# Biplot Versus Coodernadas Paralelas

Purificación Galindo Villardón[a], Purificación Vicente Galindo
*Universidad de Salamanca*
Carlomagno Araya Alpízar *Universidad de Costa Rica*

## 1. Métodos Biplot

**Un Biplot** es una representación gráfica de datos multivariantes. El objetivo del método Biplot es realizar una representación plana de la matriz $\mathbf{X}_{np}$ por medio de unos marcadores $g_1, \cdots, g_n$ para sus filas y $h_1, \cdots, h_p$ para sus columnas, elegidas de tal forma que el producto interno $g_i{}'h_j$ represente al elemento $x_{ij}$ de la matriz $\mathbf{X}$ (Gabriel, 1971).

Si los $g_i$ para $(i = 1, ..., n)$ son las filas de la matriz $\mathbf{G}$ y los $h_j$ para $(j = 1, ..., p)$ las filas de una matriz $\mathbf{H}$, el producto de estas matrices representa a la matriz de partida, de la forma:

$$\mathbf{X} = \mathbf{G}\mathbf{H}' \tag{1}$$

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \\ g_{31} & g_{32} \\ g_{41} & g_{42} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \end{bmatrix}$$

El elemento $x_{ij}$ de la matriz $\mathbf{X}$ se expresa como un producto de una fila de $G$ por una columna de $H$. Por ejemplo: $x_{11} = g_{11}h_{11} + g_{12}h_{21}$ y $x_{41} = g_{41}h_{11} + g_{42}h_{21}$. Cada elemento de la matriz de partida puede expresarse como un producto de una fila de $G$ por una columna de $H$. Se parte de una **descomposición en valores singulares** de la matriz $\mathrm{X_{nxp}}$ de rango **p**,

$$X = U\ \Sigma\ V^T \tag{2}$$

[a] pgalindo@usal.es

donde: **U** es una matriz de dimensión **(nxp)** cuyos vectores columna son ortonormales y vectores propios de $XX^T$ donde $XX^T$ tiene dimensión (nxn), V es una matriz ortogonal de dimensión (pxp) cuyos vectores columna son vectores propios de $X^TX$; donde $X^TX$ tiene dimensión (pxp) y $\Sigma$ es una matriz diagonal de dimensión (pxp) que contiene los valores singulares de X, ordenados de mayor a menor. Los valores singulares coinciden con los valores propios de $X^TX$ y $XX^T$.

Debido a que estamos aproximando una matriz de $\mathbf{X}_{np}$, de rango **r**, por una matriz de rango menor, $\mathbf{X}_q$, estamos "perdiendo información", ya que la representación Biplot es aproximada. Una forma de medir esta pérdida es a través de la **Calidad de Representación** de los puntos fila y columna, cuanto más cercano este a cien, mayor cantidad de información está siendo recogida por la representación Biplot.

Entre los métodos Biplot, nos encontramos el GH-Biplot, JK-Biplot y HJ-Biplot. El **JK-Biplot**, es una representación simultánea de individuos y variables, donde los individuos tienen máxima calidad de representación, razón por la cual se conoce RMP-Biplot (Row Metric Preserving). A este Biplot Gabriel lo denominó JK-Biplot porque utilizó **J** para denotar la matriz de marcadores fila y **K** para la matriz de marcadores columna. El **GH-Biplot**, es una representación simultánea de individuos y variables, donde las variables tienen máxima calidad de representación. A este Biplot Gabriel lo denominó GH-Biplot porque utilizó **G** para denotar la matriz de marcadores fila y **H** para la matriz de marcadores columna. El producto escalar de las columnas de **X**, coincide con el producto escalar de los marcadores columna, de ahí el hecho de que este Biplot se denomine CMP-Biplot (Column Metric Preserving) ya que preserva la métrica euclídea usual entre las columnas de **X** obteniéndose una alta calidad de representación para éstas.

El **HJ-Biplot** a diferencia de los anteriores fue propuesto por Galindo, (1985, 1986) es una representación gráfica multivariante de marcadores fila y columna, elegidos de tal forma que puedan superponerse en el mismo sistema de referencia con máxima calidad de representación. Este Biplot es muy útil en la interpretación simultánea de relaciones entre filas y columnas, no siendo su objetivo principal la aproximación de los elementos de la matriz de datos, como es el caso de los Biplot definidos por Gabriel. Los elementos de la matriz **X** están centrados por filas y columnas, por lo que la métrica introducida en el espacio de las filas es equivalente a la inversa de la matriz de covarianzas entre variables, mientras

que en el espacio de las columnas la métrica es equivalente a la inversa de la matriz de dispersión entre individuos. Dado de que en el **HJ-Biplot** se puede hacer una representación simultánea de filas y columnas se lo denomina también **RCMP-Biplot (Row Column Metric Preserving)**.

El **HJ- Biplot** permite interpretar las posiciones de las filas, de las columnas y las relaciones fila-columna a través de los factores (ejes), como en el caso del Análisis Factorial de Correspondencias (Benzecri, 1973; Greenacre, 1984) teniendo además la ventaja de que un análisis Biplot puede llevarse a cabo sobre a cualquier tipo de datos.

# 2. Coordenadas Paralelas

Las **Coordenadas Paralelas** (*Coords*‖) fueron propuestas por Alfred Inselberg (1992). Las *Coords*‖ es un sistema de visualización que permite representar n-dimensiones en un sistema bidimensional. En este sistema, cada eje vertical (ordenada) representa un atributo (dimensión) que puede ser continuo o categórico. Cada uno de los ejes verticales de un sistema de *Coords*‖ puede tener su propia escala o definirse todos con una sola escala, la primera forma nos permite la visualización de híper-superficies y el análisis del funcionamiento del conjunto de datos, con la segunda podemos hacer un análisis de las relaciones entre las variables.

Uniendo con líneas los ejes, podemos simbolizar los puntos en $n$-dimensiones. Asimismo, un punto en un espacio $n$-dimensional es transformado en una línea poligonal a través de $n$ ejes paralelos como $n-1$ segmentos de línea. De tal forma, el vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ es representado por medio de $x_1$ en la coordenada 1, $x_2$ en la coordenada 2 y así sucesivamente, hasta la $x_n$ en al coordenada $n$. A partir de la representación resultante, podemos sacar conclusiones al respecto, por ejemplo sobre la relación entre las variables.

El orden de las *Coords*‖ es una condición que puede afectar significativamente la expresividad del gráfico, variando el orden es posible abreviar el problema sin la reducción del contenido o de la modificación de los datos de alguna manera. También las correlaciones entre las variables (o dimensiones) pueden ser descubiertas concentrándose en las intersecciones de las polilíneas, al detectar grupos de observaciones con pendientes comunes en las líneas de conexión inter-variables, poniendo de relieve un determinado tipo de correlación

entre dichas variables (positiva, negativa o nula).

Las $Coords\|$ resultan útiles para captar agrupamientos ("clústeres"). Las polilíneas que tienden a estar cercanas constituirán un grupo a diferencia de aquellas que se separan y cuando hay líneas que no pertenecen a ningún grupo (fuera de los patrones) pueden considerarse como valores extremos. El descubrimiento de grupos o racimos de polilíneas diferenciadas del resto se consigue cambiando los órdenes de las dimensiones, para procurar que las relaciones de los datos puedan ser visualizadas. Es recomendable estandarizar las variables para poder compáralas y permitir un mejor descubrimiento del patrón anormal.

# 3. Biplot Versus Coodernadas Paralelas

Se pretende con este apartado, estudiar las semejanzas y diferencias entre ambos métodos de análisis de datos multivariantes. De algún modo, se intenta establecer que ambas técnicas multivariantes integradas maximizan el éxito en la interpretación de los resultados.

En los métodos Biplot la variabilidad de las variables está determinada por longitud de los vectores, mientras que en $Coords\|$ se visualiza con la dispersión de las polilíneas en los ejes.

Se observa en la Figura 1 que las variables **V4** y **V5** tienen una alta correlación positiva, tal que el ángulo entre los dos vectores es muy pequeño (en los Biplots), esto en $Coords\|$ se visualiza como líneas entre los ejes asociados a **V4** y **V5** que tienden a ser horizontales. Una correlación negativa en $Coords\|$ es por ejemplo, entre las variables **V1** y **V2**, se observa como líneas que se intersecan.

En la Figura 2 puede verse el perfil del grupo formado por los individuos **A**, **F** y **K**. Se observa que poseen valores altos en la variable **V2**, y por lo contrario valores muy pequeños en comparación a los demás individuos en las restantes variables. Las $Coords\|$ nos proporcionan un instrumento para hacer un diagnóstico de las contribuciones de las individuos y variables a los ejes factoriales de la representación Biplot.

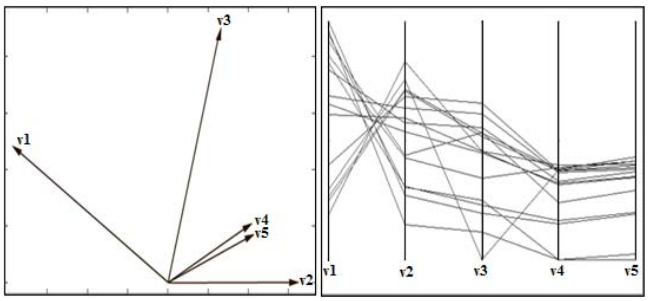


**Figura 1:** Diagnóstico de la correlación en los Biplots y *Coords*||

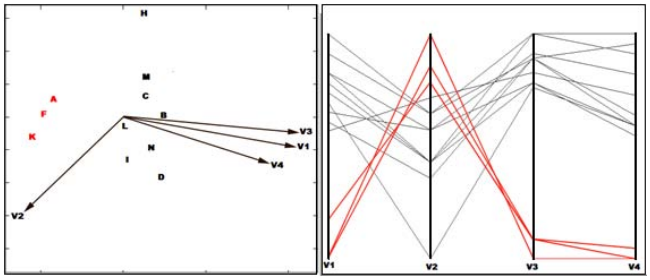


**Figura 2:** Diagnóstico de grupos en los Biplots y *Coords*||

# Referencias

[1] Benzécri, Jean Paul. 2004. *L'Analyse des Données: L'analyse des correspondances*. SParis: Dunod.

[2] Gabriel, Karl Ruben. 1971. *The Biplot Graphic Display of Matrices with Application to Principal Component Analysis*. *Biometrika*, **58**, 453-467.

[3] Galindo, María Purificación. 1985. *Contribuciones a la Representación Simultánea de datos Multidimensionales*.Tesis doctoral. Universidad de Salamanca.

[4] Galindo, María Purificación. 1986. *Una Alternativa de Representación Simultánea: HJ-biplot*. *Questió*, **10**, 13-23.

[5] Greenacre, Michael John. 1984. *Theory and Applications of Correspondence Analysis*. London: Academic Press.

[6] Inselberg, Alfred. 1992. *The Plane R2 with Coordinate Parallel*. Tel Aviv: Computer Science and Applied Mathematics Departments.