UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

LOS MÉTODOS BIPLOT COMO HERRAMIENTA DE ANÁLISIS DE INTERACCIÓN DE ORDEN SUPERIOR EN UN MODELO LINEAL/BILINEAL

Autor: Mario Varela Nualles

Tutores: José Luis Vicente-Villardón y Antonio Blázquez Zaballos

Salamanca , 2002

INTRODUCCIÓN

En determinadas situaciones prácticas, podemos estar interesados en describir los diferentes tipos de interacciones presente en tablas multivías para datos continuos; es decir, tablas en las que se cruzan N factores de variación, o lo que es lo mismo, tablas en las que cada dato aparece identificado por una combinación de niveles de cada factor.

Como es conocido, para el caso de experimentos replicados, a partir de las técnicas clásicas del Análisis de la Varianza, podemos realizar los contrastes respectivos que nos indican al menos la existencia de determinadas interacciones. Sin embargo, la interpretación puede resultar muy complicada en la medida que aumentan las dimensiones del problema.

Para datos no replicados de dos vías se han desarrollado los modelos de No Aditividad de Tukey (TUKEY (1949)) y los modelos de MANDEL (MANDEL (1961)), en los cuales se trata de modelar la interacción mediante un solo término multiplicativo, formado a partir de los efectos principales. Estos modelos son efectivos solamente en los casos de estructuras de interacción muy simple en los datos.

Una clase de modelos más versátiles son los llamados modelos AMMI (Efecto Interacción Multiplicativo y Efectos Principales Aditivos), propuestos por GAUCH en 1988, basado en la idea de GOLLOB en 1968. En estos modelos se incorporan tantos términos multiplicativos como sean necesarios para explicar la variabilidad de la interacción de segundo orden. Se basan en la descomposición en valores y vectores singulares de la matriz de residuales de interacción del Modelo Lineal asociado.

Los Modelos AMMI a su vez, permiten la utilización de las representaciones BIPLOT propuestas por GABRIEL en 1971. Son gráficos o planos factoriales que reflejan en dimensión reducida las características más relevantes de una matriz de datos. La diferencia fundamental respecto a otras representaciones, es que en este caso se logra una representación conjunta; es decir, aparecen superpuestos en el mismo gráfico los puntos fila y puntos columna; en nuestro caso, las categorías de ambos factores de variación.

Las nuevas investigaciones que surgen dentro de la Estadística Multivariante, utilizan el Biplot como método gráfico por excelencia para representar en baja dimensión los resultados. En este sentido podemos citar además de los Modelos AMMI, los modelos de Regresión Factorial en Rango Reducido (IZENMAN (1975); TER BRAAK (1994)), el MANOVA BIPLOT (GABRIEL (1972); VICENTE-VILLARDÓN (1992); AMARO (2001)), así como técnicas de integración de matrices (KROONENBERG (1983); CARLIER y KROONENBERG (1996); MARTÍN-RODRIGUEZ (1996); VAN EEUWIJK y KROONENBERG (1998); MARTÍN-RODRIGUEZ et al (2002)), las cuales generalizan el Biplot al caso de varias matrices de datos.

Para tablas de más de dos factores de variación resulta mucho más complejo explicar las interacciones, ya que por ejemplo, en el caso de tres factores, pueden generarse tres interacciones de orden dos y una interacción triple. Las de orden dos podrán ser explicadas a partir de los modelos AMMI, y para la interacción triple necesitaremos hacer una generalización de estos modelos al caso de varias matrices de datos, o lo que es lo mismo, una generalización de la descomposición en valores singulares.

En el presente trabajo de tesis utilizamos el Biplot como herramienta para explicar la interacción de orden superior asociada a un modelo lineal; específicamente haremos énfasis en las interacciones de segundo y tercer orden, para tablas de dos y tres vías, respectivamente. Veremos cómo a partir de un Biplot podemos identificar las filas y columnas responsables de la interacción.

De igual forma utilizaremos el Biplot en el diagnóstico de modelos, resultado que nos permitirá identificar los diferentes tipos de interacciones presente en las tablas multivía, sobre las cuales debemos centrar la atención.

En el primer capítulo damos la definición de Biplot y las propiedades fundamentales de los diferentes tipos de representaciones. Se ofrecen los elementos necesarios que ayudan a interpretar este tipo de gráfico, haciendo énfasis en la información relacionada con los datos que podemos obtener a partir del Biplot. Abordamos el capítulo de forma detallada debido a que constituye el primer trabajo sobre Biplot presentado en Cuba.

El segundo capítulo lo dedicamos al análisis de la interacción de segundo orden asociada a una tabla de dos vías, específicamente se hace referencia a los modelos AMMI propuestos por GAUCH en 1988.

En este capítulo incorporamos además el Análisis de Regresión en Rango Reducido (IZENMAN (1975)), conocido también con el nombre de Análisis de Componentes Principales para variables instrumentales (RAO (1964); ROBERTS y ESCOUFIER (1976)) o Análisis de Redundancia (VAN DEN WOOLLENBERG (1977); ISRAELS (1984); VAN DER

BURG y DE LEEUW (1990)). Consiste en ajustar un modelo en el que tanto la parte a explicar como la parte explicativa son matrices.

Esta técnica la utilizamos para explicar la matriz de residuales de interacción de orden dos, a partir de una matriz con información de variables externas, las cuales pueden ser medidas bien sobre los niveles del primer factor (filas) o bien sobre los niveles del segundo factor (columnas). Los parámetros del modelo se estiman combinando las técnicas de Regresión Múltiple y Técnicas de Reducción de Dimensionalidad (Biplot).

Cada capítulo va acompañado de una aplicación a datos reales relacionado con el análisis de Interacción Genotipo Ambiente, donde se pretende clasificar genotipos o variedades en estables e inestables a partir de su interacción con el ambiente (localidades o años).

El análisis de la interacción Genotipo-Ambiente ha sido un problema abordado por los mejoradores durante mucho tiempo. En sus investigaciones conducen experimentos en varios sitios y durante varios años con el objetivo de seleccionar variedades que sean capaces de mantener buenos rendimientos en condiciones climáticas adversas; contribuyendo a extender el ciclo medio del cultivo

El tercer capítulo lo dedicamos al análisis de la interacción de orden tres; se añade un nuevo factor al análisis. En tal caso, los residuales de interacción triple asociados al modelo, quedan incluidos en K matrices de orden IxJ; siendo I, J y K el número de niveles respectivos de los factores considerados.

Para cumplimentar nuestro objetivo, abordamos el Análisis de Componentes Principales de Tres Modos (KROONENBERG (1983)), en particular el modelo propuesto por TUCKER (1966), para el que KROONENBERG y DE LEEUW (1980) ofrecen un algoritmo basado en la obtención de los estimadores a partir de la minimización de la suma de cuadrados residual (TUCKALS3).

El Análisis de Componentes Principales de Tres Modos aproxima un arreglo de tres vías a partir de tres matrices de marcadores o componentes, en nuestro caso particular, obtenemos una descomposición de tres vías de los residuales de interacción de tercer orden.

Para representar los residuales de interacción triple en dimensión reducida, hacemos uso de una generalización del Biplot al caso de tres matrices de marcadores (CARLIER y KROONENBERG (1996)). En tal sentido abordamos el Biplot Interactivo; en el que se concatenan dos de los modos, y el Biplot Conjunto en el que se proyectan los marcadores de dos de los modos sobre las componentes de un tercer modo de referencia.

Consideramos además un método para seleccionar el número de componentes a retener en cada modo (TIMMERMAN y KIERS (2000)), el cual asegura la obtención de un óptimo global y no local, al aplicar el algoritmo de TUCKALS3.

En este capítulo ofrecemos además una comparación entre el modelo de TUCKALS3 y otros métodos de integración de matrices: (Meta Componentes Principales (KRZANOWSKI (1990)) y Análisis de Componentes Principales Comunes (FLURY (1995)). Igualmente se hace

un estudio comparativo con el modelo PARAFAC/CANDECOM (HARSHMAN (1970); CARROLL y CHANG (1970, 1972)).

Se introduce además una generalización de la Regresión en Rango Reducido al caso de varias matrices de datos. Este resultado nos permitirá explicar los residuales de interacción triple a partir de la información de variables externas, medidas sobre cada uno de los factores de manera independiente, o sobre combinaciones de categorías de dos de los factores de variación analizados.

Como aplicación se considera nuevamente un estudio de interacción Genotipo Ambiente, en este caso los ambientes involucran localidades y años. Los genotipos son probados en varias localidades, durante varios años. Se presentan los resultados a través del Biplot Interactivo, el cual permite representar las tres matrices de marcadores asociada a la descomposición en tres vías de los residuales de interacción triple; lo que a su vez facilita la clasificación de los genotipos en estables e inestables.

En el capítulo 4 abordamos el Biplot como herramienta para la diagnosis de modelos asociado a tablas de tres vías. Demostramos cómo a partir de la distribución geométrica de los marcadores, resultado de aplicar el algoritmo de TUCKALS3 a una tabla de tres vías, podemos detectar la presencia/ausencia de interacción triple, y en los casos de ausencia, diagnosticar el modelo que mejor se ajusta a los datos.

Este resultado nos permitirá decidir acerca de los residuales de interacción doble que debemos explicar a partir de los modelos AMMI, y decidir sobre la necesidad de analizar los residuales de interacción de tercer orden a partir de la generalización del Biplot a varias matrices de datos.

La diagnosis a partir de representaciones Biplot ha sido abordada por BRADU y GABRIEL (1978) para tablas de dos vías y por BRADU (1983, 1984) y DIAZ-LENO (1995) para diagnosticar modelos de asociación entre variables ordinales. De igual forma, DÍAZ-LENO (1995) y GABRIEL, GALINDO y VICENTE-VILLARDÓN (1998) tratan la diagnosis de modelos logarítmico lineales jerárquicos gráficos adaptado a tablas de contingencia multivía.

En este trabajo se abordará por primera vez la diagnosis de modelos para tablas de datos continuos de más de dos vías. Se darán elementos relacionados con la distribución de los marcadores en el Biplot (generalizado al caso de más de dos vías), que nos permitirá por una parte, identificar la presencia en los datos de interacciones de orden superior; y por otra parte identificar el modelo que mejor describe los datos.

C A P Í T U L O I

LOS MÉTODOS BIPLOT

1.1 INTRODUCCIÓN

La forma tradicional de presentar la información en un Análisis Multivariante, es a partir de una matriz que contiene los valores de p variables observadas en n individuos. Para poder caracterizar los individuos en función de las variables observadas, es necesario reducir la dimensionalidad del problema; es decir, representar los individuos no en el p-hiperespacio de partida, sino en un subespacio de dimensión reducida, generalmente de dimensión 2.

Un BIPLOT (GABRIEL 1971) es una representación gráfica de datos multivariantes. La característica fundamental que lo hace diferenciar de las distintas representaciones gráficas asociadas a los métodos clásicos de reducción de dimensionalidad; es que en este caso es posible una representación conjunta de filas y columnas de la matriz de datos.

El BIPLOT trata de buscar la mejor aproximación en dimensión reducida (generalmente dos) de la distribución de una muestra multivariante. Superpone sobre dicha representación, vectores que representan las variables (columnas); e indican la dirección en la que mejor se muestra el cambio individual de cada variable.

El prefijo “bi” se refiere a la superposición en la misma representación de individuos y variables (filas y columnas de la matriz de datos).

De forma más general, un BIPLOT trata de aproximar los elementos de una matriz a partir de marcadores (vectores) asociados a las filas y columnas de la misma; dichos vectores se representan en un espacio cuya dimensión va a ser menor que el rango de la matriz.

La interpretación del BIPLOT se basa en conceptos geométricos muy sencillos, así por ejemplo:

- La similitud entre individuos (filas) es una función inversa de la distancia entre los mismos.
- Las longitudes y los ángulos de los vectores que representan a las variables, se interpretan en términos de variabilidad y covariabilidad respectivamente.
- Las relaciones entre filas y columnas se interpretan en términos de producto escalar, es decir, en términos de las proyecciones de los puntos “fila” sobre los vectores “columna”.

Un BIPLOT es aplicable a cualquier matriz de datos; no necesariamente las filas representan individuos y las columnas variables. Pueden referirse a las categorías de dos factores dentro de un análisis de varianza, o simplemente puede aplicarse a una tabla de contingencia que cruza dos variables cualitativas; entre otras aplicaciones.

Desde el punto de vista algebraico, el BIPLOT se basa en el mismo principio sobre el que se sustentan la mayoría de las técnicas factoriales de reducción de dimensionalidad, es decir, hace uso de la descomposición en valores y vectores singulares de la matriz. La diferencia fundamental es que en este caso se trata de reproducir el dato y se incorpora una representación conjunta de filas y columnas.

Las dos factorizaciones BIPLOT más importantes propuestas por GABRIEL (1971) fueron denominadas: GH-Biplot y JK-Biplot. El GH-Biplot consigue una alta calidad en la representación de las columnas (variables) y no tan alta para las filas (individuos); mientras que el JK-

Biplot consigue una alta calidad de representación para las filas, y no tan alta para las columnas.

GALINDO (1985, 1986) demuestra que con una conveniente elección de los marcadores es posible representar las filas y las columnas simultáneamente sobre un mismo sistema de coordenadas, con una alta calidad de representación tanto para las filas como para las columnas. GALINDO denomina a este tipo de BIPLOT, HJ-Biplot.

El BIPLOT no solamente se utiliza con fines descriptivos; puede ser aplicado en la diagnosis de modelos **(**BRADU y GABRIEL (1974, 1978)). En este sentido, podemos ver que con una simple inspección de la posición geométrica de los marcadores se puede diagnosticar el modelo que mejor describe los datos, y por tanto la presencia o no de interacción en caso que proceda (GOWER (1990); DIAZ-LENO (1995); BLÁZQUEZ (1998)).

El BIPLOT, ha dado lugar a nuevos métodos de análisis multivariante de datos, al ser combinado con otras técnicas clásicas; en este sentido podemos citar los modelos AMMI (GOLLOB (1968)), el cual inicialmente combina las técnicas de Análisis de Varianza y Análisis de Componentes Principales, y posteriormente incorpora el BIPLOT en lugar del A.C.P **(**GABRIEL (1978); KEMPTON (1984); GAUCH (1988)). Consiste en hacer un BIPLOT a la matriz de residuales del modelo.

Otra técnica que combina el BIPLOT en este caso con la regresión lineal simple, es la Regresión a Bajo Rango **(**IZENMAN (1975); TER BRAAK, (1994)), también conocida como Análisis de Componentes Principales para variables instrumentales (RAO (1964); ROBERTS y ESCOUFIER (1976); o Análisis de Redundancia (VAN DEN WOLLENBERG (1977)). Consiste

en ajustar un modelo donde tanto la parte a explicar como la parte explicativa son matrices.

Ambos métodos han sido aplicados en problemas agrícolas, más específicamente en el Análisis de la interacción Genotipo-Ambiente. (KEMPTON (1984); GAUCH (1988); TER BRAAK (1994); VAN EEUWIJK (1995 b y c)).

GABRIEL (1972) combina el Biplot con el MANOVA; introduce algunas características del MANOVA-BIPLOT de una vía; consiste en representar mediante un BIPLOT los resultados del MANOVA. Más tarde, AMARO (2001) lo generaliza al caso de dos factores de variación; lo cual facilita el estudio de los efectos principales e interacciones para cada una de las variables analizadas.

Al igual que otras técnicas clásicas de reducción de dimensionalidad, el Biplot ha sido generalizado al caso de varias matrices de datos. En tal sentido podemos citar el Biplot Conjunto y el Biplot Interactivo ((CARLIER Y KROONENBERG (1996)); los cuales operan con tres matrices de marcadores.

Actualmente continúan las investigaciones relacionadas con el BIPLOT. Las dos tendencias iniciales de investigación siguen desarrollándose (descripción y diagnosis). Sobre todo se trata de combinar el BIPLOT con métodos clásicos del análisis de datos; surgiendo nuevas técnicas de análisis cuya información queda resumida en un BIPLOT.

1.2 DEFINICIÓN

Como en la mayoría de las técnicas de Análisis de Datos, partimos de una matriz **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 **X** en un espacio de dimensión reducida, con la pérdida mínima de información.

Si la matriz **X** es de rango dos, es posible lograr una representación exacta en dos dimensiones; en caso contrario se necesitarán tantos ejes como rango tenga **X**, para lograr un ajuste perfecto. Sin embargo, en un Biplot se sigue el mismo principio que en las técnicas factoriales de reducción de dimensionalidad, por lo que en la mayoría de los casos los últimos ejes serán residuales, es decir tendrán asociada una variabilidad despreciable, y serán eliminados. Tendremos por tanto una buena aproximación de los elementos de **X** en dimensión reducida.

Un Biplot para una matriz de datos **X** es una representación gráfica mediante marcadores (vectores): $\mathbf{a}_1,\mathbf{a}_2,...,\mathbf{a}_n$ para las filas de **X** y $\mathbf{b}_1,\mathbf{b}_2,...,\mathbf{b}_p$ para las columnas de **X**, de forma tal que el producto interno aproxime el elemento x_{ij} de la matriz de partida lo mejor posible.

Tanto los marcadores $\mathbf{a}_i$ para las filas, como los marcadores $\mathbf{b}_j$ para las columnas estarán representados en un espacio de dimensión $q \leq r$, siendo q el número de ejes retenidos y r el rango de **X**.

Si consideramos los marcadores $\mathbf{a}_1,\mathbf{a}_2,...,\mathbf{a}_n$ como filas de una matriz **A** y los marcadores $\mathbf{b}_1,\mathbf{b}_2,...,\mathbf{b}_p$ como filas de una matriz **B**, entonces podemos escribir:

$$\mathbf{X} \cong \mathbf{A}\mathbf{B}^{T}$$

La estructura de la matriz **X** puede visualizarse representando los marcadores en un espacio euclideo de q dimensiones. Generalmente se trata de tomar q lo más pequeño posible, ello estará en dependencia de sí existen o no estructuras de covariación significativa entre las columnas de **X**.

1.3 OBTENCIÓN DE MARCADORES

1.3.1 MÉTODO CLÁSICO

Se trata de buscar una matriz $\mathbf{X}_{(q)}$ de rango q, que aproxime lo mejor posible a **X**, en el sentido de los mínimos cuadrados $(\mathbf{X} \cong \mathbf{X}_{(q)} = \mathbf{A}_{(q)}\mathbf{B}_{(q)}^{T})$, más específicamente, se trata de buscar una matriz $\mathbf{X}_{(q)}$ de rango q que minimice la expresión:

$$\sum_i \sum_j (x_{ij} - x_{(q)ij})^2 = \text{traza}((\mathbf{X} - \mathbf{X}_{(q)})(\mathbf{X} - \mathbf{X}_{(q)})')$$

para todas las matrices $\mathbf{X}_{(q)}$ de rango q o menor.

El método más conocido para aproximar una matriz a bajo rango es el propuesto por ECKART y YOUNG (1936,1939) que puede encontrarse también en YOUNG y HOUSEHOLDER (1938), GABRIEL (1971), GREENACRE (1984), entre otros autores. Se basa en la descomposición en valores y vectores singulares de la matriz que deseamos aproximar.

Descomposición en valores y vectores singulares de la matriz **X**:

$$\mathbf{X = UDV'}$$

siendo **U** la matriz cuyas columnas contienen los vectores propios de **XX**' y **V** la matriz cuyas columnas corresponden a los vectores propios de **X'X**, mientras que **D** es una matriz diagonal que contiene a los valores singulares de **X**. Debe cumplirse que **U'U**=**V'V**=**I**, es decir, las columnas de **U** y **V** son ortonormales, esta propiedad asegura la unicidad de la factorización.

La mejor aproximación en rango q, $\mathbf{X}_{(q)}$ de **X** viene dada por:

$$\mathbf{X}_{(q)_{nxp}} = \mathbf{U}_{(q)_{nxq}} \mathbf{D}_{(q)_{qxq}} \mathbf{V'}_{(q)_{qxp}} = \sum_{k=1}^{q} \lambda_k \mathbf{u}_k \mathbf{v}_k'$$

donde, $\mathbf{U}_{(q)}$ y $\mathbf{V}_{(q)}$ son las matrices construidas con las q primeras columnas de **U** y **V** respectivamente, mientras que $\mathbf{D}_{(q)}$ es la matriz diagonal que contiene los q mayores valores singulares distintos de cero de **X** (λ_k).

Un algoritmo para el cálculo puede verse en GOLUB y REINSCH (1971).

Tenemos por tanto:

$$\mathbf{X = AB' = UDV'}$$

Ello implica que la elección de los marcadores para filas y columnas puede realizarse de varias maneras: Por ejemplo, tomando **A**=**UD** y **B**=**V** o **A**=**U** y **B**=**VD'** entre otras factorizaciones.

Por esta razón, varios autores proponen distintas elecciones y estudian sus propiedades de acuerdo con la factorización elegida. No obstante, la

interpretación del Biplot siempre se realiza a partir de los productos escalares, independientemente de la factorización elegida.

Comenzaremos con la descripción y propiedades de los Biplots clásicos (GABRIEL (1971)) y posteriormente nos referiremos a las modificaciones introducidas por GALINDO (1986).

La forma usual de elegir los marcadores consiste en realizar la descomposición:

$$\mathbf{A}=\mathbf{U}\mathbf{D}^{\gamma} \qquad \mathbf{B}=\mathbf{V}\mathbf{D}^{1-\gamma}$$

GABRIEL (1971) propone diversas elecciones de γ a las que da diversos nombres y para las cuales demuestra algunas de sus propiedades.

Con $\gamma=1$ obtenemos:

$$\mathbf{A}=\mathbf{U}\mathbf{D} \qquad \mathbf{B}=\mathbf{V}$$

Se verifica que $\mathbf{B}'\mathbf{B}=\mathbf{I}$ y tenemos el JK-Biplot el cual preserva la métrica para las filas.

Con $\gamma=0$ obtenemos:

$$\mathbf{A}=\mathbf{U} \qquad \mathbf{B}=\mathbf{V}\mathbf{D}$$

Se verifica que $\mathbf{A}'\mathbf{A}=\mathbf{I}$ y tenemos en este caso el GH-Biplot el cual preserva la métrica para las columnas.

De manera general hemos llamado a las matrices de marcadores, **A** para las filas y **B** para las columnas, en lo adelante la llamaremos de manera diferente en cada tipo de BIPLOT, por ejemplo:

GH-Biplot (**A**=**G** **B**=**H);** JK-Biplot (**A**=**J** **B**=**K**); HJ-Biplot (**A**=**J** **B**=**H**). Esto nos permitirá identificar los diferentes Biplots.

1.3.2 MÍNIMOS CUADRADOS ALTERNADOS

En ocasiones no es posible realizar la descomposición en valores y vectores singulares de la matriz debido a que pueden presentarse celdas vacías (datos faltantes) o bien los elementos de la matriz pueden presentar diferente ponderación o importancia. En tal caso, GABRIEL y ZAMIR (1979) ofrecen una alternativa para la estimación de los parámetros (marcadores), llamada en su trabajo "criss-cross", pero más comúnmente conocida como algoritmo de mínimos cuadrados alternados.

Como sabemos, en un Biplot ajustamos el siguiente modelo:

$$\mathbf{X} = \mathbf{AB'} + \mathbf{E}$$

donde E es una matriz de residuales.

Supongamos que las coordenadas para las filas **A** están fijadas de antemano. Tenemos entonces que las coordenadas para las columnas pueden calcularse como la matriz **B** que hace mínima la suma de cuadrados de los residuos dada por la siguiente expresión:

$$L = \left\|\mathbf{X} - \mathbf{AB}'\right\| = \text{traza}\mathbf{((X - AB')'(X - AB'))} =$$
$$\text{tr}\mathbf{(X'X)} - 2\text{tr}\mathbf{(X'AB')} + \text{tr}\mathbf{(BA'AB')}$$

La solución viene dada por la matriz:

$$\mathbf{B}' = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{X} \qquad (1)$$

es decir, las filas de **B** son los coeficientes de regresión obtenidos en la regresión de cada columna de la matriz original **X** sobre las columnas de **A**.

De la misma manera, si escribimos:

$$\mathbf{X}' = \mathbf{B}\mathbf{A}' + \mathbf{E}'$$

y fijamos los valores de **B**, podemos obtener los valores de **A** que hacen mínima la suma de cuadrados de los residuales dada por la siguiente expresión:

$$\mathrm{L} = \left\|\mathbf{X}' - \mathbf{B}\mathbf{A}'\right\| = \mathrm{traza}\mathbf{((X' - BA')'(X' - BA'))} = \mathrm{tr}\mathbf{(XX')} - 2\mathrm{tr}\mathbf{(XBA')} + \mathrm{tr}\mathbf{(AB'BA')}$$

la solución viene dada por

$$\mathbf{A}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{X}' \qquad (2)$$

Es decir, las filas de **A** son los coeficientes de regresión obtenidos en la regresión de cada columna de la matriz original **X** sobre las columnas de **B**. Por tanto, partiendo de valores iniciales arbitrarios para **A** (o **B**) y alternando las fórmulas (1) y (2), se construye un algoritmo con el que se obtienen los mismos valores esperados que con la descomposición en

valores singulares descrita anteriormente. (ver demostración en (BLÁZQUEZ (1998))).

1.4 PROPIEDADES DE LOS BIPLOTS CLÁSICOS

1.4.1 GH-BIPLOT

Si hacemos un ajuste en los marcadores asociados a un GH-Biplot (multiplicamos y dividimos por un factor de escala), es decir, tomamos:

$$\mathbf{G} = \sqrt{(n-1)}\mathbf{U} \qquad \mathbf{H} = {}^{1}\!\Big/\!_{\sqrt{n-1}}\,\mathbf{VD}$$

y trabajamos con los datos originales, centrados por columnas, para que la matriz de varianzas y covarianzas (**S**) sea proporcional a **X'X**:

$$\mathbf{S}=\mathbf{X'X}/(n-1).$$

entonces,

1- El GH-Biplot conserva la métrica para las columnas, lo cual significa que los productos escalares de los marcadores asociados a las columnas, son iguales a los productos escalares de las columnas de **X**, que son a su vez las varianzas y covarianzas.

En efecto,

$$\mathbf{S} = \frac{1}{(n-1)}\mathbf{X}'\mathbf{X} = \frac{1}{(n-1)}(\mathbf{GH}')'(\mathbf{GH}') = \frac{1}{(n-1)}\mathbf{HG}'\mathbf{GH}' = \mathbf{HU}'\mathbf{UH}' = \mathbf{HH}'$$

Además, se tiene que la descomposición espectral de la matriz de covarianzas es también su descomposición en valores singulares:

$$\mathbf{S} = \frac{1}{(n-1)}\mathbf{X}'\mathbf{X} = \frac{1}{(n-1)}\mathbf{V}\mathbf{D}^2\mathbf{V}'$$

luego, la mejor aproximación de la matriz de covarianzas en rango q es:

$$\mathbf{S} \cong \mathbf{S}_{(q)} = \frac{1}{(n-1)}\mathbf{V}_{(q)}\mathbf{D}_{(q)}\mathbf{D}_{(q)}\mathbf{V}'_{(q)} = \mathbf{H}_{(q)}\mathbf{H}'_{(q)}$$

que coincide con la que se obtiene en el Biplot de la matriz **X.**

En consecuencia, se cumplen dos propiedades muy importantes del GH-Biplot:

- La longitud al cuadrado de los vectores $\mathbf{h}_j$, aproxima la varianza de la variable, por lo tanto la longitud aproxima la desviación típica:

$$\mathbf{S}_{jj} = \mathbf{h}_j'\mathbf{h}_j$$

- El coseno del ángulo que forman dos marcadores columna, aproxima la correlación entre la variables asociadas a estas columnas:

$$\mathbf{h}_i'\mathbf{h}_j = \|\mathbf{h}_i\|\|\mathbf{h}_j\|\cos(\mathbf{h}_i,\mathbf{h}_j) \Rightarrow \cos(\mathbf{h}_i,\mathbf{h}_j) = \frac{\mathbf{h}_i'\mathbf{h}_j}{\|\mathbf{h}_i\|\|\mathbf{h}_j\|} \cong \frac{s_{ij}}{s_{ii}s_{jj}} = r_{ij}$$

Refiriéndonos a las filas,

2- En un GH-Biplot la distancia de Mahalanobis entre dos filas de **X** coincide con la distancia euclidea entre dos marcadores fila. En dimensión reducida se consigue entonces, una aproximación de la distancia de Mahalanobis.

En efecto, cada elemento de la matriz **X** puede escribirse como:

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

de forma que cada fila i de **X** puede escribirse como:

$$\mathbf{Hg}_i$$

La distancia de Mahalanobis entre dos filas i y j puede aproximarse como:

$$\delta_{ij}^2=\sum_k (x_{ik}-x_{jk})'\mathbf{S}^{-1}(x_{ik}-x_{jk})=(\mathbf{Hg}_i-\mathbf{Hg}_j)'\mathbf{S}^{-1}(\mathbf{Hg}_i-\mathbf{Hg}_j)=$$

$$(\mathbf{g}_i-\mathbf{g}_j)'\mathbf{H}'\mathbf{S}^{-1}\mathbf{H}(\mathbf{g}_i-\mathbf{g}_j)=\frac{1}{(n-1)}(\mathbf{g}_i-\mathbf{g}_j)'\mathbf{DV}'\mathbf{S}^{-1}\mathbf{VD}(\mathbf{g}_i-\mathbf{g}_j)=$$

$$\frac{1}{(n-1)}(\mathbf{g}_i-\mathbf{g}_j)'\mathbf{DV}'(n-1)(\mathbf{VD}^{-2}\mathbf{V}'\mathbf{V})\mathbf{D}(\mathbf{g}_i-\mathbf{g}_j)=(\mathbf{g}_i-\mathbf{g}_j)'(\mathbf{g}_i-\mathbf{g}_j)$$

En dimensión reducida se tiene que,

$$\delta_{ij}^2 = \sum_k (x_{ik} - x_{jk})' \mathbf{S}^{-1} (x_{ik} - x_{jk}) = (\mathbf{g}_i - \mathbf{g}_j)'(\mathbf{g}_i - \mathbf{g}_j)$$

La propiedad podría haberse enunciado en términos de los productos escalares calculados con la métrica asociada a la inversa de la matriz de covarianzas, de la forma:

$$\mathbf{XS}^{-1}\mathbf{X}=\mathbf{GG}'$$

Se consigue en dimensión reducida, una aproximación del producto escalar con la métrica de Mahalanobis.

3- El GH-Biplot proporciona una mejor aproximación para la matriz de covarianzas que para la distancia de Mahalanobis entre puntos fila.

Como ya vimos, la matriz de varianzas y covarianzas puede escribirse de la forma:

$$\mathbf{S} = \frac{1}{(n-1)} \mathbf{X}'\mathbf{X} = \frac{1}{(n-1)} \mathbf{V}\mathbf{D}^2\mathbf{V}'$$

de donde se deduce que si realizamos una aproximación a bajo rango como:

$$\mathbf{S} \cong \mathbf{S}_{(q)} = \frac{1}{(n-1)} \mathbf{V}_{(q)}\mathbf{D}_{(q)}\mathbf{D}_{(q)}\mathbf{V}'_{(q)} = \mathbf{H}_{(q)}\mathbf{H}'_{(q)}$$

tenemos una bondad de ajuste para la aproximación de la matriz de varianzas-covarianzas de:

$$\frac{\sum_{k=1}^{q}\lambda_k^4}{\sum_{k=1}^{r}\lambda_k^4}$$

Para las filas de la matriz **X** la situación es diferente. La suma de cuadrados de los elementos de $\mathbf{XS}^{-1}\mathbf{X}$ es r (el rango de **X**, que generalmente coincide con el número de columnas). Si aproximamos en dimensión q mediante:

$$\mathbf{XS}^{-1}\mathbf{X} \cong \mathbf{G}_{(q)}\mathbf{G'}_{(q)}$$

La suma de cuadrados (explicada en la aproximación) de $\mathbf{G}_{(q)}\mathbf{G'}_{(q)}$ es precisamente q, luego la bondad del ajuste de la aproximación de los productos escalares en la métrica de Mahalanobis es q/r, que por lo general, es mucho menor que la anterior.

1.4.2 JK-BIPLOT

Suponemos que los datos están centrados. Los marcadores para filas y columnas, en dimensión q, son los siguientes:

$$\mathbf{J}_{(q)}=\mathbf{U}_{(q)}\mathbf{D}_{(q)} \qquad \mathbf{K}_{(q)}=\mathbf{V}_{(q)}$$

Las propiedades más relevantes son las siguientes:

1- Los productos escalares, con la métrica identidad, de las filas de la matriz **X**, coinciden, en el espacio completo, con los productos escalares de los marcadores contenidos en **J**. La aproximación de dichos productos

escalares en dimensión reducida es óptima en el sentido de los mínimos cuadrados.

En efecto:

$$\mathbf{XX'=JK'KJ'=JV'VJ'=JJ'}$$

Además, se tiene que la descomposición espectral de la matriz de productos escalares entre las filas es también su descomposición en valores singulares:

$$\mathbf{XX'=UD^2U'}$$

luego, la mejor aproximación en rango q es:

$$\mathbf{XX'=U_{(q)}D^2_{(q)}U'_{(q)}=J_{(q)}J'_{(q)}}$$

que coincide con la que se obtiene en el Biplot de la matriz **X**.
En consecuencia, la distancia euclidea entre dos filas de **X,** coincide en el espacio completo, con la distancia euclidea entre los marcadores **J**.

Se cumple además, que los marcadores para las filas coinciden con las coordenadas de los individuos en el espacio de las componentes principales:

$$\mathbf{XV_{(q)}=UDV'V_{(q)}=U_{(q)}D_{(q)}=J_{(q)}}$$

Esta propiedad implica que podemos estudiar las similitudes entre los individuos con pérdida de información mínima, siempre que la distancia euclidea sea adecuada.

2- Los marcadores para las columnas son las proyecciones de los ejes originales (base canónica en el espacio p dimensional) en el espacio de las componentes principales.

Este resultado puede verse en LEBART et al (1995)**;** considera un JK-Biplot como un Análisis de Componentes principales con variables suplementarias.

3- La calidad de representación es mejor para las filas que para las columnas. (Demostración análoga a la realizada en el GH-Biplot).

1.5 HJ-BIPLOT. PROPIEDADES.

Como hemos comprobado en apartados anteriores, las representaciones son asimétricas en el sentido de que no obtienen la misma calidad de representación para las filas y para las columnas de la matriz de datos. Cuando el propósito es la aproximación de los elementos de la matriz original, los biplots presentados son óptimos, además en cada uno de ellos es posible representar con mejor calidad las características de las filas o de las columnas, cuando se quieren interpretar por separado.

Cuando las filas y las columnas son importantes en sí mismas, y se quieren interpretar las características de ambas manteniendo cierta relación entre las mismas; son más útiles las interpretaciones basadas en representaciones simétricas como el Análisis Factorial de Correspondencias, en el que se interpretan las posiciones de las filas, las posiciones de las columnas y las relaciones fila-columna a través de los factores, es decir, se realiza una interpretación factorial.

Sin embargo, el Análisis Factorial de Correspondencias está pensado solamente para matrices de frecuencias. Sería interesante disponer de una técnica simétrica similar, pero aplicable a cualquier conjunto de datos.

GALINDO (1986) propone el que denomina HJ-BIPLOT que responde a las características descritas en los párrafos anteriores.

Un HJ-BIPLOT para una matriz de datos **X** es una representación gráfica multivariante mediante marcadores (vectores) $\mathbf{j}_1, \mathbf{j}_2,...,\mathbf{j}_n$ para las filas y $\mathbf{h}_1,\mathbf{h}_2,...,\mathbf{h}_p$ para las columnas de **X**, elegidos de forma que ambos marcadores puedan superponerse en el mismo sistema de referencia con máxima calidad de representación.

Partimos también de la descomposición en valores singulares:

$$\mathbf{X}=\mathbf{UDV'}$$

elegimos como marcadores en dimensión reducida:

$$\mathbf{J}_{(q)}=\mathbf{U}_{(q)}\mathbf{D}_{(q)} \qquad \mathbf{H}_{(q)}=\mathbf{V}_{(s)}\mathbf{D}_{(s)}$$

Nótese que con esta factorización el dato original no se reproduce, en efecto:

$$\mathbf{X} \neq \mathbf{JH'}$$

Sin embargo, el objetivo es lograr una máxima calidad de representación para filas y columnas de **X**, para ello resulta necesario, como vimos en los

Biplots clásicos, incorporar a cada matriz de marcadores, la matriz diagonal **D**, lo cual posibilita que la descomposición espectral tanto para la matriz de varianzas y covarianzas entre columnas, como para la matriz de distancia euclidea entre filas, coincida con la descomposición en valores singulares de **X**.

Las propiedades generales del HJ-BIPLOT son las de los marcadores elegidos, añadimos aquí las propiedades relativas a las representaciones simétricas.

1- Los marcadores fila y columna se pueden representar en el mismo sistema de referencia, con la misma calidad de representación.

En el contexto de las correspondencias, GREENACRE (1984) basa esta afirmación en que ambas nubes están referidas a los mismos valores propios y por tanto están relacionadas.

El que las nubes están referidas a los mismos valores propios es obvio, ya que los valores propios de **X'X** y **XX'** son los mismos.

Las relaciones entre las nubes son las relaciones baricéntricas similares a las del Análisis Factorial de Correspondencias, concretamente:

$$\mathbf{J}_{(q)}=\mathbf{U}_{(q)}\mathbf{D}_{(q)}=\mathbf{XV}_{(q)}=\mathbf{XX'U}_{(q)}\mathbf{D}_{(q)}^{-1}=\mathbf{XH}_{(q)}\mathbf{D}^{-1}{}_{(q)}$$

$$\mathbf{H}_{(q)}=\mathbf{V}_{(q)}\mathbf{D}_{(q)}=\mathbf{X'U}_{(q)}=\mathbf{X'XV}_{(q)}\mathbf{D}^{-1}{}_{(q)}=\mathbf{X'J}_{(q)}\mathbf{D}^{-1}{}_{(q)}$$

Es decir, las coordenadas para las filas son medias ponderadas de las coordenadas de las columnas, donde las ponderaciones son los valores

originales en la matriz **X**. Lo mismo ocurre con las coordenadas de las columnas respecto de las filas.

2- Las propiedades del HJ-Biplot son las de los marcadores **J** y **H** detalladas en apartados anteriores.

1.6 SELECCIÓN DEL NÚMERO DE EJES

Los métodos factoriales gráficos presentan los resultados en forma de diagramas de dispersión, generalmente en un subespacio de dimensión 2; aunque la configuración original sea de dimensión mayor. Al proyectarse produce una pérdida de información que puede distorsionar las configuraciones iniciales.

El primer problema a tener en cuenta es el número de dimensiones necesarias para obtener una representación adecuada en dimensión reducida. Debido a que la obtención secuencial de cada uno de los ejes de la representación es idéntica a la obtenida del ajuste conjunto de todos ellos, es posible elegir el número de ejes necesarios después de realizar el cálculo de la descomposición en valores singulares.

En la literatura se presentan varios procedimientos que permiten la búsqueda del número de dimensiones necesarias para describir de forma óptima la nube de puntos. Los métodos están descritos inicialmente para el Análisis de Componentes Principales o de Correspondencias, pero pueden ser extendidos a los métodos Biplot.

Como en toda técnica factorial, debemos conocer qué parte de la variabilidad total es explicada por los q ejes o factores retenidos, o lo que

es lo mismo debemos dar una medida de cuan buena es la aproximación $\mathbf{X}_{(q)}$ de **X**.

Para ello debemos hacer una descomposición de la variabilidad total, en variabilidad explicada por el Biplot y variabilidad no explicada o residual. Sabemos que la variabilidad total asociada a una matriz **X,** se calcula por la suma de sus elementos al cuadrado, que a su vez es igual a la traza de $\mathbf{XX}'$ y se representa por:

$$\sum_{k=1}^{r} \lambda_k^2$$

por la misma razón la variabilidad asociada a $\mathbf{X}_{(q)}$; valor que representa la variabilidad explicada por el Biplot, se calcula de la siguiente forma:

$$\sum_{k=1}^{q} \lambda_k^2$$

Por tanto la variabilidad residual, que a su vez corresponde a la variabilidad asociada a la matriz ($\mathbf{X}$-$\mathbf{X}_{(q)}$) se calcula de la manera siguiente:

$$\sum_{k=q+1}^{r} \lambda_k^2$$

Por tanto:

$$\text{S.C.Total} = \text{S.C.Explicada} + \text{S.C.}\mathbf{Re}\text{sidual} \qquad \left(\sum_{k=1}^{r} \lambda_k^2 = \sum_{k=1}^{q} \lambda_k^2 + \sum_{k=q+1}^{r} \lambda_k^2\right)$$

Lo que significa que una medida de la Bondad de ajuste del Biplot puede calcularse por la cantidad:

$$\frac{\sum_{k=1}^{q}\lambda_k^2}{\sum_{k=1}^{r}\lambda_k^2} * 100\%$$

1.7 INTERPRETACIÓN DE RESULTADOS

Supongamos que hemos seleccionado un número de dimensiones suficiente para explicar correctamente el comportamiento de los datos.

En la representación Biplot en dimensión reducida, interpretaremos las distancias entre individuos como disimilaridades entre los mismos, especialmente si los individuos están bien representados; en un GH-Biplot interpretamos la longitud de los vectores que representan a las variables en términos de variabilidad y los ángulos que forman dos vectores en términos de correlación; en un JK-Biplot, no podemos hacer este tipo de interpretaciones para las variables aunque las coordenadas, nos darán una idea aproximada de cual es la relación con los ejes.

La relación individuo-variable la estudiaremos a través de la proyección de los puntos que representan a los individuos sobre los vectores que representan a las variables, esto nos permite determinar cuáles son las variables que más diferencian subconjuntos de individuos.

$$x_{ij} \cong \mathbf{a}_i'\mathbf{b}_j \quad \Rightarrow x_{ij} \cong \left\| \text{proy } \mathbf{a}_i / \mathbf{b}_j \right\| (\text{signo}) \left\| \mathbf{b}_j \right\|$$

En la representación HJ-Biplot la interpretación es la misma, sin embargo, la búsqueda de las variables que determinan las diferencias entre los individuos se realiza a través de los ejes factoriales, es decir, se interpretan las nuevas variables, combinación lineal de las de partida, y las relaciones de las mismas con las variables observadas; como se hacía en un Análisis de Componentes Principales.

La medida de la relación entre los ejes de la representación Biplot y cada una de las variables observadas es lo que se denomina Contribución Relativa del Factor al Elemento (variable), y representa la parte de la variabilidad de cada una de las variables explicada por el factor, y se interpreta de la misma manera que un coeficiente de determinación en regresión, de hecho, si los datos están centrados, es el coeficiente de determinación de la regresión de cada variable sobre el eje correspondiente.

Esta contribución nos permitirá saber cuáles son las variables más directamente relacionadas con cada eje y, por tanto, nos permite conocer las variables responsables de la colocación de los individuos sobre las proyecciones en cada uno de los ejes.

Como los ejes se construyen para que sean independientes, la contribución de cada uno de ellos a cada variable es independiente, por tanto, es posible calcular la contribución de un plano sin más que sumar las contribuciones de los ejes que lo forman.

1.7.1 CONTRIBUCIONES

Es fácil ver que la suma de cuadrados de las coordenadas principales, tanto para filas como para columnas en cada eje factorial, es igual al valor propio de la matriz de productos escalares correspondiente, o al cuadrado del valor singular:

$$\sum_{i=1}^{n} a_{il}^2 = \lambda_l^2 \qquad \sum_{j=1}^{p} b_{jl}^2 = \lambda_l^2$$

La situación general puede resumirse para las filas en la siguiente tabla:

		ejes					
		1	…	l	…	r	suma
filas	1	a_{11}^2	…	a_{1l}^2	…	a_{1r}^2	$\sum_{k=1}^{r} a_{1k}^2$
	⋮	⋮	⋱	⋮	⋱	⋮	⋮
	i	a_{i1}^2	…	a_{il}^2	…	a_{ir}^2	$\sum_{k=1}^{r} a_{ik}^2$
	⋮	⋮	⋱	⋮	⋱	⋮	⋮
	n	a_{n1}^2	…	a_{nl}^2	…	a_{nr}^2	$\sum_{k=1}^{r} a_{nk}^2$
	suma	λ_1^2	…	λ_l^2	…	λ_r^2	$\sum_{i=1}^{n}\sum_{k=1}^{r} a_{ik}^2 = \sum_{k=1}^{r} \lambda_k^2$

De esta forma, cada una de las coordenadas al cuadrado puede considerarse como la contribución absoluta a la variabilidad total.

Las contribuciones absolutas pueden convertirse en contribuciones relativas sin más que dividir por el total adecuado.

La cantidad,

$$CRT_i = \frac{\sum_{k=1}^{r} a_{ik}^2}{\sum_{k=1}^{r} \lambda_k^2}$$

se denomina, contribución relativa a la traza (variabilidad total) del elemento (fila) i; muestra la parte de la variabilidad total explicada por la fila i.

La cantidad,

$$CRE_i F_l = \frac{a_{il}^2}{\lambda_l^2}$$

se denomina contribución relativa del elemento (fila) i al factor l, y muestra la parte de la variabilidad del factor explicada por el individuo i.

La cantidad,

$$CRF_l E_i = \frac{a_{il}^2}{\sum_{k=1}^{r} a_{ik}^2}$$

se denomina, contribución relativa del factor l al elemento (fila) i, y muestra la parte de la variabilidad de la fila i, explicada por el factor l.

En consecuencia, la cantidad:

$$\frac{\sum_{k=1}^{q} a_{ik}^2}{\sum_{k=1}^{r} a_{ik}^2}$$

es una medida de la parte de la variabilidad asociada a la fila i que es explicada por los q factores retenidos, es decir explicada por el Biplot (calidad de representación).

De la misma forma es posible definir las contribuciones correspondientes a las variables:

		ejes					
		1	…	l	…	r	suma
cols	1	b_{11}^2	…	b_{1l}^2	…	b_{1r}^2	$\sum_{k=1}^{r} b_{1k}^2$
	⋮	⋮	⋱	⋮	⋱	⋮	⋮
	j	b_{j1}^2	…	b_{jl}^2	…	b_{jr}^2	$\sum_{k=1}^{r} b_{jk}^2$
	⋮	⋮	⋱	⋮	⋱	⋮	⋮
	p	b_{p1}^2	…	b_{pl}^2	…	b_{pr}^2	$\sum_{k=1}^{r} b_{pk}^2$
	suma	λ_1^2	…	λ_l^2	…	λ_r^2	$\sum_{j=1}^{p}\sum_{k=1}^{r} b_{jk}^2 = \sum_{k=1}^{r} \lambda_k^2$

$$CRT_j = \frac{\sum_{k=1}^{r} b_{jk}^2}{\sum_{k=1}^{r} \lambda_k^2}$$

representa la contribución relativa a la traza, del elemento (columna) j, y muestra la parte de la variabilidad total que es explicada por la variable j.

$$CRE_j F_l = \frac{b_{jl}^2}{\lambda_l^2}$$

representa la contribución relativa del elemento (columna) j, y muestra la parte de la variabilidad del factor l explicada por la variable j.

$$CRF_lE_j = \frac{b_{jl}^2}{\sum_{k=1}^{r} b_{jk}^2}$$

representa la contribución relativa del factor l al elemento (columna) j, y muestra la parte de la variabilidad de la variable j que es explicada por el factor l.

De igual forma, una medida de la calidad de representación de la columna j, en el espacio q-dimensional es:

$$\frac{\sum_{k=1}^{q} b_{jk}^2}{\sum_{k=1}^{r} b_{jk}^2}$$

1.7.2 ESTIMACIÓN DE FUNCIONES DE LAS OBSERVACIONES

Un Biplot nos permite además, estimar mediante proyecciones, los valores medios para filas y columnas; así como los efectos principales e interacciones en una tabla de dos vías (tabla 1.1). Estos resultados son muy utilizados cuando estamos interesados en diagnosticar modelos a partir de un Biplot.

Por ejemplo, si queremos representar la media de una de las filas a partir de los marcadores, utilizamos la siguiente igualdad:

$$x_{i.} = 1/n \sum_{j=1}^{p} x_{ij} = 1/n \sum_{j=1}^{p} a'_i b_j = a'_i \,(1/n \sum_{j=1}^{p} b_j) = a'_i \, b_.$$

donde $\mathbf{b}_.$ es el vector de medias de las coordenadas de las columnas.

La tabla 1.1. resume las funciones de las observaciones y su estimación sobre el Biplot, que serán útiles en la interpretación de modelos asociados a un diseño experimental.

Función	**Marcadores**	**Estimación Gráfica**
$\mathbf{x}_{ij}$	$a_i' b_j$	$\lVert \text{proy } a_i / b_j \rVert (s) \lVert b_j \rVert = = \lVert \text{proy } b_j / a_i \rVert (s) \lVert a_i \rVert$
$\mathbf{x}_{i.}$	$a_i' b_.$	$\lVert \text{proy } a_i / b_. \rVert (s) \lVert b_. \rVert = = \lVert \text{proy } b_. / a_i \rVert (s) \lVert a_i \rVert$
$\mathbf{x}_{.j}$	$a_.' b_j$	$\lVert \text{proy } a_. / b_j \rVert (s) \lVert b_j \rVert = = \lVert \text{proy } b_j / a_. \rVert (s) \lVert a_. \rVert$
$\mathbf{x}_{..}$	$a_.' b_.$	$\lVert \text{proy } a_. / b_. \rVert (s) \lVert b_. \rVert = \lVert \text{proy } b_. / a_. \rVert (s) \lVert a_. \rVert$
$\mathbf{x}_{i.} - \mathbf{x}_{..}$	$(a_i - a_.)' b_.$	$\lVert \text{proy}(a_i - a_.) / b_. \rVert (s) \lVert b_. \rVert$
$\mathbf{x}_{.j} - \mathbf{x}_{..}$	$a_.' (b_j - b_.)$	$\lVert \text{proy}(b_j - b_.) / a_. \rVert (s) \lVert a_. \rVert$
$\mathbf{x}_{ij} - \mathbf{x}_{i.} - \mathbf{x}_{.j} + \mathbf{x}_{..}$	$(a_i - a_.)' (b_j - b_.)$	$\lVert a_i - a_. \rVert \lVert b_j - b_. \rVert *$ $\cos((a_i - a_.), (b_j - b_.))$

Tabla 1.1.: Estimación de funciones de las observaciones mediante los marcadores del Biplot.

Una de las ventajas de la estimación de funciones de las observaciones sobre la representación Biplot es que tiene interpretaciones sencillas sobre el gráfico. Las figuras 1.1 y 1.2 muestran la interpretación de los efectos fila y columna respectivamente. El efecto correspondiente a una fila se estima proyectando el vector diferencia ($\mathbf{a}_i$-$\mathbf{a}_.$) sobre el vector que une el origen con $\mathbf{b}_.$, de forma que, salvo un factor de escala relacionado con la longitud de $\mathbf{b}_.$, es posible determinar qué filas tienen mayores efectos. Razonamiento análogo se hace para las columnas.

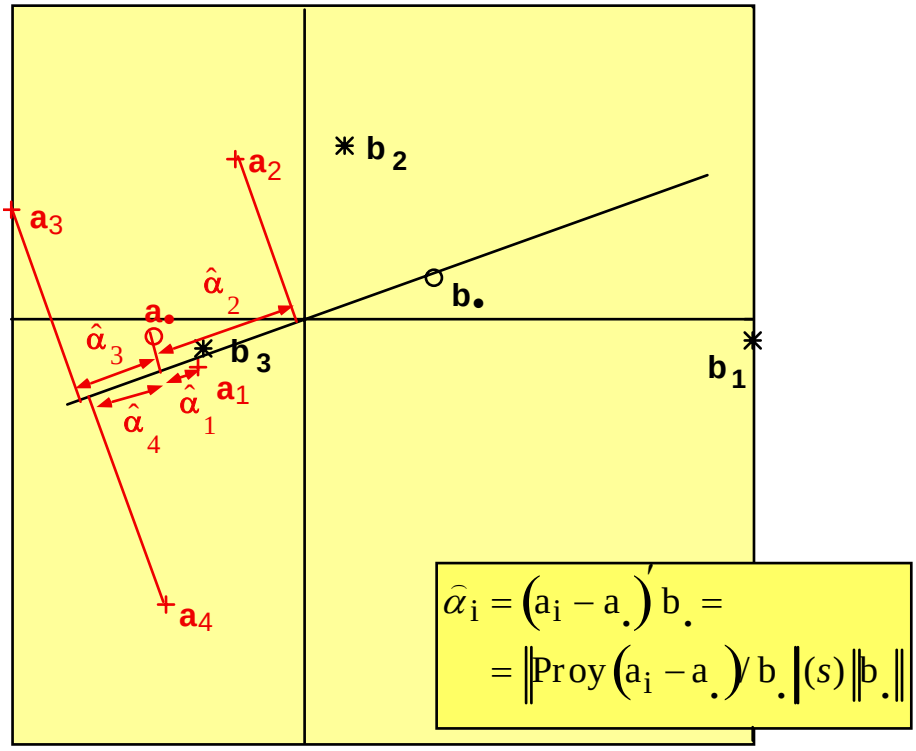


Figura. 1.1.: Estimación de los efectos fila

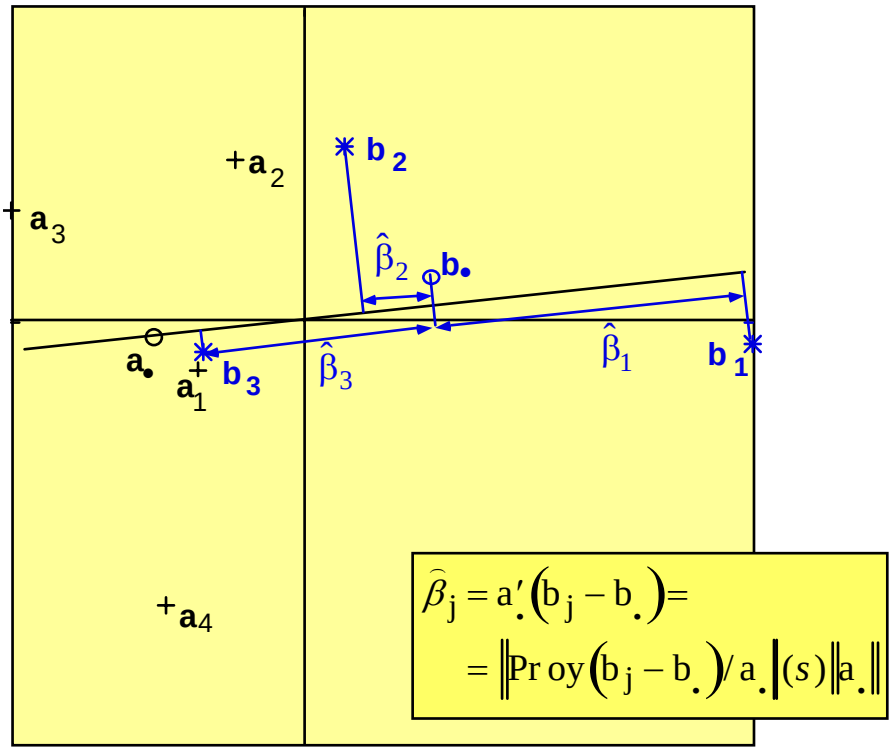


Figura. 1.2.: Estimación de los efectos columna.

1.7.3 APLICACIÓN A DATOS REALES

Se estudia el comportamiento de 10 variedades de patata teniendo en cuenta cuatro indicadores o variables en ellas observadas: Rendimiento (t/ha) y tres componentes del mismo: Peso Promedio del Tubérculo (gr), Número de Tubérculos por Planta y Altura de la Planta (cm). Los datos están enmarcados dentro del programa de mejora del cultivo desarrollado en el Instituto Nacional de Ciencias Agropecuarias de la Habana (Cuba) y corresponden a la campaña 1989-1990.

Para cumplimentar nuestro objetivo aplicamos un JK-Biplot, ya que tenemos especial interés en estudiar el comportamiento de los diferentes genotipos, como posibles variedades a ser introducidas en la producción.

En la tabla 1.2 se muestran los valores medios por genotipo en cada variable:

Genotipo	**Rend. (v1)**	**P.P. (v2)**	**# tub. (v3)**	**altura (v4)**
3-1-85 (a1)	0.40	0.06	6.43	36.00
3-87-85 (a2)	0.46	0.05	9.30	41.00
6-5-85 (a3)	0.41	0.06	7.86	32.33
6-48-85 (a4)	0.35	0.06	5.56	41.33
6-126-85 (a5)	0.40	0.07	5.56	39.66
6-423-85 (a6)	0.52	0.07	7.90	43.66
6-453-85 (a7)	0.47	0.08	6.10	51.33
Spunta (a8)	0.39	0.07	5.20	38.50
Desiree (a9)	0.30	0.05	6.23	34.00
RedPont. (a10)	0.24	0.05	5.30	28.40

Tabla 1.2.: Matriz de datos

Nótese que los genotipos del 1 al 7 se identifican con tres números, el primero se refiere al número del cruce, el segundo corresponde al número del clon y el tercero representa el año en que se obtuvo. Nótese que son todos de 1985. Los genotipos 8, 9 y 10 corresponden a variedades ya establecidas. En nuestro ejemplo constituyen por tanto controles.

El primer paso es determinar el número de ejes a retener (q). Para ello debemos hacer la descomposición en valores y vectores singulares de **X**,

Valores singulares	Inercia acum.(%)
$\lambda_1 = 4.947$	61.249
$\lambda_2 = 3.621$	94.024
$\lambda_3 = 1.482$	99.520
$\lambda_4 = 0.438$	100

Retendremos por tanto los dos primeros ejes, lo que significa que tenemos una bondad de ajuste en el Biplot del 94.024%.

Matrices de marcadores **A** y **B:**

Recordemos que en un JK-Biplot, las matrices de marcadores se obtienen como sigue:

$$\mathbf{A}=\mathbf{UD} \quad \text{y} \quad \mathbf{B}=\mathbf{V}$$

En consecuencia:

$$\mathbf{A}_{(2)} = \begin{bmatrix} -0.230 & -0.080 \\ 0.401 & -1.768 \\ -0.258 & -0.810 \\ -0.244 & 0.525 \\ 0.280 & 0.723 \\ 1.474 & -0.542 \\ 1.875 & 0.870 \\ 0.105 & 0.887 \\ -1.297 & -0.133 \\ -2.108 & 0.327 \end{bmatrix} \qquad \mathbf{B}_{(2)} = \begin{bmatrix} 0.846 & -0.376 \\ 0.718 & 0.682 \\ 0.304 & -1.159 \\ 0.823 & 0.219 \end{bmatrix}$$

Recordemos que cada fila de **A** corresponde a un genotipo y cada fila de **B** se identifica con una variable.

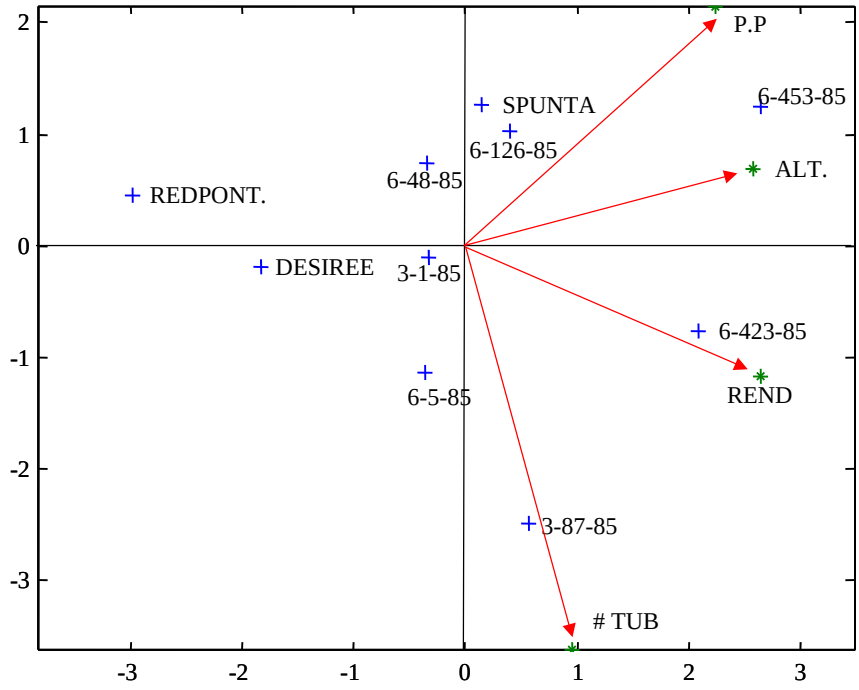


Figura 1.3.: Representación Biplot.

Si queremos obtener a partir de la representación Biplot un valor aproximado de x_{63}, basta con hacer el producto escalar correspondiente:

$$\mathbf{a}_6\mathbf{b}_3=(1.474, -0.542)(0.304,-1.159)=1.076$$
$$1.076\cong 1.047=x_{63}$$

Este resultado nos permite hacer a partir de la representación gráfica, un ordenamiento de los genotipos a partir de su proyección sobre los vectores que representan a las variables (figura 1.4).

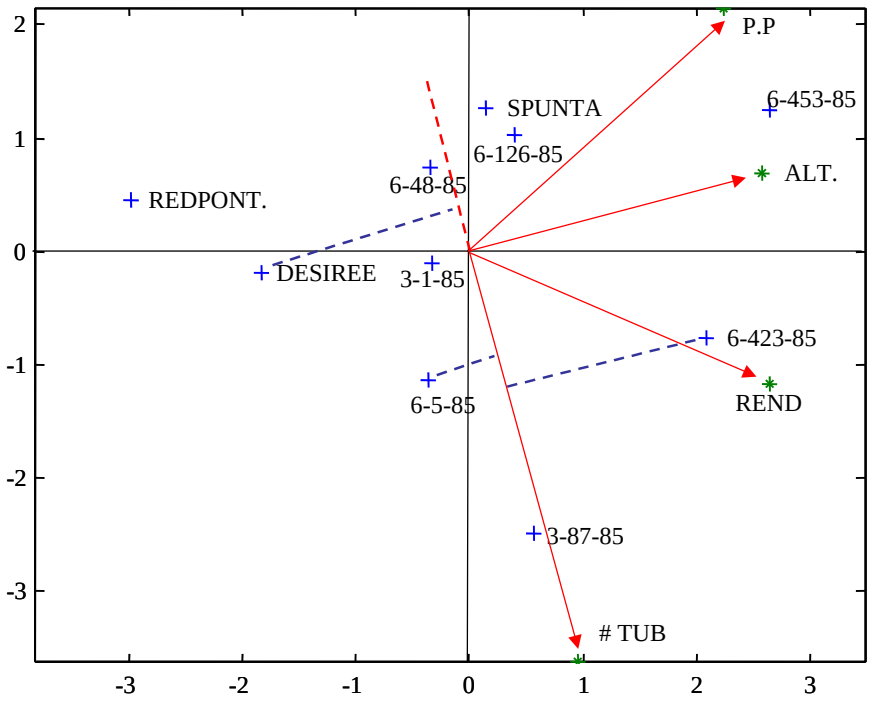


Figura 1.4.: Estimación a partir del Biplot.

Asi, el genotipo 6 (6-423-85) presenta mayor número de tubérculos ($\mathbf{b}_3$) que el genotipo 3 (6-5-85), y esta a su vez presenta mayor número de tubérculos que el genotipo 9 (Desiree). Este ordenamiento puede hacerse con el resto de genotipos y variables.

Nótese que los genotipos de l al 7 presentan valores de rendimiento superior a los controles (Spunta, Desiree y Red Pontiac). Esto lo deducimos del Biplot, ya que los genotipos del 1 al 7 se encuentran más próximos al vector que representa la variable rendimiento (figura 1.3). El primer eje diferencia los controles del resto de genotipos.

Sin embargo, en los programas de mejora no es suficiente con que un genotipo tenga buen rendimiento en condiciones específicas, es necesario además que al variar las condiciones ambientales, siga manteniendo alto rendimiento. Para ello resulta necesario repetir experimentos similares en

diferentes sitios del país e incluso a través de varios años, que permita estudiar la estabilidad de los genotipos que deseamos introducir en la producción.

Se impone entonces realizar un Análisis de Interacción Genotipo Ambiente, tema que trataremos en los próximos capítulos.

C A P Í T U L O II

LOS MÉTODOS BIPLOT COMO HERRAMIENTA DE ANÁLISIS DE INTERACCIÓN DE SEGUNDO ORDEN

2.1 MODELOS CON EFECTO INTERACCIÓN MULTIPLICATIVO (MODELOS AMMI).

2.1.1 INTRODUCCIÓN

Los modelos con término multiplicativo han sido muy utilizados para describir la interacción en tablas de dos vías; tienen la ventaja de permitir una representación Biplot (simultánea) de filas y columnas de la tabla, lo que facilita identificar las combinaciones de niveles causantes de la interacción. Estos modelos a su vez se clasifican en internos y externos (VAN EEUWIJK y KROONENBERG (1998)); internos cuando estiman la interacción haciendo uso solamente de la información contenida en la tabla inicial de datos; externos cuando utilizan además información proveniente de variables externas, ya sean medidas sobre filas, columnas o ambos factores de variación.

Entre los modelos con término multiplicativo que aparecen en la literatura podemos citar, el **Modelo Concurrente de Tukey** (TUKEY (1949)) y los **Modelos de Regresión sobre la media** (YATES y COCHRAN (1938); MANDEL (1961); FINLAY y WILKINSON (1963); EBERHART y RUSSELL (1966)). Se caracterizan por tratar de explicar la interacción a partir de un solo término multiplicativo, lo cual en muchos casos resulta insuficiente.

En muchos casos estos modelos no son adecuados para describir la interacción debido a la complejidad de la misma. En este sentido, en los modelos AMMI (GAUCH (1988)), se mantiene la descomposición en términos multiplicativos de la interacción y ésta no es forzada a tener una

característica específica, se incluyen en el modelo tantos términos como sean necesarios para explicar la variabilidad asociada a la interacción; son clasificados también como modelos internos.

Estos modelos combinan las técnicas de Análisis de Varianza y Análisis de Componentes Principales, (GAUCH y ZOBEL (1989)); tienen como objetivo explicar la interacción asociada a un ANOVA bifactorial, a partir de una representación biplot.

Los modelos AMMI han sido aplicados fundamentalmente en experimentos de campo, más específicamente en el análisis de la interacción Genotipo- Ambiente; con el objetivo de clasificar genotipos en estables e inestables a partir de su interacción con el ambiente (VAN EEUWIJK (1995 a y b); KANG y GAUCH (1996); ROMAGOSA et al (1996)).

Una variedad o genotipo es introducido en la producción cuando además de tener altos rendimientos, presenta un grado de estabilidad aceptable, es decir, reacciona favorablemente a diferentes condiciones ambientales. En las últimas etapas de los programas de mejoramiento genético, en las que las variedades han sido seleccionadas atendiendo fundamentalmente a su rendimiento en condiciones muy específicas, necesitamos conocer cuáles de ellas siguen manteniendo elevados rendimientos al variar las condiciones ambientales. Para ello los genetistas conducen experimentos a lo largo de todo el país, en diferentes épocas e incluso a través de varios años.

Una vez obtenido los datos experimentales, el análisis estadístico más comúnmente utilizado es el Análisis de la Varianza para un arreglo

bifactorial en el que los factores considerados son el genotipo y el ambiente.

Como sabemos, a partir del análisis de la varianza podemos detectar la presencia o no de interacción mediante la F de Snedecor correspondiente a esta fuente de variación. Sin embargo, una vez detectada, no la interpretamos, nos limitamos a seleccionar los genotipos con mayores valores medios, sin tener en cuenta su grado de estabilidad.

Los modelos AMMI al permitir una representación Biplot de filas (genotipos) y columnas (ambientes); dan la posibilidad de estudiar el grado de estabilidad de los genotipos al ser probados en diferentes ambientes.

Existen otras formas de hacer referencia a este tipo de modelos con término multiplicativo en la interacción y efectos principales aditivos (modelos AMMI); así por ejemplo, GABRIEL (1978) y DENIS (1991) lo denominan modelos bilineales; por otra parte, DENIS y GOWER (1992, 1994) lo llaman modelos biaditivos.

En este capítulo daremos una explicación rigurosa de estos modelos, e incorporamos un modelo de tipo externo; la técnica de Regresión Factorial en Rango Reducido (IZENMAN (1975); TER BRAAK (1994)). Será utilizada para explicar la matriz de residuales de interacción de segundo orden a partir de una matriz de variables externas medidas sobre las categorías de uno de los factores de variación considerados.

2.1.2 MODELOS AMMI. FUNDAMENTO TEÓRICO

Como sabemos el modelo lineal al que se ajustan los datos experimentales en un arreglo bifactorial bajo un Diseño Completamente Aleatorizado es de la forma:

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \qquad i=1..I \quad j=1..J \quad k=1..K$$

siendo:

I: nº de niveles del primer factor

J: nº de niveles del segundo factor

K: nº de observaciones por tratamiento o combinaciones de niveles

y_{ijk}: k ésima observación correspondiente a la combinación de niveles ij

α_i, β_j : efectos principales para filas y columnas respectivamente.

$(\alpha\beta)_{ij}$: efecto interacción.

Sabemos además que el estimador mínimo cuadrático correspondiente a la interacción se calcula de la forma:

$$\widehat{(\alpha\beta)}_{ij} = y_{ij.} - y_{i..} - y_{.j.} + y_{...}$$

En la literatura se presentan modelos en los que se trata de explicar la interacción a partir de términos multiplicativos; en dichos términos aparecen como factores los efectos principales asociados a filas, columnas o ambas fuentes de variación. Sin embargo, en muchos casos resulta imposible modelar la interacción a partir de estos términos, por la complejidad de la misma.

Estos modelos han sido desarrollados para el caso de una observación por celda, es decir no tenemos repeticiones que permitan controlar la variabilidad dentro de la celda, y por tanto no es posible estimar el error experimental.

TUKEY (1949) fue el primer investigador que propuso un modelo para el análisis de la interacción en experimentos de dos vías, para diseños no replicados.

El modelo propuesto es de la forma:

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

donde $\alpha_i\beta_j$ es el producto de los efectos principales y λ un coeficiente de regresión. Contrastar la hipótesis sobre $\lambda = 0$ será equivalente a un test de hipótesis para contrastar que el producto de los efectos no contribuye a la predicción de y_{ij} .

Luego, su finalidad es distinguir entre el modelo aditivo (solo incluye los efectos principales) y el modelo que contempla un término de interacción.

Para detalles sobre contraste y software estadístico utilizado para ajustar este modelo, consultar MILLIKEN y JOHNSON (1989).

Otro tipo de modelos con término multiplicativo muy utilizado para describir la interacción en tablas de dos vías, son los Modelos de Regresión sobre la media (YATES y COCHRAN (1938); MANDEL (1961)), los cuales se definen de la siguiente forma:

$$y_{ij} = \mu + \alpha_i + \beta_j + \lambda\gamma_i\beta_j \qquad \hat{(\alpha\beta)}_{ij} = \lambda\gamma_i\hat{\beta}_j \qquad \text{Regresión para filas}$$

$$y_{ij} = \mu + \alpha_i + \beta_j + \lambda\alpha_i\gamma_j \qquad \hat{(\alpha\beta)}_{ij} = \lambda\hat{\alpha}_i\gamma_j \qquad \text{Regresión para columnas}$$

Como su nombre lo indica, se estiman los valores de γ_i a partir de las regresiones de los efectos principales en la variable dependiente.

Consideramos oportuno señalar que en el contexto del Análisis de Interacción Genotipo-Ambiente, los modelos de **Mandel** son también conocidos como Modelos de **Finlay y Wilkinson** (FINLAY y WILKINSON (1963)).

Nótese que en los modelos anteriores la interacción es forzada a tener características muy específicas, por tanto puede darse el caso de que la descomposición en términos multiplicativos no sea suficiente para explicar la variabilidad asociada a la interacción.

En los modelos AMMI - GOLLOB (1968) Efecto Interacción Multiplitativo y Efectos Principales Aditivos-, se combinan las técnicas de Análisis de Varianza y Análisis de Componentes Principales. Se introducen al modelo tantos términos multiplicativos como sean necesarios para explicar la variabilidad de la interacción.

GABRIEL (1978) muestra la conexión entre el ajuste mínimo cuadrático de un modelo multiplicativo y la descomposición en valores singulares de una matriz. (CORNELIUS et al (1996)).

El método consiste en hacer la descomposición en valores singulares de la matriz **Z** de orden IxJ formada por los estimadores de las interacciones en el modelo anterior.

$$\mathbf{Z} = (z_{ij}) = (\alpha\hat{\beta})_{ij} = y_{ij.} - y_{i..} - y_{.j.} + y_{...}$$

Al realizar la descomposición en valores singulares de **Z** nos queda:

$$z_{ij} = (\hat{\alpha\beta})_{ij} = \sum_{m=1}^{M} \lambda_m u_{mi} v_{mj}$$

siendo M el rango de **Z.**

Llamaremos modelo AMMI de orden M a la expresión:

$$AMMI_M: \qquad E(y_{ijk}) = \mu + \alpha_i + \beta_j + \sum_{m=1}^{M} \lambda_m u_{mi} v_{mj}$$

donde:

λ_m: corresponde al valor singular de orden m de **Z'Z**

u_{mi}: coordenada i-ésima del vector singular de **ZZ´** asociado a λ_m

v_{mj}: coordenada j-ésima del vector singular de **Z'Z** asociado a λ_m

De esta forma podemos representar las filas (genotipos) y columnas (ambientes) en un subespacio de dimensión M en el que las proximidades entre genotipos van a indicar que interactúan de manera similar con el ambiente.

Así, los genotipos que se ubican cerca del origen de coordenadas serán los más estables, es decir, los que interactúan poco con el ambiente. Por otra parte aquellos genotipos que se alejan del origen serán los más inestables; tendrán altos rendimientos solamente en aquellos ambientes próximos a ellos en la representación.

Nuevamente, al estar en presencia de una técnica en la que se realiza una descomposición en valores singulares, surge la problemática de cuántos ejes elegir.

Si consideramos un solo factor el modelo será:

$$AMMI_1: \quad E(y_{ijk}) = \mu + \alpha_i + \beta_j + \lambda_1 u_{1i} v_{1j}$$

con dos factores:

$$AMMI_2: \quad E(y_{ijk}) = \mu + \alpha_i + \beta_j + \lambda_1 u_{1i} v_{1j} + \lambda_2 u_{2i} v_{2j}$$

y así sucesivamente. (GAUCH y ZOBEL (1989) ; MILLIKEN y JOHNSON (1989)).

¿ Qué modelo es el más adecuado?. ¿Cuántos términos multiplicativos deben ser incluidos en el modelo?.

Para dar respuesta a estas interrogantes, se realiza la descomposición de la suma de cuadrados de la interacción asociada al análisis de varianza para un modelo bifactorial. (VAN EEUWIJK (1995 a)), en la forma siguiente:

F.V	G.L	S.C
INT GxA	(I-1)(J-1)	$K\sum\lambda_i^2$
$AMMI_1$	(I-1)+(J-1)-1	$K\lambda_1^2$
$AMMI_2$	(I-1)+(J-1)-3	$K\lambda_2^2$
	……..	…….
$AMMI_M$	(I-1)+(J-1)-2L-1	$K\lambda_M^2$

Tabla 2.1.: Descomposición de la suma de cuadrados de la interacción.

Para conocer si un modelo con G términos multiplicativos es válido, se realiza el ANOVA, utilizando como variabilidad total la correspondiente a la interacción, se determina el residual de la interacción asociada a la parte de la variabilidad total (interacción) que no es explicada por los G términos multiplicativos, y se calcula la F de Snedecor correspondiente (tabla 2.2):

F.V	G.L	S.C	C.M	F
INT GxA	(I-1)(J-1)	$K\sum\lambda_i^2$		
$AMMI_1$	(I-1)+(J-1)-1	$K\lambda_1^2$		
$AMMI_2$	(I-1)+(J-1)-3	$K\lambda_2^2$		
	……..	…….		
$AMMI_G$	(I-1)+(J-1)-2L-1	$K\lambda_G^2$	S.C/G.L	$C.M.AMMI_G/C.M.Res$
residual	Por diferencia	Por Dif.	C.M.Res	

Tabla 2.2: ANOVA para la selección del número de términos multiplicativos.

Si la F de Snedecor es significativa, se incluye el término multiplicativo asociado a λ_G, y se pasa a analizar el término G+1. En caso contrario se elimina y no se continúa el análisis debido a la relación de orden existente entre los valores singulares.

MILLIKEN y JOHNSON (1989) dan un procedimiento para seleccionar el número de términos multiplicativos óptimo; para el caso de experimentos

no replicados. Difiere del método anterior por el hecho de que en este caso, al no tener réplicas, no es posible estimar σ^2. Se ofrecen por tanto tablas con los valores críticos necesarios para contrastar las hipótesis ($\lambda_i=0$).

Una vez definido el número de ejes a retener tendremos los marcadores asociados a los genotipos y los marcadores asociados a los ambientes, representados en un subespacio de dimensión igual a la cantidad de ejes retenidos.

$$\text{Bondad de ajuste:} \quad \frac{\sum_{q=1}^{Q} \lambda_q^2}{\sum_{m=1}^{M} \lambda_m^2} * 100\%$$

siendo Q el número de términos multiplicativos incluidos en el modelo.

2.1.3 TABLAS INCOMPLETAS E INCUMPLIMIENTO DE HIPÓTESIS DE BASE DEL MODELO

En los modelos AMMI el tratamiento para datos faltantes es similar al utilizado en el Análisis de Varianza, recordemos que en estos modelos, estimamos la matriz de residuales de interacción a partir del Análisis de Varianza Bifactorial y una vez estimada, se realiza la descomposición en valores y vectores singulares (Biplot) de la misma. Por tanto todas las técnicas conocidas para estimar los parámetros de un modelo lineal en tablas incompletas, son válidos en este contexto.

En el contexto del Análisis de Interacción Genotipo- Ambiente puede darse el caso de que algunas variedades no hayan sido probadas en determinadas localidades (años). Igualmente podemos estar en presencia de diseños

desbalanceados (observaciones perdidas). En tal caso, la estimación de los parámetros del modelo se realiza a partir de los mínimos cuadrados alternados ((GABRIEL y ZAMIR (1979); DENIS (1991); VAN EEUWIJK (1995c)).

Otro tratamiento del problema puede ser mediante el uso de modelos mixtos (SEARLE (1971); es decir considerando efectos fijos y efectos aleatorios. En tal caso ajustamos el modelo mixto a la tabla incompleta y seguidamente calculamos en la tabla completa el mejor estimador insesgado. ((VAN EEUWIJK (1995a). Los parámetros se estiman a partir del método de máxima verosimilitud de los residuos ((PATTERSON y THOMPSON (1971); SEARLE et al (1992)).

Otra situación que puede presentarse se refiere a la violación de las hipótesis de base del Modelo Lineal, recordemos que en un Análisis de Varianza asumimos que los errores siguen una distribución normal, con varianza constante entre tratamientos y aditividad de efectos. En la práctica, cuando la variable dependiente es de naturaleza continua, nos protegemos de la violación de los supuestos del modelo; sin embargo en ocasiones trabajamos con otro tipo de variables, por ejemplo, incidencia de enfermedades o variables de conteo, de las cuales sabemos que no siguen una distribución normal.

Para solucionar este problema, con frecuencia realizamos un cambio de escala o transformación a los datos; sin embargo en ocasiones esto no resuelve. Así por ejemplo, MCCULLAGH y NELDER (1991) plantean que para datos discretos donde el error sigue una distribución de Poisson, los efectos sistemáticos son multiplicativos. En tal caso, la transformación $Y^{1/2}$ da varianza constante, la transformación $Y^{2/3}$ nos da simetría o normalidad

y la transformación log(Y) produce aditividad en los efectos sistemáticos. Es por ello que una simple transformación no resuelve simultáneamente todos los supuestos del modelo.

Con los Modelos Lineales Generalizados podemos resolver este problema, ya que en los mismos las hipótesis de normalidad y homogeneidad de varianzas no son supuestos; solamente es necesario conocer la relación existente entre media y varianza en los datos (MCCULLAGH y NELDER (1991)).

Aplicando el Modelo Lineal Generalizado en el contexto de los AMMI, (VAN EEUWIJK (1995c)) se refiere a los modelos GAMMI (AMMI Generalizado). En tal sentido plantea que un AMMI no es más que un GAMMI con función link identidad y varianza constante. En su trabajo da un ejemplo para modelos logit.

2.1.4 APLICACIÓN A DATOS REALES

Usaremos unos datos en el que se evalúa el número de tubérculos por planta de 10 variedades de patata (dadas en el capítulo anterior) durante tres campañas (1989-1990, 1990-1991 1991-1992). Se utiliza un diseño Completamente Aleatorizado bajo un arreglo bifactorial, con tres observaciones por combinaciones de niveles de cada factor. Se presentan los valores medios de 10 plantas.

Este experimento forma parte de un programa de mejoramiento desarrollado en Cuba; precisamente los genotipos del 1-7 han sido obtenidos dentro del programa de mejora, se trata de estudiar su estabilidad al ser probados durante tres períodos consecutivos.

En el experimento se utilizan tres variedades controles ya establecidas: (Spunta, Desireé y Red-Pontiac). Precisamente varios de los genotipos que están siendo probados, presentan como progenitores alguna de estas variedades controles.

Como puede verse en la tabla 2.3, los genotipos se identifican mediante tres números, el primero se refiere al número del cruce, el segundo corresponde al número del clon y el tercero representa el año en que se obtuvo. Nótese que son todos de 1985.

Gen/año	**89-90**	**90-91**	**91-92**
	6.7	9.0	8.6
3-1-85 (g1)	6.0	8.6	10
	6.6	7.6	7.5
	9.8	5.9	7.0
3-87-85 (g2)	8.2	4.8	6.3
	9.9	5.8	11.4
	11.5	7.2	9.6
6-5-85 (g3)	7.6	6.4	11.1
	4.5	8.7	13
	5.6	6.1	10.2
6-48-85 (g4)	4.5	5.7	8.1
	6.6	4.4	11.3
	6.5	9.4	12.2
6-126-85 (g5)	5.8	7.3	12
	4.4	8.0	12.3
	7.1	8.5	9.8
6-423-85 (g6)	7.7	7.6	8.4
	8.9	8.3	10.6
	4.3	9.3	12.5
6-453-85 (g7)	6.5	5.8	11.3
	7.5	9.4	12.9
	4.8	7.2	8.6
Spunta (g8)	5.6	7.9	9.0
	5.2	6.0	7.3
	7.9	7.6	13.4
Desiree (g9)	6.8	5.3	12.2
	4.0	7.7	10.3
	5.0	4.3	8.0
RedPont. (g10)	4.6	5.9	6.5
	6.3	6.5	4.7

Tabla 2.3.: Matriz de datos.

El primer paso es efectuar el contraste que nos permita detectar la presencia de interacción de segundo orden:

F.V	G.L	S.C	C.M	F
genotipo	9	86.28	9.59	4.84**
ambiente	2	191.43	95.72	48.28**
interacción	18	110.64	6.15	3.10**
error	60	118.94	1.98	
total	89	507.29		

$p \leq 0.05$

El Análisis de la Varianza efectuado pone de manifiesto que existe una interacción Genotipo-Ambiente altamente significativa; se justifica por tanto el uso de los modelos AMMI.

El segundo paso es analizar cuál es el modelo más adecuado para describir la interacción. Para ello calculamos la matriz de valores residuales o interacciones calculados a partir de los estimadores mínimo cuadráticos:

$$\mathbf{Z} = (\hat{\alpha\beta})_{ij} = \begin{bmatrix} -0.12 & 1.32 & -1.18 \\ 2.92 & -1.41 & -1.48 \\ 0.31 & -0.65 & 0.35 \\ -0.09 & -0.77 & 0.88 \\ -1.80 & 0.34 & 1.47 \\ 0.65 & 0.35 & -0.98 \\ -1.44 & 0.09 & 1.36 \\ -0.35 & 0.95 & -0.58 \\ -0.83 & -0.73 & 1.57 \\ 0.84 & 0.57 & -1.39 \end{bmatrix}$$

Antes de ajustar el modelo AMMI, ajustaremos uno de los modelos internos mencionados anteriormente, el cual ha sido muy aplicado en el análisis de la interacción Genotipo-Ambiente, nos referimos al modelo de Finlay y Wilkinson. Se hará posteriormente un estudio comparativo.

Modelo de Finlay y Wilkinson:

$$y_{ijk} = \mu + \alpha_i + \gamma_i \beta_j + e_{ijk}$$

Los γ_i se obtienen a partir de regresiones de los valores de $y_{ij.}$ en $\hat{\beta}_j$.

$$\gamma_i = \sum_j (y_{ij.} - y_{.j.}) / \sum_j \hat{\beta}_j^2$$

Genotipo	Constante	Coef. (γ_i)	R^2
g1	7.84	0.49	0.69 n.s
g2	7.67	0.13	0.01 n.s
g3	8.84	1.17	0.95 n.s
g4	6.94	1.44	0.97 n.s
g5	8.65	1.84	0.92 n.s
g6	8.54	0.53	1.00 **
g7	8.83	1.76	0.95 n.s
g8	6.84	0.79	0.76 n.s
g9	8.35	1.82	0.99 **
g10	5.75	0.33	0.99 **

Tabla 2.4.: Modelos de regresión ajustados.

Para clasificar los genotipos en estables e inestables representamos en un eje de coordenadas los respectivos γ_i y los valores de rendimiento relativo $(y_{i..} / y_{...})$.

Genotipo	$y_{i..}$	$y_{i..} / y_{...}$	γ_i
g1	7.84	1.00	0.49
g2	7.67	0.97	0.13
g3	8.84	1.12	1.17
g4	6.94	0.88	1.44
g5	8.65	1.10	1.84
g6	8.54	1.09	0.53
g7	8.83	1.12	1.76
g8	6.84	0.87	0.79
g9	8.35	1.06	1.82
g10	5.75	0.73	0.33

Tabla 2.5.: Coordenadas para los genotipos.

Hacemos algo similar para los ambientes:

Ambiente	$y_{.j.}$	$y_{.j.} / y_{...}$	$\hat{\beta}_j$
a1	6.54	0.83	-1.19
a2	7.07	0.90	-0.76
a3	9.87	1.26	2.00

Tabla 2.6.: Coordenadas para los ambientes.

A continuación mostramos una representación conjunta de genotipos y años. Para los genotipos las coordenadas serán ($y_{i.}/y_{..}$, γ_i) y para los años las coordenadas serán ($y_{.j}/y_{..}$, $\hat{\beta}_j$).

De manera que el eje 1 ubicará los puntos teniendo en cuenta su valor relativo en la variable analizada, en este caso (# de tubérculos/planta); mientras que el eje 2 será una medida de la estabilidad.

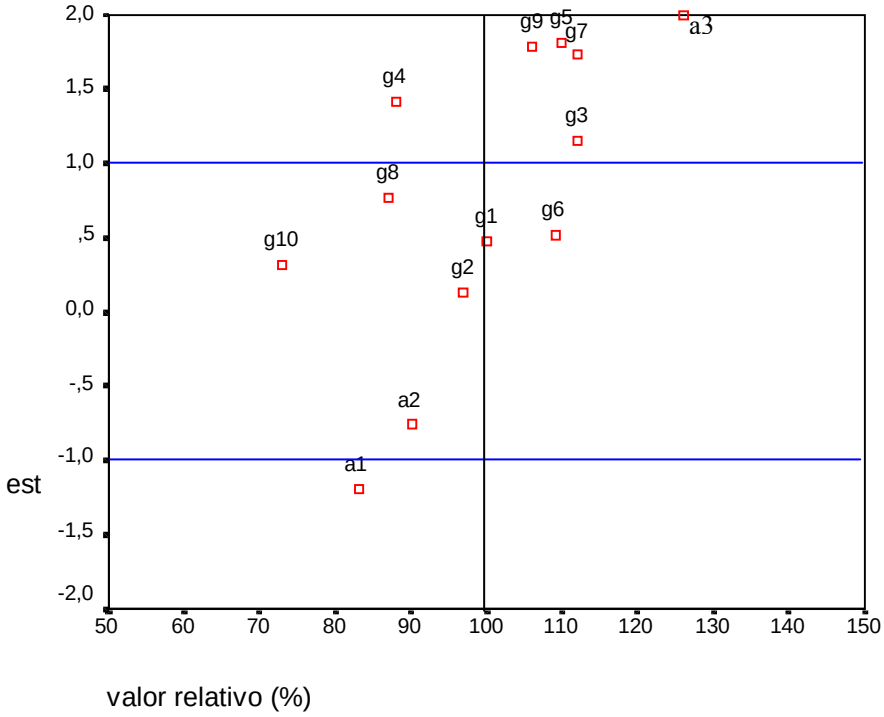


Figura 2.1.: Representación simultánea de genotipos y ambientes.

En la figura 2.1, los genotipos que se ubican dentro de la banda azul, presentan un comportamiento estable (los coeficientes de regresión oscilan entre –1 y 1), las de la parte derecha de la banda serán estables con altos valores de número de tubérculos por planta (6-423-85), mientras que las de la parte izquierda de la banda, tienen un comportamiento estable pero con valores bajos valores (3-1-85, 3-87-85, Spunta y Red Pontiac). En ambos cuadrantes superiores se encuentran las variedades que presentan altos valores de # de tubérculos en ambientes buenos, y muy malos en ambientes malos, en la parte derecha las de mayor número de tubérculos (6-5-85, 6-126-85, 6-453-85 y Desiree). Finalmente en ambos cuadrantes inferiores se encuentran las variedades que tienen buen comportamiento en ambientes desfavorables y muy malo en ambientes favorables, en nuestro caso ninguna.

Nótese que en la representación gráfica se han representado además los ambientes, ello permite realizar una interpretación en términos de proyección.

No obstante queremos destacar la poca validez de estas conclusiones debido a los valores de R^2 tan bajos que se han obtenido. Ello indica que en nuestros datos, la interacción tiene una estructura mucho más compleja, imposible de modelar con un solo término multiplicativo.

Ajustemos ahora un modelo AMMI a nuestros datos y comparemos con lo obtenido por el modelo anterior.

Consideramos nuevamente la matriz **Z** de residuales de interacción del modelo

Al realizar la descomposición en valores singulares de **Z** obtenemos los siguientes valores singulares distintos de cero:

$$\lambda_1 = 5.178 \qquad \lambda_2 = 3.164$$

Por tanto, la descomposición de la suma de cuadrados de la interacción será:

F.V	G.L	S.C
Interacción	18	110.64
$AMMI_1$	10	80,44
$AMMI_2$	8	30.20

Tabla 2.7.: Descomposición de la suma de cuadrados de la interacción.

Calculemos las contribuciones para identificar los genotipos y ambientes bien representados en el Biplot: (Ver tablas 2.8 y 2.9 respectivamente).

	Factor 1	**Factor 2**
g1	119	**891**
g2	**830**	170
g3	2	**998**
g4	265	**735**
g5	**991**	9
g6	**823**	177
g7	**1000**	0
g8	3	**997**
g9	**708**	292
g10	**773**	227

Tabla 2.8.: Contribuciones relativas del factor al elemento filas.

	Factor 1	**Factor 2**
a1	**924**	76
a2	18	**982**
a3	**847**	153

Tabla 2.9.: Contribuciones relativas del factor al elemento columnas.

De las tablas 2.8 y 2.9, concluimos diciendo con una representación plana prácticamente todos los genotipos y ambientes quedan bien representados. El eje 1 está determinado por los genotipos 3-87-85, 6-126-85, 6-423-85, 6-453-85, Desiree y Red Pontiac, en su comportamiento en las campañas

1989-1990 y 1991-1992. Ello es debido a que son los genotipos y ambientes con mayores contribuciones relativas en el primer eje factorial.

El eje 2 por su parte está relacionado con las variedades 1,3,4 y 8 en su comportamiento en la campaña 1990-1991.

Seguidamente mostramos las matrices **A** y **B** de marcadores asociadas a los genotipos y ambientes, respectivamente; las cuales nos permiten posicionar los genotipos y ambientes en el plano (ver figura 2.2).

Matrices de marcadores (JK-Biplot):

$$\mathbf{A} = \begin{bmatrix} 0.398 & 1.086 \\ 2.117 & -0.959 \\ 0.025 & -0.521 \\ -0.393 & -0.655 \\ -1.525 & 0.141 \\ 0.726 & 0.336 \\ -1.292 & -0.039 \\ 0.042 & 0.759 \\ -1.053 & -0.676 \\ 0.987 & 0.535 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1.134 & -0.532 \\ -0.103 & 1.247 \\ -1.025 & -0.714 \end{bmatrix}$$

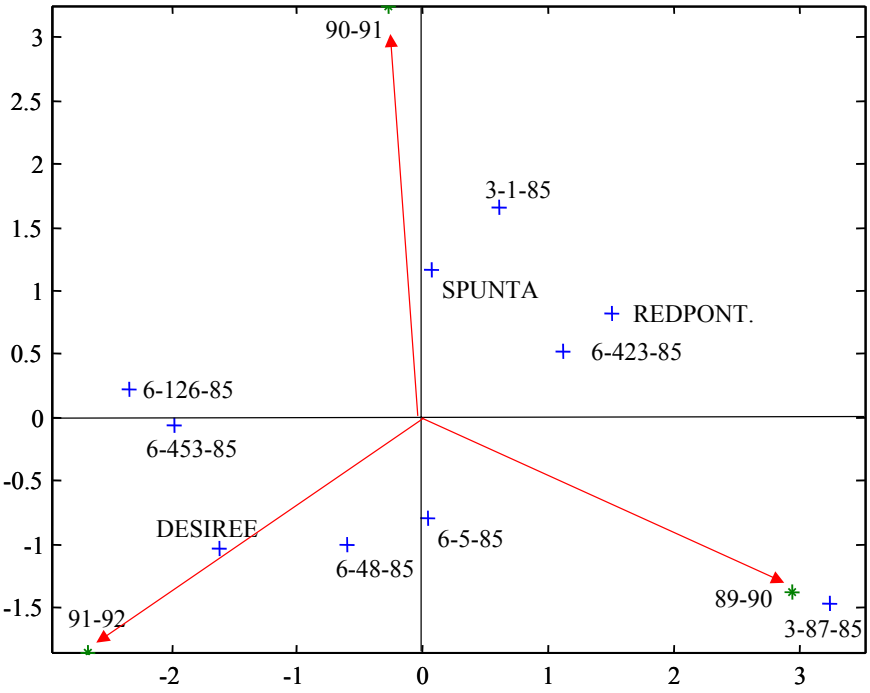


Figura 2.2.: Representación Biplot.

Como estamos realizando un Biplot a la matriz de residuales de interacción del modelo, los genotipos próximos al origen de coordenadas serán los que tendrán un comportamiento más estable al ser probados en los distintos ambientes, de igual forma aquellos con posiciones extremas, serán los responsables de la interacción altamente significativa detectada en los datos

Como se aprecia en la figura 2.2, los genotipos más inestables son: 3-1-85, 3-87-85, 6-126-85 y 6-453-85. El primer eje contrapone los genotipos 6-126-85 y 6-453-85 del genotipo 3-87-85; los dos primeros se caracterizan por presentar mayor número de tubérculos en las condiciones de la campaña 91-92, de igual forma estos dos genotipos interactúan de manera negativa en las condiciones de la campaña 89-90. El genotipo 3-87-85 se comporta de forma totalmente contraria a los genotipos 6-126-85 y 6-453-85; es decir, interactúa de manera positiva en las condiciones de la

campaña 89-90 y de forma negativa en las condiciones de la campaña 91-92.

El eje 2 destaca la interacción positiva de los genotipos 3-1-85 y Spunta fundamentalmente, en la campaña 1990-1991.

Recordemos además que proximidades entre genotipos en el gráfico Biplot indica que interactúan de manera similar con el ambiente, en ningún caso es indicativo de que presentan similar comportamiento en la variable dependiente analizada.

Podemos concluir diciendo que con la aplicación de los modelos AMMI, se introducen las representaciones Biplot y con ello se logra una clasificación de los genotipos mucho más completa. En este caso se ha logrado con una representación plana explicar el 100% de la **variabilidad de la interacción;** siendo por tanto los resultados mucho más fiables que los obtenidos a partir de la aplicación del modelo de Finlay y Wilkinson.

Podemos destacar que con la aplicación del modelo de Finlay y Wilkinson a nuestros datos se llegan incluso a obtener resultados completamente falsos como lo es la estabilidad del genotipo 2. Ello por supuesto es debido al ajuste tan malo que se produjo en la recta de regresión correspondiente a este genotipo (ver tabla 2.4).

Los modelos AMMI al permitir incorporar al modelo tantos términos como sean necesarios, evita llegar a conclusiones erróneas sobre la estabilidad de los genotipos.

2.2 REGRESIÓN EN RANGO REDUCIDO

2.2.1 INTRODUCCIÓN

Con este método al igual que en los modelos AMMI, se permite hacer una representación Biplot de filas y columnas de la matriz, con la diferencia de que ahora se puede incorporar información de variables externas, las cuales pueden ser medidas bien sobre las filas o bien sobre las columnas. En otras palabras, trataremos de explicar la matriz **Z** asociada a las interacciones a través de una matriz de variables externas.

En este caso, a diferencia del Análisis de Regresión Múltiple, en lugar de tener una variable dependiente, se trata de explicar la información contenida en una matriz. El objetivo es ajustar un modelo donde, tanto la parte a explicar como la parte explicativa, son matrices. (Modelo de Regresión Lineal Multivariante (MARDIA et al (1979)).

Cuando estimamos los parámetros del modelo a partir de técnicas de regresión múltiple y técnicas de reducción de dimensionalidad, estamos en presencia de un Modelo de Regresión Factorial en Rango Reducido (IZENMAN (1975)), este método es también conocido con el nombre de Análisis de Componentes Principales para variables instrumentales (RAO (1964); ROBERTS y ESCOUFIER (1976)); mientras que otros autores la identifican como Análisis de Redundancia (VAN DEN WOLLENBERG (1977); ISRAELS (1984) ; VAN DER BURG y DE LEEUW (1990)).

Este tipo de técnica ha sido utilizada por CÁRDENAS (2000) en el contexto de los Modelos Lineales Generalizados, en su caso trabaja con variables con distribuciones de la familia exponencial.

Incorporar variables externas es muy importante en la interpretación de la interacción ya que podemos identificar características propias de filas o columnas, causantes de la misma.

2.2.2 FUNDAMENTO TEÓRICO

Recordemos que **Z** es una matriz de orden IxJ en la que aparece en la posición ij el valor $(\hat{\alpha\beta})_{ij}$ asociado al estimador de la interacción en el modelo lineal general.

Supongamos que tenemos otra matriz **X** de información externa que puede ser de orden IxH o JxH, siendo H el número de variables externas consideradas para explicar la interacción. Estaremos en la primera situación cuando las variables externas son medidas sobre los genotipos, y nos encontraremos en el segundo caso cuando son variables externas ambientales.

Para ejemplificar, supongamos que estamos en el segundo caso, es decir la matriz **X**, es de orden JxH, o sea, son variables externas relacionadas con el ambiente.
Tenemos por tanto dos matrices:

$$\mathbf{Z} = \begin{pmatrix} \cdot & & \\ \cdot & & \\ \cdot & (\alpha\hat{\beta})_{ij} & \cdot \quad \cdot \\ \cdot & & \\ \cdot & & \end{pmatrix}_{IxJ} \qquad \mathbf{X} = \begin{pmatrix} \cdot & & \\ \cdot & x_{hj} & \cdot \quad \cdot \\ \cdot & & \end{pmatrix}_{JxH}$$

x_{hj} representa el valor que toma para el ambiente j la variable h.

¿Cómo relacionar ambas matrices?.¿ Cómo explicar **Z** a partir de la información suministrada por **X**?.

Tenemos que ajustar por tanto un modelo de regresión lineal multivariante (MARDIA et al (1979)), el cual difiere del ya conocido modelo de regresión lineal múltiple por el hecho de que ahora tanto la parte a explicar como la parte explicativa son matrices.

El problema se traduce en trabajar con el siguiente modelo :

$$\mathbf{Z'= XM + E} \qquad \mathbf{(1)}$$

M representa la matriz de coeficientes del modelo y es de orden HxI

Factorizando **M** como **M=CA'** y sustituyendo en (1), nos queda:

$$\mathbf{Z'= XCA' + E}$$

El cual se conoce como Modelo de Regresión en Rango reducido, (IZENMAN (1975) ; DAVIES y TSO (1982)).

Se puede escribir también como:

$$\mathbf{Z'= BA' + E} \quad \text{donde} \quad \mathbf{B= XC}$$

El cual es un modelo factorial en el que los factores son combinaciones lineales de las variables regresoras.

Luego, el problema se traduce en realizar un A.C.P (Biplot) a la matriz $\hat{\mathbf{Z}}$ formada por los valores ajustados a partir de modelos de regresión lineal múltiple que se realizan para cada una de las columnas de **Z'** (TER BRAAK (1994)), es decir se van a ajustar tantos modelos como filas (genotipos) se estén considerando.

$$\hat{\mathbf{Z}}' = \mathbf{X(X'X)^{-1}X'Z'}$$

Para el genotipo 1 la variable dependiente será la primera columna de **Z'**, y las variables independientes serán las columnas de la matriz **X**; para el genotipo 2 la variable dependiente será la segunda columna de **Z'**, e igualmente se toman las columnas de **X** como variables independientes, y así sucesivamente hasta llegar al genotipo I.

Una vez obtenidos los modelos de regresión se calculan los valores de $\hat{\mathbf{Z}}$ y realizamos el ACP (Biplot) correspondiente. Se van a obtener nuevas variables o factores que cumplen con la condición de ser combinaciones lineales de las variables originales

Luego, si R es el rango de $\hat{\mathbf{Z}}$, al realizar su descomposición en valores singulares, vamos a tener las siguientes dimensiones en las matrices estimadas:

$$\mathbf{C} = \begin{pmatrix} \cdot & & \\ \cdot & & \\ \cdot & c_{hr} & \cdot \quad \cdot \\ \cdot & & \\ \cdot & & \end{pmatrix}_{HxR}$$

$$\mathbf{A'} = \begin{pmatrix} \cdot & & \\ \cdot & & \\ \cdot & a_{ri} & \cdot \\ \cdot & & \end{pmatrix}_{RxI} \qquad \mathbf{B} = \begin{pmatrix} \cdot & & \\ \cdot & & \\ \cdot & b_{jr} & \cdot \\ \cdot & & \\ \cdot & & \end{pmatrix}_{JxR}$$

En la terminología del biplot en b_{jr} van a estar los marcadores para las columnas (ambientes), en a_{ri} van a estar los marcadores filas (genotipos), mientras que en c_{hr} se encuentran los marcadores para las variables ambientales, o lo que es lo mismo,

$$\mathbf{Z'} \approx \hat{\mathbf{Z}} = \sum_{r=1}^{R} \hat{\lambda}_r a'_{ri} \sum_{h=1}^{H} c_{hr} x_{hj} \qquad \text{ya que} \qquad b_{jr} = \sum_{h=1}^{H} c_{hr} x_{hj} \qquad (2)$$

donde $\hat{\lambda}_r$ representa el valor singular de orden r de $\hat{\mathbf{Z}}$.

Utilizamos el símbolo de aproximación debido a que $\mathbf{z}_{ij} = \widehat{(\boldsymbol{\alpha\beta})}_{ij}$ es estimada por $\hat{\mathbf{z}}_{ij}$; no estamos trabajando directamente con los valores de $\mathbf{Z}$.

En la practica, al realizar la descomposición en valores singulares de $\hat{\mathbf{Z}}$, vamos a obtener directamente los valores de a_{ri} y b_{jr}, no ocurriendo así con los valores de $c_{hr.}$.

Estamos en presencia de un Biplot con información externa (BLÁZQUEZ (1998)); nótese que los marcadores para las columnas (b_{jr}), son combinaciones lineales de las variables ambientales.

Para obtener los marcadores para las variables ambientales, debemos hacer uso de la ecuación (2), y por tanto ajustar las rectas de regresión correspondiente, sabiendo que la matriz **B** contiene los vectores propios de $\hat{\mathbf{Z}}\hat{\mathbf{Z}}'$

Aplicando estos resultados; el modelo lineal para un arreglo bifactorial nos queda:

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \sum_{r=1}^{R} \hat{\lambda}_r a_{ri} \sum_{h=1}^{H} c_{hr} x_{hj}$$

Finalmente hemos llegado a un modelo que permite realizar una representación biplot de tres marcadores: genotipo, ambiente y variables externas medidas sobre los ambientes. En este caso, al igual que en los modelos AMMI, podemos seleccionar el número de factores necesarios para la representación a partir de la descomposición de la suma de

cuadrados para la interacción; lo único que varía son los grados de libertad para los distintos modelos, los cuales serán I+H- 2r. (VAN EEUWIJK (1995 a)).

Bondad de ajuste.

GABRIEL (1978) plantea que en los modelos de rango reducido se realiza un doble ajuste, primero cuando estimamos los valores de **Z** a partir de las regresiones múltiples, este 1er ajuste lo denomina lineal; y luego realizamos la descomposición en valores singulares de la matriz $\hat{\mathbf{Z}}'$, el cual denomina un ajuste bilineal, es por ello que denomina esta técnica en su trabajo con el nombre de aproximación mínimo cuadrática de matrices por modelos aditivos y multiplicativos:

En el 1er ajuste sabemos que partimos de la suma de cuadrados asociada a la interacción:

$$S.C.Int = K\sum_i\sum_j (z_{ij})^2 = K\sum_{m=1}^{M}\lambda_m^2 = \text{Inercia Total}$$

¿Qué parte de la suma de cuadrados asociada a $\mathbf{Z}$ (interacción), explicamos con $\hat{\mathbf{Z}}'$ al ajustar los modelos de regresiones múltiples correspondientes.?

$$S.C.\text{ asociada a } \hat{\mathbf{Z}}' = K\sum_i\sum_j \hat{z}_{ij}^2 = K\sum_{r=1}^{R}\hat{\lambda}_r^2$$

Por tanto, el porcentaje de inercia de la suma de cuadrados de la interacción explicada en el ajuste lineal (1er ajuste) es de:

$$\text{I.E}_{\text{primer ajuste}} = \frac{\sum_i \sum_j \hat{z}_{ij}^2}{\sum_i \sum_j z_{ij}^2}.100\% = \frac{\sum_{r=1}^{R} \hat{\lambda}_r^2}{\sum_{m=1}^{M} \lambda_m^2}.100\%$$

siendo M y R el rango de las matrices $\mathbf{Z}$ y $\hat{\mathbf{Z}}$ respectivamente.

Al realizar el ajuste bilineal, es decir la descomposición en valores singulares de $\hat{\mathbf{Z}}'$, nos preguntamos:

¿Qué parte de la I.E por el primer ajuste es absorbida en el segundo ajuste?.

Sabemos que la inercia explicada al realizar la descomposición en valores singulares de $\hat{\mathbf{Z}}'$ es:

$$\text{I.E}_{\text{segundo ajuste}} = \frac{\sum_{q=1}^{Q} \hat{\lambda}_q^2}{\sum_{r=1}^{R} \hat{\lambda}_r^2}.100\%$$

siendo Q el número de ejes retenidos en el Biplot.

Por tanto la parte de la inercia absorbida por el primer ajuste que es explicada en el segundo ajuste será:

$$\frac{\sum_{q=1}^{Q}\hat{\lambda}_q^2}{\sum_{m=1}^{M}\lambda_m^2}.100\% = \text{Inercia total absorbida}$$

Por tanto, si la inercia explicada por el primer ajuste es baja, es decir si las variables regresoras están poco relacionadas con las variables dependientes, el método deja de ser efectivo ya que $\sum_{r=1}^{R}\hat{\lambda}_r^2$ será muy pequeño en comparación con $\sum_{m=1}^{M}\lambda_m^2$ y por tanto lo será más aún $\sum_{q=1}^{Q}\hat{\lambda}_q^2$.

2.2.3 APLICACIÓN PRÁCTICA

Apliquemos a nuestros datos el modelo de regresión en rango reducido.

Partimos de nuevo de la matriz **Z** de valores residuales:

$$Z' = \begin{pmatrix} -0{,}12 & 2.92 & 0.31 & -0.09 & -1.80 & 0.65 & -1.44 & -0.35 & -0.83 & 0.84 \\ 1.32 & -1.41 & -0.65 & -0.77 & 0.34 & 0.35 & 0.09 & 0.95 & -0.73 & 0.57 \\ -1.18 & -1.48 & 0.35 & 0.88 & 1.47 & -0.98 & 1.36 & -0.58 & 1.57 & -1.39 \end{pmatrix}$$

Trataremos de explicar estas interacciones a partir de una matriz de variables ambientales (**X**). Para ello utilizamos la cantidad de mm^3 de agua caídos (precipitaciones) durante los meses de Enero y Marzo, en cada uno de los períodos analizados.

$$\mathbf{X} = \begin{bmatrix} 8.8 & 45.7 \\ 31.6 & 9.5 \\ 136 & 70.9 \end{bmatrix}$$

La primera columna de **X** se refiere a los mm^3 de lluvia caídos en el mes de Enero (inicio siembra); mientras que la segunda columna de **X** se refiere a los mm^3 de lluvia caídos en el mes de Marzo (etapa cosecha). Trabajaremos con la matriz X estandarizada por columnas.

$$\mathbf{X}_{\text{est.}} = \begin{bmatrix} -0.737 & 0.119 \\ -0.401 & -1.054 \\ 1.138 & 0.935 \end{bmatrix}$$

El primer paso es calcular la matriz $\hat{\mathbf{Z}}$ formada por los valores estimados de las regresiones de **Z'** en **X**. Ajustamos 10 modelos de regresión lineal múltiple, uno para cada columna de **Z**'. En cada uno de ellos las variables independientes van a ser las columnas de **X**.

Genotipo	**Modelo**
1	Z_1=-0.027pE-1.236pM
2	Z_2=-3.515pE+2.689pM
3	Z_3=-0.298pE+2.689pM
4	Z_4=0.2360pE+0.647pM
5	Z_5=2.2570pE-1.178pM
6	Z_6=-0.872pE+0.006pM
7	Z_7=1.8330pE-0.779pM
8	Z_8=0.3200pE-1.017pM
9	Z_9=1.7110pE-0.250pM
10	Z_{10}=-1.147pE-0.098pM

donde,

pE: precipitaciones en el mes de Enero

pM: precipitaciones en el mes de Marzo

El próximo paso es ajustar cada columna de **Z'** utilizando el modelo correspondiente a cada caso. En nuestro ejemplo particular esto no es necesario porque todos los modelos anteriores son exactos, debido a que son modelos con tres observaciones y dos variables independientes (no tenemos grados de libertad en el error).

Por tanto en este ejemplo particular $\hat{\mathbf{Z}}$ coincide con **Z**'. En tal caso solo necesitamos calcular los coeficientes para las variables ambientales que permitan representarlas en el Biplot. Las matrices de marcadores **A** (genotipos) y **B** (ambientes), coinciden con las anteriores.

Recordemos que para encontrar los coeficientes para las variables externas, se deben hacer las regresiones de cada columna de **B** en **X**.

$$\mathbf{X}_{\text{est.}} = \begin{bmatrix} -0.737 & 0.119 \\ -0.401 & -1.054 \\ 1.138 & 0.935 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1.134 & -0.532 \\ -0.103 & 1.247 \\ -1.025 & -0.714 \end{bmatrix}$$

Primer eje: Y= -1.432pE+0.644pM

Segundo eje: Y= 0.501pE-1.373pM

Por tanto la matriz de marcadores **C** correspondiente a las variables ambientales será:

$$\mathbf{C} = \begin{bmatrix} -1.432 & 0.501 \\ 0.644 & -1.373 \end{bmatrix}$$

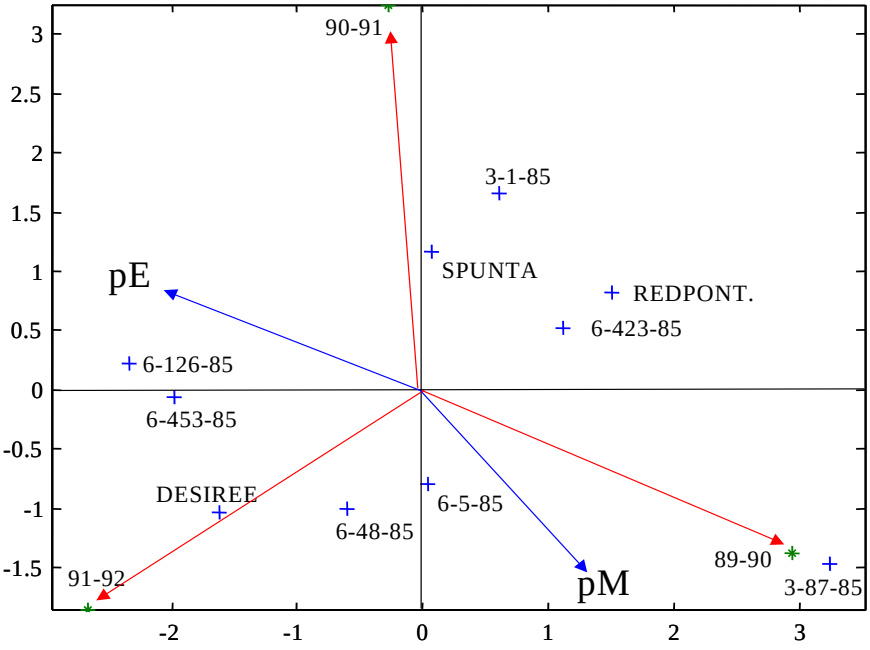


Figura 2.3.: Biplot con variables externas.

Al analizar la figura 2.3, podemos decir que los genotipos 6-126-85 y 6-453-85 interactúan positivamente en la campaña 1991-1992, es decir, altas precipitaciones en el mes de Enero. Podemos decir además que el genotipo 3-87-85 no necesita de excesiva lluvia durante el mes de Enero para presentar un valor elevado de número de tubérculos por planta durante la campaña 1989-1990.

En cuanto al segundo eje podemos decir que el genotipo 3-87-85 reacciona favorablemente a las altas precipitaciones presentadas en la campaña 1989-1990 durante el mes de Marzo, contrario al genotipo 3-1-85 el cual reacciona favorablemente en la campaña 1991-1992, caracterizada por bajas precipitaciones en el mes de Marzo.

Finalmente concluimos diciendo que con la incorporación de variables externas hemos dado una interpretación mucho más completa de la interacción Genotipo-Ambiente.

Recordemos que en nuestro caso hemos utilizado variables externas medidas sobre los ambientes; de igual forma pudo haberse utilizado variables externas medidas sobre los genotipos, e incluso medidas sobre ambas fuentes de variación.

C A P Í T U L O III

LOS MÉTODOS BIPLOT COMO HERRAMIENTA DE ANÁLISIS DE INTERACCIÓN DE ORDEN SUPERIOR

3.1 INTRODUCCIÓN

Como hemos visto en el capítulo anterior, cuando tenemos solamente dos factores de variación, la descomposición en valores singulares de la matriz de residuales de interacción **Z,** nos permite explicar la interacción de segundo orden a partir de los modelos AMMI y poner de manifiesto en la representación Biplot correspondiente qué genotipos interactúan más con el ambiente.

Cuando involucramos tres factores de variación, se incorpora al modelo un término de interacción de tercer orden; determinado por una combinación de niveles de cada factor. Es decir, si tenemos I niveles en el primer factor, J niveles en el segundo factor y K niveles en el tercer factor, los estimadores correspondiente a las interacciones triples van a estar incluidos en K matrices de orden IxJ.

En tal caso no es posible aplicar la descomposición en valores singulares clásica, ya que requeriría convertir varias matrices de datos en una sola matriz, colapsando en uno de los modos o bien fijando uno de ellos; pero en cualquier caso sería imposible estudiar la interacción de orden tres. Necesitamos por tanto obtener una generalización de este concepto a varias matrices de datos.

En otras palabras, si nuestro objetivo es reducir dimensión, y estamos trabajando con tablas de tres vías, necesitamos hacer una generalización de las técnicas anteriormente descritas que permita integrar la información de varias matrices de datos.

Los primeros trabajos de integración de matrices se dan con el Análisis Canónico (HOTELLING (1936)); es una técnica que consiste en buscar relaciones entre dos conjuntos de variables a partir de ejes canónicos. La extensión a más de dos conjuntos de variables se conoce más tarde como Análisis Canónico Generalizado (CARROLL (1968); KETTENRING (1971)).

Actualmente los trabajos de integración de matrices quedan recogidos en dos vertientes fundamentales: los métodos franceses y los métodos anglosajones.

En los métodos franceses el estudio se divide en tres etapas fundamentales:

1) Análisis de la interestructura, que consiste en la comparación global de las matrices originales.
2) Búsqueda de una matriz consenso o compromiso construida a partir de la concatenación de las matrices originales ponderadas; donde la elección de la ponderación es, en general lo que diferencia los distintos métodos.
3) Análisis de la intraestructura que consiste en un estudio más detallado de cada elemento (individuos y variables) de las matrices originales sobre un subespacio común creado en la etapa anterior.

Los primeros métodos que tratan la integración de matrices desde esta perspectiva son por orden cronológico: El Doble A.C.P (BOUROCHE y DUSSAIX (1975)), El Statis (L'HERMIER DES PLANTES (1976)) y Análisis Factorial Múltiple (ESCOFIER y PAGÈS (1984)).

Por otra parte, la escuela anglosajona se caracteriza por ajustar modelos que reproduzcan lo más fiable posible el dato original, en tal sentido podemos citar los métodos de Tuckals (KROONENBERG y DE LEEUW (1980)), basados en el modelo de TUCKER (1966); el Candecomp/Parafac (CARROLL y CHANG (1970); HARSHMAN (1970)); entre otros. Estos métodos ofrecen marcadores para los niveles de los factores o modos, lo cual facilita la interpretación de los resultados en término de representaciones Biplot.

El problema de la integración de matrices se recoge además en los métodos de Análisis Procrustes (GOWER (1975); GOWER y HAND (1996)); Meta-Componentes Principales (KRZANOWSKI (1979, 1982)) y Análisis de Componentes Principales Comunes (FLURY (1984, 1988)); los cuales se basan en la búsqueda de una configuración consenso ¨óptima¨, en el sentido de aproximar lo máximo posible las distintas configuraciones asociadas a cada matriz.

En esta misma línea de búsqueda de una configuración consenso, MARTÍN-RODRIGUEZ (1996) hace una generalización para el caso en el que las configuraciones son el resultado de aplicar un análisis biplot a cada matriz inicial de datos.

En este trabajo nos centraremos en los métodos de la escuela anglosajona, y sobre todo en los trabajos de los holandeses Kroonenberg y De Leeuw. Específicamente en los modelos de Tuckals, los cuales ofrecen marcadores para las categorías de los tres modos, lo que a su vez facilita la utilización de técnicas de representación Biplot. Constituyendo una generalización del A.C.P al caso de tres modos (varias matrices de datos).

Otro motivo para usar estos modelos es el hecho de que nuestro objetivo no es encontrar una configuración consenso, ni comparar configuraciones, sino explicar la interacción de orden superior a dos.

Estos modelos han sido aplicados en la interpretación de la interacción de tercer orden en un modelo lineal general correspondiente a un Análisis de Varianza Trifactorial (KROONENBERG y BASFORD (1989); VAN EEUWIJK y KROONENBERG (1998)); específicamente en el análisis de la interacción Genotipo-Ambiente, para el caso en que los ambientes son combinaciones de años y localidades, es decir, involucran dos factores de variación.

En este capítulo se introduce una generalización de la Regresión en Rango Reducido, al caso de varias matrices de datos. Este resultado nos permitirá explicar los residuales de interacción triple, a partir de la información de variables externas.

3.2 GENERALIZACIÓN DE LA DESCOMPOSICIÓN EN VALORES SINGULARES A TRES MODOS

A continuación daremos una serie de conceptos teóricos relacionados con el Análisis de Componentes Principales de tres vías (KROONENBERG (1983)), el cual puede ser visto como una generalización de la descomposición en valores y vectores singulares de una matriz, al caso de varias matrices de datos. Este resultado nos permitirá descomponer los residuales de interacción triple en tres matrices de marcadores, una para cada factor considerado.

Sin embargo, aunque nuestro objetivo es explicar los residuales de interacción triple, es decir los tres modos que consideramos son cada uno de los factores analizados dentro del Análisis de Varianza Trifactorial, toda la teoría que desarrollamos a continuación es aplicable a cualesquiera tres modos**.**

3.2.1 SOBRE LOS DATOS

En un Análisis de Componentes Principales de tres modos, los elementos u observaciones son clasificados de acuerdo a las categorías de tres modos. (individuo, variable y ocasión) Cada dato está relacionado con una categoría del primer modo (individuo), una categoría del segundo modo (variable) y una categoría del tercer modo (ocasión).

A su vez, existen varios tipos de datos de tres modos (KIERS (1988, 1991)):

- Datos de tres vías: Cuando existe un solo conjunto de individuos, un solo conjunto de variables y un solo conjunto de ocasiones. Es decir, la información queda recogida en K matrices de orden IxJ; siendo I, J, K la cantidad de categorías de cada modo.
- Datos de conjuntos múltiples: Cuando uno de los modos está compuesto por varios conjuntos; podemos tener varios conjuntos de individuos, un solo conjunto de variables y un solo conjunto de ocasiones, en tal caso en cada ocasión se miden las mismas variables a diferentes individuos. Tenemos por tanto K matrices de orden N_kxJ; siendo N_k el número de individuos que se evalúan en la ocasión k.

  De igual forma podemos tener varios conjuntos de variables y un mismo conjunto de individuos y ocasiones; en tal caso tenemos K

matrices de orden IxP_k ; siendo P_k el número de variables medidas en la ocasión k. Se miden en cada ocasión diferentes variables a los mismos individuos.

En nuestro caso trabajaremos con datos de tres vías, debido a que nuestro objetivo es explicar la interacción de tercer orden, en la cual los datos presentan esta estructura.

No obstante, los métodos que serán analizados son válidos para tratar con datos de conjuntos múltiples, ya que este tipo de datos puede ser llevado al primer caso, considerando matrices de productos cruzados o de productos escalares, según sea el caso (KIERS (1988)). En esta situación estamos ajustando el modelo a matrices simétricas derivadas de los datos originales, lo cual se conoce como el modelo de escalamiento de tres modos (KROONENBERG (1983); KIERS (1991); LEBART et al (1995)).

3.2.2 DIFERENCIAS CON RESPECTO A LA D.V.S. DE DOS MODOS

Cuando aplicamos un Análisis de Componentes Principales clásico (Biplot) a una matriz **Z**, estamos ajustando a los datos el siguiente modelo:

$$z_{ij} = \sum_{p=1}^{P} \lambda_{pp} u_{ip} v_{jp}$$

Cuando generalizamos a tres modos (varias matrices de datos), tratamos de ajustar a los datos el siguiente modelo:

$$z_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_{ip} b_{jq} c_{kr} \qquad \text{(Tucker (1966))}$$

Tres diferencias fundamentales pueden citarse al comparar una u otra descomposición:

- En dos vías sabemos que el rango asociado al modo fila y al modo columna de la matriz coinciden [rg(XX')=rg(X'X)], se dan las relaciones de transición entre los vectores propios en el espacio de las filas y en el espacio de las columnas; lo que a su vez permite considerar un solo conjunto de componentes principales. Para el caso de tres modos, necesitamos considerar diferentes componentes en cada modo e incluso la cantidad de componentes en cada modo (P, Q y R) no tiene por qué coincidir.
- En el caso de tres vías, la solución no puede encontrarse a partir de las primeras componentes de cada modo, como ocurría en un A.C.P clásico, donde por lo general, los primeros ejes acumulaban la mayor parte de la variabilidad.
- Para el caso de tres vías, al considerar diferente conjunto de componentes en cada modo, es necesario considerar las interrelaciones entre las componentes (g_{pqr}). Dentro de un mismo modo las componentes están incorrelacionadas, pero entre componentes de diferentes modos puede existir interrelación .

Nótese que la segunda de las diferencias es un factor muy importante a tener en cuenta, debido a que uno de los objetivos que se persigue es reducir la dimensionalidad del problema. Es por ello que necesitamos contar con un algoritmo para el cálculo de las matrices de componentes del modelo de tres vías (**A**, **B** y **C**), que nos asegure que en las primeras componentes de cada modo se concentra la mayor variabilidad de los datos, como ocurría para el caso de dos vías.

MÉTODO DE TUCKALS3

TUCKER (1966) propone un modelo para el Análisis de Componentes Principales de tres modos, en el que se contempla la reducción de dimensionalidad en los tres modos.

$$z_{ijk} = \sum_{p=1}^{P1}\sum_{q=1}^{Q1}\sum_{r=1}^{R1} g_{pqr} a_{ip} b_{jq} c_{kr} + e_{ijk} \qquad \text{(Modelo en rango reducido)}$$

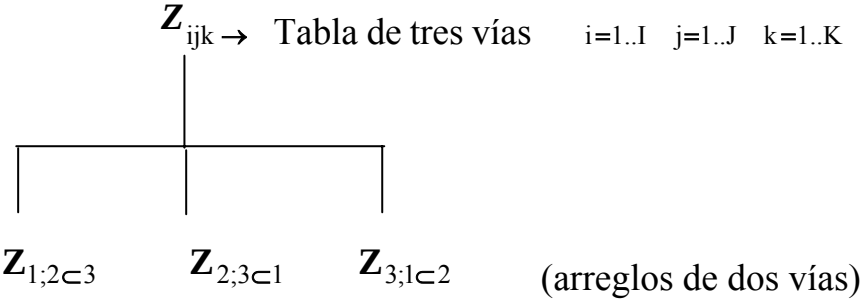
donde z_{ijk} corresponde al valor observado en la combinación de niveles ijk; P1, Q1 y R1 representan el número de componentes retenidas en cada modo; a_{ip} representa el valor que toma para el individuo i la componente p del primer modo; b_{jq} representa el valor que toma para la variable j la componente q del segundo modo; c_{kr} representa el valor que toma para la ocasión k la componente r del tercer modo y g_{pqr} es una medida de la relación entre la componente p del primer modo, la componente q del segundo modo y la componente r del tercer modo.

TUCKER (1966) propone un algoritmo para estimar las matrices **A**, **B** y **C** del modelo; sin embargo, en su trabajo plantea que las soluciones encontradas no son estimadores mínimo cuadráticos; es decir a pesar que para rango completo (tomando P1=cantidad de componentes subyacentes en el primer modo, Q1=cantidad de componentes subyacentes en el segundo modo y R1=cantidad de componentes subyacentes en el tercer modo) se logra reproducir el valor z_{ijk},; al retener las primeras componentes en cada modo, el ajuste producido por el modelo puede ser lo suficiente distante del verdadero valor de z_{ijk} como para ser considerado un mal ajuste.

Para salvar el problema en la estimación de **A**, **B** y **C,** KROONENBERG y DE LEEUW (1980) proponen un método (Tuckals3) que se basa en encontrar los estimadores para **A**, **B** y **C** de manera tal que se minimice la suma de cuadrados residual:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(z_{ijk}-\hat{z}_{ijk})^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(z_{ijk}-\sum_{p=1}^{P1}\sum_{q=1}^{Q1}\sum_{r=1}^{R1}a_{ip}b_{jq}c_{kr}g_{pqr})^2$$

Se parte de una tabla de datos de tres vías $\mathbf{Z_{ijk}}$, a partir de ella se construyen las matrices o arreglos de dos vías $\mathbf{Z}_{1;2\subset3}$, $\mathbf{Z}_{2;3\subset1}$ y $\mathbf{Z}_{3;1\subset2}$ resultado de concatenar dos de los modos. (KROONENBERG (1983)).



En las vías que se concatenan en columnas, al construir la matriz, el índice incluido varía más rápidamente que el otro; así por ejemplo, si tenemos I categorías en el primer factor , J categorías del segundo factor y K categorías en el tercer factor:

$$\mathbf{Z}_{1;2\subset3}=\begin{pmatrix} & j_1k_1 & j_2k_1 & \cdot & j_Jk_1 & \cdot & j_Jk_K \\ i_1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ i_2 & \cdot & z_{221} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ i_I & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

Descomposición simultánea: Análisis de Componentes Principales de tres vías (Modelo de TUCKALS3).

Como se dijo anteriormente, el objetivo es encontrar tres matrices de marcadores **A**, **B** y **C** que permitan aproximar simultáneamente las matrices o arreglos de dos vías anteriores:

$$\left.\begin{array}{l} \mathbf{Z}_{1;2\subset 3} = \mathbf{A}\mathbf{G}_{1;2\subset 3}(\mathbf{C}'\otimes\mathbf{B}') \\ \\ \mathbf{Z}_{2;3\subset 1} = \mathbf{B}\mathbf{G}_{2;3\subset 1}(\mathbf{A}'\otimes\mathbf{C}') \\ \mathbf{Z}_{3;1\subset 2} = \mathbf{C}\mathbf{G}_{3;1\subset 2}(\mathbf{B}'\otimes\mathbf{A}') \end{array}\right| \quad \Rightarrow z_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}a_{ip}b_{jq}c_{kr}$$

$\mathbf{A}_{IxP}$: marcadores para el modo I (Primer Factor)

$\mathbf{B}_{JxQ}$: marcadores para el modo J (Segundo Factor)

$\mathbf{C}_{KxR}$: marcadores para el modo K (Tercer Factor)

$\otimes$: producto de Kronecker

Siendo P, Q y R el rango respectivo de las matrices siguientes:

$$\mathbf{S}_{ii'} = \sum_{j=1}^{J}\sum_{k=1}^{K} z_{ijk}z_{i'jk} = \mathbf{Z}_{1;2\subset 3}\mathbf{Z}'_{1;2\subset 3}$$

$$\mathbf{L}_{jj'} = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ijk}z_{ij'k} = \mathbf{Z}_{2;3\subset 1}\mathbf{Z}'_{2;3\subset 1}$$

$$\mathbf{M}_{kk'} = \sum_{i=1}^{I}\sum_{j=1}^{J} z_{ijk}z_{ijk'} = \mathbf{Z}_{3;1\subset 2}\mathbf{Z}'_{3;1\subset 2}$$

Es decir, P, Q y R representan la cantidad de componentes principales subyacentes en cada modo respectivamente.

La solución a bajo rango se obtiene a partir de un algoritmo iterativo cuya solución inicial para **A, B** y **C** son los vectores propios asociados a los mayores valores propios de las matrices **S, L y M** respectivamente.

En el arreglo de tres vías ***G***, se encuentran las interrelaciones entre las respectivas direcciones de inercia de cada modo. Al igual que como ocurría con ***Z,*** a partir de ***G*** se construyen tres matrices o arreglos de dos vías, $G_{1;2\subset 3}$, $G_{2;3\subset 1}$ y $G_{3;1\subset 2}$, denominadas matrices de enlace; las cuales pueden ser entendidas como una generalización de la matriz de valores propios asociada a la descomposición en dos vías . En apartados posteriores se dan conceptos fundamentales para su interpretación.

Nótese que con esta representación los datos han sido tratados a partir de su estructura en tres vías, las matrices **A**, **B** y **C** de la descomposición sirven para aproximar cada una de las matrices concatenadas. De esta forma los datos no han sido forzados a tener una estructura en dos vías, lo cual haría perder generalidad.

3.2.3 ALGORÍTMO DE TUCKALS3

Si en lugar de trabajar con el modelo de rango completo, (P componentes en el primer modo, Q componentes en el segundo modo y R componentes en el tercer modo), lo hacemos con los primeros ejes de cada modo (P1, Q1 y R1) respectivamente, entonces obtendremos una solución aproximada para los valores de z_{ijk} :

$$z_{ijk} = \sum_{p=1}^{P1}\sum_{q=1}^{Q1}\sum_{r=1}^{R1} g_{pqr} a_{ip} b_{jq} c_{kr} + e_{ijk} \qquad \text{(modelo Tucker3)}$$

Como se dijo anteriormente, el algoritmo se basa en encontrar las matrices **A, B** y **C** que minimicen la expresión:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(z_{ijk} - \hat{z}_{ijk})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(z_{ijk} - \sum_{p=1}^{P1}\sum_{q=1}^{Q1}\sum_{r=1}^{R1} a_{ip} b_{jq} c_{kr} g_{pqr})^2$$

1er paso:

Solución inicial: **(A_0, B_0, C_0)** (Solución dada por TUCKER (1966)).

$\mathbf{A}_0$: P1 primeras columnas de la matriz de vectores propios de $\mathbf{S_{ii}}$,

$\mathbf{B}_0$: Q1 primeras columnas de la matriz de vectores propios de $\mathbf{L_{jj}}$,

$\mathbf{C}_0$: R1 primeras columnas de la matriz de vectores propios de $\mathbf{M_{kk}}$,

KROONENBERG (1983), demuestra que las posibles soluciones para **A, B** y **C** son $\mathbf{A}_0$, $\mathbf{B}_0$ y $\mathbf{C}_0$, o rotaciones ortonormales de las mismas; lo cual significa que la estructura de los ejes factoriales no cambia, sólo se ha rotado.

La solución se va a buscar mediante un proceso iterativo y de manera simultánea de forma tal que las soluciones encontradas sirvan para aproximar cada una de las matrices: $\mathbf{Z}_{1;2\subset3}$, $\mathbf{Z}_{2;3\subset1}$ y $\mathbf{Z}_{3;1\subset2}$ a partir de las ecuaciones dadas en el esquema anterior.

2^do paso:

Se obtiene $\mathbf{A}_1$ a partir de los P1 primeros vectores propios de:

$$(\mathbf{Z}_{1;2\subset 3}(\mathbf{C_0}\otimes\mathbf{B_0}))(\mathbf{Z}_{1;2\subset 3}(\mathbf{C_0}\otimes\mathbf{B_0}))' \quad \text{(transformación ortonormal)}$$

Se comprueba si $(\mathbf{A}_1,\mathbf{B}_0,\mathbf{C}_0)$ es solución (ver criterio de convergencia (*))

3^er paso:

Si la terna anterior no es solución, se obtiene $\mathbf{B}_1$ a partir de los Q1 primeros vectores propios de:

$$(\mathbf{Z}_{2;3\subset 1}(\mathbf{A_1}\otimes\mathbf{C_0}))(\mathbf{Z}_{2;3\subset 1}(\mathbf{A_1}\otimes\mathbf{C_0}))'$$

Se comprueba si $(\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_0)$ es solución

4^to paso:

Si la terna anterior no es solución, se obtiene $\mathbf{C}_1$ a partir de los R1 primeros vectores propios de:

$$(\mathbf{Z}_{3;1\subset 2}(\mathbf{B_1}\otimes\mathbf{A_1}))(\mathbf{Z}_{3;1\subset 2}(\mathbf{B_1}\otimes\mathbf{A_1}))'$$

Se comprueba si $(\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1)$ es solución, en caso de no serlo se pasan a calcular un nuevo $\mathbf{A}_2$, a partir de $\mathbf{B}_1$ y $\mathbf{C}_1$ y así sucesivamente hasta que el algoritmo converja.

* **Criterio de convergencia**:

La solución encontrada será aquella en la que se estabilicen los valores de **A**, **B** y **C**, que es equivalente a decir que se estabiliza la suma de cuadrados residual:

$$\left\|\mathbf{A}_i-\mathbf{A}_{i-1}\right\|^2\to 0\,,\quad \left\|\mathbf{B}_i-\mathbf{B}_{i-1}\right\|^2\to 0\,,\quad \left\|\mathbf{C}_i-\mathbf{C}_{i-1}\right\|^2\to 0$$

Una vez encontrada la solución: $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ y $\hat{\mathbf{C}}$, se pasa al cálculo de $\hat{\mathbf{G}}_{\mathbf{1;2\subset 3}}$:

$$\hat{\mathbf{G}}_{\mathbf{1;2\subset 3}} = \hat{\mathbf{A}}'\mathbf{Z}_{\mathbf{1;2\subset 3}}(\hat{\mathbf{C}}\otimes\hat{\mathbf{B}})$$

Por simetría:

$$\hat{\mathbf{G}}_{\mathbf{2;3\subset 1}} = \hat{\mathbf{B}}' \mathbf{Z}_{2;3\subset 1}(\hat{\mathbf{A}} \otimes \hat{\mathbf{C}}) \quad \text{y} \quad \hat{\mathbf{G}}_{\mathbf{3;1\subset 2}} = \hat{\mathbf{C}}' \mathbf{Z}_{3;1\subset 2}(\hat{\mathbf{B}} \otimes \hat{\mathbf{A}})$$

Nótese que en el algoritmo es necesario fijar a priori el número de componentes con las que se va a trabajar en cada modo,. (P1, Q1 y R1).

3.2.4 INTERPRETACIÓN DE LOS ELEMENTOS DE *G*

Como hemos dicho en apartados anteriores, el arreglo de tres vías ***G*** puede considerarse como una generalización de la matriz Σ de valores propios asociada a la descomposición en valores singulares de dos vías. El valor g_{pqr} se considera como una medida de la relación entre la componente p del primer modo, la componente q del segundo modo y la componente r del tercer modo; y por tanto, la cantidad:

$$\frac{(g_{pqr})^2}{\sum_{pqr}(g_{pqr})^2}$$

representa la parte de la variabilidad de los datos explicada por el análisis que es atribuida a esa combinación de componentes.

De igual forma si queremos conocer qué variabilidad absorbe cada componente en particular, basta con sumar todos los valores de $(g_{pqr})^2$ manteniendo fijo el índice asociado a la componente analizada. Por ejemplo si queremos conocer la importancia de la componente q del segundo modo, debemos calcular la cantidad:

$$\frac{\sum_{p=1}^{P1}\sum_{r=1}^{R1}(g_{pqr})^2}{\sum_{p=1}^{P1}\sum_{q=1}^{Q1}\sum_{r=1}^{R1}(g_{pqr})^2}$$

A diferencia de la matriz de valores propios asociada a la descomposición en valores singulares en tablas de dos vías, cuando generalizamos a tres vías, en ***G*** podemos encontrar valores negativos. Veremos a continuación cómo interpretar el signo de los elementos de ***G***.

Recordemos que los elementos de ***G*** representan relaciones entre componentes, es decir, entre variables continuas; por tanto su interpretación es mucho más compleja comparado con interacciones entre niveles de tres factores en un análisis de varianza o categorías en una tabla de contingencia.

Supongamos que deseamos interpretar un valor g_{pqr} positivo.

Cuando esto ocurre, KROONENBERG (1983) plantea que para las posibles combinaciones ijk de niveles de cada modo, en las ternas formada por el signo de los pesos (a_{ip} , b_{jq} , c_{kr}) pueden darse cuatro situaciones simultáneamente:

$$(+ , + , +) \quad (+ , - , -) \quad (- , + , -) \quad \text{y} \quad (- , - , +)$$

En nuestro contexto, donde los modos constituyen tres factores de variación dentro de un análisis de varianza trifactorial, cada una de estas combinaciones de signos tiene el siguiente significado:

(+ , + , +): Significa que categorías de i con altos pesos en a_{ip}, tienden a tener altos valores de la variable dependiente (interacciones de tercer orden) para combinaciones de categorías jk con altos pesos en b_{jq} y c_{kr}.

(+ , - , -): Significa que categorías de i con altos pesos en a_{ip}, tienden a tener altos valores de la variable dependiente para combinaciones de categorías jk con bajos pesos en b_{jq} y bajos pesos en c_{kr}.

(- , + , -): Significa que categorías de i con bajos pesos en a_{ip}, tienden a tener altos valores de la variable dependiente para combinaciones de categorías jk con altos pesos en b_{jq} y bajos pesos en c_{kr}.

(- , - , +): Significa que categorías de i con bajos pesos en a_{ip}, tienden a tener altos valores de la variable dependiente para combinaciones de categorías jk con bajos pesos en b_{jq} y altos pesos en c_{kr}.

El resto de combinaciones de categorías caracterizadas que no cumplan con ninguna de las cuatros combinaciones de signos anteriores, tendrán los valores más bajos de interacciones triples.

En caso contrario, es decir, si el signo de g_{pqr} es negativo, otras 4 combinaciones asociada al signo de los pesos en **A**, **B** y **C** para las distintas combinaciones , pueden darse simultáneamente en los datos:

(+ , - , +) (+ , + , -) (- , + , +) y (- , - , -)

(+ , - , +): Significa que categorías de i con altos pesos en a_{ip}, tienden a tener altos valores de la variable dependiente para combinaciones de categorías jk con bajos pesos en b_{jq} y altos pesos en c_{kr} .

(+ , + , -): Significa que categorías de i con altos pesos en a_{ip}, tienden a tener altos valores de la variable dependiente para

combinaciones de categorías jk con altos pesos en b_{jq} y bajos pesos en c_{kr}.

(- , + , +): Significa que categorías de i con bajos pesos en a_{ip}, tienden a tener altos valores de la variable dependiente para combinaciones de categorías jk con altos pesos en b_{jq} y altos pesos en c_{kr} .

(- , - , -): Significa que categorías de i con bajos pesos en a_{ip}, tienden a tener altos valores de la variable dependiente para combinaciones de categorías jk con bajos pesos en b_{jq} y bajos pesos en c_{kr} .

Recordemos que en nuestro caso particular, la variable dependiente será la interacción de tercer orden.

Al igual que en el caso de dos vías, este análisis se hace solamente con las combinaciones de categorías ijk con mayor valor absoluto en las respectivas matrices de marcadores. Solamente se trabaja con las categorías que caracterizan cada componente.

3.2.5 SELECCIÓN DEL NÚMERO DE EJES

En este caso como estamos trabajando con datos con estructura de tres vías, las soluciones no son anidadas como ocurre en un ACP; es decir, la solución 2x2x2 incluida en una 3x2x4 no coincide con la solución 2x2x2 resultante de aplicar el algoritmo, además el número de ejes fijado en cada modo no tiene por qué coincidir (TIMMERMAN y KIERS (2000)).

Como hemos dicho el criterio de convergencia del algoritmo está asociado con una estabilidad en la solución, esto implica que podemos obtener un mínimo local y no óptimo (TIMMERMAN y KIERS (2000)). Estos autores proponen un método para seleccionar el número de ejes óptimo. (DIFFIT method).

El método consiste en calcular los valores de ajuste para todas las soluciones posibles obtenidas a partir del algoritmo de TUCKALS3; recordemos que en cada solución se ajusta un modelo con el objetivo de aproximar los valores de **Z,** por tanto a cada solución estará asociado un error y un valor explicado, precisamente el ajuste coincide con la parte de **Z** explicada por cada solución.

Según TIMMERMAN y KIERS (2000), debe cumplirse para las posibles soluciones que (P1≤Q1R1), (Q1≤P1R1) y (R1≤P1Q1), esta condición es debido a que por ejemplo la solución 3x1x2 coincide con la 2x1x2, por tanto la primera se elimina por tener más ejes. Nótese que esta condición elimina soluciones redundantes.

Una vez encontrados estos valores de ajuste; en las soluciones posibles y para cada valor de S=P1+Q1+R1, se selecciona la de mejor ajuste, es decir el mayor. De las soluciones seleccionadas se pasa al cálculo de las diferencias de ajuste (DifFit) de un modelo a otro, es decir cuánto ganamos en el ajuste al aumentar el número de componentes en cada modo.

Las diferencias de ajuste de una solución a otra, desempeñan el papel de los valores propios en un A.C.P. de dos vías, los cuales estaban ordenados de

forma decreciente; significando que la ganancia en ajuste de una solución a otra era cada vez menor.

El próximo paso será por tanto, determinar un subconjunto de soluciones {s} para las cuales se cumple que todas las soluciones posteriores tienen asociada una diferencia de ajuste menor. Es decir, seleccionar aquellas soluciones que cumplen con la condición:

$$DifFit_s > DifFit_{s+n} \; n=1..S_{max}-s$$

Esto asegura que de una solución a otra se gane en ajuste cada vez menos; se trata de lograr una equivalencia con el A.C.P de dos vías, donde por lo general la mayor parte de la variabilidad queda recogida en los primeros ejes.

Para estas soluciones se calcula el cociente $DifFit_s/DifFit_{s+1}$. La solución óptima será aquella para la cual este cociente es maximal y la DifFit asociada sea mayor que el valor crítico: $\|\mathbf{Z}\|^2 /(S_{max} - 3)$, donde S_{max}=min(I,JK)+min(J,IK)+min(K,IJ).

Este método ayuda a encontrar un balance óptimo entre el número de componentes retenidas en cada modo y la variabilidad explicada por el modelo. TIMMERMAN y KIERS (2000) plantean que si el número de ejes o componentes han sido elegidos adecuadamente, rara vez el algoritmo de Tuckals3 conduce a un óptimo local.

Otros autores ofrecen contrastes de hipótesis para validar el ajuste de modelos de rango 1; es decir, modelos en los que P1=Q1=R1=1. Tratan de ajustar este tipo de modelos a datos con estructura de tres vías en

experimentos no replicados ((BOIK y MARASINGHE (1989); BOIK (1990)). Constituyen una generalización de los modelos de Mandel, al caso de tres vías.

3.3 USO DEL TUCKALS3 EN EL ANÁLISIS DE LA INTERACCIÓN DE TERCER ORDEN

A continuación veremos como explicar los residuales de interacción triple haciendo uso de toda la teoría vista hasta el momento en el capítulo, es decir, a partir de la generalización de la descomposición en valores singulares al caso de varias matrices de datos, particularmente el modelo de Tucker.

Como sabemos, en un análisis de Varianza Trifactorial, el modelo lineal que se sigue es el siguiente:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkl}$$

donde y_{ijkl} es el valor que toma la variable dependiente analizada en la repetición l, para la combinación de niveles ijk.
Sea:

$$\mathbf{Z} = (z_{ijk}) = (\hat{\alpha\beta\gamma})_{ijk} = y_{ijk.} - y_{ij..} - y_{i.k.} - y_{.jk.} + y_{i...} + y_{.j..} + y_{..k.} - y_{....}$$

Es decir, $\boldsymbol{Z}$ es una tabla de tres entradas que contiene los estimadores mínimo cuadráticos correspondientes a la interacción triple.

Al realizar la descomposición en tres vías de $\boldsymbol{Z}$, nos queda el siguiente modelo:

$$E(y_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \sum_{p=1}^{P1}\sum_{q=1}^{Q1}\sum_{r=1}^{R1} g_{pqr} a_{ip} b_{jq} c_{kr}$$

Ello permite dar una interpretación de la interacción de orden 3, e identificar las filas, columnas o celdas causantes de la misma en su interacción con las combinaciones de las restantes dos fuentes de variación.

Las interacciones de segundo orden pueden ser explicadas a partir de los modelos AMMI.

Como hemos visto, se ha logrado hacer una generalización de la descomposición en valores y vectores singulares clásica, a varias matrices de datos (tres modos). Dicho de otra forma, se ha logrado una generalización de los modelos AMMI (dos modos) al caso de tres modos; resultado que nos permite la interpretación de los residuales de interacción triple.

3.4 REPRESENTACIÓN BIPLOT

Cuando ajustamos el modelo de TUCKALS3, los residuales de interacción de orden tres se descomponen a partir de tres matrices de marcadores. Como sabemos, un Biplot permite representar simultáneamente datos con estructura de dos vías, es decir, es una representación plana de dos matrices de marcadores.

¿Cómo representar simultáneamente tres matrices de marcadores?
CARLIER y KROONENBERG (1996) proponen dos tipos de representaciones Biplot para datos con estructura de tres vías, el Biplot interactivo o Biplot con estructura multiplicativa y el Biplot conjunto. En ambos casos se parte de la descomposición en tres vías asociada al modelo

de TUCKER; aunque puede ser aplicado a cualquier otra descomposición en tres vías, obtenida a partir de otro modelo.

3.4.1 BIPLOT INTERACTIVO

Se parte del modelo de TUCKALS3 y consiste en combinar dos de los modos (J y K): Tendremos por tanto marcadores a_i y marcadores d_{jk}.

$$z_{ijk} \approx \sum_{p=1}^{P1} a_{ip} \left(\sum_{q=1}^{Q1} \sum_{r=1}^{R1} g_{pqr} b_{jq} c_{kr} \right) = \sum_{p=1}^{P1} a_{ip} d_{(jk)p}$$

Es una representación biplot con estructura multiplicativa extra en los marcadores columnas (VAN EEUWIJK y KROONENBERG (1998)). Notar que el número de ejes de la representación (P1) está determinado por la cantidad de factores retenidos en el modo I.

Este tipo de representación es aconsejable cuando uno de los modos que interactúan está ordenado (por ejemplo modo tiempo), o cuando la cantidad de niveles de los modos que se concatenan no es excesivamente grande (CARLIER y KROONENBERG (1996)).

Si el número de marcadores jk es excesivamente grande, una representación de este tipo se hace poco entendible, es aconsejable por tanto utilizar otro tipo de representación donde la información quede resumida de manera más simple.

3.4.2 BIPLOT CONJUNTO

En este caso lo que hacemos es un Biplot condicional a uno de los modos, (por ejemplo el modo K). Nuevamente se parte de la misma descomposición asociada al modelo de TUCKER3. El objetivo es hacer un Biplot para cada componente del tercer modo, en el que se representan los marcadores asociado a los otros dos modos (a_i y b_j).

Para cada componente r del tercer modo se construye la matriz:

$$\mathbf{D}_r=\mathbf{A}\mathbf{G}_r\mathbf{B}'$$

y se realiza el Biplot, es decir, la descomposición en valores singulares de $\mathbf{D}_r$, (matriz de orden IxJ). En este Biplot quedan representados las categorías del primer y segundo modo, proyectados sobre la componente r del tercer modo. En $\mathbf{G}_r$ se encuentra la parte de $\boldsymbol{G}$, relacionada con la componente r del tercer modo.

Para relacionar las categorías del tercer modo con las categorías de los dos primeros modos representadas en el gráfico, utilizamos los pesos asociados a las categorías del tercer modo en la componente r; contenidos en la matriz **C**; siendo muy importante el signo de cada peso.

Por ejemplo, si la categoría k del tercer modo tiene asociado un peso negativo alto en la componente r del tercer modo, proximidades entre marcadores a_i y b_j en el gráfico, se interpretan como que interactúan de manera negativa con la categoría k del tercer modo, de igual forma marcadores a_i y b_j distantes en la representación, indican una interacción positiva con la categoría k del tercer modo.

Cada representación gráfica la relacionamos con las categorías del tercer modo con altos pesos (en valor absoluto) en la componente del tercer modo sobre la cual se está condicionando.

Como vemos este tipo de representaciones son factibles cuando el número de categorías de los diferentes modos es elevada; aunque son más representaciones gráficas, son más simples y más fáciles de interpretar.

El Biplot interactivo ha sido utilizado por BRADU y GABRIEL (1978) y por COX y GABRIEL (1982), con la diferencia de que ellos no utilizan la descomposición en tres vías asociada al modelo de TUCKER3, en sus trabajos concatenan desde el principio del análisis dos de los modos y aplican el Biplot a la matriz concatenada; no tratan los datos a partir de su estructura de tres vías.

Bondad de ajuste:

$$\frac{\text{S.C.explicada}}{\text{S.C.Total}} * 100\% = \frac{\sum_{p=1}^{P1}\sum_{q=1}^{Q1}\sum_{r=1}^{R1}(g_{pqr})^2}{\text{Traza}((\mathbf{Z}_{1;2\subset 3})(\mathbf{Z}_{1;2\subset 3})')} * 100\%$$

3.5 IMPLEMENTACIÓN COMPUTACIONAL

Se ha elaborado un programa en MATLAB que permite ajustar el modelo de TUCKALS3 a una tabla de datos con estructura de tres vías. El programa necesita como datos de entrada, la tabla de tres vías $\boldsymbol{Z}$ (en nuestro caso los residuales de interacción de tercer orden) y la cantidad de ejes que deseamos retener en cada modo (P_1, Q_1 y R_1). Automáticamente

calcula a partir del algoritmo presentado anteriormente, las matrices de marcadores **A**, **B** y **C**, así como el arreglo de tres vías ***G***. Finalmente presenta los resultados mediante un Biplot con estructura multiplicativa en los marcadores columnas (Biplot Interactivo).

El programa se elaboró para resolver el caso particular que nos ocupaba, es decir, ajustar el modelo de Tucker a partir del algoritmo de Tuckals3. No obstante consideramos oportuno destacar que existe un software elaborado por el profesor Kroonenberg, que incluye una serie de modelos relacionados con el Análisis de Componentes Principales de tres modos.

Se muestra a continuación un diagrama que contempla las etapas fundamentales del programa; acompañado de un listado del programa.

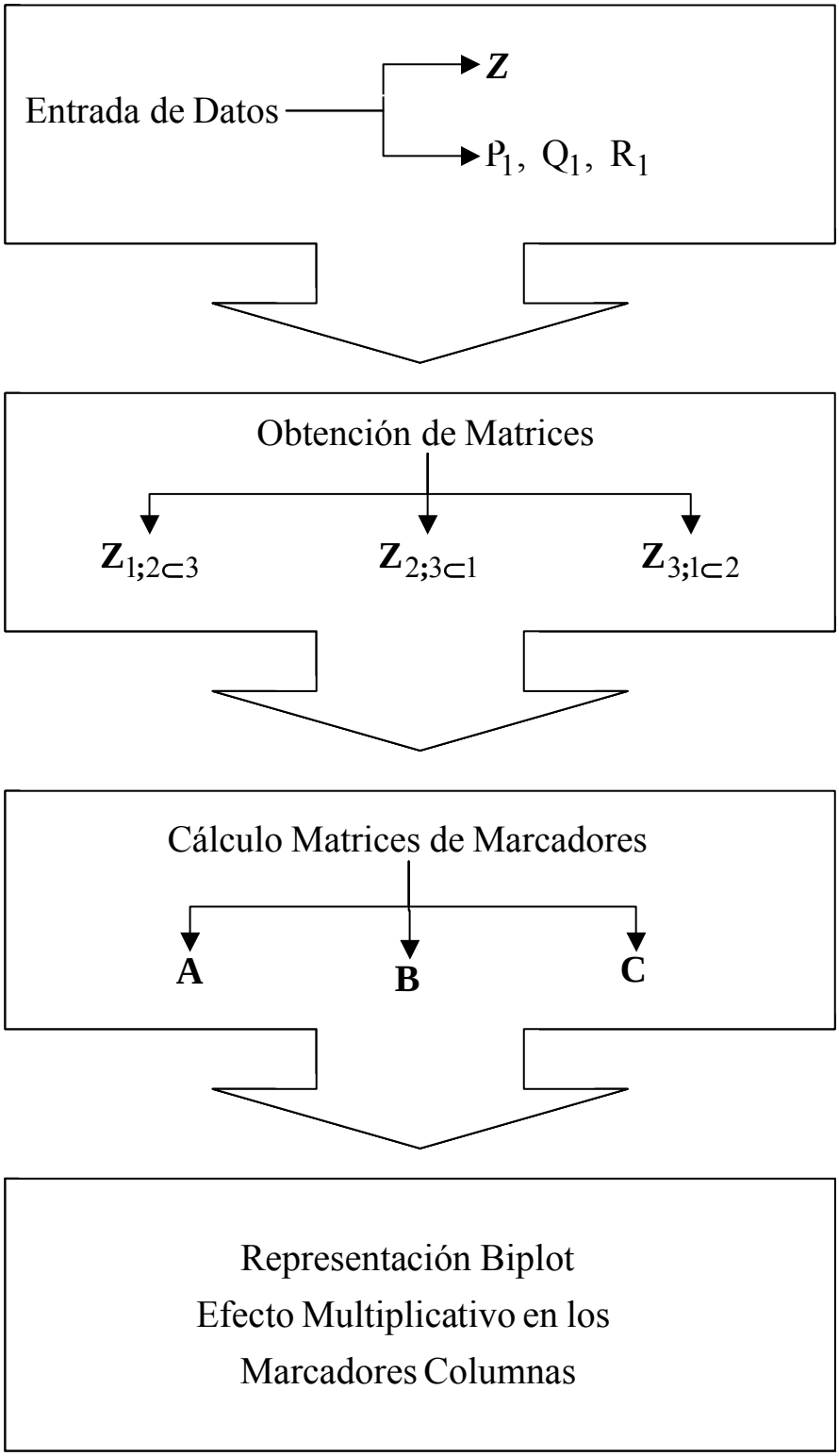


Figura 3.1.: Etapas del algoritmo asociadas al programa computacional.

```
"ANÁLISIS DE COMPONENTES PRINCIPALES DE TRES MODOS
 ALGORÍTMO DE TUCKALS3"

I=input('Teclee el número de categorías del primer factor: ');
J=input('Teclee el número de categorías del segundo factor: ');
K=input('Teclee el número de categorías del primer factor: ');

P1=input('Teclee el número de componentes del primer modo: ');
Q1=input('Teclee el número de componentes del segundo modo: ');
R1=input('Teclee el número de componentes del tercer modo: ');

 "OBTENCIÓN DE LAS MATRICES O ARREGLOS DE DOS VÍAS"
for ii=1:I
  for jj=1:J
    for kk=1:K
      disp([num2str(ii), num2str(jj),num2str(kk)]);
      m(ii,jj,kk)=input(' =  ');
      disp('    ');disp('    ');
    end
  end
end

for ii=1:I
  c=1;
  for kk=1:K
    for jj=1:J
      X1(ii,c)=m(ii,jj,kk);
      c=c+1;
    end
  end
end

for jj=1:J
  c=1;
  for ii=1:I
    for kk=1:K
      X2(jj,c)=m(ii,jj,kk);
      c=c+1;
    end
  end
end

for kk=1:K
  c=1;
  for jj=1:J
    for ii=1:I
      X3(kk,c)=m(ii,jj,kk);
      c=c+1;
```

```
    end
  end
end
```

"CÁLCULO DE LAS MATRICES DE COMPONENTES: (A,B y C)"

```
W1=0;
x=X1*X1'; p=trace(x);   %INERCIA TOTAL
y=X2*X2';
z=X3*X3';
[tam1,tam2]=size(X1);
j1=1;
[a,d1,v]=svd(x);
[b,d2,v]=svd(y);
[c,d2,v]=svd(z);
A=A(:,1:P1);
B=B(:,1:Q1);
C=C(:,1:R1);
while abs(j1)>=0.05;
  k1=Kron(C,B);
  G1=A'*X1*k1;
  r1=Kron(C',B');
  S1=A*G1*r1;
  t1=(X1-S1)*(X1-S1)';
  l1=trace(t1);
  m=l1;
  j1=l1-W1;
  k1=Kron(C,B)
  x=X1*k1;
  x=x*x';
  [a1,d2,v2]=svd(x);
  A=a1(:,1:P1);
  k1=Kron(C,B);
  G1=A'*X1*k1;
  r1=Kron(C',B');
  S1=A*G1*r1;
  t1=(X1-S1)*(X1-S1)';
  l2=trace(t1);
  m=l2;
  j1=l2-l1;
  k2=Kron(A,C);
  y=X2*k2;
  y=y*y';
  [b1,d3,v3]=svd(y);
  b=b1(:,1:Q1);
  k1=Kron(C,B);
  G1=A'*X1*k1;
  r1=Kron(C',B');
  S1=A*G1*r1;
  t1=(X1-S1)*(X1-S1)';
  l3=trace(t1);
```

```
  m=l3;
  j1=l3-l2;
  k3=Kron(B,A);
  z=X3*k3;
  z=z*z';
  [c1,d4,v4]=svd(z);
  c=c1(:,1:R1);
  k1=Kron(C,B);
  G1=A'*X1*k1;
  r1=Kron(C',B');
  S1=A*G1*r1;
  t1=(X1-S1)*(X1-S1)';
  l4=trace(t1);
  j1=l4-l3;
  W1=l4;
end;
aa=a;
```

"OBTENCIÓN DE LOS MARCADORES Djk"

```
K=Kron(C,B);
D=G1*K';
D=D';
disp(D);
 "Suma de cuadrados explicada, en porcentaje."
SCE=(p-m)/p*100
```

"REESCALAMIENTO ÓPTIMO"

```
sca=0;
scb=0;
sca=sum(sum(A.^2));
scb=sum(sum(D.^2));
sca=sca/tam1;
scb=scb/tam2;
scf=sqrt(sqrt(scb/sca));
A=A*scf;
D=D/scf;
disp(D);
stop
```

"REPRESENTACIÓN BIPLOT (BIPLOT INTERACTIVO)"

```
for i=1:R1-1
  for j=i+1:R1
    figure;
    hold
    plot(A(:,i),A(:,j),'b+')
    for k=1:tam2
      plot(D(k,i),D(k,j),'r*')
      str=['ejes' num2str(i) ' y ' num2str(j)];
      xlabel(str);
    end
```

```
    text(A(:,i),A(:,j),label1);
    text(D(:,i),D(:,j),label2);
    axis([-0.8 0.8 -0.6 0.6]);
  end;
end;
```

Seguidamente ofrecemos un esquema en el que se resume todo lo tratado hasta ahora en el capítulo: Es decir, partimos de una tabla de tres vías ***Z*** (en nuestro caso particular los residuales de interacción triple**)**; aplicamos el algoritmo de Tuckals3 para estimar los parámetros del modelo de Tucker (**A**, **B**, **C** y **G**); lo que es equivalente a obtener la descomposición en tres vías de ***Z*.** Finalmente representamos los resultados a partir de un Biplot Interactivo o un Biplot Conjunto.

Insistimos en que en nuestro caso particular, aunque partimos de una tabla ***Z*** formada por los residuales de interacción triple; la metodología desarrollada, (vista como una extensión de la descomposición en valores singulares en dos vías clásica -Análisis de Componentes Principales-) al caso de tres modos, puede ser utilizada para cualquier conjunto de datos con estructura de tres modos.

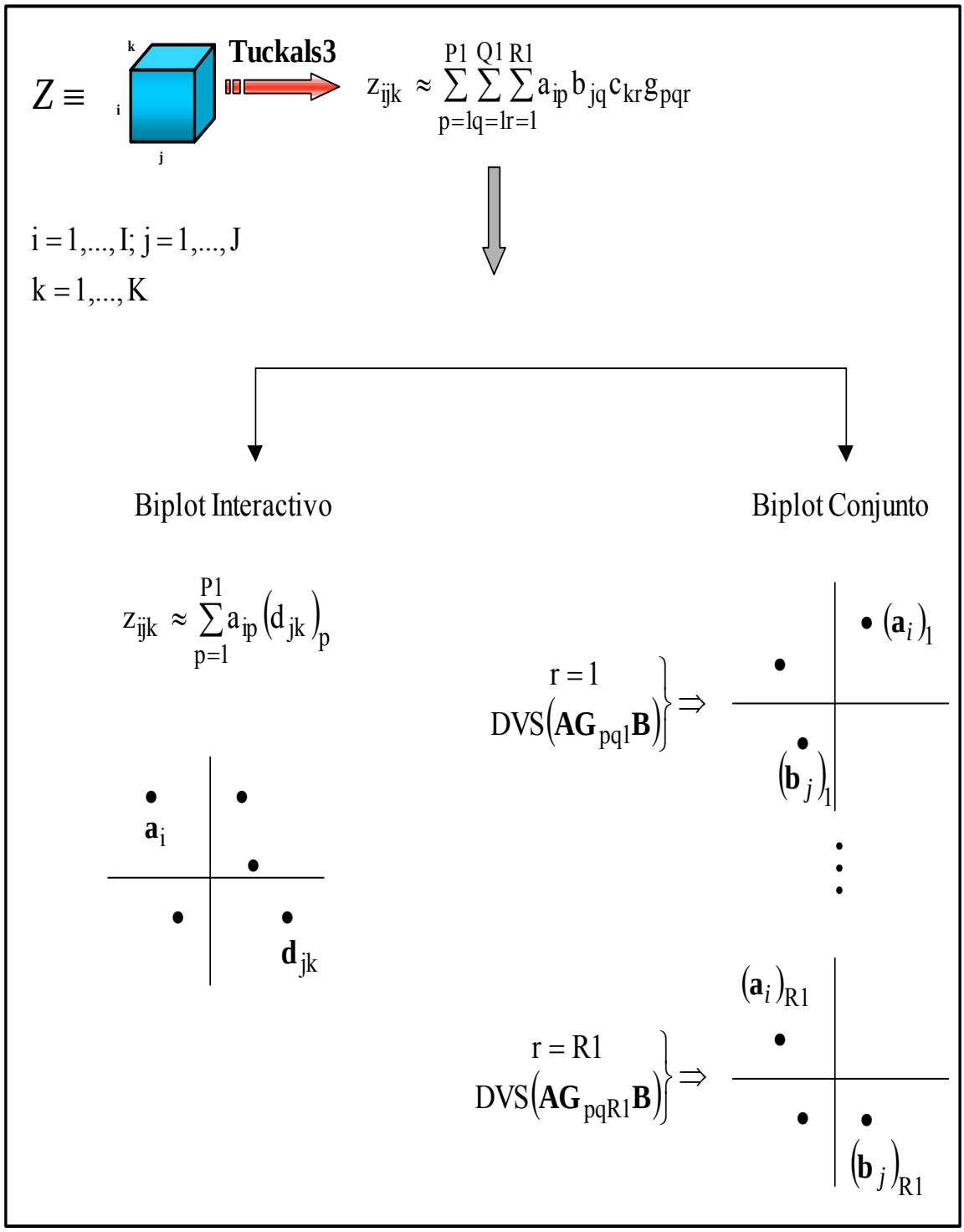


Figura 3.2.: Esquema resumen del capítulo.

3.6 COMPARACIÓN ENTRE TUCKALS3 Y OTROS MÉTODOS DE INTEGRACIÓN

CARROLL y CHANG (1970, 1972) y HARSHMAN (1970) desarrollaron paralelamente dos modelos para el análisis de datos de tres modos. Harshman llamó a su modelo PARAFAC (PARAllel FACtor Analysis); mientras que Carroll y Chang lo denominaron CANDECOMP (CANonical DECOMposition).

A diferencia del modelo de TUCKER3, el PARAFAC/CANDECOMP obtiene componentes comunes para los tres modos; no considera el arreglo de tres vías ***G***.

El PARAFAC/CANDECOMP está basado en una rotación muy simple de los datos (KIERS (1991)). La expresión del modelo es la siguiente:

$$z_{ijk} = \sum_{p=1}^{P} a_{ip} b_{jp} c_{kp}$$

Nótese que la diferencia fundamental con el TUCKALS3 es que en este caso, al no considerar el arreglo ***G***, estamos utilizando los mismos ejes en cada modo; por tanto el PARAFAC/CANDECOMP nunca superará en ajuste al modelo de TUCKASL3 (KIERS (1991)); este modelo es aplicable solamente a datos con determinada estructura; podemos decir que su uso es más limitado. En este sentido, el modelo de TUCKALS3 es aplicable a cualquier conjunto de datos de tres vías puesto que considera diferentes ejes en cada modo, se basa en proyecciones simultáneas de tres nubes de puntos.

No obstante, para los casos en que el PARAFAC/CANDECOMP logra un buen ajuste, la interpretación es mucho más sencilla que en el TUCKALS3, ya que considera menos parámetros en el modelo, siendo aconsejable su aplicación en estos casos. (KIERS (1991)).

Resumiendo, podemos decir que la ventaja fundamental del modelo de TUCKER3 es la de considerar relación entre las componentes de diferentes modos; a diferencia de la mayor parte de los métodos de integración de matrices los cuales tratan de encontrar una configuración consenso, lo cual en muchos casos es imposible debido a la estructura de covariación tan diferente presente en las distintas matrices que deseamos integrar.

En este sentido, FLURY (1995) plantea que el modelo de Krzanowski (Meta Componentes Principales), se ve sensiblemente afectado cuando se presenta inestabilidad en las matrices de vectores propios o direcciones de componentes principales. De igual forma, plantea que su modelo (Análisis de Componentes Principales Comunes) asume igualdad de todos los vectores propios o direcciones principales de inercia.

Refiriéndose a ello, KRZANOWSKI (1990) plantea que el Análisis de Componentes Principales Comunes constituye una etapa intermedia entre las técnicas que exigen matrices de varianzas y covarianzas iguales entre grupos (Análisis Canónico de Poblaciones) y técnicas para el tratamiento de datos sin estructura alguna. En otras palabras, el Análisis de Componentes Principales Comunes aunque puede ser aplicado a datos con matrices de varianzas y covarianzas entre grupos heterogéneas, éstas deben presentar determinada estructura.

Por otra parte, HARSHMAN y LUNDY (1996), plantean que el modelo de TUCKER, al considerar ejes correlacionados entre modos, da la posibilidad de estudiar variaciones mucho más complejas asociada a datos con estructura de tres vías; lo cual es imposible con un modelo tan simple como el PARAFAC.

3.7 ANÁLISIS DE INTERACCIÓN DE ORDEN SUPERIOR A TRES

Desde un punto de vista práctico, es poco usual trabajar con interacciones de orden superior a tres, debido fundamentalmente a lo complicado que resulta su interpretación; sin embargo, con el fin de darle carácter general a la metodología que desarrollamos, consideramos también este caso.

LASTOVICKA (1981) trata el caso de 4 modos, siguiendo la idea de TUCKER (1966). Refiriéndose a ello KAPTEYN et al (1986), plantean que tanto la solución dada por Tucker para tres modos, como la generalización de Lastovicka al caso de 4 modos, tienen el inconveniente de que a pesar que para rango completo reproducen el dato original, al retener las primeras componentes en cada modo el ajuste puede estar muy distante del verdadero valor.

Como vemos al generalizar a 4 modos se presenta el mismo problema que en el caso de tres modos; es por ello que KROONENBERG (1983), hace una generalización al caso de n modos y obtiene como en el caso de tres modos, soluciones mínimo cuadráticas, lo cual elimina el inconveniente que presenta la generalización dada por Lastovicka. Su modelo sigue por supuesto la idea de TUCKER (1966) y lo denomina n-Tucker citado por D'AUBIGNY y POLIT (1989).

CARROLL y CHANG (1970), hacen también una generalización al caso de n modos, pero con el mismo inconveniente de su propuesta para tres modos; es decir, consideran el mismo número de componentes en cada modo, lo cual es muy restrictivo (KAPTEYN et al (1986)).

Resumiendo podemos decir que todos los autores citados anteriormente generalizan los modelos obtenidos para tres modos, al caso de n modos, por lo que sus propuestas tendrán las mismas ventajas y desventajas que presentaban al considerar tres modos.

Por tanto, si estamos trabajando con interacciones de orden superior a tres, basta con generalizar el modelo de Tucker a más modos, en tal caso se obtendrán tantas matrices de marcadores como factores estemos considerando. El algoritmo de cálculo de los estimadores del modelo será similar, solamente será necesario incorporarle más etapas, tantas como modos sean agregados.

3.8 APLICACIÓN A DATOS REALES

Se evalúa el rendimiento (t/ha) de 10 variedades de patata, en 3 localidades durante 3 años. Las 2 primeras localidades Boyeros y San José de las Lajas, están ubicadas es la parte occidental de Cuba, mientras que la localidad de Villa Clara pertenece a la parte central del país. Son regiones de gran producción del tubérculo, con condiciones climáticas extremas.

Las variedades estudiadas fueron:

Aranka, Binella, Provento, Raja, Impala, Snowden, Granada, Desiree, Red Pontiac y Baraka.

En este caso, se quiere estudiar el comportamiento en las condiciones de Cuba, de esas 10 variedades ya establecidas en sus países de origen. Las variedades Snowden y Red-Pontiac son de procedencia canadiense, mientras que el resto son de procedencia holandesa. Se utilizó un Diseño de Bloques al azar con tres réplicas por tratamiento (combinaciones de variedad x localidad x año). Se ofrecen los valores medios para el rendimiento, los cuales se adjuntan en la tabla siguiente:

RENDIMIENTO TOTAL (T/ha)									
Localidad	Boyeros			San José de las Lajas			Villa Clara		
Var./Año	Año 1 93/94	Año 2 94/95	Año 3 95/96	Año 1 93/94	Año 2 94/95	Año 3 95/96	Año 1 93/94	Año 2 94/95	Año 3 95/96
Aranka (v1)	32.07	24.10	41.77	51.37	33.43	38.04	32.04	24.00	43.03
Binella (v2)	30.18	27.33	34.66	42.56	33.12	40.02	34.36	16.89	44.80
Provento (v3)	31.91	24.79	39.33	36.83	33.49	42.03	29.36	19.44	39.50
Raja (v4)	26.64	25.95	31.55	31.84	27.72	40.07	30.26	18.62	42.30
Impala (v5)	25.20	30.36	30.44	36.12	27.18	33.60	30.97	19.56	48.90
Snowden (v6)	27.81	20.39	30.44	24.24	26.95	37.70	29.24	18.33	29.80
Granada (v7)	28.18	25.17	30.22	38.94	37.61	31.80	28.88	20.78	27.20
Desiree (v8)	28.73	24.53	38.88	34.43	29.71	35.70	26.14	20.00	43.40
Red Pont. (v9)	27.78	27.30	34.44	30.18	23.67	43.35	22.69	16.00	41.00
Baraka (v10)	32.00	28.53	36.88	27.85	34.72	40.50	22.09	20.45	32.90

Tabla 3.1.: Matriz de datos.

ESTUDIO DE LA INTERACCIÓN DE TERCER ORDEN:

Matriz de estimadores de interacción de tercer orden:

$$\mathbf{Z}_{1;2\subset 3} = \begin{bmatrix} & \mathbf{L1A1} & \mathbf{L2A1} & \mathbf{L3A1} & \mathbf{L1A2} & \mathbf{L2A2} & \mathbf{L3A2} & \mathbf{L1A3} & \mathbf{L2A3} & \mathbf{L3A3} \\ \mathbf{V1} & -2.910 & 6.611 & -3.680 & -2.048 & -0.996 & 3.093 & 4.974 & -5.569 & 0.615 \\ \mathbf{V2} & -1.856 & 1.201 & 0.656 & 2.726 & 0.664 & -3.331 & -0.862 & -1.859 & 2.741 \\ \mathbf{V3} & 0.782 & -1.217 & 0.435 & -2.033 & 1.219 & 0.811 & 1.249 & -0.004 & -1.247 \\ \mathbf{V4} & 0.181 & -1.261 & 1.081 & 2.486 & -0.915 & -1.573 & -2.668 & 2.176 & 0.493 \\ \mathbf{V5} & -2.242 & 3.566 & -1.332 & 5.492 & -1.328 & -4.166 & -3.253 & -2.241 & 5.489 \\ \mathbf{V6} & 2.329 & -6.137 & 3.808 & -2.372 & 0.763 & 1.608 & 0.043 & 5.373 & -5.417 \\ \mathbf{V7} & -1.195 & -0.174 & 1.368 & -2.546 & 1.626 & 0.918 & 3.739 & -1.454 & -2.287 \\ \mathbf{V8} & 0.148 & 1.802 & -1.950 & -1.520 & 1.085 & 0.433 & 1.372 & -2.888 & 1.515 \\ \mathbf{V9} & 1.276 & -0.363 & -0.911 & 2.868 & -3.330 & 0.462 & -4.144 & 3.693 & 0.450 \\ \mathbf{V10} & 3.493 & -4.023 & 0.532 & -3.052 & 1.243 & 1.808 & -0.440 & 2.780 & -2.340 \end{bmatrix}$$

Hemos mostrado solamente uno de los arreglos de dos vías, porque la intención es mostrar los estimadores de interacción triple. Como sabemos la única diferencia entre las tres matrices o arreglos que se construyen a partir de la tabla de tres vías, es que en cada una de ellas, los elementos aparecen en diferente posición, son los mismos números distribuidos de forma diferente en cada matriz.

Selección del número de ejes:

A continuación mostramos los valores de ajuste para las posibles soluciones

P1	Q1	R1	S=P1+Q1+R1	Ajuste (%)
1	**1**	**1**	**3**	**52.051**
2	**2**	**1**	**5**	**65.732**
2	1	2	5	60.856
1	2	2	5	55.580
2	**2**	**2**	**6**	**90.459**
3	**2**	**2**	**7**	**97.191**
4	**2**	**2**	**8**	**99.999**

Tabla 3.2.: Valores de ajuste para distintas soluciones.

Nótese que sólo se han presentado las posibles soluciones, es decir, como estamos trabajando con interacciones, los grados de libertad para filas, columnas y celdas son (I-1), (J-1) y (K-1) respectivamente, por tanto en nuestro ejemplo P1≤9, Q1≤2 y R1≤2. Además, como se dijo anteriormente, debe cumplirse que en las posibles soluciones (P1≤Q1R1), (Q1≤P1R1) y (R1≤P1Q1), esta última condición es debido a que por ejemplo la solución 3x1x2 coincide con la 2x1x2, por tanto la primera se elimina.

A continuación seleccionamos para cada valor de S, la mejor solución es decir la que presenta un mayor ajuste o varianza explicada:

P1	Q1	R1	S	Ajuste(%)	Diferencia	Increm.
1	1	1	3	52.051	52.051	2.105
2	2	1	5	65.732	13.681	-
2	**2**	**2**	**6**	**90.459**	**24.727**	**3.673**
3	2	2	7	97.191	6.732	2.397
4	2	2	8	99.999	2.808	∞

Tabla 3.3.: Selección de los mejores ajustes.

Nótese que en la tabla 3.3. eliminamos la solución 2x2x1 debido a que existen soluciones posteriores con una diferencia de ajuste mayor.

Por tanto, la solución óptima encontrada es la 2x2x2, es para la que se encuentra un incremento mayor (P1=2, Q1=2, R1=2).

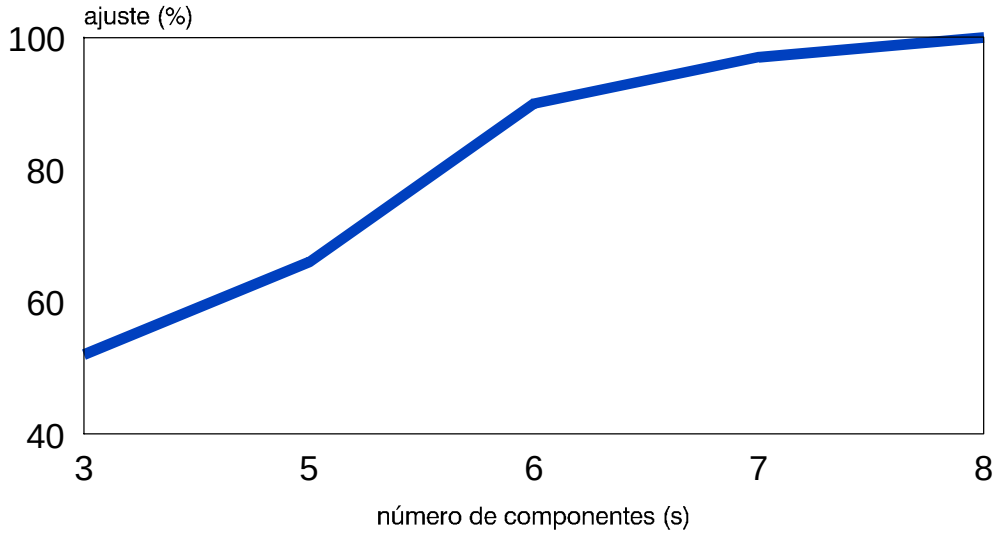


Figura 3.3.: Valores de ajuste para las mejores soluciones.

Matrices de marcadores:

Variedades

$$\mathbf{A}=\begin{bmatrix} 0.417 & -0.566 \\ 0.233 & 0.1923 \\ -0.124 & -0.159 \\ -0.053 & 0.326 \\ 0.465 & 0.429 \\ -0.597 & 0.000 \\ -0.066 & -0.350 \\ 0.163 & -0.206 \\ -0.059 & 0.395 \\ -0.379 & -0.062 \end{bmatrix} \qquad \mathbf{B}=\begin{bmatrix} 0.596 & -0.557 \\ -0.781 & -0.237 \\ 0.184 & 0.795 \end{bmatrix}$$

Localidades

Años

$$\mathbf{C}=\begin{bmatrix} 0.525 & -0.625 \\ 0.279 & 0.767 \\ -0.804 & -0.142 \end{bmatrix}$$

$$\mathbf{G}_{1;2\subset 3}=\begin{bmatrix} & \mathbf{11} & \mathbf{21} & \mathbf{12} & \mathbf{22} \\ \mathbf{1} & -12.545 & -9.611 & 9.207 & -2.696 \\ \mathbf{2} & 9.700 & -6.849 & 3.817 & -7.675 \end{bmatrix}$$

INTERPRETACIÓN DE LOS ELEMENTOS DE ***G***.

La matriz $\mathbf{G}_{1;2\subset 3}$ contiene las relaciones entre los factores o componentes de cada modo; así por ejemplo el valor $\mathbf{(-9.611)^2 = 92.371}$ indica que el primer eje del modo i, el segundo eje del modo j, y el primer eje del modo k, absorben en su conjunto esa cantidad de inercia.

Por tanto, la cantidad:

$$92.371 \big/ 556.186 * 100\% = 16.60\%$$

Representa el porcentaje de la inercia explicada por el análisis que es atribuido a la combinación de componentes 121.

A continuación interpretamos el signo de los elementos de ***G*.**

Comenzamos interpretando la combinación de componentes 111. Sabemos que g_{111}=-12.545 indica la fuerza de la relación entre las primeras componentes de cada modo.

Busquemos ahora las combinaciones de categorías de variedades, localidades y años que caracterizan esta combinación de componentes. Por supuesto, al igual que como hacemos para el caso de dos vías, nos centraremos en las categorías con mayores pesos en valor absoluto, los cuales pueden encontrarse en las matrices **A**, **B** y **C** de marcadores.

Variedades: {Aranka (v1), Impala (v5) y Snowden (v6)}
Localidades: {Boyeros (L1) y San José (L2)}
Años: {93-94 (A1) y 95-96 (A3º)}

Formemos ahora todas las combinaciones posibles de categorías y estudiemos el signo de los pesos en cada matriz de marcadores. En las ternas formadas por combinaciones de signo, cada elemento representa el signo correspondiente al peso de la categoría analizada en esa posición.

Combinación	**Signo**
Aranka x Boyeros x 93-94 (111)	(+ + +)
Aranka x Boyeros x 95-96 (113)	(+ + -)
Aranka x San José x 93-94 (121)	(+ - +)
Aranka x San José x 95-96 (123)	(+ - -)
Impala x Boyeros x 93-94 (511)	(+ + +)
Impala x Boyeros x 95-96 (513)	(+ + -)
Impala x San José x 93-94 (521)	(+ - -)
Impala x San José x 95-96 (523)	(+ - -)
Snowden x Boyeros x 93-94 (611)	(- + +)

Combinación	**Signo**
Snowden x Boyeros x 95-96 (613)	(- + -)
Snowden x San José x 93-94 (621)	(- - +)
Snowden x San José x 95-96 (623)	(- - -)

Como el signo de g_{111} es negativo, las combinaciones de categorías con cualesquiera de estas cuatro combinaciones de signo: (+ - +), (+ + -), (- + +) y (- - -), tendrán mayores valores de interacción triple al ser comparadas con el resto de combinaciones de categorías caracterizadas en esta combinación de componentes, las cuales tendrán las interacciones negativas más altas.

En nuestro caso, podemos decir que las primeras componentes de variedades, localidades y años, contraponen los grupos de categorías:

1^{er} grupo: {113, 121, 513, 521, 611 y 623}
2^{do} grupo: {111, 123, 511, 523, 613 y 621}

El primer grupo con interacciones triples positivas altas y el segundo grupo con las interacciones triples más negativas.

Analicemos ahora otro elemento de **G**, pero en este caso con signo positivo: g_{112}=9.207.

Busquemos el grupo de categorías que caracterizan la componente 1 del modo variedad, la componente 1 del modo localidad y la componente 2 del modo año:

Variedades: {Aranka (v1), Impala (v5) y Snowden (v6)}

Localidades: {Boyeros (L1) y San José (L2)}

Años: {93-94 (A1) y 94-95 (A2)}

Busquemos ahora las combinaciones de categorías y estudiemos el signo de las ternas asociadas:

Combinación	**Signo**
Aranka x Boyeros x 93-94 (111)	(+ + -)
Aranka x Boyeros x 94-95 (112)	(+ + +)
Aranka x San José x 93-94 (121)	(+ - -)
Aranka x San José x 94-95 (122)	(+ - +)
Impala x Boyeros x 93-94 (511)	(+ + -)
Impala x Boyeros x 94-95 (512)	(+ + +)
Impala x San José x 93-94 (521)	(+ - +)
Impala x San José x 94-95 (522)	(+ - +)
Snowden x Boyeros x 93-94 (611)	(- + -)
Snowden x Boyeros x 94-95 (612)	(- + +)
Snowden x San José x 93-94 (621)	(- - -)
Snowden x San José x 94-95 (622)	(- - +)

En este caso, como el signo de g_{112} es positivo, combinaciones de categorías con cualesquiera de las combinaciones de signos (+ + +), (+ - -), (- - +) y (- + -), tendrán asociado mayores valores de interacción triple en comparación al resto de combinaciones de categorías caracterizadas, las cuales tendrán los menores valores de interacción triple. Se forman por tanto los siguientes dos grupos:

1^{er} grupo: {112, 121, 512, 521, 611 y 622}

2^{do} grupo: {111, 122, 511, 522, 612 y 621}

Por tanto, la combinación de la primera componente de variedades, con la primera componente de las localidades, con la segunda componente de los años, contrapone las combinaciones de categorías del 1^{er} grupo con las combinaciones de categorías del 2^{do} grupo; el primer grupo con valores más elevados de interacción de tercer orden.

Todas estas conclusiones podrán verse más adelante en la representación Biplot con estructura multiplicativa.

Para la representación Biplot con estructura multiplicativa (Biplot Interactivo), se utilizan las matrices **A** y $\mathbf{D}_{jk}$ transformadas por un factor de escala, recordemos que $\mathbf{D}_{jk}$ se obtiene a partir de las matrices **C** , **B** y $\mathbf{G}_{1;2\subset 3}$. Cada fila de $\mathbf{D}_{jk}$ está asociada a cada una de las columnas de $\mathbf{Z}_{1;2\subset 3}$ por el orden en que aparecen; así, la tercera fila de $\mathbf{D}_{jk}$ contiene el vector de componentes asociado a L3A1.

$$\mathbf{D}_{jk} = \begin{bmatrix} -1.309 & 0.225 \\ 2.490 & -0.572 \\ -1.182 & 0.347 \\ 1.139 & 1.839 \\ -0.394 & -0.608 \\ -0.744 & -1.230 \\ 0.169 & -2.064 \\ -2.094 & 1.180 \\ 1.925 & 0.883 \end{bmatrix} \qquad \mathbf{A}_t = \begin{bmatrix} 1.752 & -2.375 \\ 0.976 & 0.8065 \\ -0.522 & -0.667 \\ -0.223 & 1.369 \\ 1.953 & 1.798 \\ -2.505 & 0.001 \\ -0.279 & -1.471 \\ 0.684 & -0.866 \\ -0.248 & 1.657 \\ -1.591 & -0.262 \end{bmatrix}$$

Bondad de ajuste:

$$\left(556.186 / 614.994\right) * 100\% = 90.437\%$$

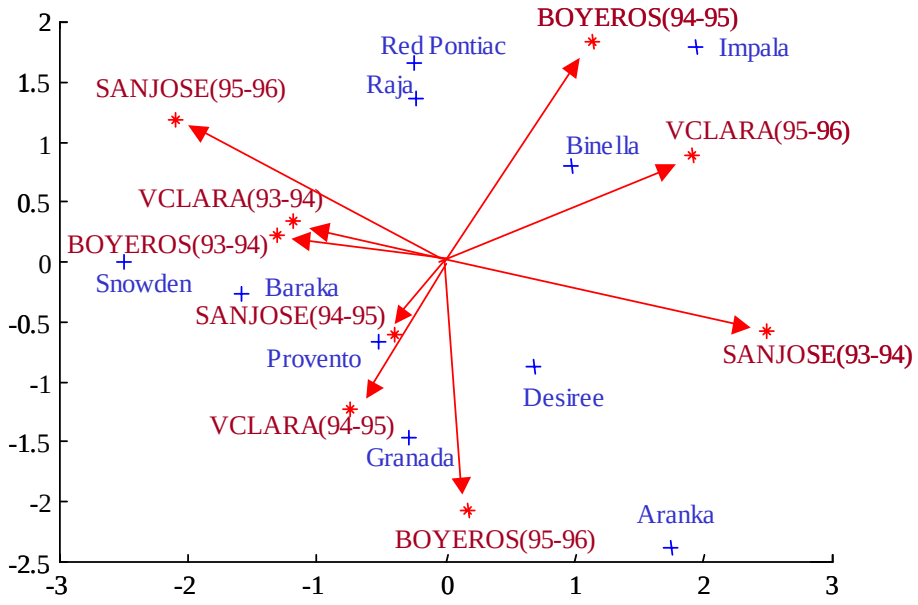


Figura 3.4.: Biplot con estructura multiplicativa (Biplot Interactivo).

Se destacan como más inestables las variedades Aranka, Impala y Snowden, (por ser las más distantes al origen de coordenadas).

Variedad 1 (Aranka): Reacciona favorablemente (altos rendimientos) en los ambientes (L1A3) y (L2A1), es decir, en la localidad Boyeros para el año 93/94, y en la localidad de San José de Las Lajas para el año 93/94. Por otra parte reacciona desfavorablemente en (L1A2), es decir en Boyeros para el año 94/95.

Variedad 5 (Impala): Reacciona favorablemente en la (L3A3) y en (L1A2), es decir, en Villa Clara para el año 95/96 y en Boyeros para el año 94/95. Tiene un comportamiento desfavorable en la (L1A3), o sea, en Boyeros para el año 95/96.

Variedad 6 (Snowden): Interactúa positivamente en las condiciones (L2A3), es decir, en San José de las Lajas en el año 95/96, presentando un mal comportamiento en (L2A1), o sea, en San José en el año 93/94.

En sentido general podemos decir que todas las variedades tuvieron un comportamiento bastante inestable.

Aunque en nuestro ejemplo fue aconsejable utilizar el Biplot con estructura multiplicativa (Interactivo) por tener pocas categorías los modos que interactúan (localidades y años); realizamos además el Biplot conjunto, con la idea de comparar los resultados de una y otra representación. Recordemos que se realiza un Biplot para cada componente del tercer modo.

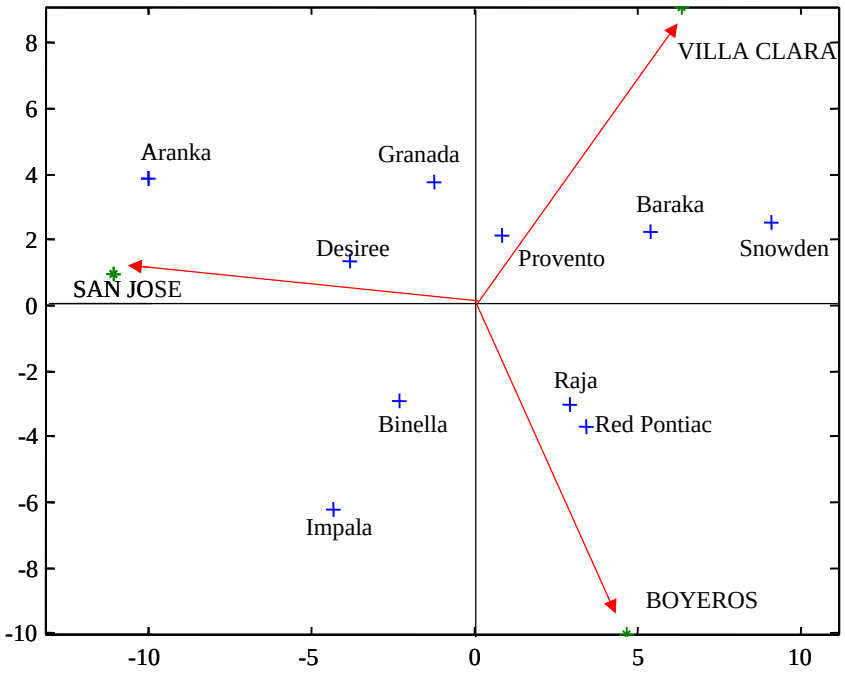


Figura 3.5.: Representación simultánea de variedades y localidades sobre la primera componente de los años (95/96(-)).

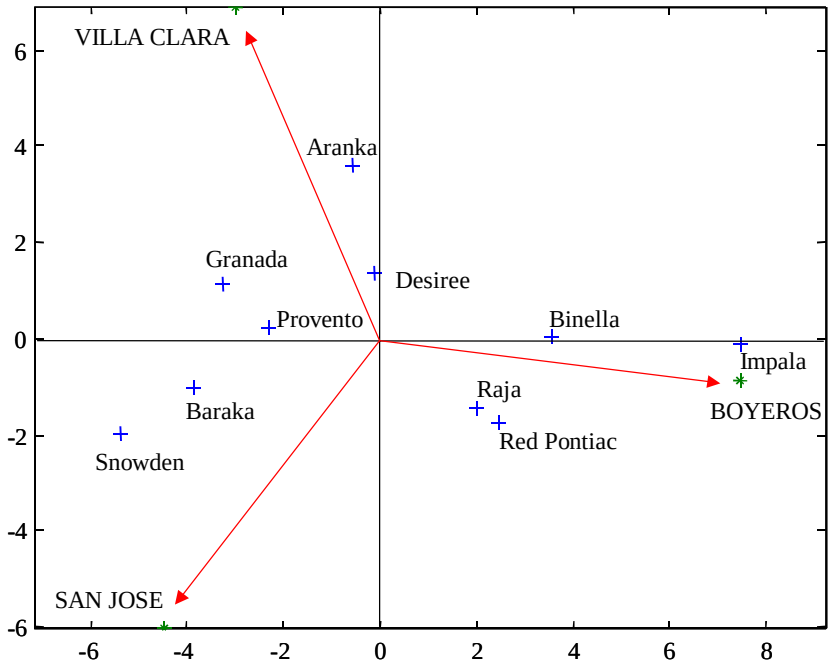


Figura 3.6.: Representación simultánea de variedades y localidades sobre la segunda componente de los años (93/94(-); 94/95(+)).

Hemos construido dos representaciones simultáneas de variedades y localidades para cada componente de los años. El primer gráfico está relacionado con el año 95/96 con peso negativo (ver matriz **C**). Por tanto proximidades entre variedades y localidades se interpretan como que interactúan de manera negativa con el año 95/96.

Podemos decir que las variedades Red Pontiac, Raja e Impala, interactúan de manera negativa en la localidad Boyeros para el año 95/96, las variedades Red Pontiac, Raja, Snowden y Baraka, interactúan de manera positiva en la localidad San José para el año 95/96; la variedad Impala interactúa de manera positiva en la localidad Villa Clara para el año 95/96; las variedades Snowden y Granada interactúan de manera negativa en la localidad Villa Clara en el año 95/96; la variedad Aranka interactúa de manera positiva en las localidad Boyeros para el año 95/96 y de manera negativa en la localidad San José para el año 95/96.

Nótese que al interpretar el primer gráfico sólo nos hemos referido al año 95/96, ya que es el de mayor peso en la primera componente del tercer modo.

Al interpretar el segundo gráfico, asociado a la segunda componente del tercer modo, vemos que en este caso las categorías más importantes son los años 93/94 y 94/95 (ver matriz **C**); el año 93/94 con peso negativo y el año 94/95 con peso positivo.

Como relaciones más importantes podemos destacar que la variedad Impala interactúa de manera positiva en la localidad Boyeros para el año 94/95 e interactúa de manera negativa en la localidad Boyeros para el año 93/94. La

variedad Aranka interactúa de manera positiva en la localidad Villa Clara para el año 94/95 y de manera negativa en la localidad Villa Clara para el año 93/94, podemos decir además que la variedad Aranka interactúa de manera positiva en la localidad San José para el año 93/94 y de manera negativa en la localidad San José para el año 94/95. La variedad Snowden interactúa de manera positiva en la localidad San José para el año 94/95 y de manera negativa en la localidad San José para el año 93/94; de la misma forma, la variedad Snowden interactúa de manera negativa en la localidad Boyeros para el año 94/95 y de manera positiva en la localidad Boyeros para el año 93/94.

Nótese que las conclusiones son similares utilizando una u otra representación Biplot.

3.9 REGRESIÓN EN RANGO REDUCIDO DE TRES MODOS

Supongamos que deseamos explicar a partir de variables externas los residuales de interacción triple asociados a una tabla de tres vías para datos continuos.

En el apartado 2.2 hemos visto que a partir de la Regresión en Rango Reducido fue posible explicar los residuales de interacción de orden dos, mediante variables externas medidas sobre los niveles de uno de los factores de variación analizados.

Se demostró que el problema era equivalente a la realización de un doble ajuste. En el primer ajuste se estimaban los residuales de interacción doble

a partir de regresiones sobre las variables ambientales y en un segundo ajuste se efectuaba un Biplot a la matriz de residuales estimada.

Este resultado nos permitió incorporar en el Biplot la información de variables externas; haciendo más enriquecedor el estudio de la interacción de segundo orden.

En el caso de tres vías podemos tener variables externas medidas sobre uno o varios de los factores considerados; e incluso, en algunos casos es posible, que las variables externas estén medidas sobre las combinaciones de dos de los factores, por ejemplo, cuando consideramos varias localidades en varios años y tenemos la medida de una variable ambiental en cada localidad y cada uno de los años.
Describiremos primero un procedimiento general para la estimación cuando tenemos variables externas en cada uno de los modos, seguido de un procedimiento que usaremos cuando las variables son medidas sobre la concatenación de dos de los modos.

3.9.1 INFORMACION EXTERNA SOBRE LOS TRES MODOS

Supongamos que disponemos de información externa adicional para los tres modos contenida en las matrices **X**, **Y** y **W** respectivamente. Las matrices son de órdenes IxL, JxM y KxN respectivamente, es decir, disponemos de L variables externas para las filas, M para las columnas y N para las celdas.

Partimos del modelo general de tres vías

$$z_{ijk} = \sum_{p=1}^{P1}\sum_{q=1}^{Q1}\sum_{r=1}^{R1} g_{pqr}a_{ip}b_{jp}c_{kr} + e_{ijk}$$

De la misma forma que en el caso general, se trata de estimar las matrices **A**, **B**, **C** y **G,** pero ahora con la restricción adicional de que **A**, **B** y **C** sean combinaciones lineales de las respectivas variables externas:

$$\mathbf{A} = \mathbf{XD}$$
$$\mathbf{B} = \mathbf{YE}$$
$$\mathbf{C} = \mathbf{WF}$$

El problema consiste, entonces, en estimar las matrices de coeficientes **D**, **E** y **F.**

La estimación se consigue mediante una generalización simple del algoritmo general de estimación.

Construimos las matrices

$$\mathbf{Z}_{1;2\subset 3} = \mathbf{AG}_{1;2\subset 3}(\mathbf{C}'\otimes\mathbf{B}') = \mathbf{XDG}_{1;2\subset 3}(\mathbf{F'W'}\otimes\mathbf{E'Y'})$$
$$\mathbf{Z}_{2;3\subset 1} = \mathbf{BG}_{2;3\subset 1}(\mathbf{A}'\otimes\mathbf{C}') = \mathbf{YEG}_{2;3\subset 1}(\mathbf{D'X'}\otimes\mathbf{F'W'})$$
$$\mathbf{Z}_{3;1\subset 2} = \mathbf{CG}_{3;1\subset 2}(\mathbf{B}'\otimes\mathbf{A}') = \mathbf{WFG}_{3;1\subset 2}(\mathbf{E'Y'}\otimes\mathbf{D'X'})$$

Las ecuaciones pueden escribirse también como:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_{1;2\subset 3} = \mathbf{DG}_{1;2\subset 3}(\mathbf{F'W'}\otimes\mathbf{E'Y'})$$
$$(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{Z}_{2;3\subset 1} = \mathbf{EG}_{2;3\subset 1}(\mathbf{D'X'}\otimes\mathbf{F'W'})$$
$$(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Z}_{3;1\subset 2} = \mathbf{FG}_{3;1\subset 2}(\mathbf{E'Y'}\otimes\mathbf{D'X'})$$

Obsérvese que las matrices $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_{1;2\subset 3}$, $(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{Z}_{2;3\subset 1}$ y $(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Z}_{3;1\subset 2}$ son los coeficientes de las regresiones de $\mathbf{Z}_{1;2\subset 3}$, $\mathbf{Z}_{2;3\subset 1}$ y $\mathbf{Z}_{3;1\subset 2}$ sobre **X**, **Y** y **W** respectivamente. Es decir, mientras que las ecuaciones iniciales aproximan los valores originales mediante marcadores que son combinaciones lineales de las variables externas, las ecuaciones transformadas aproximan los coeficientes de regresión.

Especificamos las ecuaciones transformadas ya que son importantes para la interpretación de los biplots en los que situamos sobre el gráfico los coeficientes en **D**, **E** y **F.**

El algoritmo general puede escribirse de la siguiente manera:

Paso 0: Inicio

Fijar los valores de P1, Q1 y R1 e iniciar el contador en k=0, y obtener estimadores iniciales $\mathbf{D_0}$, $\mathbf{E_0}$ y $\mathbf{F_0}$ y $\mathbf{A_0}$, $\mathbf{B_0}$ y $\mathbf{C_0}$.

Paso 0.1: Extracción de la información relacionada con las variables externas.

Se obtienen los residuales de la regresión de $\mathbf{Z}_{1;2\subset 3}$ sobre **X**

$$\mathbf{R}^{X}_{1;2\subset 3} = \mathbf{Z}_{1;2\subset 3} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_{1;2\subset 3} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Z}_{1;2\subset 3}$$

Se reorganizan los residuales en la forma $\mathbf{R}^{X}_{2;3\subset 1}$ y se calculan los residuales de la regresión de esta sobre **Y**

$$\mathbf{R}^{X,Y}_{2;3\subset 1} = \mathbf{(I - Y(Y'Y)^{-1}Y')R}^{X}_{2;3\subset 1}$$

Se reorganizan los residuales en la forma $\mathbf{R}^{X,Y}_{3;1\subset 2}$ y se calculan los residuales de la regresión de esta sobre **W**

$$\mathbf{R}^{X,Y,W}_{3;1\subset 2} = \mathbf{(I - W(W'W)^{-1}W')R}^{X,Y}_{3;1\subset 2}$$

entonces $\mathbf{R}^{X,Y,W}_{3;1\subset 2}$ contiene la parte de **Z** no explicada por las variables externas. Llamando $\mathbf{R}^{X,Y,W}$ a la organización en tres vías de $\mathbf{R}^{X,Y,W}_{3;1\subset 2}$, podemos calcular los valores ajustados para las variables externas como.

$$\hat{\mathbf{Z}} = \mathbf{Z} - \mathbf{R}^{X,Y,W}$$

El análisis de Componentes Principales de 3 vías se realiza a partir de la matriz de valores ajustados reorganizados en la forma $\hat{\mathbf{Z}}_{1;2\subset 3}$, $\hat{\mathbf{Z}}_{2;3\subset 1}$ y $\hat{\mathbf{Z}}_{3;1\subset 2}$.

Paso 0.2: Cálculo de los estimadores iniciales

Calcular $\mathbf{A_0}$,como los P1 primeros vectores propios de $\hat{\mathbf{Z}}_{1;2\subset 3}\hat{\mathbf{Z}}'_{1;2\subset 3}$

Calcular $\mathbf{B_0}$,como los Q1 primeros vectores propios de $\hat{\mathbf{Z}}_{2;3\subset 1}\hat{\mathbf{Z}}'_{2;3\subset 1}$

Calcular $\mathbf{C_0}$,como los R1 primeros vectores propios de $\hat{\mathbf{Z}}_{3;1\subset 2}\hat{\mathbf{Z}}'_{3;1\subset 2}$

Paso 1:

Aumentar el contador k=k+1.

Paso 2:

Se obtiene $\mathbf{A_k}$, como los P1 primeros vectores propios de

$$(\hat{\mathbf{Z}}_{1;2\subset 3}(\mathbf{C}'_{k-1} \otimes \mathbf{B}'_{k-1}))(\hat{\mathbf{Z}}_{1;2\subset 3}(\mathbf{C}'_{k-1} \otimes \mathbf{B}'_{k-1}))'$$

Se obtiene $\mathbf{D}_k$ como

$$\mathbf{D}_k = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}_k$$

Se comprueba si ($\mathbf{A}_k$, $\mathbf{B}_{k-1}$, $\mathbf{C}_{k-1}$) es solución, en cuyo caso se termina el proceso.

Paso 3:

Se obtiene $\mathbf{B_k}$, como los Q1 primeros vectores propios de

$$(\hat{\mathbf{Z}}_{2;3\subset 1}(\mathbf{A}'_k \otimes \mathbf{C}'_{k-1}))(\hat{\mathbf{Z}}_{2;3\subset 1}(\mathbf{A}'_k \otimes \mathbf{C}'_{k-1}))'$$

Se obtiene $\mathbf{E}_k$ como

$$\mathbf{E}_k = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{B}_k$$

Se comprueba si ($\mathbf{A}_k$, $\mathbf{B}_k$, $\mathbf{C}_{k-1}$) es solución, en cuyo caso se termina el proceso.

Paso 4:

Se obtiene $\mathbf{C_k}$, como los R1 primeros vectores propios de

$$(\hat{\mathbf{Z}}_{3;1\subset 2}(\mathbf{B}'_k \otimes \mathbf{A}'_k))(\hat{\mathbf{Z}}_{3;1\subset 2}(\mathbf{B}'_k \otimes \mathbf{A}'_k))'$$

Se obtiene $\mathbf{F}_k$ como

$$\mathbf{F}_k = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{C}_k$$

Se comprueba si ($\mathbf{A}_k$, $\mathbf{B}_k$, $\mathbf{C}_k$) es solución, en cuyo caso se termina el proceso, si no es solución se vuelve al paso 2.

Criterio de convergencia

El proceso termina cuando se estabilizan las matrices de coeficientes.

Paso final

Calcular ***G*** como

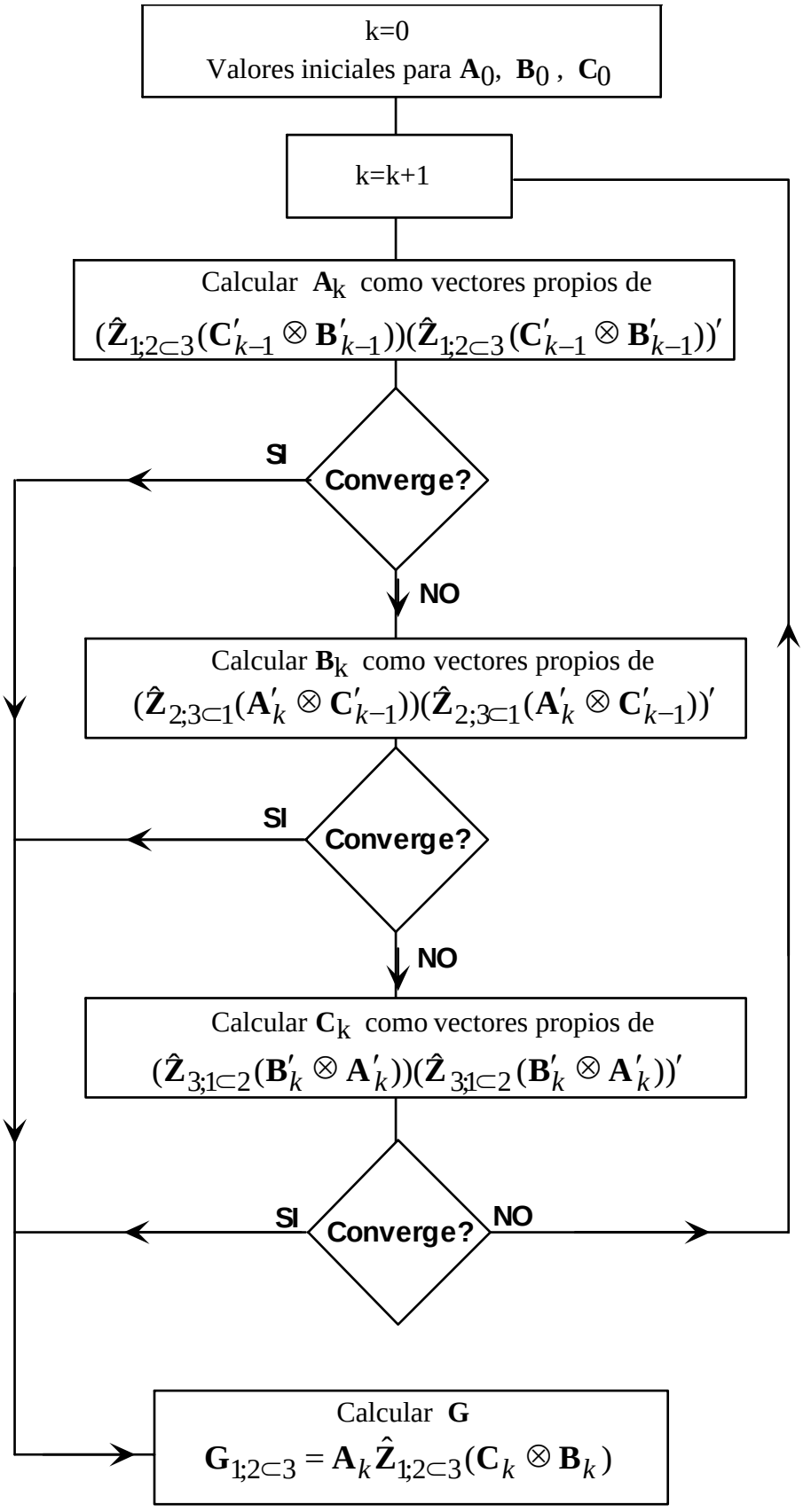
$$\mathbf{G}_{1;2\subset 3} = \mathbf{A}_k \hat{\mathbf{Z}}_{1;2\subset 3}(\mathbf{C}_k \otimes \mathbf{B}_k)$$

Si la información externa se mide solamente sobre alguna de las variables, basta con utilizar la matriz identidad como matriz de información externa donde sea adecuado. Es claro que si no hay información externa, el algoritmo coincide con el algoritmo descrito en apartados anteriores.

En las matrices **D**, **E** y **F**, estarán los coeficientes de las respectivas variables externas que se utilizarán para su representación en el gráfico. Tendremos por tanto 6 matrices de marcadores (**A**, **B**, **C**, **D**, **E** y **F**).

Evidentemente, si queremos utilizar un Biplot interactivo en el que concatenamos los marcadores asociados a **B** y **C**, de igual forma tendremos que concatenar los marcadores asociados a **E** y **F**. Tendremos por tanto, marcadores **A**, $(\mathbf{BC})_{jk}$, **D** y $(\mathbf{EF})_{jk}$

El algoritmo completo puede resumirse en la figura siguiente:


k=0
Valores iniciales para $\mathbf{A}_0$, $\mathbf{B}_0$, $\mathbf{C}_0$
k=k+1
Calcular $\mathbf{A}_k$ como vectores propios de
$(\hat{\mathbf{Z}}_{1;2\subset 3}(\mathbf{C}'_{k-1} \otimes \mathbf{B}'_{k-1}))(\hat{\mathbf{Z}}_{1;2\subset 3}(\mathbf{C}'_{k-1} \otimes \mathbf{B}'_{k-1}))'$
Converge?
SI
NO
Calcular $\mathbf{B}_k$ como vectores propios de
$(\hat{\mathbf{Z}}_{2;3\subset 1}(\mathbf{A}'_{k} \otimes \mathbf{C}'_{k-1}))(\hat{\mathbf{Z}}_{2;3\subset 1}(\mathbf{A}'_{k} \otimes \mathbf{C}'_{k-1}))'$
Converge?
SI
NO
Calcular $\mathbf{C}_k$ como vectores propios de
$(\hat{\mathbf{Z}}_{3;1\subset 2}(\mathbf{B}'_{k} \otimes \mathbf{A}'_{k}))(\hat{\mathbf{Z}}_{3;1\subset 2}(\mathbf{B}'_{k} \otimes \mathbf{A}'_{k}))'$
Converge?
SI
NO
Calcular $\mathbf{G}$
$\mathbf{G}_{1;2\subset 3} = \mathbf{A}_k \hat{\mathbf{Z}}_{1;2\subset 3}(\mathbf{C}_k \otimes \mathbf{B}_k)$

3.9.2 INFORMACION EXTERNA SOBRE LA CONCATENACIÓN DE DOS MODOS

Veremos a continuación cómo a partir del modelo de Tucker, visto como una generalización del Biplot al caso de tres modos, podemos introducir la información de variables externas en el Biplot Interactivo. En otras palabras, explicaremos los residuales de interacción triple a partir de la información de variables externas que en este caso serán medidas sobre las combinaciones de dos de los factores de variación analizados.

Se parte del siguiente modelo:

$$\mathbf{Z'}_{1;2\subset 3} = \mathbf{XM} + \mathbf{E}$$

En este caso **X** es una matriz de orden JKxH, es decir se consideran H variables externas (ambientales) medidas sobre las combinaciones de categorías jk, correspondientes al segundo y tercer factor; en nuestro caso, localidades y años.

Nótese que usar uno u otro arreglo de dos vías **Z**, es arbitrario, es una simple forma de reflejar la información de una tabla de tres vías.

El primer paso es la obtención de los estimadores para los residuales de tercer orden. Se realizan por tanto las respectivas regresiones múltiples, tomando como variables independientes las columnas de **X**, y como variable dependiente cada columna de $\mathbf{Z'}_{1;2\subset 3}$. Se ajustarán tantos

modelos como categorías tenga el primer factor (en nuestro caso I modelos).

Es decir, obtenemos los nuevos valores de interacciones de orden tres a partir de la siguiente ecuación:

$$\hat{\mathbf{Z}}'_{1;2\subset 3} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}'_{1;2\subset 3}$$

Hemos obtenido por tanto, una nueva tabla de tres vías $\hat{\boldsymbol{Z}}$, en la que cada elemento es una combinación lineal de las variables ambientales.

El próximo paso es ajustar el modelo de Tucker a la tabla de tres vías $\hat{\boldsymbol{Z}}$:

$$\hat{z}_{ijk} = \sum_{p=1}^{P1}\hat{a}_{ip}\left(\sum_{q=1}^{Q1}\sum_{r=1}^{R1}\hat{g}_{pqr}\hat{b}_{jq}\hat{c}_{kr}\right) = \sum_{p=1}^{P1}\hat{a}_{ip}\hat{d}_{(jk)p}$$

Para obtener los coeficientes de las variables ambientales, debemos ajustar un modelo de regresión múltiple para cada dimensión retenida (p):

$$\hat{d}_{(jk)p} = \sum_{h=1}^{H} f_{hp} x_{h(jk)}$$

Lo cual significa que la matriz **F**, de coeficientes para las variables ambientales puede obtenerse a partir de la siguiente fórmula:

$$\mathbf{F} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}$$

La matriz **F**, es de orden H*p, el elemento f_{hp} representa el valor correspondiente a la variable ambiental h en la componente p.

APLICACIÓN PRÁCTICA

Trataremos de explicar los residuales de interacción triple contenidos en la matriz $Z_{1;2\subset 3}$ dada en el ejemplo anterior a partir de una matriz **X** de variables externas, medidas sobre combinaciones de localidades x años.

$$\mathbf{X} = \begin{bmatrix} & \textbf{P.I} & \textbf{P.F} & \textbf{T.M} & \textbf{H.R} \\ \textbf{L1A1} & 91.5 & 48.0 & 24.6 & 74.0 \\ \textbf{L2A1} & 86.0 & 37.0 & 22.4 & 73.0 \\ \textbf{L3A1} & 33.2 & 27.2 & 23.2 & 77.8 \\ \textbf{L1A2} & 24.0 & 80.0 & 24.3 & 75.5 \\ \textbf{L2A2} & 19.5 & 48.5 & 22.9 & 77.2 \\ \textbf{L3A2} & 28.0 & 36.2 & 22.8 & 80.0 \\ \textbf{L1A3} & 152 & 79.0 & 24.5 & 74.8 \\ \textbf{L2A3} & 108 & 42.0 & 23.0 & 78.3 \\ \textbf{L3A3} & 8.5 & 86 & 23.6 & 80.5 \end{bmatrix}$$

Las variables externas analizadas fueron:

P.I: Precipitaciones (mm^3) durante los meses de Diciembre y Enero (Inicio de la campaña).

P.F: Precipitaciones (mm^3) durante los meses de Febrero y Marzo (Final de la campaña).

T.M: Temperatura promedio (^0C) durante los cuatro meses.

H.R: Humedad Relativa promedio (%) durante los 4 meses.

El próximo paso es ajustar los valores de $\mathbf{Z}_{1;2\subset 3}$, a partir de la información contenida en **X:**

$$\hat{\mathbf{Z}}'_{1;2\subset 3} = \begin{bmatrix} -2.75 & -1.08 & 0.70 & 0.41 & -2.42 & 3.38 & 0.33 & -0.95 & 0.23 & 2.13 \\ 5.19 & 0.98 & -0.68 & -1.21 & 2.23 & -4.34 & 0.52 & 1.43 & -1.08 & -3.04 \\ -3.59 & -1.44 & 0.74 & 0.69 & -2.87 & 4.23 & 0.09 & -1.32 & 0.79 & 2.68 \\ 0.12 & 2.75 & -1.09 & 0.66 & 4.68 & -4.00 & -1.38 & 0.72 & 0.00 & -2.47 \\ 0.73 & 1.18 & -0.58 & 0.16 & 2.22 & -2.17 & -0.59 & 0.41 & 0.05 & -1.41 \\ -2.08 & -1.17 & 0.50 & 0.37 & -2.09 & 2.81 & 0.06 & -0.84 & 0.64 & 1.83 \\ 1.86 & -0.79 & 0.35 & -0.89 & -1.33 & 0.09 & 0.94 & 0.41 & -0.76 & 0.08 \\ 0.27 & -2.42 & 0.96 & -0.72 & -4.08 & 3.26 & 1.28 & -0.49 & -0.11 & 2.05 \\ 0.23 & 2.01 & -0.90 & 0.50 & 3.68 & -3.26 & -1.27 & 0.63 & 0.24 & -1.86 \end{bmatrix}$$

Finalmente ajustamos el modelo de Tucker a la tabla de tres vías $\hat{\boldsymbol{Z}}$, para encontrar los marcadores $\hat{d}_{jk}$ que nos permitirán posicionar las variables externas en el Biplot Interactivo.

$$\hat{\mathbf{A}} = \begin{bmatrix} 1.16 & -2.66 \\ 1.00 & 0.99 \\ -0.48 & -0.27 \\ -0.06 & 1.05 \\ 1.87 & 1.40 \\ -2.14 & 0.15 \\ -0.34 & -1.02 \\ 0.53 & -0.45 \\ -0.18 & 0.64 \\ -1.37 & 0.18 \end{bmatrix} \qquad \hat{\mathbf{D}} = \begin{bmatrix} -1.681 & 0.37 \\ 2.035 & -0.58 \\ -1.89 & 0.54 \\ 1.67 & 0.82 \\ 1.18 & 0.40 \\ -1.17 & -0.40 \\ -0.48 & -0.69 \\ -1.57 & -0.04 \\ 1.51 & 0.06 \end{bmatrix}$$

$$\hat{\mathbf{G}}_{1;2\subset 3} = \begin{bmatrix} -14.11 & 0.05 & 0.22 & -8.69 \\ 1.30 & -3.50 & -3.24 & -2.25 \end{bmatrix}$$

El próximo paso es realizar las regresiones de **X** en **D.** Los coeficientes de las regresiones serán los elementos de la matriz **F**.

$$\mathbf{F} = \begin{bmatrix} -0.75 & -0.44 \\ 1.73 & -0.21 \\ -1.47 & 0.38 \\ -1.17 & 0.01 \end{bmatrix}$$

Tenemos por tanto las tres matrices de marcadores que nos permitirán posicionar sobre el Biplot Interactivo, las variedades, las combinaciones de localidades x año y las variables ambientales.

Antes de presentar el Biplot Interactivo, pasemos al análisis de la bondad de ajuste:

Bondad 1er ajuste:

$$\frac{\text{traza}(\hat{\mathbf{Z}}_{1;2\subset 3} * \hat{\mathbf{Z}}'_{1;2\subset 3})}{\text{traza}(\mathbf{Z}_{1;2\subset 3} * \mathbf{Z}'_{1;2\subset 3})} * 100\% = 53.5\%$$

Bondad del 2do ajuste:

$$\frac{\sum_{pqr} (\hat{g}_{pqr})^2}{\text{traza}(\hat{\mathbf{Z}}_{1;2\subset 3} * \hat{\mathbf{Z}}'_{1;2\subset 3})} * 100\% = 92.71\%$$

Por tanto, la bondad global será del 49.59%. Ello significa que sólo podrán ser explicadas algunas interacciones de tercer orden, más específicamente las relacionadas con las variedades 2,5,6 y 10, que fueron las de mayor

coeficiente de determinación (R^2) en las respectivas regresiones asociadas al 1er ajuste.

Representamos a continuación el Biplot Interactivo con variables externas. Nótese que las posiciones relativas de variedades y combinaciones de localidad x año, han variado. Estamos representando el Biplot Interactivo de la matriz de estimadores de interacciones de tercer orden ajustada a partir de las variables externas.

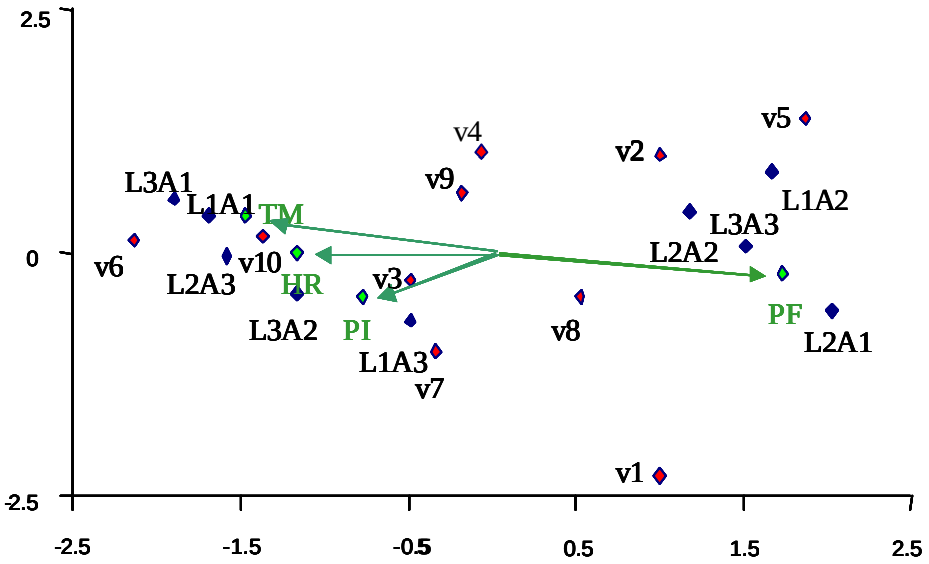


Figura. 3.7. Biplot Interactivo con variables externas

Podemos ver en la figura 3.7 que las variedades 2 y 5 (Binella e Impala) interactúan positivamente en ambientes con altas precipitaciones en la etapa final de la campaña (PF), es decir, en la L1A2 (Boyeros (94-95)) y en la L3A3 (Villa Clara (95-96)). De igual forma, estas variedades interactúan negativamente en L1A3 (Boyeros (95-96)), porque es un ambiente con altas precipitaciones en la etapa inicial de la campaña.

Podemos ver además que las variedades 6 y 10 (Snowden y Baraka), interactúan positivamente en las combinaciones de localidades y años (L3A1, L1A1 y L2A3), es decir, en Villa Clara (93-94), Boyeros (93-94) y San José (95-96), respectivamente. Son ambientes caracterizados por alta temperatura media (TM) y alta humedad relativa (HR).

C O N C L U S I O N E S

1- A partir de la exhaustiva revisión bibliográfica realizada, concluimos que donde primero se explican los residuales de interacción de segundo orden a partir de términos multiplicativos, es en el trabajo de GOLLOB (1968), en lo que denomina modelo FANOVA. Estos mismos modelos son llevados por GAUCH en 1988 al contexto del análisis de Interacción Genotipo Ambiente bajo el nombre de modelos AMMI. Por otra parte GABRIEL en 1978 y DENIS en 1991 lo denominan modelos bilineales, mientras que DENIS y GOWER en 1992 lo denominan modelos biaditivos.

HEMOS DEMOSTRADO:

2- La descomposición en valores y vectores singulares de los residuales de interacción doble, puede ser generalizada al caso de tres modos a partir del ajuste del modelo de Tucker a los residuales de interacción triple.

3- Los residuales de interacción de tercer orden pueden ser representados en dimensión reducida a partir de tres matrices de marcadores (una para cada factor), mediante un Biplot Interactivo o un Biplot Conjunto, dependiendo de la naturaleza del problema.

4- La Regresión Factorial en Rango Reducido puede ser generalizada al caso de tres modos a partir del ajuste del modelo de Tuckals3 a los residuales de interacción triple, estimados a partir de regresiones sobre variables externas, las cuales pueden ser medidas a cada uno de los factores o sobre combinaciones de dos de ellos.

5- Como consecuencia de las tres conclusiones anteriores, podemos decir que los modelos AMMI pueden ser generalizados al caso de tres modos, resultado que permite realizar estudios de Análisis de Interacción Genotipo-Ambiente, cuando los ambientes involucran dos factores de variación.

6- Los métodos estudiados para tres vías pueden ser generalizados al caso de n vías a partir del n-Tuckals; en el que se realizan n Análisis de Componentes Principales simultáneos, sobre la base del algoritmo dado para tres vías.

7- La representación gráfica para el Biplot Interactivo de la matriz de datos, nos permite diagnosticar la presencia/ausencia de interacción de tercer orden en tablas de tres vías a partir de las posiciones de los correspondientes marcadores.

8- En el caso de ausencia de interacción triple, podemos diagnosticar el modelo que mejor se ajusta a los datos, lo cual permite identificar las interacciones dobles que deben ser analizadas mediante los modelos AMMI.

a) La hipótesis relacionada con el modelo aditivo se acepta, si en el gráfico asociado al Biplot Interactivo los marcadores asociado al modo que queda aislado y los marcadores asociado a la combinación de los otros dos modos son colineales y ambas rectas son perpendiculares

b) En las hipótesis relacionadas con la ausencia de una interacción doble, si los modos asociados a esta interacción están concatenados en el Biplot Interactivo, el patrón es cristalino; en caso contrario el patrón es de líneas perpendiculares.

9- La diagnosis sigue siendo válida a nivel de subtablas.

10- Los resultados obtenidos para la diagnosis pueden ser generalizados al caso de n vías, haciendo uso de la generalización del modelo de Tuckals.

BIBLIOGRAFÍA

AMARO, R.I. (2001). *Manova-Biplot para diseños con varios factores basado en modelos lineales generales multivariantes*. Tesis Doctoral. Universidad de Salamanca.

BLÁZQUEZ, A. (1998). *Análisis Biplot basado en Modelos Lineales Generalizados*. Tesis Doctoral. Universidad de Salamanca.

BOIK, R.J. and MARASINGHE, M.G. (1989). 'Analysis of nonadditive multiway classifications'. *Journal of the American Statistical Association*, **84**. 1059-1064.

BOIK, R.J. (1990). 'A likelihood ratio test for three-mode singular values: Upper percentiles and an application to three-way ANOVA'. *Computational and Data Analysis*, **10:** 1-9.

BOUROCHE, J.M. and DUSSAIX, A.M. (1975). 'Several alternatives for three-way data analysis'. *Metra*, **14**, 299-319.

BRADU, D. (1983). 'Model Diagnosis in Two-Way Tables by means of Row and Column Euclidean Maps'. *Technical Report TWISK 307. National Research Institute for Mathematical Sciences*. Pretoria.

BRADU, D. (1984). 'Response surface model diagnosis in Two-Way Tables. Communications in Statistics'. *Theory and Methods* **13** (24), 3059-3106.

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

BRADU, D. and GABRIEL, K.R. (1978). 'The Biplot as a diagnostic tool for models of two-way tables'. *Technometrics*, **20**(1): 47-68.

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

CARLIER, A. and KROONENBERG, P.M. (1996). 'Decompositions and Biplots in Three-way Correspondence Analysis'. *Psychometrika*, **61**(2): 355-373.

CARROLL, J.D. (1968). 'A generalized of canonical correlation analysis to three or more sets of variables'. *Proceedings of 76th annual convention of the American Psychological Associations*, 227-228.

CARROLL, J.D and CHANG, J.J. (1970). 'Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition'. *Psychometrika*, **35**, 283-320.

CARROLL, J.D. and CHANG, J.J. (1972). 'IDIOSCAL (Individual Differences In Orientation Scaling): A Generalization of INDSCAL allowing idiosyncratic reference systems as well as an analytic approximation to INDSCAL'. *Artículo presentado en la Psychometric Society, Princeton*, NJ, Marzo.

CHRISTENSEN, R. (1990 a). *Log-linear Models*. Springer Verlag. New York.

CHRISTENSEN, R. (1990 b). 'Testing for nonaditivity in log-linear and logit models'. *Technical Report 4-5-90*. Department of Mathematics and Statistics. University of Mexico.

CORNELIUS, P.L.; CROSSA, J. and SEYEDSADR, M.S. (1996). 'Statistical tests and estimators of multiplicative models for genotype-by-environment interaction'. En S. MANJIT, H.G. KANG y Jr. GAUCH (eds.). *Genotype by Environment Interaction*. 199-233.

COX, C. and GABRIEL, K.R. (1982). 'Some comparisons of Biplot display and pencil-and-paper E.D.A. methods'. En R.L. LAUNER y A.F. SIEGEL (eds.). *Modern data analysis*. London: Academic Press. 45-82.

DAVIES, P.T. and TSO, M.K. (1982). 'Procedures for reduced-rank regression'. *Applied Statistics*, **31**: 244-255.

DAWID, A.P. (1979). 'Conditional Independence in statistical theory (with discussion)'. *Journal of the Royal Statistical Society*. B, **41**: 1-31.

DENIS, J.B. (1991). Ajustements de modéles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. *Revue de Statistique Appliquée,* **29(2)**, 5-24.

DENIS, J.B. and GOWER, J.C. (1992). *Biadditive models*. Technical Report. Laboratoire de Biométrie, INRA-Versailles.

DENIS, J.B. and GOWER, J.C. (1994). 'Biadditive models'. Letter to the editor. *Biometrics*, **50**, 310-311.

DIAZ-LENO, M.S. (1995). *Los métodos Biplot como herramienta de diagnóstico en la modelización de datos multidimensionales*. Tesis Doctoral. Universidad de Salamanca.

D' AUBIGNY, G. and POLIT, Z. (1989). Some optimality properties of the generalization of the Tucker method to the analysis of n-way tables with specified metrics. En R.Coppi y S.Bolasco (eds.). *Multiway data analysis*. Amsterdam: Elsevier: 39-49.

EBERHART, S.A. and RUSSELL, W.A. (1966). 'Stability parameters for comparing varieties'. *Crop Science*, **6**, 36-40.

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

ECKART, C. and YOUNG, G. (1939). 'A principal axis transformation for non-Hermitian matrices'. *Am.Math.Soc.Bull*, **45**, 118-121.

ESCOFIER, B. and PAGÉS, J. (1984). 'L' Analyse factorielle multiple: Une méthode de comparaison de groupes de variables. [Múltiple Factorial analysis: A methode to compare groups of variables]'. *Data Analysis and Informatics*, **3**, 41-55.

FINLAY, K.W. and WILKINSON, G.N. (1963). 'The analysis of adaptation in a plant breeding programme'. *Australian Journal of Agricultural Research*. **14**: 742-754.

FLURY, B.D. (1984). Common Principle Components in K groups. *Journal of the American Statistical Associations*, **79**, 892-898.

FLURY, B.D. (1988). *Common Principal Components and related multivariate models*. New York: Wiley.

FLURY, B.D. (1995). Developments in Principal Component Analysis. En W.J. Krzanowski (eds.). *Recent Advances in Descriptive Multivariate Analysis*.. Oxford Science Publications.14-33.

GABRIEL, K.R. (1971). 'The Biplot graphic display of matrices with applications to principal components analysis'. *Biometrika*, **58**(3): 453-467.

GABRIEL, K.R. (1972). 'Analysis of meteorological data by means of canonical decomposition and Biplots'. *Journal of Applied Meteorology*, **11**: 1071-1077.

GABRIEL, K.R. (1978). 'Least Squares Approximation of Matrices by Additive and Multiplicative Models'. *Journal of the Royal Statistical Society*, Series B. **40,** 186-196.

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

GABRIEL, K.R.; GALINDO, M.P. y VICENTE-VILLARDON, J.L. (1998). Use of Biplots to diagnose Independence Models in Three-Way Contingency Tables. En M.Greenacre y J.Blasius (eds.). *Visualization of Categorical Data.* Academis Press. London.

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

GALINDO, M.P. (1986). 'Una alternativa de representación simultánea: HJ-Biplot'. *Qüestiió*, 10**(1)**:13-23.

GAUCH, H.G. (1988). 'Model Selection and Validation for Yield Trials with Interaction'. *Biometrics*, **44**: 705-715.

GAUCH, H.G and ZOBEL, R.W. (1989). 'Accuracy and selection success in yield trial analyses'. *Theoretical and Applied Genetics*, 77: 473-481.

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

GOLUB, G.H. and REINSCH, C. (1971). The singular value decomposition. En J:H. Wilkinson y C. Reinsch. (eds.). *Handbook of Automatic Computation.* Springer Verlag. Berlin.

GOWER, J.C. (1975). 'Generalized Procrustes analysis'. *Psychometrika*, **40**: 33-51.

GOWER, J.C. (1990). Three-dimensional biplots. *Biometrika*, 77 (4): 773-785.

GOWER, J.C and HAND, D.J. (1996). *Biplots*. London: Chapman and Hall.

GREENACRE, M.J. (1984). *Theory and applications of Correspondence Analysis*. Academic Press. London.

HARSHMAN, R.A. (1970). 'Foundations of the PARAFAC procedure: models and conditions for an "explanatory " multi-mode factor analysis'. *UCLA Working Papers in Phonetics*, **16**: 1-84.

HARSHMAN, R.A. and LUNDY, M.E. (1996). 'Uniqueness proof for a family of models sharing features of Tucker's three mode factor analysis and Parafac Candecomp'. *Phychometrika*, **61**: 133-154.

HOTELLING, H. (1936). 'Simplified calculations of Principals Components'. *Psychometrika*, **1**: 27-35**.**

ISRAELS, A.Z. (1984). 'Redundancy analysis for qualitative variables'. *Psychometrika*, **49**: 331-346.

IZENMAN, A.J. (1975). 'Reduced-rank regression for the multivariate linear model'. *J. Mult. Analysis*, **5**: 248-264.

KANG, M.S and GAUCH, H.G. (1996). *Genotype by Environment Interaction*. CRC Press . New York.

KAPTEYN, A., NEUDECKER, H. and WANSBEEK, T. (1986). 'An approach to n-mode components analysis'. *Psychometrika*, **51**: 269-275.

KEMPTON, R.A. (1984). 'The use of Biplots in interpreting variety by environment interactions'. *Journal of Agricultural Science*. Cambridge **103:** 123-135.

KETTENRING, J.R. (1971). 'Canonical analysis of several sets of variables'. *Biometrika*, **58**: 433-460.

KIERS, H.A.L. (1988). 'Comparison of "Anglo-Saxon" and "French" Three-Mode methods'. *Statistique et Analyse des Données*, **13**: 14-32.

KIERS, H.A.L. (1991). 'Hierarchical relations among three-way methods'. *Psychometrika*, **56**: 449-470.

KROONENBERG, P.M (1983). *Three-Mode Principal Components Analysis. Theory and Applications*. Leiden, The Netherlands: DSWO-Press.

KROONENBERG, P.M. and DE LEEUW, J. (1980). 'Principal Component Analysis of Three-Mode Data by means of Alternating Least Squares Algorithms'. *Psychometrika*, **45**: 69-97.

KROONENBERG, P.M. and BASFORD, K.E. (1989). 'An investigation of multi-attribute genotype response across environments using three mode principal component analysis'. *Euphytica*, **44**: 109-123.

KRZANOWSKI, W.J. (1979). 'Between-groups comparison of principal components'. *Journal of the American Statistical Association*, **74** (367): 703-707.

KRZANOSWSKI, W.J. (1982). 'Between-group comparison of principal components. –some sampling results'. *Journal of Statistical Computation and Simulation*, **15**: 141-154.

KRZANOSWSKI; W.J. (1990). 'Between-groups analysis with heterogeneous covariance matrices. The common principal component model'. *Journal of classification*, **7**: 81-98.

LASTOVICKA. J.L. (1981). 'The extension of component analysis to four-mode matrices'. *Psychometrika*, **46**: 47-57.

LEBART, L.; MORINEAU, A.; and PIRON, M. (1995). *Statistique Exploratoire Multidimensionnelle* . Dunod. Paris.

L´HERMIER DES PLANTES, H. (1976). *Structuration Des Tableaux á trois indices de la statistique: Théorie et application d'une méthode d'analyse conjointe*. Doctoral Thesis, University of Science and Technology of Languedoc.

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

MARDÍA, K.V.; KENT, J.T. and BIBBY, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

MARTÍN-RODRIGUEZ, J. (1996). *Contribuciones a la integración de subespacios desde una perspectiva Biplot*. Tesis Doctoral.. Universidad de Salamanca.

MARTÍN-RODRIGUEZ, J. (2002). 'Comparison and integration of subspaces from a Biplot perspective'. *Journal of Statistical Planning and Inference*, **102**(2).

MCCULLAGH, P. and NELDER, J.A. (1991). *Generalized Linear Models*. Second Edition. Chapman and Hall.

MILLIKEN, G.A. and JOHNSON, D.E (1989). *Analysis of Messy Data*. Volume 2: Nonreplicated Experiments. New York: Van Nostrand Reinhold.

PATTERSON, H.D. and THOMPSON, R. (1971). 'Recovery of inter-block information when block sizes are unequal'. *Biometrika*, **58**: 545-554.

RAO, C.R. (1964). 'The use and interpretation of principal components analysis in applied research'. *Sankhya*, A**26**: 329-358.

ROBERTS, P. and ESCOUFIER, Y. (1976). 'A unifying tool for linear multivariate statistical methods: the RV-coefficient'. *Applied Statistics*, **25**: 257-265.

ROMAGOSA, I.; SE ULLRICH, F HAN and PM HAYES (1996). 'Use of the AMMI model in QTL mapping for adaptation in barley'. *Theory Applied Genetic*, **93**:30-37.

SEARLE, S.R. (1971). *Linear Models*. Wiley. New York.

SEARLE, S.R., CASELLA, G. and MCCULLOCH, C.E. (1992). *Variance Components*. Wiley. New York.

TER BRAAK, J.F. (1994). 'Biplots in Reduced Rank Regression'. *Biometrics*, **36:** 983-1003.

TIMMERMAN, M.E. and KIERS, H.A.L. (2000). 'Three-Mode principal components analysis. Choosing the numbers of components and sensitivity to local optima'. *British Journal of Mathematical and Statistical Psychology*, **53**: 1-16.

TUCKER, L.R. (1966). 'Some mathematical notes on three-mode factor analysis'. *Psychometrika*, **31**: 279-311.

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

VAN DEN WOLLENBERG, A.L. (1977). 'Redundancy analysis. An alternative for canonical correlation analysis'. *Psychometrika*, **42**: 207-219.

VAN DER BURG, E. and DE LEEUW, J. (1990). 'Non-linear redundancy analysis'. *British Journal of Mathematical and Statistical Psychology*, **43**: 217-230.

VAN EEUWIJK, F.A. (1995a). 'Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models'. *Euphytica*, **84**: 1-7.

VAN EEUWIJK, F.A. (1995b). 'Linear and bilinear models for the analysis of multi-environment trials: An application to data from the Dutch Maize Variety Trials'. *Euphytica*, **84**: 9-22.

VAN EEUWIJK. F.A. (1995c). 'Multiplicative Interaction in Generalized Linear Models'. *Biometrics*, **51**: 1017-1032.

VAN EEUWIJK, F.A. and KROONENBERG, P.M. (1998). 'Multiplicative Models for Interaction in Three-Way ANOVA, with Applications to Plant Breeding'. *Biometrics*, **54**: 1315-1333.

VICENTE-VILLARDÓN, J.L. (1992). *Una Alternativa a las Técnicas Factoriales Clásicas basada en una Generalización de los Métodos Biplot*. Tesis Doctoral. Universidad de Salamanca.

WHITTAKER, J. (1990). *Graphical Models in Multivariate Statistics*. Wiley. New York.

YATES, F. and COCHRAN, W.G. (1938). 'The analysis of groups of experiments'. *Journal of Agricultural Science*, Cambridge **28**, 556-580.

YOUNG, G. and HOUSEHOLDER, A.S. (1938). 'Discussion of a set of points in terms of their mutual distances'. *Psychometrika*, **3**: 19-22.