora que se pueden considerar como outliers. Nótese la diferencia, tan grande, entre ellos. Mientras que, con el Algoritmo Clásico (Mahalanobis, 1939), solamente se detectaron 4 outliers: el día 9 a las 12, 13 y 14 horas y el día 26 a las 10 horas. Con el Algoritmo KUR (Peña y Prieto, 2001) se detectaron 27 outliers; y con el Algoritmo MCD (Rousseeuw y van Driessen, 1999) se detectaron 40 outliers.

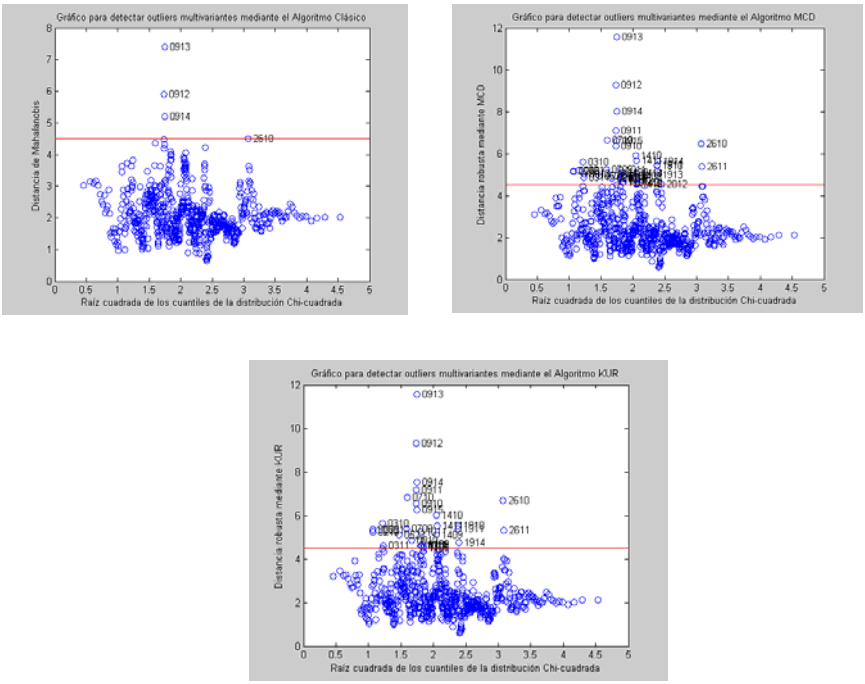

Figura 2.- Identificación de outliers mediante los tres procedimientos

4.2 Análisis exploratorio mediante el biplot clásico

En el Biplot clásico (GH-Biplot y JK-Biplot) así como en el HJ-Biplot, se obtuvo una calidad de representación global de 98.57 %. Todos los individuos (día/hora), se encontraban bien representados en el primer eje. De igual forma, todas las variables (contaminantes), se encontraban bien representadas en el primer eje.

Nos centraremos en analizar solamente dos contaminantes, el Ozono y las Partículas Suspendidas, dado que son los únicos que rebasaron el nivel máximo de puntos IMECA permitidos (100), para considerar un estado normal de contaminación.

La Figura 3 presenta el gráfico GH-Biplot, pero existen muchos día/hora (672) presentados en dicho gráfico, lo cual dificulta la identificación de aquellos día/hora que tienen mayor nivel de IMECA para cada uno de los dos contaminantes. Se presentan otros dos gráficos, en la misma Figura 3, donde se realiza un acercamiento del gráfico GH-Biplot, para visualizar en mejor forma los día/hora con mayor nivel de IMECA para el Ozono, y una adecuación al mismo gráfico GH-Biplot, para observar mejor los día/hora con mayor nivel de IMECA para las Partículas Suspendidas. Podemos observar en dicha figura que concuerdan los día/hora con mayor nivel de IMECA para el Ozono, más no así con respecto a lo que sucede con las Partículas Suspendidas, ya que no concuerdan con los día/hora con mayor nivel de IMECA, en particular dos días: 06/18 y 06/19.

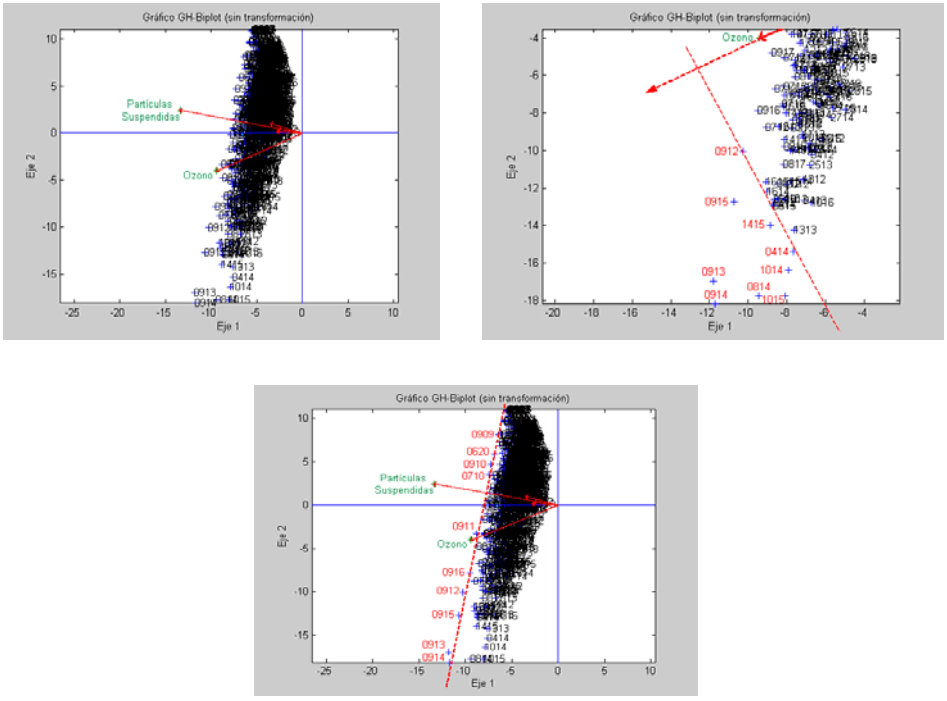

Figura 3.- Gráficos GH-Biplot

La Figura 4 presenta el gráfico JK-Biplot, además, de otros dos gráficos, donde se realizan adecuaciones del gráfico JK-Biplot, para visualizar en mejor forma los día/hora con mayor nivel de IMECA para el Ozono, y para observar mejor los día/hora con mayor nivel de IMECA para las Partículas Suspendidas. Podemos observar en dicha figura que concuerdan los día/hora con mayor nivel de IMECA para el Ozono, más no así con respecto a lo que sucede con las Partículas Suspendidas, ya que no concuerdan con los día/hora con mayor nivel de IMECA, en particular dos días: 06/18 y 06/19.

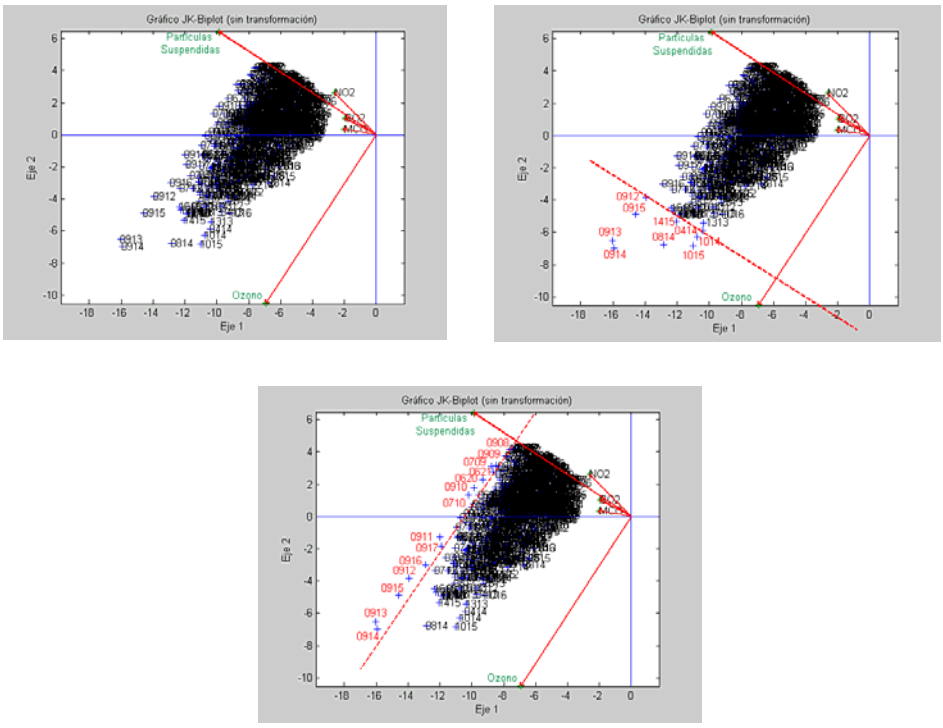

Figura 4.- Gráficos JK-Biplot

La Figura 5 presenta el gráfico HJ-Biplot, además, de otros dos gráficos, donde se realizan adecuaciones del gráfico HJ-Biplot, para visualizar en mejor forma los día/hora con mayor nivel de IMECA para el Ozono, y para observar mejor los día/hora con mayor nivel de IMECA para las Partículas Suspendidas. Se puede observar en dicha figura que concuerdan los día/hora con mayor nivel de IMECA para el Ozono, más no así con respecto a lo que sucede con las Partículas Suspendidas, ya que no concuerdan con los día/hora con mayor nivel de IMECA, e incluso, no aparece el día/hora con mayor nivel: 06/20.

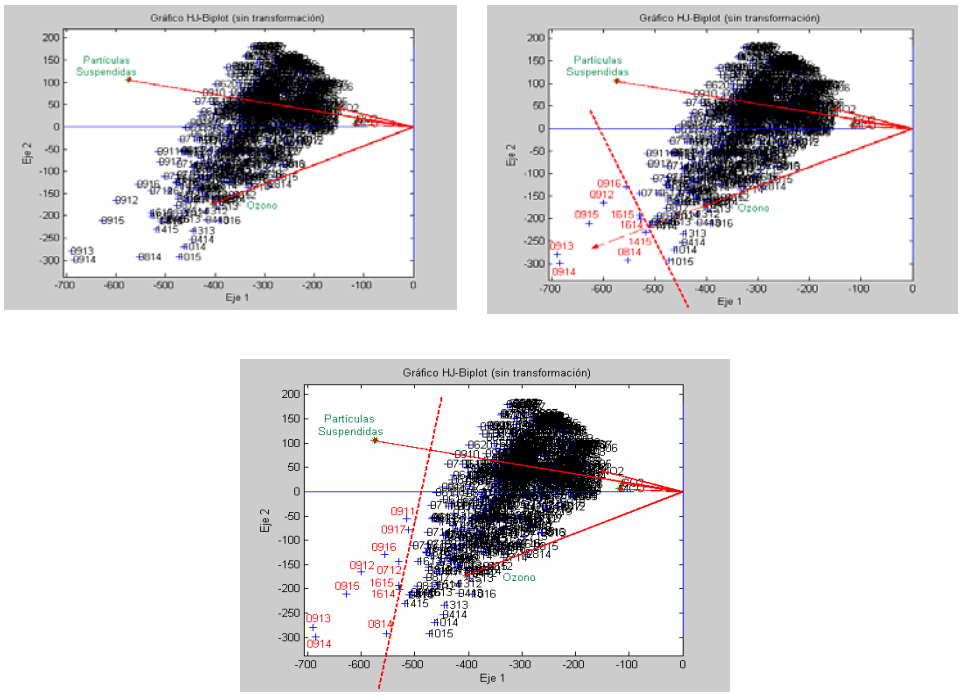

Figura 5.- Gráficos HJ-Biplot

4.3 Análisis exploratorio mediante el biplot robusto

Para iniciar se deberá determinar el número de componentes principales que se utilizarán para la obtención del Biplot Robusto. Así, la Figura 6, obtenida mediante BIPROB-2004 nos indica que se deberán de utilizar dos componentes principales.

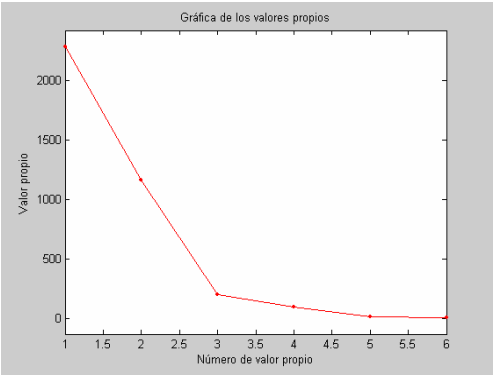

Figura 6.- Gráfica de los valores propios para determinar el número de componentes principales

Al aplicar el Biplot Robusto, tomando los dos primeros ejes, y sin hacer transformación a los datos, se obtuvo una calidad de representación global de 98.57 %, igual a la que se obtuvo con el Biplot clásico.

Todos los individuos (día/hora), se encontraban bien representados en el primer eje. Con respecto a las variables (contaminantes), el Ozono se encontraba bien representado en el eje 2, y todos los demás en el eje 1.

La Figura 7 presenta el gráfico del Biplot Robusto, además, de otros dos gráficos, donde se realiza un acercamiento del mismo, para visualizar en mejor forma los día/hora con mayor nivel de IMECA para el Ozono, y una adecuación al mismo gráfico, para observar mejor los día/hora con mayor nivel de IMECA para las Partículas Suspendidas.

Podemos observar en dicha figura que concuerdan exactamente los día/hora con mayor nivel de IMECA para el Ozono. Con respecto a las Partículas Suspendidas, concuerdan los día/hora con mayor nivel de IMECA, e incluso, ya aparece los día/hora con mayor nivel: 06/20, 06/18 y 06/19. Es decir existe una mejoría con respecto al Biplot Clásico.

5. CONCLUSIONES

En los gráficos GH-Biplot y JK-Biplot, no concuerdan los día/hora con mayor nivel de IMECA para las Partículas Suspendidas, en particular dos días: 06/18 y 06/19.En el gráfico HJ-Biplot, no concuerdan los día/hora con mayor nivel de IMECA para las Partículas Suspendidas, e incluso, no aparece el día/hora con mayor nivel: 06/20. En el gráfico Biplot Robusto, concuerdan exactamente los día/hora con mayor nivel de IMECA para el Ozono. Con respecto a las Partículas Suspendidas, concuerdan los día/hora con mayor nivel de IMECA, e incluso, ya aparece los día/hora con mayor nivel: 06/20, 06/18 y 06/19. Es decir existe una mejoría con respecto al Biplot clásico, de tal forma que se puede realizar un mejor análisis exploratorio de los datos.

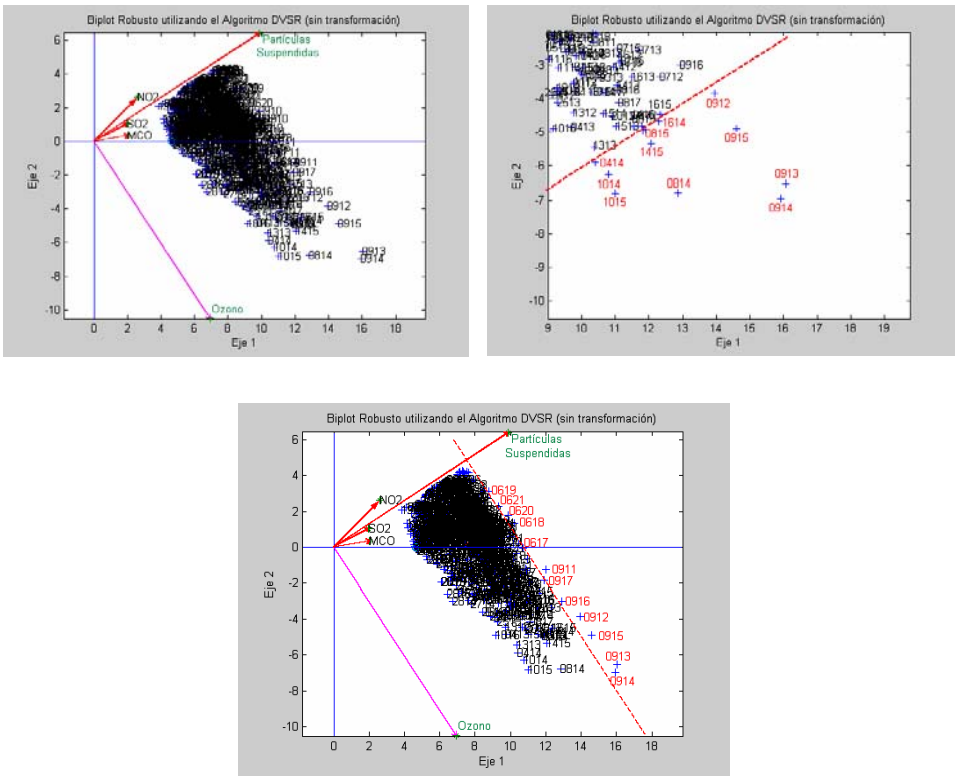

Figura 7.- Gráficos del Biplot Robusto (BIPROB)

REFERENCIAS

BRADU, D. y GABRIEL, K. R. (1974): Simultaneous statistical inference on interactions in two-way analysis of variance. **Journal of the American Statistical Association**, 29, 428-436.

ECKART, C. y YOUNG, G. (1936) : The approximation of one matrix by another of lower rank. **Psychometrika**, 1, 211-218.

GABRIEL, K. R. (1971) : The Biplot-Graphic display of matrices with application to principal component analysis. **Biometrika**, 58, 453-467.

GABRIEL, K. R. y ZAMIR, S. (1979) : Lower rank approximation of matrices by least squares with any choice of weights. **Technometrics**, 21, 489-498.

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

GALINDO, M. P. y CUADRAS, C. M. (1986) : Una extensión del método Biplot y su relación con otras técnicas. Publicaciones de Bioestadística y Biomatemática. Universidad de Barcelona, 17.

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

HERNÁNDEZ, S. (2005) : **Biplots Robustos**. Tesis Doctoral. Universidad de Salamanca. España.

MAHALANOBIS, P. C. (1939) : On the generalized distance in statistics. **Proc. Nat. Inst. Sci. India**, 2(1), 49-55.

MANDEL, J. (1961) : Non-additivity in two-way analysis of variance. **Journal of the American Statistical Associations**, 56, 878-888.

MARTENS, H.y NAES, T. (1989) : Multivariate Calibration. John Wiley & Sons, New York

PEÑA, D. y PRIETO, F. J. (2001) : Multivariate outlier detection and robust covariance matriz estimation. **Technometrics**, 43, 286-300.

RAMA (2003) : Dirección de Gestión Ambiental del Aire de la Secretaría del Medio Ambiente del Gobierno del Distrito Federal. Bases de datos del IMECA. http://148.243.232.103/imecaweb/base_datos-htm

ROUSSEEUW, P. J. y van DRIESSEN, K. (1999) : A fast algorithm for the minimum covariance determinant estimator. **Technometrics**, 41, 212-223.

TUKEY, J. W. (1949) : One degree of freedom for non-additivity. **Biometrics**, 5, 232-242.

WOLD, H. (1966a). Nonlinear estimation by iterative least squares procedures. **In: A Festschrift for F. Neyman, David, F. N. (Ed.)**, 411-444. New York: Wiley and Sons.

WOLD, H. (1966b) : Estimation of principal components and related models by iterative least squares. **Multivariate Analysis**, Academic Press, New York, 391-420.