UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADISTICA

BIPLOT CONSENSO PARA ANÁLISIS DE TABLAS MÚLTIPLES

LUZ MARY PINZÓN S.

2011

BIPLOT CONSENSO PARA ANÁLISIS DE TABLAS MULTIPLES

Memoria que para optar al Grado de Doctor por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Luz Mary Pinzón Sarmiento

Salamanca
2011

Universidad de Salamanca
Depto de Estadística

JOSÉ LUIS VICENTE VILLARDÓN

Profesor Titular del Departamento de Estadística de la Universidad de Salamanca

CERTIFICA:

Que Dña Luz Mary Pinzón Sarmiento, Licenciada en Ciencias Matemáticas, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su direccíon, el trabajo que para optar al Grado de Doctor, presenta con el título:*Biplot Consenso para análisis de tablas múltiples*; y para que conste, firma el presente certificado en Salamanca, el 20 de Octubre de 2011

A la memoria de mi Madre

por haberme dado una vida maravillosa

A Natalia y Carolina, mis hijas

por haberla llenado de amor

AGRADECIMIENTOS

Quiero expresar mis mas sinceros agradecimientos a todas las personas que de alguna manera hicieron posible la realización de este trabajo

Índice general

Índice general **1**

Introducción **I**

1. Representación Biplot **1**

1.1. Introducción 1

1.2. Los Datos 2

1.2.1. Espacio de los Individuos 2

1.2.2. Espacio de las Variables 3

1.3. Biplot 4

1.3.1. Aproximación de una matriz por otra de rango inferior 4

1.3.2. Biplots Clásicos 9

1.3.3. Biplot para análisis factorial 10

1.3.4. Biplots Generales 10

1.3.5. Propiedades de los Biplots 11

1.3.6. Biplots y regresión 15

1.3.6.1. Regresión Multivariante 15

1.3.6.2. Biplots de regresión 18

1.3.6.3. Regresiones alternadas 20

1.3.7. Indicadores de calidad de la representación 27

1.3.8. Calidad Global de representación 28
1.3.9. Correlación entre las variables originales y las componentes Estándar 28
1.3.10. Contribuciones 29

2. Tablas Múltiples **33**

2.1. Introducción 33
2.1.1. Interestructura 35
2.1.2. Consenso 35
2.1.3. Intraestructura 35
2.1.4. Trayectorias 36
2.1.5. Algunos métodos de análisis de tablas múltiples 36
2.2. Análisis Factorial Múltiple 38
2.2.1. Análisis Individuales 40
2.2.2. Análisis Global 42
2.2.2.1. Consenso 42
2.2.2.2. Trayectoria de los individuos 44
2.2.3. Interestuctura 47
2.2.3.1. Representación de cada uno de los grupos de variables 48
2.2.3.2. Medida de calidad de la representación de cada grupo de variables 50
2.3. El Método Statis 51
2.3.1. Interestructura 52
2.3.1.1. Construcción de la imágen euclidea 53
2.3.1.2. Calidad de representación 56
2.3.2. Consenso 57
2.3.2.1. Consenso a partir de objetos no normados . 57

2.3.2.2. Consenso a partir de objetos normados . . . 58
2.3.2.3. Construcción del espacio consenso 58
2.3.3. Intraestructura 59
2.3.3.1. Distancia entre dos puntos 59
2.3.3.2. Correlación de las variables iniciales con los ejes del compromiso 60
2.3.3.3. Trayectorias 60
2.4. El Método Statis Dual 61
2.4.1. Interestructura 61
2.4.1.1. Construcción de la imágen euclidea 62
2.4.1.2. Calidad de representación 62
2.4.2. Consenso 62
2.4.2.1. Consenso a partir de objetos no normados . 63
2.4.2.2. Consenso a partir de objetos normados . . . 63
2.4.2.3. Construcción del espacio consenso 64
2.4.3. Intraestructura 65
2.4.3.1. Covarianza entre variables 65
2.4.3.2. Calidad de representación 65
2.4.3.3. Trayectorias 66
2.5. El Método DACP 67
2.5.1. Interestructura 67
2.5.1.1. Construcción de la imágen euclídea 68
2.5.1.2. Calidad de representación 68
2.5.2. Consenso 69
2.5.2.1. Índices de selección de ejes del espacio Consenso 70
2.5.2.2. Consenso a partir de la selección del mejor conjunto de ejes 71

2.5.2.3. Consenso a partir de la maximización de la inercia explicada 72
2.5.2.4. Consenso a partir de la maximización de la inercia explicada con matrices reducidas . . . 72
2.5.2.5. Consenso a partir de la búsqueda secuencial de un nuevo sistema de ejes 74
2.5.3. Intraestructura 75
2.5.4. Trayectorias 76
2.6. Meta componentes 76
2.6.1. Comparación de dos subespacios 77
2.6.1.1. Ángulo mínimo entre un vector arbitrario del espacio $\mathbf{V}_1$ y el vector más próximo paralelo a este, en el espacio $\mathbf{V}_2$ 77
2.6.1.2. Base para el subespacio más cercano a los espacios $\mathbf{V}_1$ y $\mathbf{V}_2$ 78
2.6.2. Comparación de varios subespacios 80
2.7. Comparación de los métodos expuestos 82
2.7.1. Tablas de datos que se analizan 83
2.7.2. Objetos que caracterizan el análisis 83
2.7.2.1. Objetos no normados 84
2.7.2.2. Objetos normados 84
2.7.3. Objeto Consenso 86
2.7.4. Análisis de la Intraestructura 89
2.7.5. Análisis de la Interestructura 90
2.7.6. Calidad de la representación 91

3. Biplot Consenso para análisis de tablas múltiples 93
3.1. Introducción 93
3.2. Interestructura 95

3.2.1. Análisis Canónico de Poblaciones 95
3.3. Consenso 99
3.3.1. Bondad de ajuste 105
3.3.2. Procedimientos para hallar el Consenso 110
3.3.2.1. Criterio de la selección del mejor conjunto de ejes 110
3.3.2.2. Criterio de maximización de la inercia explicada 111
3.3.2.3. Criterio de maximización de inercia explicada con matrices reducidas I 112
3.3.2.4. Criterio de maximización de inercia explicada con matrices reducidas II 113
3.3.2.5. Criterio inducido por STATIS DUAL 114
3.3.2.6. Criterio inducido por Meta Componentes . . 114
3.3.3. Índices de comparación de las representaciones inducidas por los diferentes criterios 115
3.3.3.1. Pérdida de inercia 115
3.3.3.2. Proximidad entre factores 117

4. Aplicación a datos Reales **119**

4.1. Introducción 119
4.2. Los Datos 120
4.3. Estadística descriptiva 123
4.4. Análisis gráfico 124
4.5. Interestructura 127
4.5.1. Interpretación de la interestructura 128
4.6. Análisis individuales 129
4.7. Criterios 133
4.7.1. Criterio inducido por STATIS Dual-CISD 134

4.7.1.1. Espacio Consenso CISD 135
4.7.1.2. Trayectoria de los individuos 137
4.7.2. Criterio inducido por Meta Componentes-CIMC . . . 138
4.7.2.1. Espacio Consenso 140
4.7.2.2. Trayectoria de los individuos 144
4.7.3. Índices de Comparación 146
4.7.3.1. Inercia retenida en los ACPs individuales . . 146
4.7.3.2. Coeficiente RV 147
4.7.3.3. Inercia retenida por los ejes que generan los diferentes espacios consenso 148
4.7.3.4. Índices $\xi(.,\nu)$ y $\eta(\mathbf{V}_r)$ 149

5. Conclusiones **151**

Índice de tablas **153**

Índice de figuras **155**

Bibliografía **157**

Apéndices **161**

A. Tabla de datos **I**

B. Resultados para los ACP individuales **XI**

C. Biplots para los ACP individuales **XXXI**

Introducción

Teniendo en cuenta los avances en la recolección de datos, cada día es más frecuente encontrar en la práctica situaciones donde se cuenta con varias matrices de datos cuyo fin es realizar un estudio conjunto de todas ellas. Este tipo de información se puede encontrar en dos casos: un único conjunto de individuos sobre los que se miden varios conjuntos de variables de diferente naturaleza o varios conjuntos de individuos sobre los que se miden las mismas variables. Un caso especial que puede ser incluido en alguno de los anteriores es aquel donde se miden las mismas variables para un único conjunto de individuos en momentos diferentes de tiempo o en distintas situaciones experimentales. Este tipo de datos es producido, por ejemplo, para conocer el desarrollo económico de los países a lo largo del tiempo; o la evolución de las variables antropomórficas de un grupo con el cambio de edad. En la literatura, este tipo de información se conoce con el nombre de tablas multi-via y particularmente como datos cúbicos.

El principal objetivo del análisis es comparar la estructura de variabilidad y covariación dentro de los grupos.
Una forma simple de abordar el problema consiste en calcular las componentes principales de cada grupo y estudiar si la estructura de covariación dentro se mantiene en los diferentes grupos; esta estrategia aunque sencilla es excesivamente subjetiva ya que el criterio de similitud depende del investigador. Además, dos conjuntos de componentes principales aparentemente diferentes entre sí, pueden definir el mismo subespacio del espacio multivariante original.

La búsqueda de un ¨ espacio compromiso o espacio común ¨ donde quede analíticamente bien representada la estructura de covariación dentro de cada uno de los grupos, es abordada por autores como Escoufier (1973, 1980) quien propone concatenar todas las matrices de datos considerando la variabilidad entre y dentro de los grupos y realizar una representación

del conjunto completo; Krzanowsky (1979, 1982) lo resuelve a partir de la comparación de las componentes principales comunes a todos los grupos; Lavit (1988) y Bouroche (1975) proponen la diagonalización de una matriz, denominada ¨ objeto común¨ de la misma naturaleza de los grupos y Flury (1988) lo enfoca de manera más general considerando distintos modelos de similitud entre las componentes.

En este trabajo se aborda la obtención de un subespacio de proyección común (o consenso) para todas las matrices, de forma que el subespacio consenso esté, en algún sentido, lo más cerca posible de cada uno de los subespacios de máxima variabilidad individual. Se obtienen como casos particulares las diferentes soluciones dependiendo del criterio de similitud utilizado, como el Statis Dual (Escoufier, 1973; L'Hermier des Plantes, 1976; Lavit, 1988; Lavit et al.,1994), el AFM (Escofier y Pagès 1984), Meta-Biplot (Krzanowski, 1979; Martín et all, 2002) y Doble Análisis en Componentes principales-DACP (Bouroche, 1975; Dazy y Le Barzic, 1996) y se da lugar a proponer algún otro método para realizar el análisis.
Además se propone una representación biplot consenso de forma que las similitudes entre individuos de diferentes ocasiones se interpreten sobre el mismo sistema de referencia.
Para la comparación analítica entre los diferentes métodos se proponen dos indicadores que dan cuenta de la eficacia de éstos en términos de inercia explicada y cosenos de los ángulos entre los ejes individuales y ejes consenso.

Con esta idea, el primer capítulo está dedicado al desarrollo de la teoría que sustenta la representación biplot, tipos de biplot, diferentes métodos de construcción y propiedades que cumplen.

El segundo capítulo se dedica a estudiar algunos métodos de análisis de tablas múltiples, destacar sus diferencias y similitudes según el tipo de tablas que analizan, los objetos que los caracterizan, la construcción del objeto consenso que da pie al espacio común y los elementos que utiliza en el análisis, tanto de la interestructura como de la intraestructura.

En el tercer capítulo se desarrolla la teoría que sustenta la construcción del biplot consenso para análisis de tablas múltiples, sus propiedades y la obtención de los diferentes métodos a partir de este.

El trabajo se completa en el capítulo cuarto, con la aplicación a datos reales del método propuesto donde se analizan dos criterios y se aplican los indicadores para comparar los criterios propuestos en el capitulo tres.

Capítulo 1

Representación Biplot

1.1. Introducción

El Biplot es una técnica de análisis multivariante propuesta por GABRIEL (1971) como método de representación de datos de tres o más variables. La base algebráica en la que se asienta este trabajo es la descomposición en valores singulares de una matriz de datos (ECKART y YOUNG, 1936).

Esta técnica pretende aproximar los elementos de una matriz a partir de vectores denominados marcadores, asociados a las filas y columnas de la misma; dichos vectores se representan en un espacio cuya dimensión es menor que el rango de la matriz. El producto escalar de un marcador fila por un marcador columna aproxima el elemento correspondiente de la matriz original.

Desde el punto de vista algebráico, el método se basa en el mismo principio que la mayoría de las técnicas factoriales de reducción de dimensionalidad, es decir, hace uso de la descomposición espectral de la matriz. La diferencia fundamental con otras técnicas, es que se trata de reproducir el dato y se incorpora una representación conjunta de filas y columnas.

La interpretación se basa en conceptos geométricos: de similitud entre filas (individuos) que corresponde a una función inversa de la distancia entre estos; además 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.

GABRIEL en 1971, propuso dos factorizaciones que denominó: GH-Biplot (CMP) y JK-Biplot (RMP). El GH-Biplot consigue una alta calidad en la representación de las columnas (variables), mientras que el JK-Biplot consigue una alta calidad de representación para las filas.

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 lo denomina, HJ-Biplot.

GOWER (1984), propone dos tipos de biplot, los biplots de interpolación y los de predicción, mientras que GABRIEL (1971, 1981), considera fundamentalmente la predicción.

El método ha sido generalizado al caso de varias matrices de datos por CARLIER y KROONENBERG (1996).

1.2. Los Datos

Se supone que disponemos de observaciones para J variables cuantitativas sobre I individuos los cuales se representan en la matriz $\mathbf{X} = [x_{ij}]$ de tamaño $I \times J$. Las filas de la matriz $\mathbf{X}$ son vectores de tamaño J que pertenecen a un espacio que denominamos el espacio de los individuos y denotamos por R^J. Simétricamente, las columnas de la matriz $\mathbf{X}$ son vectores de dimensión I que pertenecen a un espacio que denominamos el espacio de las variables y que denotamos por R^I.

1.2.1. Espacio de los Individuos

El espacio de los individuos es el espacio vectorial R^J definido geométricamente por los J ejes correspondientes a las variables. Cada individuo es un punto de ese espacio definido por sus J coordenadas y puede ser representado como $\mathbf{x}_i$. Definimos en ese espacio una métrica Φ, que es una matriz de pesos (por lo general, diagonal) de orden J, simétrica y definida positiva, tal que el producto escalar de dos vectores $\mathbf{x}_i$ y $\mathbf{x}_j$ del espacio de los individuos viene dado por

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^{'} \Phi \mathbf{x}_j$$

la norma de un vector $\mathbf{x}_i$, que será denotada por $\|\mathbf{x}_i\|_\Phi$, por

$$\|\mathbf{x}_i\|_\Phi^2 = \langle \mathbf{x}_i, \mathbf{x}_i \rangle = \mathbf{x}_i^{'} \Phi \mathbf{x}_i$$

y la distancia entre dos individuos por

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^{'} \Phi (\mathbf{x}_i - \mathbf{x}_j)$$

Así mismo, definimos la matriz que contiene los productos escalares entre los individuos como

$$\mathbf{W} = \mathbf{X}\mathbf{\Phi}\mathbf{X}^{'}$$

En el espacio de los individuos, la tabla $\mathbf{X}$ está representada por una nube de puntos correspondientes a las I filas (individuos).

1.2.2. Espacio de las Variables

El espacio de las variables es el espacio vectorial R^I cuyos ejes en una representación geométrica son asociados a los I individuos. Cada variable j puede ser representada por el vector $\mathbf{x}_j$ del espacio. Sus elementos son los valores que la variable toma para cada uno de los I individuos.

Para estudiar la proximidad de las variables introducimos en ese espacio una métrica Ψ, que es una matriz simétrica (por lo general, diagonal) definida positiva y de orden I. Si ψ_{kl} es el elemento de la fila k y columna l de la matriz Ψ, el producto escalar de dos variables $\mathbf{x}_j$ y $\mathbf{x}_m$ se escribe:

$$\langle \mathbf{x}_j, \mathbf{x}_m \rangle = \mathbf{x}_j^{'} \Psi \mathbf{x}_m = \sum_{k=1}^{I} \mathbf{x}_{kj} \sum_{l=1}^{I} \psi_{kl} \mathbf{x}_{lm}$$

corresponde a la covarianza entre $\mathbf{x}_j$ y $\mathbf{x}_m$. Si las variables están centradas de acuerdo a la métrica, si Ψ es diagonal con sus elementos positivos y si $\sum_{k=1}^{I} \psi_{kk} = 1$ entonces

$$\|\mathbf{x}_j\|_\Psi^2 = var(\mathbf{x}_j) = Dest^2(\mathbf{x}_j) = s_j^2$$

lo que significa que la longitud de una variable es igual a su desviación típica.

Además, el coseno del ángulo — θ_{jk} — entre dos variables centradas mide la correlación entre ellas, así:

$$\rho_{jk} = \cos\theta_{jk} = \frac{\langle \mathbf{x}_j, \mathbf{x}_k \rangle_\Psi}{\|\mathbf{x}_j\|_\Psi \|\mathbf{x}_k\|_\Psi} = \frac{cov\,(\mathbf{x}_j, \mathbf{x}_k)}{s_j s_k} = \frac{s_{jk}}{s_j s_k}$$

La matriz de varianzas-covarianzas de $\mathbf{X}$ está dada por

$$\mathbf{S} = \mathbf{X}^{'}\mathbf{\Psi}\mathbf{X}$$

1.3. Biplot

Un biplot para la matriz $\mathbf{X}$ que provee información acerca de J variables medidas sobre I individuos, es una representación cartesiana mediante marcadores $\mathbf{b}_1, \ldots, \mathbf{b}_J$ para sus columnas y marcadores $\mathbf{a}_1, \ldots, \mathbf{a}_I$ para las filas, de forma que el producto escalar $\mathbf{a}_i^{'}\mathbf{b}_j$ aproxime el elemento x_{ij} lo mejor posible.

Si tanto los marcadores $\mathbf{a}_i$ para las filas, como los marcadores $\mathbf{b}_j$ para las columnas están representados en un espacio de dimensión $r \leq s$, siendo s el rango de $\mathbf{X}$, se tiene que :

$$\mathbf{X} \simeq \mathbf{A}\mathbf{B}^{'}$$

Se trata entonces de una aproximación de la matriz $\mathbf{X}$ de rango s por otra matriz $\mathbf{A}\mathbf{B}^{'}$ de rango inferior r.
Generalmente r es 2 ó 3 con el fin de visualizar la estructura de la matriz $\mathbf{X}$.

1.3.1. Aproximación de una matriz por otra de rango inferior

El método más conocido para aproximar una matriz por otra de menor rango es el propuesto por Eckart y Young (1936, 1939). La forma canónica del teorema ha sido llamada "estructura básica" por Horst (1963), y puede encontrarse en varios autores, entre ellos Gabriel (1971), Greenacre (1984) y otros más.

Curiosamente, Eckart y Young no probaron el teorema y direccionaron a los lectores a dos referencias. La primera, no tiene la prueba del teorema y en la otra, MacDuffee (1946) presenta una elegante y simple prueba pero solo para el caso donde la matriz que se reduce es cuadrada y no singular. Posteriormenente Keller (1962) obtuvo la solución por mínimos cuadrados pero no hace referencia a la contribucion original del teorema de Eckart-Young. La demostración del teorema debida a Johnson (1963) es la siguiente:

Teorema: Para cualquier matriz real $\mathbf{X}$, se pueden encontrar dos matrices $\mathbf{U}$ y $\mathbf{V}$ tales que $\mathbf{U}'\mathbf{X}\mathbf{V}$ es una matriz: real, diagonal y con elementos no negativos.

Sea $\mathbf{X}$ de orden IxJ. Sin pérdida de generalidad se asume que $I \geq J$.

El producto simétrico $\mathbf{X}\mathbf{X}'$ tiene el mismo rango que $\mathbf{X}$ y por tanto, como mucho tiene J valores propios no nulos. Seleccionando $\mathbf{U}$ como la matriz ortogonal de orden $I \times I$ que tiene en sus columnas los vectores propios de $\mathbf{X}\mathbf{X}'$ y $\mathbf{\Lambda^2}$ la matriz $I \times I$ diagonal con los valores propios no nulos de $\mathbf{X}\mathbf{X}'$ sobre la diagonal, se puede escribir:

$$\mathbf{U}'\mathbf{X}\mathbf{X}'\mathbf{U} = \begin{bmatrix} \mathbf{\Lambda^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{1.3.1}$$

Expresando el producto

$$\mathbf{U}'\mathbf{X} = \begin{bmatrix} \mathbf{F} \\ \mathbf{E} \end{bmatrix} \tag{1.3.2}$$

donde $\mathbf{F}$ es una matriz desconocida de orden $J \times J$, y $\mathbf{E}$ de orden $(I-J) \times J$ también es desconocida. La matriz $\mathbf{F}$ es aumentada con $\mathbf{E}$ en el caso que $\mathbf{X}$ y por tanto $\mathbf{U}'\mathbf{X}$ no sean cuadradas. En el caso que $\mathbf{X}$ sea cuadrada se puede omitir el aumento con $\mathbf{E}$ en (3.2.2) y las filas y columnas de ceros pueden ser omitidas del lado derecho de (3.2.1).

Multiplicando a ambos lados de (3.2.2) por su traspuesta, se obtiene:

$$\mathbf{U}'\mathbf{X}\mathbf{X}'\mathbf{U} = \begin{bmatrix} \mathbf{F} \\ \mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{F}' & \mathbf{E}' \end{bmatrix} = \begin{bmatrix} \mathbf{F}\mathbf{F}' & \mathbf{F}\mathbf{E}' \\ \mathbf{E}\mathbf{F}' & \mathbf{E}\mathbf{E}' \end{bmatrix} \tag{1.3.3}$$

Igualando (3.2.1) y (3.2.3), tenemos:

$$\begin{bmatrix} \mathbf{F}\mathbf{F}' & \mathbf{F}\mathbf{E}' \\ \mathbf{E}\mathbf{F}' & \mathbf{E}\mathbf{E}' \end{bmatrix} = \begin{bmatrix} \mathbf{\Lambda^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{1.3.4}$$

de donde

$$\mathbf{F}\mathbf{F}' = \mathbf{\Lambda^2} \tag{1.3.5}$$

y

$$\mathbf{E}\mathbf{E}' = \mathbf{0} \tag{1.3.6}$$

Si todos los elementos de $\mathbf{EE}'$ son cero, la traza de $\mathbf{EE}'$ es igual a la suma de los cuadrados de $\mathbf{E}$. Entonces

$$\mathbf{E} = \mathbf{0} \tag{1.3.7}$$

Como $\mathbf{\Lambda^2}$ es una matriz diagonal (1.3.5) indica que las filas de $\mathbf{F}$ son ortogonales y que la suma de los cuadrados de sus elementos fila es igual al correspondiente valor de la diagonal de $\mathbf{\Lambda^2}$. Esto significa que los elementos de la diagonal de $\mathbf{\Lambda^2}$ son no negativos y que existe una matriz desconocida $\mathbf{V}'$, tal que:

$$\mathbf{\Lambda V}' = \mathbf{F} \tag{1.3.8}$$

Sustituyendo (1.3.7) y (1.3.8) en (3.2.2), se tiene que:

$$\mathbf{U}'\mathbf{X} = \begin{bmatrix} \mathbf{\Lambda V}' \\ \mathbf{0} \end{bmatrix} \tag{1.3.9}$$

esta ecuación se puede escribir como:

$$\mathbf{U}'\mathbf{X} = \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{0} \end{bmatrix} \mathbf{V}' \tag{1.3.10}$$

del hecho que $\mathbf{V}'$ sea ortogonal, se tiene que:

$$\mathbf{U}'\mathbf{XV} = \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{0} \end{bmatrix} \tag{1.3.11}$$

quedando demostrado el teorema.

Para identificar más claramente las matrices $\mathbf{U}$, $\mathbf{V}$ y $\mathbf{\Lambda}$, se premultiplica (1.3.11) por su traspuesta:

$$\mathbf{V}'\mathbf{X}'\mathbf{U}\mathbf{U}'\mathbf{XV} = \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0}' \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{0} \end{bmatrix} = \mathbf{\Lambda}^2 \tag{1.3.12}$$

Sustituyendo $\mathbf{U}\mathbf{U}' = \mathbf{I}$,

$$\mathbf{V}'\mathbf{X}'\mathbf{XV} = \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0}' \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{0} \end{bmatrix} = \mathbf{\Lambda}^2 \tag{1.3.13}$$

Porque inicialmente se especificó que las columnas de $\mathbf{U}$ contienen los vectores propios de $\mathbf{XX}'$ y los elementos de la diagonal de $\mathbf{\Lambda}^2$ sus valores propios.

La ecuación (1.3.13) muestra que las columnas de $\mathbf{V}$ contienen los vectores propios de $\mathbf{X}^{'}\mathbf{X}$ y que sus valores propios también están sobre la diagonal de $\mathbf{\Lambda}^2$. Así se muestra que los valores propios de $\mathbf{X}\mathbf{X}^{'}$ y de $\mathbf{X}^{'}\mathbf{X}$ son los mismos, excepto cuando $I > J$ que $\mathbf{X}\mathbf{X}^{'}$ tiene $I-J$ valores propios nulos, sin embargo los valores propios no nulos son los mismos para las dos matrices.

La aplicación importante se deriva del hecho que en (1.3.11) premultiplicando por $\mathbf{U}$ y posmultiplicando por $\mathbf{V}^{'}$, se tiene:

$$\mathbf{X} = \mathbf{U}\begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{0} \end{bmatrix}\mathbf{V}^{'} \tag{1.3.14}$$

Es conveniente anotar que los elementos de la matriz $\mathbf{\Lambda}$ se encuentran ubicados en orden decreciente y corresponden con los vectores de las matrices $\mathbf{U}$ y $\mathbf{V}^{'}$

Ahora se define:

- $\mathbf{\Lambda}_{(r)}$ matriz de orden $r \times r$ que contiene los primeros r elementos de $\mathbf{\Lambda}$
- $\mathbf{U}_{(r)}$ submatriz de orden $I \times r$ que contiene las r primeras columnas de $\mathbf{U}$
- $\mathbf{V}^{'}_{(r)}$ submatriz de orden $r \times J$ que contiene las r primeras filas de $\mathbf{V}^{'}$

En el producto de (1.3.14), si solamente r elementos de la diagonal de $\mathbf{\Lambda}$ son no nulos, entonces

$$\begin{aligned} \hat{\mathbf{X}} &= \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{V}^{'}_{(r)} \\ &= \sum_{k=1}^{r} \lambda_k \mathbf{u}_k \mathbf{v}_k^{'} \end{aligned} \tag{1.3.15}$$

Si s el rango de $\mathbf{X}$ es mayor que r, Eckart and Young (1936) mostraron que el producto de la derecha en 1.3.15 es la aproximación de $\mathbf{X}$ por mínimos cuadrados en rango r.
Denotando $\mathbf{U}_{(s-r)}$, $\mathbf{\Lambda}_{(s-r)}$ y $\mathbf{V}_{(s-r)}{}'$ el resto de las filas ó columnas de las respectivas matrices, se tiene que:

$$\mathbf{X} = \begin{bmatrix} \mathbf{U}_{(r)} & \mathbf{U}_{(s-r)} \end{bmatrix}\begin{bmatrix} \mathbf{\Lambda}_{(r)} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{(s-r)} \end{bmatrix}\begin{bmatrix} \mathbf{V}^{'}_{(r)} \\ \mathbf{V}_{(s-r)} \end{bmatrix} \tag{1.3.16}$$

que es equivalente a :

$$\begin{aligned}\mathbf{X} &= \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{V}_{(r)}^{'} + \mathbf{U}_{(s-r)}\mathbf{\Lambda}_{(s-r)}\mathbf{V}_{(s-r)}^{'} \\ &= \sum_{k=1}^{r}\lambda_k\mathbf{u}_k\mathbf{v}_k^{'} + \sum_{k=r+1}^{s}\lambda_k\mathbf{u}_k\mathbf{v}_k^{'} \\ &= \hat{\mathbf{X}} + \mathbf{E}\end{aligned} \tag{1.3.17}$$

Luego $\hat{\mathbf{X}} = \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{V}_{(r)}^{'}$ es la mejor aproximación en rango r para $\mathbf{X}$

Keller (1962) mostró que la aproximación está dada por

$$\hat{\mathbf{X}} = \mathbf{X}\mathbf{V}_{(r)}\mathbf{V}_{(r)}^{'} \tag{1.3.18}$$

A partir de

$$\begin{aligned}\mathbf{X} &= \hat{\mathbf{X}} + \mathbf{E} \\ \mathbf{X} &= \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{V}_{(r)}^{'} + \mathbf{U}_{(s-r)}\mathbf{\Lambda}_{(s-r)}\mathbf{V}_{(s-r)}^{'}\end{aligned} \tag{1.3.19}$$

Se sabe que $\mathbf{V}_{(s-r)}^{'}\mathbf{V}_{(r)} = 0$ y postmultiplicando 1.3.19 por $\mathbf{V}_{(r)}\mathbf{V}_{(r)}^{'}$

$$\mathbf{X}\mathbf{V}_{(r)}\mathbf{V}_{(r)}^{'} = \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{V}_{(r)}^{'}\mathbf{V}_{(r)}\mathbf{V}_{(r)}^{'}$$

$$\mathbf{X}\mathbf{V}_{(r)}\mathbf{V}_{(r)}^{'} = \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{V}_{(r)}^{'}$$

La matriz *'error'* de la aproximación, es:

$$\begin{aligned}\mathbf{E} &= \mathbf{U}_{(s-r)}\mathbf{\Lambda}_{(s-r)}\mathbf{V}_{(s-r)}^{'} \\ &= \sum_{k=r+1}^{s}\lambda_k\mathbf{u}_k\mathbf{v}_k^{'}\end{aligned} \tag{1.3.20}$$

La suma de los cuadrados de los errores está dada por:

$$\begin{aligned}traza\left(\mathbf{E}^{'}\mathbf{E}\right) &= traza\left(\mathbf{V}_{(s-r)}\mathbf{\Lambda}_{(s-r)}\mathbf{U}_{(s-r)}^{'}\mathbf{U}_{(s-r)}\mathbf{\Lambda}_{(s-r)}\mathbf{V}_{(s-r)}^{'}\right) \\ &= traza\left(\mathbf{V}_{(s-r)}\mathbf{\Lambda}_{(s-r)}^{2}\mathbf{V}_{(s-r)}^{'}\right) \\ &= traza\left(\mathbf{\Lambda}_{(s-r)}^{2}\mathbf{V}_{(s-r)}\mathbf{V}_{(s-r)}^{'}\right) \\ &= traza\left(\mathbf{\Lambda}_{(s-r)}^{2}\right)\end{aligned} \tag{1.3.21}$$

Además la inercia de la matriz $\mathbf{X}$ corresponde a la suma de los cuadrados de sus elementos, que es el mismo valor que la traza de $\mathbf{X}\mathbf{X}^{'}$ y corresponde

a la suma de los valores propios λ^2, por tanto, la suma de los cuadrados de los elementos de $\mathbf{X}$ de $\hat{\mathbf{X}}$ y de $\mathbf{E} = \mathbf{X} - \mathbf{X}_{(r)}$, son respectivamente:

$$\sum_{k=1}^{s} \lambda_k^2, \qquad \sum_{k=1}^{r} \lambda_k^2 \quad y \quad \sum_{k=r+1}^{s} \lambda_k^2.$$

Una medida de la calidad de representación o inercia de $\mathbf{X}$ por medio de $\hat{\mathbf{X}}$ esta dada por:

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

Esta medida también se denomina *inercia retenida* ó *varianza absorbida.*

1.3.2. Biplots Clásicos

La forma para elegir los marcadores fila y columna para la representación de la matriz $\hat{\mathbf{X}}$, se puede hacer de muchas maneras, por tal razón varios autores proponen diversas soluciones y dan sus respectivas propiedades. Presentamos la descripción y propiedades de los biplots clásicos propuestos por GABRIEL (1971).

De manera general, los marcadores se eligen a partir de la descomposición:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}^{\gamma} \qquad \mathbf{B} = \mathbf{V}\mathbf{\Lambda}^{(1-\gamma)}$$

Con $\gamma = 1$, se obtiene

$$\mathbf{A}_{(r)} = \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)} \qquad \mathbf{B}_{(r)} = \mathbf{V}_{(r)}$$

$\mathbf{A}_{(r)}$ es la matriz de *componentes principales*, $\mathbf{B}_{(r)}$ la matriz de *coordenadas estándar* y se verifica que $\mathbf{B}_{(r)}^{'}\mathbf{B}_{(r)} = \mathbf{I}_{(r)}$, se denomina RMP-Biplot (Row Metric Preserving).

Con $\gamma = 0$, se obtiene

$$\mathbf{A}_{(r)} = \mathbf{U}_{(r)} \qquad \mathbf{B}_{(r)} = \mathbf{V}_{(r)}\mathbf{\Lambda}_{(r)}$$

$\mathbf{A}_{(r)}$ es la matriz de *coordenadas estándar*, $\mathbf{B}_{(r)}$ la matriz de *coordenadas principales* y se verifica que $\mathbf{A}_{(r)}^{'}\mathbf{A}_{(r)} = \mathbf{I}_{(r)}$, se denomina CMP-Biplot (Column Metric Preserving).

Con $\gamma = 1/2$, se obtiene

$$\mathbf{A}_{(r)} = \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}^{1/2} \qquad \mathbf{B}_{(r)} = \mathbf{V}_{(r)}\mathbf{\Lambda}_{(r)}^{1/2}$$

Se verifica que $\mathbf{A}_{(r)}^{'}\mathbf{A}_{(r)} = \mathbf{I}_{(r)}$ y $\mathbf{B}_{(r)}^{'}\mathbf{B}_{(r)} = \mathbf{I}_{(r)}$; se denomina SQRT-Biplot o Biplot simétrico.

1.3.3. Biplot para análisis factorial

Si la matriz $\mathbf{X}$ está centrada por columnas, la matriz $\mathbf{X}^{'}\mathbf{X}$ es proporcional a la matriz de covarianzas $\mathbf{S} = \mathbf{X}^{'}\mathbf{X}/(\text{n-1})$. GABRIEL (1971) propone los biplots de componentes principales, en los que ajusta los marcadores para que estén directamente relacionados con el Análisis de Componentes Principales como solución del Análisis Factorial.

Los marcadores para el CMP-Biplot están dados por:

$$\mathbf{A} = \sqrt{\mathbf{n-1}}\mathbf{U} \qquad \mathbf{B} = \frac{\mathbf{1}}{\sqrt{\mathbf{n-1}}}\mathbf{V\Lambda}$$

En dimensión completa, el producto de los $\langle \mathbf{b}_i, \mathbf{b}_j \rangle$ marcadores de las columnas, es igual a los productos escalares de las columnas de $\mathbf{X}$ y corresponden a las varianzas y covarianzas.

1.3.4. Biplots Generales

En un contexto más general, para obtener el biplot de $\mathbf{X}$ que proporcione los productos internos de sus filas con la métrica $\mathbf{\Phi}$ y los productos internos de sus columnas con la métrica $\mathbf{\Psi}$, se usa la descomposición en valores singulares generalizada (DVSG) de $\mathbf{X}$, definida como sigue: sean $\mathbf{\Psi}$ y $\mathbf{\Phi}$ matrices $I \times I$ y $J \times J$ respectivamente, definidas positivas. Considérese la DVS de $\mathbf{\Psi^{1/2}X\Phi^{1/2}}$:

$$\mathbf{\Psi^{1/2}X\Phi^{1/2}} = \mathbf{P\Lambda Q'}$$

Donde $\mathbf{P'P} = \mathbf{Q'Q} = \mathbf{I}$ y $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_s)$ con $\lambda_1 >, \ldots, \lambda_s$.

Ahora, sean: $\mathbf{R} = \mathbf{\Psi^{-1/2}P}$ y $\mathbf{T} = \mathbf{\Phi^{-1/2}Q}$ entonces $\mathbf{X} = \mathbf{R\Lambda T'}$

Donde $\mathbf{R}'\mathbf{\Psi}\mathbf{R} = \mathbf{T}'\mathbf{\Phi}\mathbf{T} = \mathbf{I}$. Esto significa que tanto las columnas de $\mathbf{R}$ como las de $\mathbf{T}$ son ortogonales con las métricas $\mathbf{\Psi}$ y $\mathbf{\Phi}$, respectivamente.

La expresión $\mathbf{X} = \mathbf{R}\mathbf{\Lambda}\mathbf{T}'$ define la DVSG de $\mathbf{X}$ con la métrica $\mathbf{\Phi}$ para las filas y la métrica $\mathbf{\Psi}$ para las columnas.

La generalización del teorema de aproximación de rango inferior r, es:

$$\hat{\mathbf{X}} = \mathbf{R}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{T}'_{(r)}$$

Que corresponde a la aproximación por mínimos cuadrados generalizados de $\mathbf{X}$ en el sentido que minimiza $traza(\mathbf{\Psi}(\mathbf{X} - \hat{\mathbf{X}})\mathbf{\Phi}(\mathbf{X} - \hat{\mathbf{X}})')$ entre todas las matrices $\hat{\mathbf{X}}$ de rango r.

La medida de la calidad de representación de $\mathbf{X}$ por medio de $\hat{\mathbf{X}}$ esta dada por:

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

Entonces, para construir el biplot de $\mathbf{X}$ con la métrica $\mathbf{\Phi}$ para las filas y la métrica $\mathbf{\Psi}$ para las columnas, se definen las matrices:

$$\mathbf{A} = \mathbf{R}\mathbf{\Lambda}^{\alpha} \quad y \quad \mathbf{B} = \mathbf{T}\mathbf{\Lambda}^{\mathbf{1}-\alpha}$$

y obtenemos el respectivo biplot de $\mathbf{X}$. En la presentación anterior, se ha seguido la idea de Greenacre (1984), pág. 346

1.3.5. Propiedades de los Biplots

Con el fin de notar la diferencia entre los biplots RMP y CMP, vamos a asumir que las métricas $\mathbf{\Phi}$ y $\mathbf{\Psi}$ son la matriz identidad.

CMP Biplot

Se supone que los datos están centrados; los marcadores filas o coordenadas estándar y los marcadores columnas o coordenadas principales, en dimensión r, son:

$$\mathbf{A} = \sqrt{\mathbf{n-1}}\mathbf{U}_{(r)} \qquad \mathbf{B} = \frac{\mathbf{1}}{\sqrt{\mathbf{n-1}}}\mathbf{V}_{(r)}\mathbf{\Lambda}_{(r)}$$

La matriz $\mathbf{A}$ contiene las coordenadas estándar y la matriz $\mathbf{B}$ las coordenadas principales. Se tiene:

- Para $\mathbf{S} = \frac{\mathbf{1}}{(\mathbf{n-1})}\mathbf{X}'\mathbf{X}$ es decir, que el producto escalar de las variables reducidas coincide con el producto escalar de los marcadores columna, de donde $\mathbf{S} = \frac{\mathbf{1}}{(\mathbf{n-1})}\mathbf{V}\mathbf{\Lambda}\mathbf{\Lambda}\mathbf{V}'$.

  Por tanto, la mejor aproximación $\hat{\mathbf{S}}$ de la matriz de covarianzas $\mathbf{S}$, en rango reducido r es:

  $$\mathbf{S} \simeq \hat{\mathbf{S}} = \frac{\mathbf{1}}{(\mathbf{n-1})}\mathbf{V}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{V'}_{(r)} = \mathbf{B}\mathbf{B}^{'}$$

  y se obtiene con el biplot de $\mathbf{X}$

  – Para $\mathbf{b}_j$ vector de $\mathbf{B}$, se tiene que: $\|\mathbf{b}_j\|^2 \simeq var(\mathbf{X}_j)$

  – Como $\|\mathbf{b}_j\|^2 = \mathbf{b}_j^{'} * \mathbf{b}_j$, entonces $\mathbf{b}_j \simeq dest(\mathbf{X}_j)$.

  Por tanto, la longitud de los marcadores columna corresponde a la desviación típica de la respectiva variable. Así, para un punto fila que se proyecte exactamente sobre el punto columna, se tendría una predicción igual a la desviación típica de la variable correspondiente.

  – El coseno del ángulo que forman dos marcadores columna $\mathbf{b}_k$ y $\mathbf{b}_j$ aproxima la correlación entre las variables:

  $$\cos(\mathbf{b}_k, \mathbf{b}_j) = \frac{\langle \mathbf{b}_k, \mathbf{b}_j \rangle}{\|\mathbf{b}_k\|\|\mathbf{b}_j\|} \simeq \frac{s_{kj}}{s_k s_j}$$

- La distancia de Mahalanobis entre dos filas de $\mathbf{X}$ coincide con la distancia euclídea entre sus respectivos marcadores fila. En dimensión reducida se consigue entonces, una aproximación de la distancia de Mahalanobis, que se escribe:

  $$\mathbf{X}\mathbf{S}^{-\mathbf{1}}\mathbf{X}^{'} \simeq \mathbf{A}\mathbf{A}^{'}$$

  Como $\mathbf{S} \simeq \mathbf{B}\mathbf{B}^{'}$ entonces $\mathbf{X}\mathbf{S}^{-\mathbf{1}}\mathbf{X}^{'} \simeq \mathbf{A}\mathbf{B}^{'}\mathbf{S}^{-\mathbf{1}}\mathbf{B}\mathbf{A}^{'} = \mathbf{A}\mathbf{A}^{'}$

- El CMP biplot proporciona una mejor aproximación para $\mathbf{S}$ matriz de varianzas y covarianzas, que para $\mathbf{XS^{-1}X}^{'}$ producto escalar con la métrica de Mahalanobis.

  La calidad global de representación de la matriz $\mathbf{X}$ en rango reducido $\mathbf{X}_{(r)}$, es:

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

  Como $\mathbf{S} \simeq \mathbf{BB}^{'} = \frac{\mathbf{1}}{\mathbf{(n-1)}} \mathbf{V}_{(r)} \mathbf{\Lambda}_{(r)} \mathbf{\Lambda}_{(r)} \mathbf{V}^{'}_{(r)}$ tenemos que su bondad de ajuste es:

  $$\sum_{k=1}^{r} \lambda_k^4 / \sum_{k=1}^{s} \lambda_k^4$$

  Que es un valor mayor que el obtenido para la matriz original.

- Para las filas de la matriz $\mathbf{X}$ la situación es diferente. La suma de los cuadrados de las filas de $\mathbf{XS^{-1}X}^{'}$ es s, rango de $\mathbf{X}$. Si aproximamos en dimensión r:

  $$\mathbf{XS^{-1}X}' \simeq \mathbf{AA}^{'}$$

  La suma de los cuadrados de las filas de $\mathbf{AA}^{'}$ es r, y su bondad de ajuste es r/s, valor menor que para las columnas.

  Luego la bondad de ajuste para la aproximación de los productos escalares con la métrica de Mahalanobis, es menor que para la matriz de varianzas covariancias y que para la matriz original. Por tal razón se denomina biplot que preserva la métrica de las columnas.

RMP Biplot

Se supone que los datos están centrados; los marcadores filas o *componentes principales* y los marcadores columnas o *cordenadas estandar*, en dimensión r son:

$$\mathbf{A} = \mathbf{U}_{(r)} \mathbf{\Lambda}_{(r)} \qquad \mathbf{B} = \mathbf{V}_{(r)}$$

La matriz $\mathbf{A}$ contiene las coordenadas sobre las componentes principales y la matriz $\mathbf{B}$ las coordenadas estándar.

Las propiedades son:

- Los productos escalares de las filas de la matriz $\mathbf{X}$, con la métrica euclidea, coinciden con los productos escalares de sus respectivos marcadores en el espacio completo. La aproximación de dichos productos escalares en dimensión reducida es:

$$\mathbf{X}\mathbf{X}^{'} \simeq \mathbf{A}\mathbf{B}^{'}\mathbf{B}\mathbf{A}^{'} = \mathbf{A}\mathbf{A}^{'}$$

- Los marcadores para las filas coinciden con las coordenadas de los individuos en el espacio de las componentes principales. Teniendo en cuenta que $\mathbf{V}$ es la matriz que contiene los vectores propios de la matriz de covarianzas, entonces las coordenadas sobre las r primeras componentes principales, son:

$$\mathbf{X}\mathbf{V}_{(r)} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{'}\mathbf{V}_{(r)} = \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)} = \mathbf{A}$$

  Implica que se pueden estudiar las diferencias y similitudes entre los individuos con pérdida de información mínima, si la distancia euclidea es la apropiada en la matriz original.

- Las coordenadas para las variables son las proyecciones de los ejes originales en el espacio de las componentes principales. Teniendo en cuenta que las coordenadas de los vectores que forman la base canónica se encuentran en la matriz $\mathbf{I_J}$, la proyección de las mismas sobre las componentes principales, es:

$$\mathbf{I_J}\mathbf{V}_{(r)} = \mathbf{V}_{(r)} = \mathbf{B}$$

  Así, las coordenadas para las columnas marcan la unidad para las escalas de interpolación.

- La similitud entre las columnas se aproxima utilizando como métrica la inversa de la matriz de dispersión entre los individuos. No se hace la demostración de esta propiedad ya que las distancias no presentan un interés práctico, además no es posible interpretar los ángulos en términos de correlaciones.

- La calidad de representación es mejor para las filas que para las columnas.

$$\mathbf{X}\mathbf{X}^{'} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{'}\mathbf{V}\mathbf{\Lambda}\mathbf{U}^{'} \simeq \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{U}^{'}_{(r)}$$

  Por tanto, la calidad global de representación de las filas de la matriz $\mathbf{X}$ en rango reducido $\mathbf{X}_{(r)}$, es:

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

La calidad de representación de las columnas se calcula mediante:

$$\mathbf{V}_{(r)}^{'}\mathbf{V}_{(r)} = \mathbf{I}_{(r)}$$

y $traza(\mathbf{V}_{(r)}^{'}\mathbf{V}_{(r)}) = r$, así la calidad de representación de las columnas de $\mathbf{X}$ es r/s, valor menor que la calidad de representación de sus filas.

- La calidad de aproximación para las filas de la matriz $\hat{\mathbf{X}}$ por los marcadores para las filas, viene dada por:

$$\sum_{k=1}^{r} \lambda_k^4 / \sum_{k=1}^{s} \lambda_k^4$$

1.3.6. Biplots y regresión

En esta sección se muestra como lograr el biplot a partir de una regresión multivariante, posteriormente con el planteamiento de modelos se encuentran el CPM y el RPM biplot y finalmente se considera un modelo bilineal donde se utilizan regresiones alternadas para resolverlo y cuya solución genera el biplot.

1.3.6.1. Regresión Multivariante

En términos generales vamos a considerar el modelo definido por:

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

Donde $\mathbf{X}_{(I\times J)}$ es una matriz de J variables observadas para cada uno de los I individuos, $\mathbf{A}_{(I\times s)}$ es una matriz conocida y $\mathbf{B}_{(s\times J)}$ es una matriz desconocida de parámetros de la regresión. $\mathbf{E}$ es una matriz no observada de errores aleatorios, incorrelada con las filas de la matriz $\mathbf{X}$, con media cero y matriz de covarianza $\mathbf{S}$.

Cuando $\mathbf{A}$ representa una matriz de s variables independientes observadas para cada uno de los I individuos, la ecuación es llamada modelo de regresión multivariante. Las columnas de $\mathbf{X}$ representan las variables dependientes las cuales son explicadas en términos de las variables independientes dadas por las columnas de $\mathbf{A}$. Si $\mathbf{A}$ es una matriz de variables aleatorias los estimadores y valores esperados son interpretados como condicionales de $\mathbf{A}$.

Los estimadores para esta regresión se derivan de la regresión múltiple.

Las columnas de $\mathbf{X}$ representan las variables dependientes las cuales son explicadas en términos de las variables independientes de $\mathbf{A}$, entonces:

$$\mathbf{E}(\mathbf{X}) = \mathbf{AB}$$

Asumiendo que las variables de $\mathbf{A}$ están centradas y que $I > J$. Para que el estimador de $\mathbf{B}$ sea único se debe suponer que $\mathbf{A}$ es de rango completo s, así la inversa de $(\mathbf{A}^{'}\mathbf{A})^{-\mathbf{1}}$ existe.

Sea $\mathbf{P} = \mathbf{I} - \mathbf{A}(\mathbf{A}^{'}\mathbf{A})^{-\mathbf{1}}\mathbf{A}^{'}$ de orden $I \times I$, simétrica e idempotente, de rango $I - s$, proyector sobre las columnas de $\mathbf{R}^{\mathbf{I}}$, ortogonal a las columnas de $\mathbf{A}$ entonces se tiene que $\mathbf{PA} = \mathbf{0}$.

Los estimadores de máxima verosimilitud de $\mathbf{B}$ y $\mathbf{S}$, están dados por:

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

$$\widehat{\mathbf{S}} = \tfrac{1}{I}\mathbf{X}'\mathbf{P}\mathbf{X}$$

Donde $\mathbf{A}^{'}\mathbf{A}$ es la matriz de covarianzas entre las variables independientes y $\mathbf{A}'\mathbf{X}$ la covarianza entre las variables independientes y las dependientes.

En regresión multivariante los estimadores OLS y GLS son los mismos y coinciden para $\mathbf{B}$.
Por tanto:

- El modelo ajustado se escribe $\widehat{\mathbf{X}} = \mathbf{A}\widehat{\mathbf{B}} = \mathbf{A}(\mathbf{A}^{'}\mathbf{A})^{-\mathbf{1}}\mathbf{A}^{'}\mathbf{X}$, de donde

  $$\mathcal{H} = \mathbf{A}(\mathbf{A}^{'}\mathbf{A})^{-\mathbf{1}}\mathbf{A}^{'}$$

  denominada *Hat Matriz* de orden $I \times I$, hace la transformación de $\mathbf{X}$ a $\widehat{\mathbf{X}}$

- La covarianza entre una variable independiente $\mathbf{a}_s$ y una variable dependiente $\mathbf{x}_j$ es:

  $$E(\mathbf{a}_s^{'}\mathbf{x}_j) = E(\mathbf{a}_s^{'}\mathbf{a}_s\mathbf{B}_j) = E(\mathbf{a}_s^{'}\mathbf{a}_s)\mathbf{B}_j$$

- El cuadrado del coeficiente de correlación múltiple entre una variable $\mathbf{x}$ y la matriz $\mathbf{A}$, suponiendo que $\mathbf{A}$ está centrada, se define a partir de:

$$\begin{bmatrix} s_{11} & \mathbf{s_{12}} \\ \mathbf{s_{21}} & \mathbf{S_{22}} \end{bmatrix} = \frac{1}{I}\begin{bmatrix} \mathbf{x}' \\ \mathbf{A}' \end{bmatrix}\begin{bmatrix} \mathbf{x} & \mathbf{A} \end{bmatrix} = \frac{1}{I}\begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{A} \\ \mathbf{A}'\mathbf{x} & \mathbf{A}'\mathbf{A} \end{bmatrix}$$

Por tanto:

$$\widehat{\mathbf{B}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{X} = \mathbf{S_{22}^{-1}s_{21}}$$

Luego, la correlación entre $\mathbf{x}$ y $\mathbf{A}\widehat{\mathbf{B}}$ está dada por:

$$cor = (\mathbf{x}, \mathbf{A}\widehat{\mathbf{B}}) = \frac{\mathbf{s_{12}S_{22}^{-1}s_{21}}}{\left[\mathbf{x}'\mathbf{x}\ \mathbf{s_{12}S_{22}^{-1}s_{21}}\right]^{1/2}}$$

Que es equivalente a:

$$cor^2 = (\mathbf{x}, \mathbf{A}\widehat{\mathbf{B}}) = \frac{\left[\mathbf{s_{12}S_{22}^{-1}s_{21}}\right]^2}{\mathbf{x}'\mathbf{x}\left[\mathbf{s_{12}S_{22}^{-1}s_{21}}\right]}$$

Y también:

$$cor^2 = (\mathbf{x}, \mathbf{A}\widehat{\mathbf{B}}) = \frac{\mathbf{s_{12}S_{22}^{-1}s_{21}}}{\mathbf{x}'\mathbf{x}}$$

Es decir, que:

$$cor^2 = (\mathbf{x}, \mathbf{A}\widehat{\mathbf{B}}) = \frac{\left[\widehat{\mathbf{B}}'\mathbf{s_{21}}\right]^2}{\mathbf{x}'\mathbf{x}\ \widehat{\mathbf{B}}'\mathbf{S_{22}^{-1}}\widehat{\mathbf{B}}}$$

- La medida de correlación entre vectores que representa la proporción de varianza explicada por el modelo, puede ser obtenida a partir de la matriz $\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1}\widehat{\mathbf{E}}'\widehat{\mathbf{E}}$. Donde $\widehat{\mathbf{E}}'\widehat{\mathbf{E}} = \mathbf{X}'\mathbf{P}\mathbf{X}$ que toma valores entre cero, cuando toda la variación de $\mathbf{X}$ es explicada por el modelo y $\mathbf{X}'\mathbf{X}$ cuando el modelo no explica nada de la variación de $\mathbf{X}$. Luego $\mathbf{I} - \mathbf{D}$ varía entre la matriz identidad y la matriz cero.

  Así las siguientes medidas de correlación multivariante varían entre cero y uno e involucran los coeficientes de la estimación.
  Hopper (1959, pp.249-250) propone:

  - *traza de correlación* $r_T^2 = \frac{1}{J}traza(\mathbf{I} - \mathbf{D})$

- *determinante de correlación* $r_D^2 = det\ (\mathbf{I} - \mathbf{D})$.

Hotelling (1936) sugiere el *vector coeficiente alienación* $r_A = det\ \mathbf{D}$

Además:

$$\widehat{\mathbf{E}}'\widehat{\mathbf{E}} = \mathbf{X}'\mathbf{P}\mathbf{X} = \mathbf{X}'(\mathbf{I} - \mathcal{H})\mathbf{X} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathcal{H}\mathbf{X}$$

De donde

$$\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathcal{H}\mathbf{X}) = \mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{H}\mathbf{X}$$

Luego los indicadores se pueden escribir como:

- $r_T^2 = \frac{1}{J} traza((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{H}\mathbf{X})$
- $r_D^2 = det((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{H}\mathbf{X})$
- $r_A = det(\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{H}\mathbf{X})$

Las medidas de correlación son invariantes bajo conmutación y se obtienen los mismos resultados si $\mathbf{D} = \widehat{\mathbf{E}}'\widehat{\mathbf{E}}(\mathbf{X}'\mathbf{X})^{-1}$.

- La matriz de covarianzas de $\widehat{\mathbf{B}}$ es $\widehat{\mathbf{B}}'\widehat{\mathbf{B}}$.

1.3.6.2. Biplots de regresión

Aplicando los conceptos de la regresión o modelo lineal multivariante, podemos entender el biplot como un modelo bilineal multivariante. El término bilineal procede del hecho que fijando $\mathbf{A}$, $\mathbf{B}$ se obtiene de un modelo lineal y viceversa.

A partir de la aproximación de la matriz $\mathbf{X}$ por medio de la matriz $\hat{\mathbf{X}} = \mathbf{A}\mathbf{B}'$ en rango reducido r, se tiene que:

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

donde:

$$\widehat{\mathbf{S}} = \frac{1}{I}\mathbf{X}'\mathbf{P}\mathbf{X}$$

Además, las matrices $\mathbf{A}$ y $\mathbf{B}$ son de rango completo. Por tanto, se pueden considerar dos modelos:

1. Conociendo los valores de la matriz $\mathbf{A}$, se puede estimar $\mathbf{B}$, a partir de:

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

2. Conociendo los valores de la matriz $\mathbf{B}$, se puede estimar $\mathbf{A}$

En este caso se parte del modelo:

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

Y el estimador es:

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

- *Individuos y variables externas*
  Tanto los individuos como las variables que no intervienen en el análisis se denominan individuos externos o suplementarios y variables externas o suplementarias y en ocasiones es necesario obtener su representación en el biplot.

  Usando las ecuaciones de regresión en el modelo bilineal es posible proyectar en una representación biplot tanto individuos como variable suplementarias.

  Si $\mathbf{x}_n$ es un vector que contiene las medidas de las J variables para un nuevo individuo, buscamos sus coordenadas $\mathbf{a}_n$ sobre el biplot de forma que sus valores originales se aproximen lo mejor posible en el biplot. Considerando $\mathbf{B}$ fijo:

  $$\mathbf{x'}_n = \mathbf{a'}_n\mathbf{B'} + \mathbf{e'}_n \qquad \text{ó} \qquad \mathbf{x}_n = \mathbf{B}\mathbf{a}_n + \mathbf{e}_n$$

  es decir, la suma de cuadrados de los residuales ha de ser mínima. La solución es inmediata:

  $$\mathbf{a}_n = (\mathbf{B'B})^{-1}\mathbf{B'x}_n$$

  La bondad de ajuste o calidad de representación de la aproximación viene dada por el coeficiente de determinación de la regresión.

  De la misma forma, si $\mathbf{y}_n$ es un vector que contiene las medidas para los I individuos en una nueva variable, buscamos las coordenadas $\mathbf{b}_n$ de la nueva variable sobre el biplot de forma que los correspondientes productos escalares aproximen lo mejor posible los valores en el biplot, considerando $\mathbf{A}$ fijo:

  $$\mathbf{y}_n = \mathbf{A}\mathbf{b}_n + \mathbf{e}_n$$

procediendo como antes

$$\mathbf{b}_n = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y}_n$$

y el coeficiente de determinación de la regresión puede utilizarse como una medida de la calidad de representación

1.3.6.3. Regresiones alternadas

Se parte del hecho que una matriz $\mathbf{X}$ se puede descomponer en el producto de dos matrices $\mathbf{A}$ y $\mathbf{B}$, sin embargo no se conoce ninguna de esta matrices. Luego tenemos el modelo bilineal

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

donde el propósito es encontrar las dos matrices $\mathbf{A}$ y $\mathbf{B}$

La demostración debida a Gabriel (1995a) presenta el algoritmo en dimensión 1 y obtiene el primer valor y vector singulares, a partir del ajuste por mínimos cuadrados de regresiones alternadas de los parámetros $\widehat{\mathbf{A}}^{'}$ y $\widehat{\mathbf{B}}^{'}$

Cabe resaltar que el algoritmo converge pero puede hacerlo a un mínimo local. Es el caso, si la solución inicial está próxima a cualquiera de los vectores singulares, caso en que converge a dicho vector y no necesariamente al primero.

Ajuste para rango 1

A partir de cualesquier vectores columna $\mathbf{a}$ y $\mathbf{b}$ de tamaño Ix1 y Jx1, respectivamente, se consigue:

$$\mathbf{X} \simeq \mathbf{a}\mathbf{b}^{'}$$

El proceso se inicia con el modelo:

$$\mathbf{X} = \mathbf{a}\mathbf{b}^{'} + \mathbf{E}$$

De donde

$$\widehat{\mathbf{b}}^{'} = (\mathbf{a}^{'}\mathbf{a})^{-\mathbf{1}}\mathbf{a}^{'}\mathbf{X}$$

Posteriomente se tiene en cuenta que:

$$\widehat{\mathbf{a}}^{'} = (\mathbf{b}^{'}\mathbf{b})^{-\mathbf{1}}\mathbf{b}^{'}\mathbf{X}^{'}$$

El vector $\mathbf{a}$ se puede expresar como combinación lineal de las columnas de $\mathbf{U}$, matriz de la DSV de $\mathbf{X}$

$$\mathbf{a} = \sum_{\mathbf{k=1}}^{\mathbf{s}} \alpha_{\mathbf{k}} \mathbf{u_k}$$

Entonces

$$\mathbf{a}^{'}\mathbf{a} = \sum_{\mathbf{k=1}}^{\mathbf{s}} \alpha_{\mathbf{k}}^{\mathbf{2}}$$

Luego $\widehat{\mathbf{b}}'$, es:

$$\widehat{\mathbf{b}}' = (\mathbf{a}^{'}\mathbf{a})^{-\mathbf{1}}\mathbf{a}^{'}\mathbf{X} = \frac{1}{\sum_{k=1}^{s} \alpha_k^2} \sum_{k=1}^{s} \alpha_k \mathbf{u}_k^{'} \sum_{k=1}^{s} \lambda_k \mathbf{u}_k \mathbf{v}_k^{'}$$

$$= \frac{1}{\sum_{k=1}^{s} \alpha_k^2} \sum_{k=1}^{s} \alpha_k \lambda_k \mathbf{v}_k^{'}$$

De lo anterior se deduce que el vector $\widehat{\mathbf{b}}'$ es combinación lineal de las columnas de $\mathbf{V}$ con coeficientes $\alpha_k \lambda_k$, salvo el factor de escala $\sum_{k=1}^{s} \alpha_k^2$.

Ahora,

$$\widehat{\mathbf{b}}'\widehat{\mathbf{b}} = \frac{1}{\left[\sum_{k=1}^{s} \alpha_k^2\right]^2} \sum_{k=1}^{s} \alpha_k^2 \lambda_k^2$$

$$\widehat{\mathbf{b}}'\mathbf{X} = \frac{1}{\sum_{k=1}^{s} \alpha_k^2} \sum_{k=1}^{s} \alpha_k \lambda_k \mathbf{v}_k^{'} \sum_{k=1}^{s} \lambda_k \mathbf{v}_k \mathbf{u}_k^{'}$$

$$= \frac{1}{\sum_{k=1}^{s} \alpha_k^2} \sum_{k=1}^{s} \alpha_k \lambda_k^2 \mathbf{u}_k^{'}$$

Con está estimación de $\widehat{\mathbf{b}}'\mathbf{X}^{'}$ calculamos $\widehat{\mathbf{a}}'$

$$\widehat{\mathbf{a}}' = (\mathbf{b}^{'}\mathbf{b})^{-\mathbf{1}}\mathbf{b}^{'}\mathbf{X}^{'} = \frac{\left[\sum_{k=1}^{s} \alpha_k^2\right]^2}{\sum_{k=1}^{s} \alpha_k^2 \lambda_k^2} \frac{1}{\sum_{k=1}^{s} \alpha_k^2} \left[\sum_{k=1}^{s} \alpha_k \lambda_k^2 \mathbf{u}_k^{'}\right]$$

Comparando el valor inicial con el obtenido mediante el proceso, se tiene:

$$\mathbf{a} = \sum_{\mathbf{k=1}}^{\mathbf{s}} \alpha_{\mathbf{k}} \mathbf{u_k}$$

$$\widehat{\mathbf{a}} = p_1 \left[\sum_{\mathbf{k=1}}^{\mathbf{s}} \alpha_{\mathbf{k}} \lambda_{\mathbf{k}}^{\mathbf{2}} \mathbf{u_k} \right]$$

Para p_1 constante obtenida en la primera iteración.

Si el proceso itera nuevamente

$$\widehat{\mathbf{a}}^{(\mathbf{2})} = p_2 \left[\sum_{\mathbf{k=1}}^{\mathbf{s}} \alpha_{\mathbf{k}} \lambda_{\mathbf{k}}^{\mathbf{4}} \mathbf{u_k} \right]$$

Y en general las sucesivas estimaciones para $\widehat{\mathbf{a}}$ serán de la forma:

$$\widehat{\mathbf{a}}^{(\mathbf{q})} = p_q \left[\sum_{\mathbf{k=1}}^{\mathbf{s}} \alpha_{\mathbf{k}} \lambda_{\mathbf{k}}^{\mathbf{2q}} \mathbf{u_k} \right]$$

Como los valores singulares, en la descomposición de la matriz $\mathbf{X}$, son positivos y se encuentran ordenados, la iteración sucesiva de este procedimiento da mayor importancia al primer valor singular. Significa que el vector $\mathbf{a}$ que buscamos, es proporcional al primer vector singular.

Sea $\theta^{(q)}$ el ángulo que forman $\widehat{\mathbf{a}}^{(\mathbf{q})}$ y $\mathbf{u_1}$, se va a demostrar que este ángulo converge a cero; lo que permite asegurar que el proceso iterativo de la búsqueda de $\mathbf{a}$ y $\mathbf{b}$ es finito.

El ángulo se puede calcular por medio del coseno:

$$\cos\theta^{(q)} = \frac{\widehat{\mathbf{a}}^{(q)'}\mathbf{u}_1}{\left\|\widehat{\mathbf{a}}^{(q)}\right\| \|\mathbf{u}_1\|} = \frac{p_q}{p_q\sqrt{\sum_{k=1}^{s} \alpha_k^2 \lambda^4 q_k}} = \frac{1}{\sqrt{\sum_{k=1}^{s} \frac{\alpha_k^2 \lambda_k^{4q}}{\alpha_1^2 \lambda_1^{4q}}}}$$

$$= \frac{1}{\sqrt{1 + \sum_{k=1}^{s} \left(\frac{\alpha_k}{\alpha_1}\right)^2 \left(\frac{\alpha_k}{\alpha_1}\right)^{4q}}}$$

Expresión que converge a 1, por tanto el ángulo θ converge a cero.

Estimando **b** a partir del resultado encontrado para el vector **a**, se tiene:

$$\widehat{\mathbf{b}}' = (\mathbf{a}'\mathbf{a})^{-\mathbf{1}}\mathbf{a}'\mathbf{X} = \frac{1}{p^2}p\mathbf{u}_1\sum_{k=1}^{s}\lambda_k\mathbf{u}_k\mathbf{v}_k' = \frac{1}{p}\lambda_1\mathbf{v}_1'$$

Es decir, que **b** es proporcional al producto entre el primer valor y vector singular.

En resumen:

$$(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}) \rightarrow \left(p\mathbf{u}_1, \frac{1}{p}\lambda_1\mathbf{v}_1'\right)$$

Las constantes de proporcionalidad se pueden evitar normalizando en cada paso el vector **a**, es decir que $\sum \mathbf{a}_k^2 = 1$. También es posible hacer la normalización al final del proceso de tal forma que $\sum \mathbf{a}_k^2 = p$

Entonces:

$$\mathbf{b} = \frac{1}{p}\lambda_1\mathbf{v}_1' \qquad \lambda_1 = p\sqrt{\sum \mathbf{b}_j^2} \qquad \mathbf{v}_1 = \frac{b}{\sqrt{\sum \mathbf{b}_j^2}}$$

Ajuste para rango 2

La demostración debida a Blázquez (1988) se realiza siguiendo el proceso descrito para el ajuste de rango 1.

A partir del modelo

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

Con **A** y **B** matrices de rango 2. Sean:

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2) \qquad \mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2)$$

tales que las columnas de **A** y **B** son linealmente independientes.

Se trata de demostrar que los vectores obtenidos a partir de cualquier par de matrices generan el mismo subespacio que los dos primeros vectores singulares, o que los vectores que forman las columnas de **A** son combinaciones lineales de los dos primeros valores singulares, por la izquierda (es decir, de **U**), mientras los vectores que forman las columnas de **B** son combinaciones lineales de los dos primeros vectores singulares por la derecha (es decir, de **V**).

De manera análoga al caso anterior, cada vector columna de **A** se puede escribir como combinación lineal de las columnas de **U** porque estas forman

una base ortonormal del espacio I dimensional.
Entonces:

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2) = \left[\sum_{k=1}^{s} \alpha_k \mathbf{u}_k, \sum_{k=1}^{s} \beta_k \mathbf{u}_k \right]$$

Supongamos que $\mathbf{A}$ es conocido y vamos a estimar $\mathbf{B}$. Iniciando con:

$$\mathbf{A}'\mathbf{A} = \begin{bmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{bmatrix} [\mathbf{a}_1, \mathbf{a}_2] = \begin{bmatrix} \mathbf{a}_1'\mathbf{a}_1 & \mathbf{a}_1'\mathbf{a}_2 \\ \mathbf{a}_2'\mathbf{a}_1 & \mathbf{a}_2'\mathbf{a}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_k \alpha_k^2 & \sum_k \alpha_k\beta_k \\ \sum_k \alpha_k\beta_k & \sum_k \beta_k^2 \end{bmatrix}$$

De donde:

$$(\mathbf{A}'\mathbf{A})^{-1} = \frac{1}{\sum_k \alpha_k^2 \sum_k \beta_k^2 - \left(\sum_k \alpha_k\beta_k\right)^2} \begin{bmatrix} \sum_k \beta_k^2 & -\sum_k \alpha_k\beta_k \\ -\sum_k \alpha_k\beta_k & \sum_k \alpha_k^2 \end{bmatrix}$$

$$= \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}$$

Con el resultado anterior, se puede estimar $\mathbf{B}$ a partir de:

$$\mathbf{B}' = \begin{bmatrix} \mathbf{b}_1' \\ \mathbf{b}_2' \end{bmatrix} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'X$$

$$= \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} \sum_k \alpha_k \mathbf{u}_k' \\ \sum_k \beta_k \mathbf{u}_k' \end{bmatrix} \left[\sum_k \alpha_k \mathbf{u}_k \mathbf{v}_k' \right]$$

$$= \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} \sum_k \alpha_k\lambda_k \mathbf{v}_k' \\ \sum_k \beta_k\lambda_k \mathbf{v}_k' \end{bmatrix} \begin{bmatrix} \sum_k \lambda_k(p_{11}\alpha_k + p_{12}\beta_k)\mathbf{v}_k' \\ \sum_k \lambda_k(p_{12}\alpha_k + p_{22}\beta_k)\mathbf{v}_k' \end{bmatrix}$$

Del hecho que las columnas de $\mathbf{A}$ sean ortonormarles, la expresión se simplifica a:

$$\mathbf{B}' = \begin{bmatrix} \mathbf{b}_1' \\ \mathbf{b}_2' \end{bmatrix} = \begin{bmatrix} \frac{1}{\sum_k \alpha_k^2} \sum_k \alpha_k\lambda_k \mathbf{v}_k' \\ \frac{1}{\sum_k \beta_k^2} \sum_k \beta_k\lambda_k \mathbf{v}_k' \end{bmatrix}$$

Simplificando la notación para los coeficientes, las columnas de $\mathbf{B}$ son combinaciones lineales de las columnas de $\mathbf{V}$, y por tanto:

$$\mathbf{b}_1 = \sum_k \lambda_k \gamma_k \mathbf{v}_k$$

$$\mathbf{b}_2 = \sum_k \lambda_k \delta_k \mathbf{v}_k$$

Para valores de $\mathbf{B}$ se tiene que:

$$\mathbf{B}'\mathbf{B} = \begin{bmatrix} \mathbf{b}_1'\mathbf{b}_1 & \mathbf{b}_1'\mathbf{b}_2 \\ \mathbf{b}_2'\mathbf{b}_1 & \mathbf{b}_2'\mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \sum\limits_k \lambda_k^2\gamma_r^2 & \sum\limits_k \lambda_k^2\gamma_k\delta_k \\ \sum\limits_k \lambda_k^2\gamma_k\delta_k & \sum\limits_k \lambda_k^2\delta_k^2 \end{bmatrix}$$

siendo su inversa:

$$(\mathbf{B}'\mathbf{B})^{-1} = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}$$

$$= \frac{1}{\sum\limits_k \lambda_k^2\gamma_r^2 \sum\limits_k \lambda_k^2\delta_k^2 - \left[\sum\limits_k \lambda_k^2\gamma_k\delta_k\right]^2} \begin{bmatrix} \sum\limits_k \lambda_k^2\gamma_r^2 & -\sum\limits_k \lambda_k^2\gamma_k\delta_k \\ -\sum\limits_k \lambda_k^2\gamma_k\delta_k & \sum\limits_k \lambda_k^2\delta_k^2 \end{bmatrix}$$

Recalculado para $\mathbf{A}$:

$$\mathbf{A}' = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{X}' = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} \begin{bmatrix} \sum\limits_k \lambda_k\gamma_k\mathbf{v}_k' \\ \sum\limits_k \lambda_k\delta_k\mathbf{v}_k' \end{bmatrix} \begin{bmatrix} \sum\limits_k \lambda_k\mathbf{v}_k\mathbf{u}_k' \end{bmatrix}$$

$$= \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} \begin{bmatrix} \sum\limits_k \lambda_k^2\gamma_k\mathbf{u}_k \\ \sum\limits_k \lambda_k^2\delta_k\mathbf{u}_k \end{bmatrix} = \begin{bmatrix} \sum\limits_k \lambda_k^2(q_{11}\gamma_k + q_{12}\delta_k)\mathbf{u}_k \\ \sum\limits_k \lambda_k^2(q_{12}\gamma_k + q_{22}\delta_k)\mathbf{u}_k \end{bmatrix}$$

$$= \begin{bmatrix} \sum\limits_k \lambda_k^2\alpha_k^{(1)}\mathbf{u}_k \\ \sum\limits_k \lambda_k^2\beta_k^{(1)}\mathbf{u}_k \end{bmatrix}$$

Comparando los coeficientes iniciales con los calculados en el paso anterior, se tiene que:

$$(\mathbf{a}_1, \mathbf{a}_2) = \left[\sum_{k=1}^{s} \alpha_k\mathbf{u}_k, \quad \sum_{k=1}^{s} \beta_k\mathbf{u}_k\right]$$

$$(\mathbf{a}_1^{(1)}, \mathbf{a}_2^{(1)}) = \left[\sum_k \lambda_k^2 \alpha_k^{(1)} \mathbf{u}_k, \ \sum_k \lambda_k^2 \beta_k^{(1)} \mathbf{u}_k \right]$$

Iterando el proceso q veces, se tiene:

$$(\mathbf{a}_1^{(q)}, \mathbf{a}_2^{(q)}) = \left[\sum_k \lambda_k^{2q} \alpha_k^{(q)} \mathbf{u}_k, \ \sum_k \lambda_k^{2q} \beta_k^{(q)} \mathbf{u}_k \right]$$

Es decir, que los dos primeros valores singulares adquieren mayor importancia en cada paso, esto implica que el procedimiento converge a dos vectores no proporcionales, que son combinaciones lineales de los dos primeros vectores singulares.

$$\begin{aligned}(\mathbf{a}_1, \mathbf{a}_2) &= (\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2, \beta_1 \mathbf{u}_1 + \beta_2 \mathbf{u}_2) \\ &= (\mathbf{u}_1, \mathbf{u}_2) \begin{bmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{bmatrix} = (\mathbf{u}_1, \mathbf{u}_2)\mathbb{A}\end{aligned}$$

siendo:

$$\mathbb{A} = \begin{bmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{bmatrix}$$

Luego los valores obtenidos para $(\widehat{\mathbf{a}}_1, \widehat{\mathbf{a}}_2)$ generan el mismo subespacio que las columnas de $\mathbf{U}$, debido a que son combinaciones lineales de estas y son linealmene independientes. También se tiene que:

$$\mathbf{A}'\mathbf{A} = \mathbb{A}'\mathbf{U}'_{(2)}\mathbf{U}_{(2)}\mathbb{A} = \mathbb{A}'\mathbb{A}$$

A partir de $\mathbb{A}$ los estimadores para de las coordenadas de la matriz original, son:

$$\begin{aligned}\mathbf{B}' &= \begin{bmatrix} \mathbf{b}'_1 \\ \mathbf{b}'_2 \end{bmatrix} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'X = (\mathbb{A}'\mathbb{A})^{-1}\mathbb{A}'\mathbf{U}'_{(2)}X \\ &= (\mathbb{A}'\mathbb{A})^{-1}\mathbb{A}'\mathbf{U}'_{(2)}\mathbf{U}\mathbf{\Lambda}\mathbf{V} = (\mathbb{A}'\mathbb{A})^{-1}\mathbb{A}'\mathbf{\Lambda}_{(2)}\mathbf{V}'_{(2)}\end{aligned}$$

Es decir que :

$$\mathbf{B}' = \begin{bmatrix} \mathbf{b}'_1 \\ \mathbf{b}'_2 \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{2} \lambda_k \gamma_k \mathbf{v}_k \\ \sum_{k=1}^{2} \lambda_k \delta_k \mathbf{v}_k \end{bmatrix} = \begin{bmatrix} \gamma_1 & \gamma_2 \\ \delta_1 & \delta_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}$$

Es decir que las columnas de $\mathbf{B}$ son combinaciones lineales de los dos primeros vectores singulares, por la derecha. La matriz $\mathbf{X}$ se puede escribir de la forma:

$$\mathbf{X} = \mathbf{A}^{'}\mathbf{B} = \mathbf{U}_{(2)}\mathbb{A}(\mathbb{A}^{'}\mathbb{A})^{-1}\mathbb{A}^{'}\mathbf{\Lambda}_{(2)}\mathbf{V}_{(2)}^{'} = \mathbf{U}_{(2)}\mathbf{\Lambda}_{(2)}\mathbf{V}_{(2)}^{'}$$

Puesto que

$$\mathbb{A}(\mathbb{A}^{'}\mathbb{A})^{-1}\mathbb{A}^{'} = I$$

Con este procedimiento queda demostrado que el algoritmo de mínimos cuadrados alternados conduce a la descomposición en valores singulares.

1.3.7. Indicadores de calidad de la representación

A lo largo de este capítulo se ha mostrado de diferentes maneras, las bondades de aproximación del biplot a una matriz original. De igual manera se han deducido propiedades que con llevan a mostrar la eficiencia de su representación. En esta sección se hace un resumen de lo que se consideran los indicadores que muestran la calidad del gráfico obtenido con esta metodología.

La metodología biplot puede ser expresada de manera sencilla considerando la matriz $\mathbf{X}$ de tamaño $I \times J$, asumiendo que los datos están centrados por la media de cada variable y usando su descomposición en valores singulares

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$$

Retomamos este análisis a partir del *RCP biplot* y de manera análoga se hace para el *CRP biplot*.
Las I filas se encuentran en el espacio $\mathbf{R}^J$, la matriz $\mathbf{V}_{JJ}$ es cuadrada y ortogonal ($\mathbf{V}^{'}\mathbf{V} = \mathbf{I}$).

Luego la matriz $\mathbf{XV} = \mathbf{U\Lambda}$ denominada matriz de componentes principales, contiene la proyección de las I filas de la matriz original sobre el espacio $\mathbf{R}^J$ generado por las columnas de $\mathbf{V}$. Por tanto, si reducimos la matriz a $\mathbf{XV}_{(r)}$ esta contiene I filas en el espacio $\mathbf{R}^r$ subespacio de $\mathbf{R}^J$ e implica que a partir de la matriz $\mathbf{XV}_{(r)}$ se pueden estudiar las diferencias y similitudes entre los individuos con pérdida de información mínima, si la distancia euclidea es la apropiada en la matriz original.

Una expresión equivalente que muestra la relación con el biplot de regresión está dada a partir del teorema de Eckart-Young:

- Si el rango de $\mathbf{X}$ es mayor que r, la matriz $\widehat{\mathbf{X}} = \mathbf{U}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{V}_{(r)}^{'}$ es la aproximación por mínimos cuadrados en rango r (ver 1.3.17) donde las columnas de $\mathbf{V}_{(r)}$ corresponden a los r vectores propios asociados a los r primeros valores propios de $\mathbf{X}\mathbf{X}^{'}$, $\mathbf{\Lambda}_{(r)}$ matriz diagonal con sus r primeros valores propios y $\mathbf{U}_{(r)}$ matriz de los r vectores propios asociados a la matriz $\mathbf{X}^{'}\mathbf{X}$
- Además, Keller (1962) mostró que $\widehat{\mathbf{X}} = \mathbf{X}\mathbf{V}_{(r)}\mathbf{V}_{(r)}^{'}$ (ver 1.3.18)

1.3.8. Calidad Global de representación

Teniendo en cuenta que la matriz $\widehat{\mathbf{X}}$ es la mejor aproximación por mínimos cuadrados de $\mathbf{X}$ entonces la suma de los residuales al cuadrado corresponde al valor:

$$traza(\mathbf{X}-\widehat{\mathbf{X}})^{'}(\mathbf{X}-\widehat{\mathbf{X}}) = traza\left[\mathbf{X}(\mathbf{I}-\mathbf{V}_{(r)}\mathbf{V}_{(r)}^{'})\mathbf{X}^{'}\right]$$

$$= traza\left(\mathbf{\Lambda}^2 - \mathbf{\Lambda}_{(r)}^2\right) = \sum_{k=r+1}^{s}\lambda_k^2$$

Esta valor comparado con la *traza* de $\mathbf{\Lambda}^2 = \sum_{k=1}^{s}\lambda_k^2$ es lo que se denomina calidad *Global de representación* y corresponde a la proporción de información que explica el biplot en r dimensiones.

1.3.9. Correlación entre las variables originales y las componentes Estándar

La coordenada de cada individuo sobre la *k-ésima* componente estándar del vector $\mathbf{v}$, denominada *componente principal*, está dada por $\mathbf{a}_k = \mathbf{X}\mathbf{v}_k$ donde $\mathbf{X}^{'}\mathbf{X}\mathbf{v}_k = \lambda_k^2\mathbf{v}_k$. Denotando la variable h de la matriz original por $\mathbf{x}_h$, se tiene que la correlación entre $\mathbf{x}_h$ y $\mathbf{a}_k$, es

$$cor^2(\mathbf{x}_h, \mathbf{a}_k) = cor^2(\mathbf{x}_h, \mathbf{X}\mathbf{v}_k) = \frac{\mathbf{s_{12}}\mathbf{S_{22}^{-1}}\mathbf{s_{21}}}{\mathbf{x}_h^{'}\mathbf{x}_h}$$

Entonces:

$$cor^2(\mathbf{x}_h, \mathbf{X}\mathbf{v}_k) = \frac{(\mathbf{x}_h^{'}\mathbf{X}\mathbf{v}_k)^2}{\mathbf{x}_h^{'}\mathbf{x}_h} = \frac{\lambda_k^4\mathbf{v}_{kh}^2}{s_{hh}\lambda_k^2} = \frac{\lambda_k^2\mathbf{v}_{kh}^2}{s_{hh}}$$

Donde $\mathbf{v}_{kh}$ es la componente h del vector $\mathbf{v}_k$ y s_{hh} el *h-ésimo* elemento de la matriz $\mathbf{X}'\mathbf{X}$. Cuando los datos originales se encuentran estándarizados, se tiene que $\frac{1}{n}\mathbf{X}'\mathbf{X}$ es la matriz de correlación de las variables de la matriz original, por tanto $s_{hh} = n$ para todo h .

Así $r^2 = \lambda_k^2 \mathbf{v}_{kh}^2$ de donde $r = \lambda_k \mathbf{v}_{kh}$, es la correlación entre las observaciones para la variable h de la matriz original y la *k-ésima* componente estándar.

Esto significa, que si la matriz original se encuentra estandarizada; en la columna k de la matriz $\mathbf{V\Lambda}$ se encuentran las correlaciones entre las variables originales y la *k-ésima* componente.

Si este valor es 1, indica que la variable original está correlacionada con la respectiva componente, que el ángulo que forman es cero y por tanto, la proyección de la variable original se hace exactamente sobre esta componente.

Si este valor es cercano a 1, indica que la variable original está correlacionada con la respectiva componente y que el ángulo que forman es muy pequeño, es decir que la variable se ubica muy cerca de la componente.

Para una variable externa $\mathbf{X}_m$, se tiene que:

$$cor^2(\mathbf{X}_m, \mathbf{c}_k) = cor^2(\mathbf{X}_m, \mathbf{X}\mathbf{v}_k) = \frac{(\mathbf{X}'_m \mathbf{X}\mathbf{v}_k)^2}{\mathbf{X}'_m \mathbf{X}_m}$$

Y este valor indica que tan próxima se encuentra la variable a la componente.

1.3.10. Contribuciones

- *Indicadores generados por las filas*
  Tanto para las componentes principales en el *RPM*, como para las coordenadas principales en el *CPM*, se tiene que:

  $$\mathbf{A}\mathbf{A}' = \mathbf{\Lambda}^2$$

  Para $\mathbf{\Lambda}^2$ matriz de valores propios.

  A continuación se ilustra la matriz $\mathbf{A}$ de componentes principales (coordenadas principales) elevadas al cuadrado, donde se puede observar que:

Filas/Ejes	eje 1	...	eje r	...	eje R	suma
1	a_{11}^2	...	a_{1r}^2	...	a_{1R}^2	$\sum_{\alpha=1}^{R} a_{1\alpha}^2$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
p	a_{p1}^2	...	a_{pr}^2	...	a_{pR}^2	$\sum_{\alpha=1}^{R} a_{p\alpha}^2$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	a_{I1}^2	...	a_{Ir}^2	...	a_{IR}^2	$\sum_{\alpha=1}^{R} a_{I\alpha}^2$
suma	λ_1^2	...	λ_r^2	...	λ_R^2	$\sum_{i=1}^{I}\sum_{\alpha=1}^{R} a_{i\alpha}^2 = \sum_{\alpha=1}^{R} \lambda_\alpha^2$

Esto significa, que cada una de las coordenadas al cuadrado puede considerarse como la contribución absoluta, de cada individuo a la variabilidad total del respectivo eje.

Para comparar la contribución del individuo en los diferentes ejes, se calcula su contribución relativa a cada eje, mediante el cociente:

$$CRE_iF_\alpha = \frac{a_{i\alpha}^2}{\lambda_\alpha^2}$$

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

Adicionalmente, a la cantidad:

$$CRT_i = \frac{\sum_{\alpha=1}^{R} a_{i\alpha}^2}{\sum_{\alpha=1}^{R} \lambda_{i\alpha}^2}$$

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

A la cantidad

$$CRF_\alpha E_i = \frac{a_{i\alpha}^2}{\sum_{\alpha=1}^{R} a_{i\alpha}^2}$$

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

Este valor también corresponde al coseno al cuadrado del ángulo formado por la fila i y el factor α.

- *Indicadores generados por las coordenadas principales*
  Análogo a las filas, para las componentes estándar en el *RPM*, como para el *CPM*, se tiene que:

$$\mathbf{B}\mathbf{B}^{'} = \mathbf{I}$$

Lo que implica que los vectores que forman la matriz, son normados y por tanto sus valores al cuadrado corresponden a la contribución de la variable a la formación del factor.

Así, el valor:

$$\mathbf{b}_{j\alpha}^2 * 100$$

indica la proporción de inercia que la variable j aporta a la construcción del factor α

La cantidad:

$$CRT_j = \frac{\sum_{\alpha=1}^{R} b_{j\alpha}^2}{\sum_{\alpha=1}^{R} \lambda_{j\alpha}^2}$$

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.

La cantidad:

$$CRE_jF_\alpha = \frac{b_{j\alpha}^2}{\lambda_\alpha^2}$$

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

La cantidad:

$$CRF_\alpha E_j = \frac{b_{j\alpha}^2}{\sum_{\alpha=1}^{R} b_{j\alpha}^2}$$

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

Capítulo 2

Tablas Múltiples

2.1. Introducción

Desde el punto de vista estadístico lo más simple es medir una variable a un conjunto de individuos, con esta información se construye un vector de observaciones y se tiene entonces una entrada: individuos. Si además se considera un conjunto de más de una variable, la información se puede estructurar como una matriz de datos, y se tienen dos entradas: individuos y variables.

Si para cada matriz anterior se hace una repetición de mediciones en ocasiones diferentes, estaríamos ante un arreglo de tres dimensiones; se tienen tres entradas: individuos, variables y ocasiones. Finalmente se puede pensar en hacer repeticiones en condiciones o espacios diferentes, de lo medido anteriormente hasta construir un arreglo de cuatro, cinco o más entradas, que en forma general puede denominarse como un K arreglo de datos.

El análisis desde el punto de vista de la explicación de los datos, en el caso de una entrada, lo que se modela, principalmente, es la distribución de los datos; en dos vías se pueden construir modelos explicativos o de co-variación y para tres, cuatro, o en general K entradas, una de las posibilidades es la construcción de modelos que se han llamado de tres, cuatro ó K vías. También es clásico asociar el conjunto de datos a una representación gráfica que permita reconocer y mostrar la tendencias esenciales de los fenómenos que se estudian.

Nuestro propósito fundamental es analizar tablas múltiples, donde a las

observaciones, por ejemplo: objetos o individuos, se les han hecho varias mediciones (variables) en varias ocasiones. El témino ocasiones se puede referir a momentos en el tiempo ó a condiciones de medición diferentes.

Consideraremos dos tipos de datos de tablas múltiples:

- observaciones de diferentes conjuntos de individuos, en un número fijo de variables, en diferentes ocasiones;
- observaciones de los mismos individuos en diferentes conjuntos de variables en diferentes ocasiones.

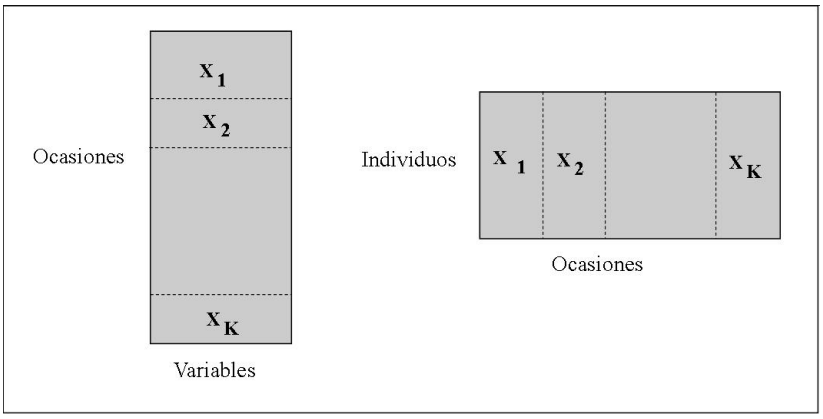

Figura 2.1: Esquema de los datos múltiples

Un caso intermedio es considerar el mismo conjunto de individuos a los que se les ha medido el mismo conjunto de variables en diferentes ocasiones, este tipo de tablas se denominan "datos Cúbicos".

En los últimos años se han planteado diferentes modelos para el análisis, tanto de datos cúbicos como de tablas múltiples. Cuando interesa compararlos, una forma es plantear modelos que minimicen una cierta función de pérdida.

Es importante notar que las técnicas para el análisis de datos múltiples pueden ser utilizadas para analizar los datos de cúbicos, ya que estos siempre pueden verse como supermatrices. Sin embargo, no es posible analizar los conjuntos de datos múltiples como datos cúbicos, porque no siempre se tiene la información para completar el cubo que caracteriza estos datos.

Cuando se cuenta con tablas múltiples, inicialmente se procede a analizarlas de manera individual e indendiente pero se corre el riesgo de quedar inmerso en un análisis de numerosas representaciones.

Por está razon nace la necesidad de buscar un conjuno único de representación que se convierta en un resúmen global y que es denominado 'compromiso' del conjunto de tablas.

A continuación se presentan los conceptos que se tienen en cuenta en este tipo de análisis.

2.1.1. Interestructura

El objetivo de la interestructura, también llamado 'análisis global', es comparar las tablas entre ellas y reconocer grupos homogéneos. Para hacer esta comparaciones se debe tener en cuenta el concepto de proximidad entre ellas, además buscar un conjunto de representación gráfica sobre la cual se pueda interpretar la proximidad entre dos puntos como correspondencia a dos tablas similares o parecidas en el sentido de la distancia considerada.

Con este análisis se muestran las proximidades entre las tablas sin poder interpretar la descripcion 'detallada' de los elementos que generan las similitudes o difrencias entre las tablas.

2.1.2. Consenso

Consiste en resumir las K tablas de datos en una sola denominada *compromiso o consenso*, de la misma naturaleza que las tablas; donde la selección de ésta depende del método de trabajo a utilizar. El espacio generado por el compromiso permite un resumen global del conjunto de tablas.

2.1.3. Intraestructura

Como el análisis global no es suficiente para comparar las tablas con mayor detalle, el compromiso obtenido en la etapa anterior permite representar las *posiciones consenso* de cada uno de los elementos (individuos y/o variables) que constituyen las diferentes tablas. Las posiciones compromiso de los elementos corresponden a las posiciones medias de los mismos. Este procedimiento es denominado 'análisis de la intraestructura'

2.1.4. Trayectorias

Este concepto tiene orígen en los estudios donde las tablas son generadas a partir del tiempo y lo que se busca es describir la evolución del fenómeno. Sin embargo el concepto es aplicable a las tablas generadas a partir de diferentes experimentos ó a cualquier otro tipo de tablas donde este concepto sea interpretable.

El objetivo se puede aplicar de manera general para conocer la evolución del fenómeno que se analiza y a nivel de detallado, estudiar la evolución de cada uno de los elementos que componen las tablas. Es la razón para que se hable de análisis de las trayectorias.

2.1.5. Algunos métodos de análisis de tablas múltiples

Las tablas múltiples han sido analizadas por medio de diferentes métodos entre los que se tiene:

- El AFM (Análisis Factorial Múltiple) fue introducido por ESCOFIER y PAGES en 1984. Este método se utiliza para analizar conjuntos de datos en los que un mismo conjunto de individuos se describen por varios conjuntos de variables. Dentro de una tabla las variables deben ser del mismo tipo (continuas o nominales) pero es posible tener tablas con variables de diferente tipo.
- Los Métodos STATIS y STATIS DUAL introducidos por el equipo de ESCOUFIER (1980 y 1985) y desarrollado posteriormente por LAVIT (1988); permite explorar simultáneamente varias tablas de datos. Se aplica a tablas de datos cuantitativos que han sido coleccionados en diferentes ocasiones, sobre los mismos individuos, cuyas variables pueden ser eventualmente diferentes.
- El DACP (Doble Análisis de Componentes Principales), planteado por BOUROCHE (1975), es un método aplicable a conjuntos de datos donde a los mismos individuos se les mide las mismas variables en diferentes ocasiones
- EL ASCM (Análisis de Series Cronológicas Multidimensionales) fue planteado por TENNEHAUS y PRIEURET (1974), es aplicable a tablas de datos temporales con los mismos individuos y las mismas variables.

- Meta Componentes, planteado por KRAZNOWSKI (1979), es un método que permite comparar subespacios obtenidos por componente principales de varias tablas de datos con las mismas variables.

- El análisis de Componentes Principales Comunes, fue introducido por FLURY (1984 y 1988) y permite comparar matrices de covarianzas de la misma dimensión.

- Una generalización del Análisis de Componentes Principales a K conjuntos de variables. Este enfoque fue planteado por CASIN, (2001) como una generalización del Análisis de Componentes Principales (ACP) a varias tablas de datos y con la intención de tener en cuenta tanto la estructura de correlación dentro de las tablas de datos como la relación entre éstos. Plantea un integración entre la técnica del Análisis Canónico General postulado por CARROL, (1968) y el ACP. Con el primero se capitaliza exclusivamente las relaciones entre las tablas y se ignora la correlación dentro de ellas y con el segundo se describe la estructura de correlación pero se aplica solo a una tabla.
  El propósito del método es encontrar tanto direcciones comunes en conjuntos de variables, como describir cada uno de éstos. El método proporciona componentes en diferentes conjuntos de variables, las cuales están mutuamente incorrelacionadas dentro de estos conjuntos, y están relacionadas con componentes de los otros conjuntos de variables.

- Para análisis exploratorio de tablas multiples existen los métodos fundamentales propuestos por TUCKER (1964 y 1966), que los llamó Análisis Factorial de Tres Modos. Los modelos son: Tucker 3 cuando se quieren reducir los tres modos y Tucker 2 cuando uno de los tres modos no se reduce. Están basados en la descomposición en valores singulares de tres modos y son los modelos derivados del análisis de componentes principales, si se trabaja a rango completo se logra reproducir el dato, pero si se retienen las primeras componentes para cada modo no se garantiza un buen ajuste.

- KROONENBERG y de LEEUW en 1980, desarrollaron un algoritmo para encontrar las matrices de las componentes para cada modo de tal manera que se minimiza la suma de cuadrados residual; es decir, que si no se trabaja a rango completo se obtiene una solución aproximada. El procedimiento se basa en la descomposición en valores singulares de tres modos por lo que se adaptan muy bien para el tratamiento

de datos de tres vías, ya que los marcadores se estiman a partir del análisis simultáneo de tres matrices concatenadas, y por lo tanto los datos no son forzados a tener una estructura de dos vías. Esta forma de determinación de las componentes de cada modo la denominaron Tuckals 3 y Tuckals 2 respectivamente.

La diferencia entre los modelos planteados por TUCKER y los de KROONENBERG y DE LEEUW es la forma de estimación de los marcadores para cada uno de los modos.

- KROONENBERG en 1983 propone el ajuste al modelo de escalamiento de tres modos (TUCKER, 1972) que es utilizado cuando dos de los tres modos coinciden, es decir, es un modelo derivado del Tucker 3 y por lo tanto la estimación del mismo. Si se aplica el procedimiento descrito por KROONENBERG y de LEEUW, se puede considerar una derivación del Tuckals 3.

- KROONENBERG (1983), presenta una aplicación a datos reales para integrar las matrices de correlación de varias tablas. Compara los resultados obtenidos en los análisis de componentes principales de dos vías realizados individualmente en cada matriz, con los que obtiene mediante la aplicación del Tuckals 2. A este procedimiento lo denomina análisis de tres modos de las matrices de correlación.

- KIERS (1991) menciona la utilización de los métodos de componentes principales de tres modos en el análisis de datos de conjuntos múltiples a partir de las matrices de productos cruzados.

En este trabajo centraremos nuestra atención en algunos métodos como el AFM, STATIS, STATIS DUAL, DACP y Meta componentes, con el fin de mostrar algunos enfoques que dan base para el desarrollo del tema que nos proponemos.

2.2. Análisis Factorial Múltiple

El Análisis Factorial Múltiple es un análisis (Escofier y Pagès, 1984; 1990) adaptado al tratamiento de datos donde un mismo conjunto de individuos se describe a través de varios grupos de variables.
Por lo tanto la información de partida esta constituida por K tablas de I individuos asociados a las filas y descritas por J variables asociadas a las

columnas.

Una de las ventajas de este método es que las variables pueden ser cuantitativas, cualitativas o de ambos tipos, interviniendo todas ellas como activas en el análisis global.
La única restricción que impone el AFM para el tratamiento de tablas mixtas es que las variables sean homogéneas al interior de cada grupo.
La definición de grupos proviene del hecho que las variables que constituyen una tabla se refieren a un mismo tema, a una misma fecha, etc.

El objetivo es encontrar una estructura común o representativa de todas las tablas

Por tanto, denotando $\mathbf{X}$ la tabla que cruza el conjunto I de individuos con el conjunto J de todas las variables sin diferenciar las J_k variables de cada tabla y $\mathbf{X}_k$ la tabla que cruza I con todas las variables del grupo J_k; por tanto $J{=}\cup_k J_k$. Se asume que I, J, K, J_k designan a la vez el conjunto y su cardinal.

La matriz de pesos para los individuos se denota $\mathcal{D}$, si todos los individuos desempeñan el mismo papel, todos tendrán el mismo peso $1/I$. Sin embargo, hay ocasiones en las que es necesario dar pesos diferentes a los individuos, por ejemplo cuando los individuos representan diferentes poblaciones, entonces se asigna a cada individuo un peso proporcional de acuerdo a la población que representa. Estos pesos intervienen en el cálculo de las medias de cada variable y en la medida de asociación entre estas. La matriz de los pesos para las variables se denota $\mathbf{M}$ para la tabla global $\mathbf{X}$ y $\mathbf{M}_k$ para las variables de la tabla $\mathbf{X}_k$. En la mayoría de los casos las variables intervienen de la misma manera con peso igual a 1, sin embargo si se quiere destacar o reducir la influencia de alguna variable, se cambiará su peso. Estos pesos afectan las distancias entre individuos debido a que los pesos de las variables ponderan la influencia en cada columna de las tablas.

Con el fin de analizar la tabla $\mathbf{X}_k$ se considera el estudio $(\mathbf{X}_k, \mathbf{M}_k, \mathcal{D})$

En el análisis de la tabla $\mathbf{X}$, la introducción simultánea de varios grupos necesita equilibrar la influencia de éstos, lo que se logra al multiplicar los pesos iniciales de todas las variables del grupo J_k por un coeficiente α_k.

Entonces el AFM se desarrolla, en forma general, en dos etapas:

Primera etapa: se realizan Análisis Individuales de cada grupo de variables y se obtiene el mayor valor propio, que servirá para ponderar las variables de las respectivas tablas, que serán analizadas en la segunda etapa.

Segunda Etapa: se realiza ACP del conjunto J de variables, previamente ponderadas, es decir, un Análisis Global, que da paso a la comparación de las tablas y a las trayectorias de los individuos.

A continuación detallamos las dos etapas que conforman el proceso y los análisis que conllevan.

2.2.1. Análisis Individuales

Esta etapa consiste en realizar un análisis factorial separado de cada una de las tablas $\mathbf{X}_k$, es decir, que se hacen K análisis, uno para cada tabla.

Estos análisis pueden ser:

- Análisis de Componentes Principales (ACP) normados o no normados en los grupos de variables cuantitativas.

- Análisis de Correspondencias Múltiples (ACM) si las variables son cualitativas.

Estos análisis se denominarán análisis parciales, de donde se obtiene el mayor valor propio $\lambda^2_{1_k}$, cuyo inverso servirá de ponderador α_k para el grupo J_k de variables, de la tabla X_k; también se obtienen las coordenadas estándar que serán tratadas en la segunda etapa como variables suplementarias, es decir, que no intervendrán en el análisis pero serán proyectadas sobre el respectivo subespacio.

La razón para ponderar se debe a que en el tratamiento simultáneo de varios grupos de variables conlleva el problema que un grupo preponderante determine el análisis global, esto es, que la determinación de una componente se apoye solamente en la inercia de una dirección principal.

Normalizar para igualar el peso de los diferentes grupos, implica que los grupos participen de un modo equilibrado en la determinación de las componentes. Pero por otra parte, es importante respetar la estructura múltiple que presenta cada una de las tablas.

El AFM utiliza como ponderador el inverso del primer valor propio de cada matriz, es decir $1/\lambda^2_{1_k}$. Por tanto las variables de un mismo grupo (tabla)

reciben la misma ponderación.

Esta ponderación logra que la inercia de la primera dirección principal de cada tabla sea igual a 1 y optimiza en razón a 1, la inercia de las otras direcciones. Sin embargo, no equilibra la inercia total de cada tabla, ya que es un simple cambio de escala, es decir, que considera la naturaleza múltiple de los datos.

Es importante conocer si una matriz tiene estructura multidimensional porque podría tener una influencia preponderante en el análisis global. Se debe aclarar, no obstante, que el hecho de que un grupo tenga estructura multidimensional, no implica que necesariamente participe en la formación de uno o varios ejes del espacio consenso, pero a diferencia de un grupo unidimensional puede intervenir en la formación de varios ejes. Por tanto, es de utilidad considerar una medida de dimensionalidad de cada grupo.

A partir de $\mathbf{W}_k\mathcal{D} = \mathbf{X}_k\mathbf{M}_k\mathbf{X}_k^{'}\mathcal{D}$, se define una medida de asociación entre los grupos de variables J_k y J_l a partir del producto escalar de Hilbert Schmidt.

Se define el producto escalar de Hilbert-Schmidt como:

$$\langle \mathbf{W}_k\mathcal{D}, \mathbf{W}_l\mathcal{D}\rangle_{HS} = traza(\mathbf{W}_k\mathcal{D}\mathbf{W}_l\mathcal{D})$$

El cuadrado de la norma de $\mathbf{W}_k\mathcal{D}$

$$\|\mathbf{W}_k\mathcal{D}\|^2 = \langle \mathbf{W}_k\mathbf{D}, \mathbf{W}_k\mathcal{D}\rangle_{HS} = \sum_s \lambda_{s_k}^4$$

Siendo los $\lambda_{s_k}^2$ los s valores propios de de la matriz $\mathbf{W}_k\mathcal{D}$, es decir, los obtenidos cuando se realiza el análisis de la tabla $\mathbf{X}_k$.

Debido a la ponderación de las variables del grupo X_k por $1/\lambda_{1_k}^2$, el cuadrado de la norma de $\mathbf{W}_k\mathcal{D}$, es:

$$[N_g(J_k)]^2 = \left\| \frac{\mathbf{W}_k\mathcal{D}}{\lambda_{1_k}^2} \right\|^2 = \sum_s \left[\frac{\lambda_{s_k}^2}{\lambda_{1_k}^2} \right]^2 = 1 + \sum_{s>1} \left[\frac{\lambda_{s_k}^2}{\lambda_{1_k}^2} \right]^2$$

Por tanto la norma depende de la estructura del grupo y su valor se interpretan como un indicador de multidimensionalidad. Si $N_g(J_k) \simeq 1$ implica que

la estructura del grupo es prácticamente unidimensional. Entre más grande sea $N_g(J_k)$ mayor estructura multidimensional del grupo.

2.2.2. Análisis Global

Con este análisis se obtiene la estructura consenso y los ejes de la Intra-estructura para poner en evidencia los principales factores de variabilidad de los individuos, en los distintos grupos de variables.

2.2.2.1. Consenso

Se trata de encontrar un objeto que genere un espacio común donde se representen de la mejor manera, según un criterio, las K tablas para ser analizadas.

El objeto corresponde al estudio ($\mathbf{XM}\mathcal{D}$) al que se le realiza un análisis en componentes principales (ACP) normado, donde:

$$\mathbf{X} = \left[\begin{array}{cccccc} [\mathbf{X}_1] & \dots & [\mathbf{X}_k] & \dots & [\mathbf{X}_K] \end{array} \right]$$

Utilizando la métrica:

$$\mathbf{M} = \left[\begin{array}{ccccc} \frac{1}{\lambda^2_{1_1}}\mathbf{M}_1 & \dots & \dots & \dots & 0 \\ \vdots & \ddots & & & \vdots \\ \vdots & \dots & \frac{1}{\lambda^2_{1_k}}\mathbf{M}_k & \dots & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \frac{1}{\lambda^2_{1_K}}\mathbf{M}_K \end{array} \right]$$

donde $\mathbf{M}_k$ es la matriz de pesos para las variables de la tabla $\mathbf{X}_k$

Este procedimiento es equivalente a realizar un ACP Normado de la matriz $\mathbf{X}$ que es la matriz formada por la yuxtaposición de las K tablas, ponderadas por el inverso de su respectivo mayor valor propio, es decir:

$$\mathbf{X} = \left[\begin{array}{cccccc} \frac{\mathbf{X}_1}{\lambda^2_{1_1}} & \dots & \frac{\mathbf{X}_k}{\lambda^2_{1_k}} & \dots & \frac{\mathbf{X}_K}{\lambda^2_{1_K}} \end{array} \right]$$

usando como métrica $\mathbf{M}$, la matriz

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \dots & \dots & \dots & 0 \\ \vdots & \ddots & & & \vdots \\ \vdots & \dots & \mathbf{M}_k & \dots & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \mathbf{M}_K \end{bmatrix}$$

El ACP ponderado de la tabla $\mathbf{X}$ permite representar la estructura compromiso en un espacio de dimensión menor y obtener la Imagen Euclídea Compromiso, que es una *Imagen Media* de la tabla (es decir, las tablas originales yuxtapuestas).

La interpretación es idéntica a la de un ACP, por lo que tendremos:

–una visión del conjunto de individuos medios y de sus trayectorias sobre la *imagen euclídea compromiso.*

–la representación de las variables, como siempre sucede en un ACP, puede ser considerada indirectamente, como una ayuda a la interpretación de la imagen euclídea de la nube de los individuos y como una representación óptima de las correlaciones entre las variables y los factores.

Como las variables en el AFM pueden ser mixtas se tendrá que: si las variables son continuas se miden las correlaciones entre ellas y los factores; si las variables son cualitativas, se representa para cada modalidad la coordenada del centro de gravedad de los individuos que presentan esa modalidad o categoría.

A partir del ACP anterior es posible realizar una comparación de los individuos medios, pero también es importante observar si dos individuos, que son próximos desde un punto de vista general, también lo son desde un punto de vista parcial (en cada tabla). Es decir, se quiere comparar la posición relativa de cada individuo visto desde las diferentes tablas.

Para exponer la teoría, nos centraremos en tres espacios dentro los cuales se puede presentar el AFM según se trate de representar los individuos, las variables o los grupos de variables.

2.2.2.2. Trayectoria de los individuos

Para conocer la trayectoria de los individuos, se realiza el análisis de la nube de individuos en el espacio $\mathbf{R}^J$

En este espacio se busca una representación de la nube N_I de los individuos caracterizados por las variables del conjunto total J y una representación superpuesta de las K nubes N_I^k, ubicadas en los subespacios $\mathbf{R}^{J_k}$ (de $\mathbf{R}^J$) de los individuos caracterizados unicamente por las variables del grupo J_k.

Con este objetivo el AFM realiza una representación superpuesta de cada una de las nubes parciales de individuos sobre los ejes del análisis global, proyectándolas como individuos suplementarios o ilustrativos.

Es importante aclarar que, si bien el procedimiento de proyección es igual al de los elementos suplementarios, los grupos de individuos parciales no son exactamente elementos suplementarios, ya que sus valores han contribuido a la construcción de los ejes de la Intra-estructura. Los grupos de individuos parciales son grupos activos en el análisis global.
Por lo tanto cada individuo aparecerá varias veces en dicho análisis, como individuo medio y como individuo parcial.

Para obtener una buena reprentación simultánea de las nubes N_I^k de individuos en $\mathbf{R}^J$, es necesario que la representación posea buenas propiedades, a saber:

– La representación de cada nube debe ser una proyección ortogonal de la misma.
– Cada nube de individuos debe estar bien representada.
– Los puntos que representan a un mismo individuo en las diferentes tablas deberán estar próximos los unos de los otros.

Veamos como se hace la representación para que las propiedades se cumplan:

Representación de las K nubes de individuos N_I^k en el espacio $\mathbf{R}^J$ y de la nube promedio

Para representar simultáneamente las nubes N_I^k en $\mathbf{R}^J$, se realiza un ACP de la tabla $\mathbf{X}$ introduciendo como elementos suplementarios las K tablas; para esto se considera que los individuos parciales están contenidos en las tablas $\tilde{\mathbf{X}}_k$ de dimension IxJ, que corresponde a:

$$\tilde{\mathbf{X}}_k = \begin{bmatrix} \mathbf{0} & \mathbf{X}_k & \mathbf{0} \end{bmatrix}$$

Proyectando las $\tilde{\mathbf{X}}_k$ en el espacio $\mathbf{R}^J$, se cumple la primera condición.

Veamos que las otras dos condiciones se cumplen:

Se denota N_I^* a la nube de los puntos que conforman los centros de gravedad $i^* = \frac{1}{K}\sum_k i^k$, correspondientes a los K puntos i^k que representan el mismo individuo i en las N_I^k : N_I^* es la nube promedio y su representación se logra realizando un ACP de la tabla $\mathbf{X}$ con las variables ponderadas.

La nube N_I^* es homotética (con razón $1/K$) de la nube N_I por ser la nube de los centros de gravedad.

Representación simultánea de las nubes de individuos

El objetivo del análisis consiste en representar los individuos de cada tabla y como referencia la nube promedio, con el fin de poner en evidencia las similitudes o diferencias. Para esto se debe tener en cuenta las dos siguientes condiciones:

- Cada nube N_I^k debe estar bien representada. Dado que la representación de N_I^k es una proyección ortogonal, se busca para las N_i^k las proyecciones de inercia importantes y por consecuencia, aquellas que maximicen la inercia proyectada de $N_I^K = \cup_k N_I^k$.
- Las representaciones de las nubes se deben parecer entre sí, esto significa que los puntos homólogos deben ser lo más próximos posibles. La nube N_I^K se ha particionado en K nubes, donde cada una contiene los I individuos analizados a través de un solo grupo de variables. Ahora se introduce otra partición de N_I^K, sea esta: I nubes, denotadas N_i^K, cada una con K puntos que representan el mismo individuo i a través de cada grupo de variables y cuyo centro de gravedad es i^*.

Aplicando el principio de Huygens a las proyecciones $N_I^K = \cup_{i=1}^I N_i^k$, se tiene:

$$\text{inercia total de } N_I^K = \text{inercia Intra} + \text{inercia Inter}$$

donde la inercia Intra corresponde a la inercia de las N_i^K respecto a i^* y la inercia Inter a la inercia de N_I^*; para que los puntos que representan un

mismo individuo estén cerca, se debe minimizar la inercia proyectada de cada N_i^K, es decir, minimizar la inercia Intra de N_I^K.

Como según la primera condición, se quiere maximizar el primer miembro, que es la Inercia Total y ahora se quiere minimizar el primer término del segundo miembro que es la Inercia Intra, las dos condiciones son incompatibles y no se pueden satisfacer a la vez.
Por tanto, dándoles la misma importancia, las mejores representaciones son obtenidas al maximizar la Inercia Inter, por las proyecciones en el plano generado por los dos primeros vectores propios de la nube N_I^* de los centros de gravedad.

Así, si se realiza un ACP ponderado de la nube N_I^* se cumplen la segunda y tercera condición, ya que en dicho análisis se maximiza la Inercia Inter.
Teniendo en cuenta que la nube N_I^* es homotética de la nube N_I la imagen euclídea óptima así obtenida será la misma que si se realiza un ACP ponderado de la tabla $\mathbf{X}$ (que es lo que se hace en el Análisis Global).

Luego se justifica la proyección de las K nubes de individuos en la estructura obtenida a partir del análisis global.

Indicador de la calidad de las proyecciones de los individuos parciales

Como el AFM busca reducir la inercia intra (a través de la inercia inter) de las nubes N_i^K, para que los puntos que representan al mismo individuo i estén lo más próximos los unos de los otros, parece lógico tomar como medida de la semejanza entre las proyecciones de las nubes N_I^k, la inercia intra en relación a la inercia total.

Por tanto, para cada eje es importante calcular el cociente:

$$\frac{Inercia\,Inter}{Inercia\,Total} = 1 - \frac{Inercia\,Intra}{Inercia\,Total}$$

Si el cociente es próximo a 1, significa que todas las nubes N_I^k tienen muchas características comunes sobre ese eje lo que validaría la realización de un estudio fino, o detallado, de sus diferencias sobre ese eje.

Si por el contrario es próximo a cero, eso indica que las diferencias de formas entre las nubes en ese eje no son importantes y por lo tanto no se justifica un estudio detallado de las trayectorias.

2.2.3. Interestuctura

El objetivo de esta etapa del AFM es realizar una interpretación de la posición de las distintas tablas o grupos de variables, sin preocuparse del grupo de pertenencia, con el fin de comprobar si los distintos grupos definen estructuras diferenciadas sobre el conjunto de individuos.

Primero vamos a considerar una medida de la relación entre dos tablas y posteriormente veremos cómo se realiza el estudio para comparar los K grupos de variables.

Para representar el grupo J_k se toma el operador $\mathbf{W}_k\mathcal{D} = \mathbf{X}_k\mathbf{M}_k\mathbf{X}_k^{'}\mathcal{D}$, que se puede considerar un elemento del espacio $\mathbf{R}^{I^2}$, siendo $\mathbf{M}_k$ la matriz que contiene las ponderaciones que recibe cada variable en el AFM.
Se sabe que al diagonalizar $\mathbf{W}_k\mathcal{D}$ se puede obtener una perfecta reconstitución de la estructura de la nube N_I^k de los individuos de la tabla $\mathbf{X}_k$.

La expresión representativa del *k-ésimo* grupo es:

$$\frac{\mathbf{W}_k}{\lambda_{1_k}^2}$$

En $\mathbf{R}^{I^2}$ se define el producto escalar de Hilbert-Schmidt, para dos grupos de variables $\mathbf{J}_k$ y $\mathbf{J}_l$ como:

$$\left\langle \frac{\mathbf{W}_k\mathcal{D}}{\lambda_{1_k}^2}, \frac{\mathbf{W}_l\mathcal{D}}{\lambda_{1_l}^2} \right\rangle_{HS} = traza\left[\frac{\mathbf{W}_k\mathcal{D}\mathbf{W}_l\mathbf{D}}{\lambda_{1_k}^2 \times \lambda_{1_l}^2}\right]$$

Entonces para comparar dos grupos de variables, se utiliza el indicador $L_g(J_k, J_l)$, que se define como:

$$L_g(J_k, J_l) = traza\left[\frac{\mathbf{W}_k\mathcal{D}\mathbf{W}_l\mathbf{D}}{\lambda_{1_k}^2 \times \lambda_{1_l}^2}\right]$$

Si el valor $L_g(J_k, J_l)$ es grande, implica que las variables de los dos grupos están correlacionadas y por tanto tienen direcciones de inercia común, de lo contrario, si el valor es cercano a cero cada variable del grupo J_k no está correlacionada con las variables del grupo J_l.

Este indicador no tiene un límite superior; es más grande cuanto más multidimensionales sean los grupos k y l; y presentan mayor cantidad de dimensiones comunes y próximas a las direcciones de inercia más importantes de cada grupo.

El índice que mide la magnitud de la relación entre dos tablas o dos grupos de variables en el AFM, es el coeficiente RV definido por Escoufier en 1976:

$$RV(J_k, J_l) = \frac{L_g(J_k, J_l)}{N_g(J_k)N_g(J_l)}$$

Este índice toma valores entre cero y 1, se interpreta como un valor de similitud entre las estructuras comunes de los grupos.
Toma el valor de 1 si las nubes de los individuos generadas por los dos grupos J_k, J_l son homotéticas, es decir que si el valor es cercano 1 los grupos son más similares; si el valor es cercano a cero los grupos son más diferentes.

2.2.3.1. Representación de cada uno de los grupos de variables

En el AFM se proyectan las tablas sobre un sistema de ejes que son los mismos que los obtenidos en la Intra-estructura.

Al grupo J_k de variables, se asocia la nube de variables N_J^k en $\mathbf{R}^I$, se puede calcular la matriz de distancias entre individuos y a esta matriz de distancias le corresponde un punto en el espacio $\mathbf{R}^{I^2}$ y al conjunto N_J^K le corresponde una nube de K puntos en $\mathbf{R}^{I^2}$. En esta nube, dos puntos estarán próximos si las estructuras conferidas por los dos grupos de variables para los I individuos se parecen.

La imagen euclídea de los grupos de variables construida por el AFM se obtiene de proyectarlos sobre los vectores de $\mathbf{R}^{I^2}$, inducidos por los factores del análisis global, y no sobre sus ejes de inercia.

Para representar los grupos de variables, se impone a los ejes de que sean elementos simétricos de rango 1, es decir matrices asociadas a grupos compuestos de una sola variable; por tanto, a cada vector de $\mathbf{R}^{I^2}$ le corresponderá una variable de $\mathbf{R}^I$ de manera que al proyectar el grupo, se pueda interpretar en función de sus variables originales.

El objetivo es encontrar un subespacio ortonormado de $\mathbf{R}^{I^2}$, que se ajuste lo mejor posible al conjunto de tablas.

Para buscar factores comunes a los diferentes grupos, se habla del concepto de Análisis multicanónico debido a Carroll, que procede en dos etapas:

- Buscar una secuencia de variables, ortogonales entre sí, llamadas variables generales, que resuman la tendencia general de los grupos.

- Buscar para cada variable general y en cada grupo, una variable llamada canónica, combinación lineal de las variables del grupo, que maximiza un criterio de asociación entre la variable general y el grupo.

El criterio utilizado por Carroll es el coeficiente de correlación múltiple, que tiene el inconveniente de ser inestable, es decir sensible a pequeñas variaciones de las variables, en el caso que las variables del grupo estén correlacionadas entre ellas.

En el AFM, se procede de la siguiente manera:

Se busca un vector de $\mathbf{R}^{I^2}$ que denominaremos $\nu_1^{I^2}$ asociado a un único vector normado ν_1 de $\mathbf{R}^I$, de forma que: $\nu_1^{I^2} = \nu_1 {\nu_1}' \mathcal{D}$ siendo ν_1 el primer vector propio obtenido en el análisis global. Estos vectores pueden ser vistos como un grupo de una sola variable, entonces se le puede asociar a una matriz de distancias entre individuos y a un vector de $\mathbf{R}^{I^2}$.

El criterio consiste en maximizar la suma de las proyecciones de las $\mathbf{W}_k$ sobre $\nu_1^{I^2}$. Luego se buscará un segundo vector $\nu_2^{I^2}$ ortogonal al primero y así sucesivamente.

La proyección de $\mathbf{W}_k$ sobre $\nu_1^{I^2}$, es:

$$\left\langle \mathbf{W}_k\mathcal{D}, \nu_1^{I^2} \right\rangle_{HS} = \left\langle \mathbf{W}_k\mathcal{D}, \nu_1\nu_1'\mathcal{D} \right\rangle_{HS}$$

$=$ proyección de $\mathbf{W}_k\mathcal{D}$ sobre $\nu_1^{I^2}$= coordenada de $\mathbf{W}_k\mathcal{D}$ sobre $\nu_1\nu_1'\mathcal{D}$

$= traza(\mathbf{X}_k\mathbf{M}_k{\mathbf{X}_k}'\mathcal{D}\nu_1\nu_1'\mathcal{D}) = traza(\nu_1'\mathcal{D}\mathbf{X}_k\mathbf{M}_k{\mathbf{X}_k}'\mathcal{D}\nu_1) = \left\|\mathbf{X}_k'\mathcal{D}\nu_1\right\|^2_{M_k}$

$=$ la proyección del *k-ésimo* grupo sobre ν_1 primer vector propio del análisis global.

Luego podemos decir que el grupo formado por la variable $\nu_1 \in \mathbf{R}^I$ tiene su representación en $\mathbf{R}^{I^2}$ a través del operador de rango 1.

Además si dos variables son ortogonales en $\mathbf{R}^I$, los vectores asociados en $\mathbf{R}^{I^2}$ también lo son.

El mismo proceso se repite para los R primeros vectores normados del Análisis Global, esto es: $\left\|\mathbf{X}_k'\mathcal{D}\nu_r\right\|^2_{M_k}$ = la proyección del *k-ésimo* grupo sobre ν_r para *r=1,...,R.*

Una característica importante es que los ejes de la Inter-estructura en el AFM son los mismos que los obtenidos en el estudio de la Intra-estructura, ya que los puntos que representan cada tabla o grupo de variables son los obtenidos en el Análisis Global y por lo tanto los mismos pueden ser interpretados en términos de las variables originales.

2.2.3.2. Medida de calidad de la representación de cada grupo de variables

El criterio utilizado por medir la asociación entre una variable ν_r y un grupo de variables J_k se define por:

$$L(\nu_r, J_k) = \left\| \mathbf{X}_k' \mathcal{D} \nu_r \right\|_{M_k}^2$$

y vimos que corresponde a la proyección del *k-ésimo* grupo sobre ν_r, para *r=1,...,R.*

El valor máximo que puede tomar este índice es 1, dada la normalización del AFM, y alcanzará este valor sólo cuando se lo considere con el primer vector propio del análisis general, es decir cuando *r=1.*

Es importante observar que si este índice se calcula para todos los grupos de variables, se tendrá la variabilidad de todos los grupos que es explicada por el *r-ésimo* vector:

$$L(\nu_r, J_k) = \sum_{k=1}^{K} \left\| \mathbf{X}_k' \mathcal{D} \nu_r \right\|_{M_k}^2$$

Esta expresión representa la inercia en $\mathbf{R}^I$ de la proyección de todos los grupos de variables sobre ν_r, es decir, la variabilidad de $\mathbf{X}$ que es explicada por el *r-ésimo* factor.

Se debe observar que si las variables son continuas, centradas y reducidas este índice es:

$$L(\nu_r, J_k) = \sum_{k=1}^{K} \{r(\nu_r, J_k)\}$$

Si las variables son cualitativas:
$L(\nu_r, J_k)$ = Inercia proyectada de $\mathbf{X}_k$ sobre ν_r.

En resumen para comparar la contribución de cada grupo en los diferentes ejes, se calcula la contribución de cada grupo J_k relativa a los ejes α y se puede hallar mediante alguna de las siguientes expresiones puesto que dan el mismo valor:

- en $\mathbf{R}^J$ está dada por:

$$\frac{\sum_{i=1}^{I_k} a_{i_k\alpha}^2}{\sum_i a_{i\alpha}^2}$$

  para a_{i_k} la coordenada del individuo i del grupo J_k

- en $\mathbf{R}^I$ el valor corresponde a:

$$\frac{\sum_{j=1}^{J_k} b_{j_k\alpha}^2}{\sum_{j=1}^{J} b_{j\alpha}^2}$$

  donde b_{j_k} es la coordenada de la *j-ésima* variable del grupo J_k.

- en $\mathbf{R}^{I^2}$ el valor se calcula mediante:

$$\frac{a_{k\alpha}}{\sum_{k=1}^{K} a_{k\alpha}}$$

  donde a_k es la coordenada del grupo J_k.

2.3. El Método Statis

El Método STATIS: Structuration de Tableaux A Trois Indices de la Statisque fue desarrollado por L'HERMIER des PLANTES (1976), LAVIT (1988) y LAVIT et al. (1994). Las bases teóricas de este método fueron desarrolladas por ESCOUFIER (1973, 1976).

El método permite la exploración simultánea de tablas múltiples, donde al mismo conjunto de individuos I, se les ha medido en K condiciones o situaciones experimentales diferentes; y las variables pueden ser las mismas o no, en las distintas condiciones o tablas.

El objetivo del anális es encontrar una estructura común o representativa a todas las tablas o grupos de variables. Dado que las tablas tienen los mismos individuos este análisis privilegia la posición de estos.

Al igual que en el AFM se consideran las matrices de pesos, $\mathbf{M}_k$ para las variables de la tabla $\mathbf{X}_k$ y $\mathcal{D}$ para los individuos, es decir que estamos considerando el estudio de $(\mathbf{X}_k, \mathbf{M}_k, \mathcal{D})$

Dado que se tiene los mismos individuos en las K tablas, el estudio se va a caracterizar por las matrices $\mathbf{W}_k$ de tamaño IxI, a los que posteriormente nos referimos como objetos, donde $\mathbf{W}_k = \mathbf{X}_k\mathbf{M}_k\mathbf{X}_k^{'}$ va a representar la tabla k. El método se basa sobre estos objetos dado que contienen todas las asociaciones inter individuos.

Otro objeto que se considera dentro del análisis corresponde a $\frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_{HS}}$, para la norma definida por el producto escalar de Hilbert Schmith. El objetivo de trabajar con los objetos normados es eliminar el efecto preponderante que, las tablas o grupos de variables con norma elevada, puedan tener en la construcción de la estructura consenso.

2.3.1. Interestructura

A partir de los productos escalares entre los objetos $\mathbf{W}_k$ se construye la matriz $\mathcal{S}$ de tamaño KxK:

$$\mathcal{S} = \begin{bmatrix} \ddots & & \ddots \\ & S_{kl} = \langle \mathbf{W}_k, \mathbf{W}_l \rangle_{HS} & \\ \ddots & & \ddots \end{bmatrix}$$

Cuando las tablas son representadas por los objetos normados $\frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_{HS}}$, se tiene:

$$\tilde{\mathcal{S}} = \begin{bmatrix} 1 & \dots & RV(1,k) & \dots & RV(1,K) \\ \vdots & & \vdots & & \vdots \\ RV(k,1) & \dots & 1 & \dots & RV(k,K) \\ \vdots & & \vdots & & \vdots \\ RV(K,1) & \dots & RV(K,k) & \dots & 1 \end{bmatrix}$$

Los índices RV son útiles en la interpretación de la interestructura debido a que:

- Su valor corresponde a una medida de similitus entre dos estudios normados
- Cuando $RV(k,l)$ es uno, significa que los estudios $\mathbf{X}_k\mathbf{M}_k\mathbf{X}_k'$ y $\mathbf{X}_l\mathbf{M}_l\mathbf{X}_l'$ tienen imagen euclidea equivalente para los individuos, es decir, que las nubes de individuos son homotéticas.
- Además, si $RV(k,l)$=0, las variables de $\mathbf{X}_k$ tienen covarianza nula con las variables de $\mathbf{X}_l$.

Entonces, las matrices $\mathcal{S}$ y $\tilde{\mathcal{S}}$ permiten predecir la proximidad dos a dos, entre las K matrices de datos, dando una medida de la similitud entre dos grupos de variables, para el mismo grupo de individuos. Hecho que permite una interpretación similar a la de la Matriz de Correlaciones entre variables.

2.3.1.1. Construcción de la imágen euclidea

Consiste en buscar un subespacio que reproduzca los productos escalares entre los objetos que representan las tablas.

Entonces se realiza la descomposición espectral de:

$$\mathcal{S} = \mathbf{U}_{\mathcal{S}}\mathbf{\Lambda}_{\mathcal{S}}^{\mathbf{2}}\mathbf{U}_{\mathcal{S}}'$$

donde la matriz $\mathbf{A}_{W_K} = \mathbf{U}_{\mathcal{S}}\mathbf{\Lambda}_{\mathcal{S}}$ contiene las coordenadas de los objetos $\mathbf{W}_1 \dots \mathbf{W}_K$ representandos en la imagen euclidea. Si se desea dar diferente importancia a las tablas, se utiliza la matriz de pesos Δ, matriz diagonal de tamaño KxK, de lo contrario esta matriz, es la identidad.

Para visualizar los resultados se hace una aproximación de la imagen en el plano, lo que se logra restringiendo la matriz $\mathbf{A}_{W_K}$ a sus dos primeras coordenadas. Está representación es tal, que la distancia entre dos puntos $\mathbf{A}_{W_k}$ y $\mathbf{A}_{W_l}$ es la mejor con la distancia de Hilbert Schmidt, entre los objetos $\mathbf{W}_k$ y $\mathbf{W}_l$ representativos de las tablas.

Esta representación permite visualizar la interestructura en el plano, y por tanto, si dos puntos están proximos indica una estructura común de los individuos de las tablas correspondientes.

Considerando el teorema de Frobenius: Una matriz simétrica que tiene todos sus términos positivos admite un primer vector propio con todas sus componentes positivas; y aplicándolo a la matriz $\mathcal{S}$, significa que la diferenciación de las $\mathbf{A}_{W_K}$ imagenes euclideas, está dada solo por el eje 2 porque en el primer eje, todas las coordenadas de los puntos son positivas. (ver figura 2.2).

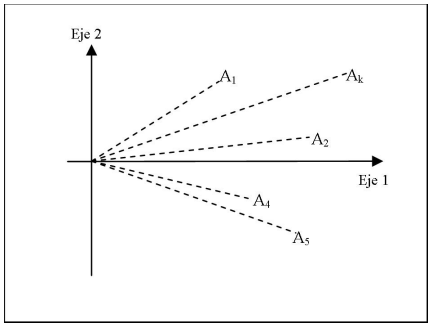

Figura 2.2: Representación euclidea de las matrices iniciales

Si conectamos cada punto con el origen de coordenadas, tendremos una estimación de la correlación o de los cosenos de los ángulos que forman los objetos representativas de cada tabla.

Es posible incluir en la imagen euclidea, tablas que no intervienen en el análisis ni en la determinación de los ejes. Para este fin se realiza el análisis incluyendo estas tablas y dandoles un peso nulo, lo que hace que no intervengan en la construcción de la imagen.

Teniendo en cuenta que el análisis de la interestructura por medio de la representación gráfica, permite tomar decisiones acerca de normar los objetos y de la consistencia del compromiso, vamos a considerar algunos ejemplos para ilustrar diferentes situaciones.

Los estudios considerados en la figura 2.3 son caracterizados por objetos $\mathbf{W}_k$ de normas similares y coeficientes RV elevados. Se puede concluir que hay una estructura de individuos común a las tablas y que la estructura queda bien descrita por el consenso.

La figura 2.4 muestra que la tabla $\mathbf{X}_1$ tiene una estructura de individuos

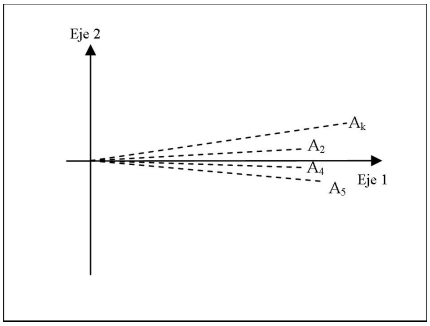

Figura 2.3: Presencia de estructura comun entre las tablas

diferente a las otras tablas y por tanto el objeto $\mathbf{W}_1$ interviene poco en la construcción del consenso. Lo que justifica que se denomine consenso mayoritario.

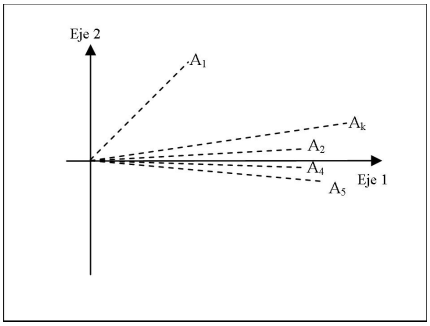

Figura 2.4: El objeto $\mathbf{W}_1$ no está bien representado por el consenso

Graficamente se puede apreciar que las normas de los objetos no tienen valores similares y que solo los objetos de normas elevadas contribuyen a la construcción del consenso.
En este caso es preferible realizar el análisis de la interestructura con los objetos normados: $\frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_{HS}}$

En el siguiente caso (ver figura 2.6) los objetos son diferentes y no presentan estructura común para los individuos. Los cosenos o coeficientes RV son pequeños y el consenso es solo una media ponderada de los objetos, que está correlacionado con estos pero que no los caracteriza.

Como se trata de representar gráficamente la información en un subespacio

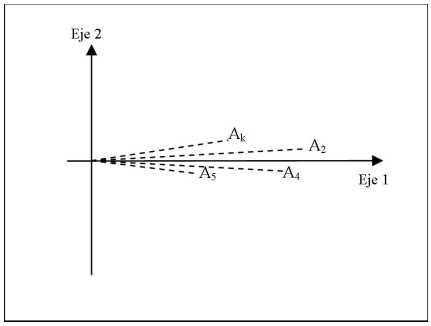

Figura 2.5: El consenso no es buen resumen de los objetos debido a que sus normas son muy diferentes

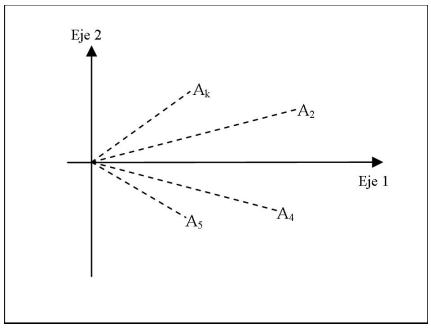

Figura 2.6: El consenso no es buen resumen de los objetos debido a que las tablas no presentan estructura común

de dimensión menor, sólo se pueden interpretar los elementos que están bien representados, por tanto, es importante evaluar la calidad de representación de las configuraciones representativas de las distintas tablas.

2.3.1.2. Calidad de representación

Dado que el proceso que se realiza para hallar los ejes correspondientes a la representación de la Interestructura, es un análisis en componentes principales, se puede calcular la calidad del representación del objeto $\mathbf{W}_k$, en los dos primeros ejes, mediante:

$$CRT_k = \frac{a_{1_k}^2 + a_{2_k}^2}{\sum_{\alpha=1}^{K} a_{\alpha_k}^2}$$

donde $a_k = (a_{1_k}, a_{2_k})$ es la coordenada del objeto k sobre los dos primeros ejes, y $\sum_{\alpha=1}^{K} a_{\alpha_k}^2$ es la suma de las coordenas del objeto k en todos los ejes.

Como los ejes de la interestructura no son interpretables, la contribución de los objetos a la formación de los ejes no se calcula.

2.3.2. Consenso

El análisis de la Interestructura permite ver la semejanza de las diferentes tablas. Si efectivamente las tablas se parecen porque contienen estructuras de covariación similares entre sus variables, o estructuras de similaridad entre sus individuos comparables, es importante la construcción de un espacio consenso, de la misma naturaleza de los objetos para representarlos, ya que el consenso debe ser un buen resumen de los mismos.

En el Statis el consenso se define como una media ponderada de los objetos.

2.3.2.1. Consenso a partir de objetos no normados

Denotando $\mathbf{W}$ al objeto que genera el espacio consenso, se tiene que:

$$\mathbf{W} = \sum_{k=1}^{K} \alpha_k \mathbf{W}_k$$

Donde los factores de ponderación α_k, están dados por:

$$\alpha_k = \frac{1}{\lambda_{1_{W_k}}} \left[\sum_{k=1}^{K} \sqrt{s_{kk}} \right] \mathbf{u}_{1k_s}$$

Donde:
$\lambda_{1_{W_k}}$: es la raiz cuadrada del primer valor propio del objeto $\mathbf{W}_k$
s_{kk}: es el *k-ésimo* valor de la diagonal de la matriz $\mathcal{S}$.
$\mathbf{u}_{1k_s}$: es la *k-ésima* componente del primer vector propio de la matriz $\mathcal{S}$.

El objeto consenso $\mathbf{W}$ es el más correlacionado con todos los objetos $\mathbf{W}_k$, de acuerdo al producto escalar de Hilbert-Schmidt y es de la misma naturaleza que estos.

En caso que el consenso se construya con objetos $\mathbf{W}_k$ que tengan norma alta, el compromiso estará influenciado por estos objetos, en cambio cuando

se construye con los objetos normados, no tendrá en cuenta la variabilidad de las distintas tablas.

2.3.2.2. Consenso a partir de objetos normados

Si en el anális intervienen los objetos normados el consenso está definido como una media ponderada de los objetos: $\frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_{HS}}$

Y el objeto consenso es:

$$\mathbf{W} = \sum_{k=1}^{K} \alpha_{k'} \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_{HS}}$$

reemplando el valor de la norma:

$$\mathbf{W} = \sum_{k=1}^{K} \alpha_{k'} \frac{\mathbf{W}_k}{\sqrt{\sum_s (\lambda_s^2)^k}}$$

Y los factores de ponderación $\alpha_{k'}$, están dados por:

$$\alpha_{k'} = \frac{1}{\lambda_{1_{W_k}}} \mathbf{u}_{1k_s}$$

Donde:
$\lambda_{1_{W_k}}$: es la raiz cuadrada del primer valor propio del objeto $\mathbf{W}_k$
$\mathbf{u}_{1k_s}$: es la *k-ésima* componente del primer vector propio de la matriz $\mathcal{S}$.

2.3.2.3. Construcción del espacio consenso

El objeto consenso $\mathbf{W}$ es una matriz IxI, con $\mathcal{D}$ matriz de pesos para los individuos y diagonalizando :

$$\mathbf{W}\mathcal{D} = \mathbf{U}_W \mathbf{\Lambda}_W^2 \mathbf{U}_W^{'}$$

se obtiene:

- la matriz $\mathbf{U}_W$ de las componentes estándar que generan el subespacio.
- La matriz $\mathbf{A}_W = \mathbf{U}_W \mathbf{\Lambda}_W$ que contiene las imagenes euclideas consenso de los individuos.

Además

$$\mathbf{W}\mathcal{D} = [\mathbf{U}_W\mathbf{\Lambda}_W]\,[\mathbf{U}_W\mathbf{\Lambda}_W]^{'} = [\mathbf{U}_W\mathbf{\Lambda}_W]^2$$

$$= \sum_{k=1}^{K} \alpha_k \mathbf{W}_k\mathcal{D} = \sum_{k=1}^{K} [\sqrt{\alpha_k}\mathbf{U_k}\mathbf{\Lambda_k}]^2$$

Por tanto:

$$[\mathbf{U}_W\mathbf{\Lambda}_W]^2 = \sum_{k=1}^{K} [\sqrt{\alpha_k}\mathbf{U}_k\mathbf{\Lambda}_k]^2$$

y a partir de la DVS de las tablas, se tiene que:

$$\sum_{k=1}^{K} \sqrt{\alpha_k}\mathbf{X_k} = \sum_{k=1}^{K} \sqrt{\alpha_k}\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k^{'}$$

Lo que significa que la imágen euclidea $\mathbf{A}_W = \mathbf{U}_W\mathbf{\Lambda}_W$, también se puede obtener a partir del análisis en componentes principales de tabla construída por yuxtaposición de la tablas iniciales ponderadas por $\sqrt{\alpha_k}$, es decir, de la tabla:

$$\left[\ \sqrt{\alpha}_1\mathbf{X}_1 \quad \ldots \quad \sqrt{\alpha}_k\mathbf{X}_k \quad \ldots \quad \sqrt{\alpha}_K\mathbf{X}_K\ \right]$$

2.3.3. Intraestructura

La imagen consenso de los individuos está compuesta de los puntos $A_1, \ldots, A_I$ cuyas coordenadas están contenidas en la matriz $\mathbf{A}_W = \mathbf{U}_W\mathbf{\Lambda}_W$. Escribiendo las coordenadas $\mathbf{A}_W$ en términos del objeto consenso, se tiene

$$\mathbf{W}\mathcal{D} = \mathbf{U}_W\mathbf{\Lambda}_W^2\mathbf{U}_W^{'}$$

$$\mathbf{W}\mathcal{D}\mathbf{U}_W\mathbf{\Lambda}_W^{-1} = \mathbf{U}_W\mathbf{\Lambda}_W$$

$$\mathbf{W}\mathcal{D}\mathbf{U}_W\mathbf{\Lambda}_W^{-1} = \mathbf{A}_W$$

2.3.3.1. Distancia entre dos puntos

En esta imágen, la distancia compromiso entre dos puntos A_i, A_j se puede interpretar como la distancia promedio entre los individuos originales x_i, x_j en las ocasiones estudiadas.

Esta distancia, expresada en función de los x_i, x_j, es:

$$d^2(A_i, A_j) = \sum_{k=1}^{K} \alpha_k \left\| \mathbf{x}_{i_k} - \mathbf{x}_{j_k} \right\|^2_{M_k}$$

Donde los α_k son los coeficientes de ponderación de cada objeto que conforman el objeto compromiso.

2.3.3.2. Correlación de las variables iniciales con los ejes del compromiso

Para las J_k variables iniciales, centradas y reducidas de la tabla $\mathbf{X}_k$, la correlación entre estas y las componentes estándar que generan el espacio compromiso, está dada por:

$$corr(J_k, \mathbf{U}) = \mathbf{X}_k^{'} \mathcal{D} \mathbf{U}_W$$

Estas correlaciones se resumen en un gráfico, donde cada variable J_k está representada por un punto en el eje q, cuya coordenada es la correlación entre la variable y el eje q.

2.3.3.3. Trayectorias

La representación de las trayectorias en la imagen euclidea consenso, consiste en representar en esta imagen las K nubes de individuos. Así se obtiene una representación de los IK puntos de individuos, donde los K puntos de un individuo corresponden a su trayectoria.

El método utilizado se deduce de la obtencion de las coordendas de los puntos compromiso $A_1, \ldots, A_I$ que se hace por:

$$\mathbf{W} \mathcal{D} \mathbf{U}_W \mathbf{\Lambda}_W^{-1} = \mathbf{A}_W$$

Considerando que cada objeto se puede ubicar como suplementario, las coordenadas para los individuos de la tabla $\mathbf{X}_k$, se obtienen por medio de:

$$\mathbf{W}_k \mathcal{D} \mathbf{U}_W \mathbf{\Lambda}_W^{-1} = \mathbf{A}_k$$

De esta manera se ubican los individuos de la tabla $\mathbf{X}_k$ en el espacio consenso, sin que influencien la construcción del espacio consenso.

Además, el individuo compromiso A_i, es el centro de gravedad de los puntos $A_{i_1}, \ldots, A_{i_K}$ ponderados por los coeficientes $\alpha_1, \ldots, \alpha_K$ y esta propiedad se conserva en la proyección.

2.4. El Método Statis Dual

El Statis Dual es un método que a diferencia del Statis privilegia las posiciones relativas de las variables.

La información de análisis está constituida por K tablas denotadas $\mathbf{X}_k$ que tienen las mismas J variables cuantitativas medidas sobre I individuos eventualmente diferentes. De manera similar al STATIS, el objetivo del anális es encontrar una estructura común a todas las tablas o grupos de individuos.

Se consideran las matrices de pesos, $\mathbf{M}$ para las variables y $\mathcal{D}_k$ para los individuos de la tabla $\mathbf{X}_k$, es decir que estamos considerando el estudio de $(\mathbf{X}_k, \mathbf{M}, \mathcal{D}_k)$

Dado que se tiene los mismas variables en las K tablas, el estudio se va a caracterizar por las matrices $\mathbf{S}_k$ de tamaño J x J. Donde $\mathbf{S}_k = \mathbf{X}_k^{'}\mathcal{D}_k\mathbf{X}_k$ corresponde a la matriz de varianzas y covarianzas de la tabla $\mathbf{X}_k$.

Otro objeto que se considera dentro del análisis corresponde a $\frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}}$, para la norma definida por el producto escalar de Hilbert Schmith. De manera similar al STATIS, el consenso asociado a los objetos $\mathbf{S}_k$ está influenciado por aquellos que tienen norma elevada, y el consenso asociado a las configuraciones normadas, no tiene en cuenta las diferencias de estructuras entre las distintas tablas.

2.4.1. Interestructura

A partir de los productos escalares entre los objetos $\mathbf{S}_k$ se construye la matriz $\mathcal{Z}$ de tamaño KxK:

$$\mathcal{Z} = \begin{bmatrix} \ddots & & \ddots \\ & Z_{kl} = \langle \mathbf{S}_k, \mathbf{S}_l \rangle_{HS} & \\ \ddots & & \ddots \end{bmatrix}$$

Igualmente se podrá trabajar con los objetos normados $\frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}}$, se tiene:

$$\tilde{\mathcal{Z}} = \begin{bmatrix} \ddots & & \ddots \\ & \tilde{Z}_{kl} = \left\langle \frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}}, \frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}} \right\rangle_{HS} & \\ \ddots & & \ddots \end{bmatrix}$$

Se debe observar que en este caso la matriz no es la matriz de los coeficientes de Correlación Vectorial RV, ya que dicho coeficiente mide el grado de relación entre dos objetos respecto al mismo conjunto de individuos.

2.4.1.1. Construcción de la imágen euclidea

Se trata de buscar un subespacio que reproduzca las productos escalares entre los objetos, es decir, que reproduzca la matriz de varianzas y covarianzas de cada una de las K tablas.

Con este objetivo se diagonaliza $\mathcal{Z}$ ó $\tilde{\mathcal{Z}}$

Si trabajamos con $\mathcal{Z}$ y conociendo que su descomposición espectral es:

$$\mathcal{Z} = \mathbf{V}_{\mathcal{Z}} \mathbf{\Lambda}_{\mathcal{Z}}^2 \mathbf{V}_{\mathcal{Z}}^{'}$$

donde la matriz $\mathbf{A}_{S_K} = \mathbf{V}_{\mathcal{Z}} \mathbf{\Lambda}_{\mathcal{Z}}$ contiene las coordenadas de los objetos $\mathbf{S}_1 \dots \mathbf{S}_K$ representandos en la imagen euclidea. Si se desea dar diferente importancia a las tablas, se utiliza la matriz de pesos Δ, matriz diagonal de tamaño KxK, de lo contrario esta matriz, es la identidad.

2.4.1.2. Calidad de representación

De manera análoga al método STATIS, se calcula la calidad de representación del objeto $\mathbf{S}_k$, en los dos primeros ejes, mediante:

$$CRT_k = \frac{a_1^2 + a_2^2}{\sum_{\alpha=1}^{K} a_\alpha^2}$$

donde $a = (a_1, a_2)$ es la coordenada del objeto k sobre los dos primeros ejes, y $\sum_{\alpha=1}^{K} a_\alpha^2$ es la suma de las coordenas del objeto k en todos los ejes.

Como los ejes de la interestructura no son interpretables, la contribución de los objetos a la formación de los ejes no se calcula.

2.4.2. Consenso

El objeto consenso se define como una media ponderada de los objetos:

$$\mathbf{S}_k \text{ ó } \frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}}$$

Como el objeto consenso es es la combinación lineal más correlacionada con los objetos $\mathbf{S}_k$, de acuerdo al producto escalar de Hilbert-Schmidt, entonces el consenso es una matriz de varianzas y covarianzas consenso entre las variables que denotaremos $\mathbf{S}$.

2.4.2.1. Consenso a partir de objetos no normados

El consenso se define como:

$$\mathbf{S} = \sum_{k=1}^{K} \beta_k \mathbf{S}_k$$

Donde los factores de ponderación β_k, están dados por:

$$\beta_k = \frac{1}{\lambda_{1_{S_k}}} \left[\sum_{k=1}^{K} \sqrt{Z_{kk}} \right] \mathbf{v}_{1k_z}$$

Donde:

$\lambda_{1_{S_k}}$: es la raiz cuadrada del primer valor propio del objeto $\mathbf{S}_k$
Z_{kk}: es el *k-ésimo* valor de la diagonal de la matriz $\mathcal{Z}$.
$\mathbf{v}_{1k_z}$: es la *k-ésima* componente del primer vector propio de la matriz $\mathcal{Z}$.

2.4.2.2. Consenso a partir de objetos normados

Cuando es necesario normar los objetos, el consenso esta dado por:

$$\mathbf{S} = \sum_{k=1}^{K} \beta_k \frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}}$$

Y los factores de ponderación β_k, están dados por:

$$\beta_k = \frac{1}{\lambda_{1_{S_k}}} \mathbf{v}_{1k_s}$$

Donde:

$\lambda_{1_{S_k}}$: es la raiz cuadrada del primer valor propio del objeto $\mathbf{S}_k$
$\mathbf{v}_{1k_s}$: es la *k-ésima* componente del primer vector propio de la matriz $\mathcal{Z}$.

En el Statis Dual el compromiso es el objeto que representa mejor las relaciones entre las variables definidas en el conjunto de tablas originales y puede considerarse como una matriz de varianzas y covarianzas promedio, sobre las distintas tablas.

2.4.2.3. Construcción del espacio consenso

El objeto consenso $\mathbf{S}$ es una matriz JxJ, con $\mathbf{M}$ matriz de pesos para las variables y diagonalizando :

$$\mathbf{SM} = \mathbf{V}_S \mathbf{\Lambda}_S^2 \mathbf{V}_S^{'}$$

se obtiene:

- la matriz $\mathbf{V}_S$ de las componentes estándar que generan el subespacio.
- La matriz $\mathbf{B}_S = \mathbf{V}_S \mathbf{\Lambda}_S$ que contiene las imagenes euclideas consenso para las variables.

Además

$$\mathbf{SM} = [\mathbf{V}_S \mathbf{\Lambda}_S] [\mathbf{V}_S \mathbf{\Lambda}_S]^{'} = [\mathbf{V}_S \mathbf{\Lambda}_S]^2$$

$$= \sum_{k=1}^{K} \beta_k \mathbf{S}_k \mathbf{M} = \sum_{k=1}^{K} \left[\sqrt{\beta_k} \mathbf{V}_k \mathbf{\Lambda}_k \right]^2$$

Por tanto:

$$[\mathbf{V}_S \mathbf{\Lambda}_S]^2 = \sum_{k=1}^{K} \left[\sqrt{\beta_k} \mathbf{V}_k \mathbf{\Lambda}_k \right]^2$$

y a partir de la DVS de las tablas iniciales traspuestas y ponderadas, se tiene que:

$$\sum_{k=1}^{K} \sqrt{\beta_k} \mathbf{X}_k^{'} = \sum_{k=1}^{K} \sqrt{\beta_k} \mathbf{V}_k \mathbf{\Lambda}_k \mathbf{U}_k^{'}$$

Lo que significa que la imágen euclidea $\mathbf{B}_S = \mathbf{V}_S \mathbf{\Lambda}_S$, también se puede obtener a partir del análisis en componentes principales de tabla construída

por superposición de la tablas iniciales ponderadas por $\sqrt{\beta_k}$, es decir, de la tabla:

$$\begin{bmatrix} \sqrt{\beta}_1\mathbf{X}_1 \\ \vdots \\ \sqrt{\beta}_k\mathbf{X}_k \\ \vdots \\ \sqrt{\beta}_K\mathbf{X}_K \end{bmatrix}$$

2.4.3. Intraestructura

La imagen consenso de las variables, corresponde a los puntos $B_1, \ldots, B_I$ cuyas coordenadas están contenidas en la matriz $\mathbf{B}_S = \mathbf{V}_S\mathbf{\Lambda}_S$. Escribiendo las coordenadas $\mathbf{B}_S$ en términos del objeto consenso, se tiene

$$\mathbf{SM} = \mathbf{V}_S\mathbf{\Lambda}_W^2\mathbf{V}_S^{'}$$

$$\mathbf{SMV}_S\mathbf{\Lambda}_S^{-1} = \mathbf{V}_S\mathbf{\Lambda}_S$$

$$\mathbf{SMV}_S\mathbf{\Lambda}_S^{-1} = \mathbf{B}_S$$

2.4.3.1. Covarianza entre variables

En esta imágen, la covarianza entre dos variables compromiso B_l, B_m se puede interpretar como la covarianza promedio de las variables de cada tabla $\mathbf{x}_l, \mathbf{x}_m$ en las ocasiones estudiadas.

Esta covarianza, expresada en función de las variables iniciales $\mathbf{x}_l, \mathbf{x}_m$, es:

$$cov(B_i, B_j) = \sum_{k=1}^{K} \beta_k cov(\mathbf{x}_l, \mathbf{x}_m)$$

Donde los β_k son los coeficientes de ponderación de cada objeto que conforman el objeto compromiso.

2.4.3.2. Calidad de representación

Considerando que el proceso que se realiza para hallar los ejes correspondientes a la representación de la Interestructura, es un análisis en componentes

principales, se puede calcular la calidad del representación del objeto $\mathbf{S}_k$, por medio de su imagen euclidea B_k en los dos primeros ejes, mediante:

$$CRT_k = \frac{b_1^2 + b_2^2}{\sum_{\alpha=1}^{K} b_\alpha^2}$$

donde $b = (b_1, b_2)$ es la coordenada del objeto k sobre los dos primeros ejes, y $\sum_{\alpha=1}^{K} b_\alpha^2$ es la suma de las coordenas del objeto k en todos los ejes.

También se puede encontrar la contribución relativa del elemento compromiso B_l, al *q-ésimo* factor del espacio de representación de la intraestructura y se interpreta como la parte de la variabilidad del factor q explicada por la variable compromiso B_l.

$$CRF_q E_l = \frac{b_q^2}{{\lambda_q}^2}$$

2.4.3.3. Trayectorias

En una sección precedente vimos como se obtienen las coordenadas para la representación de las variables compromiso, pero en realidad el objetivo es proyectar el conjunto de las JK variables originales, es decir, se quiere observar el comportamiento de las variables observadas a través de las distintas tablas.

El Statis Dual considera las variables de cada tabla como elementos suplementarios, para obtener las trayectorias de las variables sobre la imagen euclídea compromiso.

El método utilizado se deduce de la obtencion de las coordendas de los puntos compromiso $B_1, \ldots, B_I$ que se hace por:

$$\mathbf{SMV}_S\mathbf{D}_S^{-1} = \mathbf{B}_S$$

Considerando que cada objeto se puede ubicar como suplementario, las coordenadas para las variables de la tabla $\mathbf{X}_k$, se obtienen por medio de:

$$\mathbf{S}_k\mathbf{MV}_S\mathbf{D}_S^{-1} = \mathbf{B}_k$$

De esta manera se ubican los individuos de la tabla $\mathbf{X}_k$ en el espacio consenso, sin que influencien la construcción del espacio.

2.5. El Método DACP

El Doble Análisis en componentes principales, que se denota con las siglas DACP fué introducido por J. M. Bouroche en 1975 y retomado por F. Dazy y J. F. Barzic en 1996

Este método se aplica a datos donde el mismo conjunto de variables cuantitativas ha sido medido sobre el mismo conjunto de individuos, en diferentes instantes. Es decir, se aplica al análisis de datos denominados cubicos.
Bouroche propone este análisis para el estudio de datos donde la tercera dimensión es el tiempo y donde busca evolución de los datos.

La información de análisis está constituida por K tablas denotadas $\mathbf{X}_k$ que tienen las mismas J variables cuantitativas medidas sobre los mismos I individuos. El objetivo del DACP, al igual que los métodos que estudian tablas múltiples, es comparar globalmente la evolución del grupo de variables y la evolución del conjunto de individuos.

Dado que se tienen las mismas variables y los mismo individuos para todas las tablas, se consideran las matrices de pesos, $\mathbf{M}$ para las variables y $\mathcal{D}$ para los individuos, es decir que estamos considerando el estudio de $(\mathbf{X}_k, \mathbf{M}, \mathcal{D})$

En el DACP el estudio se va a caracterizar por:

- Las matrices $\mathbf{S}_k$ de tamaño $J \times J$. Donde $\mathbf{S}_k = \mathbf{X}_k^{'}\mathcal{D}\mathbf{X}_k$ corresponde a la matriz de varianzas y covarianzas de la tabla $\mathbf{X}_k$ previamente centrada.

- Las matrices $\mathbf{W}_k$ de tamaño $I \times I$. Donde $\mathbf{W}_k = \mathbf{X}_k\mathbf{M}\mathbf{X}_k^{'}$ corresponde a la matriz de asociaciones entre los individuos de la tabla $\mathbf{X}_k$ previamente centrada.

2.5.1. Interestructura

En esta parte el DACP busca estudiar la evolución de las tablas por medio del centro de gravedad de cada una de ellas.

Para la tabla $\mathbf{X}_k$, denotamos $\mathbf{g}_k^{'} = \bar{x}_{1_k}, \ldots, \bar{x}_{J_k}$ el vector que contiene los centros de gravedad para las J variables.

La matriz $\bar{\mathbf{X}}$ de tamaño $K \times J$, que contiene los centros de gravedad de las

K tablas, es:

$$\bar{\mathbf{X}} = \begin{bmatrix} \mathbf{g}_1^{'} \\ \vdots \\ \mathbf{g}_k^{'} \\ \vdots \\ \mathbf{g}_K^{'} \end{bmatrix} = \begin{bmatrix} \bar{x}_{1_1} & \dots & \bar{x}_{J_1} \\ \vdots & & \vdots \\ \bar{x}_{1_k} & \dots & \bar{x}_{J_k} \\ \vdots & & \vdots \\ \bar{x}_{1_K} & \dots & \bar{x}_{J_K} \end{bmatrix}$$

2.5.1.1. Construcción de la imágen euclídea

Se busca el subespacio de representación de los centros de gravedad de la matriz $\bar{\mathbf{X}}$.

Esto se logra mediante un ACP de la matriz $\bar{\mathbf{X}}$, es decir, mediante la descomposición espectral de:

$$\bar{\mathbf{X}}\bar{\mathbf{X}}' = \mathbf{U_g}\mathbf{\Lambda_g^2}\mathbf{U_g'}$$

donde el primer eje se explica en términos de la evolución global en el tiempo: los centros de gravedad $\mathbf{g}_k$ en general, varían de manera continúa sobre este eje.

La matriz $\mathbf{A_{g_K}} = \mathbf{U_g}\mathbf{\Lambda_g}$ contiene las coordenadas de los objetos $\mathbf{g}_1, \dots, \mathbf{g}_k$, para visualizar los resultados en el plano se restringe la matriz $\mathbf{A_{g_K}}$ a sus dos primeras coordenadas.

2.5.1.2. Calidad de representación

Dado que el proceso que se realiza para hallar los ejes correspondientes a la representación de la Interestructura, es un análisis en componentes principales, se puede calcular la calidad del representación de cada uno de los $\bar{x}_{j_k}$, en los dos primeros ejes, mediante:

$$CRT_{j_k} = \frac{a_{1(j_k)}^2 + a_{2(j_k)}^2}{\sum\limits_{\alpha=1}^{J} a_{\alpha(j_k)}^2}$$

donde $a_{j_k} = (a_{1(j_k)}, a_{2(j_k)})$ es la coordenada de $\bar{x}_{j_k}$ sobre los dos primeros ejes, y $\sum\limits_{\alpha=1}^{J} a_{\alpha(j_k)}^2$ es la suma de las coordenas de $\bar{x}_{j_k}$ en todos los ejes.

Para comparar la contribución de $\bar{x}_{j_k}$ en los diferentes ejes, se calcula su contribución relativa a cada eje, mediante el cociente:

$$CRE_{j_k}F_\alpha = \frac{a^2_{\alpha(j_k)}}{\lambda^2_\alpha} = \frac{a^2_{\alpha(j_k)}}{\sum_{\alpha=1}^{J} a^2_{\alpha(j_k)}}$$

A este valor se denomina contribución relativa del elemento $\bar{x}_{j_k}$ al factor α, y muestra la parte de la variabilidad del factor explicada por el $\bar{x}_{j_k}$.

2.5.2. Consenso

Este método centra su interés en la construcción del elemento que genera el subespacio de representación común.

Una vez analizada la Interestructura que permite ver la evolución de las diferentes tablas a través de tiempo, se propone estudiar estructuras de similaridad entre sus individuos a partir de la deformación de las nubes alrededor de sus centros de gravedad.

Esto consiste en realizar un ACP para las K nubes de individuos, centradas con respecto al centro de gravedad de sus variables, lo que permite eliminar el fenómeno de evolución en el tiempo y hacer un análisis de sus estructuras.

Entonces se efectúan K ACP de las tablas de tamaño $I \times J$, definidas por:

$$\mathbf{Y}_k = \begin{bmatrix} \ddots & & \ddots \\ & y_{ij_k} = x_{ij_k} - \bar{x}_{j_k} & \\ \ddots & & \ddots \end{bmatrix}$$

Para $k = 1, \ldots, K$

Estos K Análisis proveen dos tipos de información:

- Analizar cada una de las tablas a partir de la interpretación del respectivo ACP
- Notando $R < Min(J, I)$, el número de ejes retenido en cada uno de los ACP, se tiene:
  -K matrices $\mathbf{V}_k$ de coordenadas estándar, de tamaño $J \times R$, cuyos vectores de tamaño J serán denotados $(\mathbf{v}_{r_k})_{r=1,\ldots,R}$. Estos vectores

corresponden a los vectores propios de la descomposición espectral de la matriz $\mathbf{S}_k = \mathbf{Y}_k^{'}\mathcal{D}\mathbf{Y}_k$, asociados a los R mayores valores propios $\lambda^2_{r_k\ r=1,\ldots,R}$

-K matrices $\mathbf{A}_k$ de componentes principales, de tamaño $I \times R$, donde $\mathbf{A}_k = \mathbf{Y}_k\mathbf{V}_k$ en términos de la matriz inicial.
También se tiene que $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Lambda}_k$ donde $\mathbf{U}_k$ corresponde a la matriz de vectores propios de la descomposición espectral de la matriz $\mathbf{W}_k = \mathbf{Y}_k M \mathbf{Y}_k^{'}$

2.5.2.1. Índices de selección de ejes del espacio Consenso

A diferencia de los análisis expuestos en los capítulos anteriores donde el objeto que genera el espacio consenso es único para cada método; el DACP propone varios criterios para la construcción del objeto aplicando los dos indicadores que presentamos a continuación.

- *Inercia explicada por un factor*

  Se tiene que para $\mathbf{S}_k = \mathbf{V}_k\mathbf{\Lambda}_k^2\mathbf{V}_k^{'}$ y por tanto $\mathbf{V}_k^{'}\mathbf{S}_k\mathbf{V}_k = \mathbf{\Lambda}_k^2$ entonces $\mathbf{v}_{r_k}^{'}\mathbf{S}_k\mathbf{v}_{r_k} = \lambda^2_{r_k}$ representa la inercia explicada por el eje r para $\mathbf{S}_k$.

  Definición de la inercia explicada por el factor $\mathbf{v}$
  La inercia de $\mathbf{S}_k$ explicada por un factor $\mathbf{v}$ corresponde a la cantidad $\mathbf{v}'\mathbf{S}_k\mathbf{v}$

  Para un conjunto de ejes $(\mathbf{v}_r)_{r=1,\ldots,R}$, y alguna tabla k, se define el índice $\xi(k,\mathbf{v})$, como:

$$\xi(k,\mathbf{v}) = \frac{\sum_{r=1}^{R}\lambda^2_{r_k} - \sum_{r=1}^{R}\mathbf{v}_r^{'}\mathbf{S}_k\mathbf{v}_r}{\sum_{r=1}^{R}\lambda^2_{r_k}}$$

  Este índice mide el porcentaje de inercia que pierde la nube N_I^k de los individuos de la tabla k cuando se proyecta sobre el subespacio definido por $(\mathbf{v}_r)_{r=1,\ldots,R}$ en cambio de proyectarla sobre sus R primeras componentes estándar.

  Esto significa, que cuando la nube de individuos se proyecta sobre el subespacio generado por los ejes $\mathbf{v}$ diferente a los genrados por sus vectores propios, la inercia disminuye en un porcentaje de $\xi(k,\mathbf{v})$

- *Proximidad entre dos factores* $\mathbf{v}$ *y* ν

  La proximidad entre dos factores puede ser medida por el angulo entre los dos vectores $\mathbf{v}$ y ν de dimensión J, más precisamente por el coseno de su ángulo.
  El índice que mide la proximidad entre dos conjuntos de ejes $(\nu_r)_{r=1,\ldots,R}$ y $(\mathbf{v}_{r_k})_{r=1,\ldots,R}$, está dado por:

$$\eta(k, \nu_1, \ldots, \nu_r) = \sum_{r=1}^{R} cos^2(\nu_r, \mathbf{v}_{r_k})$$

  Por tanto, el índice que mide la proximidad entre el conjunto de ejes $(\nu_r)_{r=1,\ldots,R}$ y los K conjuntos de ejes $(\mathbf{v}_{r_k})_{r=1,\ldots,R}$ está dado por:

$$\eta(\nu_1, \ldots, \nu_r) = \sum_{k=1}^{K} \sum_{r=1}^{R} cos^2(\nu_r, \mathbf{v}_{r_k})$$

2.5.2.2. Consenso a partir de la selección del mejor conjunto de ejes

Consiste en seleccionar dentro de los K conjuntos de ejes encontrados, el conjunto que de la menor pérdida de inercia al proyectar todas las tablas sobre el.

Sea κ una de la K tablas y $\mathbf{v}_{R(\kappa)}$ su conjunto de coordenadas estándar, entonces :

$$\xi(., \kappa) = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{\sum_{r=1}^{R} \lambda_{r_k}^2 - \sum_{r=1}^{R} \mathbf{v}'_{r(\kappa)} \mathbf{S}_k \mathbf{v}_{r(\kappa)}}{\sum_{r=1}^{R} \lambda_{r_k}^2} \right]$$

donde:

$$\xi(., \kappa) = \frac{1}{K} \sum_{k=1}^{K} [\xi(k, \kappa)]$$

Entonces el valor $\xi(., \kappa)$ mide el promedio de pérdida de inercia de todas las nubes N_I^k al ser proyectadas sobre los ejes definidos por κ.

Luego, el criterio consiste en seleccionar la tabla κ de entre las K tablas, de tal manera que para el conjunto $\mathbf{v}_{r(\kappa)}$ se tenga, que:

$$\xi(., \kappa) = Min_{(k=1,\ldots,K)} \xi(., k)$$

El espacio consenso es el generado por vectores propios de la matriz $\mathbf{V}_\kappa$ y corresponde al espacio de representación de la tabla κ.

2.5.2.3. Consenso a partir de la maximización de la inercia explicada

Sea $(\mathbf{v}_r)_{r=1,\ldots,R}$ el conjunto buscado, entonces para algún k la inercia de la nube de sus individuos explicada por el sistema, está dada por la cantidad $\sum_{r=1}^{R} \mathbf{v}_r^{'}\mathbf{S}_k\mathbf{v}_r$

El criterio consiste en encontrar el conjunto $(\mathbf{v}_r)_{r=1,\ldots,R}$ que maximice la inercia de todas las nubes al ser proyectadas sobre el. Esto significa resolver

$$Max_{\mathbf{v}_1,\ldots,\mathbf{v}_r} \sum_{k=1}^{K}\sum_{r=1}^{R} \mathbf{v}_r^{'}\mathbf{S}_k\mathbf{v}_r = Max_{\mathbf{v}_1,\ldots,\mathbf{v}_r} \sum_{r=1}^{R} \mathbf{v}_r^{'}\mathbf{S}\mathbf{v}_r$$

para

$$\mathbf{S} = \sum_{k=1}^{K} \mathbf{S}_k$$

La solución consiste en hallar la descomposición espectral de $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{'}$ donde la matriz $\mathbf{V}$ es el conjunto de vectores buscado.

La matriz $\mathbf{S}$ que genera el espacio consenso corresponde a la matriz de inercia de la nube $N_I = \cup_{k\in K} N_I^k$ y puede denominarse *matriz de varianzas y covarianzas consenso*; entonces el espacio consenso está dado por un ACP de la matriz $\mathbf{Y}$ (ver figura 2.7)

2.5.2.4. Consenso a partir de la maximización de la inercia explicada con matrices reducidas

Este criterio planteado por Dazy y Le Barzic (1996), se deduce de los dos anteriores.

el criterio de minimizar la inercia explicada por un conjunto $(\mathbf{v}_r)_r = 1, \ldots, R$

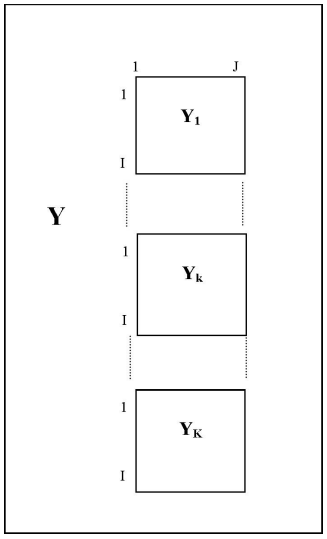

Figura 2.7: Tablas de datos centradas y superpuestas

se expresa mediante el índice:

$$\xi(.,\mathbf{v}) = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{\sum_{r=1}^{R}\lambda_r^2 - \sum_{r=1}^{R}\mathbf{v}_r^{'}\mathbf{S}_k\mathbf{v}_r}{\sum_{r=1}^{R}\lambda_r^2}\right]$$

A partir de $\xi(.,\mathbf{v})$ se deduce el índice que denominaremos $\varphi(.,\mathbf{v})$:

$$\varphi(.,\mathbf{v}) = \sum_{k=1}^{K}\left[\frac{\sum_{r=1}^{R}\mathbf{v}_r^{'}\mathbf{S}_k\mathbf{v}_r}{\sum_{r=1}^{R}\lambda_r^2}\right]$$

donde $\varphi(k,\mathbf{v}) = \dfrac{\sum_{r=1}^{R}\mathbf{v}_r^{'}\mathbf{S}_k\mathbf{v}_r}{\sum_{r=1}^{R}\lambda_r^2}$ representa el porcentaje de inercia de la nube N_I^k explicada por el conjunto $(\mathbf{v}_r)_r = 1,\ldots,R$

Entonces, buscar minimizar $\xi(.,\mathbf{v})$ es equivalente a maximizar $\varphi(.,\mathbf{v})$, que

se puede escribir como:

$$\varphi(.,\mathbf{v}) = \sum_{r=1}^{R} \mathbf{v}_r^{'} \left[\sum_{k=1}^{K} \frac{\mathbf{S}_k}{\sum_{r=1}^{R} \lambda_r^2} \right] \mathbf{v}_r$$

Por tanto el criterio consiste en maximizar $\varphi(.,\mathbf{v})$, donde:

$$\mathbf{S} = \sum_{k=1}^{K} \frac{\mathbf{S}_k}{\sum_{r=1}^{R} \lambda_r^2}$$

es la suma de las matrices $\mathbf{S}_k$ normadas por la aproximación de orden R de su traza. Este método es interesante cuando las matrices $\mathbf{S}_k$ con normas elevadas influencien de manera considerable el sistemas de ejes retenidos. Además esta "ponderación de variables" es similar a la utilizada por el AFM.

2.5.2.5. Consenso a partir de la búsqueda secuencial de un nuevo sistema de ejes

Para conseguir el espacio consenso, este criterio se base en el índice $\eta(\nu_1,\ldots,\nu_r)$ que mide la proximidad entre el conjunto de ejes $(\nu_r)_{r=1,\ldots,R}$ y los K conjuntos de ejes $(\mathbf{v}_{r_k})_{r=1,\ldots,R}$

$$\eta(\nu_1,\ldots,\nu_r) = \sum_{k=1}^{K} \sum_{r=1}^{R} cos^2(\nu_r, \mathbf{v}_{r_k})$$

La búsqueda del conjunto $(\nu_r)_{r=1,\ldots,R}$ se hace de forma secuencial, llevando los siguientes pasos:
Primero se busca ν_1 tal que $\sum_{k=1}^{K} cos^2(\nu_1, \mathbf{v}_{1_k})$ sea máximo.
En el segundo paso: se busca ν_2 del tal manera que $\sum_{k=1}^{K} cos^2(\nu_2, \mathbf{v}_{2_k})$ sea máximo y sujeto a ortogonalidad con ν_1 y asi sucesivamente.
Y finalmente, en el paso R se busca ν_R que maximice $\sum_{k=1}^{K} cos^2(\nu_R, \mathbf{v}_{R_k})$ sujeto a ortogonalidad con el subespacio generado por $(\nu_1,\ldots,\nu_{R-1})$.

Este proceso se puede formalizar de manera analítica:

- Se busca ν_1 tal que

$$\eta(\nu_1) = \sum_{k=1}^{K} cos^2(\nu_1, \mathbf{v}_{1_k}) \text{ sea máximo.}$$

$\eta(\nu_1)$ se puede escribir como:

$$\eta(\nu_1) = \sum_{k=1}^{K} \nu_1^{'} \mathbf{v}_{1_k} \mathbf{v}_{1k}^{'} \nu_1 = \nu_1^{'} \left[\sum_{k=1}^{K} \mathbf{v}_{1k} \mathbf{v}_{1_k}^{'} \right] \nu_1$$

denotando $\mathbf{V}_1 = [\mathbf{v}_{1_1}, \ldots, \mathbf{v}_{1_K}]$ la matriz que tiene por columnas el primer factor de cada una de las K tablas, se tiene que:

$$\eta(\nu_1) = \nu_1^{'} \mathbf{V}_1 \mathbf{V}_1^{'} \nu_1$$

por tanto, ν_1 es el vector propio de $\mathbf{V}_1 \mathbf{V}_1^{'}$ asociado al mayor valor propio.

- En la etapa R, se busca $\eta(\nu_R)$ tal que:
–$\eta(\nu_R/\nu_1, \ldots, \nu_{R-1}) = \sum_{k=1}^{K} cos^2(\nu_R, \mathbf{v}_{R_k})$ sea máximo
– y ortogonal al subespacio generado por $(\nu_1, \ldots, \nu_{R-1})$

Entonces ν_R es el vector propio asociado al mayor valor propio de la matriz

$$\left[\prod_{r=1}^{R-1} (I_J - \nu_r \nu_r^{'}) \right] \mathbf{V}_R \mathbf{V}_R^{'} \left[\prod_{r=1}^{R-1} (I_J - \nu_r \nu_r^{'}) \right]^{'}$$

donde I_J es la matriz identidad de tamaño J y $\mathbf{V}_R = [\mathbf{v}_{R_1}, \ldots, \mathbf{v}_{R_K}]$ la matriz $J \times K$ que tiene por columnas los factores R de las K tablas.

2.5.3. Intraestructura

- Para el método que selecciona el consenso de entre la K tablas, se asume que κ es la tabla seleccionada como objeto genrador del espacio consenso. En este espacio las posiciones consenso de las variables corresponden a las obtenidas en el ACP de la tabla κ y sus contribuciones a cada factor son los calculados en el análisis de está.

- Para los demás métodos:
Las posiciones consenso de las variables se obtienen de la correlación entre los ejes del espacio consenso con las variables; las contribuciones de estas a cada factor, son la coordenada al cuadrado.

2.5.4. Trayectorias

La representación de las trayectorias en la imagen euclidea consenso, consiste en representar en esta imagen las K nubes de individuos.

- Para el método que selecciona el consenso de entre la K tablas:
  Sea κ la matriz consenso, para obtener la representación de los IK puntos de individuos, estos se dejan como individuos suplementarios de la tabla κ, es decir que sus R coordenadas se hallan mediante $\mathbf{A}_k = \mathbf{X}_k\mathbf{V}_{R\kappa}$

  Los individuos i_κ de la tabla κ, corresponden a los individuos consenso con los que se puede comparar la desviación de las trayectorias de los individuos de las otras tablas.

- Para los demás métodos:
  La posición de los IK individuos se calcula considerándolos como individuos suplementarios en este espacio. Denotando $\mathbf{V}_R$ los R primeros vectores propios del objeto consenso, las coordenadas para los individuos de la tabla k, se encuentran en la matriz $\mathbf{A}_k = \mathbf{X}_k\mathbf{V}_R$

En este espacio no hay individuos consenso para comparar la desviación de las trayectorias. Dazy y Le Barzic proponen ubicarlos en el promedio de las trayectorias de los individuos.
Esto significa, que si $(a_{i_k})_\alpha$ es el conjunto de las K coordenadas del individuo i sobre el eje α; el valor $i^*_\alpha = \frac{1}{K}\sum_{k=1}^{K}(a_{i_k})_\alpha$ es la coordenada del individuo consenso sobre dicho eje.

2.6. Meta componentes

Este método fué presentado por W. J. Krzanowski en 1979 y su objetivo es encontrar el espacio consenso para el análisis de K tablas de datos conformadas por I individuos, iguales o diferentes en cada tabla, a los cuales se les ha medido las mismas J variables.

Centra su interés en la construcción del espacio consenso, que realiza a partir de las componentes principales de las tablas que se desea analizar y no propone análisis de la interestructura.

2.6.1. Comparación de dos subespacios

Para ejemplificar el procedimiento se considera el caso donde se tienen dos tablas de datos $\mathbf{X}_1$ y $\mathbf{X}_2$ con I_1, e, I_2 individuos respectivamente, sobre los que se han medido las mismas J variables y los datos se encuentran centrados.

Se realiza un ACP individual de las tablas centradas. Notando $R < Min(J, I)$, el número de ejes retenido en cada uno de los ACP, se tiene:

- Dos matrices $\mathbf{V}_1$ y $\mathbf{V}_2$ de coordenadas estándar, de tamaño $J \times R$, cuyos vectores de tamaño J serán denotados $(\mathbf{v}_{r_k})_{r=1,\ldots,R}$. Estos vectores corresponden a los vectores propios de la descomposición espectral de la matriz $\mathbf{S}_k = \mathbf{X}_k^{'}\mathcal{D}\mathbf{X}_k$, asociados a los R mayores valores propios $\lambda^2_{r_k\ r=1,\ldots,R}$.
- Dos matrices $\mathbf{A}_1$ y $\mathbf{A}_2$ de componentes principales, de tamaño $I \times J$, donde $\mathbf{A}_k = \mathbf{X}_k\mathbf{V}_k$ en términos de la matriz inicial.
  También se tiene que $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Lambda}_k$ donde $\mathbf{U}_k$ corresponde a la matriz de vectores propios de la descomposición espectral de la matriz $\mathbf{W}_k = \mathbf{X}_k\mathbf{M}\mathbf{X}_k^{'}$

  Las matrices $\mathbf{A}_1$ y $\mathbf{A}_2$ corresponden a las coordenas de la dos nubes de individuos representadas en los subespacios de dimension R generados por $\mathbf{V}_1$ y $\mathbf{V}_2$, de ejes rotados respecto a los sitemas de ejes originales, donde se ubican los individuos originales.

2.6.1.1. Ángulo mínimo entre un vector arbitrario del espacio $\mathbf{V}_1$ y el vector más próximo paralelo a este, en el espacio $\mathbf{V}_2$

Consiste en encontrar el ángulo del vector bisector entre las nubes de individuos generadas por $\mathbf{V}_1$ y $\mathbf{V}_2$
Se considera un vector unitario arbitrario $\mathbf{a}$ en $\mathbf{R}^J$, entonces, para algún vector $\mathbf{x}$ en $\mathbf{R}^J$, se tiene que:

$$\begin{aligned}
\mathbf{x}^{'}\mathbf{V}_1 &= \mathbf{a}^{'} \\
\mathbf{x}^{'}\mathbf{V}_1\mathbf{V}_1^{'} &= \mathbf{a}^{'}\mathbf{V}_1^{'} \\
\mathbf{V}_1\mathbf{V}_1^{'}\mathbf{x} &= \mathbf{V}_1\mathbf{a} \\
\nu &= \mathbf{V}_1\mathbf{a}
\end{aligned}$$

El vector ν contiene las coordenadas de $\mathbf{a}$ referidas a los J ejes ortogonales originales.

La proyección de ν sobre el subespacio $\mathbf{V}_2$ está dada por:

$$\begin{aligned}\nu^* &= \mathbf{V}_2(\mathbf{V}_2'\mathbf{V}_2)^{-1}\mathbf{V}_2'\nu \\ &= \mathbf{V}_2\mathbf{V}_2'\nu\end{aligned}$$

Denominando θ el ángulo entre ν y ν^*

$$\begin{aligned}cos^2\theta &= \nu'\nu^* \\ &= \nu'\mathbf{V}_2\mathbf{V}_2'\nu \\ &= \mathbf{a}'\mathbf{V}_1'\mathbf{V}_2\mathbf{V}_2'\mathbf{V}_1\mathbf{a}\end{aligned}$$

Notando $\mathbb{S}_{12} = \mathbf{V}_1'\mathbf{V}_2\mathbf{V}_2'\mathbf{V}_1 = \left[\mathbf{V}_1'\mathbf{V}_2\right]\left[\mathbf{V}_1'\mathbf{V}_2\right]'$

$$cos^2\theta = \mathbf{a}'\mathbb{S}_{12}\mathbf{a}$$

Minimizar θ es equivalente a maximizar $cos^2\theta$;
luego el vector $\mathbf{a}$ que maximiza $\mathbf{a}'\mathbb{S}_{12}\mathbf{a}$ corresponde al primer vector propio de $\mathbb{S}_{12}$, entonces

$$cos^2\theta = \mathbf{a}'\mathbb{S}_{12}\mathbf{a}$$

de donde:

$$\theta = cos^{-1}(\lambda_1)$$

2.6.1.2. Base para el subespacio más cercano a los espacios $\mathbf{V}_1$ y $\mathbf{V}_2$

Sea λ_r el *r-ésimo* mayor valor propio de $\mathbb{S}_{12}$ asociado al vector propio $\mathbf{a}_r$, y $\nu_r = \mathbf{V}_1\mathbf{a}_r$ *r=1,...,R*, es decir que ν_r es el vector cuyas componentes son las coordenadas de $\mathbf{a}_r$ referenciadas a los J ejes originales.
La proyección de ν_r sobre $\mathbf{V}_2$, es $\nu_r^* = \mathbf{V}_2\mathbf{V}_2'\nu_r$

Como los vectores $\mathbf{a}_r$ son ortogonales, se tiene que:

$$\nu_r'\nu_j = \mathbf{a}_r'\mathbf{V}_1'\mathbf{V}_1\mathbf{a}_j = \mathbf{a}_r'\mathbf{a}_j = 0$$

Por tanto, los vectores ν_r también son ortogonales.
además

$$\begin{aligned}
\nu_r^{'*}\nu_j^* &= \left[\mathbf{V}_2\mathbf{V}_2^{'}\nu_r\right]^{'}\left[\mathbf{V}_2\mathbf{V}_2^{'}\nu_j\right] \\
&= \nu_r^{'}\mathbf{V}_2\mathbf{V}_2^{'}\nu_j \\
&= \mathbf{a}_r^{'}\mathbf{V}_1^{'}\mathbf{V}_2\mathbf{V}_2^{'}\mathbf{V}_1\mathbf{a}_j \\
&= \mathbf{a}_r^{'}\mathbb{S}_{12}\mathbf{a}_j \qquad como\ \mathbb{S}_{12}\mathbf{a}_j = \lambda_j^2\mathbf{a}_j \\
&= \lambda_j^2\mathbf{a}_r^{'}\mathbf{a}_j = 0
\end{aligned}$$

implica que los vectores ν_r^* tambien son ortogonales.

Entonces ν_1 y ν_1^* son los dos vectores más próximos en el espacio original, bajo la condición que uno esté en el subespacio de las componentes principales de $\mathbf{X}_1$ y el otro en el subespacio de las componentes principales de $\mathbf{X}_2$. Continuando con los vectores y valores propios de $\mathbb{S}_{12}$ se tiene que ν_2 y ν_2^* dan direcciones ortogonales a la precedentes y determinan el menor ángulo entre los dos subespacios en dicha dirección.

Por tanto, $\nu_1, \ldots, \nu_R$, es un conjunto de vectores ortogonales en el espacio generado por las variables de $\mathbf{X}_1$ y $\nu_1^*, \ldots, \nu_R^*$ forman el correspondiente conjunto de vectores ortogonales en el espacio $\mathbf{X}_2$. Además el ángulo entre ν_r y ν_r^* está dado por $cos^{-1}(\lambda_r)$ para *r=1,. . . ,R.*

Sea θ_{rj} el ángulo entre la *r-ésima* componente principal de $\mathbf{X}_1$ y la *j-ésima* componente principal de $\mathbf{X}_2$, entonces $cos\theta_{rj}$ es el elemento (r, j) de la matriz $\mathbb{H}_{12} = \mathbf{V}_1^{'}\mathbf{V}_2$, y

$$\sum_{r=1}^{R}\lambda_r^2 = traza\ \mathbb{S}_{12} = traza\ \mathbb{H}_{12}\mathbb{H}_{12}^{'} = \sum_{r=1}^{R}\sum_{j=1}^{R} cos^2\theta_{rj}$$

Lo que significa que la suma de valores propios de la matriz $\mathbb{S}_{12}$ es igual a la suma de los cosenos al cuadrado de los ángulos entre los vectores propios de los dos grupos. Esta expresión tiene valor mínimo cero, si los dos espacios son ortogonales; toma el máximo valor R si los dos subespacios coinciden. Estos resultados muestran que la similutud entre los dos subespacios se puede expresar mediante los vectores R y ν_r^* donde λ_r^2 mide la contribución del *r-ésimo* par de vectores a la similitud total.

De lo expuesto se pueden obtener los siguientes resultados:

- Si los dos subespacios definidos por las componentes principales de $\mathbf{X}_1$ y $\mathbf{X}_2$ se intersectan en un subespacio r dimensional del espacio J dimensional original, entonces los r primeros valores propios de $\mathbb{S}_{12}$ tienen valor 1 y el conjunto $\{\nu_1, \ldots, \nu_r\}$, es una base del espacio intersección.
- Si $\mathbf{X}_1$ y $\mathbf{X}_2$ han sido caracterizados por r_1 y r_2 componentes respectivamente y $r_1 \neq r_2$, entonces $\mathbb{H}_{12}$ es una matriz $r_1 \times r_2$ de rango $k = min(r_1, r_2)$. Por tanto, la matriz de mayor rango entre $\mathbb{H}_{12}\mathbb{H}_{12}^{'}$ y $\mathbb{H}_{12}^{'}\mathbb{H}_{12}$ tendrá $|r_1 - r_2|$ valores propios igual a cero.

Con este análisis, el ángulo entre los subespacios generados por el conjunto de componentes principales provee una medida del grado de diferencia entre estos. Como los pares de vectores ν_r y ν_r^* asociados a los valores propios λ_r^2 están definidos con respecto a los J ejes originales, las diferencias pueden ser intrerpretadas con relacion a las variables originales.

Una forma natural de encontrar en el espacio original, el vector más próximo a los dos definidos, es la bisectriz entre ellos y está dada por:

$$\mathbf{c}_i = \left\{1/(1 + 3\lambda_i^2)^{1/2}\right\} (\mathbf{I} + \mathbf{V}_2^{'}\mathbf{V}_2)\nu_i$$

El conjunto $\mathbf{c}_1, \ldots, \mathbf{c}_R$ define el subespacio R dimensional del espacio original, más proximo a los subespacios generados por las componentes principales de cada grupo.

2.6.2. Comparación de varios subespacios

Consideramos que tenemos K tablas de datos $\mathbf{X}_1, \ldots, \mathbf{X}_K$ que contienen las medidas para las mismas J variables de los individuos $I_1, \ldots, I_K$, respectivamente.
Notando $R < Min(J, I_k)$, el número de ejes retenido en cada uno de los ACP, se tiene:

- K matrices $\mathbf{V}_k$ de coordenadas estándar, de tamaño $J \times R$, cuyos vectores de tamaño J serán denotados $(\mathbf{v}_{rk})_{r=1,\ldots,R}$. Estos vectores corresponden a los vectores propios de la descomposición espectral de las matrices $\mathbf{S}_k = \mathbf{X}_k^{'}\mathcal{D}\mathbf{X}_k$, asociados a los R mayores valores propios $\lambda^2_{r_k\ r=1,\ldots,R}$.

- K matrices $\mathbf{A}_k$ de componentes principales, de tamaño $I_k \times J$, donde $\mathbf{A}_k = \mathbf{X}_k\mathbf{V}_k$ en términos de la matriz inicial.
  También se tiene que $\mathbf{A}_k = \mathbf{U}_k\mathbf{D}_k$ donde $\mathbf{U}_k$ corresponde a la matriz de vectores propios de la descomposición espectral de la matriz $\mathbf{W}_k = \mathbf{X}_k\mathbf{M}\mathbf{X}_k^{'}$

De manera análoga a la comparación de dos grupos, se hace la comparación entre las K tablas.
Sea ν el vector que buscamos en el espacio J dimensional original, de tal manera que el ángulo θ entre ν y el vector más próximo, casi paralelo a él en el espacio generado por las R componentes principales del *k-ésimo grupo*, sea mínimo.
Para alguna tabla k, se tiene que $cos^2\theta_k = \nu^{'}\mathbf{V}_k\mathbf{V}_k^{'}\nu$, luego el vector ν que se busca es tal, que satisface:

$$\begin{aligned} Max \sum_{k=1}^{k} cos^2\theta_k &= \sum_{k=1}^{k} \nu^{'}\mathbf{V}_k\mathbf{V}_k^{'}\nu \\ &= \nu^{'}\sum_{k=1}^{k}\left[\mathbf{V}_k\mathbf{V}_k^{'}\right]\nu \\ &= \nu^{'}\mathbb{H}\nu \quad para\ \mathbb{H} = \sum_{k=1}^{k}\mathbf{V}_k\mathbf{V}_k^{'} \end{aligned}$$

El vector ν buscado corresponde al primer vector propio de $\mathbb{H}$ que denominamos ν_1 asociado el primer valor propio λ_1^2.

Entonces para alguna tabla k, se tiene que $cos^2\theta_k = \nu_1^{'}\mathbf{V}_k\mathbf{V}_k^{'}\nu_1$ es una medida de similaridad entre el vector ν_1 y las R componentes principales del grupo k.

Y por tanto, el valor:

$$\frac{1}{K}\sum_{k=1}^{K} cos^2\theta_k = \frac{1}{K}\sum_{k=1}^{K}\nu_1^{'}\mathbf{V}_k\mathbf{V}_k^{'}\nu_1 = \frac{1}{K}\nu_1^{'}\mathbb{H}\nu_1$$

que varía entre cero y uno, se puede considerar una de medida de similaridad ente en vector ν_1 y las componentes principales de todas las tablas.

El siguiente vector del espacio J dimensionsl más cercano a los K subespacios y ortogonal a ν_1 e s el vector ν_2 correspondiente al segundo vector propio

de $\mathbb{H}$. Continuando con la descomposición en valores y vectores propios de $\mathbb{H} = \mathbf{V}_H \mathbf{\Lambda}_H^2 \mathbf{V}_H^{'}$, los pimeros R vectores de la matriz $\mathbf{V}_H$ generan el espacio más cercano a todos los K subespacios.

Si el número de componentes principales r_k retenidas en los ACP de cada una de las tablas no es igual, el subespacio consenso va a tener $r = min(r_1, \ldots, r_k)$ dimensiones. Si tomanos la dimensión r_{k+1} este vector propio será ortogonal por lo menos a uno de los subespacios.

2.7. Comparación de los métodos expuestos

En esta parte, además de hacer un resúmen de los métodos presentados se compararan entre sí a partir del esquema en que se hizo la presentación de cada capitulo y se deducen ventajas y límites de cada método.

Los métodos analizados son:

- AFM-Análisis Factorial Múltiple
- STATIS: Structuration de Tableaux A Trois Indices
- STATIS DUAL
- DACP-Doble Análisis en Componentes Principales
  - Selección del mejor conjunto de ejes
  - Maximización de la inercia explicada
  - Maximización de la inercia explicada con matrices reducidas
  - Búsqueda secuencial de un nuevo sistema
- Meta componentes

La comparación se efectúa para los aspectos donde son más relevantes las diferencias y son:

- Tablas de datos que se analizan
- Objetos que caracterizan el análisis
- Objeto consenso

- Análisis de la intraestructura
- Análisis de la interestructura
- Calidad de representación

2.7.1. Tablas de datos que se analizan

Los métodos expuestos permiten el análisis de tablas de datos que tienen información de variables cuantitativas.

El STATIS complementado por la versión dual, permite tratar la mayoría de tablas cuantitivas con los mismos individuos, o el dual, con las mismas variables.

El DACP es más restrictivo que los otros métodos, exigiendo que todas las tablas sean identicas en número de individuos y de variables.

Tanto el AFM como el Meta componentes, además de posibilitar el análisis de tablas de datos cuantitativos, permiten analizar tablas de datos cualitativos y de tablas mixtas, es decir, que permiten comparar tablas de datos cualitativos con tablas de datos cuantitativos. Sin embargo el AFM requiere que los individuos sean los mismos en las diferentes tablas, mientras que en el Meta componentes se deben tener las mismas variables en las distintas tablas.

2.7.2. Objetos que caracterizan el análisis

La notacion usada para los objetos es:

$$\mathbf{W}_k = \mathbf{X}_k\mathbf{M}_k\mathbf{X}_k^{'} \qquad \mathbf{S}_k = \mathbf{X}_k^{'}\mathcal{D}\mathbf{X}_k$$

$$\langle \mathbf{W}_k\mathcal{D}, \mathbf{W}_l\mathcal{D}\rangle_{HS} = traza(\mathbf{W}_k\mathcal{D}\mathbf{W}_l\mathcal{D})$$

$$\langle \mathbf{S}_k\mathbf{M}_k, \mathbf{S}_l\mathbf{M}_l\rangle_{HS} = traza(\mathbf{S}_k\mathbf{M}_k\mathbf{S}_l\mathbf{M}_l)$$

Los métodos Statis, Statis Dual, y DACP: maximización de inercia explicada y maximización de inercia explicada con matrices reducidas; permiten

trabajar con los objetos normados y no normados.

El AFM trabaja con los objetos normados, el Meta componentes y las demás variaciones del DACP: selección del mejor conjunto de ejes, búsqueda secuencial del nuevo sistema de ejes; utilizan los objetos no normados.

La normalización propuesta para los objetos, difiere de un método a otro.

2.7.2.1. Objetos no normados

Tanto el Statis Dual como el DACP: maximización de inercia explicada, selección del mejor conjunto de ejes, búsqueda secuencial del nuevo sistema de ejes y el Meta componentes; utilizan los objetos: $\mathbf{S}_k$

El Statis utiliza los objetos no normados $\mathbf{W}_k$

Cuando se aplican los métodos de análisis utilizando objetos no normados se corre el riesgo que ciertos objetos influencien de manera importante la construcción del espacio consenso.

2.7.2.2. Objetos normados

El Statis, Statis Dual, el AFM y DACP: maximización de inercia explicada con matrices reducidas, utilizan los siguientes objetos normados:

$$STATIS: \qquad \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_{HS}} = \frac{\mathbf{W}_k}{\sqrt{\sum_s \lambda_{s_k}^4}}$$

$$STATIS\ DUAL: \qquad \frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}} = \frac{\mathbf{S}_k}{\sqrt{\sum_s \lambda_{s_k}^4}}$$

$$AFM: \qquad \frac{\mathbf{W}_k}{\lambda_{1_k}^2}$$

$$DACP\,(Max\,de\,inercia\,explicada): \qquad \frac{\mathbf{S}_k}{\sum_{r=1}^{R} \lambda_{r_k}^2}$$

Para el Statis, la normalización hace que los objetos tengan norma 1, eliminando la estructura múltiple al interior de las tablas, es decir, que desaparece

la noción de asociación entre variables de una misma tabla.
En el DACP: maximización de inercia explicada con matrices reducidas, la normalización pretende igualar la inercia total de las nubes de individuos definidas por los K grupos, haciendo desaparecer la noción de asociación entre variables de una misma tabla.

Estas dos normalizaciones tienden relativamente a debilitar la inercia en cada una de las direcciones de un grupo compuesto de variables independientes y además refuerza las direcciones de un grupo compuesto de pocas variables muy correlacionadas.

En el AFM, la ponderación por $1/\lambda_{1_k}^2$ para cada grupo, no altera la estructura múltiple de los diferentes grupos de variables porque la inercia total no interviene. Es igual a 1 para la primer dirección de cada grupo y permite comparar con 1 la inercia en las demás direcciones. Además, con está ponderación ningun grupo puede influenciar de manera preponderante el espacio consenso.

Por tanto, la ponderación del AFM pretende mejorar la calidad del espacio consenso comparado con los otros métodos que utilizan objetos normados.

Sin embargo, hay que notar que la ponderación por $1/\lambda_{1_k}^2$ no nace de un concepto teórico que justifique que sea óptima. Esta es arbitaria y de la misma manera Dazy y Le Barzic (1996) proponen ponderar por $1/\left(\lambda_{1_k}^2 + \lambda_{2_k}^2\right)$, que presenta las siguientes características:

- Iguala a 1 la inercia del primer plano factorial de cada tabla.
- Permite comparar con 1 la inercia de cada tabla en cualquier dimensión.
- Ninguna de las nubes puede influenciar de manera preponderante el primer plano factorial del espacio consenso, donde generalmente se representa la información.
- Permite mantener la noción de asociación entre variables de un mismo grupo.

Comparando la ponderación de Statis con la del AFM, se puede observar que si todas las variables están fuertemente correlacionadas entre ellas, es decir que existe un efecto de tamaño el primer valor propio es grande comparado con los otros, entonces para Statis tenemos que:

$$\frac{\mathbf{W}_k}{\sqrt{\sum_s \lambda_{s_k}^4}} = \frac{\mathbf{W}_k}{\sqrt{\lambda_{1_k}^4 \sum_s \frac{\lambda_{s_k}^4}{\lambda_{1_k}^4}}} = \frac{\mathbf{W}_k}{\lambda_{1_k}^2 \sqrt{1 + \sum_{s=2} \frac{\lambda_{s_k}^4}{\lambda_{1_k}^4}}}$$

y por aproximación de primer orden, se obtiene que:

$$\frac{\mathbf{W}_k}{\sqrt{\sum_s \lambda_{s_k}^4}} \simeq \frac{\mathbf{W}_k}{\lambda_{1_k}^2} \qquad \textit{encontrando el objeto del AFM}$$

Para el Statis Dual y el DACP: maximización de inercia explicada con matrices reducidas:

$$\frac{\mathbf{S}_k}{\sqrt{\sum_s \lambda_{s_k}^4}} = \frac{\mathbf{S}_k}{\sqrt{\lambda_{1_k}^4 \sum_s \frac{\lambda_{s_k}^4}{\lambda_{1_k}^4}}} = \frac{\mathbf{S}_k}{\lambda_{1_k}^2 \sqrt{1 + \sum_{s=2} \frac{\lambda_{s_k}^4}{\lambda_{1_k}^4}}}$$

y

$$\frac{\mathbf{S}_k}{\sum_{r=1}^{R} \lambda_{r_k}^2} = \frac{\mathbf{S}_k}{\lambda_{1_k}^2 \sum_{r=1}^{R} \frac{\lambda_{r_k}^2}{\lambda_{1_k}^2}} = \frac{\mathbf{S}_k}{\lambda_{1_k}^2 \left[1 + \sum_{r=2}^{R} \frac{\lambda_{r_k}^2}{\lambda_{1_k}^2}\right]}$$

por aproximación de primer orden:

$$\frac{\mathbf{S}_k}{\sum_{r=1}^{R} \lambda_{r_k}^2} \simeq \frac{\mathbf{S}_k}{\lambda_{1_k}^2} \simeq \frac{\mathbf{S}_k}{\sqrt{\sum_s \lambda_{s_k}^4}}$$

Luego las normalizaciones de orden 1 son iguales y en general, para cualquier otro orden difieren poco.

2.7.3. Objeto Consenso

Para el AFM el objeto consenso se obtiene sumando los objetos representativos:

$$\mathbf{W} = \sum_{k=1}^{K} \frac{\mathbf{W}_k}{\lambda_{1_k}^2}$$

Al igual que para el DACP: maximización de inercia explicada y maximización de inercia explicada con matrices reducidas; el objeto consenso se obtiene sumando los objetos representativos de cada tabla. Este objeto tiene en cuenta las asociaciones entre los individuos ó entre las variables.

Maximización de la inercia explicada:

$$\mathbf{S} = \sum_{k=1}^{K} \mathbf{S}_k$$

Maximización de la inercia explicada con matrices reducidas:

$$\mathbf{S} = \sum_{k=1}^{K} \frac{\mathbf{S}_k}{\sum_{r=1}^{R} \lambda_r^2}$$

El objeto consenso para el Statis y el Statis Dual difiere de los anteriores. En este caso se busca una combinación lineal de los objetos representativos de cada tabla que sea lo más correlacionada con los objetos según el producto escalar de Hilbert Schmidt. Además el objeto consenso es de la misma naturaleza que los objetos representativos, ya sean normados o no.

El consenso para el Statis se obtiene como:

$$\mathbf{W} = \sum_{k=1}^{K} \alpha_k \mathbf{W}_k \qquad \textit{para objetos no normados}$$

$$\mathbf{W} = \sum_{k=1}^{K} \alpha'_k \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_{HS}} \qquad \textit{para objetos normados}$$

Para el Statis Dual el consenso está dado por:

$$\mathbf{S} = \sum_{k=1}^{K} \beta_k \mathbf{S}_k \qquad \textit{para objetos no normados}$$

$$\mathbf{S} = \sum_{k=1}^{K} \beta'_k \frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}} \qquad \textit{para objetos normados}$$

Considerando que un buen consenso el espacio donde ninguna tabla es preponderante, entonces para el método Statis los coeficientes α_k son próximos

los unos de los otros, y se puede escribir $\alpha_k \simeq \alpha$, respectivamente $\alpha'_k \simeq \alpha'$. Para el Statis Dual $\beta_k \simeq \beta$, respectivamente $\beta'_k \simeq \beta'$.

Entonces el consenso para el Statis se puede aproximar por :

$$\mathbf{W} \simeq \sum_{k=1}^{K} \alpha \mathbf{W}_k \simeq \alpha \sum_{k=1}^{K} \mathbf{W}_k$$

y para los objetos normados

$$\mathbf{W} \simeq \alpha' \sum_{k=1}^{K} \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_{HS}}$$

y el objeto consenso para el Statis Dual se puede aproximar por:

$$\mathbf{S} \simeq \beta \sum_{k=1}^{K} \mathbf{S}_k$$

y para los objetos normados

$$\mathbf{S} \simeq \beta' \sum_{k=1}^{K} \frac{\mathbf{S}_k}{\|\mathbf{S}_k\|_{HS}}$$

Los factores $\alpha, \alpha', \beta, \beta'$ no tienen influencia en la búsqueda de los valores propios de $\mathbf{W}$ y $\mathbf{S}$, por tanto los objetos consenso son los mismos que se obtendrían al hacer la suma de los objetos representativos.

Para el método Meta componentes, el objeto consenso está definido por:

$$\mathbf{H} = \sum_{k=1}^{K} \mathbf{V}_k \mathbf{V}'_k$$

A partir de la descomposición en valores singulares de la tabla $\mathbf{X}_k$ se tiene:

$$\mathbf{X}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}'_k$$

$$\mathbf{D}_k^{-1} \mathbf{U}'_k \mathbf{X}_k = \mathbf{V}'_k$$

$$\mathbf{X}'_k \mathbf{U}_k \mathbf{D}_k^{-1} = \mathbf{V}_k$$

De donde

$$\mathbf{V}_k \mathbf{V}'_k = \mathbf{X}'_k \left(\mathbf{X}_k \mathbf{X}'_k \right)^{-1} \mathbf{X}_k$$

expresión que corresponde a la matriz que proyecta sobre las filas de $\mathbf{X}_k$, para alguna inversa condicional $\left(\mathbf{X}_k\mathbf{X}_k'\right)^{-1}$. Por tanto el objeto consenso para el metacomponentes es la suma de las matrices proyectoras sobre las filas para cada una de las matrices:

$$\mathbf{H} = \sum_{k=1}^{K} \mathbf{X}_k' \left(\mathbf{X}_k\mathbf{X}_k'\right)^{-1} \mathbf{X}_k$$

En caso de presencia del efecto tamaño, es decir que todas las variables en cada grupo estén altamente correlacionadas, el valor propio de cada tabla será mucho mayor que los demás valores propios, en ese caso, este método encontrará un espacio consenso similar que el DACP: búsqueda secuencial de un nuevo sistema de ejes. Sin embargo, si en alguna de las tablas la explicación de los ejes se encuentra rotada, el Meta componentes captará dicha rotación y los dos métodos encuentran espacios consensos diferentes, siendo óptima la solución para el Meta componentes.

2.7.4. Análisis de la Intraestructura

Las posiciones compromiso y las trayectorias de los individuos.

El Statis y el AFM proponen une representación de las posiciones compromiso que corresponde a la media de los individuos sobre el periodo de análisis.

Para el Statis la representación se obtiene efectuando un ACP de las tablas yuxtapuestas, ver figura2.8.

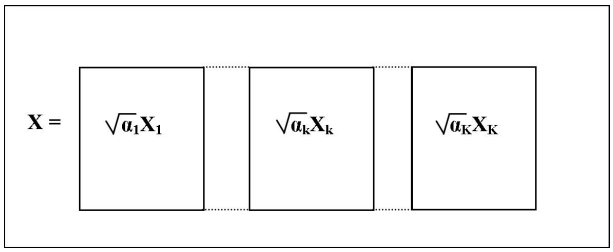

Figura 2.8: posiciones compromiso de los individuos en Statis

Para el AFM, las trayectorias de los individuos son obtenidas proyectando los individuos definidos por cada tabla de datos sobre los ejes del espacio

consenso. Esto se logra proyectándolos como elementos externos al efectuar un ACP de las tablas yuxtapuestas (ver figura2.9).

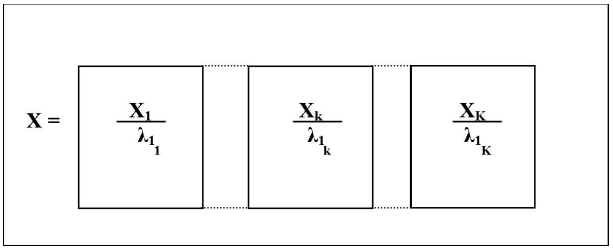

Figura 2.9: posiciones compromiso de los individuos en el AFM

Observación: De acuerdo a la deducción de los objetos compromiso, para estos dos métodos las posiciones compromiso serán muy próximas si se trabaja con los objetos normados

Para el DACP: Selección del mejor conjunto de ejes; representa las trayectorias de los individuos y las posiciones compromiso.

Mientras que para DACP: maximización de inercia explicada, maximización de inercia explicada con matrices reducidas y búsqueda secuencial de un nuevo sistema de ejes; se representan unicamente la trayectoria de los individuos.

2.7.5. Análisis de la Interestructura

En esta etapa se busca representar cada tabla por un punto con el fin de encontrar similitudes y diferencias entre tablas.

Para el AFM se proyectan las objetos representativos de las tablas sobre un sistema de ejes de la forma $\nu_{\mathbf{i}}\nu_{\mathbf{i}}^{'}$ donde los $\nu_{\mathbf{i}}$ son los ejes de la intraestructura. La calidad de representación es menor que en el STATIS pero la imagen en fácil de interpretar.

Para el Statis y el Statis Dual se efectúa el ACP de la tabla de productos escalares entre los objetos representativos. Obteniendo la representación de la proximidad de las tablas y su calidad de representación es mejor que para

el AFM. Por otra parte, los ejes de la interestructura no son interpretables.

El DACP presenta un análisis de interestructura diferente a los otros métodos. De hecho el AFM y el Statis analizan globalmente las posiciones de los individuos centrados, mientras que en el DACP se efectúa un ACP de los centros de gravedad de las tablas.

2.7.6. Calidad de la representación

Además de los indicadores de calidad propios del ACP, el DACP utiliza dos índices numéricos que miden la calidad de la imagen euclidea del espacio consenso.
El primer índice $\xi(., \mathbf{v})$ mide el porcentaje promedio de la perdida de inercia del conjunto de las K nubes cuando se proyectan sobre el subespacio generado por $(\mathbf{v}_r)_{r=1,...,R}$

$$\xi(., \mathbf{v}) = \frac{1}{K} \sum_{k=1}^{K} [\xi(k, \mathbf{v})]$$

para

$$\xi(k, \mathbf{v}) = \frac{\sum_{r=1}^{R} \lambda_{r_k}^2 - \sum_{r=1}^{R} \mathbf{v}_r' \mathbf{S}_k \mathbf{v}_r}{\sum_{r=1}^{R} \lambda_{r_k}^2}$$

El segundo índice que mide la proximidad entre el conjunto de ejes $(\nu_r)_{r=1,...,R}$ y los K conjuntos de ejes $(\mathbf{v}_{r_k})_{r=1,...,R}$, está dado por:

$$\eta(\nu_1, \ldots, \nu_r) = \sum_{k=1}^{K} \sum_{r=1}^{R} cos^2(\nu_r, \mathbf{v}_{r_k})$$

El método Meta componentes hace uso del índice η y define:

$$\frac{1}{K} \sum_{k=1}^{K} cos^2 \theta_k = \frac{1}{K} \sum_{k=1}^{K} \nu_1' \mathbf{V}_k \mathbf{V}_k' \nu_1 = \frac{1}{K} \nu_1' \mathbf{H} \nu_1$$

que varía entre cero y uno, es una de medida de similaridad ente en vector ν_1 del espacio consenso y las componentes principales de todas las tablas.

Este índice se puede generalizar para las R primeras componentes del espacio consenso, como:

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{R} cos^2(\nu_r, \mathbf{V}_k) = \frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{R} \nu_r^{'}\mathbf{V}_k\mathbf{V}_k^{'}\nu_r$$

y su valor vará entre 1 y R dando la similitud entre el espacio consenso R dimensional y las componentes principales de todas las tablas.

Cuando los valores de los índices no son favorables se puede afirmar que la relación entre los individuos no es estable de una tabla a otra y se hace necesario analizar por separado cada tabla. Estos valores se convierten en una herramienta útil para identificar cuando el espacio consenso no refleja una estructura común para los individuos.

Capítulo 3

Biplot Consenso para análisis de tablas múltiples

3.1. Introducción

En los métodos desarrollados en los capítulos anteriores para análisis de tablas múltiples, cuando se tiene el mismo número de individuos, de manera general se yuxtaponen las K tablas de datos originando una matriz $\mathbf{X}$ de tamaño $I \times JK$. Si las tablas tienen las mismas variables se aplilan las columnas obtiendo una matriz $\mathbf{X}$ de tamaño $\sum_k I_k \times J$ y posteriormente se efectúa un ACP a la matriz $\mathbf{X}$.

Como todos los métodos de análisis multivariante se busca reducir el espacio original para representar los datos en un espacio de dimensión menor, con el fin de perder un poco de información pero ganar en análisis; por tal razon es importante evaluar el porcentaje de información que se pierde al reducir la dimensión del espacio, así como evaluar la calidad de representación de las filas y columnas de la matriz yuxtapuesta.

Por tanto el objetivo que persiguen estos métodos es encontrar un espacio común ó *Espacio Consenso* para la representación de los elementos de filas y columnas de las diferentes tablas, de tal manera que su representación sea óptima bajo algún criterio.

Para los métodos presentados, el Statis y el Statis Dual aportan una buena calidad de representación de las tablas, dando una representación de la interestructura que permite observar la proximidad entre las tablas. Sin embargo

el proceso de búsqueda del espacio consenso no facilita la interpretación de los resultados.

El AFM ofrece una fácil interpretación de los resultados. En la Intraestructura los ejes del espacio consenso son interpretables de acuerdo a su correlación con los ejes iniciales, además las posiciones de los individuos y de los individuos consenso permiten la comparación de estos a través de las tablas. En la interestructura permite conocer las proximidades entre tablas, sin embargo está limitado a tablas que contienen los mismos individuos.

El DACP es un método que presenta una alternativa interesante, sin embargo está se presenta para el análisis de tablas cúbicas.

El Meta compontentes tiene la ventaja de detectar tablas con ejes rotados, sin embargo no presenta un proceso para el análisis de la interestructura.

Todos los métodos anteriormente descritos tienen la limitación que no representan de manera conjunta individuos y variables.

A partir de estas limitaciones para los métodos presentados surge la idea de *Biplot consenso para análisis de tablas múltiples*, que tiene bases en el método Biplot propuesto por Gabriel, 1971, tema ampliamente tratado en el capítulo: *Representación Biplot*, pg:1.

En este trabajo buscamos un espacio de representación común para todas las tablas. Este espacio generaliza el análisis de tablas múltiples con variables comunes y en particular, el análisis de tablas cúbicas, poniéndolas en un marco de referencia común.

Este método permite:

- Una formulación común de los métodos de tablas múltiples para la integración de objetos procedentes de la relación entre variables (o individuos) y los métodos de integración de subespacios.
- Obtener muchas de las técnicas de la literatura como casos particulares, como también la posibilidad de definir nuevos criterios.
- Dotar de interpretación biplot a técnicas que hasta el momento no la tenían. En la literatura se ha encontrado la interpretación biplot parcial de algunos de ellos pero nunca en un marco común ni desde el punto de vista de la obtención de un subespacio consenso.

- Además de un marco de referencia común, la comparación del funcionamiento de las distintas técnicas aplicadas a un mismo conjunto de datos, con el objeto de seleccionar la más adecuada. En la literatura hemos encontrado comparaciones por pares de metodologías, por ejemplo AFM y STATIS y desde un punto de vista diferente al que se plantea en este trabajo.

La exposición seguirá el mismo esquema utilizado en la presentación de los anteriores métodos.

3.2. Interestructura

Para ver la relación entre los grupos que se analizan se realiza una análisis canónico de poblaciones (ACPob). Este análisis forma parte de una serie de técnicas destinadas a clasificar o a explicar grupos de individuos, caracterizados por cierto número de variables numéricas o nominales. Puede considerarse un método tanto descriptivo como predictivo.

La técnica permite estudiar la estructura de varios grupos de individuos con respecto a un conjunto de variables observadas, proyectando el espacio Euclídeo generalizado en un espacio Euclídeo, de tal manera que la separación entre los distintos grupos sea máxima, con respecto a la variabilidad dentro de los grupos. Es considerado como un Análisis de Componentes Principales de una matriz cuyas filas corresponden a los centros de gravedad de los grupos en el espacio de las variables. La matriz de métricas para los individuos, es la inversa de la matriz de varianzas-covarianzas dentro de los grupos y la de las variables es la matriz diagonal, cuyos términos sobre la diagonal son los tamaños muestrales de cada uno de los grupos (Lebart et al, 1995).

En este contexto, se aplica el ACPob desde el punto de vista descriptivo que no requiere hipótesis sobre la distribución de los datos.

3.2.1. Análisis Canónico de Poblaciones

Se consideran K tablas con las mismas J variables y diferente número de individuos, notados $I_1, \ldots, I_K$ de tal manera que el conjunto total de individuos está dado por $I = \sum_k I_k$.

Apilando las matrices respecto a sus columnas se obtiene una matriz $\mathbf{X}$ de

tamaño $I \times J$ que contiene toda la información de las K tablas. Se construye la matriz $\mathbb{Z}$ de tamaño $I \times K$ de indicadoras de los individuos al grupo, donde

$$\mathbb{Z}' = \begin{bmatrix} \mathbf{1}'_{I_1} & \dots & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{1}'_{I_k} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \dots & \mathbf{1}'_{I_K} \end{bmatrix}$$

Sin pérdida de generalidad, se supone que las columnas de $\mathbf{X}$ están centradas y se define:

- La matriz $\mathbf{N}$ de tamaño $K \times K$ como

$$\mathbf{N} = \mathbb{Z}'\mathbb{Z} = diag\,(I_1, \dots, I_k, \dots, I_K)$$

- La matriz $\bar{\mathbf{X}}$ de tamaño $K \times J$, que contiene los centros de gravedad de las K tablas, con $\mathbf{g}_k^{'} = \bar{x}_{1_k}, \dots, \bar{x}_{J_k}$ el vector que contiene los centros de gravedad para las J variables, es:

$$\bar{\mathbf{X}} = \begin{bmatrix} \mathbf{g}_1^{'} \\ \vdots \\ \mathbf{g}_k^{'} \\ \vdots \\ \mathbf{g}_K^{'} \end{bmatrix} = \begin{bmatrix} \bar{x}_{1_1} & \dots & \bar{x}_{J_1} \\ \vdots & & \vdots \\ \bar{x}_{1_k} & \dots & \bar{x}_{J_k} \\ \vdots & & \vdots \\ \bar{x}_{1_K} & \dots & \bar{x}_{J_K} \end{bmatrix}$$

  y se expresa mediante:

$$\bar{\mathbf{X}} = \mathbf{N^{-1}}\mathbb{Z}'\mathbf{X}$$

- La matriz de varianzas-covarianzas dentro de los grupos:

$$\mathbf{S}_d = \frac{1}{(I-K)}\left(\mathbf{X'X} - \bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}\right)$$

- La matriz de covarianzas entre los grupos:

$$\mathbf{S}_e = \frac{1}{(K-I)}\bar{\mathbf{X}}'\mathbf{N}\bar{\mathbf{X}}$$

El objetivo es separar al máximo los grupos a partir de las medias de los J variables, es decir, hallar el vector $\mathbf{v}$ que maximice:

$$g(\mathbf{v}) = \frac{\mathbf{v'S}_e\mathbf{v}}{\mathbf{v'S}_d\mathbf{v}} \tag{3.2.1}$$

bajo la condición que $\mathbf{v}'\mathbf{S}_d\mathbf{v} = 1$ con el fin que la solución sea única. El planteamiento es equivalente a encontrar $\mathbf{v}$ que maximice

$$L(\mathbf{v}) = \mathbf{v}'\mathbf{S}_e\mathbf{v} - \lambda^2(\mathbf{v}'\mathbf{S}_d\mathbf{v} - 1)$$

y que es solución para:

$$(\mathbf{S}_e - \lambda^2\mathbf{S}_d)\mathbf{v} = \mathbf{0} \tag{3.2.2}$$

$$\mathbf{S}_e\mathbf{v} = \lambda^2\mathbf{S}_d\mathbf{v} \tag{3.2.3}$$

de donde:

$$\mathbf{v}'\mathbf{S}_e\mathbf{v} = \lambda^2\mathbf{v}'\mathbf{S}_d\mathbf{v}$$

$$\mathbf{v}'\mathbf{S}_e\mathbf{v} = \lambda^2$$

además, siempre que $\mathbf{S}_d$ sea no singular, (3.2.2) se puede escribir como:

$$(\mathbf{S}_d^{-1}\mathbf{S}_e - \lambda^2\mathbf{I})\mathbf{v} = \mathbf{0} \tag{3.2.4}$$

Luego el vector $\mathbf{v}$ que maximiza $g(\mathbf{v})$ es el vector propio de $(\mathbf{S}_d^{-1}\mathbf{S}_e)$ asociado al mayor valor propio λ^2.

Como la matriz $(\mathbf{S}_d^{-1}\mathbf{S}_e)$ es no simétrica, su descomposición espectral se halla de la siguiente manera (Cuadras 1996 y Gittins 1985):

A partir de

$$\begin{aligned} (\mathbf{S}_e - \lambda^2\mathbf{S}_d)\mathbf{v} &= \mathbf{0} \\ \mathbf{S}_d^{-1/2}(\mathbf{S}_e - \lambda^2\mathbf{S}_d)\mathbf{v} &= \mathbf{0} \\ (\mathbf{S}_d^{-1/2}\mathbf{S}_e - \lambda^2\mathbf{S}_d^{1/2})\mathbf{v} &= \mathbf{0} \end{aligned} \tag{3.2.5}$$

que es equivalente a

$$\begin{aligned} (\mathbf{S}_d^{-1/2}\mathbf{S}_e\mathbf{S}_d^{-1/2}\mathbf{S}_d^{1/2} - \lambda^2\mathbf{S}_d^{1/2})\mathbf{v} &= \mathbf{0} \\ (\mathbf{S}_d^{-1/2}\mathbf{S}_e\mathbf{S}_d^{-1/2} - \lambda^2\mathbf{I})\mathbf{S}_d^{1/2}\mathbf{v} &= \mathbf{0} \end{aligned} \tag{3.2.6}$$

con $\mathbf{q} = \mathbf{S}_d^{1/2}\mathbf{v}$ se tiene que:

$$(\mathbf{S}_d^{-1/2}\mathbf{S}_e\mathbf{S}_d^{-1/2} - \lambda^2\mathbf{I})\mathbf{q} = \mathbf{0}$$

Por tanto $\mathbf{q}$ es vector propio de

$$\mathbf{S}_d^{-1/2}\mathbf{S}_e\mathbf{S}_d^{-1/2}$$

correspondiente al valor propio λ^2

Además $\mathbf{v} = \mathbf{S}_d^{-1/2}\mathbf{q}$
y

$$\begin{aligned}\mathbf{v}'\mathbf{S}_d\mathbf{v} &= \mathbf{q}'\mathbf{S}_d^{-1/2}\mathbf{S}_d\mathbf{S}_d^{-1/2}\mathbf{q} \\ &= \mathbf{q}'\mathbf{q} = 1\end{aligned}$$

Entonces todas las soluciones para (3.2.4) se pueden expresar de la forma:

$$\mathbf{S}_e\mathbf{V} = \mathbf{S}_d\mathbf{V}\mathbf{\Lambda}$$

Donde $\mathbf{V}$ es la matriz de vectores propios y $\mathbf{\Lambda}$ los respectivos valores propios en orden no ascendente. Los vectores de $\mathbf{V}$ son ortogonales y normados bajo $\mathbf{S}_d$, luego

$$\mathbf{V}'\mathbf{S}_d\mathbf{V} = \mathbf{I}$$

Para $\mathbf{v}_1$, vector propio asociado la primer valor propio, la expresión:

$$\mathbf{y}_1 = \mathbf{X}\mathbf{v}_1$$

es la combinación lineal que maximiza el ratio de varianza, significa que se obtiene la transformación ortogonal que hace máxima la separación entre los grupos, relativa la variabilidad dentro de estos. El vector $\mathbf{y}_1$ es la transformación de $\mathbf{X}$ asociada al vector $\mathbf{v}_1$ y se denomina primera variable canónica.

Entonces

$$\mathbf{Y} = \mathbf{X}\mathbf{V}$$

es la transformación lineal de la nube de individuos representados por las variables originales, en la nube de individuos representados por las variables canónicas y a este espacio se le denomina espacio canónico. La proyección de las medias de cada grupo sobre el espacio canónico está dada por:

$$\bar{\mathbf{Y}} = \bar{\mathbf{X}}\mathbf{V}$$

Además

$$\bar{\mathbf{Y}}\bar{\mathbf{Y}}' = \bar{\mathbf{X}}\mathbf{V}\mathbf{V}'\bar{\mathbf{X}}' = \bar{\mathbf{X}}\mathbf{S}_d^{-1}\bar{\mathbf{X}} \qquad \textit{por que } \mathbf{V}\mathbf{V}' = \mathbf{S}_d^{-1}$$

implica que en el espacio canónico, las distancias euclídeas entre las medias de las variables canónicas coinciden con la distancia de Mahalanobis en el

espacio de las variables originales. Por tanto el espacio canónico se puede pensar como un espacio Euclídeo.

Como en los análisis multivariantes es posible representar el conjunto de individuos en el espacio con pocas dimensiones. Si se decide retener r variables canónicas, la nube de individuos se representan en el biplot mediante:

$$\mathbf{A}_{(r)} = \mathbf{X}\mathbf{V}_{(r)}$$

y sus centroides serán reprentados por medio de la proyección sobre las variables canónicas, como:

$$\bar{\mathbf{A}}_{(r)} = \bar{\mathbf{X}}\mathbf{V}_{(r)}$$

Cuando $r = 2$ ó 3, la representación del espacio canónico proporciona un resumen visual de las relaciones entre los grupos; estas representaciones requieren algunas suposiciones mínimas para su construcción, una de ellas es la elección adecuada de la distancia a usar en las relaciones entre los grupos.

El gráfico muestra la forma y dispersión de cada grupo de individuos.

3.3. Consenso

En el análisis de la interestructura se da a conocer la varibilidad entre los grupos, considerando que los individuos pertenecen a poblaciones claramente diferenciadas.

Ahora se quiere encontrar un método para comparar las tablas de acuerdo a su variabilidad dentro de los grupos.

Una forma simple de abordar el problema consiste en calcular las componentes principales de cada tabla y estudiar si la estructura de variabilidad se mantiene entre las distintas tablas. Esta estrategia aunque sencilla es extremadamente subjetiva ya que es el mismo investigador quien decide la similitud. Además, dos conjuntos de componentes que son bastante diferentes en apariencia pueden definir el mismo subespacio multivariante original.

Otra forma de ver el problema es representar todos los individuos sobre un espacio de componentes comunes que recoja la mayor cantidad posible de la estructura de covariación dentro de las tablas.

Trataremos ambas formas mediante la construcción de biplots comunes para todos los grupos, dotando de este tipo de interpretación a la integración de

los subespacios y proponiendo una representación biplot común para todas las tablas de tal manera que permita la interpretación de cada una de ellas sobre la misma estructura.

Consideramos K tablas de datos $\mathbf{X}_k$, con I_k individuos y J columnas, para $k = 1, \ldots, K$; cuyos datos están centrados de forma que cada nube de individuos tiene como referencia el orígen.

La descomposicióon en valores singulares de cada tabla, es:

$$\mathbf{X}_k = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^{'} \tag{3.3.1}$$

donde $\mathbf{U}_k$ define un conjunto de vectores ortogonales que sirven como base del subespacio definido por las componentes principales. $\mathbf{V}_k$ contiene los vectores propios de la matriz de covarianzas, esto significa que:

$$\mathbf{X'}_k \mathbf{X}_k = \mathbf{V}_k \mathbf{\Lambda}_k^2 \mathbf{V}_k^{'}$$

$$\begin{aligned} \mathbf{S}_k &= \frac{1}{(I_k - 1)} \mathbf{X'}_k \mathbf{X}_k \\ &= \frac{1}{(I_k - 1)} \mathbf{V}_k \mathbf{\Lambda}_k^2 \mathbf{V}_k^{'} \end{aligned}$$

A efectos prácticos puede tomarse la matriz de productos cruzados $\mathbf{X'}_k \mathbf{X}_k$ o la de covarianza $\mathbf{S}_k$ sin que se modifiquen los resultados pues solo los valores propios difieren por un factor escala de $\frac{1}{(I_k - 1)}$.

Es conocido que un biplot para $\mathbf{X}_k$ puede obtenerse de (3.3.1), expresado como:

$$\mathbf{X}_k = \mathbf{A}_k \mathbf{B}_k^{'}$$

con $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Lambda}_k^{\gamma}$ y $\mathbf{B}_k = \mathbf{V}_k \mathbf{\Lambda}_k^{\gamma - 1}$ para $0 \leq \gamma \leq 1$

Para nuestro propósito tomaremos $\gamma = 1$, por su relación con las componentes principales, es decir:

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{\Lambda}_k \quad , \quad \mathbf{B}_k = \mathbf{V}_k.$$

El biplot en dimensión reducida se obtiene a partir de las primeras columnas de ambas matrices. Tenemos entonces que

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{\Lambda}_k = \mathbf{X}_k \mathbf{V}_k$$

son las coordenadas de los individuos sobre las componentes principales.

Si tomamos la base canónica en el espacio $\mathbf{R}^J$ completo, representada por la matriz identidad $\mathbf{I}_J$, tenemos que

$$\mathbf{B}_k = \mathbf{I}_J \mathbf{V}_k$$

Es decir, las coordenadas de las variables son las proyecciones de los vectores unitarios en la dirección de los ejes iniciales sobre las componentes principales.

Sobre el biplot tenemos varias formas de medir la bondad de ajuste, todas ellas equivalentes. La bondad de ajuste se conoce en el contexto de las componentes principales como la absorción de inercia retenida en el análisis.

- La primera forma de expresarla, es simplemente como cociente de valores propios de la bondad de ajuste en dimension reducida r es

$$\frac{\sum_{j=1}^{r} \lambda_{j_k}^2}{\sum_{j=1}^{J} \lambda_{j_k}^2} \tag{3.3.2}$$

  donde $\lambda_{j_k}^2$ es el *j-ésimo* valor propio de la *k-ésima* tabla.

- Los valores ajustados de la aproximación a bajo rango son

$$\hat{\mathbf{X}}_k = \mathbf{X}_k \mathbf{V}_k \mathbf{V'}_k \qquad \text{en dimesión reducida}$$

  (Se han omitido los índices porque están implícitos en la exposición y faciltan la misma).
  Podemos entonces, descomponer la suma de cuadrados total en dos partes, la explicada por la aproximación y la residual.

$$\|\mathbf{X}_k\| = \left\|\hat{\mathbf{X}}_k\right\| + \left\|\mathbf{X}_k - \hat{\mathbf{X}}_k\right\| \tag{3.3.3}$$

  y obtener la bondad del ajuste como

$$\frac{\left\|\hat{\mathbf{X}}_k\right\|}{\|\mathbf{X}_k\|} = \frac{Traza\left(\hat{\mathbf{X}}'_k \hat{\mathbf{X}}_k\right)}{Traza\left(\mathbf{X'}_k \mathbf{X}_k\right)}$$

  Es inmediato observar que (3.3.2) y (3.3.3) son iguales.

El hecho de obtener una bondad de ajuste global buena no significa que todos los individuos, o variables, tengan una calidad de representación adecuada y

es conveniente separar la calidad para cada uno de los individuos. La forma habitual de medir la calidad de representación de un individuo consiste en calcular el coseno al cuadrado del ángulo que forman el vector que representa al individuo en el espacio completo y el vector proyectado en el espacio de las componentes principales.

$$\begin{aligned} Cos^2(\mathbf{x}_{i_k}, \hat{\mathbf{x}}_{i_k}) &= \frac{(\mathbf{x}_{i_k}, \hat{\mathbf{x}}'_{i_k})^2}{\|\mathbf{x}_{i_k}\| \, \|\hat{\mathbf{x}}_{i_k}\|} \\ &= \frac{(\mathbf{x}_{i_k} \mathbf{V}_{j_k(r)} \mathbf{V'}_{j_k(r)} \mathbf{x'}_{i_k})^2}{(\sum_j \mathbf{x}^2_{ij_k})(\sum_j \hat{\mathbf{x}}^2_{ij_k})} \\ &= \frac{(a'_{i_{k(r)}} a_{i_{k(r)}})^2}{(a'_{i_{k(s)}} a_{i_{k(s)}})(a'_{i_{k(r)}} a_{i_{k(r)}})} \\ &= \frac{a'_{i_{k(r)}} a_{i_{k(r)}}}{a'_{i_{k(s)}} a_{i_{k(s)}}} \\ &= \frac{\sum_{\alpha=1}^{r} a^2_{i_k\alpha}}{\sum_{\alpha=1}^{s} a^2_{i_k\alpha}} \end{aligned}$$

donde $\mathbf{x}_{i_k}$ es un vector fila que contiene los valores del individuo i para las J variables, $\hat{\mathbf{x}}_{i_k}$ es un vector fila que contiene los valores ajustados, $a_{i_{k(s)}}$ es el vector fila con las coordenadas del individuo sobre todas las componentes principales y $a_{i_{k(r)}}$ es el vector que contiene las coordenadas sobre las r componentes principales retenidas y todo ello para la tabla k.

Observese que la calidad de representación es una medida aditiva que se obtiene mediante contribuciones separadas de cada uno de los ejes de la representación. A la cantidad

$$CRF_\alpha E_i = \frac{a^2_{i\alpha}}{\sum_{\alpha=1}^{s} a^2_{i\alpha}}$$

La denominaremos contribución relativa del factor α al elemento i. Al tratarse de una cantidad aditiva puede utilizarse para buscar que "eje" o combinación de ejes permite explicar la variabilidad del individuo y por tanto interpretar su posición en el gráfico factorial. En este contexto, la medida obtenida sirve para interpretar las distancias entre individuos y por tanto es útil para la caracterización de los mismos.

Es inmediato comprobar que la contribución es también la correlación al cuadrado entre $\mathbf{x}_{i_k}$ y $\hat{\mathbf{x}}_{i_k\alpha}$, es decir, puede interpretarse en el contexto de la regresión presentada en el primer capítulo.

Los valores esperados para el individuo i, $\hat{\mathbf{x}}_{i_k}$ pueden escribirse como:

$$\hat{\mathbf{x}}_{i_k} = a_{i_k}\mathbf{B'}_k$$

donde utilizando las ecuaciones de regresión, a_{i_k} se obtienen como:

$$a_{i_k} = \mathbf{x}_{i_k}\mathbf{B}_k(\mathbf{B'}_k\mathbf{B}_k)^{-1}$$

El coeficiente de determinación para esta regresión es entonces, lo que hemos denominado "calidad de la representación" de los individuos. Esta última medida es más general en el sentido que no depende de la selección de las coordenadas, mientras que la relación con la calidad de representación se produce cuando los individuos se representan en "coordenadas principales".

En conclusión:

El coeficiente de determinación de la regresión para cada fila de la matriz de datos es independiente de la elección de los marcadores seleccionados a partir de la Descomposición en Valores Singulares, además, si las filas se representan mediante coordenadas principales, los coeficientes de determinación son los cosenos al cuadrado entre los vectores en el espacio multidimensional $\mathbf{R}^J$ y sus proyecciones sobre las componentes principales.

Desde el punto de vista práctico esto implica que las tradicionales "calidades de representación" pueden interpretarse en cualquier biplot, sea cual sea el valor de γ, como la bondad del ajuste en la predicción de los valores que toma el individuo en cada una de las variables, es decir, sirven para interpretar las relaciones fila-columna. Si además las filas están representadas en coordenadas principales sirven para interpretar las relaciones fila-fila, es decir, la aproximación de los productos escalares y distancias entre filas.

De la misma forma, es posible medir la bondad de ajuste de la aproximación de cada una de las columnas de $\mathbf{X}$. Si $\mathbf{x}_{j_k}$ es un vector columna con los valores en la variable j para todos los individuos. Los valores ajustados $\hat{\mathbf{x}}_{j_k}$ para la columna se pueden escribir como

$$\hat{\mathbf{x}}_{j_k} = \mathbf{A}_k\mathbf{b'}_{j_k}$$

donde $\mathbf{b}_{j_k}$ es la *j-ésima* fila de $\mathbf{B}_k$ que se obtiene por regresión como

$$\mathbf{b}_{j_k} = \mathbf{x'}_j\mathbf{A}_k(\mathbf{A'}_k\mathbf{A}_k)^{-1}$$

La correlación al cuadrado entre $\mathbf{x}_{j_k}$ y $\hat{\mathbf{x}}_{j_k}$ es el coeficiente de determinación de la regresión y puede tomarse como medida de bondad de ajuste de cada columna.

Obsérvese que la medida es también independiente de la elección particular del biplot a partir de la descomposición en valores singulares, es decir, puede ser utilizada en cualquiera de ellos. Si seleccionamos el CMP biplot, coincide además con los cosenos al cuadrado de los vectores que representan a las variables en el espacio multidimensional y el espacio en dimensión reducida. En el RPM biplot, esto no es así pero la medida puede seguir siendo utilizada para estudiar la relación fila-columna.

Tenemos entonces caracterizados los biplots en términos de la descomposición en valores singulares y de la regresión y algunas de sus propiedades referidas a los índices de intrpretación. Trataremos de generalizar esta situación al caso en que tenemos tablas múltiples y deseamos un subespacio de proyección comun a todos ellos.

El problema es entonces, dadas K tablas $\mathbf{X}_k$ buscar un conjunto de vectores ortogonales y de norma 1 que formen una base de $\mathbf{R}^J$ y que sean óptimos en algún sentido, para el conjunto de las K tablas. Llamaremos $\mathbf{V}$ a este conjunto de vectores ortonormales que normalmente tomaremos en dimensión reducida r.

Por el momento vamos a suponer que tenemos dicho conjunto, estudiaremos las propiedades del biplot resultante y después analizaremos diversas soluciones dependiendo del criterio de optimalidad seleccionado.

Llamaremos al subespacio definido por las r primeras columnas de $\mathbf{V}$, "subespacio consenso", "subespacio compromiso", ó "subespacio común".

Cualquier conjunto de vectores ortonormales que define un subespacio genera una representación biplot que se obtiene proyectando las matrices de datos en dicho subespacio. Las propiedades del biplot resultante siguen siendo similares a las descritas anteriormente para el biplot de cada una de las tablas.

Sean $\mathbf{X}_k$ $k = 1, \ldots, K$ las tablas múltiples, todas ellas con el mismo número de columnas, es decir, se dispone de I_k nubes de puntos en $\mathbf{R}^J$. Cada una de las nubes tiene sus propias componentes principales e incluso, aunque los individuos procedan de la misma población, las componentes principales diferirán en algún sentido. El estudio separado de las componentes de cada nube es generalmente difícil de comparar entre las distintas tablas.

Si construimos un sistema de referencia común obtendremos un biplot para cada una de las tablas en el que las coordenadas de las columnas son comunes a todas ellas y definidas por $\mathbf{V}$, es decir,

$$\mathbf{X}_k = \tilde{\mathbf{A}}_k \mathbf{V}' + \tilde{\mathbf{E}}_k \tag{3.3.4}$$

donde

$$\tilde{\mathbf{A}}_k = \mathbf{X}_k \mathbf{V}$$

en la dimensión adecuada, es decir, las coordenadas de las filas de cada tabla son simplemente la proyección de la nube de puntos en $\mathbf{R}^J$ sobre el subespacio definido por $\mathbf{V}$. De la misma manera que los biplots individuales las coordenadas podrían obtenerse por regresión, teniendo en cuenta que

$$\mathbf{X'}_k = \mathbf{V}\tilde{\mathbf{A}}'_k + \mathbf{E}_k$$

tenemos que

$$\tilde{\mathbf{A}}'_k = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{X'}_k \tag{3.3.5}$$

y teniendo en cuenta que los vectores de $\mathbf{V}$ forman un conjunto de vectores ortonormales, la expresión en (3.3.5) es igual a

$$\tilde{\mathbf{A}}_k = \mathbf{X}_k \mathbf{V}$$

que coincide con el resultado anterior.

Los valores ajustados en la nueva aproximación son entonces

$$\tilde{\mathbf{X}}_k = \tilde{\mathbf{A}}_k \mathbf{V}' = \mathbf{X}_k \mathbf{V}\mathbf{V}'$$

es decir, las coordenadas de los puntos proyectados pero en el sistema de referencia de $\mathbf{R}^J$.

3.3.1. Bondad de ajuste

De nuevo podemos medir la bondad de ajuste para los elementos que intervienen en el biplot definido en el espacio consenso.

- De la misma forma que se hizo en (3.3.3), la suma de cuadrados total para cada matriz puede descomponerse como

$$\|\mathbf{X}_k\| = \left\|\tilde{\mathbf{X}}_k\right\| + \left\|\mathbf{X}_k - \tilde{\mathbf{X}}_k\right\|$$

donde la inercia explicada por la solución obtenida, se puede medir como

$$\sigma_k(\mathbf{X}_k, \mathbf{V}) = \frac{\left\|\tilde{\mathbf{X}}_k\right\|}{\|\mathbf{X}_k\|} = \frac{Traza\left(\tilde{\mathbf{X}}'_k\tilde{\mathbf{X}}_k\right)}{Traza\left(\mathbf{X}'_k\mathbf{X}_k\right)} \tag{3.3.6}$$

En general $Traza(\hat{\mathbf{X}}'_k\hat{\mathbf{X}}_k) \geq Traza(\tilde{\mathbf{X}}'_k\tilde{\mathbf{X}}_k)$ y solo serán iguales cuando las componentes principales de todas las matrices sean las mismas, es decir que $\mathbf{V}_k = \mathbf{V}$ para $\mathbf{V}_{k=1,\ldots,K}$. Esto se debe a que $\mathbf{V}$ no es óptimo para cada una de las tablas, sino que es un compromiso de todas ellas.

- Si consideramos la "super" matriz $\mathbf{X}$, en la que se concatenan todas las tablas individuales usando la dimensión común

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \\ \vdots \\ \mathbf{X}_K \end{bmatrix} \tag{3.3.7}$$

Obtenemos una representación biplot para la "super" matriz, como:

$$\begin{aligned} \mathbf{X} &= \tilde{\mathbf{A}}\mathbf{V}' + \mathbf{E} \\ &= \tilde{\mathbf{X}} + \mathbf{E} \end{aligned} \tag{3.3.8}$$

con $\tilde{\mathbf{A}} = \mathbf{XV}$, donde la matriz $\tilde{\mathbf{A}}$ es simplemente la concatenación de las $\mathbf{A}_k$, es decir,

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}}_1 \\ \vdots \\ \tilde{\mathbf{A}}_k \\ \vdots \\ \tilde{\mathbf{A}}_K \end{bmatrix}$$

A partir de esta representación global, es posible medir la bondad de ajuste, o inercia total absorbida; aplicando un razonamiento similar al anterior

$$\begin{aligned} \sigma(\mathbf{X}, \mathbf{V}) &= \frac{\left\|\tilde{\mathbf{X}}\right\|}{\|\mathbf{X}\|} = \frac{Traza\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)}{Traza\left(\mathbf{X}'\mathbf{X}\right)} \\ &= \sum_k \frac{Traza(\tilde{\mathbf{X}}'_k\tilde{\mathbf{X}}_k)}{Traza(\mathbf{X}'\mathbf{X})} \end{aligned}$$

Multiplicando y dividiendo por $Traza(\mathbf{X'}_k\mathbf{X}_k)$, se obtiene:

$$= \sum_k \frac{Traza(\tilde{\mathbf{X}}'_k\tilde{\mathbf{X}}_k)}{Traza(\mathbf{X'}_k\mathbf{X}_k)} \frac{Traza(\mathbf{X'}_k\mathbf{X}_k)}{Traza(\mathbf{X'X})}$$

$$= \sum_k \delta_k \sigma_k(\mathbf{X}_k, \mathbf{V})$$

donde $\delta_k = \frac{Traza(\mathbf{X'}_k\mathbf{X}_k)}{Traza(\mathbf{X'X})}$ es la contribución de cada nube a la inercia global.

En algunos casos podemos considerar nubes normadas para eliminar el tamaño de cada una de ellas de forma que la inercia global es simplemente la media de las inercias de cada nube.

En el biplot inicial, las columnas de $\mathbf{A}_k$ son incorreladas ya que proceden de las columnas de $\mathbf{U}_k$ obtenidas de las descomposición en valores singulares; en el biplot consenso, las columnas de $\tilde{\mathbf{A}}_k$ no son necesariamente incorreladas, la correlación entre las mismas muestra la desviación de la nube de puntos con respecto a los ejes consenso.

- De la misma forma que antes es posible definir la bondad de ajuste para cada individuo en cada una de las tablas, como el coseno al cuadrado entre el vector que representa al individuo en $\mathbf{R}^J$ y su proyección en el espacio consenso

$$\begin{aligned} Cos^2(\mathbf{x}_{i_k}, \tilde{\mathbf{x}}_{i_k}) &= \frac{(\mathbf{x}_{i_k}, \tilde{\mathbf{x}}'_{i_k})^2}{\|\mathbf{x}_{i_k}\| \, \|\tilde{\mathbf{x}}_{i_k}\|} \\ &= \frac{(\mathbf{x}_{i_k}\mathbf{V}\mathbf{V}'\mathbf{x}_{i_k})^2}{(\sum_j \mathbf{x}^2_{ij_k})(\sum_j \tilde{\mathbf{x}}^2_{ij_k})} \\ &= \frac{(\tilde{a}'_{i_{k(r)}}\tilde{a}_{i_{k(r)}})^2}{(\tilde{a}'_{i_{k(s)}}\tilde{a}_{i_{k(s)}})(\tilde{a}'_{i_{k(r)}}\tilde{a}_{i_{k(r)}})} \\ &= \frac{\tilde{a}'_{i_{k(r)}}\tilde{a}_{i_{k(r)}}}{\tilde{a}'_{i_{k(s)}}\tilde{a}_{i_{k(s)}}} \\ &= \frac{\sum_{\alpha=1}^{r} \tilde{a}^2_{i_k\alpha}}{\sum_{\alpha=1}^{s} \tilde{a}^2_{i_k\alpha}} \end{aligned} \tag{3.3.9}$$

- Es posible también definir la calidad de representación para las variables a partir de la descomposición global en (3.3.8) en términos de la correlación al cuadrado entre cada columna de $\mathbf{X}$ y la correspondiente en $\tilde{\mathbf{X}}$.
  Sea $\mathbf{x}_j$ la variable j de la matriz $\mathbf{X}$, la variable j de la matriz $\tilde{\mathbf{X}}$ está formada por las coordenadas de los individuos sobre la *j-ésima* componente de $\mathbf{V}$ que se encuentran en el vector $\mathbf{a}_j = \mathbf{X}\mathbf{v}_j$ donde $\mathbf{X}'\mathbf{X}\mathbf{v}_j = \rho_j \mathbf{v}_j$

  $$cor^2(\mathbf{x}_j, \mathbf{a}_j) = cor^2(\mathbf{x}_j, \mathbf{X}\mathbf{v}_j)$$

  Entonces:

  $$cor^2(\mathbf{x}_j, \mathbf{X}\mathbf{v}_j) = \frac{(\mathbf{x}_j^{'}\mathbf{X}\mathbf{v}_j)^2}{(\mathbf{x}_j^{'}\mathbf{x}_j)(\mathbf{v}'_j\mathbf{X}'\mathbf{X}\mathbf{v}_j)} = \frac{\rho_j^2 v_{jj}^2}{s_{jj}}$$

  Donde v_{jj} es la componente j del vector $\mathbf{v}_j$ y s_{jj} el elemento j de la matriz $\mathbf{X}'\mathbf{X}$.

  La correlación entre las dos variables es la coordenada v_{jj}^2 afectada por un factor ρ_j^2 y por la desviación estándar de la variable original.

- Si los individuos son los mismos en todas las tablas, las coordenadas en cada tabla definen una trayectoria cuyos elementos pueden interpretarse todos en el mismo sistema de referencia, el definido por $\mathbf{V}$ donde las correlaciones definidas previamente pueden utilizarse para la interpretación global de los datos. La trayectoria para el individuo i, está dada por:

  $$\mathbf{T}_i = \begin{bmatrix} \mathbf{a}_{i_1} \\ \vdots \\ \mathbf{a}_{i_k} \\ \vdots \\ \mathbf{a}_{i_K} \end{bmatrix}$$

  La bondad de ajuste de la trayectoria para el individuo i se puede expresar en terminos del promedio de la bondad de ajuste de los K individuos que la conforman. De (3.3.9) se obtiene que:

  $$\frac{1}{K}\sum_{k=1}^{K} Cos^2(\mathbf{x}_{i_k}, \tilde{\mathbf{x}}_{i_k}) = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{\alpha=1}^{r} \tilde{a}_{i_k\alpha}^2}{\sum_{\alpha=1}^{s} \tilde{a}_{i_k\alpha}^2}$$

- También es posible definir pseudo trayectorias para las variables en cada ocasión proyectando pseudo-muestras sobre el biplot. Las pseudo-muestras serían como una de las tablas individuales, completando con cero las que no intervienen. Denotando por $\hat{\mathbf{B}}_k$ las coordenadas de las variables, estas se calculan mediante regresión. Buscamos una matriz $\hat{\mathbf{B}}_k$ de coeficientes de regresión que aproxime lo mejor posible las pseudo muestras

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{X}_k \\ \vdots \\ \mathbf{0} \end{bmatrix} \simeq \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_k \\ \vdots \\ \mathbf{A}_K \end{bmatrix} \hat{\mathbf{B}}'_k = \mathbf{A}\hat{\mathbf{B}}'_k$$

  Luego, a partir de la regresión

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

  y con la parte correspondiente al biplot

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

  Aunque estos dos resultados no tienen porqué ser idénticos.

- Otra ayuda para interpretar la solución final podría ser la proyección de las componentes de cada ocasión sobre las consenso e incluso calcular sus ángulos.

  La componente α de la *k-ésima* ocasión, definida por el vector $\mathbf{v}_{\alpha_k}$ de $\mathbf{V}_k$ puede proyectarse directamente sobre el sistema de referencia definido por $\mathbf{V}$. Las coordenadas que dan las direcciones de las componentes de cada tabla, se proyectan en el consenso a partir de

$$\mathbf{P}_k = \mathbf{V}'_k\mathbf{V}$$

  Para $\mathbf{v}_\delta$ vector de $\mathbf{V}$ que da la direccion de la componente δ del espacio consenso, el coseno al cuadrado del ángulo

$$Cos^2(\mathbf{v}_{\alpha_k}, \mathbf{v}_\delta) = \frac{\mathbf{v}'_{\alpha_k}\mathbf{v}_\delta}{\|\mathbf{v}_{\alpha_k}\| \, \|\mathbf{v}_\delta\|} = \mathbf{v}'_{\alpha_k}\mathbf{v}_\delta$$
$$= \sum_j \mathbf{v}'_{j\alpha_k}\mathbf{v}_{j\delta}$$

  Es conveniente conservar el signo para ver el sentido de la asociación.

Todo lo expuesto anteriormente está pensado para biplots del tipo RMP.

3.3.2. Procedimientos para hallar el Consenso

Vamos a presentar distintas formas de hallar el espacio consenso aplicando los conceptos y definiciones presentados en la sección precedente.

3.3.2.1. Criterio de la selección del mejor conjunto de ejes

Si seleccionamos como sistema de referencia común la matriz $\mathbf{V}_\kappa$ que genera el espacio de representacion de la tabla $\mathbf{X}_\kappa$ para alguna de las K tablas, obtenemos un biplot para cada una de las tablas en este sistema, mediante:

$$\mathbf{X}_k = \tilde{\mathbf{A}}_{k_\kappa} \mathbf{V'}_\kappa + \tilde{\mathbf{E}}_{k_\kappa}$$

donde las coordenadas de las columnas son comunes a todas las tablas y

$$\tilde{\mathbf{A}}_{k_\kappa} = \mathbf{X}_k \mathbf{V}_\kappa$$

las coordenadas de las filas de cada tabla son la proyección de la nube de individuos en $\mathbf{R}^J$ sobre el subespacio definido por $\mathbf{V}_\kappa$. Estas coordenadas se pueden obtener por regresión, a partir de

$$\hat{\mathbf{A}}'_{k_\kappa} = (\mathbf{V'}_\kappa \mathbf{V}_\kappa)^{-1} \mathbf{V'}_\kappa \mathbf{X'}_k$$

que es equivalente a

$$\tilde{\mathbf{A}}_{k_\kappa} = \mathbf{X}_k \mathbf{V}_\kappa$$

Los valores ajustados en esta aproximación son:

$$\tilde{\mathbf{X}}_{k_\kappa} = \tilde{\mathbf{A}}_{k_\kappa} \mathbf{V'}_\kappa = \mathbf{X}_k \mathbf{V}_\kappa \mathbf{V'}_\kappa$$

De manera similar que en las ecuaciones (3.3.3) y (3.3.6), la suma de cuadrados total para cada matriz se puede descomponer como

$$\|\mathbf{X}_k\| = \left\|\tilde{\mathbf{X}}_{k_\kappa}\right\| + \left\|\mathbf{X}_k - \tilde{\mathbf{X}}_{k_\kappa}\right\|$$

donde la inercia de la tabla $\mathbf{X}_k$ explicada por el subespacio generado por $\mathbf{V}_\kappa$, se puede expresar como:

$$\sigma_{k_\kappa}(\mathbf{X}_k, \mathbf{V}_\kappa) = \frac{\left\|\tilde{\mathbf{X}}_{k_\kappa}\right\|}{\|\mathbf{X}_k\|} = \frac{Traza\left(\tilde{\mathbf{X}}'_{k_\kappa} \tilde{\mathbf{X}}_{k_\kappa}\right)}{Traza\left(\mathbf{X'}_k \mathbf{X}_k\right)}$$

El valor $\sigma_{k_\kappa} = 1$ cuando el sistema de referencia de la tabla $\mathbf{X}_k$ está generado por $\mathbf{V}_\kappa$.

Definiendo

$$\sigma_\kappa = \frac{1}{K}\sum_{k=1}^{K}\sigma_{k_\kappa}$$

Se obtiene la cantidad de inercia promedio que retiene el subespacio generado por $\mathbf{V}_\kappa$ al proyectar todas las tablas sobre el.

El espacio consenso estará generado por la matriz $\mathbf{V}_{k*}$ correspondiente al sistema de referencia de la tabla $\mathbf{X}_{k*}$, que satisaface

$$\sigma_{k*} \geq \sigma_k \qquad \forall k = 1, \ldots, K$$

Donde los individuos de la tabla $\mathbf{X}_{k*}$ sirven de elementos de comparación para analizar la dispersión de las trayectorias.

A partir de las matrices $\tilde{\mathbf{A}}_{k_{k*}}$ y $\mathbf{V}_{k*}$ se construye la representación biplot

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \\ \vdots \\ \mathbf{X}_K \end{bmatrix} \simeq \begin{bmatrix} \tilde{\mathbf{A}}_{1_{k*}} \\ \vdots \\ \tilde{\mathbf{A}}_{k_{k*}} \\ \vdots \\ \tilde{\mathbf{A}}_{K_{k*}} \end{bmatrix} \mathbf{V}'_{k*} = \mathbf{A}\mathbf{B}'$$

3.3.2.2. Criterio de maximización de la inercia explicada

Partiendo de la "super" matriz $\mathbf{X}$, definida en (3.3.7), en la que se concatenan todas las tablas usando la dimensión común; la configuración biplot de $\mathbf{X} = \mathbf{A}\mathbf{B}'$ se puede conseguir:

- Por medio de regresiones alternadas como se vio en el capítulo I.
- Otra forma de encontrar la representación biplot es considerado

  $$\mathbf{X}'\mathbf{X} = \sum_k \mathbf{X}'_k\mathbf{X}_k = \sum_k \mathbf{S}_k = \mathbf{S}$$

  donde la descomposición espectral de

  $$\mathbf{S} = \mathbf{V}_s\mathbf{\Lambda}_s^2\mathbf{V}'_s \tag{3.3.10}$$

Luego la matriz $\mathbf{V}_s$ genera el subespacio consenso y las coordenadas de los individuos se calculan por medio de la proyección de las tablas sobre este espacio, es decir:

$$\mathbf{A}_s = \mathbf{X}\mathbf{V}_s \tag{3.3.11}$$

Así, la representación biplot está dada por:

$$\mathbf{X} = \mathbf{A}_s\mathbf{V'}_s$$

- Teniendo en cuenta que la descomposición en valores singulares de $\mathbf{X}$ se puede expresar como

$$\mathbf{X} = \mathbf{U}_s\mathbf{\Lambda}_s\mathbf{V'}_s \tag{3.3.12}$$

las matrices $\mathbf{A}_s = \mathbf{U}_s\mathbf{\Lambda}_s$ y $\mathbf{V'}_s$ conforman la configuración biplot

3.3.2.3. Criterio de maximización de inercia explicada con matrices reducidas I

En el análisis simultanéo de tablas estas no intervienen de la misma forma, dado que poseen estructura diferente. Por tanto, se puede tener una tabla con fuerte estructura, es decir que sus variables esten muy correlacionadas y esta va a influenciar mas la construcción del espacio consenso.

La forma de equilibrar la influencia de las tablas está inspirada en el concepto que formula el AFM y es dando a cada variable un peso, este peso debe ser el mismo para todas las variables de una misma tabla con el fin de conservar la estructura interna. El peso que se va a dar a cada una de las variables de una tabla es el inverso de su respectivo primer valor singular; es decir, $\frac{1}{\lambda_{k_1}}$ para las variables de la tabla $\mathbf{X}_k$.

Esto significa que el análisis se realiza a partir de la "super" matriz

$$\mathbf{X}_I = \begin{bmatrix} \frac{1}{\lambda_{1_1}}\mathbf{X}_1 \\ \vdots \\ \frac{1}{\lambda_{k_1}}\mathbf{X}_k \\ \vdots \\ \frac{1}{\lambda_{K_1}}\mathbf{X}_K \end{bmatrix} \tag{3.3.13}$$

Luego la representación biplot se puede realizar a partir de $\mathbf{X}_I$ ó teniendo en cuenta que:

$$\mathbf{X'}_I\mathbf{X}_I = \sum_k \frac{\mathbf{X'}_k\mathbf{X}_k}{\lambda_{k_1}^2} = \sum_k \frac{\mathbf{S}_k}{\lambda_{k_1}^2} = \mathbf{S}_I$$

y la representación está dada por:

$$\mathbf{X}_I = \mathbf{A}_I\mathbf{V'}_I$$

3.3.2.4. Criterio de maximización de inercia explicada con matrices reducidas II

Consiste en tener en cuenta los r primeros valores propios no nulos de cada tabla y equilibrar su influencia ponderando las variables de la misma tabla por el inverso de la suma de los r valores propios.

El análisis se realiza a partir de

$$\mathbf{X}_{II} = \begin{bmatrix} \frac{1}{\sqrt{\sum_r \lambda_{1_r}^2}}\mathbf{X}_1 \\ \vdots \\ \frac{1}{\sqrt{\sum_r \lambda_{k_r}^2}}\mathbf{X}_k \\ \vdots \\ \frac{1}{\sqrt{\sum_r \lambda_{K_r}^2}}\mathbf{X}_K \end{bmatrix} \tag{3.3.14}$$

Donde:

$$\mathbf{X'}_{II}\mathbf{X}_{II} = \sum_k \frac{\mathbf{X'}_k\mathbf{X}_k}{\sum_r \lambda_{k_r}^2} = \sum_k \frac{\mathbf{S}_k}{\sum_r \lambda_{k_r}^2} = \mathbf{S}_{II}$$

Este criterio es interesante en el caso que las tablas tengan "normas" diferentes. Entendiendo como norma la suma de los r primeros valores propios.

Para que el caso que nos ocupa, donde el análisis y representación de la información se hace generalmente en dos dimensiones, ponderar por el inverso de los dos primeros valores propios, seria suficiente para ver el comportamiento en el plano .

En el caso que la "norma" de cada tabla coincida con la norma de Hilbert Schmidt, estariamos generando el método Statis Dual con objetos normados. Ver 2.4.2.2 página 63.

De manera general, para este caso la representación biplot se denotada por

$$\mathbf{X}_{II} = \mathbf{A}_{II}\mathbf{V'}_{II}$$

3.3.2.5. Criterio inducido por STATIS DUAL

Para este caso se parte de la "super" matriz

$$\mathbf{X}_{\beta} = \begin{bmatrix} \sqrt{\beta_1}\mathbf{X}_1 \\ \vdots \\ \sqrt{\beta_k}\mathbf{X}_k \\ \vdots \\ \sqrt{\beta_K}\mathbf{X}_K \end{bmatrix} \tag{3.3.15}$$

Para β definido en la seccion 2.4.2.1 página 63, y se tiene que:

$$\mathbf{X'}_{\beta}\mathbf{X}_{\beta} = \sum_k \beta_k \mathbf{X'}_k\mathbf{X}_k = \sum_k \beta_k \mathbf{S}_k = \mathbf{S}_{\beta}$$

Similar a los casos anteriores la representación biplot está dada por

$$\mathbf{X}_{\beta} = \mathbf{A}_{\beta}\mathbf{V'}_{\beta}$$

3.3.2.6. Criterio inducido por Meta Componentes

Se trata de encontrar los vectores ν que conforman el espacio $\mathbf{V}$ de manera que el ángulo entre ν y el vector más próximo en el espacio generado por las componentes principales de los grupos sea mínimo.

Como se vió en 2.6.2 página 80, la matriz que contiene los cosenos al cuadrado de los ángulos que forman las direcciones principales de todas las tablas es:

$$\mathbb{H} = \sum_{k=1}^{k} \mathbf{V}_k\mathbf{V}_k'$$

Luego el sistema $\mathbf{V}$ de vectores ν buscado, es la matriz de vectores propios de la descomposición espectral de

$$\mathbb{H} = \mathbf{V}_H\mathbf{\Lambda}_H^2\mathbf{V'}_H$$

La matriz $\mathbf{A}_H$ se puede conseguir de dos maneras:

- Como proyección de la "super" matriz $\mathbf{X}$ sobre el sistema $\mathbf{V}_H$, así

$$\mathbf{A}_H = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \\ \vdots \\ \mathbf{X}_K \end{bmatrix} \mathbf{V}_H$$

- A partir de estimacion de $\mathbf{A}_H$ por mínimos cuadrados, en el modelo:

$$\mathbf{X} = \hat{\mathbf{A}}_H \mathbf{V}_H + \mathbf{E}_H$$

Luego la representación biplot es:

$$\mathbf{X} = \mathbf{A}_H \mathbf{V'}_H$$

3.3.3. Índices de comparación de las representaciones inducidas por los diferentes criterios

Con los criterios vistos en la sección precedente se obtiene el espacio consenso y la respectiva representación biplot de las K tablas que se desean analizar.

La pregunta que naturalmente nace es, si los criterios satisfacen las propiedades que el análisis multivariante propone, entonces ¿cuál es el mejor criterio para aplicar a los datos?

A esta pregunta se ha tratado de dar respuesta mediante comparaciones entre los objetos y técnicas que cada método impone para su desarrollo, sin embargo no hay una herramienta que permita conocer la comparación entre 2 o más técnicas.

A continuación se proponen dos indicadores para comparar la calidad de representación que cada criterio hace de la información.

3.3.3.1. Pérdida de inercia

A partir de 2.5.2.1 *Índices de selección de ejes del espacio Consenso* pag 70, donde se vió que para cada tabla $\mathbf{X}_k$ se tiene que $\mathbf{S}_k = \mathbf{V}_k \mathbf{\Lambda}_k^2 \mathbf{V}_k^{'}$ y por tanto $\mathbf{V}_k^{'} \mathbf{S}_k \mathbf{V}_k = \mathbf{\Lambda}_k^2$ entonces $\mathbf{v}_{r_k}^{'} \mathbf{S}_k \mathbf{v}_{r_k} = \lambda_{r_k}^2$ representa la inercia explicada por el eje r para $\mathbf{S}_k$.

Al realizar la representación de la tabla k en los R primeros ejes factoriales, el valor de la inercia explicada corresponde a $\sum_{r=1}^{R} \lambda_{r_k}^2$, que se puede expresar como:

$$traza(\mathbf{V}_{r_k}^{'}\mathbf{S}_k\mathbf{V}_{r_k}) = traza(\mathbf{\Lambda}_{r_k}^2)$$

para $\mathbf{V}_{r_k}$ y $\mathbf{\Lambda}_{r_k}^2$ reducidas a los primeros R vectores.

Ademas:

$$\begin{aligned} traza(\mathbf{V}_{r_k}^{'}\mathbf{S}_k\mathbf{V}_{r_k}) &= traza(\mathbf{V}_{r_k}^{'}\mathbf{X'}_k\mathbf{X}_k\mathbf{V}_{r_k}) \\ &= traza([\mathbf{V}_{r_k}^{'}\mathbf{X'}_k][\mathbf{X}_k\mathbf{V}_{r_k}] \\ &= traza(\hat{\mathbf{A}}'_{r_k}\hat{\mathbf{A}}_{r_k}) \end{aligned}$$

Por tanto, la inercia de la tabla $\mathbf{X}_k$ explicada por los primeros R factores se puede expresar como $traza(\hat{\mathbf{A}}'_{r_k}\hat{\mathbf{A}}_{r_k})$, para $\hat{\mathbf{A}}_{r_k}$ matriz de coordenadas de los individuos en rango reducido R.

De manera análoga se puede conocer la inercia explicada por los R primeros ejes factoriales inducidos por el espacio consenso para la tabla k. Para $\mathbf{V}_r$ matriz de los primeros R vectores propios del espacio consenso, se tiene:

$$\begin{aligned} traza(\mathbf{V}_{r}^{'}\mathbf{S}_k\mathbf{V}_{r}) &= traza(\mathbf{V}_{r}^{'}\mathbf{X'}_k\mathbf{X}_k\mathbf{V}_{r}) \\ &= traza([\mathbf{V}_{r}^{'}\mathbf{X'}_k][\mathbf{X}_k\mathbf{V}_{r}] \\ &= traza(\tilde{\mathbf{A}}'_{r_k}\tilde{\mathbf{A}}_{r_k}) \end{aligned}$$

Donde $\tilde{\mathbf{A}}_{r_k}$ es la matriz de coordenadas de los individuos de la tabla k en el espacio consenso en dimension reducida.

Comparando estos valores, se tiene que:

$$traza(\hat{\mathbf{A}}'_{r_k}\hat{\mathbf{A}}_{r_k}) \geq traza(\tilde{\mathbf{A}}'_{r_k}\tilde{\mathbf{A}}_{r_k})$$

Debido a que la mejor representación de los datos se da en el espacio generado por los vectores propios de la respectiva tabla.

Y por tanto, se tiene que:

$$0 \leq 1 - \frac{traza(\tilde{\mathbf{A}}'_{r_k}\tilde{\mathbf{A}}_{r_k})}{traza(\hat{\mathbf{A}}'_{r_k}\hat{\mathbf{A}}_{r_k})}$$

Este índice es el mismo definido en 2.5.2.1 pag 70, y denotado $\xi(k,\nu)$ que mide el porcentaje de inercia que pierde la nube N_I^k de los individuos de la

tabla k cuando se proyecta sobre el subespacio definido por $(\nu_r)_{r=1,\ldots,R}$ en cambio de proyectarla sobre sus R primeras componentes estándar.

Toma valores entre [0,1), cero en el caso que los factores generados por el espacio consenso coincidan con los factores generados por la tabla

Entonces al proyectar todas las tablas sobre el espacio consenso, el índice:

$$\xi(.,\nu) = \frac{1}{K}\sum_{k=1}^{K}\left[1 - \frac{traza(\tilde{\mathbf{A}}'_{r_k}\tilde{\mathbf{A}}_{r_k})}{traza(\hat{\mathbf{A}}'_{r_k}\hat{\mathbf{A}}_{r_k})}\right]$$

mide el promedio de pérdida de inercia de todas las nubes N_I^k al ser proyectadas sobre los ejes definidos por el espacio consenso.

Este índice tiene la ventaja que se puede calcular a partir de las cooordenadas de los individuos en el espacio consenso y en el espacio generado por la tabla, en la dimensión respectiva.

3.3.3.2. Proximidad entre factores

La proximidad entre dos factores se puede medir por el ángulo que ellos forman o más exactamente por el coseno al cuadrado de dicho ángulo.
Así una forma de medir la similitud entre el espacio consenso y cada una de las tablas es conociendo el valor del coseno al cuadrado de los ángulos que forman sus factores.

De acuerdo a 2.6.1.2 pag 78, para $\mathbf{V}_r$ y $\mathbf{V}_{r_k}$ matrices en dimensión reducida r de los vectores propios que generan el espacio consenso y la tabla k respectivamente, se tiene que para θ_{rj} el ángulo entre la *r-ésima* componente de $\mathbf{V}_r$ y la *j-ésima* componente de $\mathbf{V}_{r_k}$

$$\begin{aligned}\sum_{r=1}^{R}\sum_{j=1}^{R} cos^2\theta_{rj} &= traza(\mathbf{V}'_r\mathbf{V}_{r_k}\mathbf{V}'_{r_k}\mathbf{V}_r)\\ &= traza(\mathbf{B}'_r\mathbf{B}_{r_k}\mathbf{B}'_{r_k}\mathbf{B}_r)\end{aligned}$$

Esta expresión toma el valor mínimo cero cuando los dos espacios son ortogonales y valor máximo R cuando los dos subespacio coinciden. Luego la expresión:

$$\eta(\mathbf{V}_r) = \frac{1}{R}traza(\mathbf{B}'_r\mathbf{B}_{r_k}\mathbf{B}'_{r_k}\mathbf{B}_r)$$

toma valores entre cero y uno y permite conocer el ángulo promedio entre la tabla k y el espacio consenso.

Entonces para medir la similitud entre los factores de un espacio consenso $\mathbf{V}_r$ con los factores de todas las tablas, vamos a utilizar el índice:

$$\eta(\mathbf{V}_r) = \frac{1}{RK} \sum_{k=1}^{K} traza(\mathbf{B'}_r \mathbf{B}_{r_k} \mathbf{B'}_{r_k} \mathbf{B}_r)$$

que toma valores entre cero y uno, y se calcula a partir de las coordenadas de las variables en el espacio consenso y de las coordenadas de las variables de las respectivas tablas, en dimension reducida.

Capítulo 4

Aplicación a datos Reales

4.1. Introducción

Los datos que se presentan a continuación son la base de la aplicación del método que se desarrolla en este trabajo. Se busca dar a conocer respuestas metodológicas y no se pretende realizar un análisis económico de la información.

La Organización para la Cooperación Europea Económica (la OECE), fue formada en 1948 para administrar la ayuda americana y canadiense en el marco del Plan Marshall para la reconstrucción de Europa después de la segunda Guerra Mundial.

En los años 1950 la OECE proporcionó el marco para negociaciones que apuntaban a la determinación de condiciones para establecer el libre Comercio europeo y atraer otros miembros de OECE en una base multilateral.

En 1957 después de firmarse en Roma, dos tratados que daban existencia a la Comunidad Económica Europea (CEE) y a la Comunidad de la Energía Atómica (EURATOM); fue elaborada la reforma de la OECE por medio de la convención que se firmó en diciembre de 1960 donde la Organización para la Cooperación Económica y el Desarrollo (OCDE) oficialmente reemplazó la OECE en septiembre de 1961. Los firmantes fueron los representantes de Francia, Bélgica, Luxemburgo, Países Bajos, Italia y la Republica Federal de Alemania. En 1973, tres nuevos países ingresaron en la CEE: el Reino Unido, Dinamarca e Irlanda. Nacía la "Europa de los Nueve". La actual Organización para la Cooperación Económica y el Desarrollo (OCDE) es hoy

una organización económica internacional conformada por 32 países cuyos representantes se reúnen para intercambiar información y armonizar políticas con el objetivo de maximizar su crecimiento económico y coayudar a su desarrollo y al de los países no miembros. Se considera que la OCDE agrupa a los países más avanzados y desarrollados del planeta, siendo apodada como club de países ricos. Los países miembros son los que proporcionan el 70 % del mercado mundial y representan el 80 % del PNB mundial (Rapport annuel 2007, OCDE).

Objetivos

La OCDE se ha constituido en uno de los foros mundiales más influyentes, en el que se analiza y se establecen orientaciones sobre temas de relevancia internacional como economía, educación y medioambiente. El principal requisito para ser país miembro de la OCDE es liberalizar progresivamente los movimientos de capitales y de servicios. Los países miembros se comprometen a aplicar los principios de: liberalización, no discriminación, trato nacional y trato equivalente. Sus principales objetivos son:

- Contribuir a una sana expansión económica en los países miembros, así como no miembros, en vías de desarrollo económico.
- Favorecer la expansión del comercio mundial sobre una base multilateral y no discriminatoria conforme a las obligaciones internacionales.
- Realizar la mayor expansión posible de la economía y el empleo y un progreso en el nivel de vida dentro de los países miembros, manteniendo la estabilidad financiera y contribuyendo así al desarrollo de la economía mundial.

4.2. Los Datos

Después de 1963 el observatorio de la OCDE publica anualmente indicadores económicos de los países miembros de la organización.

El perfil estadístico por país es un resumen seleccionado de más de 40 bases de datos disponibles en la biblioteca de la OCDE. Describe las áreas de: Población y migración, producción e ingreso, globalización, precios, energía, empleo, ciencia y tecnología, medio ambiente, educación, finanzas públicas

y calidad de vida. Las áreas de ciencia y tecnología y medio ambiente son las más recientes y no todos los países tienen valores en las variables.

Muchos de los índices presentados en los perfiles son redundantes y para algunos no todos los países tienen valor; por tanto seleccionamos 10 que consideramos no redundantes y para los que faltan muy pocos datos. Se registra el área al que pertenece el índice en el perfil del país, la sigla con que se distinguirá en los análisis y la descripción de este.

Área	Sigla	Descripción
Producción	PIB	Producto interno Bruto en US por habitante al precio y tasa de cambio corriente
Producción	PIN	Porcentaje del producto interno bruto proveniente de la industria
Producción	PSE	Porcentaje del producto interno bruto proveniente de servicios
Productividad	EAG	Porcentaje de población empleada en agricultura
Productividad	EIN	Porcentaje de población empleada la industria
Globalización	EBS	Exportación de bienes en billones de dólares
Globalización	IBS	Importación de bienes en billones de dólares
Globalización	BCC	Balance de la cuenta corriente como porcentaje del PIB
Precios	TIN	Tasa de interés a largo plazo
Energía	ELC	Generación de electricidad- terawatios por hora

Tabla 4.1: Variables bajo análisis.

De los 32 países que están registrados, seleccionamos 17 que tienen completa la mayoría de la información. A continuación se presentan los nombres de los países y las siglas con las que serán representados en el análisis.

El periodo que describen los índices corresponde a 9 años comprendidos entre el 2000 al 2008, último año de actualización de la información para la mayoría de los índices y países.

En las salidas de las tablas y de los gráficos, la sigla del país se completa con 1, 2, . . .,9, para indicar la tabla a la que pertenecen. Así: F1 corresponde a Francia para los valores correspondientes al año 2000 y G9 a Grecia para los valores del 2008.

Sigla	Nombre del país
Al	Alemania
Au	Australia
B	Bélgica
C	Canadá
D	Dinamarca
E	España
USA	Estados Unidos
F	Francia
G	Grecia
Ir	Irlanda
It	Italia
J	Japón
N	Noruega
PB	Países Bajos
P	Portugal
Ru	Reino Unido
S	Suecia

Tabla 4.2: Sigla y nombre de los países que intervienen en el análisis

La estructura de la información es la siguiente:
Para cada uno de los nueve años se considera una tabla con 10 variables y 17 países, (ver información en el apéndice A) Los análisis estadísticos que utilizan las técnicas descritas en esta tesis doctoral los hemos obtenido con programas escritos en MATLAB específicamente para este fin (Ver Vicente-Villardón, 2011 a), o con programas realizados en R, ya que por tratarse de métodos no descritos hasta ahora no están contemplados en paquetes estándar. Cuando se trata de técnicas estándar, se ha utilizado el programa SPAD versión 7 y el paquete FactoMineR.

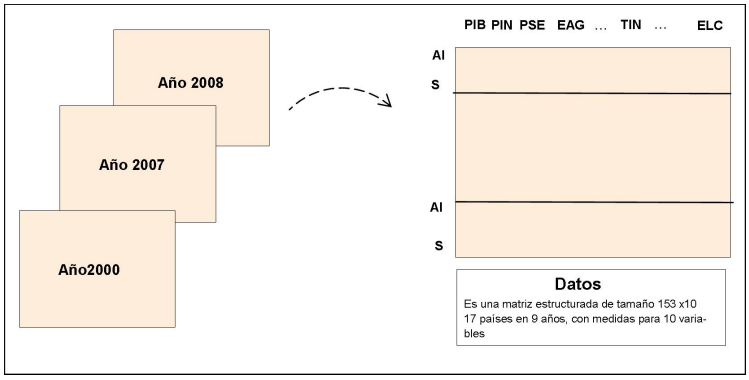

Figura 4.1: Esquema de los datos para los 9 años

4.3. Estadística descriptiva

Para conocer acerca de la información que nos proponemos analizar presentamos algunas estadísticas que permiten hacerse una idea de los datos.

Año	Est	Producción			Productiv		Globalización			Pr	En
		PIB	PIN	PSE	EAG	EIN	EBS	IBS	BCC	TIN	ELC
2000	Prom	26825,98	29,2	68,22	5,26	26,91	220,89	240,81	0,06	5,4	489,91
	Des	4866,2	5,57	5,45	4,16	4,63	211,75	296,86	5,52	0,98	953,13
2001	Prom	27822,73	28,51	68,99	5,02	26,62	212,76	231,58	0,36	4,93	479,34
	Des	4799,81	5,33	5,23	3,96	4,53	200,51	279,07	5,57	0,98	908,78
2002	Prom	28861,62	27,8	69,88	4,86	26,08	219,36	237,48	0,67	4,84	493,72
	Des	4740,6	5,38	5,28	3,81	4,63	199,3	283,38	4,82	1	952,63
2003	Prom	29548,27	27,24	70,55	4,74	25,66	251,95	273,85	0,99	4,12	496,4
	Des	4920,75	5,12	4,97	3,83	4,56	222,58	310,26	4,9	0,85	958,9
2004	Prom	31103,65	27,28	70,55	4,46	25,17	297,34	325,27	1,09	4,09	507,91
	Des	5556,21	5,27	5,14	3,29	4,28	262,17	363,48	5,36	0,71	980,8
2005	Prom	32620,85	27,41	70,69	4,39	24,95	322,14	362,37	0,69	3,46	521,22
	Des	6287,03	5,66	5,51	3,22	4,06	283,01	410,5	6,57	0,64	1008,7
2006	Prom	34528,78	27,61	70,59	4,27	24,83	362,38	410,26	0,5	3,85	521,68
	Des	6947,52	5,88	5,76	3,15	4,04	323,23	458,82	7,54	0,62	1010,5
2007	Prom	36223,27	27,37	70,84	4,16	24,71	413,38	457,71	0,12	4,25	529,07
	Des	7174,09	5,61	5,46	3,09	4,03	371,76	485,53	7,61	0,7	1022,4

2008	Prom	36892,27	27,4	70,91	4,07	24,1	452,78	503,7	-0,43	4,12	526,87
	Des	7749,52	6,29	6,1	3,05	3,85	410,66	527,8	8,12	0,76	1022,1
Total	Prom	31603,05	27,8	70,14	4,58	25,5	305,89	338,11	0,45	4,34	507,36
	Des	6792,89	5,47	5,37	3,46	4,28	290,43	391,48	6,18	0,98	954,5

Tabla 4.3: Promedio y desviación típica de las variables para cada año y total

Como los datos están medidos en diferentes escalas, para la aplicación de los diferentes métodos se escoge trabajar con los datos centrados y reducidos al interior de cada tabla. Cada individuo estará afectado por el mismo peso.

4.4. Análisis gráfico

Con el fin de apoyar los análisis posteriores se presentan las graficas de las variables a lo largo del tiempo.

Variables de producción. EL producto interno bruto (PIB) y el porcentaje

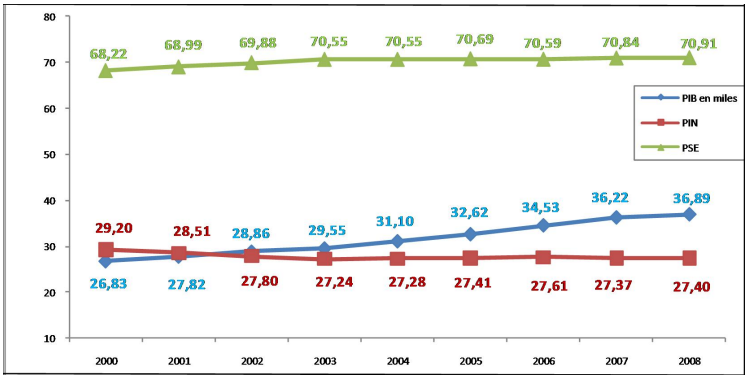

Figura 4.2: Tendencia de las variables de producción

proveniente de servicios (PSE) se incrementan a lo largo del periodo, mientras que el porcentaje proveniente de la industria (PIN) decrece.

Variables de productividad. La tendencia de las dos variables consideradas, porcentaje de la población empleada en industria y población empleada en agricultura, presentan tendencia decreciente a lo largo de los 9 años bajo estudio.

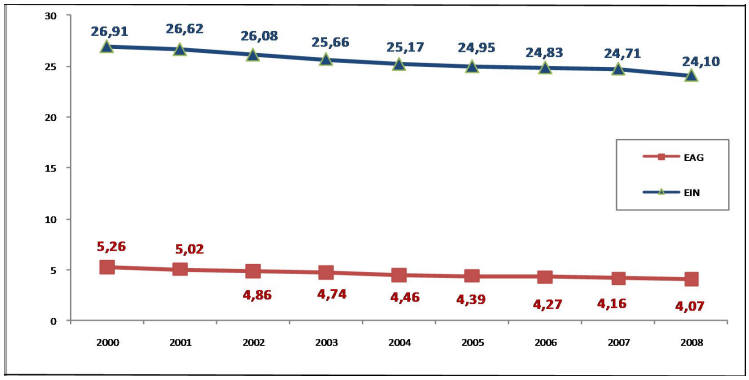

Figura 4.3: Tendencia de las variables de productividad

Variable de precio. La tasa de interés a largo plazo presenta tendencia creciente hasta el año 2005, a partir de dicha fecha inicia la tendencia creciente hasta el final del periodo

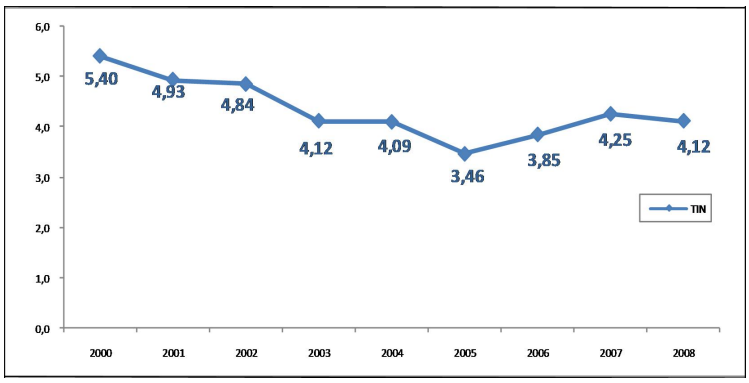

Figura 4.4: Tasa de interés a largo plazo

Variables de globalización. Tanto las importaciones como las exportaciones de bienes y servicios presentan tendencia creciente, mientras que el Balance de la cuenta corriente como porcentaje del PIB (BCC) tiene forma de parábola, creciente desde el año 2000 hasta el año 2004 y decrece aceleradamente presentando valores negativos en el 2008.

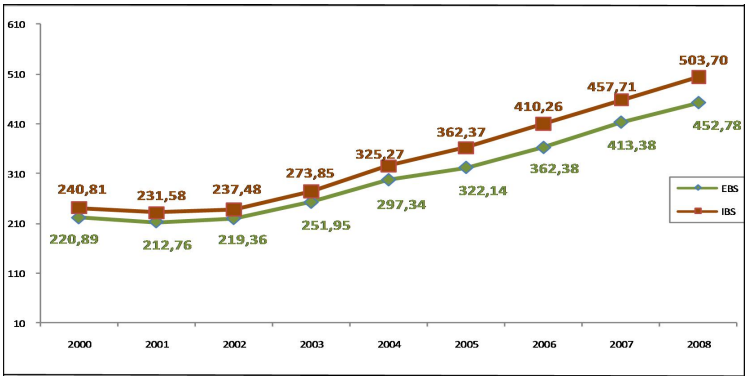

Figura 4.5: Tendencia de las variables de importación y exportación de bienes y servicios

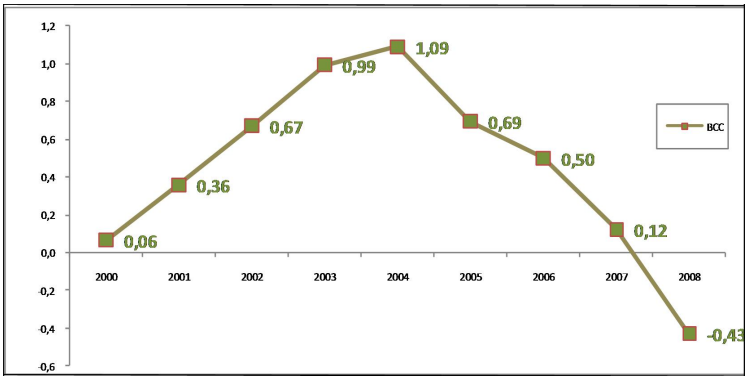

Figura 4.6: Balance de la cuenta corriente como porcentaje del PIB

Variable de electricidad. Generación de electricidad en terawatios por hora presenta una tendencia creciente a lo largo del periodo

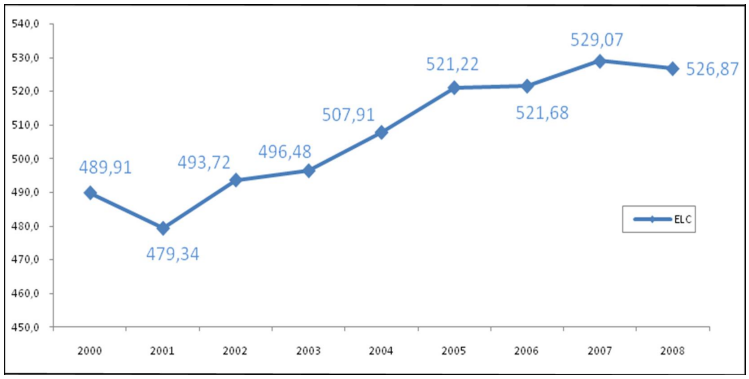

Figura 4.7: Generación de electricidad en terawatios por hora

4.5. Interestructura

Consiste en representar por medio del análisis canónico de varianzas, las medias de las variables de cada año, en un subespacio de baja dimensión cuyos ejes dan la máxima discriminación entre grupos. La representación de regiones de confianza permite visualizar e identificar los años que son significativamente diferentes, a partir de la información multivariante aportada por las variables en estudio.

Los análisis individuales muestran que en general las variables presentan bajo nivel de inercia que explica la varianza entre los grupos. Las variables PIB (39,7), TIN (51,5) y EBS (12,6) son las que explican en mayor proporción la diferencia a través del tiempo. Mientras que BBC, EAG , PIN y PSE presentan alto nivel de inercia residual y baja explicación para diferenciar los años.

Los dos primeros ejes presentan diferencias significativas respecto a los demás y explican el 98.8 % es decir, casi la totalidad de la información.

Variable	Total	Explained	Residual	F	sig.
BCC	152	0.822	151.178	0.098	0.99924
EAG	152	1.905	150.095	0.228	0.98516
EBS	152	12.687	139.313	1.639	0.11861
EIN	152	6.521	145.479	0.807	0.59758
ELC	152	0.049	151.951	0.006	1
IBS	152	9.18	142.82	1.157	0.32938
PIB	152	39.795	112.205	6.384	0
PIN	152	2.033	149.967	0.244	0.98165
PSE	152	4.126	147.874	0.502	0.85308
TIN	152	51.58	100.42	9.245	0

Tabla 4.4: Anovas individuales

Dimension	Eigenv.	Porcent.	Porc Acum.	F	p-val
1	2.094	91.803	91.803	78.933	0
2	0.582	7.093	98.896	6.099	0
3	0.202	0.852	99.748	0.733	0.663
4	0.077	0.125	99.873	0.108	0.999
5	0.063	0.083	99.955	0.071	1
6	0.042	0.036	99.992	0.031	1
7	0.018	0.007	99.998	0.006	1
8	0.009	0.002	100	0.001	1
9	0	0	100	0	1

Tabla 4.5: Valores propios e inercia de la interestructura determinada por el análisis canónico de varianza

4.5.1. Interpretación de la interestructura

Para realizar el análisis de la interestructura se presenta el biplot de la interestructura

El primer eje se interpreta como un factor de tiempo, de hecho se constata la una evolución temporal cuasi-lineal a través del tiempo sobre este eje. Lo que significa que las variables EAG, EIN y ELC (ver fig.4.3 y 4.7) que estan correlacionadas con este eje varian de forma lineal de acuerdo al tiempo. Es

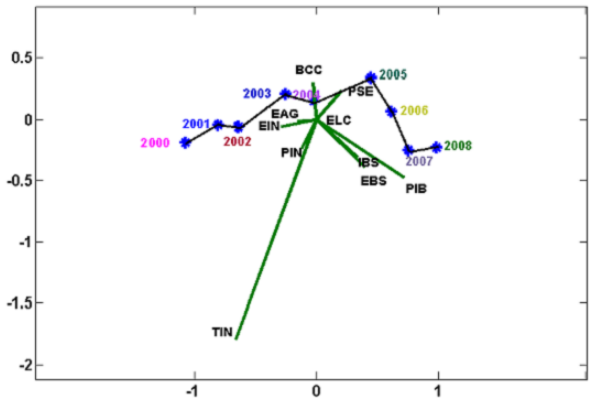

Figura 4.8: Representación de las tablas en el plano 1-2 (interestructura)

decir que la productividad va en receso (correlación negativa) mientras que la producción de energia aumenta al transcurrir el tiempo.

Las coordenadas de las tablas sobre el segundo eje varian de manera mas amplia. El eje 2 está correlacionado principalmente con las variables TIN, PIB, EBS e IBS.

4.6. Análisis individuales

Es interesante analizar de manera individual las tablas para conocer el comportamiento de los datos en cada uno de los años, de tal manera, que en el análisis del espacio consenso sea más facil la interpretación de las variables, de las trayectorias y en general de los resultados obtenidos (ver resultados en el apéndice B).

Se puede observar que las tablas presentan estructura equilibrada sobre los dos primeros ejes, debido a que ninguna los influencia de manera preponderante.
Los dos primeros valores propios son similares para las tablas hasta el 2004, a partir del 2005 el segundo valor propio aumenta y por tanto la inercia que retiene el primer eje tiende a disminuir mientras que en el segundo eje tiende a aumentar, manteniendose estable la inercia retenida en el plano, ver cuadro 4.6

Tabla	Valores propios		Porc. de Inercia		Porc. acumulado	
	eje 1	eje 2	eje 1	eje 2	eje 1	eje 2
2000	7.53	6.93	35.40	30.00	35.40	65.40
2001	7.67	6.88	36.90	29.60	36.90	66.40
2002	7.63	6.77	36.40	28.60	36.40	65.10
2003	7.56	6.74	35.80	28.40	35.80	64.20
2004	7.40	6.84	34.30	29.30	34.30	63.50
2005	7.44	7.06	34.60	31.12	34.60	65.80
2006	7.35	7.20	33.80	32.40	33.80	66.20
2007	7.30	7.07	33.30	31.30	33.30	64.60
2008	7.51	7.30	35.30	33.40	35.30	68.70

Tabla 4.6: Dos primeros valores propios e inercia que retienen para los ACP's individuales en los 9 años

Respecto a la variables, en términos generales, se pueden distinguir dos grupos que definen el análisis a lo largo del periódo.

Las variables IBS, ELC y EBS relacionadas con comercio exterior y electricidad, tienen alta calidad de representación en el eje 1 hasta el 2005, a partir de este año su representación sobre este eje cambia y en el 2008 se representan mejor sobre el eje 2.

Con las variables PIN, PIB y BCC relacionadas con el producto interno bruto, ocurre algo análogo, en los primeros años se explican mejor en el eje 2 y en el 2008 las variables PIB y BCC se explican en el eje 1 mientras que PIN continúa representada en el eje 2.

La variables PSE y EAG que en los primeros años se representan en el segundo eje, a partir del 2005 la variable PSE incrementa de manera importante la representación en el eje 1 mientras que la variable EAG se representa bien sobre el eje 1.

Las variables EIN y TIN tienen muy baja representación en el plano.

Para visualizar el análisis se ilustran las variables obtenidas en los diferentes ACP's, donde se puede ver la evolución de los dos grupos que influyen en la construcción de los ejes y las variaciones que estos generan.

En relación a los países, a manera de ejemplo, se presentan los valores para

Explicación	Año 2000		Año 2005		Año 2008	
Columnas	eje 1	eje 2	eje 1	eje 2	eje 1	eje 2
IBS	91.341	0.831	88.863	1.688	27.637	65.505
ELC	81.679	0.253	76.808	0.6	26.188	51.981
EBS	83.855	0.079	65.201	0.059	32.033	45.059
BCC	0.009	81.523	0.027	68.461	57.211	22.465
PIB	23.128	67.235	18.309	69.776	71.013	11.354
PIN	11.878	61.881	10.176	73.25	26.268	61.966
PSE	24.778	45.951	16.841	62.136	20.365	65.675
EAG	29.006	31.865	37.539	27.567	60.327	1.505
EIN	6.744	10.36	21.814	7.929	15.031	0.516
TIN	1.471	0.269	10.71	0.084	16.592	7.85

Tabla 4.7: Calidad de representación de las variables en los dos primeros ejes de los ACP's individuales, para los años 2000, 2005 y 2008

las variables Exportación de bienes en billones de dólares (EBS) y Porcentaje del producto interno bruto proveniente de la industria (PIN), que son dos variables que cambian su representación sobre los ejes factoriales a través del tiempo.

Para los años comprendidos entre el 2000 y el 2006, se preciben las diferencias gráficas sobre las variables TIN y EAG mientras que las mayores diferencias se notan en los años 2007 y 2008 donde los ángulos entre las variables y los ejes cambian de tamaño, mostrando el efecto de rotación de las variables sobre los ejes definidos por el respectivo ACP, en comparación con los primeros años. Los biplots para cada uno de los años incluyendo los países se pueden ver en el apéndice B.

La variable EBS depende del tamaño del cada país, sin embargo se observa que el incremento entre el 2000 y el 2005 en promedio fue del 56 % respecto al 2000. El incremento del 2005 al 2008 fue del 41 % respecto al 2005. Esto significa que el incremento en todo el periodo 2000-2008 en promedio fue del 1,2 %, los paises con menor incremento son Irlanda, Canadá, Reino Unido, Japón y USA.

Para la variable Porcentaje del producto interno bruto proveniente de la industria (PIN), la variación del 2000 al 2005 respecto del 2005 fue del -6,2 % y entre el 2005 y el 2008 respecto del 2008 fué del -0,3 %. Para el periódo total fue del -6,4 % respecto del 2000.

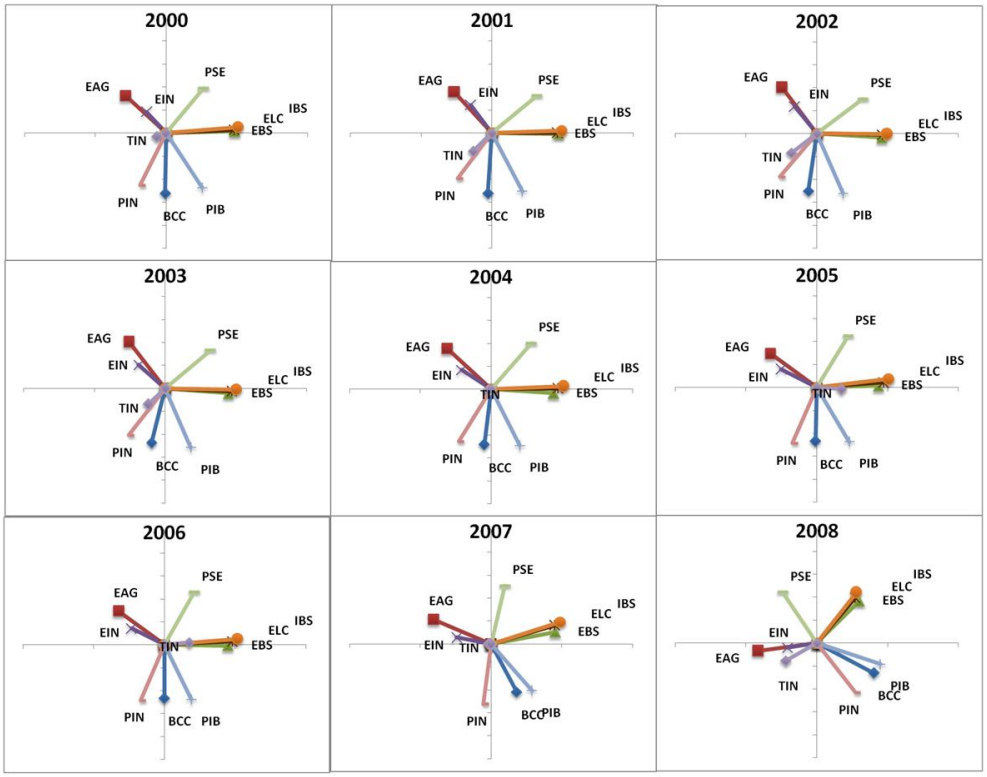

Figura 4.9: Representación de las variables en el plano 1-2, para los ACP's individuales

Irlanda, Bélgica, Portugal, y Rusia son los países que presentan mayor tendencia decreciente en esta variable.
Por otra parte, Noruega, Australia y Países Bajos, que habían disminuido su porcentaje en el 2005, presentan incremento para la variable en el 2008.

Reiteramos que esta aplicación, aunque es con datos reales, se hace con el fin de mostrar y comparar los métodos estadísticos propuestos y esta lejos de ser un análisis económico de los países que intervienen o de los años que se analizan.

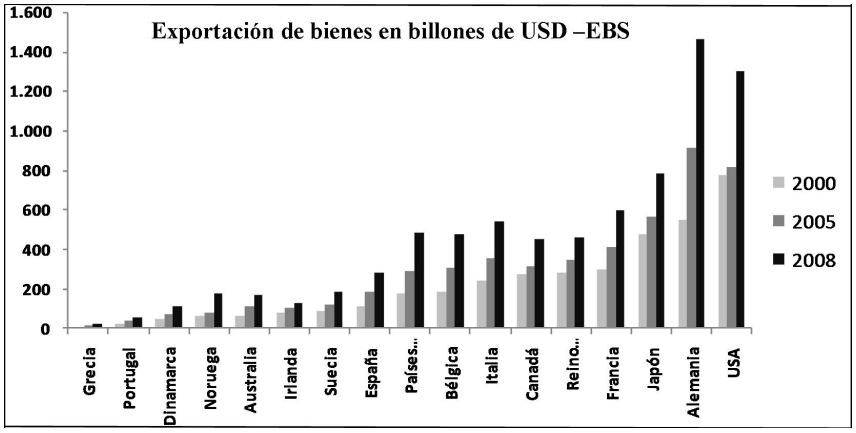

Figura 4.10: Variable EBS para los años 2000, 2005 y 2008 por país

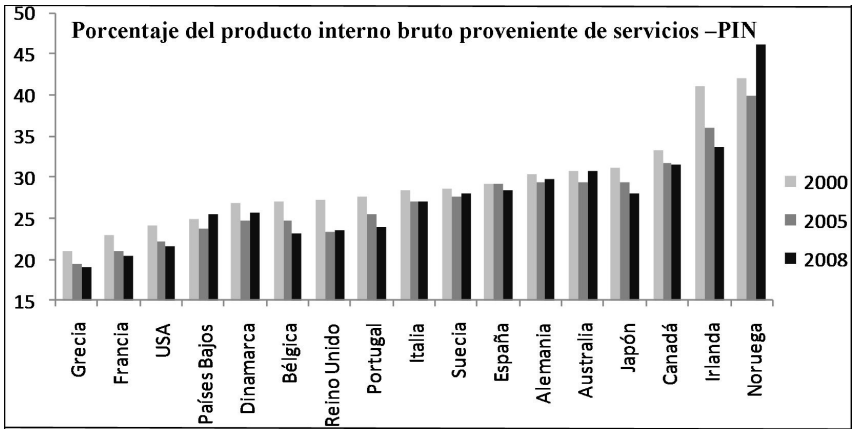

Figura 4.11: Variable PIN para los años 2000, 2005 y 2008 por país

4.7. Criterios

Se presenta el análisis para dos criterios con el fin de evidenciar analíticamente las diferencias y similitudes al analizar la intraestructura de las tablas bajo estudio.

4.7.1. Criterio inducido por STATIS Dual-CISD

Este criterio consigue la representación de las tablas con un índice $\xi(.,\nu)$= 1,28 %, que mide el porcentaje de inercia que pierden en promedio las K nubes de individuos, cuando se proyectan sobre el subespacio consenso.

Año	% inercia consenso	% inercia tabla	% de pérdida
2000	65,15	65,40	0,4 %
2001	65,64	66,40	1,1 %
2002	64,06	65,10	**1,6 %**
2003	63,78	64,20	0,7 %
2004	63,45	63,50	0,1 %
2005	64,48	65,80	**2,0 %**
2006	64,73	66,20	**2,2 %**
2007	64,19	64,60	0,6 %
2008	66,76	68,70	**2,8 %**

Tabla 4.8: Inercia explicada por el consenso CISD, el ACP individual y pérdida de inercia sobre el consenso

Las tablas que presentan mayor pérdida de inercia al ser proyectadas sobre este espacio consenso son las correspondientes a los años 2002, 2005, 2006 y 2008.

Con respecto a los ángulos entre los ejes factoriales de las tablas individuales y las del espacio consenso, se presenta la tabla con el calculo de los cosenos entre los respectivos ejes.

Encontramos que a partir del 2005 el primer eje cambia de sentido, como indica el signo negativo y el ángulo disminuye como lo indica el coseno.
En el 2007 los dos ejes cambian de sentido.
En el 2008 además que los ejes cambian de sentido, los ángulos disminuyen de forma importante lo que implica que los ejes están rotados. Hecho que se verifica al calcular el coseno entre el ángulo del eje 1 del espacio consenso y el eje 2 de la tabla cuyo valor es 0.80 (37 grados); y el coseno del ángulo entre el eje 2 del espacio consenso y el eje 1 de la tabla que es 0.796 (37,2 grados).

El indice que mide la similitud entre los factores del espacio consenso y los factores de toda las tablas, en rango reducido 2 es $\eta(\mathbf{V}_2)$ =0.981. Muestra

Coseno de Ángulos

Año	Consenso Eje 1 año Eje 1	Consenso Eje 2 año Eje 2	Consenso Eje 3 año Eje 3
2000	0,996	-0,998	0,995
2001	-0,993	0,984	-0,971
2002	0,985	-0,977	-0,965
2003	0,991	0,986	0,893
2004	0,998	0,999	0,997
2005	-0,969	0,996	-0,986
2006	-0,971	0,995	0,964
2007	-0,942	-0,941	0,976
2008	-0,553	-0,588	0,958

Tabla 4.9: Coseno de los ángulos entre los 3 primeros ejes factoriales de cada tabla y los respectivos ejes del espacio consenso CISD

que en promedio el ángulo entre los dos primeros ejes factoriales del espacio consenso y los dos primeros de las tablas individuales es 7.92 grados.

Los índices $\xi(.,\nu)$ de pérdida de inercia y $\eta(\mathbf{V}_2)$, muestran que el espacio consenso explica buena parte de la inercia de las tablas y los ángulos entre los ejes factoriales individuales y el espacio consenso diferen poco; indicando un buen ajuste del espacio consenso respecto a las tablas bajo análisis.

4.7.1.1. Espacio Consenso CISD

En este análisis, el primer eje corresponde al 34,3 % de explicación, el eje 2 al 30.4 % y el eje 3 al 15.5 %. Por tanto el primer plano factorial explica el 64,7 % de la intraestructura de las tablas.

Las variables EBS, ELC e IBS que dependen del tamaño del país y están relacionadas con *comercio exterior y electricidad* se encuentran explicadas en el eje 1, mientras que BCC, PIB y PIN que definen la *situación interna de un país* se encuentran explicadas por el eje 2.

La variable EAG tiene el mismo porcentaje de explicación en los dos ejes, mientras que PSE está mejor representada en el eje 2 (57,8 %).

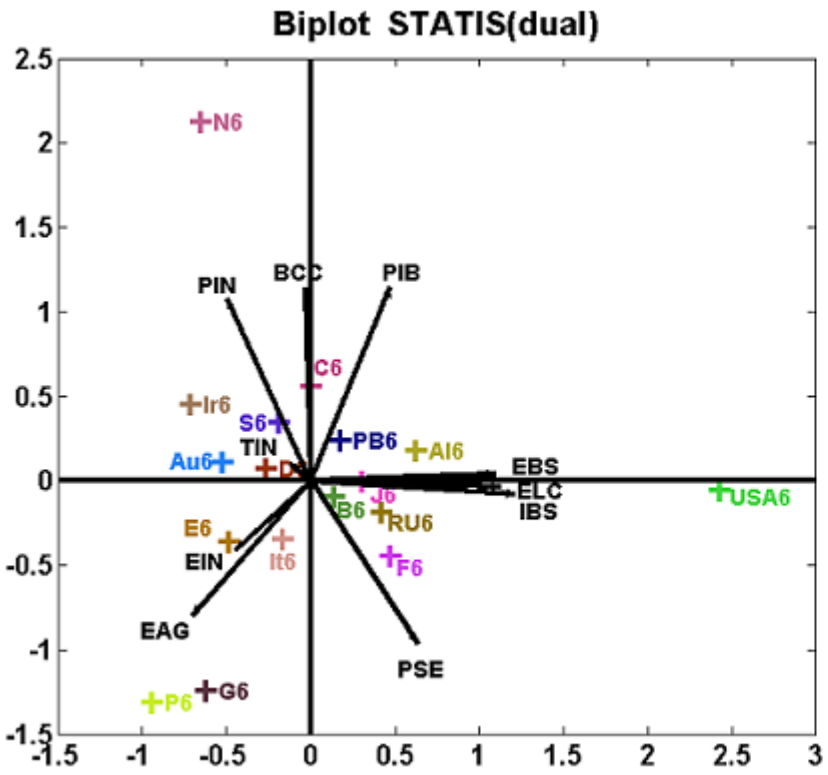

Figura 4.12: Representación consenso CISD en el plano 1-2

Para el análisis inicial se identifica cada país con las coordenadas del año 2005. El primer eje *comercio exterior y electricidad* ordena los países de mayor a menor (derecha a izquierda), siendo USA y Alemania los que conservan los mayores valores, mientras que Grecia y Portugal presentan los menores valores.

En el segundo eje la *situación interna de un país* caracteriza a Noruega que tiene los mayores valores para las tres variables: BCC, PIB y PIN.

Con relación a las variables, a excepción de TIN y EIN, las demás están bien explicadas en el primer plano factorial.

Las variables Porcentaje del producto interno bruto proveniente de la industria-PIN y Porcentaje del producto interno bruto proveniente de servicios-PSE, tienen fuerte correlación negativa (-0,98). Lo que implica que Noruega e

Irlanda presentan los menores porcentajes del producto interno bruto proveniente de servicios, mientras que Grecia, Portugal y Francia tienen los mayores porcentajes en esta variable.

Grecia y Portugal son los países con mayor Porcentaje de población empleada en agricultura-EAG.

Con este análisis se obtiene información suficiente para conocer la relación entre casi todas las variables y los países que se estudian. Sin embargo, se conoce que el tercer eje conserva el 15.5 % de la inercia total, porcentaje poco despreciable en comparación con los primeros ejes. Además no se tiene análisis de las variables EIN y TIN; por esta razón analizaremos el plano 1-3.

La información representada en este plano1-3 es del 49.8 %. Las variables EBS, ELC e IBS tienen mayor explicación en este plano que en el 1-2 y las variables Porcentaje de la población empleada la industria-EIN y Tasa de interés a largo plazo-TIN se explican en el eje 3 con porcenajes de 52.8 % y 59.9 % respectivamente.

La variable EIN caracteriza principalmente a España, Alemania e Italia con los mayores valores, mientras que la variable TIN posiciona a Japon como el país con menor tasa de interes a largo plazo en el trancurso de los años analizados.

Entonces, con el análisis en las 3 primeras dimensiones del espacio consenso se encuentran las caracterizaciones de los países por las variables propuestas.

4.7.1.2. Trayectoria de los individuos

En general las trayectorias de los países tienen buena representación en el plano 1-2. Los países con menor representación son Japon, Dinamarca, Bélgica, y Países Bajos
Los países con mayor dispersión a lo largo del eje 1 son Alemania e Irlanda. Y los que menos varían son Francia, Italia y Países bajos.

En el Eje 2 los países con mayor dispersión son Irlanda, Noruega y Bélgica, y los que varían menos son Portugal, Japon y España.

Para conocer las variaciones en la trayectorias es necesario volver a la matriz de datos originales donde se encuentra que la trayectoria de Noruega esta definida por las variables BCC, PIN y PIB que para los años 2006, 2007 y

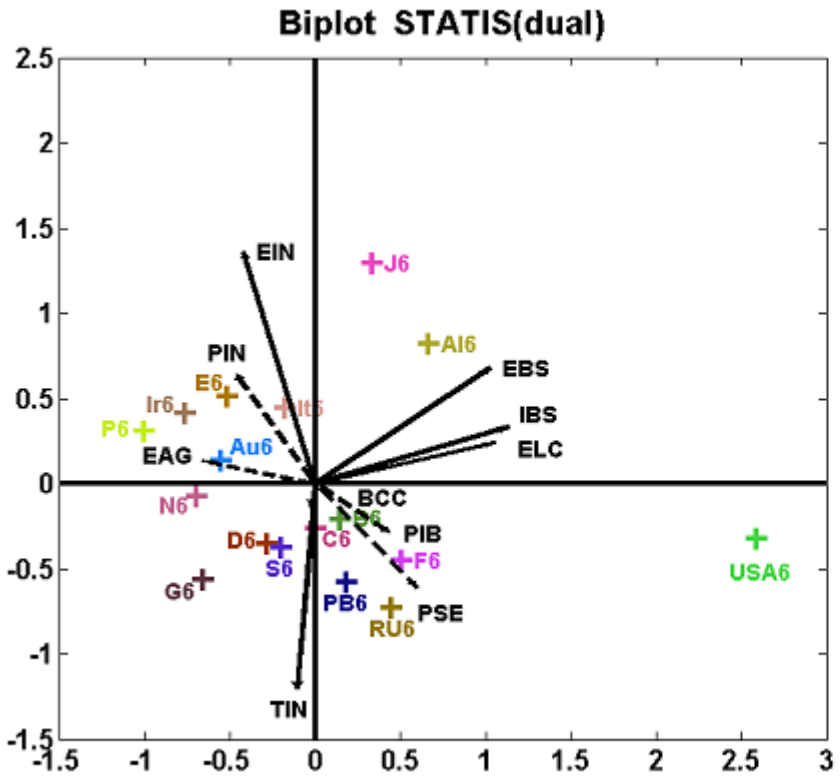

Figura 4.13: Representación consenso CISD en el plano 1-3

2008 presenta los mayores valores.

Irlanda a lo largo del periódo y para todos los años ha incrementado el PIB, sin embargo a partir del 2004 el Porcentaje del producto interno bruto proveniente de la industria-PIN disminuyó y desde el 2006 el Balance de la cuenta corriente como porcentaje del PIB (BCC) disminuyó notablemente.

En la trayectoria para Alemania se destaca por los altos porcentajes para la variable EBS y el año 2008, con el mejor IBS y BCC.

4.7.2. Criterio inducido por Meta Componentes-CIMC

Este criterio presenta un índice $\xi(.,\nu)$= 1,29 % igual al generado por el criterio inducido por STATIS Dual (CISD) y con perfil de pérdida similar

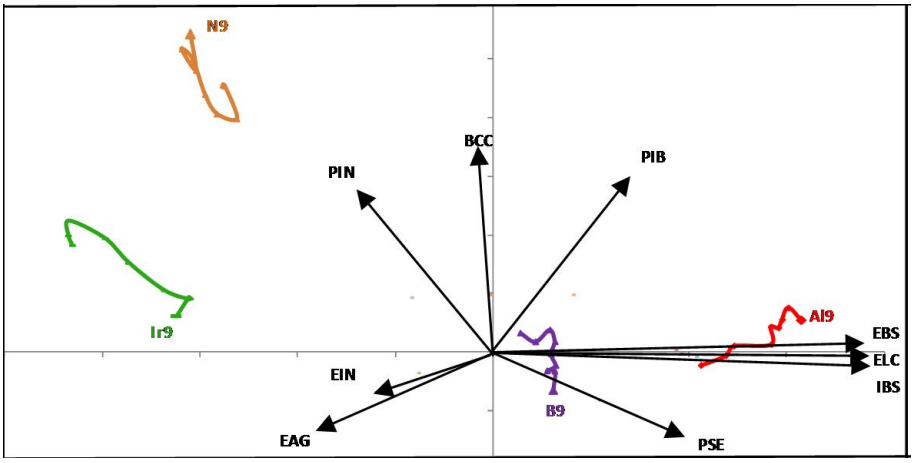

Figura 4.14: Trayectoria de Noruega, Irlanda, Belgica y Alemania con el CISD

entre los respectivos años.

Año	% inercia consenso	% inercia tabla	% de pérdida
2000	65,14	65,40	0,4 %
2001	65,69	66,40	1,1 %
2002	64,09	65,10	**1,5 %**
2003	63,80	64,20	0,6 %
2004	63,45	63,50	0,1 %
2005	64,54	65,80	**1,9 %**
2006	64,75	66,20	**2,2 %**
2007	64,15	64,60	0,7 %
2008	66,59	68,70	**3,1 %**

Tabla 4.10: Inercia explicada por el CIMC, el ACP individual y pérdida de inercia sobre el consenso

Para el análisis de los ángulos entre los ejes factoriales, se presenta la tabla con el calculo de los cosenos entre los ejes invertidos de las tablas vs los del espacio consenso. En este espacio consenso las tablas son representadas mediante los ejes rotados, a excepcíon del año 2008 que tiene coseno del ángulo entre el eje 1 del espacio consenso y el eje 1 de la tabla, con valor de 0,813 (31 grados); y el coseno del ángulo entre el eje 2 del espacio consenso y el eje 2 de la tabla es 0.872 (29 grados).

Coseno de ángulos

Año	Consenso Eje 1 año Eje 2	Cosenso Eje 2 año Eje 1
2000	0,985	0,983
2001	-0,986	-0,991
2002	0,987	-0,990
2003	-0,994	0,996
2004	-0,993	0,992
2005	-0,975	-0,951
2006	-0,976	-0,954
2007	0,887	-0,890
2008	0,475	-0,439

Tabla 4.11: Coseno de los ángulos entre los 2 primeros ejes factoriales de cada tabla y los 2 ejes del espacio consenso CIMC

El indice $\eta(\mathbf{V}_2)$ =0.910 en rango reducido 2, muestra que el ángulo promedio entre los dos primeros ejes factoriales del espacio consenso y los dos primeros de las tablas individuales es 17,45 grados.

Los índices $\xi(.,\nu)$ y $\eta(\mathbf{V}_2)$, muestran que este espacio consenso, al igual que en el criterio anterior, hay buen ajuste del espacio consenso respecto a las tablas bajo análisis.

4.7.2.1. Espacio Consenso

En este análisis, el primer eje corresponde al 49,34 % de explicación y el eje 2 al 48,82 %. Lo que implica que el primer plano factorial explica el 98,17 % de la intraestructura de las tablas
En el plano factorial 1-2 las variables están representadas en el 99.9 % a excepción de TIN que se explica en un 85.9 %.

En el eje 1, las variables BCC y PIB que dependen del tamaño del país, están representadas con porcentajes del 97 % y 94 %, respectivamente. La representación de las variables EAG, EIN y PIN es menor que las anteriores pero sigue siendo importane sobre este eje, 69,5 %, 60 % y 71 %.

En el eje 2 las variables IBS, ELC y EBS presentan la mayor correlación, por tanto su repesentatividad es alta, con porcentajes del 99,5 %, 99 % y 97 %.

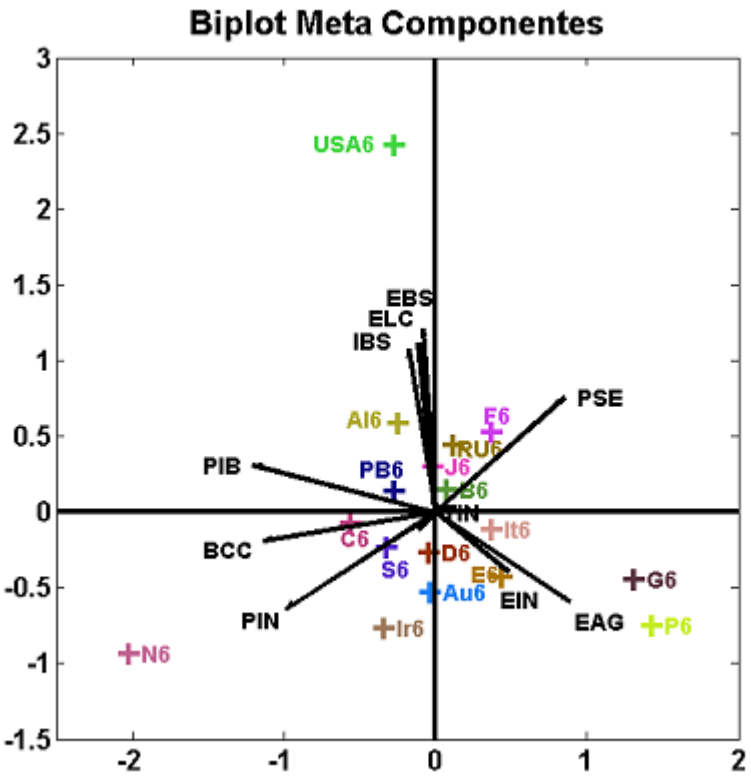

Figura 4.15: Representación consenso CIMC en el plano 1-2

Las variables PSE y TIN tienen porcentajes de explicación similar en los dos ejes.

La diferencia de representación de las variables en los dos primeros ejes del consenso, entre el CISD y el CIMC, se debe a la forma como captan las correlaciones de las variables.

En el cuadro 4.12 se encuentra la tabla de correlación entre variables, distribuidas de acuerdo a la representación en los ejes del consenso. Los valores resaltados con negrita corresponden al conjunto de variables que el CISD representa bien y los valores entre doble linea, el conjunto que el CIMC representa mejor.

Se observa que CISD representa bien en los ejes 1-2 las variables con correlación positiva y con valores próximos. Por tanto la variable EIN que solo

Repres.	Eje consenso					Eje consenso		
Var	BCC	PIB	PIN	EAG	EIN	IBS	ELC	EBS
BCC		**0,52**	**0,54**	-0,55	-0,30	-0,14	-0,18	0,06
PIB	**0,52**		**0,41**	-0,56	-0,52	0,35	0,31	0,31
PIN	**0,54**	**0,41**		-0,13	0,20	-0,30	-0,24	-0,18
EAG	-0,55	-0,56	-0,13		0,33	-0,41	-0,28	-0,47
EIN	-0,30	-0,52	0,20	0,33		-0,18	-0,24	-0,04
IBS	-0,14	0,35	-0,30	-0,41	-0,18		**0,89**	**0,90**
ELC	-0,18	0,31	-0,24	-0,28	-0,24	**0,89**		**0,67**
EBS	0,06	0,31	-0,18	-0,47	-0,04	**0,90**	**0,67**	
PSE	**-0,45**	-0,30	**-0,98**	-0,03	-0,23	**0,39**	**0,30**	0,28
TIN	-0,07	-0,10	0,05	0,06	-0,13	-0,16	-0,10	-0,29

Tabla 4.12: Matriz de correlación entre variables

está correlada significativamente con PIB no tiene buena calidad de representación. La variable TIN que no presenta correlación significativa con el grupo de variables tampoco la representa en este plano.

El CIMC toma las variables cuya correlación en valor absoluto es próxima y es menos restrictivo, permitiendo representaciones de variables que tienen correlación alta con algunas variables del grupo y no con todas, como es el caso de las variables EAG e EIN. La variable TIN se encuentra bien representada en el plano aunque no tiene correlación significativa con las demás variables bajo estudio.

De manera análoga que para el CISD, cada país se identifica con las coordenadas del año 2005

En el primer eje las variables BCC, PIB y PIN caracteriza a Noruega. La variable EAG caracteriza a todos los años de Grecia, Portugal y los primeros 4 años de España, mientras que la variable EIN caracteriza los primeros 5 años de Portugal, los primeros 6 de Italia y los primeros 3 de España.

El segundo eje ordena los países de mayor a menor (derecha a izquierda), de la misma forma que lo hace el CISD.

Al igual que el CISD, este espacio capta la correlación fuerte y negativa (-0,98) entre las variables Porcentaje del producto interno bruto proveniente de la industria-PIN y Porcentaje del producto interno bruto proveniente de servicios-PSE. Lo que implica que Noruega presenta el menor porcentaje

del producto interno bruto proveniente de servicios, mientras que Grecia, Portugal y Francia tienen los mayores porcentajes.

La variable TIN aunque tiene buena representación en el plano, gráficamente no es posible asignarle países que caracterice.

De los resultados de este análisis se pueden concluir muchas de las relaciones variables países, sin embargo queda pendiente la caracterización de la variable TIN.

Entonces vamos a trabajar con este espacio en 3 dimensiones y analizamos los cambios que se dan.

Como se vió en 3.3.2.6 página 114, la descomposición espectral de la matriz $\mathbb{H}$ que genera las dimensiones del espacio consenso es

$$\mathbb{H} = \sum_{k=1}^{k} \mathbf{V}_k \mathbf{V}_k^{'}$$

Al variar la dimension de los espacios $\mathbf{V}_k$ de 2 a 3 dimensiones, cambia la matriz $\mathbb{H}$, su descomposición espectral

$$\mathbb{H} = \mathbf{V}_H \mathbf{\Lambda}_H^2 \mathbf{V'}_H$$

el modelo que genera

$$\mathbf{X} = \hat{\mathbf{A}}_H \mathbf{V}_H + \mathbf{E}_H$$

la representación biplot

$$\mathbf{X} = \mathbf{A}_H \mathbf{V'}_H$$

y por ende los indicadores.

En este espacio de dimensión 3, el primer eje explica el 33,19 % , el segundo 32,91 % y el tercero 32 % de la información para un total del 98,49 % de explicación. El plano 1-2 explica el 66 % y el plano 1-3 el 65,6 %. Se revisa el análisis en el primer plano factorial con el fin de conocer los cambios que genera la variación de dimensión.

La variable EIN bien representada en dimesión reducida 2, se representa con mayor claridad en el eje 3 de la dimensión reducida 3. La variable TIN con buena representación en el plano 1-2 en dimensión reducida 2 adquiere importancia sobre el eje 3 en dimensión reducida 3.
Graficamente no hay variaciones en el plano 1-2. Se presenta el plano 1-3 para analizar las variables que se encuentran bien representadas

Var	Espacio reducido dimensión 2		Espacio reducido dimensión 3		
BCC	**97,43**	2,56	**92,77**	4,79	2,31
EAG	**69,46**	30,52	**64,07**	19,81	16,10
EBS	2,68	**97,31**	3,86	**91,79**	4,33
EIN	**60,03**	39,91	3,26	1,35	**95,37**
ELC	0,97	**99,02**	1,48	**95,70**	2,79
IBS	0,41	**99,58**	0,83	**98,33**	0,82
PIB	**93,81**	6,18	**80,91**	0,97	18,10
PIN	**71,03**	28,96	**63,18**	7,49	29,29
PSE	56,51	43,47	52,31	14,17	33,47
TIN	41,25	44,69	0,13	21,38	**78,37**

Tabla 4.13: Calidad de representación de las variables en espacio reducido 2 y 3 -CISD

La variable EIN caracteriza principalmente a España, Alemania, e Italia con mayores valores y USA que presenta valores bajos para los últimos 6 años. La variable TIN posiciona a Japon como el país con menor tasa de interes a largo plazo en el trancurso de los años analizados y USA que presenta el menor valor para el año 2009.

De los anteriores análisis se puede observar que el CIMC en cualquier dimensión va a obtener una alta explicación de la información. En dimensión reducida 2, se obtiene buena representación de las variables e individuos en general. Si la correlación de las variables no es significativa y la dimensión de representación es baja, el método reagrupa las variables que tengan alguna correlación enmascarando la posibilidad de análisis en una dimensión más alta.
El CIMC y el CISD en dimensión reducida 3, generan los mismos resultados.

4.7.2.2. Trayectoria de los individuos

Este análisis se realiza en el plano 1-2 en dimensión reducida 3, que explica más las variables y se permite hacer la comparación con las trayectorias obtenidas para el CISD.

Los países que presentan baja calidad de representación son: Belgica, Italia y Países Bajos mientras que Grecia, Noruega, Portugal y USA presentan la

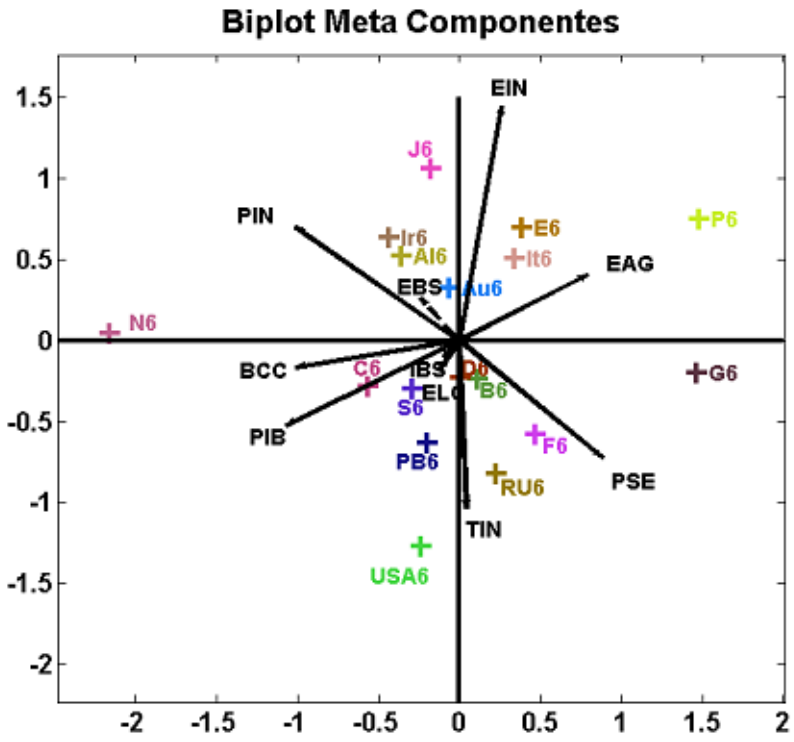

Figura 4.16: Representación consenso CIMC en el plano 1-3

mayor representación
Los países con mayor dispersión a lo largo del eje 1 son USA, Irlanda y Alemania y los que menos varían son Francia, Italia y Países Bajos.

En el Eje 2 los países con mayor dispersión son USA, Irlanda y Noruega y los que varían menos son Francia, Belgica y Alemania.
Para comparar los dos criterios se presentan los gráficos de las trayectorias de los mismos países que para el CISD.
Graficamente no se perciben diferencias en las trayectorias y el análisis es el mismo que para las trayectorias de CISD.

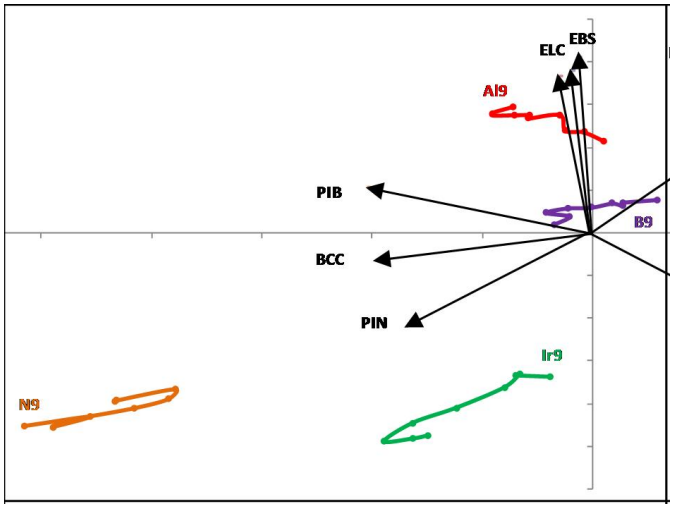

Figura 4.17: Trayectoria para Noruega, Irlanda, Bélgica y Alemania con el CIMC

4.7.3. Índices de Comparación

A lo largo del trabajo hemos visto que algunos métodos proponen índices para conocer la similutud entre las tablas, para conocer la calidad de representación en un espacio determinado y para conocer el desempeño entre los espacios consenso. En esta sección, a modo de resumen los retomamos y mostramos como cada uno aporta al análisis del espacio consenso de las tablas que estamos analizando.

4.7.3.1. Inercia retenida en los ACPs individuales

Un primer resultado que a la vista nos permite conocer la influencia de las tablas en el análisis de la intraestructura es el porcentaje de inercia que retiene cada ACP individual, en cada uno de sus ejes.

Se puede notar que las tablas poseen estructura similar y no hay tablas con fuerte estructura en términos de la inercia retenida en cada uno de los ejes para los ACP individuales. Esto hace pensar que los criterios que buscan equilibrar la influencia de las tablas, ponderándolas por algun valor o normando las matrices **S**, no tendrán un efecto importante en comparación con los criterios que no lo hacen.

Año	Eje 1	Acum Eje 2	Acum Eje 3
2000	35,40	65,40	81,20
2001	36,90	66,40	81,90
2002	36,40	65,10	81,20
2003	35,80	64,20	80,90
2004	34,30	63,50	80,50
2005	34,60	65,80	80,30
2006	33,80	66,20	80,30
2007	33,30	64,60	79,50
2008	35,30	68,70	82,10

Tabla 4.14: Inercia retenida en los tres primeros ejes de los ACP individuales

Como se observó en la sección precedente, a partir del 2005 la inercia del primer eje disminuye y aumenta en el segundo eje. Además, la tabla correspondiente al 2008 es la que acumula mas inercia en los dos y tres primeros ejes.

4.7.3.2. Coeficiente *RV*

Un indicador de gran importancia en el desarrollo del AFM y que permite comparar las tablas dos a dos, es el coeficiente *RV* (ver 2.2.3 página 48) propuesto por Escoufier (1973), quien inicialmente lo definió como medida de asoción entre dos vectores aleatorios pertenecientes al mismo espacio de probabilidad. Se presenta el calculo del coeficiente RV sin ponderar las matrices por el primer valor propio, con el fin de analizar la relacion de las tablas y su posible influencia en la construcción del espacio consenso.
Se observan dos grupos de tablas similares, las tablas correspondientes a los años 2001, 2002 y 2003, y las tablas del 2005 y 2006. Las tablas correspondientes al 2006, 2007 y 2008 son las más disímiles al conjunto total. El año 2004 mantiene similaridad con todas las tablas, lo que explica que en el método "selección de la mejor tabla", este año sea seleccionado como espacio consenso.

	2000	2001	2002	2003	2004	2005	2006	2007	2008
2000	1								
2001	0,99	1							
2002	0,98	**1**	1						
2003	0,99	**0,99**	**1**	1					
2004	0,99	0,98	0,97	0,99	1				
2005	0,96	0,93	0,92	0,94	0,98	1			
2006	0,95	0,91	0,90	0,93	0,97	**1**	1		
2007	0,97	0,95	0,94	0,96	0,98	0,98	0,98	1	
2008	0,94	0,92	0,92	0,92	0,94	0,90	0,92	0,97	1

Tabla 4.15: Coeficiente *RV* para las tablas bajo análisis

4.7.3.3. Inercia retenida por los ejes que generan los diferentes espacios consenso

Respecto a los espacio consenso, la primera información que se obtiene es el porcentaje de inercia retenido en cada uno de los ejes que generan el espacio y permite hacer una comparación entre los diferentes criterios.

	Mejor tabla	Maxima Varianza	Statis Dual	Meta Comp.
Eje 1	34,286	34,291	34,299	33,191
Eje 2	29,256	30,401	30,399	32,906
Subtotal	63,542	64,692	64,698	66,097
Eje 3	16,925	15,462	15,46	32,394
Total	80,467	80,154	80,158	98,491

Tabla 4.16: Inercia retenida en los 3 primeros ejes que generan los espacios consenso inducidos por los diferentes criterios No normados

Los espacios que generan menor diferencia son el inducido por la mejor tabla (CIMT) y el inducido por el Meta Componentes (CIMC). El CIMT retiene menos inercia el primer plano, sin embargo en dimensión reducida tres es similar a los otros criterios. El CIMC es el que retiene mas varianza debido a que se contruye directamente sobre las 3 dimensiones. Sin embargo, cuando se realiza el análisis en dimensión reducida 2, (ver 4.7.2.1 página 140) se obtiene un nivel de explicación del 98,17 % y es el espacio que mas explica pero a su vez desconoceríamos que es el que más pierde en la tercera dimensión. Se presentan los resultados con 3 decimales para que mostrar que las dife-

	M Var Estan	M Var/λ_1	M Var/$(\lambda_1 + \lambda_2)$	Statis D Norm
Eje 1	34,285	34,253	34,3	34,291
Eje 2	30,401	30,431	30,373	30,391
Subtotal	64,686	64,684	64,673	64,682
Eje 3	15,462	15,45	15,478	15,467
Total	80,148	80,134	80,151	80,149

Tabla 4.17: Inercia retenida en los 3 primeros ejes que generan los espacios consenso inducidos por los diferentes criterios Normados

rencias, a excepción del CIMC, son muy pequeñas y se dan a nivel de las décimas.
Respecto a los criterios que norman o estandarizan la supermatriz $\mathbf{X}$, en general las diferencias son mucho más pequeñas y se dan al nivel de las centésimas. Esto implica que para los datos que estamos analizando no es necesario normar o estandarizar las tablas porque ninguna influye de manera preponderante en la construcción del espacio consenso.

4.7.3.4. Índices $\xi(.,\nu)$ y $\eta(\mathbf{V}_r)$

Los índices propuestos en 3.3.3, página 115, permiten un análisis completo para comparar el nivel de ajuste de los diferentes espacios consenso con las tablas que se analizan.

El índice $\xi(.,\nu)$ que mide la inercia promedio que pierden las tablas al ser proyectadas sobre el respectivo espacio consenso, indica que la representación biplot de las tablas se expresa con mayor precisión por el Criterio de Máxima Varianza, el Statis Dual o para alguno de los criterios que norman la super matriz.
Con el índice $\eta(\mathbf{V}_r)$ que mide el ángulo promedio entre los ejes del espacio consenso y los ejes de las tablas, se obtiene resultados similares que con el índicador $\xi(.,\nu)$.

Se debe resaltar que la información para el criterio inducido por el Meta Componentes para la dimensión 2, es un "sub-índice" extraído de la representación en 3 dimensiones. Realizando el mismo análisis para este criterio en dimension reducida 2, se tiene que:

Índice	$\xi(.,\nu)$		$\eta(\mathbf{V}_r)$		Ángulo promedio en°	
Criterio	Dim 2	Dim 3	Dim 2	Dim 3	Dim 2	Dim 3
Mejor tabla	1,43	1,01	0,979	0,982	8,33	7,71
Máx Varianza	1,28	0,89	0,981	0,984	7,92	7,27
Statis Dual	1,28	0,89	0,981	0,984	7,92	7,27
Meta Comp.	5,42	0,91	0,910	0,984	17,45	7,27
M Var Estan	1,28	0,89	0,981	0,984	7,92	7,27
M Var/λ_1	1,29	0,89	0,981	0,984	7,92	7,27
M Var/$(\lambda_1+\lambda_2)$	1,28	0,89	0,981	0,984	7,92	7,27
Statis D Norm	1,28	0,89	0,981	0,984	7,92	7,27

Tabla 4.18: Índices de ajuste de los espacios consenso inducidos por los diferentes criterios en dimensión reducida 2 y 3

Índice	$\xi(.,\nu)$	$\eta(\mathbf{V}_r)$	Ángulo promedio en°
Meta Comp.	1,3	0,981	7,92

Tabla 4.19: Índices de ajuste para el espacio consenso inducido por el criterio Meta Componentes en dimensión reducida 2

Estoas índices igualan a los obtenidos para el espacio consenso inducido por los criterios Máxima varianza y Statis Dual.

Capítulo 5

Conclusiones

1. Es posible describir el biplot clásico en términos de la proyección de subespacios de dimensión reducida, así como en términos de regresiones alternadas, demostrando que ambas aproximaciones son equivalentes. La utilización conjunta de ambas aproximaciones permite enriquecer la interpretación de los índices de bondad de ajuste clásicos para cada individuo y cada variable.

2. Cuando se dispone de varias matrices de datos, bien sea un único conjunto de individuos sobre los que se miden varios conjuntos de variables de diferente naturaleza, varios conjuntos de individuos sobre los que se miden las mismas variables, o el caso especial donde se miden las mismas variables para un único conjunto de individuos en momentos diferentes de tiempo; el objetivo principal consiste en encontrar una representación que permita interpretar el conjunto completo de tablas en un *Espacio Consenso*.

3. Para el estudio simultáneo de tablas se consideran tres aspectos principales:

   La *Inter-estructura* ó estudio de las similitudes y diferencias entre grupos.

   La *Intra-estructura de las variables*, ó estudio de las similitudes en la estructura de covariación o correlación entre las variables.

   La *Intra-estructura de los individuos*, ó estudio de las similitudes entre individuos y su modificación a través de las diferentes tablas.

En este trabajo se ha tratado especialemente el estudio de la *Intra-estructura de las variables*, donde se introduce un punto de vista original con respecto a la literatura revisada.

4. En la literatura se han encontrado diversos métodos para la obtención del *Espacio Consenso* pero ninguno de ellos, salvo el Meta-biplot, presenta los resultados mediante la forma biplot.

5. A partir de la proyección sobre subespacios de dimensión reducida proponemos la obtención de un subespacio de referencia común para todas las tablas y estudiamos sus propiedades generalizando a esta situación los índices de bondad de ajuste de los biplots clásicos.

6. Se proponen índices para la comparación de distintos subespacios consenso, basados en los ángulos entre los subespacios ó la cantidad de inercia retenida por cada uno de ellos.

7. Muchas de las técnicas clásicas para el estudio de tablas múltiples pueden desarrollarse dentro de la propuesta adaptando su formulación original. Entre dichas técnicas podemos mencionar: AFM Dual, Statis, DUAL, variaciones del DACP, Meta- componentes, entre otros.

8. En la aplicación a datos reales se puede comprobar la potencia del método desarrollado al igual que el desempeño de los índices propuestos.

9. En la aproximación biplot propuesta, basada en el concepto de regresión se aplica la norma euclídea o de Frobenius. La elección de esta norma se justifica desde el punto de vista geométrico, y estadístico; pero sobre todo porque esta norma es diferenciable, lo que implica que el proceso de mínimos cuadrados conduce a las llamadas ecuaciones normales. Entonces pensar en conocer el cambio que produce la selección de otra norma no es un camino sencillo de abordar.

Índice de tablas

4.1. Variables bajo análisis. 121

4.2. Sigla y nombre de los países que intervienen en el análisis . . 122

4.3. Promedio y desviación típica de las variables para cada año y total 124

4.4. Anovas individuales 128

4.5. Valores propios e inercia de la interestructura determinada por el análisis canónico de varianza 128

4.6. Dos primeros valores propios e inercia que retienen para los ACP's individuales en los 9 años 130

4.7. Calidad de representación de las variables en los dos primeros ejes de los ACP's individuales, para los años 2000, 2005 y 2008 131

4.8. Inercia explicada por el consenso CISD, el ACP individual y pérdida de inercia sobre el consenso 134

4.9. Coseno de los ángulos entre los 3 primeros ejes factoriales de cada tabla y los respectivos ejes del espacio consenso CISD . 135

4.10. Inercia explicada por el CIMC, el ACP individual y pérdida de inercia sobre el consenso 139

4.11. Coseno de los ángulos entre los 2 primeros ejes factoriales de cada tabla y los 2 ejes del espacio consenso CIMC 140

4.12. Matriz de correlación entre variables 142

4.13. Calidad de representación de las variables en espacio reducido 2 y 3 -CISD 144

4.14. Inercia retenida en los tres primeros ejes de los ACP individuales 147

4.15. Coeficiente *RV* para las tablas bajo análisis 148

4.16. Inercia retenida en los 3 primeros ejes que generan los espacios consenso inducidos por los diferentes criterios No normados . 148

4.17. Inercia retenida en los 3 primeros ejes que generan los espacios consenso inducidos por los diferentes criterios Normados . . . 149

4.18. Índices de ajuste de los espacios consenso inducidos por los diferentes criterios en dimensión reducida 2 y 3 150

4.19. Índices de ajuste para el espacio consenso inducido por el criterio Meta Componentes en dimensión reducida 2 150

Índice de figuras

2.1. Esquema de los datos múltiples 34

2.2. Representación euclidea de las matrices iniciales 54

2.3. Presencia de estructura comun entre las tablas 55

2.4. El objeto $\mathbf{W}_1$ no está bien representado por el consenso . . . 55

2.5. El consenso no es buen resumen de los objetos debido a que sus normas son muy diferentes 56

2.6. El consenso no es buen resumen de los objetos debido a que las tablas no presentan estructura común 56

2.7. Tablas de datos centradas y superpuestas 73

2.8. posiciones compromiso de los individuos en Statis 89

2.9. posiciones compromiso de los individuos en el AFM 90

4.1. Esquema de los datos para los 9 años 123

4.2. Tendencia de las variables de producción 124

4.3. Tendencia de las variables de productividad 125

4.4. Tasa de interés a largo plazo 125

4.5. Tendencia de las variables de importación y exportación de bienes y servicios 126

4.6. Balance de la cuenta corriente como porcentaje del PIB . . . 126

4.7. Generación de electricidad en terawatios por hora 127

4.8. Representación de las tablas en el plano 1-2 (interestructura) 129

4.9. Representación de las variables en el plano 1-2, para los ACP's individuales 132

4.10. Variable EBS para los años 2000, 2005 y 2008 por país . . . 133

4.11. Variable PIN para los años 2000, 2005 y 2008 por país 133

4.12. Representación consenso CISD en el plano 1-2 136

4.13. Representación consenso CISD en el plano 1-3 138

4.14. Trayectoria de Noruega, Irlanda, Belgica y Alemania con el CISD 139

4.15. Representación consenso CIMC en el plano 1-2 141

4.16. Representación consenso CIMC en el plano 1-3 145

4.17. Trayectoria para Noruega, Irlanda, Bélgica y Alemania con el CIMC 146

C.1. Representación del año 2001 en el plano 1-2 xxxi

C.2. Representación del año 2002 en el plano 1-2 xxxii

C.3. Representación del año 2003 en el plano 1-2 xxxii

C.4. Representación del año 2004 en el plano 1-2 xxxiii

C.5. Representación del año 2006 en el plano 1-2 xxxiii

C.6. Representación del año 2007 en el plano 1-2 xxxiv

C.7. Representación del año 2008-1 en el plano 1-2 xxxiv

Bibliografía

[1] ANDERSON T.W. (1984). *An introduction to multivariate statistical analysis*. Wiley, New York.

[2] BACCALÁ N. (2004). 'Contribuciones al análisis de matrices de datos multivía: tipología de las variables'. Tesis doctoral. Universidad de Salamanca.

[3] BLÁZQUEZ A. (1988). 'Análisis biplot basado modelos lineales generalizados'. Tesis doctoral. Universidad de Salamanca.

[4] BOUROCHE J.M. (1975). 'Analyse des données ternaires: la double analyses en composantes principales'. Tesis doctoral. Universidad de Paris VI.

[5] CALIER A. AND KROONEMBERG P.M. (1996) 'Decompositions and biplots in three-way correspondence analysis', *Psychometrika* **61**, 355-373.

[6] CARROLL J.D. (1968). 'A generalization of canonical correlation analysis to three or more sets of variables'. *Proceedings of 76th annual convention of the American Psychological Association* pp 227-228.

[7] 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.

[8] CASIN P. (2001). ' A generalization of principal component analysis to K sets of variables'. *Computational Statistics and Data Analysis* **35**, 417-428

[9] CUADRAS C. M. (1996). *Métodos de Análisis Multivariante*. EBU, S.L. Barcelona. PPU, S.A.

[10] DAZY F. ET LE BARZIC J. F. (1996). *L'analyse des données évolutives. Méthodes et applications*. Éditions Technip. Paris.

[11] ECKART C. AND YOUNG G. (1936). 'The aproximation of one matrix by another of lower rank' *Psychometrika* **1**, 211-218.

[12] ECKART C. AND YOUNG G. (1939). 'A Principal Axis Transformation for Non-hermitian Matrices' *Am. Math. Soc. Bull* **45**, 118-121.

[13] ESCOFIER B. ET PAGES J. (1984). 'Analyse factorielle multiple'. *Cahiers du BURO* 2. ISUP. Paris.

[14] ESCOFIER B. ET PAGES J. (1990). *Analyses Factorielles Simples et Multiples: Objectifs, Méthodes et Interprétation*. Dunod, Paris.

[15] ESCOUFIER Y. (1973). 'Le traitement des variables vectorielles'. *Biometrics* **2**, 750-760.

[16] ESCOUFIER Y. (1976). 'Opérateur associé à un tableau de données'. *Annales de l'Insee* **22-23**, 165-178.

[17] ESCOUFIER Y. (1980). 'L'Analyse Conjointes de Plusieurs Matrices de Données'. *Biométrie et temps*, Jolivet, M. (ed.), 59-72. Paris: Société Francaise de Biométrie

[18] ESCOUFIER Y. (1985). 'Objectifs et Procédures de l' Analyse Conjointe de Plusieurs Tableaux'. *Statiste et Analyse des Données:***10** 1, 1-10.

[19] ESCOUFIER Y. (1987). *The duality diagram: a means of better practical*. In P. Legendre, and L. Legendre, (eds.). Developments in Numerical Ecology, pp.: 139-156. NATO advanced Institute, Serie G. Springer Verlag, Berlin.

[20] FLURY B. (1984). 'Common Principal Components in K groups'. Wiley, New York. *Journal of the American Statistical Association* **79** 892-898.

[21] FLURY B. (1988). *Common Principal Components and Related Multivariate Models*. Wiley, New York.

[22] FLURY B. (1995). *Developments in principal component analysis*. In: W. J. Krzanowski, (ed.). Recent Advances in Descriptive Multivariate Analysis, pp.: 90-112. Clarendon Press, Oxford.

[23] GABRIEL K.R. (1971). 'The biplot graphic display of matrices with application to principal component analysis' *Biometrika* **58**, 3 453-467

[24] GABRIEL K.R. (1995a). 'Generalised Linear Models' Seminar. Universidad de Salamanca.

[25] GABRIEL K.R. (1981). *'Biplot Display of Multivariate Matrices for Inspection of Data Diagnosis* In Barnett, V. (ed.). Interpreting Multivariate Data, pp.: 147-174. Wiley. Chichester, UK,

[26] GALINDO M.P. (1985). 'Contribuciones a la representación de datos multidimensionales'. Tesis Doctoral. Universidad de Salamanca.

[27] GALINDO M.P. (1986). 'Una alternativa de representación simultánea: HJ-Biplot' *Questíio* **10**, 1 13-23.

[28] GOWER J.C. (1984). *Biplots*. Chapman and Hall, London.

[29] GREEN P. E AND CARROLL J. D. (1976). *Mathematical Tools for Applied Multivariate Analysis*. Academic Press, N.Y.

[30] GREENACRE M. (1984). *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, London.

[31] GITTINS R. (1985). *Canonical Analyses*. Springer-Verlag, New York.

[32] 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.

[33] HORST P. (1963). *Matrix Algebra for social scientists* Holt, Rinehart and Winston Inc, New York.

[34] JOHNSON R. M. (1963). 'On a theorem stated by Eckart and Young *Psychometrika* **28**,3, 259-262.

[35] KELLER J. B. (1962). 'Factorization of matrices by least squares' *Biometrika* **49**, 239-242

[36] KIERS H. A. L. (1988). 'Comparison of Anglo-Saxon and French Three-Mode Methods' *Statistique et Analyse des Données* **13**, 14-32.

[37] KIERS H. A. L. (1991). 'Hierarchical relations among three-way methods' *Psychometrika* **56**, 449-470.

[38] KIERS H. A. L., CLÉROUX R. AND TEN BERGE J. (1994). 'Generalized canonical analysis based on optimising matrix correlations and a relation with IDIOSCAL' *Computational Statistics and Data Analysis* **18**, 331-340.

[39] KRZANOWSKI W.J. (1979). 'Between-groups comparison of principal components' *Journal of the American Statistical Association* **74**, 703-707.

[40] KRZANOWSKI W.J. (1982). 'Between-groups comparison of principal components. Some sampling results' *Journal of Statistical Computation and Simulation* **15**, 141-154.

[41] KROONENBERG P. M. (1983). *Three-Mode Principal Components Analysis. Theory and Applications.* DSWO-Press. Leiden, The Netherlands

[42] KROONENBERG P. M. AND LEEUW J. (1980). 'Principal component analysis of three-mode data by means of alternating least squares algorithms.' *Psychometrika* **45**, 69-97.

[43] LAVIT C. (1988) *Analyse Conjointe de Tableaux Quantitatifs*. Massons, París.

[44] LAVIT C., ESCOUFIER Y., SABATIER R. AND TRAISSAC P. (1994) 'The ACT (STATIS method)'. *Computational Statistics and Data Analysis* **18**, 97-119.

[45] LEBART L., MORINEAU A. ET PIRON M. (1995 a) *Statistique Exploratoire Multidimensionnelle*. Dunod. Paris.

[46] LEBART L., MORINEAU A. Y WARNICK K. M. (1995 b) *Multivariate Descriptive Statistical Analysis*. Wiley, New York.

[47] L'HERMEIR DES PLANTES H. (1976) 'Structuration des Tableaux a Trois Indices de la Statistique: Theorie et Application d'une Méthode d'Analyse Conjointe'. Tesis Doctoral. Université des Sciences et Techniques du Languedoc.

[48] MACDUFFEE C. C. (1946) *The theory of matrices*. (corrected reprint of first edition) Chelsea Publishing Co, New York.

[49] MARDÍA K.V., KENT J.T. AND BIBBY J.M. (1979) *Multivariate Analysis*. Academic Press, London.

[50] MARTÍN-RODRÍGUEZ J., GALINDO-VILLARDÓN M.P. AND VICENTE-VILLARDÓN J.L. (2002) 'Comparison and Integration of Subspaces from a Biplot Perspective'. *Journal of Statistical Planning and Inference* **102**, 411-423.

[51] OECD. Bibilioteca en línea. http://stats.oecd.org/Index.aspx?lang=en. Consulta: Sep 17, 2010.

[52] OECD. Bibilioteca en línea. http://www.oecd-ilibrary.org/fr/economics. Rapport annuel de l'OCDE-2007.

[53] TENNENHAUS M. ET PRIEURET (1974) 'Analyse de séries chronologiques multidimensionnelles' *Revue RAIRO* **2**, 5-16.

[54] TUCKER L.R. (1964). 'The extension of Factor Analysis to Three dimensional Matrices'. *In H. Gullikson and N. Frederiksen (Eds.)* Contributions to Mathematical Psychology. pp.: 46-57. New York.

[55] TUCKER L.R. (1966). 'Some mathematical notes on three-mode factor analysis'. *Psychometrika* **31**, 279-311.

[56] TUCKER L.R. (1972). 'Relation between multidimensional scaling and three-mode factor analysis'. *Psychometrika* **37**, 3-27.

[57] VALLEJO A. (2004). 'Análisis de datos multivía con estructura de grupos'. Tesis doctoral. Universidad de Salamanca.

[58] VICENTE VILLARDON J. L. (2011). 'MULTIPLOT: un Programa para Análisis de Datos Multivariantes Basado en Biplot'. Departamento de Estadística. Universidad de Salamanca.

Apéndices

Apéndice A

Tabla de datos

Datos para el año 2000										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al1	-1,7	2,64	550,2	33,66	572,3	495,4	25952,23	30,3	68,5	5,27
Au1	-0,7	5,84	62,3	30,56	59,9	67,4	28773,1	30,8	67,2	5,56
B1	4	1,78	185,2	26,3	82,8	171,7	27627,82	27	71,6	5,57
C1	2,7	3,3	277,6	22,53	605,6	240	28484,97	33,2	64,5	5,93
D1	1,6	3,34	49,6	26,44	36	44,4	28825,85	26,8	70,6	5,66
E1	-4	6,66	113,3	31,15	222,2	152,9	21322,67	29,2	66,4	5,53
USA1	-4,2	2,58	780,3	23,01	4025,7	1258,1	35050,79	24,2	74,6	6,03
F1	1,6	3,71	295,6	22,32	536,1	304	25275,5	22,9	74,3	5,39
G1	-7,8	17,39	11	22,6	53,4	29,8	18412,46	21	72,5	6,11
Ir1	-0,4	7,86	76,3	28,62	23,7	50,7	28647,39	41,1	55,5	5,48
It1	-0,5	5,36	239,9	32,41	269,9	238,1	25597,47	28,4	68,8	5,58
J1	2,6	5,05	479,2	31,22	1048,6	379,7	25607,72	31,1	67,2	1,74
N1	15	4,14	59,9	21,9	139,6	34,4	36130,12	42	56	6,22
PB1	1,9	3,03	180,1	20,21	89,7	174,7	29409,14	24,9	72,4	5,41
P1	-10,2	12,78	24,4	34,71	43,4	39,9	17088,7	27,6	68,6	5,6
RU1	-2,6	1,54	282,9	25,18	374,4	339,4	26074,49	27,3	71,7	5,33
S1	3,8	2,38	87,4	24,57	145,2	73,1	27761,14	28,6	69,4	5,37

Datos para el año 2001										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al2	0,2	2,61	572	33,05	581,9	486,3	26859,4	29,7	69	4,8
Au2	-0,8	5,71	64,7	29,92	60,7	69	28803,77	30,3	67,7	5,08
B2	3,4	1,65	190,3	26,03	78,6	178,7	28493,32	26,1	72,6	5,06
C2	2,3	2,82	261,1	22,43	589,8	221,6	29331,82	31,9	65,9	5,48
D2	2,6	3,29	50,1	25,42	37,7	44,3	29442,2	25,8	71,4	5,09
E2	-3,9	6,5	116,1	31,46	233,2	155	22594,97	29,2	66,6	5,12
USA2	-3,9	2,45	731	22,46	3838,6	1180,1	35871,41	23	75,8	5,02
F2	1,9	3,57	289,6	22,2	545,7	293,9	26649,07	22,4	74,7	4,94
G2	-7,3	16,11	10,3	23,03	53,1	28,2	19931,83	21,4	72,2	5,3
Ir2	-0,7	7,02	77,4	29,08	24,6	51,1	30475,78	40,6	56,3	5,02
It2	-0,1	5,29	244,3	32,11	271,9	236,1	27131,74	28,1	69,2	5,19
J2	2,1	4,88	402,6	30,48	1029,8	348,6	26156,16	29,8	68,6	1,32
N2	16,1	3,93	59	21,77	119,2	33	37097,61	40,2	58	6,24
PB2	2,4	2,89	175,5	20,38	93,8	169,9	30793,14	24,7	72,8	4,96
P2	-9,9	12,82	24,1	34,24	46,2	39,5	17802,9	27,3	69,1	5,16
RU2	-2,1	1,39	272,6	24,62	382,4	338	27582,67	26,3	72,8	4,93
S2	3,8	2,26	76,3	23,77	161,6	63,5	27968,45	27,9	70,1	5,11

Datos para el año 2002										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al3	2,1	2,52	615,6	32,53	582	490,1	27587,15	29,1	69,7	4,78
Au3	2,7	5,72	71,3	29,64	60,3	71,4	30231,09	29,5	68,5	4,97
B3	4,6	1,71	215,8	25,4	80,9	198,1	30013,75	25,5	73,3	4,89
C3	1,7	2,77	252,6	22,5	601,2	222,4	29911,32	31	66,9	5,3
D3	2,9	3,16	55,7	24,53	39,3	49,3	30756,33	25,6	72,2	5,06
E3	-3,3	6,01	125,9	31,26	241,6	165,9	24066,5	28,9	67	4,96
USA3	-4,3	2,54	693,2	21,58	4026,1	1202,3	36764,5	22,4	76,6	4,61
F3	1,4	3,48	304,9	21,81	553,9	303,8	27776,42	21,8	75,5	4,86
G3	-6,8	15,47	10,8	22,78	53,9	32,5	21597,59	19,5	74,6	5,12
Ir3	-0,9	7,04	88,3	27,69	24,8	52,3	33000,36	41,1	56,2	4,99
It3	-0,8	5,07	254,2	32,07	277,5	246,6	26803,97	27,8	69,7	5,03
J3	2,9	4,67	416,7	29,68	1048,4	337,6	26804,93	29,1	69,3	1,26
N3	12,5	3,79	59,6	21,81	130,3	34,9	37051,94	37,9	60,3	6,38
PB3	2,5	2,63	175,3	19,01	96,1	163,4	31943,49	24,1	73,6	4,89
P3	-8,1	12,41	25,8	34,03	45,7	40	18446,84	26,8	69,9	5,01
RU3	-1,7	1,39	280,6	23,88	384,6	359,4	28887,58	25	74,1	4,9
S3	4	2,13	82,9	23,13	146,7	67,1	29003,77	27,5	70,6	5,3

Datos para el año 2003										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al4	1,9	2,52	748,5	31,86	601,5	601,8	28563,29	28,9	70,2	4,07
Au4	1,7	5,6	89,2	29,55	57,7	91,5	31076,69	29,5	68,6	4,15
B4	4,1	1,76	255,5	24,83	83,6	234,8	30237,91	24,8	74,1	4,15
C4	1,2	2,78	272,1	22,21	589,5	240,2	31268,55	31,2	66,7	4,8
D4	3,5	3,07	64,6	24,05	46,2	56,2	30423,68	25	73,1	4,31
E4	-3,5	5,75	156,3	30,81	257,9	209,7	24745,16	28,9	67,1	4,13
USA4	-4,7	1,65	723,7	20,01	4054,4	1305,1	38142,73	22	76,8	4,02
F4	0,8	3,41	358,1	21,6	561,8	362,5	27396,09	21,2	76,3	4,13
G4	-6,6	15,29	13,7	22,53	57,9	44,9	22698,91	19	75,5	4,27
Ir4	-0,75	6,5	92,9	27,81	24,9	54,2	34460,67	37,8	59,7	4,13
It4	-1,3	4,92	299,5	32,16	286,3	297,4	27134,11	27,1	70,4	4,3
J4	3,2	4,63	472	29,25	1037,5	383,5	27487,12	29	69,4	1
N4	12,3	3,68	67,9	21,6	106,8	39,9	38294,03	37,8	60,7	5,05
PB4	5,5	2,89	227,3	19,22	96,8	209	31698,58	23,9	73,8	4,12
P4	-6,1	12,69	31,8	32,56	46,5	47,1	18788,91	25,8	70,9	4,18
RU4	-1,6	1,25	307,7	23,34	395,5	393,5	29844,97	23,9	75,2	4,53
S4	7,2	2,1	102,4	22,72	135,4	84,2	30059,05	27,2	70,9	4,64

Datos para el año 2004										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al5	4,7	2,43	911,8	31,51	608,5	718,2	29895,28	29,3	69,6	4,04
Au5	2,2	5,03	110,8	27,78	61,6	111,3	32592,22	29,4	68,7	4,15
B5	3,5	1,98	306,4	24,88	84,4	285,4	31146,04	24,6	74,3	4,06
C5	2,3	2,66	317,2	22,3	599,9	273,8	32845,62	31,7	66,1	4,58
D5	2,3	3,12	74,7	23,72	40,4	66,8	32295,54	24,7	73,4	4,3
E5	-5,3	5,52	182,7	30,55	277,2	259,3	25953,44	29,1	67,3	4,1
USA5	-5,3	1,6	817,9	19,95	4147,7	1525,3	40267,19	22,2	76,5	4,27
F5	0,6	3,37	413,7	21,33	569,1	434,2	28268,56	20,9	76,6	4,1
G5	-5,9	12,59	15,2	22,47	58,8	52,8	24155	19,4	75,7	4,26
Ir5	-0,6	6,36	104,3	27,61	25,2	62,3	36444,8	36	61,5	4,06
It5	-0,9	4,47	353,5	31,01	295,8	355,3	27411,17	27	70,5	4,26
J5	3,7	4,51	565,7	28,37	1067,2	455,2	29020,89	29,3	69,1	1,49
N5	12,7	3,49	82,5	20,94	110,2	48,5	42250,01	39,9	58,5	4,37
PB5	7,5	3,12	290,5	19,11	100,8	257,7	33202,9	23,8	74	4,1
P5	-7,6	12,14	35,7	31,36	44,8	54,9	19167,75	25,4	71,4	4,14
RU5	-2,1	1,27	349	22,29	391,2	468,1	31785,29	23,4	75,6	4,88
S5	6,7	2,13	123,2	22,64	151,7	100,5	32060,21	27,6	70,6	4,43

Datos para el año 2005										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al6	5,1	2,38	977,1	29,98	613,4	779,8	31365,57	29,1	70	3,35
Au6	2,2	5,52	117,7	27,62	63	120	33409,34	29,5	68,9	3,39
B6	2,6	2,04	334	24,68	85,7	320,2	32140,92	24,1	75,1	3,37
C6	1,9	2,71	360,6	21,99	626	314,4	35105,98	31,63	66,17	4,07
D6	4,3	3,06	83,3	24,14	36,2	75	33195,88	25,5	73,1	3,4
E6	-7,4	5,29	192,8	29,85	288,9	289,6	27376,76	29,7	67,1	3,39
USA6	-6	1,55	904,3	19,8	4268,4	1732,3	42493,81	22,3	76,4	4,29
F6	-0,4	3,29	434,4	21,11	571,5	476	29692,12	20,7	77	3,41
G6	-7,4	12,41	17,5	22,43	59,4	54,9	24640,64	19,6	75,5	3,59
Ir6	-3,5	5,9	110	27,95	25,6	70,3	38569,6	35	63,1	3,32
It6	-1,7	4,24	373	31,11	296,8	384,8	28144,03	26,9	70,9	3,56
J6	3,6	4,43	594,9	27,92	1088,4	515,9	30311,52	29,1	69,4	1,35
N6	16,3	3,31	103,8	20,91	137,2	55,5	47318,78	42,9	55,6	3,75
PB6	7,3	3,15	320,1	19,55	100,2	283,2	35110,65	24,2	73,7	3,37
P6	-9,5	11,9	38,1	30,75	46,2	61,2	20656,24	24,5	72,6	3,44
RU6	-2,6	1,35	384,4	22,27	395,4	515,8	32724,4	23,5	75,9	4,41
S6	7	2	130,3	22,02	158,4	111,4	32298,09	27,7	71,2	3,38

Datos para el año 2006										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al7	6,5	2,27	1122	29,8	629,4	922,2	32904,62	29,9	69,2	3,76
Au7	2,8	5,53	134,2	28,23	60,8	134,3	35311,68	29,7	68,6	3,8
B7	2	1,94	369,1	24,68	84,3	353,7	33364,8	24	75,1	3,81
C7	1,4	2,64	388,2	21,99	615,9	350	36903,65	31,44	66,39	4,21
D7	3	2,97	91,6	23,63	45,6	85,3	35183,02	26,1	72,5	3,81
E7	-9	4,8	214,1	29,67	295,5	330	29637,93	29,8	67,4	3,78
USA7	-6	1,52	1037	19,94	4274,3	1919	44629,54	22,7	76,3	4,79
F7	-0,5	3,17	479	20,96	569,3	529,9	30813,18	20,4	77,5	3,8
G7	-11,3	12,03	20,9	22,03	60,2	63,7	26355,61	20	76,1	4,07
Ir7	-3,5	5,62	108,8	27,59	27,1	76,6	41143,16	34	64,4	3,79
It7	-2,6	4,31	417,2	30,46	307,7	442,6	29517,13	27,2	70,7	4,05
J7	3,9	4,26	646,7	28,03	1093	579,1	31935,34	28,9	69,7	1,74
N7	17,3	3,28	122,2	20,89	121,2	64,3	52041,27	44,8	53,7	4,08
PB7	9,3	3,12	370,2	19,3	98,4	331,5	37173,16	24,6	73,2	3,78
P7	-10	11,77	43,4	30,74	48,6	66,7	21662,08	24,3	72,9	3,91
RU7	-3,3	1,3	448,4	22,12	394	598,4	34084,65	23,6	75,7	4,5
S7	8,5	1,98	147,4	22,04	143,3	127,1	34328,39	27,9	70,6	3,7

Datos para el año 2007										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al8	7,9	2,26	1328,8	30,03	629,5	1059,3	34683,36	30,2	68,9	4,22
Au8	3,6	5,75	156,6	27,34	60,9	156,1	36838,85	30,4	67,8	4,3
B8	2,2	1,85	430,9	24,43	87,5	413,6	34665,16	23,8	75,4	4,33
C8	1	2,5	420,2	21,59	639,7	380,4	38499,55	31,47	66,36	4,27
D8	1,5	2,9	101,6	23,51	39,2	98	36325,62	25,8	73	4,29
E8	-10	4,56	253,8	29,4	300,2	391,2	31469,28	29,2	68,1	4,31
USA8	-5,2	1,43	1162,5	19,78	4322,9	2017,1	46433,81	21,8	76,9	4,63
F8	-1	3,01	539,4	20,79	564,4	611,1	32493,67	20,4	77,4	4,3
G8	-14,5	11,55	23,5	22,37	62,7	76,1	27793,07	19,6	76,5	4,5
Ir8	-5,3	5,42	122	27,53	27,9	86,7	44296,29	33,6	64,8	4,33
It8	-2,4	4,01	500,2	30,48	308,2	511,9	31015,81	27,5	70,4	4,49
J8	4,8	4,24	714,3	27,86	1123,5	622,2	33634,58	28,5	70,1	1,67
N8	14,1	2,84	136,4	21,12	136,4	80,3	53671,95	42,7	55,9	4,77
PB8	8,7	2,97	476,8	19,1	103,2	421,3	39593,77	24,5	73,4	4,29
P8	-9,4	11,71	51,3	30,72	46,9	78,2	22638,4	24,5	73	4,42
RU8	-2,7	1,37	440	22,35	392,3	624,7	34957,07	23,1	76,1	5,01
S8	8,8	2,24	169,1	21,62	148,8	152,8	36785,28	28,3	70,3	4,17

Datos para el año 2008										
etique	BCC	EAG	EBS	EIN	ELC	IBS	PIB	PIN	PSE	TIN
Al9	6,7	2,27	1466,1	29,9	626,7	1204,2	35432,12	29,8	69,3	3,98
Au9	3,2	5,59	172,8	26,07	64,1	175,4	37857,51	30,7	67,6	4,26
B9	-2,5	1,8	477,2	24,65	83,1	470,7	35287,79	23,1	76,2	4,4
C9	0,5	2,35	455,7	21,54	632,6	408,3	39014,33	31,54	66,28	3,6
D9	2,2	2,71	115,8	22,9	36,4	110,8	36808,27	25,6	73,2	4,28
E9	-9,6	4,35	279,2	27,92	306,5	418,7	31455,41	28,4	69	4,36
USA9	-4,9	1,49	1299,9	19,05	4329,4	2164,8	47185,65	21,49	77,26	3,67
F9	-2,3	2,89	594,9	20,64	569,9	695,5	33052,36	20,4	77,6	4,23
G9	-14,5	11,32	25,5	22,3	58,6	89,3	28896,41	19	77,3	4,8
Ir9	-5,3	5,77	126,9	25,75	28,5	84,1	41493,22	33,56	65,24	4,55
It9	-3,4	3,86	539,6	30,03	312,4	553,2	31195,24	27	71	4,68
J9	3,2	4,19	781,4	27,34	1078,1	762,5	34131,54	28,07	70,64	1,47
N9	18,5	2,79	177,6	21,1	141,7	94,5	58598,61	46,2	52,6	4,46
PB9	4,8	2,58	485,6	18,02	107,7	437,5	41063,18	25,5	72,8	4,23
P9	-12,1	11,52	55,9	29,43	45,4	90,1	23286,97	23,9	73,8	4,52
RU9	-1,6	1,47	459,3	21,33	386,2	636	35620,02	23,6	75,2	4,59
S9	9,8	2,19	183,9	21,64	149,5	167,3	36789,84	28	70,5	3,89

Apéndice B

Resultados para los ACP individuales

```
GROUP/OCCASSION : 2000
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)

Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot

Eigenvalues and variance explained

Inertia
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1      7.525      0.354      0.354
Axis 2      6.931        0.3      0.654
Axis 3      5.033      0.158      0.812
Axis 4      4.036      0.102      0.914
Axis 5      2.963      0.055      0.969
Axis 6       1.52      0.014      0.984
Axis 7      1.329      0.011      0.995
Axis 8      0.854      0.005      0.999
Axis 9      0.258          0          1
Axis 10     0.239          0          1

Row Coordinates ----------
Axis
Row         Axis 1   Axis 2      Axis 3
  Al1       1.154      0.326      1.586
  Au1      -1.037     -0.189      0.221
```

```
  B1       0.115     -0.357     -0.696
  C1        0.19     -1.341     -0.476
  D1      -0.532     -0.212     -0.885
  E1      -1.138      1.032      0.627
  USA1     5.979      0.413     -0.282
  F1       0.966      0.669      -1.18
  G1      -1.774      3.224     -1.661
  Ir1     -2.042     -1.805      1.242
  It1     -0.265      0.498       0.61
  J1       1.059     -0.019       3.21
  N1      -1.267     -4.609     -0.557
  PB1      0.422     -0.245     -1.602
  P1      -2.214      2.904      0.935
  RU1      0.759      0.389     -0.334
  S1      -0.375     -0.679     -0.757

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC     -0.005     -0.521       -0.1
  EAG     -0.286      0.326      0.016
  EBS      0.487      0.016      0.258
  EIN     -0.138      0.186      0.625
  ELC       0.48      0.029      0.082
  IBS      0.508      0.053      0.114
  PIB      0.256     -0.473     -0.124
  PIN     -0.183     -0.454      0.326
  PSE      0.265      0.391      -0.29
  TIN     -0.064      -0.03     -0.554

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT

Row Contributions ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al1     22.631      1.803     42.754
  Au1     51.886      1.723      2.356
  B1       0.638       6.11     23.201
  C1       1.283     63.778       8.05
  D1      12.665      2.011     35.089
  E1      38.928     32.009     11.825
  USA1     90.15      0.431        0.2
  F1      23.377     11.229     34.919
  G1      16.227     53.617     14.227
  Ir1     35.287     27.571     13.059
```

```
  It1      4.343     15.297     22.971
  J1       6.534      0.002      59.98
  N1       6.571     86.964      1.272
  PB1      4.221      1.426      60.79
  P1      32.195      55.38      5.739
  RU1     29.804       7.83      5.768
  S1       6.591     21.601     26.809

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      0.009     81.523      1.589
  EAG     29.006     31.865      0.039
  EBS     83.855      0.079     10.549
  EIN      6.744      10.36     61.856
  ELC     81.679      0.253      1.055
  IBS     91.341      0.831      2.056
  PIB     23.128     67.235      2.447
  PIN     11.878     61.881     16.839
  PSE     24.778     45.951     13.294
  TIN      1.471      0.269      48.62
*********************************
GROUP/OCCASSION : 2001
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)

Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot

Eigenvalues and variance explained
Inertia
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1      7.679      0.369      0.369
Axis 2      6.878      0.296      0.664
Axis 3      4.973      0.155      0.819
Axis 4      3.861      0.093      0.912
Axis 5      2.928      0.054      0.966
Axis 6      1.578      0.016      0.981
Axis 7       1.43      0.013      0.994
Axis 8      0.909      0.005      0.999
Axis 9        0.3      0.001          1
Axis 10     0.257          0          1

Row Coordinates ----------
Axis
Row Axis 1 Axis 2 Axis 3
```

```
  Al2     -1.216     -0.185     -1.628
  Au2      1.136     -0.001     -0.265
  B2      -0.222      0.315        0.8
  C2      -0.081      1.295      0.276
  D2       0.474      0.285      1.081
  E2       1.167     -1.112     -0.697
  USA2    -5.958      0.024     -0.227
  F2      -1.131     -0.418      1.374
  G2       1.533     -3.017      1.615
  Ir2      2.294       1.57     -1.845
  It2      0.257     -0.436     -0.585
  J2      -1.139     -0.673     -2.615
  N2       1.698      4.571      0.176
  PB2     -0.482      0.524      1.641
  P2       2.155     -3.182     -0.676
  RU2     -0.923     -0.224      0.584
  S2       0.439      0.664      0.992

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      0.027      0.521       0.09
  EAG      0.267     -0.359     -0.033
  EBS      -0.47      0.006     -0.267
  EIN       0.15     -0.244     -0.595
  ELC     -0.464     -0.004     -0.145
  IBS     -0.494      -0.02     -0.161
  PIB     -0.216      0.505      0.039
  PIN      0.243      0.395     -0.426
  PSE     -0.317     -0.323      0.387
  TIN      0.133      0.158      0.431

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT

Row Contributions ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al2     22.607      0.524     40.489
  Au2     66.491          0      3.612
  B2       2.463      4.972     32.042
  C2        0.26     66.658       3.02
  D2        9.44      3.428     49.158
  E2      36.758     33.409     13.101
  USA2    90.272      0.001      0.131
  F2      32.574      4.436     48.004
```

```
  G2      13.699     53.059     15.206
  Ir2     40.266     18.875     26.052
  It2          4      11.49      20.69
  J2       8.143      2.842     42.936
  N2      11.446     82.923      0.122
  PB2      5.779      6.833     67.135
  P2      28.725     62.647      2.828
  RU2     39.116      2.296     15.628
  S2       8.387     19.145      42.79

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      0.272       80.1       1.24
  EAG     26.336     38.199       0.17
  EBS     81.565      0.012     11.024
  EIN      8.332     17.541     54.688
  ELC     79.521      0.005      3.267
  IBS     90.024      0.117      4.005
  PIB     17.243     75.433      0.234
  PIN     21.704     46.038     28.096
  PSE     37.004     30.853     23.111
  TIN      6.568      7.362     28.723

*********************************
GROUP/OCCASSION : 2002
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)

Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot

Eigenvalues and variance explained
Inertia
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1      7.633      0.364      0.364
Axis 2       6.77      0.286      0.651
Axis 3      5.082      0.161      0.812
Axis 4       3.89      0.095      0.907
Axis 5      2.972      0.055      0.962
Axis 6      1.697      0.018       0.98
Axis 7      1.468      0.013      0.993
Axis 8      0.973      0.006      0.999
Axis 9      0.282          0          1
Axis 10      0.24          0          1
```

```
Row Coordinates ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al3      1.196     -0.089     -1.653
  Au3     -1.182     -0.274     -0.216
  B3       0.257     -0.548       0.77
  C3      -0.033     -1.102      0.144
  D3      -0.508     -0.414      1.143
  E3      -1.105      1.085     -0.848
  USA3     5.888     -0.158     -0.342
  F3       1.159      0.398      1.307
  G3      -1.175      3.373      1.699
  Ir3     -2.381      -1.68     -1.901
  It3     -0.334      0.661      -0.77
  J3       1.218      0.595      -2.45
  N3      -1.863     -3.933      0.336
  PB3      0.458     -0.651      1.766
  P3      -2.048      3.343      -0.77
  RU3      1.015      0.179      0.707
  S3      -0.563     -0.785      1.078

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC     -0.059       -0.5      0.087
  EAG     -0.246      0.403     -0.044
  EBS      0.458     -0.035     -0.288
  EIN     -0.158      0.237     -0.582
  ELC      0.464     -0.011     -0.153
  IBS      0.496     -0.003     -0.165
  PIB      0.185      -0.52      0.063
  PIN      -0.26     -0.376     -0.435
  PSE      0.326      0.301      0.408
  TIN     -0.177     -0.168      0.398

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT

Row Contributions ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al3     19.612       0.11     37.429
  Au3     63.722      3.427      2.125
  B3       2.957     13.446     26.544
  C3       0.056     63.396      1.076
  D3      10.585      7.026     53.641
```

```
  E3        32.5      31.35     19.124
  USA3    89.299      0.065      0.302
  F3      36.523      4.315     46.436
  G3       7.619     62.766      15.92
  Ir3     37.176     18.499     23.681
  It3      5.377     21.023     28.468
  J3       9.628      2.296     38.952
  N3      16.984     75.696      0.551
  PB3      4.648      9.386      69.16
  P3       25.03     66.716       3.54
  RU3     41.226      1.283     19.995
  S3      12.249     23.824      44.88

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      1.287      71.69      1.212
  EAG      22.07     46.561      0.313
  EBS      76.48      0.349     13.361
  EIN      9.104     16.042     54.671
  ELC      78.48      0.033       3.79
  IBS      89.61      0.003      4.375
  PIB     12.426     77.338      0.648
  PIN     24.584     40.512     30.595
  PSE     38.623     25.917     26.818
  TIN     11.452      8.049     25.603

*********************************
GROUP/OCCASSION : 2003
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)

Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot

Eigenvalues and variance explained
Inertia
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1      7.568      0.358      0.358
Axis 2      6.746      0.284      0.642
Axis 3      5.168      0.167      0.809
Axis 4      3.847      0.093      0.902
Axis 5      3.126      0.061      0.963
Axis 6      1.636      0.017       0.98
Axis 7      1.376      0.012      0.992
Axis 8      1.101      0.008      0.999
```

```
Axis 9      0.291      0.001          1
Axis 10      0.25          0          1

Row Coordinates ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al4      1.317      0.198      1.801
  Au4     -1.258      0.229      0.384
  B4       0.286      0.291     -0.761
  C4      -0.177      1.253     -0.264
  D4      -0.539      0.213     -1.194
  E4      -1.092     -0.942      1.024
  USA4     5.913      0.295      0.168
  F4       1.241     -0.759      -1.18
  G4      -1.036     -3.465      -1.57
  Ir4     -2.131      1.582      1.454
  It4     -0.324     -0.756      0.874
  J4       0.886     -0.376      2.797
  N4      -1.893      3.908      -0.37
  PB4      0.407       0.65     -1.603
  P4      -2.022     -3.164      0.647
  RU4      1.058     -0.243     -0.965
  S4      -0.636      1.084     -1.242

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC     -0.096      0.476     -0.126
  EAG     -0.257     -0.408      0.041
  EBS      0.446      0.043      0.303
  EIN     -0.189     -0.206      0.585
  ELC       0.47      0.024      0.136
  IBS      0.505       0.01      0.163
  PIB      0.182      0.516     -0.066
  PIN     -0.257      0.409      0.381
  PSE       0.32     -0.332     -0.366
  TIN     -0.117      0.137     -0.469

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT

Row Contributions ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al4     20.439      0.462     38.254
  Au4     67.719      2.247      6.298
```

```
  B4        4.09      4.236     29.072
  C4       1.105     55.542      2.473
  D4      11.457      1.792     56.341
  E4      29.695     22.068     26.123
  USA4    90.741      0.226      0.073
  F4      37.173     13.898     33.622
  G4       5.993     67.081     13.764
  Ir4     38.352     21.142     17.866
  It4      3.967     21.615     28.938
  J4       4.912      0.884     48.963
  N4      17.773     75.781      0.679
  PB4      3.813      9.745     59.249
  P4      26.735     65.451      2.736
  RU4     35.815      1.881     29.754
  S4      10.494     30.487     40.045

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      3.268     64.351      2.648
  EAG     23.592     47.378      0.283
  EBS     71.191       0.53     15.313
  EIN      12.74     12.046     57.064
  ELC     78.941      0.165      3.109
  IBS     91.163      0.026      4.451
  PIB     11.832     75.797      0.731
  PIN     23.681     47.485     24.292
  PSE     36.732     31.342     22.342
  TIN       4.86      5.305     36.683

*********************************
GROUP/OCCASSION : 2004
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)

Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot

Eigenvalues and variance explained
Inertia
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1      7.407      0.343      0.343
Axis 2      6.842      0.293      0.635
Axis 3      5.204      0.169      0.805
Axis 4      3.872      0.094      0.898
Axis 5      3.283      0.067      0.966
```

```
Axis 6      1.762      0.019      0.985
Axis 7      1.153      0.008      0.993
Axis 8      0.953      0.006      0.999
Axis 9      0.281          0          1
Axis 10     0.246          0          1

Row Coordinates ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al5         1.3      0.487      1.984
  Au5      -1.165      0.391      0.073
  B5        0.306     -0.017     -0.551
  C5       -0.084      1.314     -0.377
  D5       -0.471     -0.031      -1.23
  E5       -1.153     -0.953       1.05
  USA5      5.793     -0.206      0.165
  F5        1.182     -1.043     -0.943
  G5       -1.277     -3.107     -1.245
  Ir5      -1.898      1.478      0.825
  It5      -0.365     -0.746      0.833
  J5        0.539      0.182      3.133
  N5       -1.566      4.335     -0.407
  PB5         0.48      0.584     -1.505
  P5       -2.303     -3.148      0.781
  RU5       1.154     -0.488     -1.423
  S5       -0.474      0.968     -1.162

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      -0.047      0.483     -0.087
  EAG      -0.309     -0.355      0.097
  EBS       0.441      0.035      0.342
  EIN      -0.209     -0.169      0.579
  ELC       0.475     -0.009      0.149
  IBS       0.513     -0.027      0.179
  PIB       0.203      0.491     -0.142
  PIN      -0.228      0.456      0.288
  PSE       0.286     -0.401     -0.283
  TIN      -0.017      0.013     -0.541

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT

Row Contributions ----------
Axis
```

```
Row Axis 1 Axis 2 Axis 3
  Al5     17.023      2.384     39.643
  Au5     72.341      8.159      0.285
  B5       5.309      0.016     17.201
  C5       0.247     60.375      4.978
  D5       8.999       0.04     61.413
  E5      27.927     19.108     23.184
  USA5    89.931      0.114      0.073
  F5      32.339     25.177     20.569
  G5      10.888     64.425     10.349
  Ir5     40.761     24.733        7.7
  It5      5.121     21.413      26.68
  J5       1.781      0.203     60.286
  N5       10.86     83.254      0.734
  PB5      4.781      7.068     46.918
  P5       31.73     59.286      3.647
  RU5     28.277       5.06     42.981
  S5       7.092     29.628     42.695

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      0.768     68.182      1.281
  EAG     32.714     36.783      1.606
  EBS     66.548      0.363     19.743
  EIN     15.023      8.339     56.801
  ELC     77.336      0.025      3.782
  IBS     90.282      0.219      5.441
  PIB     14.191     70.643      3.392
  PIN     17.834     60.846     14.058
  PSE     28.066     47.113     13.547
  TIN        0.1      0.046       49.6

*********************************
GROUP/OCCASSION : 2005
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)

Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot

Eigenvalues and variance explained
Inertia
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1      7.444      0.346      0.346
Axis 2       7.06      0.312      0.658
```

```
Axis 3       4.82      0.145      0.803
Axis 4      4.062      0.103      0.906
Axis 5      3.026      0.057      0.963
Axis 6       1.81       0.02      0.984
Axis 7      1.187      0.009      0.993
Axis 8      1.018      0.006      0.999
Axis 9       0.28          0          1
Axis 10     0.225          0          1

Row Coordinates ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al6     -1.295      0.294     -2.089
  Au6      1.235      0.353     -0.094
  B6      -0.251     -0.291      0.346
  C6      -0.349       1.29      0.612
  D6       0.609      0.177       0.84
  E6       1.206     -0.734     -0.903
  USA6    -5.921     -0.588     -0.113
  F6       -1.02     -1.164      0.753
  G6       1.542     -2.786       1.45
  Ir6      1.607      1.238     -0.557
  It6      0.475     -0.792     -0.939
  J6       0.086      0.042     -2.923
  N6       1.012       5.12      0.522
  PB6     -0.436      0.495      1.164
  P6        2.42     -2.866      -0.33
  RU6     -1.298       -0.6      1.391
  S6       0.379      0.812       0.87

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      0.009      0.469      0.044
  EAG      0.329     -0.297     -0.003
  EBS     -0.434     -0.014     -0.404
  EIN      0.251      -0.16     -0.598
  ELC     -0.471     -0.044     -0.157
  IBS     -0.507     -0.074     -0.215
  PIB      -0.23      0.473      0.111
  PIN      0.171      0.485     -0.231
  PSE     -0.221     -0.447      0.213
  TIN     -0.176      0.016      0.543

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT
```

```
Row Contributions ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al6     18.753      0.968     48.814
  Au6     77.998       6.39      0.453
  B6       3.443      4.628      6.553
  C6         3.86     52.599     11.838
  D6      16.379      1.387     31.163
  E6      31.318     11.599     17.557
  USA6     90.81      0.894      0.033
  F6      24.714      32.16     13.443
  G6       16.57     54.092      14.65
  Ir6      36.21     21.485      4.345
  It6      7.375     20.507     28.826
  J6       0.056      0.013     64.696
  N6       3.564     91.264      0.948
  PB6      4.848      6.235     34.474
  P6      37.395     52.452      0.696
  RU6     31.545      6.745     36.248
  S6       4.829     22.141     25.399

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC      0.027     68.461      0.286
  EAG     37.539     27.567      0.001
  EBS     65.201      0.059     23.727
  EIN     21.814      7.929     51.948
  ELC     76.808        0.6      3.587
  IBS     88.863      1.688        6.7
  PIB     18.309     69.776      1.798
  PIN     10.176      73.25      7.735
  PSE     16.841     62.136      6.569
  TIN      10.71      0.084     42.862

********************************
GROUP/OCCASSION : 2006
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)

Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot

Eigenvalues and variance explained
Inertia
```

```
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1      7.353      0.338      0.338
Axis 2      7.201      0.324      0.662
Axis 3      4.753      0.141      0.803
Axis 4      4.097      0.105      0.908
Axis 5      2.925      0.053      0.962
Axis 6        1.9      0.023      0.984
Axis 7      1.215      0.009      0.993
Axis 8       0.98      0.006      0.999
Axis 9      0.266          0          1
Axis 10     0.173          0          1

Row Coordinates ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al7     -1.418      0.524      2.188
  Au7      1.298      0.327      0.198
  B7       -0.21     -0.398     -0.181
  C7      -0.249      1.114     -0.589
  D7       0.647      0.135     -0.733
  E7        1.07     -0.697       0.76
  USA7    -5.885      -0.44     -0.367
  F7      -0.998     -1.229     -0.598
  G7       1.462     -2.943     -1.536
  Ir7      1.504      1.048      0.083
  It7      0.382     -0.787      0.815
  J7       0.066      0.137      3.099
  N7       1.147      5.268      -0.77
  PB7     -0.392        0.6     -0.983
  P7       2.322     -2.857      0.287
  RU7     -1.207     -0.664     -1.062
  S7       0.461      0.861     -0.612

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC          0      0.466      0.042
  EAG      0.322     -0.296     -0.041
  EBS      -0.45      0.007        0.4
  EIN      0.233     -0.145       0.61
  ELC      -0.48      -0.03      0.094
  IBS     -0.518     -0.053      0.187
  PIB     -0.192      0.476     -0.181
  PIN      0.168      0.484      0.162
  PSE     -0.211     -0.458     -0.137
```

```
  TIN     -0.176     -0.021     -0.586

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT

Row Contributions ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al7      20.93      2.858      49.84
  Au7     73.458      4.666      1.711
  B7        2.43      8.759      1.805
  C7       2.784     55.699     15.579
  D7      21.144       0.92     27.111
  E7      26.659     11.306     13.452
  USA7    91.005      0.508      0.355
  F7      22.194     33.634      7.974
  G7      14.234     57.711     15.713
  Ir7     40.638     19.743      0.124
  It7       5.18     21.984     23.549
  J7       0.031      0.135     68.959
  N7       4.255     89.709      1.914
  PB7      3.656      8.581     23.044
  P7      34.974     52.978      0.534
  RU7     34.462     10.432     26.669
  S7       6.755     23.585     11.892

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC          0     70.283      0.243
  EAG     35.132     28.306      0.237
  EBS     68.427      0.017     22.635
  EIN     18.303      6.827       52.5
  ELC      77.93      0.301       1.26
  IBS     90.777        0.9      4.922
  PIB     12.424     73.406      4.645
  PIN      9.516     76.014      3.685
  PSE     15.014     67.854      2.645
  TIN     10.421      0.144     48.437

********************************
GROUP/OCCASSION : 2007
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)

Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot
```

```
Eigenvalues and variance explained
Inertia
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1        7.3      0.333      0.333
Axis 2      7.074      0.313      0.646
Axis 3      4.893       0.15      0.795
Axis 4      3.952      0.098      0.893
Axis 5      3.237      0.066      0.959
Axis 6      1.971      0.024      0.983
Axis 7      1.228      0.009      0.992
Axis 8      1.068      0.007      0.999
Axis 9      0.251          0          1
Axis 10     0.171          0          1

Row Coordinates ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al8     -1.757     -0.335       2.07
  Au8      1.168     -0.912      0.261
  B8      -0.203      0.487      -0.39
  C8      -0.375     -1.149     -0.327
  D8       0.633     -0.144     -0.777
  E8         1.2      0.385      0.717
  USA8    -5.253      1.977     -0.504
  F8      -0.715      1.507     -0.847
  G8       2.411      2.684     -1.182
  Ir8      1.261     -1.489      0.111
  It8       0.55      0.462      0.918
  J8      -0.634       0.01      3.227
  N8      -0.105      -5.02     -0.841
  PB8     -0.758      -0.38     -1.115
  P8       3.097      2.043      0.668
  RU8     -0.647      0.944     -1.306
  S8       0.125     -1.071     -0.684

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC     -0.182     -0.418      0.059
  EAG      0.404      0.214      0.074
  EBS     -0.453      0.104      0.337
  EIN      0.242      0.056      0.621
  ELC     -0.451      0.164      0.067
  IBS     -0.491      0.192      0.139
```

```
  PIB     -0.289     -0.407     -0.216
  PIN      0.054     -0.526      0.186
  PSE     -0.099      0.509     -0.175
  TIN      0.015          0     -0.598

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT

Row Contributions ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al8     27.575          1     38.287
  Au8     52.688     32.129      2.623
  B8       2.055     11.888       7.62
  C8        6.26     58.693      4.744
  D8       19.978      1.028     30.139
  E8       34.319      3.538     12.249
  USA8     77.837     11.027      0.715
  F8       11.185     49.611     15.657
  G8       35.887     44.473       8.63
  Ir8      25.093     34.946      0.195
  It8      10.348      7.286     28.836
  J8        2.523      0.001     65.421
  N8        0.041     93.105      2.614
  PB8       13.48      3.383     29.185
  P8       61.458     26.728      2.861
  RU8      10.105      21.53     41.223
  S8        0.475     34.858     14.229

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC     11.055     54.703      0.521
  EAG      54.25      14.26      0.823
  EBS     68.226      3.356     16.946
  EIN     19.527      0.982     57.645
  ELC     67.648      8.455      0.665
  IBS     80.189     11.474      2.876
  PIB     27.899     51.869      7.001
  PIN      0.983     86.472      5.152
  PSE      3.244     81.162      4.577
  TIN      0.076          0     53.426

*******************************
GROUP/OCCASSION : 2008
BIPLOT ANALYSIS (PRINCIPAL COMPONENTS ANALYSIS)
```

```
Transformation of the raw data: Column standardization
Type of Biplot: RMP - Biplot

Eigenvalues and variance explained
Inertia
Axis Eigenvalue Expl. Var.  Cummulative
Axis 1      7.512      0.353      0.353
Axis 2      7.309      0.334      0.687
Axis 3      4.634      0.134      0.821
Axis 4      3.522      0.078      0.898
Axis 5      3.299      0.068      0.966
Axis 6      1.786       0.02      0.986
Axis 7      1.209      0.009      0.995
Axis 8      0.808      0.004      0.999
Axis 9       0.25          0          1
Axis 10     0.157          0          1

Row Coordinates ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al9     -1.551      1.067      2.109
  Au9      0.347     -1.393      0.355
  B9       0.512      0.593     -0.441
  C9      -1.028     -0.569     -0.035
  D9       0.428     -0.604       -0.8
  E9       1.196     -0.191      0.729
  USA9    -3.132      4.583      -0.87
  F9       0.352       1.46     -1.037
  G9       3.518      0.489     -1.072
  Ir9      0.523     -1.765      0.141
  It9      0.915      0.131      1.026
  J9       -1.15      1.106      2.533
  N9      -3.446     -4.488     -0.354
  PB9     -0.715     -0.115     -1.526
  P9       3.684     -0.042      0.973
  RU9      0.064      0.692     -1.149
  S9      -0.517     -0.952     -0.582

Column Coordinates ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC     -0.403     -0.259      0.024
  EAG      0.414     -0.067       0.14
  EBS     -0.301      0.367      0.276
```

```
  EIN      0.206     -0.039      0.739
  ELC     -0.273      0.395     -0.027
  IBS      -0.28      0.443      0.085
  PIB     -0.449     -0.184     -0.242
  PIN     -0.273     -0.431      0.227
  PSE       0.24      0.444     -0.221
  TIN      0.217     -0.153     -0.438

RELATIVE CONTRIBUTIONS OF THE FACTOR TO THE ELEMENT

Row Contributions ----------
Axis
Row Axis 1 Axis 2 Axis 3
  Al9     20.858      9.866     38.591
  Au9      5.052     81.183      5.284
  B9      11.755     15.741      8.727
  C9      44.802     13.696      0.052
  D9       8.639     17.167     30.142
  E9      44.156       1.13     16.412
  USA9    28.176     60.329      2.172
  F9       3.116     53.668     27.104
  G9      79.632      1.538        7.4
  Ir9      5.632     64.201      0.413
  It9     22.632      0.464      28.42
  J9       9.206      8.513     44.672
  N9      35.574     60.338      0.374
  PB9     13.375      0.349     60.998
  P9      86.821      0.011      6.058
  RU9      0.155      18.29     50.463
  S9       7.769     26.309      9.836

Column Contributions ----------
Axis
Column Axis 1 Axis 2 Axis 3
  BCC     57.211     22.465      0.076
  EAG     60.327      1.505      2.628
  EBS     32.033     45.059     10.188
  EIN     15.031      0.516     73.205
  ELC     26.188     51.981      0.095
  IBS     27.637     65.505      0.975
  PIB     71.013     11.354      7.886
  PIN     26.268     61.966      6.887
  PSE     20.365     65.675      6.562
  TIN     16.592       7.85     25.705
```

Apéndice C

Biplots para los ACP individuales

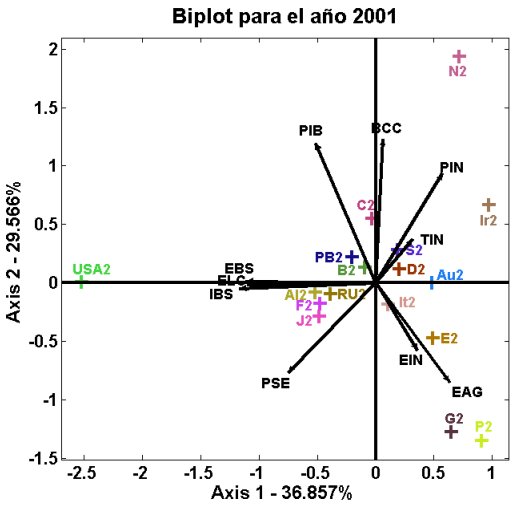

Figura C.1: Representación del año 2001 en el plano 1-2

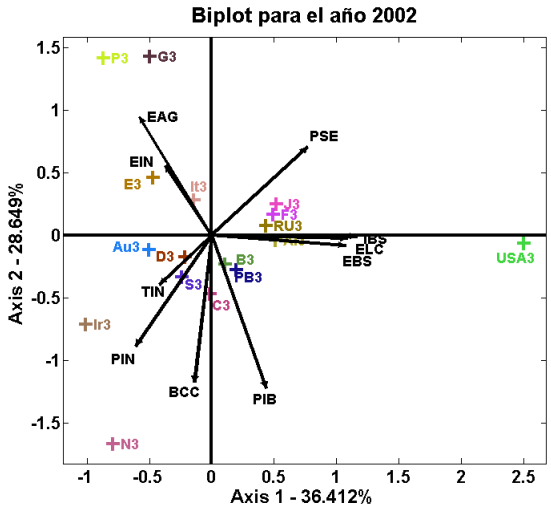

Figura C.2: Representación del año 2002 en el plano 1-2

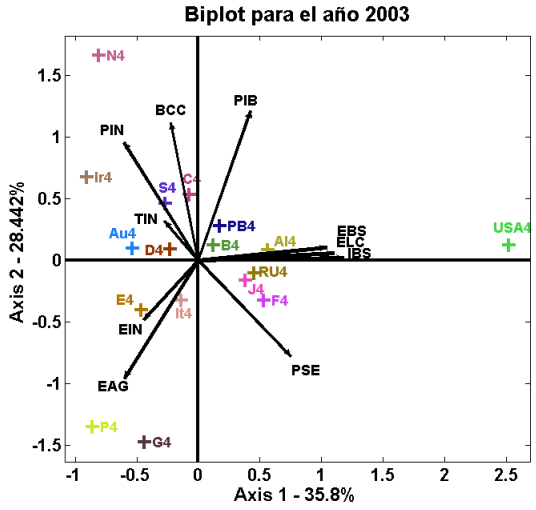

Figura C.3: Representación del año 2003 en el plano 1-2

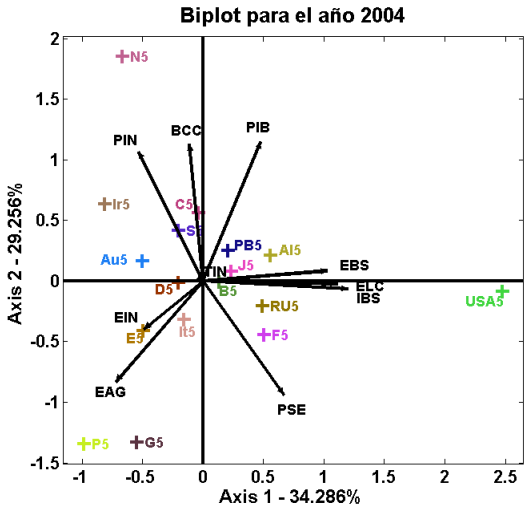

Figura C.4: Representación del año 2004 en el plano 1-2

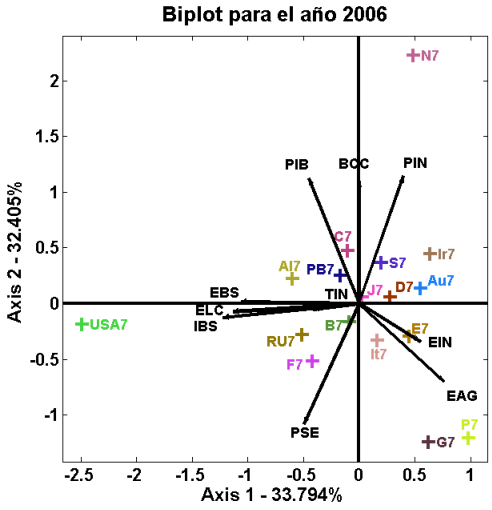

Figura C.5: Representación del año 2006 en el plano 1-2

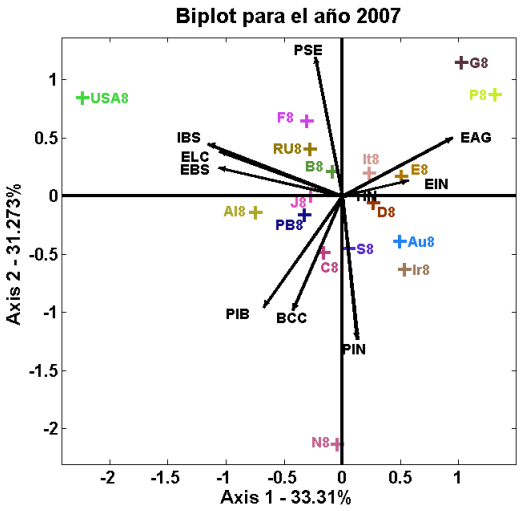

Figura C.6: Representación del año 2007 en el plano 1-2

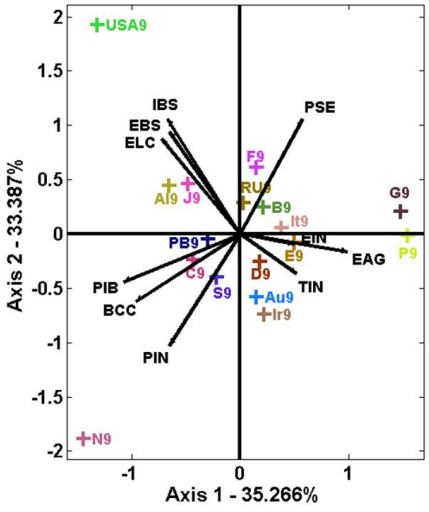

Figura C.7: Representación del año 2008-1 en el plano 1-2