Capítulo 4

Representación Biplot

4.1. Introducción.

La representación Biplot es una representación gráfica de datos multivariantes en dos o tres dimensiones. Las representaciones de las variables son normalmente vectores y los individuos se representan por puntos. El prefijo 'bi' se refiere a la superposición en la misma representación de individuos y variables.

Definición 4.1 Representación Biplot. *Si partimos de la DVS de una matriz A en un espacio euclídeo ponderado: $A_{I\times J} = N_{I\times K}(\mathbb{D}_\alpha)_{K\times K}M'_{K\times J}$, las filas de $F_{[K^*]} = V_{[K^*]}\mathbb{D}_{\alpha[K^*]}$ se representan en el subespacio S, normalmente de 2 o 3 dimensiones, para observar la configuración multidimensional de las filas de la matriz de datos A, y las filas de $G_{[K^*]} = U_{[K^*]}\mathbb{D}_{\alpha[K^*]}$ se representan en el subespacio S', para observar la configuración multidimensional de las columnas de A. Sin embargo es posible representar las filas de $G_{[K^*]}$ como $G_{[K^*]} = U_{[K^*]}$ en el mismo espacio que el de las filas de $F_{[K^*]}$, esta particular representación común de puntos fila y columna de la matriz A, se llama representación biplot (jk-biplot como veremos más adelante).*

En una representación biplot $A_{[K^*]} = F_{[K^*]}G'_{[K^*]}$ luego el producto escalar de la i-ésima fila de F y la j-ésima fila de G, es un valor aproximado del elemento a_{ij} de la matriz:

$$a_{ij} \approx f_i'g_j = |f_i| \cdot |g_j| \cos(f_i, g_j) \tag{4.1}$$

Los datos a_{ij} de la matriz A se centran usualmente, respecto a la media de las columnas, la matriz A, se define como la matriz de filas centradas de Y: $A = Y - 1\bar{y}'$:

$$\sum_{i=1}^{I} w_i(a_i - x_i)'\mathbb{D}_q(a_i - x_i) = \left[\sum_i \omega_i \left(\underbrace{y_i - \bar{y}}_{a_i} - \sum_{k=1}^{K^*} f_{ik}u_k\right)'\right] \mathbb{D}_q \left(y_i - \bar{y} - \sum_{k=1}^{K^*} f_{ik}u_k\right)$$

donde $\bar{y}$ es el centroide de la nube de puntos fila de A.

Por lo tanto una desviación $a_{ij} > 0$ (valores de la matriz por encima del valor medio (cero), $a_{ij} = y_{ij} - \bar{y}_j > 0 \Longrightarrow y_{ij} > \bar{y}_j$), se indica por vectores f_i y g_j formando un ángulo agudo y $a_{ij} < 0$ por vectores bajo un ángulo obtuso.

4.2. Tipos de Biplot

Si tenemos una matriz $X_{n\times p}$ la descomposición en valores singulares de esa matriz es

$$X = V\mathbb{D}_\alpha U' = V\mathbb{D}_\lambda^{1/2}U' = V\mathbb{D}_\lambda^{c/2}\mathbb{D}_\lambda^{(1-c)/2}U' = FG' \qquad \text{siendo } \ 0 \le c \le 1$$

por lo tanto $F = V\mathbb{D}_\lambda^{c/2}$ y $G = U\mathbb{D}_\lambda^{(1-c)/2}$.

Gabriel y Odoroff en 1986, propusieron 3 elecciones del parámetro c

- SQ-biplot ($c = 0{,}5$), en este caso

$$F = V\mathbb{D}_\lambda^{1/4} \qquad G = U\mathbb{D}_\lambda^{1/4}$$

En estas representaciones punto-vector, donde las filas se representan por vectores y las columnas por puntos o viceversa, cada uno de los objetos representados por los vectores resumen el espacio, proyectando los objetos representados por puntos sobre una única dimensión orientada hacia la región dominante del espacio. Los vectores representan direcciones y las proyecciones de los puntos representan distancias. El ángulo entre un vector y cada uno de los ejes principales (definidos por la DVS), mide la importancia de la contribución que cada dimensión realiza sobre los datos de los que se deriva el vector.
Dentro de cada conjunto de puntos F y G, las distancias entre dos puntos están relacionadas de forma similar a la relación que tengan en la matriz inicial. Puesto que son representaciones punto-vector, nunca se pueden interpretar distancias entre puntos de F y G.
La intersección de las proyecciones de los puntos de F, perpendiculares al vector que representa la primera, segunda o tercera fila de G, g_1, g_2 y g_3 corresponden a la primera, segunda o tercera columna de X.
Si consideramos la matriz X, las coordenadas de las filas y columnas F y G son en este caso

$$X = \begin{pmatrix} -2 & -2 & -1 \\ 0 & 1 & 0 \\ 2 & 2 & 2 \\ 0 & -1 & -1 \end{pmatrix} \qquad F = \begin{pmatrix} 1{,}36 & -0{,}41 \\ -0{,}3 & -0{,}41 \\ -1{,}58 & 0 \\ 0{,}53 & 0{,}82 \end{pmatrix} \qquad G = \begin{pmatrix} -1{,}25 & 0{,}82 \\ -1{,}42 & -0{,}41 \\ -1{,}07 & -0{,}41 \end{pmatrix}$$

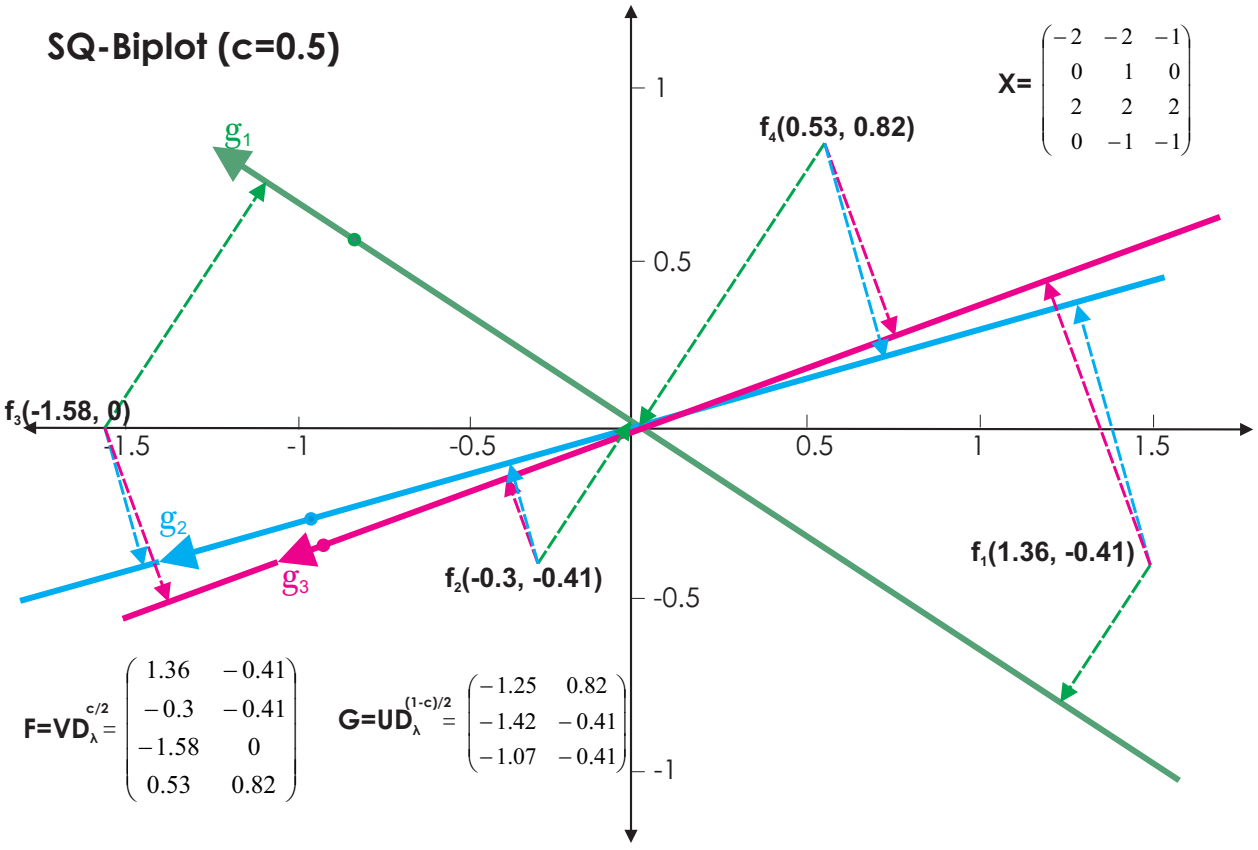

Figura 4.1: SQ-biplot-1

Se puede observar que la proyecciones de G sobre f_3 son aproximadamente iguales correspondiendo a los valores de la 3ª fila. La correspondencia debería ser exacta, pero hay una pequeña cantidad de

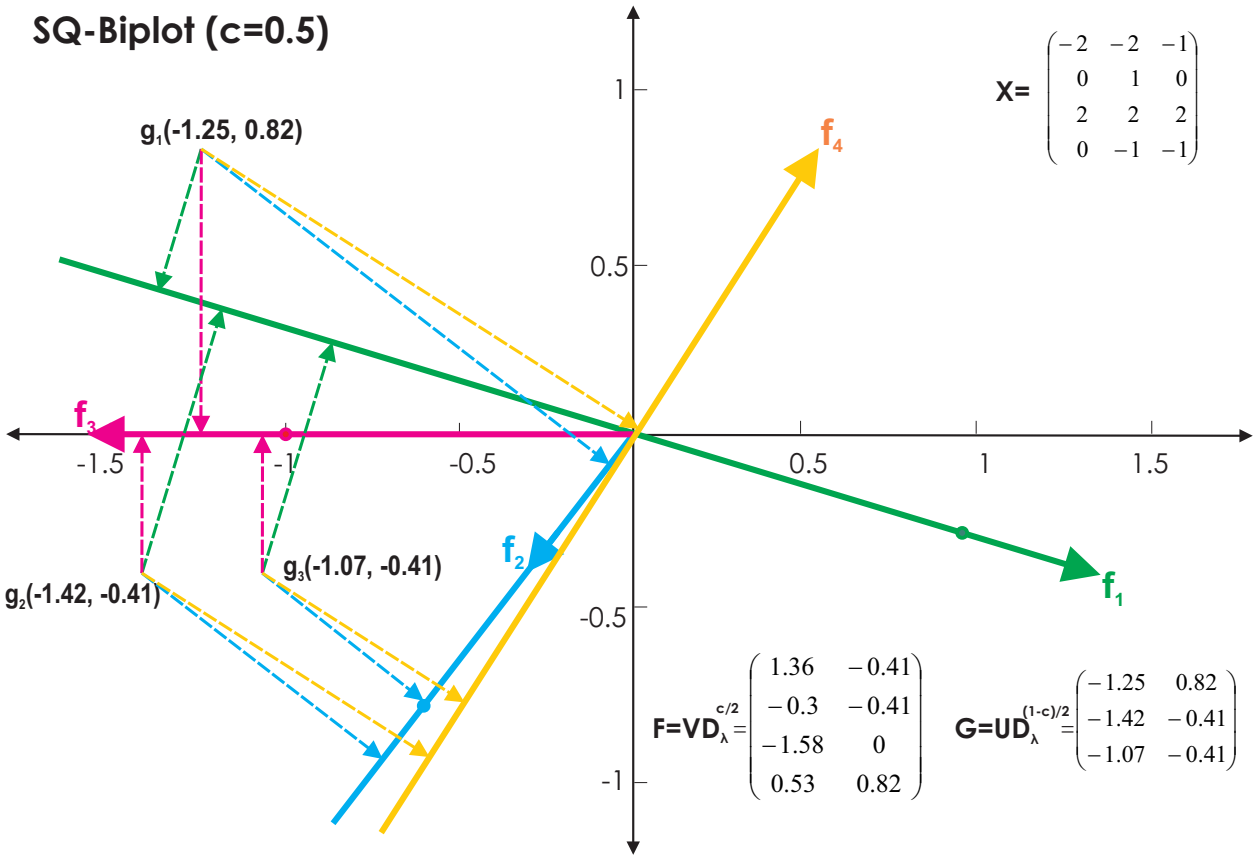

Figura 4.2: SQ-biplot-2

variabilidad explicada por la tercera dimensión que no está representada en el biplot. Los puntos g_2 y g_3 están próximos, lo que se traduce en que la 2ª y 3ª columna de X son casi iguales. El hecho de que f_1 y f_3 estén muy separados, se traduce en que la 1ª y 3ª fila de X son muy diferentes.

Los cálculos intermedios los obtenemos en R:

```
> X <- matrix(c(-2, 0, 2, 0, -2,1,2,-1,-1,0,2,-1), nrow=4)
> X
     [,1] [,2] [,3]
[1,]   -2   -2   -1
[2,]    0    1    0
[3,]    2    2    2
[4,]    0   -1   -1
> meanX <- apply(X,2,mean)
> meanX
[1] 0 0 0
> sdX <- apply(X,2,sd)
> sdX
[1] 1.632993 1.825742 1.414214
> t(X)%*%X
     [,1] [,2] [,3]
[1,]    8    8    6
[2,]    8   10    7
[3,]    6    7    6
```

```
> X%*%t(X)
     [,1] [,2] [,3] [,4]
[1,]    9   -2  -10    3
[2,]   -2    1    2   -1
[3,]  -10    2   12   -4
[4,]    3   -1   -4    2
> svd(t(X)%*%X)
$d
[1] 22.2819293  1.0000000  0.7180707

$u
           [,1]       [,2]        [,3]
[1,] -0.5735637  0.8164966  0.06601555
[2,] -0.6544159 -0.4082483 -0.63645361
[3,] -0.4927115 -0.4082483  0.76848471

$v
           [,1]       [,2]        [,3]
[1,] -0.5735637  0.8164966  0.06601555
[2,] -0.6544159 -0.4082483 -0.63645361
[3,] -0.4927115 -0.4082483  0.76848471

> svd(X%*%t(X))
$d
[1] 2.228193e+01 1.000000e+00 7.180707e-01 5.416122e-16

$u
           [,1]          [,2]       [,3] [,4]
[1,] -0.6246689  4.082483e-01  0.4394567  0.5
[2,]  0.1386365  4.082483e-01 -0.7510747  0.5
[3,]  0.7290486 -6.938894e-16  0.4674271  0.5
[4,] -0.2430162 -8.164966e-01 -0.1558090  0.5

$v
           [,1]          [,2]       [,3] [,4]
[1,] -0.6246689  4.082483e-01  0.4394567 -0.5
[2,]  0.1386365  4.082483e-01 -0.7510747 -0.5
[3,]  0.7290486 -1.110223e-16  0.4674271 -0.5
[4,] -0.2430162 -8.164966e-01 -0.1558090 -0.5

> lambda <- diag(sqrt(svd(t(X)%*%X)$d))
> lambda2 <- lambda[-3,1:2]
> lambda2
         [,1] [,2]
[1,] 4.720374    0
[2,] 0.000000    1
> lambda2^(1/2)
         [,1] [,2]
[1,] 2.172642    0
[2,] 0.000000    1
> V <- svd(X%*%t(X))$v[,1:2]
> V
           [,1]          [,2]
[1,] -0.6246689  4.082483e-01
```

```
[2,]  0.1386365  4.082483e-01
[3,]  0.7290486 -1.110223e-16
[4,] -0.2430162 -8.164966e-01
> F <- V%*%lambda2^(1/2)
> round(F,2)
      [,1]  [,2]
[1,] -1.36  0.41
[2,]  0.30  0.41
[3,]  1.58  0.00
[4,] -0.53 -0.82
> U <- svd(t(X)%*%X)$u[,1:2]
> G <- U%*%lambda2^(1/2)
> round(G,2)
      [,1]  [,2]
[1,] -1.25  0.82
[2,] -1.42 -0.41
[3,] -1.07 -0.41
```

- GH-biplot ($c = 0$), en este caso

$$F = V \qquad G = U\mathbb{D}_{\lambda}^{1/2}$$

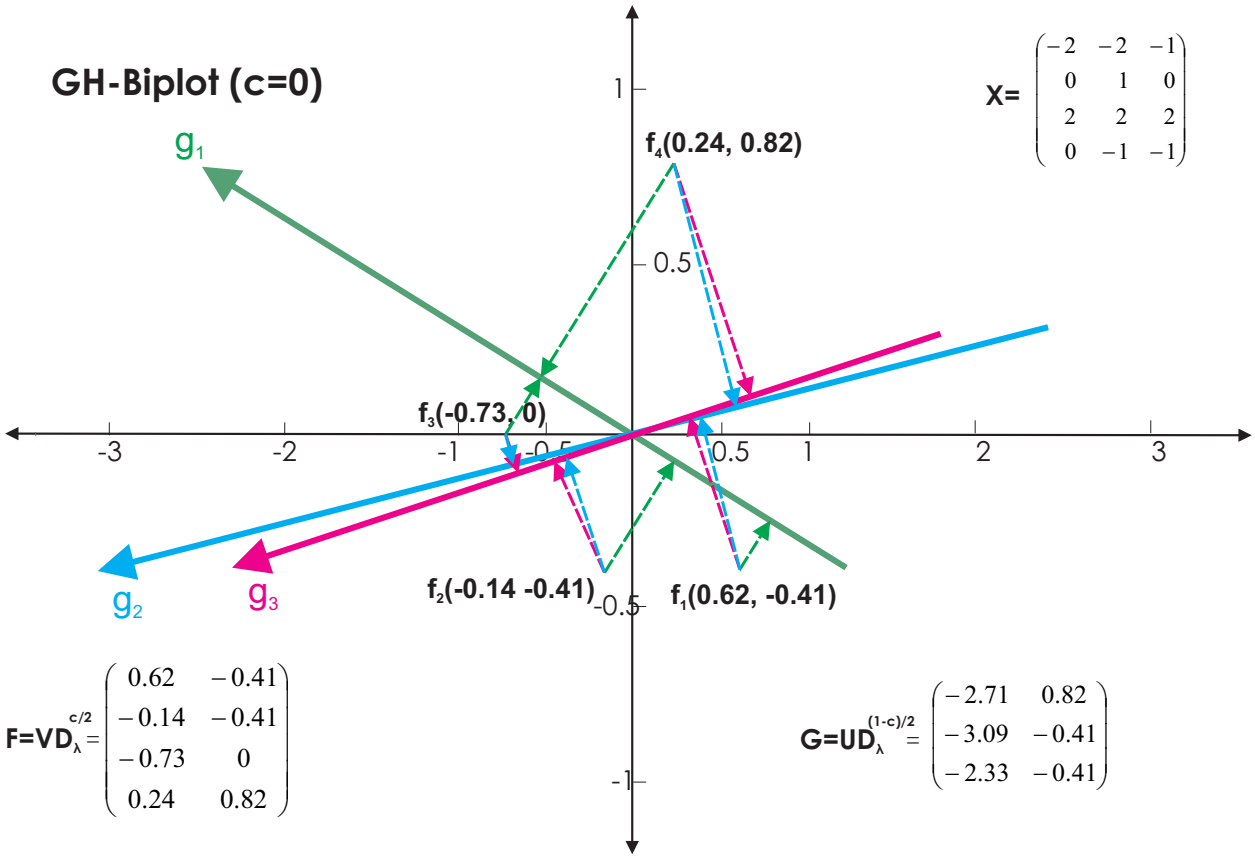

Figura 4.3: GH-biplot-1

Este Biplot preserva la métrica de las columnas. Si representamos F con vectores y G con puntos, las distancias entre puntos de G son euclídeas, ya que están en las mismas unidades que los datos originales. Si representamos G con vectores y F por puntos, las distancias entre puntos de F son distancias

de Mahalanobis (ponderadas). El coseno del ángulo entre 2 vectores de G, g_j y $g_{j'}$, aproximará la correlación entre ambas columnas j y j'. La longitud del vector g_j se relaciona aproximadamente en términos de la desviación típica de la j-ésima columna de X.

Las coordenadas de las filas y columnas F y G son

$$F = \begin{pmatrix} 0{,}62 & -0{,}41 \\ -0{,}14 & -0{,}41 \\ -0{,}73 & 0 \\ 0{,}24 & 0{,}82 \end{pmatrix} \qquad G = \begin{pmatrix} -2{,}71 & 0{,}82 \\ -3{,}09 & -0{,}41 \\ -2{,}33 & -0{,}41 \end{pmatrix}$$

```
> F <- V
> round(F,2)
      [,1]  [,2]
[1,] -0.62  0.41
[2,]  0.14  0.41
[3,]  0.73  0.00
[4,] -0.24 -0.82
> G <- U%*%lambda2
> round(G,2)
      [,1]  [,2]
[1,] -2.71  0.82
[2,] -3.09 -0.41
[3,] -2.33 -0.41
```

- JK-biplot ($c = 1$), en este caso

$$F = V\mathbb{D}_{\lambda}^{1/2} \qquad G = U$$

este Biplot preserva la métrica de las filas, luego las distancias entre puntos de F son euclídeas y las distancias entre puntos de G son distancias de Mahalanobis. La longitud del vector f_i se relaciona con la desviación típica de la i-ésima fila de X.

Las coordenadas de las filas y columnas F y G son

$$F = \begin{pmatrix} 2{,}95 & -0{,}41 \\ -0{,}66 & -0{,}41 \\ -3{,}44 & 0 \\ 1{,}15 & 0{,}82 \end{pmatrix} \qquad G = \begin{pmatrix} -0{,}57 & 0{,}82 \\ -0{,}65 & -0{,}41 \\ -0{,}49 & -0{,}41 \end{pmatrix}$$

```
> F <- V%*%lambda2
> round(F,2)
      [,1]  [,2]
[1,] -2.95  0.41
[2,]  0.65  0.41
[3,]  3.44  0.00
[4,] -1.15 -0.82
> G <- U
> round(G,2)
      [,1]  [,2]
[1,] -0.57  0.82
[2,] -0.65 -0.41
[3,] -0.49 -0.41
```

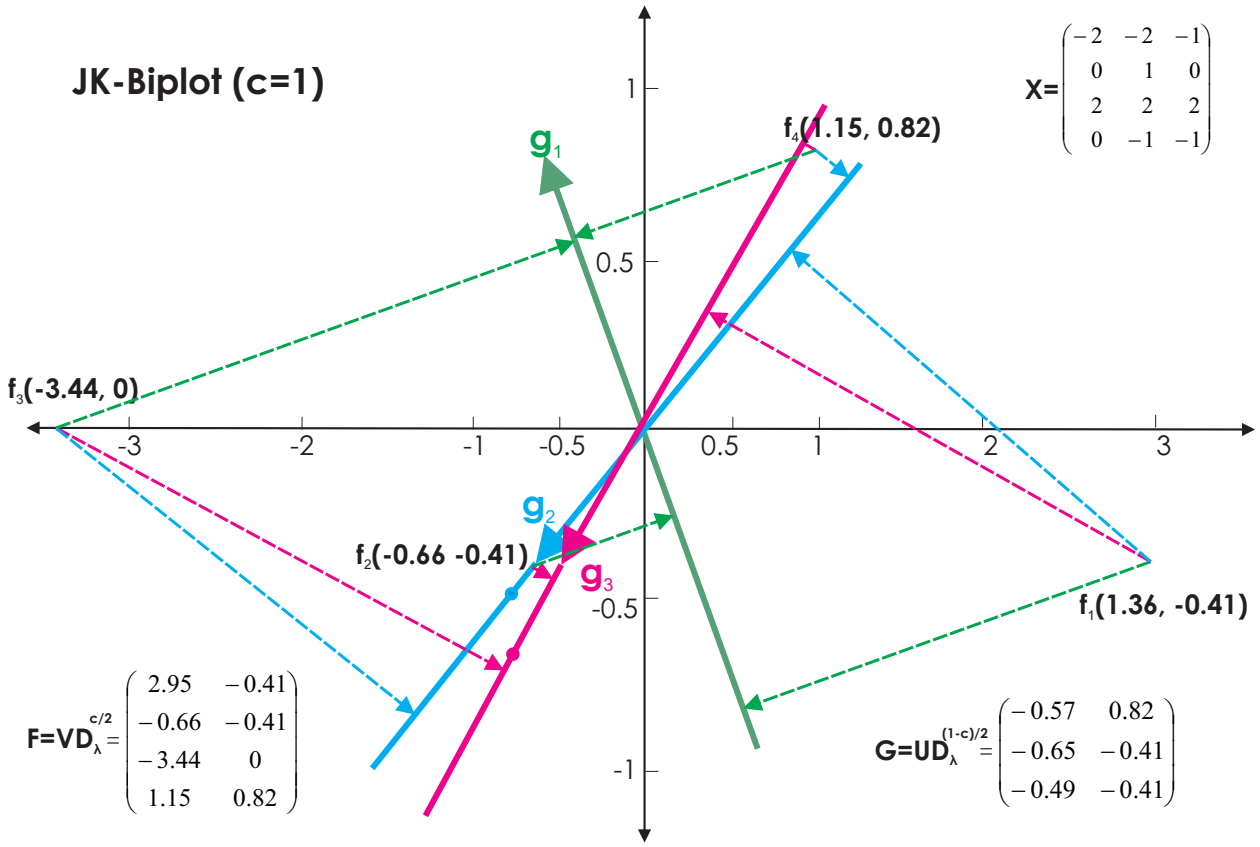

Figura 4.4: JK-biplot-1

Como es lógico, por definición los tres Biplot producen la misma estimación de los datos originales, difieren solo en la representación geométrica. en este caso la estimación obtenida es

$$X = \begin{pmatrix} -2 & -1{,}8 & -1{,}3 \\ 0 & 0{,}6 & 0{,}5 \\ 2 & 2{,}3 & 1{,}7 \\ 0 & -1{,}1 & -0{,}9 \end{pmatrix}$$

```
> X <- -F%*%t(G)
> round(X,1)
     [,1] [,2] [,3]
[1,]   -2 -1.8 -1.3
[2,]    0  0.6  0.5
[3,]    2  2.3  1.7
[4,]    0 -1.1 -0.9
```