Análisis Biplot

I. Artículos seminales de Gabriel K.R. & Odoroff C.F.
II. Descomposición por valor singular

I. Artículos seminales de Gabriel K.R. & Odoroff C.F.

Gabriel K.R. **The biplot graphic display of matrices with application to principal component analysis**. Biometrika (1971), 58, 3, pp. 453-467

Gabriel K.R. & Odoroff, C.L. (1986) **The biplot for Exploration and Diagnosis: examples and software**, Dept. of Stat. Rep. #86/03, University of Rochester, Rochester, New York.

Any $n \times m$ matrix $\mathbf{Y}$ of rank r can be factorized as

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

into a $n \times r$ matrix $\mathbf{G}$ and a $m \times r$ matrix $\mathbf{H}$, both necessarily of rank r (Rao, 1965*a*, 1*b*.2.3).

(...)

Factorization (1) may be written as

$$y_{ij} = \mathbf{g}_i'\mathbf{h}_j$$

for each i and j, where y_{ij} is the element in the ith row and jth column of $\mathbf{Y}$, $\mathbf{g}_i'$ is the ith row of $\mathbf{G}$ and $\mathbf{h}_j$ is the jth row of $\mathbf{H}$. In this form, the factorization assigns the vectors $\mathbf{g}_1, \ldots, \mathbf{g}_n$, one to each of the n rows of $\mathbf{Y}$ and the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_m$, one to each column of $\mathbf{Y}$.

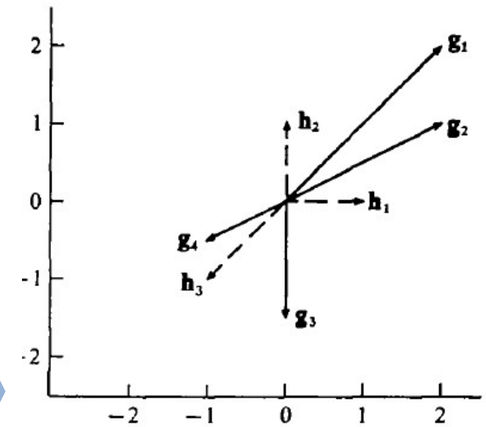
(...)

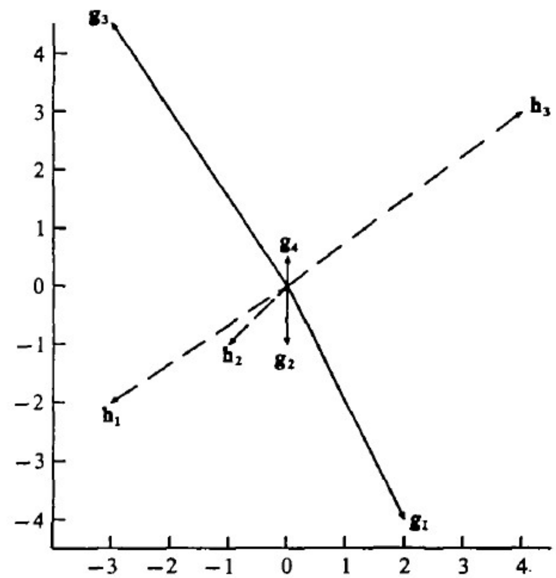
In a matrix of rank two, the effects $\mathbf{g}_1, \ldots, \mathbf{g}_n$ and $\mathbf{h}_1, \ldots, \mathbf{h}_m$ are vectors of order two. These $n+m$ vectors may be plotted in the plane, giving a representation of the nm elements of $\mathbf{Y}$ by means of the inner products of the corresponding row effect and column effect vectors. Such a plot will be referred to as a *biplot* since it allows row effects and column effects to be plotted jointly. In the rest of this section only matrices $\mathbf{Y}$ of rank $r = 2$ will be considered.

$$\begin{bmatrix} 2 & 2 & -4 \\ 2 & 1 & -3 \\ 0 & -\frac{3}{2} & \frac{3}{2} \\ -1 & -\frac{1}{2} & \frac{3}{2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 1 \\ 0 & -\frac{3}{2} \\ -1 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -4 \\ 0 & -1 \\ -3 & 4\frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} -3 & -1 & 4 \\ -2 & -1 & 3 \end{bmatrix}.$$

No unicidad de factorización
(depende de la factorización utilizada)





$$\mathbf{Y} = (\mathbf{GR}')(\mathbf{HR}^{-1})'$$

… son rotaciones

for any nonsingular $\mathbf{R}$. To examine this nonuniqueness, consider the singular value decomposition of $\mathbf{R}'$,

$$\mathbf{R}' = \mathbf{V}'\boldsymbol{\Theta}\mathbf{W},$$

where $\mathbf{V}$ and $\mathbf{W}$ are 2×2 orthonormal matrices and $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \theta_2)$ and the transposed inverse is

$$\mathbf{R}^{-1} = \mathbf{V}'\boldsymbol{\Theta}^{-1}\mathbf{W}$$

In choosing, as in (1), factors **G** and **H** of $\mathbf{Y}_{(2)}$ for biplotting, one may use the factorization provided by the singular decomposition (38). Writing

$$\mathbf{p}'_\alpha = (p_{\alpha 1}, \ldots, p_{\alpha n}), \quad \mathbf{q}'_\alpha = (q_{\alpha 1}, \ldots, q_{\alpha m}),$$

one obtains

$$\mathbf{Y}_{(2)} = \begin{bmatrix} p_{11} & p_{21} \\ \vdots & \vdots \\ p_{1n} & p_{2n} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} q_{11} \cdots q_{1m} \\ q_{21} \cdots q_{2m} \end{bmatrix}.$$

One choice of **G** and **H** would be in terms of rows

$$\mathbf{g}'_i = (\sqrt{\lambda_1} p_{1i}, \sqrt{\lambda_2} p_{2i}) \quad (i = 1, \ldots, n),$$
$$\mathbf{h}'_j = (\sqrt{\lambda_1} q_{1j}, \sqrt{\lambda_2} q_{2j}) \quad (j = 1, \ldots, m).$$

Other choices of **G** and **H** are obtained by defining

$$\mathbf{g}'_i = (p_{1i}, p_{2i}) \quad (i = 1, \ldots, n),$$
$$\mathbf{h}'_j = (\lambda_1 q_{1j}, \lambda_2 q_{2j}) \quad (j = 1, \ldots, m),$$

$$\mathbf{G}'\mathbf{G} = \mathbf{I}_2$$

or as

$$\mathbf{g}'_i = (\lambda_1 p_{1i}, \lambda_2 p_{2i}) \quad (i = 1, \ldots, n),$$
$$\mathbf{h}'_j = (q_{1j}, q_{2j}) \quad (j = 1, \ldots, m),$$

$$\mathbf{H}'\mathbf{H} = \mathbf{I}_2,$$

$$S = \frac{1}{n} Y'Y$$

$$d^2_{i,c} = (y_i - y_c)' S^{-1} (y_i - y_c)$$

Principal component analysis consists of singular decomposition of such a matrix **Y** (Whittle, 1952). Note that (28) becomes

$$nSq_\alpha = \lambda_\alpha^2 q_\alpha, \qquad (46)$$

the usual form of the equations for principal component analysis, except that the factor n is often omitted and λ_α^2/n computed instead of λ_α^2.

The singular decomposition (24) shows that matrix **Y** can be factorized as

$$Y = (p_1, \ldots, p_r)(\lambda_1 q_1, \ldots, \lambda_r q_r)'.$$

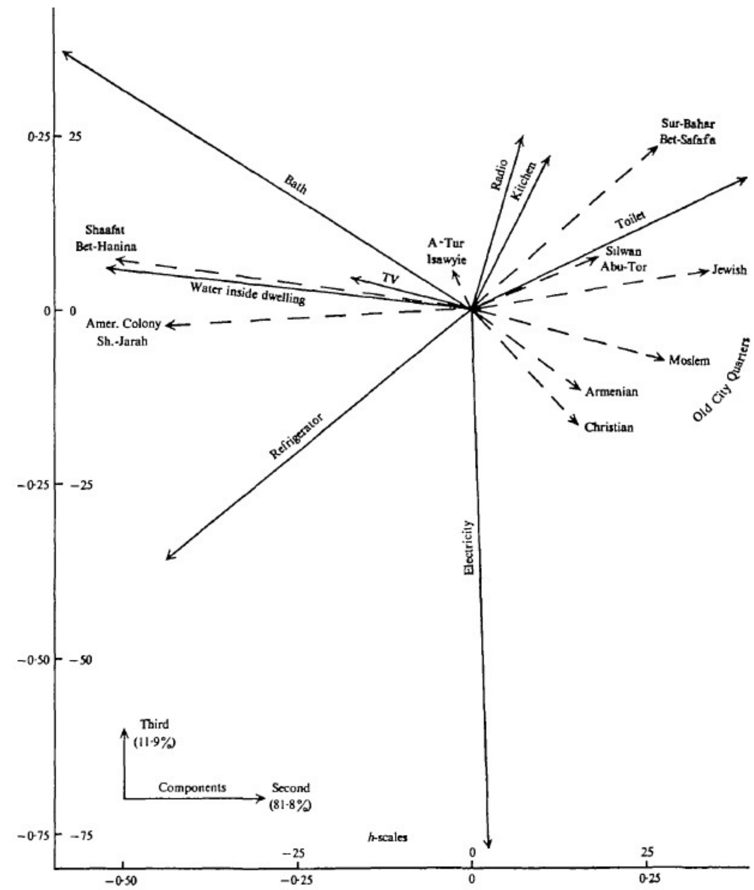
Factorizando…

$$(p_1, \ldots, p_r)'(p_1, \ldots, p_r) = I_r,$$

$$(p_1, \ldots, p_r)(p_1, \ldots, p_r)' = \frac{1}{n} Y S^{-1} Y',$$

$$(\lambda_1 q_1, \ldots, \lambda_r q_r)'(\lambda_1 q_1, \ldots, \lambda_r q_r) = \operatorname{diag}(\lambda_1, \ldots, \lambda_r),$$

$$(\lambda_1 q_1, \ldots, \lambda_r q_r)(\lambda_1 q_1, \ldots, \lambda_r q_r)' = nS.$$



$$G = (p_1, p_2)\sqrt{n},$$

$$H = \frac{1}{\sqrt{n}}(\lambda_1 q_1, \lambda_2 q_2),$$

$$y_{ij} \sim g_i' h_j,$$

Aproximación a una observación individual y_{ij}

II. Descomposición por Valor Singular DVS

2.1 Es una forma alternativa para obtener las raíces y vectores característicos

2.2 Usar la técnica DVS para analizar datos en dos dimensiones

DESCOMPOSICIÓN POR VALOR SINGULAR (DVS)

Los datos originales se pueden obtener de las CP:

$$\mathbf{X} = \mathbf{Y}\mathbf{V}' \qquad \text{(i)}$$

O también en términos de las transformaciones:

Descomposición por valor singular (DVS)

$$\boxed{\mathbf{X} = \mathbf{U}^*\mathbf{L}^{1/2}\mathbf{U}'} \qquad \text{(ii)}$$

En DVS, una matriz de datos **X** se descompone como el producto de vectores característicos de **X’X**, los vectores característicos de **XX’** y una función de sus raíces características **L**.

$$\mathbf{X} = \mathbf{Y}\mathbf{L}^{1/2}\mathbf{U}' = \mathbf{U}^*\mathbf{L}^{1/2}\mathbf{Y}^*$$

$$\begin{array}{cc} & \text{variables} \\ \mathbf{X} = & \begin{bmatrix} -2 & -2 & -1 \\ 0 & 1 & 0 \\ 2 & 2 & 2 \\ 0 & -1 & -1 \end{bmatrix} \text{ observaciones} \\ \text{media} \longrightarrow & \begin{matrix} 0 & 0 & 0 \end{matrix} \end{array}$$

Conservando columnas o variables

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 8 & 8 & 6 \\ 8 & 10 & 7 \\ 6 & 7 & 6 \end{bmatrix}$$

Tr(X'X)=24

Conservando filas u observaciones

$$\mathbf{X}\mathbf{X}' = \begin{bmatrix} 9 & -2 & -10 & 3 \\ -2 & 1 & 2 & -1 \\ -10 & 2 & 12 & -4 \\ 3 & -1 & -4 & 2 \end{bmatrix}$$

Tr(XX')=24

Raíces características de **X'X**

$$l_1 = 22.2819$$
$$l_2 = 1.0000$$
$$l_3 = .7181$$

X´X>0, r(X'X)=3

Raíces características de **XX'**

$$l_1 = 22.2819$$
$$l_2 = 1.0000$$
$$l_3 = .7181$$
$$l_4 = 0$$

XX´≥0, r(XX')=3

Vectores características de **X'X**

$$\mathbf{U} = \begin{bmatrix} -.574 & .816 & -.066 \\ -.654 & -.408 & .636 \\ -.493 & -.408 & -.768 \end{bmatrix}$$

Vectores características de **XX'**

$$\mathbf{U}^* = \begin{bmatrix} .625 & -.408 & -.439 \\ -.139 & -.408 & .751 \\ -.729 & 0 & -.467 \\ .243 & .816 & .156 \end{bmatrix}$$

$$\mathbf{L} = \begin{bmatrix} 22.2819 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & .7181 \end{bmatrix} \qquad \mathbf{L}^{1/2} = \begin{bmatrix} 4.7204 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & .8474 \end{bmatrix}$$

Vectores característicos escalados por $\mathbf{L}^{1/2}$=diag($\sqrt{\lambda}$)

$$\mathbf{V} = \mathbf{U}\mathbf{L}^{1/2} = \begin{bmatrix} -2.707 & .816 & .056 \\ -3.089 & -.408 & -.539 \\ -2.326 & -.408 & .651 \end{bmatrix} \qquad \mathbf{V}^* = \mathbf{U}^*\mathbf{L}^{1/2} = \begin{bmatrix} 2.9549 & -.408 & -.372 \\ -.656 & -.408 & .636 \\ -3.441 & 0 & -.396 \\ 1.147 & .816 & .132 \end{bmatrix}$$

Vectores característicos escalados por $\mathbf{L}^{-1/2}$=diag($1/\sqrt{\lambda}$)

$$\mathbf{W} = \mathbf{U}\mathbf{L}^{-1/2} = \begin{bmatrix} -.122 & .816 & -.078 \\ -.139 & -.408 & .751 \\ -.104 & -.408 & -.906 \end{bmatrix} \qquad \mathbf{W}^* = \mathbf{U}^*\mathbf{L}^{-1/2} = \begin{bmatrix} .132 & -.408 & -.518 \\ -.029 & -.408 & .886 \\ -.154 & 0 & -.552 \\ .051 & .816 & .184 \end{bmatrix}$$

Scores de las observaciones (CP):

$$\mathbf{Y} = \mathbf{XW} = \mathbf{U}^*$$

(nxp)

*"Los scores de las **n** observaciones (CP) son iguales a los vectores característicos obtenidos de las matriz suma de cuadrados de observaciones"*

Scores de las variables (CP):

$$\mathbf{Y}^* = \mathbf{X}'\mathbf{W}^* = \mathbf{U}$$

(pxp)

*"Los scores de las **p** variables (CP) son iguales a los vectores característicos obtenidos de la matriz suma de cuadrados de variables"*

De (ii):

$$\mathbf{U}^*\mathbf{L}^{1/2}\mathbf{U}' = \begin{bmatrix} .625 & -.408 & -.439 \\ -.139 & -.408 & .751 \\ -.729 & 0 & -.467 \\ .243 & .816 & .156 \end{bmatrix} \begin{bmatrix} 4.72 & 0 & 0 \\ 0 & 1.00 & 0 \\ 0 & 0 & .85 \end{bmatrix} \begin{bmatrix} -.574 & -.654 & -.493 \\ .816 & -.408 & -.408 \\ -.066 & .636 & -.768 \end{bmatrix}$$

$$= \begin{bmatrix} -2 & -2 & -1 \\ 0 & 1 & 0 \\ 2 & 2 & 2 \\ 0 & -1 & -1 \end{bmatrix} = \mathbf{X}$$

Gracias