Revista de Matemática: Teoría y Aplicaciones 2000 **7**(1-2) : 43–70
cimpa – ucr – ccss issn: 1409-2433

REGRESIÓN Y ANÁLISIS FACTORIALES*

Roger Lafosse**

Recibido: 15 Julio 2000

Resumen

Este curso introduce algunos resultados recientes que se han aportado en análisis factorial en el contexto de "multi-sets"(incluyendo cubos de datos). Los análisis factoriales más clásicos se basan en la noción de descomposición en valores singulares. Por ello, los análisis de 2 tablas se introducen aquí a partir de un criterio cualitativo, el de no redundancia de las relaciones parciales entre "factores comunes". Una proposición de extensión de la noción de regresión lineal simple, considerada aquí entre los dos conjuntos de individuos que definen los dos conjuntos de variables, conduce a medidas específicas para cada una de las tablas. Entonces, ciertas yuxtaposiciones de gráficos pueden ser justificadas. Todo el procedimiento es prolongado para analizar la dependencia de K tablas con una $(K+1)$-ésima. El ACOM de Chessel y Hanafi (1996) es interpretado en este contexto como un ACP de conjuntos de variables.

Palabras clave: dependencia entre conjuntos de variables, no redundancia, simultaneidad, biplots, mínimos cuadrados ordinarios, correlación, regresión lineal simple.

Abstract

Some recent developments in factor analysis of multi-sets are introduced in this short course. The more usual factor analyses are based on the singular values de-composition. So, the analyses of two matrices here are introduced from a qualitative criterion, for a non redundancy of partial relations between çommon factors". A proposal for extending the simple linear regression, here considered between the two sets of the individuals which define the two sets of variables, lead to specific measures for each matrix. Some juxtapositions of graphics then are justified. The whole previous approach is extended for analyzing the dependency of K matrices with one more matrix. The ACOM of Chessel and Hanafi (1996) then is considered as a PCA of sets of variables.

Keywords: dependence between sets of variables, non redondancy, simultaneity, biplots, ordinary least squares, correlation, simple linear regression.

Mathematics Subject Classification: 62-07, 62H25

*Notas del minicurso ofrecido durante el XII SIMMAC.

**Laboratoire de Statistique et Probabilités, UMR C55830, Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse Cedex 4, France; E-Mail: lafosse@cict.fr

1. Análisis en componentes principales

Se trata de resumir datos numéricos para describir, compactar o cortar la información. Técnica de base raramente utilizada sola, aparece en varios enfoques. En análisis exploratorio, se le usa de manera interactiva sobre todo con varias clasificaciones en subgrupos, ligámenes entre ventanas, rotaciones en 3 ó 4 dimensiones, con el fin de mostrar estructuras particulares.

1.1. Los datos analizados, notaciones y algunas definiciones

Los datos numéricos analizados se presentan bajo la forma de una tabla X (una matriz con n filas y p columnas). Las n filas son los "individuos" o unidades estadísticas. Las p columnas son las variables, es decir son las medidas hechas sobre los individuos, son los parámetros estudiados. Comúnmente n es muy superior a p, pero como no se hace aquí ninguna referencia al contexto probabilístico, nada impide que sea de manera diferente (más tarde, un valor pequeño de n podrá ser compensado por la observación de varias tablas en lugar de uno solo).

Las variables son vectores de $\mathbb{R}^n$. Este espacio está provisto de la métrica euclídea D. Se confunden en este curso las notaciones de las aplicaciones lineales con las de sus representaciones matriciales en las bases canónicas, aquí de $\mathbb{R}^p$ o de $\mathbb{R}^n$. La matriz D de dimensión n es diagonal, los términos de la diagonal, no nulos, son los pesos p^k asociados a los individuos: $\sum p^k = 1$. Lo más común es $p^k = 1/n$.

Los individuos son vectores de $\mathbb{R}^p$. Este espacio está provisto de la métrica euclídea M. A menudo $M = I_p$, matriz identidad $p \times p$. A veces se trata de una matriz diagonal por bloques. La pareja (X, M) puede denotar la *nube* de los individuos en $\mathbb{R}^p$, es decir el de puntos filas de X representado por M. La pareja (X, D) puede denotar la *estructura* de X, es decir el conjunto de las asociaciones entre las variables columnas de X. A veces se dice que el ACP consiste en hacer el análisis del *triplete estadístico* (X, M, D).

Un vector columna se denota u, y se denota por el vector transpuesto u' cuando es un vector fila. Cuando un vector *normado* u define un *eje* de representación en $(\mathbb{R}^p, M)$, la *componente* XMu correspondiente es el vector de $\mathbb{R}^n$ que contiene las coordenadas de los individuos de X proyectados sobre este eje. Una componente es una combinación lineal de las variables de X, los coeficientes de esta combinación se encuentran en el Mu; vista así, una componente es a veces llamada como una *variable latente*.

Si las columnas de la matriz U están constituidas de varios ejes de $\mathbb{R}^p$, las columnas de XMU constituyen la misma cantidad de componentes correspondientes.

Cuando todas las variables de X son centradas (de medias nulas), entonces X se dice centrado, el centro de gravedad de la nube de puntos-individuos está en el origen, las componentes XMU también están centradas. En lo que sigue las tablas se suponen centradas.

Cuando U corresponde en una base de vectores $\{u_i\}$ de $\mathbb{R}^p$, la suma de las inercias (de los momentos de inercia con respecto a O) de la nube (X, M) proyectada sobre los p ejes es igual a la inercia total de la nube, es decir $tr(X'DXM) = \sum_{i=1}^{p} var(XMu_i)$.

$X'DX$ es la *matriz de covarianzas* de las variables de X, y la de correlaciones cuando las variables de X son reducidas o estandarizadas. Cuando $M = I_p$, la cantidad $tr(X'DX) = \sum_{i=1}^{p} var(Xu_i)$, denota *varianza total* de X y entonces la descomposición de inercia corresponde también a una descomposición de la varianza total.

Denotamos *diag* la operación que consiste en extraer de una matriz cuadrada los términos que se encuentran sobre la diagonal. Se llama *contribuciones de las variables* de X (contribuciones a la variabilidad total) la lista de las varianzas de las variables dada por $diag(X'DX)$.

Llamamos *contribuciones de los individuos* la lista de las partes de inercias aportadas por cada $diag(XMX'D)$. Las contribuciones pueden permitir condensar la información contenida en una tabla X reordenando filas y columnas: el nuevo orden se obtiene ordenando los valores de las dos

listas en orden decreciente, haciendo seguir los nombres de las variables y los de los individuos. A veces, lo esencial de la información de la tabla puede condensado en la parte superior derecha. También se puede redondear (por ejemplo, a lo sumo un decimal más que los datos reducidos) y reemplazar los valores demasiado pequeños o nulos (por ejemplo, entre $-$,2 y $+$,2 para datos reducidos) por espacios en blanco (o guiones). Entonces, puede suceder que las filas y las columnas, que constituyen la parte no visible de una gran tabla, sean relativas a una información despreciable, mientras que la esquina superior izquierda de la primera página sea legible ya que contiene pocos valores.

Las contribuciones se expresan a menudo en forma de porcentaje (valores redondeados al entero), es decir por listas $100\, diag(X'DX)/tr(X'DX)$ y $100\, diag(XMX'D)/tr(XMX'D)$. Para los individuos, tomar por ejemplo 1000 en lugar de 100 cuando n es grande.

La inercia total es descomponible según una base M-ortonormada $\{u_i\}$ cualquier de $\mathbb{R}^p$. La contribución parcial de un individuo corresponde a su contribución relativa a un eje. Para un eje i, las *contribuciones parciales* de las filas son las contribuciones totales de las filas de la tabla estimada por proyección XMu_iu_i'

$$diag[XMu_iu_i'M(XMu_iu_i')'D] = diag(XMu_iu_i'MX'D),$$

o bien, denotando w_i la componente estandarizada asociada a XMu_i,

$$var(XMu_i)\, diag(w_iw_i'D).$$

Observaciones. Puede suceder que n y p sean muy grandes, al punto que el progrrama computacional y el equipo de cálculo consuman demasiado tiempo, impidiendo alnzarse en un proceso de análisis verdaderamente interactivo, o incluso al punto de sobrepasar la capacidad de almacenamiento. En muchos casos, esta situación proviene de la "torpeza" del estadístico: un número de variables muy grande significa a veces redundancia de la información, ausencia de un administrador de bases de datos, falta de uso de métodos de clasificación o de selección, etc.

Un número de individuos muy grande significa a veces que no se usan agrupamientos, clasificaciones, la noción de validación cruzada. Cuando se tiene más de 250 individuos, normalmente se extraen al azar 250; se somete este juego de datos al análisis; se comparan los resultados a los obtenidos sobre el mismo análisis efectuado sobre otra muestra de 250 individuos extraídos también al azar. La validez del análisis es confirmada si los dos conjuntos de resultados no difieren de manera significativa. La cifra 250 se da aquí a título indicativo, pero a menudo no se debe sobrepasar.

1.2. El análisis en componentes principales o ACP

1.2.1. Introducción

Bertin [2] analizaba una tabla X con el objetivo de poner en evidencia las relaciones que pueden existir entre las filas y las columnas. Se trata de definir simultáneamente paquetes de variables y subgrupos de individuos, ciertos subgrupos pudiendo ser particularmente explícitos para algunos paquetes y viceversa. Para hacer este trabajo, se servía del conocimiento profundo que tenía de los datos y de permutaciones, tanto sobre las filas como sobre las columnas. Realizando mutuamente estas permutaciones, cada valor se pone individualmente sobre un cartón, se servía de ganchos para enganchar ya sea una fila de cartones para mover, ya sea de una columna. Hoy este trabajo se hace fácilmente con una hoja electrónica, reemplazando antes los valores por espacios en blanco si son despreciables, o por símbolos $+++$, $++$, $+$, $-$, $--$ y $---$, para traducir los otros valores, desde los mayores hasta los más pequeños, de manera que se obtenga una tabla fácil para leer. A la izquierda de la tabla, se trata entonces de condensar, mediante permutaciones, la relación

fila-columna más fuerte. A la derecha permanece la que parece más débil. En general, la tabla que se obtiene al final es explícita, con más sentido que la obtenida por permutaciones hechas de acuerdo con la importancia de las contribuciones, pues se tienen dos objetivos simultáneamente en el cerebro del usuario de este método poco rápido: el objetivo de resumir y el de clasificar (variables e individuos).

Conociendo una tabla de datos observados X, el ACP de un triplete estadístico (X, M, D) puede corresponder a varias motivaciones y algunas de ellas están contenidas en la manera de plantear el problema de optimización que define el análisis. En lo que sigue, nos proponemos solamente de definir el ACP satisfaciendo un criterio de no redundancia de corte del ligamen entre la nube (X, M) y la estructura (X, D) (siguiendo de esta manera el espíritu de Bertin). Cuando M no es la métrica identidad, o cuando n ó p son grandes, el enfoque de Bertin no puede ser abordado.

Sin embargo, es raro hacer un ACP de los datos brutos, ya que un pretratamiento es en general necesario.

1.2.2. Pretratamiento de los datos

A menudo las variables constituyen un conjunto no homogéneo. Las escalas de medida son independientes entre una variable y la otra, y a veces son arbitrarias. Por ejemplo, se puede proponer el logaritmo de estas medidas tanto como las medidas mismas. Ahora bien, los resultados del ACP dependerán de las escalas. Para enfrentar esta situación, en particular para dar "pesos iguales" a las variables y obtener un centro de gravedad de la nube de individuos más o menos cerca del centro de simetría (de manera que sea natural de posicionarse con respecto a él), y también para que los cálculos de las varianzas sean resistentes de cara a valores demasiado excepcionales, podemos inspirarnos de las indicaciones dadas abajo durante el pretratamiento de los datos. Otras indicaciones, sobre todo acerca de la manera de estandarizar en análisis multitablas, pueden ser encontradas en Harshman & Lundy [16].

a) Se examina de manera gruesa la distribución de cada variable. El problema de las transformaciones se plantea sobretodo si el número de los individuos es bastante grande para que la noción de distribuión pueda ser precisada. Si las variables tienen distribuciones de aspecto muy diferente, se puede pensar en transformarlas de manera que se vuelvan estas distribuciones más o menos simétricas y así bastante análogas. De esta forma, las correlaciones entre variables que miden una analogía de disposiciones relativas de los individuos sean menos dependientes de la naturaleza diferente de estas distribuciones. Para hacer la transformación, se usan funciones monótonas, como las funciones potencia o logaritmo (transformaciones de Box & Cox o transformaciones de Tukey). Una distribución común, posible en el caso en que las distribuciones iniciales aparezcan muy diferentes siendo cada una con una ley conocida, es la distribución uniforme: las transformaciones son entonces los inversos de las funciones de distribución. Otra distribución común e interesante es la distribución normal o gaussiana, ya que los criterios de mínimos cuadrados del ACP pueden ser acompañados de justificaciones probabilísticas. Eventualmente, en caso de duda sobre la pertinencia de transformar, se pueden hacer 2 análisis: con transformación y sin ella.

b) Observar para cada variable si algunos valores verdaderamente excepcionales existen para algunos individuos. Puesto que las distribuciones son bastante parecidas después de la etapa a), la supresión eventual de individuos excepcionales es considerada de manera homogénea para todas las variables.

c) Centrar las variables. Lo más común, se divide cada variable por su desviación estándar. Es entonces sobre esta tabla centrada (y a menudo estandarizada), denotado siempre X, que se considera efectuar el ACP.

1.2.3. ACP del triplete (X, M, D)

Un problema que se plantea a menudo en ACP es el de la condensación de la información en subespacios de dimensión reducida. Varios criterios equivalentes pueden ser asociados a esta preocupación. Se puede definir un primer *eje principal* normado u en vista de:

- maximizar la dispersión de la nube (X, M) proyectada sobre este eje

$$f(u) = u'MX'DXMu = var(XMu).$$

- definir la componente (principal) XMu más correlacionada con las variables x_j de X, para revelar las covariaciones más intensas entre variables, las covariaciones de las variables con la componente implica covariaciones entre variables

$$g(u) = \sum cov^2(XMu,\, x_j) = u'M(X'DX)^2Mu.$$

El conjunto de soluciones sucesivas, obtenidas cuando se replantea la búsqueda de un nuevo vector normado colocándose en el subespacio ortogonal a las soluciones anteriores, constituyen los ejes principales y las componentes principales del ACP, y su número es igual a $r = rang(X)$. Se obtienen así mejores resúmenes sucesivos.

Además de definir resúmenes, otro objetivo constituye en definir u de tal forma que la pareja $(u,\, XMu)$ formada por vectores no nulos, es decir autónoma, pueda caracterizar una parte de lo que liga a los individuos con las variables, porque esta parte es sin redundancia con las otras partes. Esta no redundancia implica la definición de un número mínimo de parejas. Un explorador tiene a menudo sus propias preocupaciones, para una variable particular o para un subgrupo de individuos particulares, y son los planos del ACP (interesantes a priori debido a esta no redundancia) más cercanos de esta variable, o de este subgrupo, que entonces usará. Esto nos conduce a interesarnos a las parejas que verifican

$$\rho^2(XMu,\, XMu^*) = 0,\ \forall u^* \perp u.$$

De manera equivalente, podemos escribir

$$(I_p - uu'M)X'DXMu = 0$$

o lo que es lo mismo

$$X'DXMu = (u'MX'DXMu)\, u.$$

Todas las parejas principales del ACP de (X, M, D) son entonces las únicas soluciones. La familia de estas parejas descompone a relación entre filas y columnas, es decir la aplicación lineal de $(\mathbb{R}^p, M)$ en $(\mathbb{R}^n, D)$ representada por la matriz X, como lo indica el cálculo por descomposición en valores singulares de X (SVD de X), matriz de rango r,

$$X = C\Delta U' = \sum_{j=1}^{r} s_j c_j u_j'.$$

Las columnas de C, que verifican $C'DC = I_r$, son las componentes principales reducidas (estandarizadas) respectivamente asociadas a los ejes principales, columnas de U que verifican $U'MU = I_r$, y a los cuadrados de los r valores singulares s_j, iguales a las varianzas a las varianzas de las componentes XMU.

Nótese que esta última definición global de los términos del ACP. Hacemos notar que esta última definición global de los términos del ACP no implica un ordenamiento de las parejas. Pero implica desde el inicio dos sistemas de vectores ortogonales mientras que los criterios f ó g no implican

ortogonalidad (sería necesario definir criterios sucesivos definiendo las acumulaciones constituidas por las partes distintas gracias a la ortogonalidad).

La mayor parte de los algoritmos de diagonalización o de cálculo de SVD dan el conjunto de vectores propios o singulares en desorden. No es sino en un segundo momento que el ordenamiento de los valores propios o singulares se hace, para la representación usual de los resultados. Estos cálculos parecen estar más intrínsecamente ligados a un corte de la información por dimensiones generadas por los vectores, que a los valores propios o singulares.

1.2.4. Biplots

Clásicamente en ACP, la representación de las variables se hace proyectándolas (ortogonalmente en el sentido de D, las variables son estandarizadas si se quiere trazar los círculos de correlaciones) sobre los planos generados por las parejas de ejes definidos por dos componentes principales reducidas. Pero revelar las correlaciones entre variables de esta manera no es siempre práctico y la lectura de la matriz de las correlaciones (o de covarianzas) se vuelve más precisa, y más fácil si uno da una representación simplificada por la codificación de los valores y permutaciones de variables.

La representación usual de los individuos se hace por proyección en el sentido de M sobre los subespacios generados por 2, 3, incluso 4, ejes principales. Cuando se usan dos ejes, se yuxtaponen a menudo la representación de los individuos con las de las variables proyetadas sobre el plano formado por componentes correspondientes. La lectura es entonces simultánea, los individuos colocados en una dirección que pasa por O y que están asociados a las variables muy correlacionadas con la misma dirección que pasa por O considerada en la representación de las variables.

Sin embargo, existe otra manera de proceder, más justificada que esta última cuando se desea realizar esta simultaneidad. Se trata de los biplots de Gabriel [10] y [12], que consisten en hacer una representación superpuesta de los individuos y de las variables de manera que se revelen la relación mutua que los individuos tienen con las variables (para evitar la sobrecarga sobre un mismo gráfico, a veces se yuxtaopnen en lugar de sobreponerlos). Es un gráfico adaptado al espíritu mismo de la última definición dada del ACP. Trataremos de proponer solamente planos que no contengan información redundante con respecto a esta relación mutua. Es decir, por ejemplo, una vez fijada una clasificación de los ejes, se prefiere mirar el planos de los ejes 3-4 después del plano 1-2, en lugar de los planos 1-3 y 1-4. Un biplot se realiza a menudo en dos dimensiones, pero a veces se realiza en tres [30].

Los biplots más usuales en ACP corresponden a la representación usual de los individuos, las direcciones asociadas a las variables no son obtenidas por proyección (coordenadas en $U\Delta$), sino el trazo de sus cosenos directores (coordenadas en U). La última definición dada de los ejes del ACP, que no tiene por objetivo resumir sino cortar, es independiente de ciertas transformaciones de X por similitudes, y así se puede desear que la interpretación de los ejes por las variables no dependa tampoco de ellas. Esto implica una representación de las direcciones de las variables independientes de los valores singulares. La interpretación de los ejes por las variables, realizada via los biplots usuales, permite entonces esta independencia.

Sea U_{ij} la matriz $p \times 2$ de las columnas u_i y u_j de U. El biplot asociado a estos dos ejes corresponde a dar una representación exacta de los valores de la matriz

$$B = XMU_{ij}U'_{ij}.$$

La dirección de la variable k set define por la recta que pasa por el origen O y el punto v_k de $\mathbb{R}^2$, columna k de U'_{ij}. El individuo l se representa por los puntos i_l' de $\mathbb{R}^2$, línea l de XMU_{ij}. Los vectores v_k y i_l de $\mathbb{R}^2$ se escriben superpuestos en una misma base I-ortonormada, se tiene entonces el elemento b_{lk} de B igual al producto escalar

$$b_{lk} = (i_l,\, v_k).$$

Es por ello que la importancia de la distancia al origen del proyectado ortogonalmente (en el sentido usual (del individuo l sobre la dirección k que vale b_{lk}, indica la importancia de esta variable para este individuo (en el subespacio considerado).

Se añade al gráfico la información sobre las contribuciones de las variables. Las longitudes de las flechas se definen por las raíces cuadradas de las sumas de las dos participaciones parciales relativas a las dos dimensiones. La lista de estas sumas constituye la diagonal de una matriz simétrica semidefinida positiva

$$diag(X'DB).$$

En efecto, en [19] la definición de las *participaciones parciales* de las variables por un eje principal i se propone para medir las intensidades de ligamen entre las variables de X y la nube proyectada sobre este eje. Así, se les llama participaciones (en lugar de contribuciones), se prefiere decir participaciones de las variables de X al estimado por proyección XMu_iu_i') y se definen por la lista de valores

$$diag(X'DXMu_iu_i') = s_i^2\, diag(u_iu_i').$$

Estas participaciones descomponen $tr(X'DX)$, puesto que la suma de todos los proyectores parciales $u_iu_i'M$ reconstituye la matriz identidad cuando $r = p$. Cuando $M = I_p$, la suma de las participaciones de las variables iguala a la suma de las contribuciones de los individuos, relativamente a cada eje (y entonces esta definción de las participaciones parciales, coincide con la definición usual de las contribuciones parciales de las variables). El análisis de la nube se convierte en el análisis de la estructura.

Una flecha orientada de O hacia el punto define el coseno director de una variable, indica la dirección de los valores positivos de esta variable.

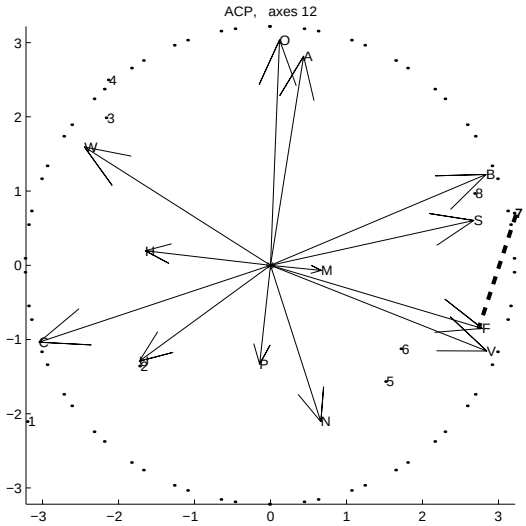
Una misma escala se conserva para todos los ejes, y así una misma escala se conserva para todas las longitudes de las flechas de todos los biplots.

Nota: Los datos, así como los que se consideran en lo que sigue, provienen del Instituto de Ciencias del Mar ISMER, Universidad de Quebec en Rimouki (UQAR) [5].

En este ejemplo, se miden 13 variables un día dado en 8 estanques. Este biplot de ACP de $(X, I_8, D_{1/13})$ *contiene 64 % de la información sobre la estructura de los datos. Permite una comparación de los estanques y evaluar el rol de ciertas variables en esta compración. El círculo trazado no es un "círculo de correlaciones", sino una referencia trazada para faciliar la lectura (el ojo no posee la misma objetividad en el eje de una diagonal que en el de una vertical). Todas las variables han sido reducidas o estandarizadas, las asociadas a las flechas más cortas son las peor representadas en este plano. La variable "C" está casi totalmente representada (esto ha sido verificado aparte en un diagrama de bastones), así entonces las variables que tienen una flecha de longitud similar. Se evita considerar las proyecciones de los números de estanque sobre las direcciones de las variables mal representadas, cuando se quiere sacar de la sola lectura del plano de los ejes 1-2 conclusiones definitivas sobre los datos mismos.*

El proyectado del estanque 7 sobre la variable "F" está bastante alejado del origen en el sentido indicado por la flecha, lo cual revela un gran valor de "F" para este estanque (comparativamente al conjunto de los estanques). Así mismo, un pequeño valor de "W" existe para este estanque, puesto que el proyectado sobre la dirección "W" (mentalmente prolongada) también está alejada del origen. De una forma más general, la variable "F" opone los 4 primeros estanques a los 4 siguientes y esto resume bastante bien de qué manera esta variable participa en el análisis.

El proyectado de 7 sobre "O" es bastante cercano del origen, lo que indica un valor de esta variable próximo de la media para el estanque 7. Para una variable mal representada como "M", la tabla de las participaciones por eje permite designar un subconjunto minimal de ejes suficientes para englobar lo esencial de su participación. Si hicieran falta 3 ejes, un biplot 3D permitiría proponer por rotación el mejor plano de visualizatión de esta variable y entonces de los estanques que contribuyen.



1.2.5. Observaciones

Los ejes tienen orientaciones arbitrariamente escogidas: un programa calcula los vectores propios normados (de norma 1) y da tanto la solución u_i como la solución -u_i. Una mirada demasiado rápida podría llevar al lector a difarenciar gráficos que no deben serlo. Más generalmente, la representación de los individuos es independiente de toda isometría R' que se podría aplicar a los individuos antes del análisis: el ACP de (X, I, D) da la misma representación que el de (XR, I, D). En efecto, una configuración de n puntos no cambia la naturaleza por rotación. Pero sobre biplots las representaciones de las columnas respectivas no serán idénticas, unas son columnas de X, otras de XR, de manera que se obtiene así una lectura de lo que diferencia las dos tablas. En ACP no se miran los valores de las tablas, sino las disposiciones relativas de individuos que producen estos valores. Ahora bien, el regreso a los valores inciales se produce en una representación biplot por la representación de las variables.

Cuando $M = I_p$, el análisis de la nube de individuos es una manera de llevar el estudio de

la estructura, asÍ como el análisis de los ligámenes entre variables e individuos. Toda base I_p-ortonormada de $\mathbb{R}^p$ permite abordar toda la estructura de X, y así todo sistema otonormado de rango r que genera el subespacio generado por los ejes principales. El sistema de los ejes del ACP de (X, I_p, D) es el único que permite el estudio sin redundancia de la estructura.

La visualización de los individuos se enriquece fuertemente en una representación 3D en lugar de una 2D, con colores diferentes para los subgrupos provenientes de una clasificación de los individuos, o provenietes de un corte, por ejemplo por cuantiles sobre una variable suplementaria, o añadiendo ciertos puntos sucesivos por líneas...

¿Cual es la escogencia de la métrica M? A menudo M es diagonal y luego de una transformación correspondiente de los datos, se podría deducir que es inútil considerar otras métricas aparte de la métrica identidad. Así, por ejemplo, para la métrica diagonal de Joreskog en la que se divide cada variable por su desviación estándar multiplicada por la raíz cuadrada de "1 menos el cuadrado del coeficiente de correlación múltiple de la variable con todas las otras variables". Sin embargo, es la comparación entre resultados provenientes de diferentes ACP obtenidos cuando M cambia y realizados guardando siempre la misma tabla X (es decir las mismas variables), que puede ser deseada. El caso más clásico de métrica no diagonal es cuando M es diagonal por bloques, asociada a una partición del conjunto de variables en K subconjuntos. En particular, los bloques diagonales podrían ser las métricas de Mahalanobis respectivas de subconjuntos de la partición (las inversas de las matrices de covarianzas).

Sobre todo cuando los términos de la diagonal de D son iguales, un estudio que haga intervenir un triplete (X, I_p, D) puede inducir otro estudio, variante del anterior, habiendo reemplazado el triplete estadístico por el triplete (X, I_p, I_n). Otros consideran el procedimiento recíporco, que evita entonces introducir las métricas en sus notaciones. Además una propuesta de análisis relativos a las filas de X puede constituir entonces también una proposición de análisis para las columnas, y recíprocamente. Es así por ejemplo que ellos consideran de buena gana el "producto cruzado" $X'X$ que la matriz de covarianzas $X'DX$. Sin embargo, este intercambio de los papeles no es siempre evidente, por ejemplo cuando los individuos son considerados como independientes, con X observación de un vector aleatorio multivariado.

2. Análisis de la dependencia de dos nubes

2.1. Introducción

Sean X $n \times p$ y Y $n \times q$ dos matrices. Los conjuntos respectivos de p y q columnas definen dos conjuntos de variables centradas medidas sobre un mismo conjunto de n individuos.

Se trata de analizar la dependencia de los tripletes estadísticos (X, M, D) y (Y, N, D), más precisamente de un estudio de la dependencia entre las nubes (X, M) y (Y, N).

Para obtener el análisis, se descompone la dependencia global en dependencias parciales no redundantes entre componentes. Las escogencias más usuales de métricas conducen a la definición de los análisis factoriales de las 2 tablas más clásicas.

2.2. Definición de las parejas de componenes monógamas

Entre todas las componentes que se pueden pensar para poner en evidencia la dependencia de (X, M) con (Y, N) se definen las que forman parejas (XMa_i, YNb_i) tales que la componente XMa_i se asocia particularmente con la componente YNb_i, pues los ligámenes de correlación entre XMa_i y las variables de Y y aquéllos entre YNb_i y las variables de X se reducen a una única correlación no nula: aquélla entre XMa_i y YNb_i.

Se dice que una pareja (XMa_i, YNb_i) forma *componentes monógamas* cuando la correlación de la componente YNb_i con toda posible componente diferente de XMa_i es nula, y cuando la correlación de la componente XMa_i con toda componente posible diferente de YNb_i es nula también, la correlación no nula de las dos componentes de la pareja que asocia dos partes de inercia, una al interior de $(\mathbb{R}^p, M)$ y otra de $(\mathbb{R}^q, N)$. Más precisamente, en este análisis de dependencia se definen parejas (a_i, b_i) de $\mathbb{R}^p \times \mathbb{R}^q$ que verifican

$$\rho^2(XMa_i, YNb^*) = 0, \ \forall\, b^* \perp b_i,$$

$$\rho^2(XMa^*, YNb_i) = 0, \ \forall\, a^* \perp a_i,$$

$$\rho^2(XMa_i, YNb_i) \neq 0,$$

o también

$$X'DYNb_i = (a_i' MX'DYNb_i)\, a_i,$$

$$Y'DXMa_i = (a_i' MX'DYNb_i)\, b_i,$$

$$a_i' MX'DYNb_i \neq 0,$$

de manera que, siendo r el rango de $X'DY$, la familia de las parejas de ejes $(a_i,\, b_i)$ que son soluciones es la de los r vectores singulares de la matriz $X'DY$ considerada como aplicación lineal de $(\mathbb{R}^q, N)$ en $(\mathbb{R}^p, M)$

$$X'DY = A\Delta B', \ \ A'MA = B'NB = I_r.$$

Salvo por el orden de multiplicidad de los valores singulares y los signos de éstos, una pareja $(a_i,\, b_i)$ puede ser reemplzada por una pareja $(-a_i, -b_i)$, las parejas soluciones sin redundancia están entonces dadas como provenientes de una descomposición de ese tipo. La definición 2.2 no está asociada a la optimización de un criterio numérico y entonces no implica orden en el ordenamiento de las parejas de soluciones.

El sistema $\{XMa_i\}$ es en general formado por las componentes correlacionadas, como también lo es el sistema $\{YNb_i\}$. Esto quiere decir, en el caso de métricas identidad, que cada uno de los sistemas no permite hacer un estudio sin redundancia de las dos estructuras respectivas, puesto que solo los ejes de un ACP lo permiten. Sin embargo, estos dos sistemas permiten subir cada una de las dos partes de estructuras implicadas en la dependencia, siendo esta última analizada sin redundancia. Y, por ejemplo, cuando $r = p$, toda la información interna a X es cubierta, puesto que entonces

$$\sum_{i=1}^{r} XMa_i a_i' = X.$$

Observación: La definición anterior es a la que nos referiremos más frecuentemente en lo que sigue. La definición siguiente está basada en un criterio numérico que define una clasificación a priori de las parejas, y esta clasificación corresponde al general dado por los programas computacionales. Las parejas monógamas se definen más frecuentemente como soluciones sucesivas del problema de optimización, bajo restricciones respectivas de ortonormalidad, del criterio f

$$f(a, b) = cov(XMa, YNb),$$

o incluso del criterio g

$$g(a, b) = cov^2(XMa, YNb),$$

los óptimos sucesivos de f constituyen los r valores singulares anteriores ordenados en orden decreciente.

Claramente, la definición global 2.2 tiene por objeto asociar pares de partes de inercia distintas, de manera que no se tenga que justificar la ortogonalidad de cada uno de los dos sistemas de ejes.

Con el criterio f nos podemos preguntar por qué decidir sobre soluciones sucesivas con ejes ortogonales. Dicho de otro modo, por qué no soluciones sucesivas con, por ejemplo, componentes XMa_i de correlaciones nulas dos a dos, como en regresión PLS [31].

Más interesante a considerar para justificar las soluciones de la definición 2.2 es el criterio g, pues la noción de sumas del criterio asociadas a las ortogonalidades consideradas corresponde a soluciones encajadas, provenientes de la diagonlización de una matriz simétrica semidefinida positiva.

2.3. Ligamen con los análisis factoriales clásicos

Las componentes principales de un triplete estadístico (X, M, D) son definidas como componentes endógamas, siendo las únicas que no están correlacionadas con ellas mismas descomponiendo la inercia total de X en partes distintas. Es así con $(Y, N, D) = (X, M, D)$, las dos componentes de una pareja monógama indentificándose entonces con una componente principal de ACP.

Definida con estas métricas cualesquiera a partir del criterio f, el análisis ha sido llamado análisis de coinercia [8], lo que indica que se trata en el caso general del análisis entre nubes de individuos. Con las métricas identidad, se le llama análisis procrusto, o *análisis de Tucker* [29], de Cliff [7], de Green [11]. Con las métricas de Mahalanobis, *análisis canónico* (clásico). Con una de Mahalanobis y otra identidad, se le llama: análisis predictivo, *ACPVI* de Rao [26]-Johansson [17], ACP bajo restricciones...

El hecho de haber centrado las dos tablas, y por lo tanto de tener un centro de gravedad común para las dos nubes de individuos, corresponde a un ajuste por mínimos cuadrados entre dos conjuntos de individuos apareados, realizado por traslación. De esta manera, las parejas monógamas son independientes de las traslaciones.

Así mismo, estas parejas son independientes de toda isometría realizada en $(\mathbb{R}^p, M)$ o en$(\mathbb{R}^q, N)$. Es más, la definición de los ejes es independiente de afinidades ortogonales que se hubieran realizado en todas las direcciones de estos ejes (solo los valores singulares serían modificados). En la definición 2.2, las correlaciones entre las dos componentes de parejas son independientes de todas estas transformaciones y constituyen las medidas más intrínsecas para caracterizar el ligamen de dos nubes de individuos. El análisis del ligamen entre 2 nubes se fundamenta en estas correlaciones, se trata de un análisis de dependencia.

El análisis de Tucker (métricas identidad) aparece como análisis de la dependencia de 2 estructuras, puesnubes de individuos o conjuntos de variables definen, entonces, la misma información. Este análisis corresponde también a un problema de ajuste por rotación ortogonal de la nube de individuos: tomando X como objetivo, la rotación procrusta R' de $(\mathbb{R}^q, I)$ en $(\mathbb{R}^p, I)$ es definida por

$$R = BA' = Y'DX(X'DYY'DX)^{-1/2}.$$

Ella asocia el sistema I-ortogonal B al sistema I-ortogonal A, lo que explica el nombre "rotación ortogonal". Las coordenadas de las filas de X en el sistema de coordenadas A están en la matriz XA mientras que las de la tabla transformada YR expresadas en este sistema de coordenadas común A se encuentran en YB. Nótese así que las representaciones superpuestas de las dos nubes no se modificarían al cambiar el objetivo (representando las filas de XR' y de Y, en el sistema de coordenadas común B).

Cuando $q \leq p$, se puede completar respectivamente los sistemas B y A por $q - r$ vectores I-ortonormados, para obtener los sistemas B_1 y A_1. El sistema B_1 es entonces una base de $(\mathbb{R}^q, I)$. Con $R_1 = B_1 A_1'$ se obtienen todas las soluciones del problema de búsqueda de isometrías R_1' sobre $(\mathbb{R}^q, I)$ que minimizan en el sentido de mínimos la distancia entre individuos apareados según el criterio

$$f(R_1') = d^2(X, YR_1) = tr(X'DX + Y'DY - 2X'DYR_1),$$

y entonces que maximizan $tr(X'DYR_1)$, para dar el óptimo $tr[(X'DYY'DX)^{1/2}]$.

En cuanto al análisis canónico permite analizar la dependencia de dos subespacios, los generados por las variables. Trata sobre todo cuando la noción misma de estructura de tablas importa poco. Por ejemplo, se quiere comparar dos gamas de productos que compiten probados sobre los mismos individuos. Si no hay ningún sentido en la búsqueda de una analogía entre "lo que liga entre ellos los productos de una de las gamas" y "lo que liga entre ellos los productos del otro", entonces es mejor hacer un análisis canónico para analizar el ligamen de las dos gamas.

Las componentes del análisis de Tucker de dos conjuntos de variables que no definirían más que dos subespacios y nada más, es decir tales que cada uno de ellos esté constituido de variables reducidas no correlacionadas (pueden definir así dos proyectores de $(\mathbb{R}^n, D)$), son las componentes del análisis canónico de estos dos conjuntos [28]. Así se puede obtener el análisis canónico de dos tablas X y Y realizando el análisis de Tucker de los dos conjuntos de variables sintéticas constituidas de los dos juegos respectivos de componentes principales reducidas.

El análisis canónico es sin duda el análisis más utilizado en la medida en que, formando tablas de variables indicadoras de modalidades, se convierte en *análisis discriminante* (una de las dos tablas es la de las indicadoras de la pertenencia a los diferentes subgrupos), o *análisis de correspondencias* (con las dos tablas de indicadoras de modalidades de dos variables cualitativas). Estos análisis a menudo son objeto de un programa particular, frecuentemente más rápido de ejecución y con desarrollo que les son propios.

En lo que sigue, se desarrolla el análisis de dependencia basándose en lo que tiene de más intrínseco: los ejes del análisis, las correlaciones de las componentes monógamas, el criterio del cuadrado de la covarianza. El enfoque hace posible una evaluación de lo que separa el análisis de Tucker, análisis canónico, y ACPVI.

Cuando se yuxtapone o superpone la nube de filas de X proyectadas sobre el plano de los ejes $\{a_i, a_j\}$ a la nube de filas de Y proyectadas sobre el plano de los ejes $\{b_i, b_j\}$, se ponen en evidencia ciertas diferencias y similitudes, pero sin saber de hecho en cual medida las inercias representadas contribuyen a la dependencia (aún cuando todos los ejes a_i y todos los ejes b_i generan respectivamente los espacios $\mathbb{R}^p = \mathbb{R}^q = \mathbb{R}^r$).

Para obtener un análisis de dependencia entre estructuras, se puede desear que las partes de inercia representadas coincidan con las partes de varianzas de las tablas que se implican en la dependencia. Para alcanzar este objectivo, se propone en lo que sigue gráficos realizados proyectando otros "individuos" diferentes de las filas de las tablas X y Y.

2.4. Definición de las contribuciones de los individuos a la dependencia

2.4.1. Contribuciones a la correlación de dos variables

Se describe en esta sección una manera de definir y representar las contribuciones de los individuos, contribuciones con un cuadrado de correlación lineal entre dos variables. El procedimiento será en lo que sigue extendido al caso de dos conjuntos de variables.

Sean x y y dos variables centradas medidas sobre los mismos n individuos, de correlación lineal $\rho(x, y)$. Cada una de las dos variables ha contribuido a su manera al valor de esta medida de ligamen y se quiere evaluar para cada una las contribuciones de los individuos. Se consideran así las dos regresiones lineales simples, la de y en x, y la de x en y

$$y_i = \widehat{a}x_i + e_i,\ \forall i = 1 \cdots n,$$

$$x_i = \widehat{b}y_i + f_i,\ \forall i = 1 \cdots n.$$

Las dos igualdades siguientes detallan cómo los individuos contribuyen al cuadrado de la correlación a partir de los valores (nombrados contribuciones) que descomponen cada una de las dos partes de

inercia implicadas en el ligamen

$$\rho^2 = \frac{\sum (\widehat{a}x_i)^2}{var(y)} = \frac{\sum (\widehat{b}y_i)^2}{var(x)}.$$

Así, para visualizar sobre un eje las contribuciones absolutas de y, se toman los valores $\widehat{a}x_i$. Para comparar, sobre otro eje paralelo y orientado en el mismo sentido, las contribuciones de x tomadas corresponden a los valores $\widehat{b}y_i$ o $-\widehat{b}y_i$, según el signo de ρ.

De hecho, para tomar en cuenta el signo, en el contexto más general que sigue, dos vectores $\overrightarrow{a}$ y $\overrightarrow{b}$ son calculados. Los valores $\widehat{a}x_i$ están así en una dirección referenciada por un vector normado $\overrightarrow{a}$ de $\mathbb{R}$ y los valores $\widehat{b}y_i$ en una dirección paralela referenciada por un vector normado $\overrightarrow{b}$ de $\mathbb{R}$. Cuando la correlación es positiva $< \overrightarrow{a}, \overrightarrow{b} >= 1$, y cuando la correlación es negativa, $< \overrightarrow{a}, \overrightarrow{b} >= -1$.

Interesarse simultáneamente a los dos sistemas de contribuciones, es interesarse tanto a la suma $\rho^2 var(y)$ como a la suma $\rho^2 var(x)$. Entonces la cantidad $cov^2(x, y)$ constituye una medida compromiso de los dos sistemas de contribuciones. De allí viene nuestra interpretación del criterio g.

2.4.2. Contribuciones a la dependencia de dos nubes

Se quiere caracterizar las contribuciones de la nube (Y, N) a la dependencia, dependencia considerada con la nube (X, M). La caracterización se hace por una matriz de misma dimensión que la matriz Y, matriz ajustada a Y a partir de X, para evaluar la inercia de Y implicada en la dependencia.

Cliff [7] ha propuesto el cálculo de una imagen (de X), ajustada a Y en el sentido de mínimos cuadrados, que es la solución procrusta $XMAB'$. Sin embargo el criterio a optimizar no es una medida de dependencia que se aparenta a un cuadrado de correlación. Es más, el valor del criterio a la solución depende de afinidades ortogonales hechas sobre X de manera que su imagen no pueda caracterizar contribuciones de Y a la dependencia.

Así se considera, antes que la transformación BA', la familia de las aplicaciones lineales $B\Delta_c A'$ de $(\mathbb{R}^p, M)$ en $(\mathbb{R}^q, N)$, donde Δ_c es matriz diagonal definida positiva conteniendo los coeficientes de las afinidades posibles. La imagen Z a determinar se escribe

$$Z = XMA\Delta_c B',$$

y se plantea entonces el problema de la optimización del coseno (en el sentido N y D)

$$f(\Delta_c) = \cos^2(Y, Z) = \frac{tr^2[(YNB)'D(XMA)\Delta_c]}{tr(Y'DYN)\, tr[(XMA)'D(XMA)\Delta_c^2]} = \frac{tr^2(\Delta\Delta_c)}{tr(Y'DYN)\, tr(\Delta_X\Delta_c^2)},$$

nombrando Δ_X la matriz diagonal de las varianzas de las componentes XMA.

Sea v el vector desconocido cuyas componentes son los elementos diagonales de Δ_c y u el vector cuyas componentes son los elementos diagonales de Δ. El criterio puede escribirse entonces con dos productos escalares

$$f(v) = \frac{(u,\, v)^2}{(v,\, v)_{\Delta_X}}$$

Se plantea $\widetilde{v} = \Delta_X^{\frac{1}{2}} v$. Bajo restricciones de norma, el problema se reduce a maximizar el criterio

$$g(\widetilde{v}) = \widetilde{v}'\Delta_X^{-\frac{1}{2}} uu' \Delta_X^{-\frac{1}{2}} \widetilde{v} = (u'\Delta_X^{-\frac{1}{2}} \widetilde{v})^2$$

Entonces, una solución es obtenida para

$$\widetilde{v} = \Delta_X^{-\frac{1}{2}} u / \left\| \Delta_X^{-\frac{1}{2}} u \right\|.$$

Finalmente, $v = \Delta_X^{-1} u / \left\| \Delta_X^{-\frac{1}{2}} u \right\|$ y como el problema es invariante por homotecia, $v = \Delta_X^{-1} u$ es también una solución, de manera que $\Delta_c = \Delta_X^{-1} \Delta = \Delta_a$. La matriz Δ_a contiene así en la diagonal los coeficientes de las regresiones simples de los YNb_j sobre los XMa_j respectivos. Se obtiene la imagen solución, llamada *imagen concordante*

$$Y_C = XMA\Delta_a B'.$$

Así mismo, la imagen concordante X_C definida para evaluar las contribuciones a la dependencia de las filas de X, vale

$$X_C = YNB\Delta_b A'.$$

El índice global correspondiente al óptimo, llamado LAI (como *linear agreement index*) de X con respecto a Y

$$\begin{aligned} LAI[(X, MD)/(Y, N, D)] &= \cos^2(X, X_C) \\ &= \frac{tr^2(X'DX_CM)}{tr(X'DXM)\, tr(X_C'DX_CM)} \\ &= \frac{tr(X'DX_CM)}{tr(X'DXM)} = \frac{tr(X_C'DX_CM)}{tr(X'DXM)}, \end{aligned}$$

mide el acuerdo de X con la nube de sus contribuciones a la dependencia, por una parte de inercia de X implicada en la dependencia con Y. En la última forma, este índice hace pensar en la noción de cociente de correlación, en la medida en que, cuando M es identidad, el numerado es suma de varianzas explicadas y el denominador es una varianza total.

2.4.3. Cuadriplots de concordancia

Se trata de dos biplots yuxtapuestos para leer las relaciones de dependencia.

Para empezar se yuxtaponen los ejes apareados. Así un sistema de coordenadas $\{a_i, a_j\}$ se yuxtapone al sistema de coordenadas $\{b_i, b_j\}$. Entonces se trazan en estos sistemas respectivos las coordenadas de las filas de X_C y las de las filas de Y_C. Estas coordenadas se encuentran en las matrices $YNB\Delta_b$ y $XMA\Delta_a$, respectivamente. La yuxtaposición se justifica pues concierne a dos sistemas de contribuciones a una dependencia común, definida por las correlaciones entre componentes monógamas $\rho^2(YNb_i, XMa_i)$ y $\rho^2(YNb_j, XMa_j)$. Cuando las dos correlaciones están muy cercanas a 1, la dependencia calculada es muy fuerte, y las dos nubes se parecen mucho pues los individuos están en disposiciones relativas bastante análogas.

Como sea, habiendo seguido tanto como sea posible una analogía de misma naturaleza con respecto a los individuos, la comparación se prolonga mostrando cómo las variables se sitúan respecto a estas nubes. Ahora bien, la asociación de los dos ejes i y la de los dos ejes j conduce a poder asociar una direccíon cualquiera atravesando una de las dos nubes con la misma dirección y atravesando la otra nube de contribuciones.

La dependencia obtenida por las correlaciones entre partes de inercia sirve a evaluar una analogía de valores de las tablas (de estructuras) cuando las métricas son métricas identidad, con un control regresando a los datos iniciales. Para obtener este control, a cada nube de contribuciones se superpose una representación de las variables, poniendo así en relación las variables de X y de Y con las nubes.

Las coordenadas de las filas de X_C se ponen en relación con los cosenos directores de las variables de X que se encuentran en la matriz A. Es decir que nos interesamos en la descomposición de X_C en producto de 2 matrices

$$X_C = (YNB\Delta_b)\, A',$$

y más precisamente para un sistema de coordenadas de ejes i y j, al producto de las dos columnas respectivas de $YNB\Delta_b$ por las dos filas respectivas de A', que llevan a una representación exacta de los términos de la matriz $(X_C)_{ij}$. Efectivamente, por ejemplo para el eje i, se tiene

$$X'DX_CMa_ia_i' = \rho^2(YNb_i, XMa_i)var(XMa_i)a_ia_i'.$$

Los cosenos directores corresponden a las columnas de A, como lo indican las coordenadas de las variables de X sobre el eje del vector X_CMa_i. Tomando los términos de la diagonal de la matriz positiva de arriba, se definen las participaciones parciales positivas para las variables columnas de X. La suma de las dos participaciones parciales de una variable asociada a los dos ejes i y j, define el cuadrado de la longitud de esta variable sobre el biplot.

Cuando $M = I_p$ y para un eje dado, la suma de las participaciones parciales

$$tr(X'DX_Ca_ia_i') = \rho^2(YNb_i, Xa_i)var(Xa_i) = a_i'X_C'DX_Ca_i.$$

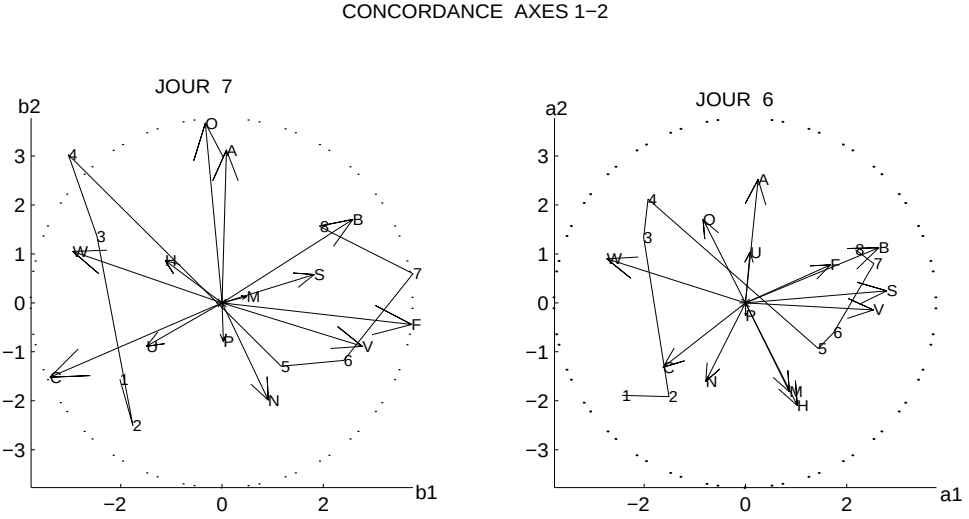
representa una parte de la variabilidad de X igual a la suma de las contribuciones de los individuos de X a la dependencia. Cuando además $N = I_q$, es la dependencia de las dos estructuras que se descompone por ejes.

Finalmente, se yuxtapone a este biplot el relativo a la misma pareja de ejes y a una descomposición de la imagen concordante Y_C análoga a la de X_C. En el caso de métricas identidades, la lectura revela diferencias y similitudes entre las dos tablas, estando estas dos nociones imbricadas.

Observaciones: El análisis se sigue en concordancia siguiendo la analogía de las dos nubes de contribuciones. La clasificación de los ejes debería entonces hacerse según las contribuciones totales por eje. Sin embargo, dos totales existen para cada eje, según que se miren las contribuciones de X o las de Y. Se hace un compromiso para una clasificación de los ejes a priori tomando la clasificación qeu se obtiene del criterio $g = cov^2(XMa_i, YNb_i)$.

Otras analogÍas entre tablas han sido propuestas para dos tablas o más, sobre todo cuando las mismas variables sirven a definir cada tabla. Entre dos nubes de individuos para los métodos procrustos (GPR), con un sistema de componentes comunes provenientes del mismo sistema de ejes producto a partir de una tabla (PCW) o provenientes de dos sistemas análogos de correlaciones entre variables y componentes producto desde una tabla (PCS), o sus variantes producidas desde el conjunto de las tablas (SCA-W, SCA-S), o a partir de un sistema de ejes modelo (SCA-P). (ver especialmente H. Kiers & J. Ten Berge, 1994 [18]). Algoritmos particulares, que tratan de obtener primeras componentes con fuerte varianza, se implementan.

2.4.4. Ejemplo



El análisis de dependencia de las dos estructuras permite el estudio de las analogías y diferencias entre el día 6 y el día 7. Los dos biplots anteriores corresponden a 13 variables medidas en 8 estanques los días 6 y 7, suponiendo que 7 está asociado al presente. Cada variable ha sido dividida por una desviación estándar media, raíz cuadrada de la media de las 2 varianzas calculadas los días 6 y 7. Así una variable tiene el mismo peso que otra en el análisis, pero sus dos representaciones conservan la misma relación inicial entre las dos varianzas. Las mismas escalas han sido utilizadas sobre los biplots, tanto para las contribuciones de los individuos como para las participaciones de las variables. Se ve que las variables que han sido implicadas en la analogía han aumentado globalmente en variabilidad el día 7. Esta visión global, perceptible por la dilatación de la nube de los estanques, no subraya los cambios importantes variable por variable, que se producen. Estos últimos cambios se revelan del hecho que los 2 conjuntos de variables son calcados sobre las 2 nubes casi idénticas, por construcción. Se ve entonces lo que ha cambiado entre las dos tablas, no encontrando las variables en la misma posición con la misma longitud.

Más de 90 % de variabilidad de cada una de las dos tablas se implica en la dependencia de

estructuras. Más 75 % de la contribución a la dependencia está cubierta por los primeros 3 ejes, para cada día. Sin embargo, es solamente el plano 1-2 que debe ser privilegiado aquí para la lectura de lo que difiere o no. En efecto, los estanques 1 y 2 constituyen una repetición de dos experimentos efectuados en exactemente las mismas condiciones, y a priori no deberían diferenciarse. Así mismo para los estanques 3 y 4, 5 y 6, 7 y 8, de manera que de hecho solo 4 condiciones experimentales diferentes se comparan. Ahora bien el tercer eje es debido tanto a diferencias entre los estanques 1 y 2 o 3 y 4 que entre los estanques 2 y 3 o 6 y 7, es decir tanto por diferencias sintomáticas de un "ruido" como a condiciones experimentales diferentes. Así el tercer eje parece poco creíble para juzgar sobre las diferencias entre los 4 experimentos, contrariamente a los dos primeros en que los agrupamientos por parejas de experimentos repetidos están mejor marcadas.

Notamos sobre el plano 1-2 que el aumento en variabilidad del día 7 proviene sobre todo de desviaciones grandes entre repeticiones. Estas desviaciones se sitúan más bien a lo largo del 1 para los cuatro últimos estanques. La variación del día 7 de silicato ("S") que se encuentra más o menos en esta dirección de las dos variables.

Este fenómeno, pues ella disminuye en lugar de aumentar en longitud y esto revela entonces una disminución marcada de silicato en los últimos estanques (o a un aumento en los primeros). En cambio, la ausencia relativa de cilles ("C") en los últimos estanques puede parecer más importante el día 7, pero esto puede ser debido sobre todo a las desviaciones entre repeticiones.

La variable "F" que designaba claramente los estanques 7 y 8 como los que contenían más flagelos, parece que ahora opone los 4 primeros estanques a los otros 4.

De manera general, los cuatro experimentos difieren según una dosis creciente de rayos ultravioletas, de los primeros estanques a los últimos. Enconces lo que opone los primeros a los últimos, a los largo del eje 1, constituye lo que es más directamente interpretable. Sin embargo la desviación más evidente entre experimentos se sitúa pasando de los estanques 3-4 a los estanques 5-6, creciendo el día 7. Corresponde al paso de una dosis natural de rayos ultravioletas para 3-4 a una sobredosis para 5-6. Se podría así considerar que la dirección, pasando entre el punto medio entre 3 y 4, y el punto medio entre 5 y 6, caracteriza aún mejor el cambio a estudiar que la dirección del eje 1.

Más que continuar comentando las diferencias, se introducen ahora complenmentos de análisis.

2.5. Definición de contribuciones a la independencia

Una imagen concordante hace referencia a una dependencia medida por cuadrados de correlaciones $\rho^2(YNb_j, XMa_j)$. Por ello, todo lo que induce la no igualdad a 1 de estas correlaciones es del campo de la independencia entre las dos nubes consideradas.

2.5.1. Anti-imagen

Guttman define dos anti-imágenes, para separar las contribuciones respectivas a la dependencia y a la independencia de dos matrices. Se trata de dependencia entre espacios generados, que no dependen de la manera en que los espacios son generados.

La matriz $P_Y = Y(Y'DY)^{-1}Y'D$ es la del proyector de R^n sobre el subespacio generado por las variables de Y, la anti-imagen X_A se define por

$$X_A = X - P_Y X,$$

por oposición a la imagen de Guttman $P_Y X$ definida para extraer de X sus contribuciones a la dependencia con Y. La anti-imagen Y_A se define también como

$$Y_A = Y - P_X Y.$$

Como un proyector es idempotente,

$$(P_Y X)'DX_A = X'DP_Y(X - P_Y X) = 0,$$

$$Y'DX_A = 0,$$

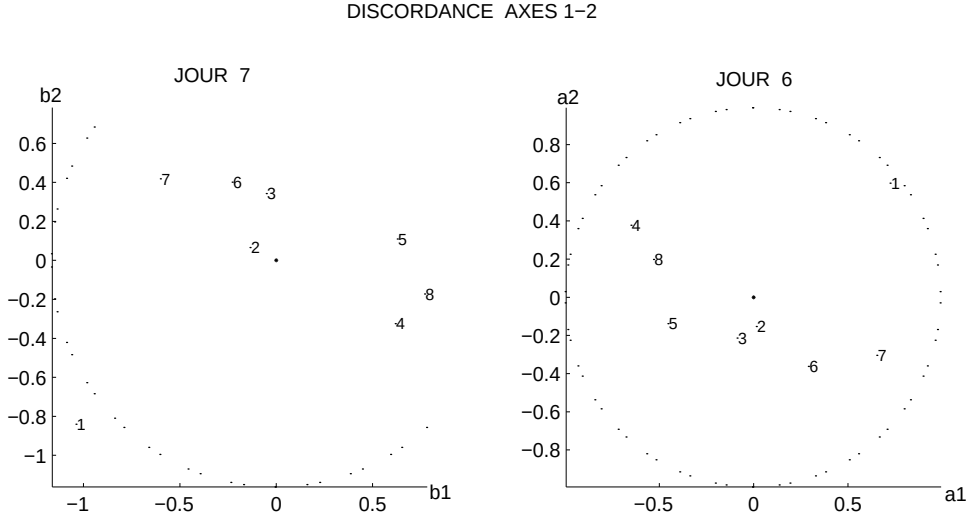
y estas igualdades a matrices nulas indican que una anti-imagen no tiene ligamen ni con otra tabla ni con su complemento al interior de su tabla. Una anti-imagen designa una parte de estructura propia a una tabla que puede ser aislada y analizada separadamente.

Como se tiene

$$X'DX_A = X'_A DX_A,$$

las participaciones de las variables de X a la anti-imagen X_A son también las de las columnas de la anti-imagen. Así el ACP de (X_A, M, D) permite el análisis de la relación entre las variables de X y la nube (X_A, M).

La imagen concordante X_C permanece sin cambio cuando, en lugar de definirla entre los tripletes (X, M, D) y (Y, N, D), se define entre los tripletes $(P_Y X, M, D)$ y (Y, N, D), habiendo entonces reemplazado X por su parte $P_Y X$, dependiente de Y. La imagen concordante constituye una parte de la imagen de Guttman.



Se nota que las dos imágenes dan sensiblemente la misma información. Dos estanques asociados

al mismo experimento están más bien alejados entre ellos. Esto indica que la discordancia puesta en evidencia proviene más bien de las duplicaciones de los experimentos. Por ello, se podría pensar en guardar solo una de las dos repeticiones, las que inducen menos discordancia (2, 3, 5 y 8), y recomenzar el análisis sin las otras 4. Prácticamente todas las analogías y diferencias entre tablas se volverían legibles a partir de los biplots de concordancia. Otra solución se reduce a hacer la media de dos repeticiones antes de lanzar el análisis.

2.5.2. Imagen discordante

Después de la imagen concordante y la anti-imagen, se considera una tercera imagen, llamada *imagen discordante.* La anti-imagen constituye la diferencia entre nubes proveniendo de espacios generados por las variables que son diferentes, correspondientes a definir dos partes que no pueden ser puestas en relación por linealidad.

Durante un análisis de Tucker, dos espacios generados pueden ser idénticos sin que las estructuras internas puedan identificarse (aún después de haber considerado todas las isometrías posibles). Con los biplots, la imagen concordante permite analizar si dos estructuras son cercanas o no, y así ciertas diferencias se ponen en evidencia. Los ejes se definen para permitir la comparación de las dos nubes de contribuciones y el análisis se lleva a cabo como si fueran idénticos. De hecho no lo son y la imagen discordante pone mejor en evidencia esta diferencia complementaria ya visible en concordancia.

Las imágenes concordantes y discordantes se conciben sobre todo para dar un desarrollo al análisis de Tucker (aún si la exposición sea producida con métricas cualesquiera, tratándose de un análisis de dependencia entre nubes en el caso general). En el caso en que las dos métricas M y N son idénticas, la observación de una sola de las dos imágenes discordantes basta a menudo para la descripción de la desviación entre las dos nubes de contribuciones.

En el análisis de Tucker, cuando las dos estructuras se definen por dos juegos diferentes de variables, la imagen discordante permite hacer la medida de la incapacidad que tiene un juego para poder reproducir el sistema de covariaciones del otro. Cuando las dos estructuras se definen por el mismo juego, es la misma cosa. Pero en este caso la imagen concordante permite analizar además si la reproducción se hace de la misma manera a partir de las mismas variables respectivas o no. Puede suceder que dos subconjuntos diferentes de variables produzcan una misma estructura parcial. Esta situación que no puede descubrirse en discordancia lo será en concordancia. Es así que 2 tablas pueden ser idénticas, salvo por una permutación de los nombres de variables: la imagen discordante es nula pues las dos estructuras generadas son idénticas. Sin embargo no son generadas de la misma forma a partir de las mismas variables, y son los biplots de concordancia los que hacen descubrir la importante diferencia que existe entre las dos tablas.
Se define la imagen discordante relativa a X por la matriz $n \times p$

$$X_D = P_Y X - X_C.$$

Entonces se ha descompuesto la imagen de Guttman $P_Y X$ en dos imágenes, y la tabla X en tres imágenes

$$X = X_C + X_D + X_A,$$

con la descomposición correspondiente de la inercia total de X

$$tr(X'DXM) = tr(X'_C DX_C M) + tr(X'_D DX_D M) + tr(X'_A DX_A M).$$

La forma del índice LAI nos conduce a considerar la igualdad

$$1 = \frac{tr(X'_C DX_C M)}{tr(X'DXM)} + \frac{tr(X'_D DX_D M)}{tr(X'DXM)} + \frac{tr(X'_A DX_A M)}{tr(X'DXM)},$$

para proponer un índice global de discordancia y un índice global para la anti-imagen.

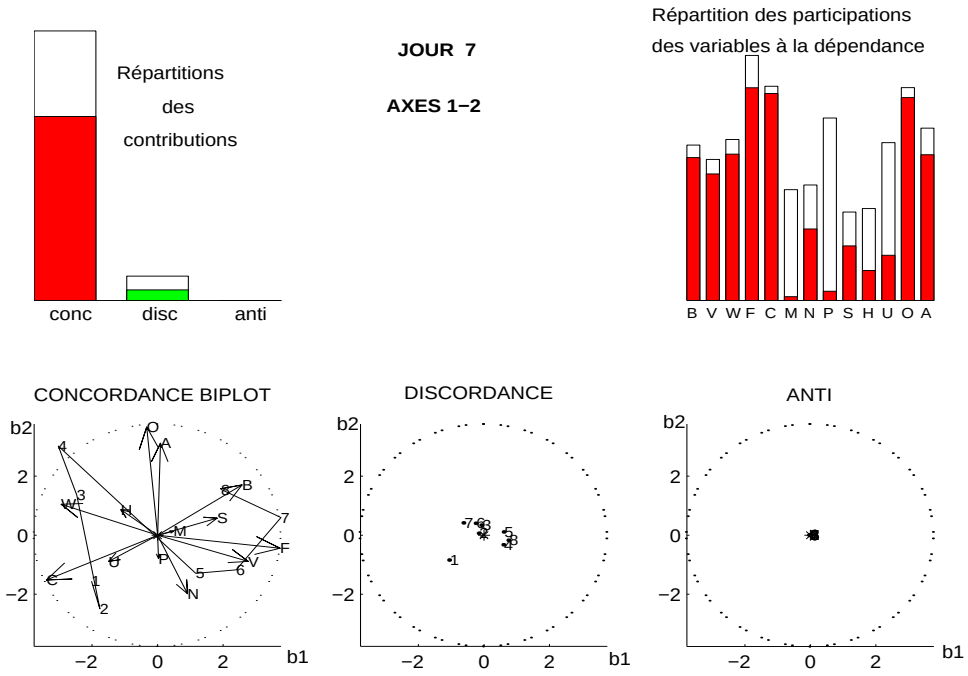
Se puede verificar que la anti-imagen X_A, que es independiente de la imagen de Guttman $P_Y X = X_C + X_D$, es también independiente de las dos imágenes que la descomponen, pues se tienen las igualdades

$$X'_C D X_A = X'_D D X_A = 0.$$

Toda la imagen concordante X_C está englobada por el sistema de ejes $\{a_i\}$ y se muestra que es lo mismo para la imagen discordante X_D. Se pueden entonces usar sistemas de coordenadas definidos en concordancia para representer la discordancia.
La interpretación de los ejes dada por los biplots de concordancia sirve en discordancia y para la anti-imagen. Un mismo sistema de coordenadas (a_i, a_j) puede servir para visualizar tres sistemas de contribuciones complementarias, yuxtapuestas debido a una complementaridad de inercias de nubes proyectadas sobre cada eje. En efecto, $\forall \imath$, se verifica

$$a'_i M X' D X M a_i = a'_i M X'_C D X_C M a_i + a'_i M X'_D D X_D M a_i + a'_i M X'_A D X_A M a_i.$$

En general, el sistema de ejes de concordancia $\{a_i\}$ no permite captar completamente la anti-imagen X_A. Sin embargo, para conocer la importancia relativa de esta anti-imagen para cada eje de concordancia, se quiere yuxtaponer la anti-imagen al lado de la imagen discordante y de la imagen concordante. Por ejemplo, cuando la anti-imagen es asimilable a un ruido, puede ser interesante constatar que este ruido no es importante en el subespacio considerado (generado por ejes de concordancia).



La parte sombreada de las barras de los diagramas corresponde a definir la importancia de las partes de inercia reveladas por solo las dos dimensiones consideradas. La suma de las inercias totales del diagrama de bastones de la izquierda es igual a la inercia total de la tabla JOUR (DIA) 7 y entonces a la suma de las varianzas totales de las variables del JOUR (DIA) 7 representadas sobre el diagrama de la derecha (las métricas son identidad).

La suma de las inercias representadas sobre los 3 plots es igual a la inercia de la nube de los estanques del JOUR (DIA) 7 proyectado sobre el plano considerado. La anti-imagen es nula, lo que se produce a menudo cuando se tiene como aquí más variables que individuos. Gráficos análogos podrían ser asociados a JOUR (DIA) 6.

Propiedad: La discordancia es nula cuando la concordancia es buscada con la otra tabla sin la restricción de estructura sobre esta tabla.

En un estudio de dependencia entre estructuras, se suprime toda restricción proveniente de la estructura de una tabla cuando se escoge por métrica la de Mahalanobis en lugar de la métrica identidad (otra manera de suprimir la estructura de una tabla consistiría en reemplazarla por la tabla de sus componentes principales reducidas; esto conduce a las mismas consecuencias). Así por ejemplo cuando $N = (Y'DY)^{-1}$, entonces $X_C = P_Y X$, y la discordancia X_D es nula. Esta propiedad indica que la discordancia medida sobre una tabla está ligada a la estructura interna de la otra tabla.

Cuando $M = I_p$, reemplazar $N = I_q$ por $N = (Y'DY)^{-1}$, es pasar del análisis de Tucker al ACPVI. La parte dependiente X_C se vuelve la parte máxima $(P_Y X)$ de variabilidad de X pudiendo explicarse a partir de Y o a partir de cualquier conjunto de variables que puedan generar un subespacio idéntico. El índice $\text{LAI}[(X,\, I_p)/(Y,\, (Y'DY)^{-1})] = \frac{tr[X'DP_Y X]}{tr(X'DX)}$, es conocido bajo el nombre de índice de redundancia de Stewart y Love [27]. Entonces, hacer el ACP de $(P_Y X, I, D)$, con biplots, se reduce a representar los biplots de concordancia asociados a X_C. El ACP de tal triplete a sido preconizado por D'Ambra L. & Lauro N.C. (1982).

3. Análisis de la dependencia de K tripletes con otro triplete

3.1. Introducción

Un estudio de la dependencia simultánea entre una tabla y K tablas es considerada [20] y [15]. Desde el punto de vista de la regresión múltiple, se trata de una extensión en que cada variable de esta regresión sería reemplazada por un conjunto de variables. Desde el punto de vista del álgebra lineal, se trata de una proposición para extender la descomposición en valores singulares de una aplicación lineal a la descomposición de un ligamen creado por K aplicaciones lineales entre un espacio métrico y K otras espacios, que tratan sobre el corte sin redundancia de este ligamen. Desde el punto de vista del análisis factorial de tablas múltiples, se trata de una extensión no simétrica de los análisis factoriales clásicos de dos tablas, una tabla sirve de referencia para K otras tablas, como en el estudio de dependencia entre una variable cualitativa y K otras [4], ciertos análisis canónicos [1], ciertas regresiones PLS [31], ciertos análisis procrustos generalizados o incluso los análisis PCS, PCA, SCA-P [18], etc.

Se deducen varias ventajas por la presencia de una tabla de referencia. Cuando K tablas dependen simultáneamente de una $K+1$-ésima de una misma manera, es que ellas dependen entre ellas también. Así la tabla de referencia solo puede estar presente para expresar una restricción para todas las otras tablas, con el fin de realizar un estudio de su interdependencia bajo esta restricción. O bien la referencia es un compromiso de las K tablas de manera que el análisis se convierte en un análisis de K tablas. Muchos análisis de K tablas crean de hecho un compromiso necesario cuando se implementan y el usuario no puede, en general, modificar este compromiso inherente al método. Aquí el usuario puede fabricar una referencia compromiso como él quiera, por media ponderada de las K tablas, por concatenación ponderada, etc. O incluso la referencia sirve para crear un "efecto lupa", como en el ejemplo tratado...

El principio de un corte sin redundancia de una aplicación lineal entre espacios métricos ha permitido introducir el ACP de un triplete y los análisis factoriales clásicos de 2 tripletes, el cálculo se hacía por descomposición en valores singulares de una matriz. Para ser más explícito sobre este asunto:

Sea f una aplicación lineal entre espacios métricos representada por la matriz A. Se dice que una pareja de vectores normados (a, b) que verifican $f(a) = b$ participa en el corte de f cuando toda redundancia es evitada con todas las otras parejas (c, d) que verifican $f(c) = d$, o sea cuando se tiene $f(a) = sb,\ s \neq 0$, y $f(a^{\perp}) \subset b^{\perp}$.

Como $\{f(a^{\perp}) \subset b^{\perp}\} \iff \{f'(b) = sa\}$, se puede deducir que las únicas parejas que participan en el corte de f son los pares de vectores singulares de A que se asocian a un valor singular. Escogiendo una pareja (a_i, b_i) por valor singular s_i, se obtiene el corte de f por pedazos $f = \sum s_i b_i a_i'$.

Estos pares también se pueden definir a partir del problema de la optimización sucesiva del criterio s (o, más fecuentemente considerado, del criterio s^2), así este criterio se puede asociar a la noción misma de corte de f. Sin embargo, la definición de las parejas por el criterio pedido, que justifiquen las restricciones de sucesión escogidas, no están implícitamente definidas por el criterio. El criterio como tal no conduce a definir una sola pareja, mientras que la noción de corte conduce a considerarlos todos. Para lo que sigue, como no hemos podido encontrara un noción de corte suficiente, hemos escogido el camino de una definición paso a paso de las soluciones sucesivas, a partir de una extensión del criterio s^2.

3.2. Definición de las componentes

Los datos son constituidos de $K+1$ matrices Y y $X_i,\ i = 1 \ldots K$, correspondientes a $K+1$ conjuntos de variables centradas, medidas sobre los mismos n individuos. La matriz Y es de

dimensión $n \times q$ y las X_i son de dimensiones respectivas $n \times m_i$. Se denota $m = \sum m_i$ y X la tabla $n \times m$ forma por concatenación de las X_i

$$X = [X_1\ X_2\ \cdots\ X_K].$$

Las K tablas X_i constituyen una partición de X.

Se propone una generalización del análisis de la dependencia de 2 tablas, idéntico a este último cuando $K = 1$.

Las S soluciones buscadas están formadas de $(K+1)$-uplas $(b_j,\ a_{1j}, a_{2j}, \ldots, a_{K_j})$, $j = 1 \ldots S$, donde los K vectores normados a_{ij} de $\mathbb{R}^{m_i}$, $i = 1, \ldots, K$, se asocian al vector normado b_j de $\mathbb{R}^q$.

La métrica euclídea sobre $\mathbb{R}^q$ se denota N, las de $\mathbb{R}^{m_i}$ se denotan respectivamente M_i y la de $\mathbb{R}^m$ formada por bloques diagonales M_i se denota M.

Se hubiera querido que las K-uplas verifiquen

$$\rho(YNb_j,\ X_iM_ia_{ij}) \neq 0,\ \forall i,\ \forall j,$$

$$\rho(YNb_j,\ X_iM_ia_i^*) = 0,\ \forall i,\ et\ \forall a_i^* \perp a_{ij},$$

$$\rho(YNb^*,\ X_iM_ia_{ij}) = 0,\ \forall i,\ et\ \forall b^* \perp b_j,$$

para definir, para cada j fijo, K componentes $X_iM_ia_{ij}$ todas socias de la componente YNb_j, sin que estos socios están correlacionados a los YNb_k para $k \neq j$. Pero se sabe que es imposible obtener en general, pues entonces haría falta que b_j sea vector singular de K matrices $Y'DX_i$ simultáneamente.

Se propone entonces definir $(K+1)$-uplas sucesivas, según el criterio

$$f(b, u_1, u_2, \ldots, u_K) = \sum_{i=1}^{K} cov^2(YNb,\ X_iM_iu_i).$$

Esta escogencia corresponde en efecto a una generalización del criterio f considerado cuando $K = 1$, es decir para dos tablas (nota de la sección 2.2). Se muestra que su optimización bajo restricciones de norma equivale a la optimización bajo restricciones de norma del criterio g

$$g(b, a) = cov^2(YNb,\ XMa),$$

en la medida en que el óptimo se obtiene para el mismo vector b, el mismo valor del máximo para f y g, los vectores soluciones u_i son los vectores normados sub-bloques del vector solución a, de dimensiones respectivas m_i. Es como decir que la obtención de la primera $(K+1)$-upla $(b_1, a_{11}, a_{21}, \ldots, a_{K1})$ se deduce de los dos vectores singulares asociados al mayor valor singular de la matriz $Y'DX$, considerada como la de una aplicación lineal de $(\mathbb{R}^m, M)$ en $(\mathbb{R}^q, N)$. Una primera solución de este tipo ha sido considerada a menudo en análisis multi-tablas basándose en el criterio g (por ejemplo para encontrar un código de K variables cualitativas que pueda asociarse a un código de una $(K+1)$-ésima [4]).

Como $X_i'DYNb_1 = s_i\, a_{i1}$, $\forall i$, esta primera solución posee la propiedad deseada siguiente

$$\forall i:\ \rho(YNb_1,\ X_iM_ia_i^*) = 0,\ \forall a_i^* \perp a_{i1},$$

de manera que las componentes socias $X_iM_ia_{i1}$ serán las únicas en poder asociarse con la componente YNb_1. Es más, la naturaleza del criterio f indica que las correlaciones $\rho(YNb_1, X_iM_ia_{i1})$ serán frecuentemente todas no nulas. Finalmente, alq uerer extender la noción de dependencia de dos tablas, se obtiene para esta primera solución muchas propiedades deseadas.

Las soluciones siguientes se calculan como la primera, pero después de ser colocadas en los ortogonales de las soluciones anteriores, ortogonalidades implícitamente indicadas en la propiedad

anterior. El cálculo de una solución j, $j \geq 2$, se hace así después de haber reemplazado cada matriz X_i por la matriz

$$X_i - \sum_{k=1}^{j-1} X_i M_i a_{ik} a'_{ik}.$$

Estas deflaciones, y la búsqueda de una nueva solución, se detienen cuando los rangos de estas matrices se vuelven todos nulos. El número S de soluciones es así al menos igual a máx$\{rang(Y'DX_i)\}$. En efecto, puede ser superior a este máximo cuando las soluciones a_j poseen algunos subvectores a_{ij} nulos.

Notando B, A_i y A, las matrices de dimensiones respectivas $q \times S$, $m_i \times S$ y $m \times S$, contienen los ejes respectivos b_j, a_{ij} y a_j en columnas, las soluciones así calculadas tienen las propiedades suiguiente

$$B'NB = A'_i M_i A_i = A'MA = I_S.$$

$$\rho(YNb_j,\, X_i M_i a_{ik}) = 0,\ \forall i,\ et\, \forall k \geq j+1.$$

$$Y'DXMAA' = B\Delta A'.$$

La última igualdad indica que la descomposición en valores singulares de $Y'DXMAA'$, matriz de una aplicación lineal de $(\mathbb{R}^m, M)$ en $(\mathbb{R}^q, N)$ restringida al subespacio definido por el proyector $AA'M$, permite encontrar A y B, es decir las $(K+1)$-uplas soluciones. Un corte real de dependencia es asociado a estas soluciones y esto es fundamental para poder conducir la continuación del análisis. Es por esta razón que su desarrollo está fundamentado en el análisis de dependencia de las dos nubes (Y, N) y $(XMAA', M)$.

Los K sistemas ortonormados A_i pueden ser superpuestos al sistema ortonormado B por rotación ortogonal (por isometría parcial). Esto se reduce a transformar cada tabla X_i en una tabla $X_i M_i A_i B'$ cuyas filas pueden representarse en el sistema de coordenadas de ejes $\{b_j\}$ contenidos en B, así como las filas de la tabla Y. El efecto de tales rotaciones, que corresponden a una proposición de solución a un problema procrusto generalizado donde el objetivo sería Y, puede visualizarse en el sistema de coordenadas común, las coordenadas de las filas se encuentran entonces en las matrices YNB y $X_i M_i A_i$. Esto ha sido propuesto para hacer corresponder las imágenes dentales 3D sobre una imagen dada, las métricas siendo las métricas identidad [3].

3.3. Imágenes concordantes

Se quiere asociar a cada una de las $K+1$ tablas una imagen concordante que contenga las contribuciones de sus filas a la dependencia simultánea estudiada.

Se definen primero las dos imágenes concordantes que resultan del análisis de dependencia de las nubes (Y, N) y $(XMAA', M)$.

Se obtienen así las imágenes $X_C = YNB\Delta_b A'$ de dimensión $n \times m$ y $Y_C = XMA\Delta_a B'$ de dimensión $n \times q$.

Se definen entonces las imágenes concordantes X_{C_i} relativas a las tablas X_i por los subbloques de X_C de dimensiones $n \times m_i$. La imagen X_C se llama *imagen concordante sintética*. Se prueba

$$X_{C_i} = YNB\Delta_{b_i} A'_i,$$

cada matriz diagonal Δ_{b_i} contiene los coeficientes de regresión simples de las componentes $X_i M_i a_{ij}$ sobre YNb_j.

Ponemos $\rho^2_{ij} = \rho^2(YNb_j,\, X_i M_i a_{ij})$ y $\rho^2_j = \rho^2(YNb_j,\, XMa_j)$.

La inercia de la nube de las filas de la imagen X_{C_i} proyectada sobre el eje j vale

$$\rho^2_{ij}\, var(X_i M_i a_{ij}),$$

es decir la acumulación para el eje j de las contribuciones de las filas de X_i a una dependencia con Y medida por ρ_{ij}^2. Cuando $M_i = I_{m_i}$ esta inercia representa también una parte de la varianza total de X_i. Según la igualdad

$$cov^2(YNb_j,\, XMa_j) = \sum_{i=1}^{K} cov^2(YNb_j,\, X_iM_ia_{ij}),$$

se deduce

$$\rho_j^2 var(XMa_j) = \sum_{i=1}^{K} \rho_{ij}^2 var(X_iM_ia_{ij}),$$

de manera que para cada eje j las contribuciones a la dependencia $\rho_j^2 var(XMa_j)$, provenientes del corte y de la imagen concordante sintética X_C, se descomponen para precisar la parte tomada por cada una de las tablas X_i. Como la contribución total a la dependencia entre Y y $XMAA'$ se descompone por tablas según la igualdad

$$\sum_{j=1}^{S} \rho_j^2 var(XMa_j) = \sum_{i=1}^{K}\sum_{j=1}^{S} \rho_{ij}^2 var(X_iM_ia_{ij}),$$

el valor

$$\max_i \{\sum_{j=1}^{S} \rho_{ij}^2 var(X_iM_ia_{ij})\}$$

permite llamar la tabla X_i cuya contribución a la dependencia es mayor. Esto puede ser usado para designar entre las estructuras X_i la que sería "más cercana" de la estructura Y. Este criterio puede servir en la selección de variables, las tablas X_i corresponden entonces a diferentes hipótesis "explicativas", para crear procedimientos de escogencia paso a paso [24]. Los valores

$$\sum_{j=1}^{S} \rho_{ij}^2 var(X_iM_ia_{ij})$$

son los numeradores de índices llamados aquí *de concordancia parcial*

$$LAI_p[(X_i, M_i),\, (Y, N)] = \cos^2_{(D,M_i)}(X_i, X_{C_i}) = \frac{tr(X'_{C_i}DX_{C_i}M_i)}{tr(X_i'DX_iM_i)}.$$

Se verifica el ligamen entre índice de concordancia sintética e índices de concordancia parcial

$$LAI[(XMAA', M),\, (Y, N)] = \sum_{i=1}^{K} \frac{tr(X_i'DX_iM_i)}{tr(X'DXM)} LAI_p[(X_i, M_i),\, (Y, N)].$$

Las medidas relativas representadas por los índices de concordancia parcial podrían eventualemente considerarse en lugar de las medidas absolutas indicadas antes, en ciertos problemas de selección y de reconocimiento de patrones.

En los sistemas de coordendas asociados yuxtapuestos, se representan las nubes respectivas de las contribuciones. Las filas de Y_C son representadas en un sistema de coordenadas $\{b_k, b_j\}$ al lado de las de X_{C_i} representadas en los sistemas de coordenadas respectivos $\{a_{ik}, a_{ij}\}$. El parecido entre nubes (X_{C_i}, M_i) puede ser importante, las coordenadas respectivas se encuentran en las matrices $YNB\Delta_{b_i}$. Esto es comprensible sabiendo que cada tabla X_i debe "concordar" con la misma tabla Y. Toda dirección cualquiera fijada que pasa por el origen y que atraviesa uno de estas $K+1$ nubes

puede entonces ser asociada, en esta relación de dependencia, a la misma dirección que atraviesa cada una de las K otras nubes.

Nota. Las componentes XMa_j aparecen como medias ponderadas de las componentes socias

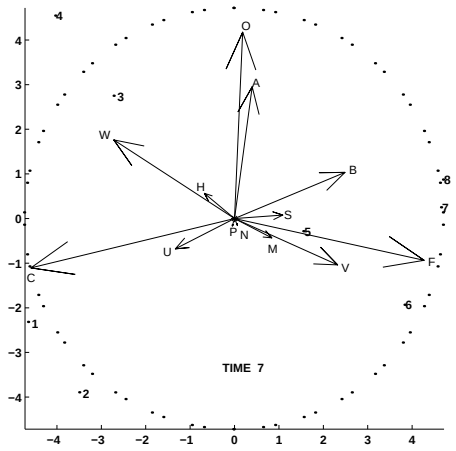
$$XMa_j = \sum_{i=1}^{K} p_{ij} X_i M_i a_{ij},$$

con los números positivos

$$p_{ij} = \frac{cov(X_i M_i a_{ij}, YNb_j)}{cov(XMa_j, YNb_j)},$$

$$\sum_{i=1}^{K} p_{ij}^2 = 1.$$

Esta última igualdad precisa, después de simplificar por $var(YNb_j)$, que la suma de las contribuciones de las componentes socias a la dependencia iguala la contribución de la componente media. Esto indica el tipo de compromiso constituido por esta componente media.



Sobre los biplots yuxtapuestos anteriores, el día 7 ha sido tomado como referencia y el análisis ha sido hecho entre estructuras (métricas identidad). Por construcción las escalas de longitudes de

flechas y de trazo de las contribuciones de los estanques son autónomas para la referencia y pueden ser administradas aparte. Para los otros 6 días, se adoptan escalas comunes. Son los días 5 y 6 los que dependen mucho del día 7, como revela la dilatación relativamente importante de las dos nubes respectivas de contribuciones. Globalmente, en un primer análisis parece iluso querer ir más lejos en la explicación del día 7. De hecho, esta salida, que ha producido un efecto lupa sobre día 7 y los dos días anteriores, no presenta mayor interés para estos datos, ya que la previsión del día 8 no es un objetivo. Otro estudio para analizar las contribuciones de los días, y no de los estanques, pone en evidencia que hacia el día 4 se produce un cambio radical en la evolución. Entonces, para estos datos ha sido preferible tomar el día 4 como referencia. El efecto lupa se produce entonces alrededor del día 4.

3.4. Biplots

Las $K+1$ nubes de contribuciones a la dependencia pueden ser yuxtapuestas en sus sistemas de coordenadas respectivos. Así las filas de Y_C se pryectan sobre el plano (b_k, b_l) y las de X_{C_i} se proyectan sobre el plano (a_{ik}, a_{il}), yuxtapuesto, para $i = 1, \ldots, K$.

No existe ningún problema en sobreponer las filas de Y a la nube de contribuciones de Y_C.

Pero cuando $X_i M_i A_i A_i' \neq X_i$, la definición de las participaciones de las variables de X_i plantea un problema pues la lista de los valores

$$diag(X_i' D X_{C_i} M_i a_{ik} a_{ik}'),$$

contiene valores negativos. La desviación tiene la situación en que todos los valores son positivos y nulos en promedio,

$$tr([X_i - X_i M_i A_i A_i']' D X_{C_i} M_i a_{ik} a_{ik}') = 0,$$

y los valores negativos aparecen en general para las variables subrepresentadas, de manera que el fenómeno podría ignorarse a menudo. Nos proponemos más bien definir las participaciones por medio de los valores absolutos

$$\mid diag(X_i' D X_{C_i} M_i a_{ik} a_{ik}') \mid,$$

considerando que para obtener efectivamente estos números positivos ha sido necesario modificar, cada vez que era necesario, los signos correspondientes de los valores del coseno director contenidos en a_{ik}'. Es decir que cada modificación de signo conlleva una modificación de la dirección de la variable representada.

En el análisis de la dependencia entre estructuras, las inercias representadas por eje no serán iguales a las varianzas totales representadas para los mismos ejes. Una redundancia de variabilidad aparece, aún si a menudo es baja, que traduce el hecho que para poner realmente en relación la nube con cada una de las variables sea necesario sobre-representarla.

3.5. Imágenes discordantes y anti-imágenes

Prolongando la sección anterior, la anti-imagen sintética es definida por $X_A = X - P_Y X$, de donde las anti-imágenes parciales definidas como sub-bloques matriciales $X_A = [X_{A_1} X_{A_2} \ldots X_{A_K}]$. Esto se reduce a definir estas anti-imágenes parciales por $X_{A_i} = X_i - P_Y X_i$.

Se considera el proyector P_Z de $\mathbb{R}^n$ sobre el subespacio generado por las componentes medias XMA. La anti-imagen relativa a Y se define por $Y_A = Y - P_Z Y$. De las definiciones anteriores, se deduce una imagen discordante sintética $X_D = P_Y X - X_C$, cuyos sub-bloques $X_D = [X_{D_1}\ X_{D_2}\ \ldots\ X_{D_K}]$ corresponden a la definición $X_{D_i} = P_Y X_i - X_{C_i}$.

Igualmente, $Y_D = P_Z Y - Y_C$.

Nos proponemos representar una imagen dicordante parcial con los sistemas de coordenadas de los ejes A_i, de manera que todo sucede como si las imágenes discordantes hubieran sido más bien definidas por $X_D = P_Y XMAA' - X_C$.

Entonces la interpretación de ese tipo de imagen se parece a la dada en análisis de dependencia de 2 tablas. Sobre todo, con las métricas identidad, una imagen parcial muestra su diferencia de estructura con la de Y.

3.6. CONCOR y ACOM

Se llama CONCOR el análisis de $K+1$ tablas que acaba de ser desarrollado [23]. Para $K=1$ el análisis CONCOR es un análisis de co-inercia de dos tablas. El ACOM se llama análisis de co-inercia múltiple [6] de K tablas X_i, pero para $K=2$ no es el análisis de co-inercia de dos tablas.

Si se considera el análisis CONCOR con la matriz $Y=X$ como referencia, las K tablas son las tablas X_i analizadas por el ACOM, y si además se toma como métrica $N=(Y'DY)^{-1}$, entonces las componentes $X_iM_ia_{ij}$ son las componentes explicadas del ACOM (explicadas por una variable llamada auxiliar que no es otra que YNb_j) [21]. Conociendo esta escogencia de métrica para Y y la naturaleza de esta tabla, el ACOM con métricas identidad es un análisis de dependencia con el subespacio generado por todas las variables, o incluso un análisis de K subconjuntos de variables bajo la restricción que sean dependientes del subespacio generado por el conjunto. En vista de la naturaleza de esta restricción, ésta no tiene ningún efecto particular, de manera que este ACOM se reduce a un análisis de la interdependencia entre K paquetes de variables, igual que el ACP es un análisis de la interdependencia entre K variables.

El ACOM presentado aquí aparece como un análisis CONCOR particular con una matriz por invertir (susceptiple de ser de gran tamaño). De hecho no es así. En el caso general, en el análisis CONCOR, las S soluciones son dadas por S cálculos SVD sucesivas, la primera SVD es la de $Y'DX$. En el caso general, en ACOM, las S soluciones son dadas desde S cálculos SVD sucesivos, la primera SVD es la de X. Desde este punto de vista, también es claro que si CONCOR es una generalización del análisis de dependencia de dos tripletes, entonces el ACOM es una generalización del análisis de la interdependencia entre variables de un mismo triplete. CONCOR $K+1$ tablas y ACOM K tablas son las extensions respectivas de dos situaciones "$K=1$" no identificables.

Nota: La redacción "personal" sigue a una serie de colaboraciones, sobre todo con M. Hanafi y J. Ten Berge, y de una sensibilización al enfoque exploratorio por medio de la Escuela EDA dirigida por J. Vanpoucke y E. Horber [9]. En un futuro se espera poner este curso en acceso libre en la red. Se agradece la traducción al español de Javier Trejos.

Referencias

[1] Balbi, S.; Esposito, V. (1999) "Rotated canonical analysis onto a reference subspace", *Computational Statistics and Data Analysis*, North Holland.

[2] Bertin, J. (1967) *Sémiologie Graphique.* Mouton/Gauthier-Villars, Paris/La Haye.

[3] Boucays, F.; Lafosse, R.; Madrid, C.; Treil, J. (1998) "Spatial concordance of a 3D cephalometric analysis of 29 subjects", #3136 *J. Dent. Res.* **77** (IADR Abstracts).

[4] Cazes, P.; Baumerder, A.; Bonnefous, S.; Pagès, J.P. (1977) "Codage et analyse des tableaux logiques. Introduction à la pratique des variables qualitatives", *Cahiers du BURO* **27**, Université de Paris VI.

[5] Chanut, J.P.; Lafosse, R.; Demers, S.; Mostajir, B.; Chatila, K.; Belzile, C.; Roy, S.; Gosselin, M. (2000) "Analyse de concordance entre traitements: cas d'une communauté planctonique sous rayonnement ultraviolet-B", *XXXII Journées de Statistique, SFdS*, Fez, Maroc.

[6] Chessel, D.; Hanafi, M. (1996) "Analyses de la co-inertie de K nuages de points", *Revue de Statistique Appliquée* **44**(2): 35–60.

[7] Cliff, N. (1966) "Orthogonal rotation to congruence", *Psychometrika* **31**: 33–42.

[8] D'Ambra, L.; Lauro, N.C. (1982) "Analisi in componenti principali in rapporto a un sottospazio di riferimento", *Riv. Statist. Appl.* **15**: 51–67.

[9] Dolédec, S.; Chessel, D. (1994) "Co-inertia analysis: an alternative method for studying species-environment relationships", *Freshwater Biology* **31**: 277–294.

[10] EDA Summer School, http://www.unige.ch/ses/sococ/mirage/

[11] Gabriel, K.R. (1971) "The biplot graphic display of matrices with application to principal component analysis", *Biometrika* **58**: 453–457.

[12] Green, B.F.(1952) "The orthogonal approximation of an oblique estructura in factor analysis", *Psychometrika* **17**: 429–440.

[13] Gower, J.C.; Hand, D.J. (1996) *Biplots.* Chapman & Hall, London.

[14] Guttman, L. (1953) "Image theory for the structure of quantitative variates", *Psychometrika* **18**: 277–296.

[15] Hanafi, M. (1997) *Structure de l'Ensemble des Analyses Multivariées des Tableaux de Données à Trois Entrées: Eléments Théoriques et Appliqués.* Thèse de Doctorat, Université Claude Bernard, Lyon I.

[16] Hanafi, M.; Lafosse, R. (in press) "Généralisations de la régression linéaire simple pour analyser la dépendance de K ensembles de variables avec un K+1 ème", *Revue de Statistique Appliquée.*

[17] Harshman, R.; Lundy (1984) "Data prepocessing and extended PARAFAC model", in: H.G. Law, C.W. Snyder, J.A. Hattie & R.P. McDonald (Eds.) *Research Methods for Multimode Data Analysis*, Praeger, New York: 216–284.

[18] Johansson, J.K. (1981) "An extension of Wollenberg's redundancy analysis", *Psychometrika* **46**: 93–103.

[19] Kiers, H.; Ten Berge, J. (1994) "Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure", *British Journal of Mathematical and Statistical Psychology* **47**: 109–126.

[20] Lafosse, R. (1997) "Analyse de concordance de deux tableaux: monogamies, simultanéités et découpages", *Revue de Statistique Appliquée* **45**(3): 45–72.

[21] Lafosse, R.; Hanafi, M. (1997) "Concordance d'un tableau avec K tableaux: définition de K+1 uples synthétiques", *Revue de Statistique Appliquée* **45**(4): 111–126.

[22] Lafosse, R.; Hanafi, M. (1997) "Concordances d'un tableau avec K tableaux et analyse de co-inertie de K tableaux", *XXIX Journées de Statistique, ASU-SSF*, Carcassonne.

[23] Lafosse, R.; Chessel, D. Hanafi, M. (1997) "Analogies de structures de vins de Cahors", *5èmes Journées Européennes Agro-Industrie et Méthodes Statistiques* **25**: 1–10.

[24] Lafosse, R. (1998) "Programme de l'analyse CONCOR en langage MATLAB", version 1.3. ftp: //ftp.mathworks.com/pub/contrib/v5/stats/concor/ or E-Mail to: lafosse@cict.fr.

[25] Lafosse, R. (1999) "Analysis of concordance between matrices and proposals for selecting variables", *IX International Symposium on Applied Stochastic Models and Data Analysis*, Lisboa.

[26] Lafosse, R. (1999) "Analyses of some relation between arrays and graphics", *Proc. Int. Conf. on Probability and Statistics and their Applications*, Hanoi.

[27] Rao, C.R. (1964) "The use and the interpretation of principal component analysis in applied research", *Sankhya (series A)* **26**: 329–358.

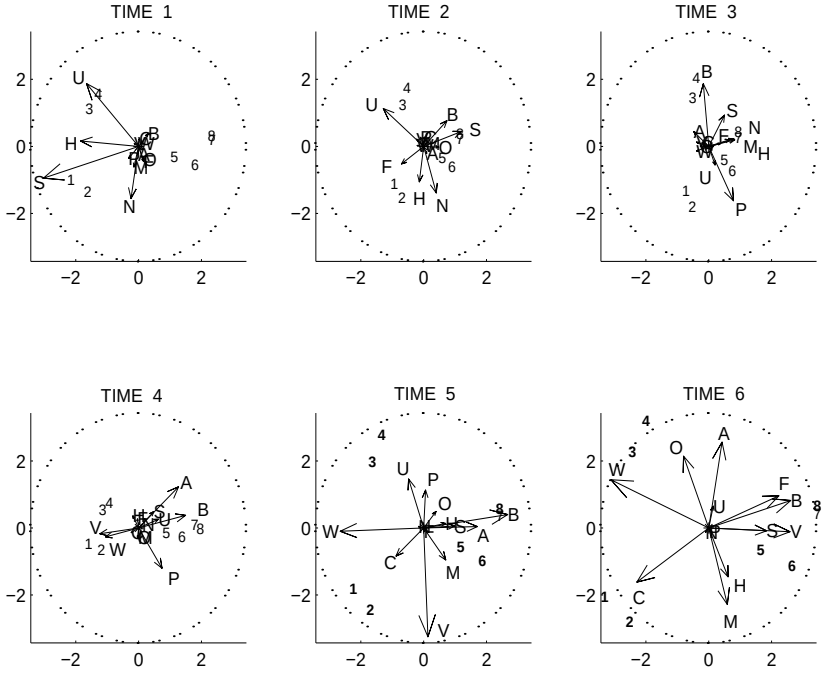
[28] Stewart, D.; Love, W. (1968) "A general canonical correlation index", *Psychological Bulletin* **70**: 160–163.

[29] Ten Berge, J.M.F. (1977) *Optimizing Factorial Invariance.* Doctoral Thesis, Groningen.

[30] Tucker, L.R. (1958) "An interbattery method of factor analysis", *Psychometrika* **23**: 111–136.

[31] Free software, with Xlisp.stat, http://forest.psych.unc.edu/researh/ViStaF.html

[32] Wold, S.; Kettaneh, N.; Tjessem, K. (1996) "Hierarchical multiblock PLS and PC models for easier interpretation and as an alternative to selection variable", *Journal of Chemometrics* **10**: 453–482.


TIME 1
TIME 2
TIME 3
TIME 4
TIME 5
TIME 6