ANÁLISIS CONJUNTO DE MÚLTIPLES TABLAS DE DATOS MIXTOS MEDIANTE PLS

Por

VICTOR MANUEL GONZALEZ ROJAS

TESIS DOCTORAL EN ESTADÍSTICA

Director

Ph.D TOMÀS ALUJA BANET
Profesor Titular UPC

Departament d'Estadística
i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Barcelona, Agosto de 2014

AGRADECIMIENTOS

Doy mis más sinceros agradecimientos a la Universidad del Valle – Escuela de Estadística por haber apoyado mi comisión de estudios doctoral. En especial a mi director de tesis Tomàs Aluja Banet por sus precisos aportes y apreciados consejos que hicieron posible el desarrollo de esta investigación enmarcada en el campo de los métodos PLS.

Quiero resaltar el apoyo incondicional de mi querida y amada esposa María Victoria, quien me acompañó durante todo el desarrollo de este proyecto dándome la fortaleza necesaria para hacer realidad este, el más grande de mis sueños académicos.

Dedico este logro a mi hijo Carlos Alvaro que está formándose profesionalmente; a mis padres Victor Manuel (Q.E.P.D) y Ana Virginia quien siempre ha estado pendiente de la culminación del mismo, acompañándome permanentemente con sus oraciones.

A mis demás familiares, amigos y colegas por acompañarme en cada momento. También agradezco a quienes de una u otra forma me proporcionaron ayudas técnicas e hicieron de este proyecto una realidad.

RESUMEN

El contenido fundamental de esta tesis corresponde al desarrollo de los métodos GNM-NIPALS, GNM-PLS2 y GNM-RGCCA para la cuantificación de las variables cualitativas a partir de las primeras *k* componentes proporcionadas por los métodos apropiados en el análisis de J matrices de datos mixtos. Estos métodos denominados GNM-PLS (General Non Metric Partial Least Squares) son una extensión de los métodos NM-PLS que toman sólo la primera componente principal en la función de cuantificación.

La trasformación de las variables cualitativas se lleva a cabo mediante procesos de optimización maximizando generalmente funciones de covarianza o correlación, aprovechando la flexibilidad de los algoritmos PLS y conservando las propiedades de pertenencia grupal y orden si existe; así mismo se conservan las variables métricas en su estado original excepto por estandarización.

GNM-NIPALS ha sido creado para el tratamiento de una ($J = 1$) matriz de datos mixtos mediante la cuantificación vía reconstitución tipo ACP de las variables cualitativas a partir de una función agregada de *k* componentes. GNM-PLS2 relaciona dos ($J = 2$) conjuntos de datos mixtos $Y \sim X$ mediante regresión PLS, cuantificando las variables cualitativas de un espacio con la función agregada de las primeras H componentes PLS del otro espacio, obtenidas por validación cruzada bajo regresión PLS2. Cuando la matriz endógena Y contiene sólo una variable de respuesta el método se denomina GNM-PLS1.

Finalmente para el análisis de más de dos bloques ($J > 2$) de datos mixtos $Y \sim X_1 + \ldots + X_{J-1}$ a través de sus variables latentes (LV) se implementa el método NM-RGCCA basado en el método RGCCA (Regularized Generalized Canonical Correlation Analysis) que modifica el algoritmo PLS-PM implementando el *nuevo modo A* y especifica las funciones de maximización de covarianzas o correlaciones asociadas al proceso. La cuantificación de las variables cualitativas en cada bloque X_j se realiza mediante la función *inner* $Z_j = \sum_j e_j Y_j$ de dimensión J debido a la agregación de las estimaciones *outer* Y_j. Tanto Z_j como Y_j estiman la componente ξ_j asociad al *j*-ésimo bloque.

Palabras claves: Análisis de múltiples tablas de datos mixtos, General Non Metric, Partial Least Squares, Regression, Path Modeling, Nonlinear estimation by Iterative, Regularized Generalized Canonical Correlation Analysis.

ABSTRACT

The fundamental content of this thesis corresponds to the development of the GNM-NIPALIS, GNM-PLS2 and GNM-RGCCA methods, used to quantify qualitative variables parting from the first *k* components given by the appropriate methods in the analysis of *J* matrices of mixed data. These methods denominated GNM-PLS (General Non Metric Partial Least Squares) are an extension of the NM-PLS methods that only take the first principal component in the quantification function.

The transformation of the qualitative variables is done through optimization processes, usually maximizing functions of covariance or correlation, taking advantage of the flexibility of the PLS algorithms and keeping the properties of group belonging and order if it exists; The metric variables are keep their original state as well, excepting standardization.

GNM-NIPALS has been created for the purpose of treating one ($J = 1$) mixed data matrix through the quantification via ACP type reconstruction of the qualitative variables parting from a *k* components aggregated function. GNM-PLS2 relates two ($J = 2$) mixed data sets $Y \sim X$ through PLS regression, quantifying the qualitative variables of a space with the first *H* PLS components aggregated function of the other space, obtained through cross validation under PLS2 regression. When the endogenous matrix *Y* contains only one answer variable the method is denominated GNM-PLS1.

Finally, in order to analyze more than two blocks ($J = 2$) of mixed data $Y \sim X_1 + \ldots + X_{J-1}$ through their latent variables (LV) the GNM-RGCCA was created, based on the RGCCA (Regularized Generalized Canonical Correlation Analysis) method, that modifies the PLS-PM algorithm implementing the *new mode A* and specifies the covariance or correlation maximization functions related to the process. The quantification of the qualitative variables on each X_j block is done through the *inner* $Z_j = \sum_j e_j Y_j$ function, which has *J* dimension due to the aggregation of the *outer* Y_j estimations. Z_j, as well as Y_j estimate the ξ_j component associated to the *j*-th block.

Keywords: Analysis of multiple tables of mixed data, General Non Metric, Partial Least Squares, Regression, Path Modeling, Nonlinear estimation by Iterative, Regularized Generalized Canonical Correlation Analysis.

TABLA DE CONTENIDO

CAPÍTULO 1

INTRODUCCIÓN

Dentro de los métodos clásicos de análisis multivariado de datos, se encuentra el Análisis de Componentes Principales (ACP), para el tratamiento de una matriz de datos continuos; el Análisis Canónico (AC) junto al Análisis Interbaterias (AIB) y el Análisis de Redundancias (AR) entre otros para el tratamiento de dos tablas de datos; y el Análisis Factorial Múltiple (AFM), el Análisis Canónico Generalizado (ACG) y STATIS entre otros, para el tratamiento de *J* tablas de datos relacionando las variables observadas con las variables latentes (LV).

También, y a partir de los años 70, aparecieron los métodos denominados Partial Least Squares (PLS). Los métodos PLS son procedimientos algorítmicos basados en productos escalares entre vectores que iteran hasta la convergencia, permitiendo obtener las componentes o LV de una matriz o las componentes más relacionadas de varias matrices de datos métricos.

Para el tratamiento de una matriz de datos surgió Nonlinear estimation by Iterative PLS (NIPALS), para el tratamiento de dos matrices PLS Regression (PLSR) y para el análisis de interrelaciones de *J* bloques el PLS Path Modeling (PLS-PM).

Estas aportaciones metodológicas han sido desarrolladas dentro del marco de variables continuas (métricas); otros métodos como el Análisis de Correspondencias tratan activamente sólo variables cualitativas. Sin embargo, en la practica el investigador encuentra datos de naturaleza mixta, es decir, coexisten variables tanto cuantitativas como cualitativas. Generalmente, han existido dos formas de tratar las variables cualitativas (no métricas) en el análisis factorial de datos mixtos:

1. Asociando la tabla disyuntiva completa (TDC) conformada por tantas indicadoras como categorías contenga la variable cualitativa; a su vez, las indicadoras pueden ser transformadas o no.

   Sin embargo, reemplazar cada variable cualitativa con su correspondiente matriz indicadora trae consigo algunos inconvenientes. Los pesos asociados a cada indicadora no son comparables con los pesos asignados a las otras variables, genera el efecto de matrices esparcidas incrementando

innecesariamente la dimensionalidad e inercia, y además, dificulta la interpretación al perderse la integralidad de la variable cualitativa.

2. Cuantificando cada variable no métrica de forma óptima tal que conserven las propiedades de pertenencia y orden si existe. Eventualmente también se transforman las variables métricas dentro del proceso de optimización.

   Se desarrollaron varios métodos para estos procesos de cuantificación, dentro de los cuales se encuentran los denominados Non Metric PLS (NM-PLS) tales como: Dual Scaling (Enfoque Clásico, HOMALS, …), Kruskal and Shepard, Prinqual, Princals entre otros (Gifi 1990, Linting et al. 2007) y los más recientes denominados NM-NIPALS, NM-PLS1, NM-PLS2 y NM-PLSPM de Russolillo 2009. En general estos métodos se caracterizan por cuantificar las variables cualitativas tomando sólo la primera componente principal en la transformación.

   Cuadras 1989, Boj et al. 2007, presentarón métodos de clasificación y regresión basados en distancia, con datos mixtos.

En este trabajo de tesis se presentan los métodos GNM-PLS (General Non Metric PLS) como una generalización de los métodos NM-PLS, ya que se cuantifica mediante un criterio de optimización cada variable cualitativa a partir de k componentes, permitiendo un mayor poder descriptivo en el análisis conjunto de J tablas de datos mixtos. Las variables cuantitativas conservan su escala original (excepto por estandarización) y también se mantienen las propiedades de pertenencia grupal y orden (si existe) durante el proceso de cuantificación.

Así, se desarrolla GNM-NIPALS el cual permite cuantificar las variables cualitativas de una matriz de datos mixtos mediante una función lineal de k componentes principales, tipo reconstitución, maximizando la inercia en el plano k-dimensional asociado al ACP de la matriz cuantificada.

Los métodos GNM-PLS Regression (GNM-PLSR) tienen por objetivo realizar una regresión PLS en dos conjuntos de datos mixtos $Y_{n.r}$ e $X_{n.p}$, conteniendo q_r y q_p variables cualitativas en los conjuntos respectivos. Las variables cualitativas de un grupo son cuantificadas con una función de h componentes agregadas del otro grupo y viceversa.

GNM-PLSR se diferencia en que cuantifica óptimamente cada una de las variables cualitativas Y_q tomando una función agregada de las primeras H

componentes derivadas de la validación cruzada bajo regresión PLS, de la forma $\gamma_t = c_{q1}t_1 + c_{q2}t_2 + \cdots + c_{qH}t_H$. Los ponderadores c_{qh} se obtienen como los coeficientes de regresión entre la variable cuantificada $\hat{Y}_q$ y la componente t_h conllevando máxima covarianza agregada por cada etapa *h*: $\sum_h^H cov^2\,(t_h, u_h)$.

Las u_h son las componentes en el espacio de las $\hat{Y}$ en cada etapa h, y conforman la función agregada γ_u para la cuantificación de variables cualitativas X_q. Dado que $\hat{Y}$ ahora contiene todas las variables métricas y cuantificadas, entonces para *H* componentes en la función agregada, el máximo corresponde a

$$\sum_k^r r^2(Y_k, \gamma_t)\ .$$

Este resultado corresponde a la primera fase del AIB con máximo valor propio λ_h en cada etapa *h*. Si $r \geq 1$ el método general de cuantificación GNM-PLR se denomina GNM-PLS2, el cual a su vez contiene como caso particular al método GNM-PLS1 cuando *r* = 1.

Finalmente, para cuantificar y relacionar más de dos bloques de datos mixtos se ha creado el programa NM-RGCCA bajo el entorno R, como versión mejorada y complemento del NM-PLSPM.

PLS-PM estima la red de relaciones denominadas *outer* entre las variables manifiestas (MV) contenidas en cierto grupo X_j y su propia LV ξ_j , e *inner* entre las LV de *Q* bloques de datos. Se relacionan las componentes *outer* $Y_j = X_j w_j$ e *inner* $Z_j = \sum_j e_j\, Y_j$ que estiman ξ_j, para obtener las condiciones de estacionariedad y estimar los pesos *outer* w_j, mediante los modos *A* y *B* que implementan regresiones simples y múltiples respectivamente.

NM-PLS PM se crea aprovechando la flexibilidad del algoritmo PLS-PM para intervenirlo inmediatamente después de la fase de obtención de las componentes *inner* Z_j de dimensión *J* y a partir de éstas cuantificar las variables cualitativas del grupo respectivo; sin embargo no es clara la función de optimización.

El método RGCCA (Regularized Generalized Canonical Correlation Analysis) de Tenenhaus and Tenenhaus 2011, modifica el modo *A* de PLS-PM e implementa el *nuevo modo A* y constituye un marco general para el análisis multibloques proporcionando ahora algoritmos óptimos monótonamente convergentes a funciones de covarianzas o correlaciones bien definidas.

NM-RGCCA se crea, para cuantificar óptimamente las variables cualitativas de cada bloque con la respectiva componente *inner* proporcionada por RGCCA, heredando la convergencia monótona creciente y la maximización de las funciones de covarianza o correlación según el *nuevo modo A* o *B* elegido.

1.1. ESTRUCTURA DE LA TESIS

Este trabajo de tesis está estructurado en seis capítulos. El estado del arte se ha dividido en dos capítulos, el segundo y el tercero.

El capítulo dos recoge una importante revisión de los principales métodos clásicos multivariantes de datos métricos relacionados con los métodos PLS, tales como el ACP para el tratamiento de una matriz, el cual constituye la base fundamental en la comprensión de casi todos los demás métodos.

Para el estudio de dos matrices de datos se presentan los siguientes métodos: AC, AIB, AR y la Regresión Multivariante. El ACG trata el estudio de más de dos grupos.

En el capítulo tres se presentan los principales métodos PLS desarrollados por Herman Wold: NIPALS, PLS1, PLS2 y PLS-PM. Estos métodos se caracterizan porque pueden funcionar con datos faltantes, matrices esparcidas ($p > n$) y presencia de multicolinealidad.

El algoritmo NIPLAS básicamente desarrolla un ACP permitiendo obtener los valores y vectores propios (componentes principales) de una matriz de datos. Para estudiar la relación entre dos grupos de variables Y *e* X, Wold creó los métodos de regresión, PLS-R, cuyos parámetros son estimados por medio de los algoritmos PLS1 en el caso de una sola variable respuesta o PLS2 en el caso de respuesta múltiple.

Los métodos PLS-PM descritos por Wold (1975, 1982, 1985) en el análisis de Modelos de Ecuaciones Estructurales (SEM) para estudiar J bloques basados en componentes, operan sin supuestos distribucionales, lo cual los hacen muy atractivos para la estimación de modelos tipo regresión. PLS-PM es entonces una interesante alternativa a los SEM basados en el Análisis de Covarianza bajo multinormalidad, a través de su programa LISREL (Joreskog 1970).

El método PLS-PM es modificado mediante el *nuevo modo A* dando lugar al RGCCA (Regularized Generalized Canonical Correlation Analysis) de Tenenhaus (2011), el cual constituye un marco general para el análisis multibloques ($J > 2$) mediante algoritmos que asocian funciones de maximización monótonamente convergentes.

A partir de una red de conexiones entre los bloques, el objetivo del RGCCA es encontrar componentes PLS que expliquen sus propios bloques y que estén altamente correlacionadas con las de los bloques conectados precisando las funciones de maximización asociadas.

Al final de este capítulo tres se presenta una síntesis del proceso de cuantificación de cada variable cualitativa, que básicamente descansa en los conceptos de proyección ortogonal de la matriz indicadora asociada, de la maximización de la razón de correlación y en la regresión monótona.

Para el tratamiento de datos mixtos se han constituido los capítulos cuatro, cinco y seis que tratan respectivamente una, dos y más matrices de datos cuanti-cualitativos. Cada uno de estos capítulos recoge en primer lugar las versiones NM-PLS presentadas por Russolillo (2009), y luego sus correspondientes versiones generalizadas GNM-PLS objeto de esta tesis, para la transformación de las variables cualitativas a variables métricas conservando propiedades de pertenencia grupal y orden si es necesario.

Así, el capítulo cuatro presenta los conceptos fundamentales del algoritmo NM-NIPALS y luego su extensión al GNM-NIPALS que cuantifica las variables cualitativas de *una* matriz de datos con una función agregada de k (> 1) componentes proporcionada por ACP de los datos cuantitativos, ganando mayor poder descriptivo de las relaciones de variabilidad especialmente en el primer plano factorial.

Análogamente el capítulo cinco, inicia presentando los métodos de regresión NM-PLS1 y NM-PLS2 que luego son generalizados respectivamente con las versiones GNM-PLS1 y GNM-PLS2 que cuantifican las variables cualitativas por funciones agregadas de k componentes tipo regresión suministradas por los propios algoritmos PLS1 y PLS2. El método GNM-PLS1q se presenta como principal caso de interés para el cual la única variable respuesta es cualitativa.

El capítulo seis presenta la versión NM-PLSPM que relaciona J bloques conteniendo datos mixtos. Los procesos de cuantificación se desarrollan con las

funciones *inner* que constituyen funciones agregadas de las *J* estimaciones *outer* de cada uno de los bloques. Ya que en PLS-PM no son claras las funciones de maximización, Tenennhaus (2011) presentó el método RGCCA donde se muestra la convergencia monótona de las funciones de maximización asociadas con las componentes.

Se interviene adecuadamente las fases del algoritmo RGCCA para a partir de las estimaciones *inner* realizar la cuantificación de variables cualitativas en cada bloque; se obtiene así el método denominado NM-RGCCA considerado un caso de dimensión *J* dentro del hipotético GNM-RGCCA.

Todos los ejemplos de aplicación GNM-PLS toman parcial o totalmente la base de datos créditos denominada *cred.438* y el software utilizado ha sido desarrollado en el entorno de programación R; algunas de estas instrucciones se han escrito especialmente en los capítulos de aplicación por considerarse importantes e ilustrativos. Los programas asociados a los métodos GNM-NIPALS, GNM-PLS1, GNM-PLS2 y NM-RGCCA se encuentran adjuntos en los Anexos A, B, C y D.

La convergencia de las funciones en todos los escenarios es rápida, se requirieron no más de 15 iteraciones, sin embargo se definió 100 iteraciones en cada proceso para tener un amplio margen de resultados; para presentar las gráficas se tomó un número fijo de 20 iteraciones.

CAPÍTULO 2

MÉTODOS FACTORIALES CLASICOS

Con los métodos factoriales (Lebart 2006, Aluja 1999, Tenenhaus 1998, Escofier 1992) se propone obtener representaciones sintéticas de grandes conjuntos de datos, describiendo las asociaciones entre individuos y entre variables en espacios de dimensiones menores. Los métodos factoriales de análisis multivariante exploratorio de datos más utilizados se derivan de dos técnicas fundamentales que son el *ACP* y el Análisis de Correspondencias.

La matriz inicial de datos $R_{n,p}$ con *n* filas-nube individuos y *p* columnas-nube variables (puede ser una tabla de contingencia), generalmente es transformada (centrada: $X_{n,p}$, estandarizada o en perfiles) y sometida al análisis de datos. Su término general es x_{ij} ($i_{ésima}$ observación de la $j_{ésima}$ variable); $i = 1, \ldots, n$; $j = 1, \ldots, p$; note que un vector fila x_i', pertenece al espacio R^p y un vector columna X_j pertenece al espacio R^n .

La posición de los puntos en la nube está dada por el conjunto de *distancias* entre todos los puntos y determina la *forma* de la nube, la cual caracteriza la naturaleza e intensidad de las relaciones entre los individuos y entre las variables y revela las estructuras de la información contenida en los datos. Los cálculos de las proximidades geométricas entre puntos fila y entre puntos columna traducen en efecto las asociaciones estadísticas.

Las nubes compuestas por montones de puntos pueden ser alargadas (lineales), no lineales, esféricas (ausencia de relación), triangulares etc. Una manera de darse cuenta visualmente de la forma de una nube es proyectando los puntos sobre nuevos ejes (cambio de base) o planos que minimicen las deformaciones de las *distancias* debidas a la proyección: "Búsqueda de los principales ejes de representación, en los que la inercia (distancias ponderadas) de la nube proyectada sea máxima eje por eje".

En R^p, los vectores ortogonales $u_\propto$ que determinan la dirección de los ejes factoriales garantizan el máximo alargamiento de la nube de individuos N_I.

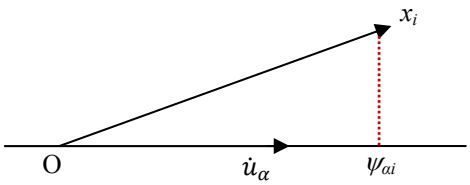
2.1. ANÁLISIS FACTORIAL GENERAL: LAS MATRICES X, M, N

En el cálculo de la distancia entre los puntos *i* y *l*, $d^2(i,l) = \sum_j m_j(x_{ij} - x_{lj})^2$, es posible considerar la importancia de las columnas mediante los coeficientes m_j que ponderan la influencia de cada columna *j* y se encuentran en la diagonal de la matriz *M* denominada *métrica*. Así, haciendo $r = i - l$ entonces

$$< r, r >_M = r'Mr = d_M^2(i,l) \ .$$

2.1.1. Proyección e inercia de la nube de individuos

La coordenada de la proyección de un punto *i* sobre el vector director $\dot{u}$ del eje α con métrica canónica *I* es $x_i'I\dot{u}_\alpha = x_i'\dot{u}_\alpha$, mientras que con una métrica cualquiera *M* está dada por el producto escalar $x_i'M\dot{u}_\alpha = \psi_{i\alpha}$ (gráfica 2.1). Las coordenadas de todos los puntos *i* sobre este mismo eje están contenidas en el factor $\psi_\alpha = XM\dot{u}_\alpha$. El vector $\dot{u}_\alpha$ bajo la métrica *M* es unitario; esto es, $\dot{u_\alpha}M\dot{u}_\alpha = 1$.



Gráfica 2.1. Proyección ortogonal del individuo *i* sobre $\dot{u}$

La inercia (~varianza) como medida de dispersión alrededor del centro de gravedad *G* para la nube de individuos en el hiperespacio es $IG_I = \sum_{i=1}^{n} p_i d_M^2(i,G)$. Los pesos p_i de los individuos se encuentran en la matriz diagonal *N*.

El objetivo geométrico del análisis factorial es buscar un nuevo sistema de ejes ortogonales en los que se *proyecte esta inercia* tal que los primeros ejes concentren la mayor parte de la misma y en forma decreciente. Generalmente, se quiere analizar una matriz de datos estandarizados $Z_{n,p}$ que garantiza propiedades interesantes como el de no afectación de las *p* variables por las unidades de medida. Si la matriz Z es de rango completo entonces

$$r(Z) = r(Z') = r(ZZ') = r(Z'Z) = IG_I = p \ .$$

Así, con datos de partida centrados, $x_{ij} = r_{ij} - \bar{r}_j$ el origen *O* coincide con el centro de gravedad *G* de la nube, y los elementos diagonales de la métrica $m_j = 1/s_j^2$ completarán el cálculo de distancia e inercia de los individuos (con variables estandarizadas). Las proyecciones de los individuos sobre $\dot{u}_\propto$ esta dada por $\psi_\alpha = XM\dot{u}_\alpha$ con media $E(\psi_\alpha) = 0$ debido al centramiento de las variables, y la *inercia proyectada* respecto al origen

$$IO_\alpha = \sum_i p_i[\psi_{\alpha i}]^2 = \psi'_\alpha N\psi_\alpha = \dot{u}'_\alpha MX'NXM\dot{u}_\alpha$$

es la cantidad a maximizar en cada eje *α,* bajo la restricción $\dot{u}'_\alpha M\dot{u}_\alpha = 1$. La derivada de la ecuación lagrangiana asociada, conduce a la descomposición espectral de la matriz $X'NXM$, la cual no es simétrica y por tanto no garantiza la ortogonalidad de los ejes. Sin embargo, haciendo la transformación $u = M^{1/2}\dot{u}_\alpha$ en la ecuación de la inercia IO_α , la matriz simétrica a diagonalizar es ahora $M^{1/2}X'NXM^{\frac{1}{2}}$ $bajo$ $u'u = 1$.

Para el caso en el que los datos han sido normalizados, $z_{ij} = (r_{ij} - \bar{r}_j)/s_j$, el centro de gravedad de N_I también coincide con el origen, $\psi_\alpha = Zu_\alpha$, $u'_\propto u_\propto = 1$ y por tanto la inercia de la nube de individuos es

$$IO_I = u'_\alpha Z'NZu_\alpha = \sum_{i=1}^n p_i d^2(i,O) = \sum_i \frac{1}{n} \sum_j z_{ij}^2 = \sum_j v(Z_j) = traza\,(Z'NZ) = p\ .$$

Posteriormente se mostrará que la inercia de la nube de individuos es idéntica a la nube de variables cuyo centro de gravedad no coincide con el origen; esto es:

$$IO_I = IG_J = p\ .$$

2.2. ANÁLISIS EN COMPONENTES PRINCIPALES – ACP

Presentado e integrado a la estadística matemática por Harold Hotelling en 1933; el ACP para la estadística clásica trata de la búsqueda de nuevos ejes principales (reducción), y desde el punto de vista reciente del análisis de datos se considera como una técnica de representación (descripción) de los datos que se utiliza sin referencia a hipótesis ni a un modelo estadístico en particular.

El ACP se aplica a tablas de datos donde las columnas son variables cuantitativas y donde las filas son los individuos u objetos de observación.

Las proximidades entre variables se interpretan en términos de *correlación* y las proximidades entre individuos en términos de *similitudes (semejanzas)*.

2.2.1. Análisis de la nube de los individuos

En R^p con datos no transformados, la distancia euclidiana entre parejas de individuos es: $d^2(i,i') = \sum_{j=1}^{p}(r_{ij} - r_{i'j})^2$. Dos puntos individuos están cerca si son caracterizados o toman valores casi iguales en cada una de las variables.

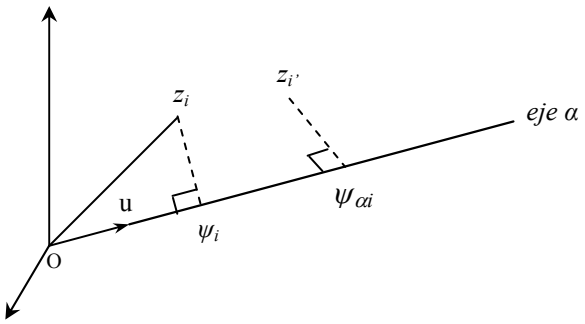
Para lograr la mejor representación, se busca inicialmente el *subespacio* en R^p de dimensión uno, un vector u en el eje α (u_α), sobre el que se maximice la suma de los cuadrados de las distancias entre todas las parejas de puntos (*i, i'*) proyectados: $\max_u \sum_i^n \sum_{i'}^n d^2(i,i')$.

De acuerdo con Lebart (1984), encontrar el máximo de la suma de cuadrados de las distancias entre todas las parejas de individuos es lo mismo que maximizar sobre u la suma de cuadrados de las distancias entre cada punto y el centro de gravedad de la nube: $\max_u \sum_i^n d^2(i,G)$. De ahí que el ACP requiera al menos centrar los datos, $x_{ij} = r_{ij} - \bar{r}_j$ lo cual equivale *trasladar el Origen a G;* se ha hecho coincidir el origen de los ejes y el centro de gravedad de la nube N_I centrada.

En el cálculo de la distancia entre individuos pueden existir variables con escalas muy diversas (segundos, metros, kilogramos …) y se desea hacer jugar a cada variable un papel idéntico en la definición de las proximidades entre individuos. Para ello se corrigen las escalas reduciendo las variables ya centradas así: $z_{ij} = (r_{ij} - \bar{r}_j)/s_j$; en este caso $Z_j \sim (0, 1)$.

Entonces bajo esta transformación (normalización) que ya contiene los pesos de las columnas, con métrica la matriz idéntica $M = I$ y N la matriz de pesos de los individuos, se desarrolla el análisis en *componentes principales normado* diagonalizando $Z'NZ$ que es la matriz de correlaciones entre las variables j, j' .

Geométricamente:



Gráfica 2.2. Proyección del individuo *i* sobre *u* ($\psi_{i=Zi.u}$) en el nuevo eje α

El teorema de Pitágoras aplicado a cada uno de los *n* triángulos rectángulos del tipo $Oz_i\psi_i$ de la gráfica 2.2 conduce a la relación

$$\sum_{i=1}^{n} z_i\psi_i^2 = \sum_i Oz_i^2 - \sum_i O\psi_i^2 \ .$$

Premultiplicando por *1/n*, y ya que $\frac{1}{n}\sum_i Oz_i^2$ es la cantidad fija (observada), minimizar $\frac{1}{n}\sum_i z_i\psi_i^2$ (proyección óptima) es equivalente a maximizar la cantidad $\frac{1}{n}\sum_i O\psi_i^2 = \psi' N\psi = u'Z'NZu$ que es la *inercia (varianza) de la nube* N_I *proyectada* sobre el eje *α* de dirección *u,* pues debido al centrado en $X, E(\psi) = 0$.

La longitud de la proyección ortogonal del vector z_i sobre el subespacio de una dimensión con vector director unitario *u* es $\|O\psi_i\| = \|c_i.u\| = \|c_i\| = |z_i'u|$; por tanto $c_i = z_i'u$ es el escalar[1] que dilata o contrae *u* hasta alcanzar el punto ψ_i ; $u \in R^p$.

Buscar el máximo de $\frac{1}{n}\sum_i O\psi_i^2$ equivale a encontrar *u* tal que

$$\max_u(u'Z'NZu) \quad \text{bajo} \quad u'u = 1 \ .$$

La derivada de la función lagrangiana asociada $L(u) = u'Z'NZu - \lambda(u'u - 1)$, $\partial L/\partial u = 0$, conduce a resolver el sistema de valores y vectores propios

$$Z'NZu = \lambda u \ . \qquad [2.1]$$

[1] El producto escalar ortogonal $(z_i\text{-}c_iu)'(c_i.u) = 0 \Rightarrow c_iz_i'u - c_i^2u'u = 0 \Rightarrow z_i'u = c_i$.

Sea u_1 el vector propio correspondiente al mayor valor propio λ_1 de $Z'NZ$ que da ese máximo. El subespacio en R^p de dos dimensiones $(u_1\ ,\ u_2\ )$ que mejor se ajusta a la nube contiene ortogonalmente $(u_1'u_2 = 0)$ al subespacio u_2 con su correspondiente $\lambda_2 \leq \lambda_1$ bajo $u_2'u_2 = 1$. Se busca de manera análoga el mejor subespacio de dimensión $q(\leq p)$ que recoja gradual y decrecientemente las proporciones de inercia proyectadas.

El análisis efectúa una *traslación* y *rotación* alrededor del origen y obtiene un sistema de vectores ortonormados $u_1\ ,\ u_2\ ,\dots,\ u_p$ que pasan lo más cerca de la nube; es decir, se diagonaliza la matriz de correlaciones $Z'NZ = UDU'$; U es ortogonal $(U'U = I)$ conteniendo los vectores propios $u_\propto$ y D es diagonal con los valores propios λ_α de $Z'NZ$.

Las coordenadas de los *n* puntos individuos proyectados sobre el eje 'canónico' u_α son los *n* componentes del vector (variable) $\psi_\alpha = Zu_\alpha$ el cual es una combinación lineal de las variables iniciales Z_j .

Si $\alpha = 1$, $\psi_1 = u_{11}Z_1 + u_{21}Z_2 + \cdots + u_{p1}Z_p$ es la primera componente principal, los pesos $u_{p\alpha}$ asociados a las variables, las ubicarán al lado del eje con mayor correlación (coordenada) y los individuos con valores máximos o mínimos en estas variables seguirán sus direcciones ubicándose en los extremos correspondientes. Note que como $E(\psi_\alpha) = 0$, $v(\psi_\alpha) = \psi_\alpha' N\psi_\alpha = \lambda_\alpha$ así, cada *componente* explica una proporción de variabilidad total $\lambda_1 + \lambda_2 + \cdots + \lambda_p = p$.

2.2.2. Análisis de la nube de puntos variables (Estandarizadas)

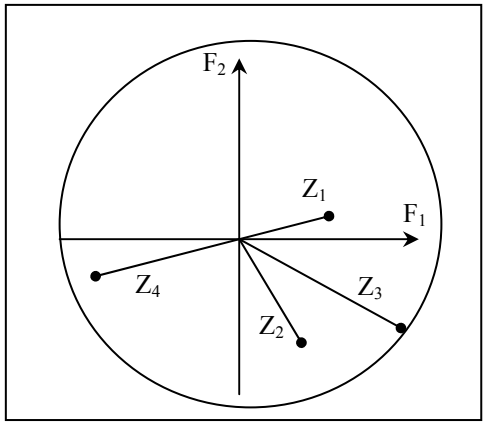
En R^n el análisis de los puntos variable se hace con referencia al origen *O*; la distancia $r = j - k$ entre dos variables es

$$d^2(j,k) = r'Nr = \sum_{i=1}^{n} \frac{1}{n}(z_{ij} - z_{ik})^2 = 2(1 - r_{jk})\ ,$$

lo cual implica[2] $0 \leq d^2\ (j,k) \leq 4$. Como las variables fueron estandarizadas están a una distancia 1 del origen *O* esto es $d^2(j,O) = j'Nj = \|j\| = 1$, por tanto todos los puntos variables están sobre una hiperesfera de radio 1 centrada en el origen.

[2] Dos variables fuertemente correlacionadas están muy próximas la una de la otra *(r_{kj} =1),* o tan alejadas como sea posible *(r_{kj} = -1)* según la relación lineal.

Los planos de ajuste cortarán la hiperesfera siguiendo círculos al interior de los cuales se encuentran proyectados los puntos variables y su estructura de correlaciones (ver gráfica 2.3). Así, los puntos mejor representados son los más próximos al círculo de correlaciones sobre el plano factorial:



Gráfica 2.3. Círculo de correlación en el plano F_1, F_2

El *coseno* del ángulo θ de dos vectores variables ('estandarizadas') es el coeficiente de *correlación* entre estas variables, y como tienen norma 1, tal coseno es el producto escalar:

$$cos\theta = \frac{k'Nj}{\|k\| \, \|k\|} = k'.j = cor(k,j) \quad .$$

Procediendo en forma análoga al espacio de los individuos, en R^n las variables con métrica N se proyectan sobre el vector $\dot{v}$ tal que $\varphi = Z'N\dot{v}$ e inercia $\dot{v}'NZZ'N\dot{v} = \lambda$ a maximizar bajo la restricción $\dot{v}'Nv = 1$, lo cual implica diagonalizar la matriz no simétrica asociada al sistema $ZZ'N\dot{v} = \lambda\dot{v}$.

Sin embargo, bajo la transformación $v = N^{1/2}\dot{v}$ se debe resolver el sistema-simétrico

$$N^{1/2}ZZ'N^{1/2}v = \lambda v, \quad \text{bajo} \quad v'v = 1 \quad . \qquad [2.2]$$

2.2.3. Relaciones de transición

Realizando el cambio $Y = N^{1/2}Z$ en [2.1] se tiene $Y'Yu = \lambda u$, y en [2.2] $YY'v = \lambda v$ esto es $Y'Y\underbrace{Y'v} = \lambda\underbrace{Y'v}$, con lo cual $u = kY'v$, y por tanto $u'u = k^2v'YY'v \rightarrow 1 = k^2\lambda,\ k = 1/\sqrt{\lambda}$ entonces

$$u = \tfrac{1}{\sqrt{\lambda}}Y'v = \tfrac{1}{\sqrt{\lambda}}\ Z'N^{1/2}v$$

y se tiene que la solución al sistema con el mismo valor propio λ_α está dada por:

$$v_\alpha = \tfrac{1}{\sqrt{\lambda_\alpha}}N^{\frac{1}{2}}Zu_\alpha = \tfrac{1}{\sqrt{\lambda_\alpha}}N^{\frac{1}{2}}\psi_\alpha \quad ,\ v \in R^n \ .$$

Igual que en el espacio fila, se busca la combinación lineal más relacionada con el conjunto de variables originales que maximiza la inercia proyectada, obteniendo las coordenadas de los puntos variables como sus proyecciones sobre el α-ésimo eje factorial v_α , mediante

$$\varphi_\alpha = Z'N^{1/2}v_\alpha = \tfrac{1}{\sqrt{\lambda_\alpha}}Z'NZu_\alpha = \sqrt{\lambda_\alpha}u_\alpha \ . \qquad [2.3]$$

Retomando la parte central de [2.3], la coordenada de la *j*-ésima variable es

$$\varphi_{j\alpha} = \tfrac{1}{\sqrt{\lambda_\alpha}}Z'N\psi_\alpha = \textstyle\sum_{i=1}^{n}\frac{r_{ij}-\bar{r}_j}{s_j}\ \frac{1}{n}\ \frac{\psi_{\alpha i}}{\sqrt{\lambda_\alpha}} = cor(j,\psi_\alpha)\ .\ \text{Además}\ |\varphi| \leq 1\ .$$

La coordenada de una variable sobre un eje α es el *coeficiente de correlación* de la variable con el factor ψ_α . Así, este nuevo vector estará los más correlacionado posible con las variables de origen, con lo cual se sabe que en el eje α $v(\varphi_\alpha) = \varphi'_\alpha\varphi_\alpha = \sum_j cor^2(\psi_\propto, Z_j) = \lambda_\alpha$ es máxima.

- Otras relaciones de interés:

$$u_\alpha = \tfrac{1}{\sqrt{\lambda_\alpha}}Z'N^{1/2}v_\alpha \quad ,\quad u_{j\propto} = \tfrac{1}{\sqrt{\lambda_\lambda}}\varphi_{j\alpha} = \tfrac{1}{\sqrt{\lambda_\lambda}}cor(j,\psi_\alpha)\ ,$$

$$\psi_\alpha = Zu_\alpha = \tfrac{1}{\sqrt{\lambda_\alpha}}ZZ'N^{1/2}v_\alpha = \sqrt{\lambda_\alpha}N^{-1/2}v_\alpha \ .$$

Observe que los factores φ_α e ψ_α son colineales a los vector propio u_α e v_α .

La tabla 2.1, presenta una síntesis de las fases del desarrollo factorial para el Análisis de Componentes Principales normado.

ACP normalizado $z_j = (r_j - \bar{r}_j)/s_j$: Resumen Factorial		
	Nube individuos $\in R^p$	**Nube variables $\in R^n$**
Métrica , pesos	$M = I \ , \ N = diag(1/n)$	$N \ , \ M = I$
Componente Principal	$\psi = ZMu = Zu$	$\varphi = Z'N^{1/2}v$
Inercia (I_O) proyectada	$\psi'N\psi = u'Z'NZu$	$\varphi'M\varphi = \varphi'\varphi = v'N^{1/2}ZZ'N^{1/2}v$
Matriz a diagonalizar	$Z'NZ$	$N^{1/2}ZZ'N^{1/2}$
Relaciones transición	$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}Z'N^{1/2}v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\varphi_\alpha$	$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}N^{1/2}Zu_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}N^{1/2}\psi_\alpha$
Valores propios	$IO = \lambda_1 + \lambda_2 + \cdots + \lambda_p = p$	

Tabla 2.1. Resumen factorial del ACP normalizado. Note que $z_j'z_j$ ya contiene la métrica $1/s_j^2$

2.2.4. Reconstitución de los datos

Teniendo los valores propios λ_α ordenados en forma decreciente, la tabla Z se puede reconstituir exactamente a partir de $\psi = ZU, \ Z = \ \psi U' = \lambda^{1/2}N^{-1/2}VU'$ mediante la fórmula:

$$N^{1/2}Z = \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}\, v_\alpha\, u_\alpha' = VDU' \quad \text{ó} \quad Z = \sum_\alpha \psi_\alpha u_\alpha' \ ,$$

que es la descomposición de la matriz $N^{1/2}Z$ en valores singulares, $\sqrt{\lambda_\alpha}$, contenidos en la diagonal de la matriz *D*. *V* e *U* contienen *p* vectores propios v_α e u_α respectivamente tal que $V'V = I_p$ e $UU' = U'U = I_p$.

Si los *p-q* valores propios más pequeños se juzgan "despreciables" se puede limitar la suma a los *q* primeros términos correspondiente a los valores propios $\lambda_1 , \lambda_2 , \ldots , \lambda_q$:

$$N^{1/2}Z \approx N^{1/2}Z^* = \sum_{\alpha=1}^{q}\sqrt{\lambda_\alpha}\, v_\alpha\, u_\alpha' \ .$$

La calidad de la reconstitución se puede evaluar por la cantidad

$$\tau_q = \frac{\text{traza } Z^{*\prime} N Z^*}{\text{traza } Z' N Z} = \frac{\sum_{\alpha=1}^{q} \lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha} \ .$$

El coeficiente τ_q , inferior o igual a 1, se llamará *tasa de inercia* o aún *porcentaje de varianza relativa a los q primeros factores*. Observe que $\Sigma\, \lambda_\alpha = p$.

2.2.5. Elementos suplementarios

Intervienen a posteriori del análisis activo para enriquecer y caracterizar aún más los ejes. Para ubicar los individuos suplementarios con respecto a los otros en el espacio R^p también deben ser transformados mediante $z_{+ij} = (r_{+ij} - \bar{r}_j)/s_j$. Las coordenadas de estos puntos suplementarios se obtienen con el producto $Z_+ u_\alpha$.

En el caso de variables suplementarias continuas es necesario efectuar la transformación $z_{ij}^+ = (r_{ij}^+ - \bar{r}_j^+)/s_j^+$ con el fin de interpretar las distancias en términos de correlaciones. Las coordenadas de las variables suplementarias son entonces las componentes del vector $\varphi_\alpha^+ = Z^{+\prime} N^{1/2} v_\alpha$ y cada una corresponde a la correlación entre la variable y el factor.

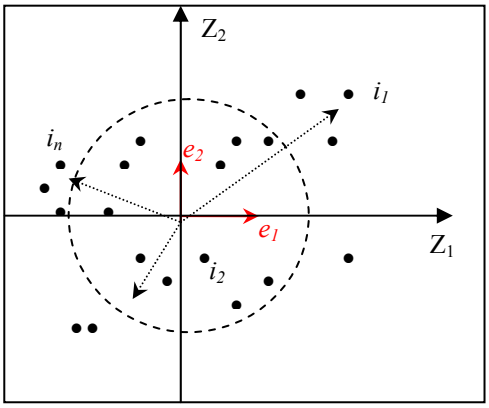
Si la variable es nominal con *m* modalidades, se procederá a representar como individuos suplementarios los centros de gravedad de las clases de individuos conformadas por cada modalidad. Si se tiene millares de individuos conviene hacer un clúster y proyectar los individuos medios de cada clase como suplementarios.

2.2.6. Representación simultánea

Las proximidades entre individuos se interpretan en términos de similitudes de comportamiento con respecto a las variables. No se debe interpretar la distancia que separa un individuo de una variable ya que estos puntos no parten de un mismo espacio. La superposición de los dos planos factoriales carece de sentido.

No obstante si no se toman los puntos variables y por tanto sus correlaciones sino más bien sus direcciones en el gráfico de representación de los individuos, se puede considerar la representación simultánea, en este espacio R^p, a la vez a los individuos y a las variables representadas como vectores directores (gráfica 2.4).

La posibilidad de una representación simultánea reside entonces en la proyección (en fila suplementaria) del antiguo eje canónico (en R^p) e_j sobre el nuevo eje u_α, cuya coordenada vale $e_j' u_\alpha = u_{j\alpha}$.



Gráfica 2.4. Individuos y variables como vectores directores a ser representados conjuntamente

De las relaciones de transición, el análisis de la nube de variables se deduce del de la nube de individuos. Recuérdese que en R^n la coordenada de la variable *j* sobre el eje α es $\varphi_{j\alpha} = \sqrt{\lambda_\alpha}\, u_{j\alpha}$.

Las dos nubes de variables se diferencian por una dilatación $\sqrt{\lambda_\alpha}$ definida sobre cada eje. La interpretación de distancia entre dos variables sólo se puede hacer en R^n. Sobre la representación simultánea, es lícito comparar las posiciones relativas de dos individuos con respecto al conjunto de variables, o dos variables respecto al conjunto de individuos.

La dirección de una variable define zonas para los individuos: de un lado, aquellos que toman valores fuertes para esta variable y, al lado opuesto, aquellos que toman valores bajos. Nos interesará el alargamiento de los individuos en la dirección de la variable.

2.2.7. Elementos de ayuda para la interpretación

Los ejes factoriales permiten obtener de hecho en R^p, la mejor visualización aproximada de las distancias entre los individuos de una parte y de otra parte las

variables que a su vez califican y caracterizan tanto a los ejes como a los individuos.

Recuerde que el valor propio es la varianza o inercia proyectada de la nube asociada a un factor; en un ACP normado, la suma de las inercias es igual al número *p* de variables y por tanto la inercia media vale 1. Interesarán en general aquellos ejes cuya inercia sea "notablemente" superior a uno; en este caso se observará un decrecimiento *irregular* de los primeros valores propios.

Los porcentajes de inercia de los ejes definen los "poderes explicativos" de los factores: ellos representan la parte de la varianza total tenida en cuenta por cada factor. Su validez debe entonces tener en cuenta el número de variables y su forma de codificación. Un % de inercia del 10% es débil si la tabla contiene 10 variables pero es fuerte en el caso de 100 variables. La importancia de un factor también puede depender de variables exógenas o suplementarias.

Las variables fuertemente correlacionadas con un eje van a contribuir a la definición (calificación) de ese eje. Interesarán entonces aquellas variables que presenten coordenadas altas y que se sitúen más cerca del círculo de correlación, interpretando las componentes principales en función del reagrupamiento de ciertas variables y la oposición de otras. La correlación (*cos(θ)*) entre dos variables será conservada en proyección según la calidad del ajuste. No se debe interpretar la distancia entre variables alejadas del círculo de correlación.

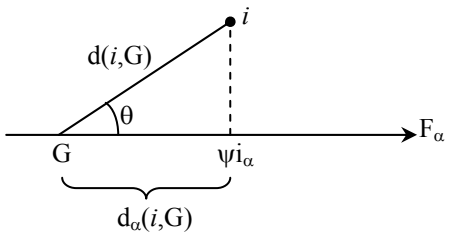
De la gráfica 2.5 se puede deducir para las *variables* el coseno (cateto adyacente/ hipotenusa), como sigue:

$$cos^2(j,\alpha) = \frac{\varphi_{j\alpha}^2}{d^2(j,G)} = \varphi_{j\alpha}^2 = cor^2(j,\alpha) \ .$$

La inercia proyectada sobre el eje α de R^n es $\lambda_\alpha = \sum_j \varphi_{j\alpha}^2$ por tanto interesa que la contribución de una variable *j* a la inercia del eje $Ca_\alpha(j) = \varphi_{j\alpha}^2/\lambda_\alpha$ sea alta.

Si *los individuos* no son anónimos para el estudio, interesa conocer quienes participaron en la formación de los ejes.

Se consideran las *contribuciones* relativamente más *fuertes* de los puntos *i* (de masa p_i) a la inercia del eje α: $Ca_\alpha(i) = p_i\, \psi_{i\alpha}^2/\lambda_\alpha$. Los individuos que más *contribuyen* a la determinación del eje son los más excéntricos y el examen de las coordenadas factoriales o la lectura de la gráfica 2.5 son suficientes.



Gráfica 2.5. Representación de las contribuciones y cosenos de los individuos

La *calidad* de la representación está definida por:

$$cos^2(i,\alpha) = \frac{d_\alpha^2(i,G)}{d^2(i,G)} = \frac{\psi_{i\alpha}^2}{d^2(i,G)} \ .$$

Note que $\sum_i Ca_\alpha(i) = 1$, $\sum_\alpha cos^2(i,\alpha) = 1$; la suma de las contribuciones de todos los individuos y de los cosenos sobre todos los ejes α es 1.

Interesan para el análisis de individuos y variables, aquellos puntos con contribuciones y cosenos altos; son puntos excéntricos de alta calidad responsables en buena parte de la formación de los ejes.

Cuando todas las variables están correlacionadas positivamente entre ellas, se sitúan del mismo lado de un eje factorial. Esta característica que aparece con más frecuencia sobre el primer eje, se llama *Factor tamaño* y permite el cálculo de otros índices estadísticos.

2.3. ANÁLISIS CANÓNICO (AC)

El AC estudia las relaciones entre dos grupos de variables. Así, contiene como caso particular la regresión múltiple si uno de los dos grupos está reducido a una sola variable numérica, al análisis discriminante cuando la variable a explicar es nominal y al análisis de correspondencias cuando los dos grupos son nominales.

El método de AC desarrollado por Hotelling (1936) busca sintetizar las interrelaciones existentes entre dos grupos de variables, identificando la combinación lineal de las variables del primer grupo más correlacionada a la combinación lineal del segundo grupo.

La tabla de datos *R* con *n* líneas y $p + q$ columnas es particionada en dos subtablas *X* e *Y* con *p* y *q* columnas respectivamente.

$$R = [X \vdots Y] \ .$$

Se supone que las variables están *centradas-reducidas;* entonces la *matriz de correlaciones* de las $p + q$ variables se escribe

$$V(R) = \frac{1}{n} R'R = \frac{1}{n} \begin{bmatrix} X'X & X'Y \\ Y'X & Y'Y \end{bmatrix} \ .$$

Sean *a* y *b* dos vectores con *p* y *q* componentes, definiendo las combinaciones lineales

$$Xa = \sum_{j=1}^{p} a_j X_j = a_1 X_1 + \cdots + a_p X_p$$

$$Yb = \sum_{k=1}^{q} b_k Y_k = b_1 Y_1 + \cdots + b_q Y_q \ .$$

Los vectores $Xa=t$ e $Yb=u$ pertenecen a R^n ya que son combinaciones de las columnas de X_j e Y_k que también $\epsilon\ R^n$ y engendran los espacios V_X e V_Y .

Se propone encontrar en principio las dos combinaciones lineales *t* e *u más correlacionadas* sobre el conjunto de datos. Como las variables están centradas las combinaciones lineales también lo están.

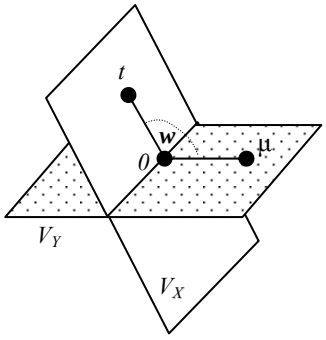
Se impone a las dos combinaciones tener una *varianza unitaria*; esto es:

$$V(t) = V(Xa) = \frac{1}{n} a'X'Xa = 1 \ , \ V(u) = V(Yb) = \frac{1}{n} b'Y'Yb = 1 \ . \qquad [2.4]$$

Bajo estas condiciones, se maximiza *r* el coeficiente de correlación entre las combinaciones lineales, que coincide con la covarianza, es decir

$$r = cor(Xa, Yb) = cor(t, u) = E(t'.u) = \frac{1}{n} a'X'Yb \ .$$

Ya que los datos en ambos grupos están centrados, el coeficiente de correlación no es otro que el coseno del ángulo *w* entre los subespacios V_X e V_Y. La búsqueda de los coeficientes (vectores o variables canónicas) *a* y *b* conducen a minimizar el ángulo *w* entre los subespacios V_X e V_Y. (Ver gráfica 2.6).



Gráfica 2.6. Angulo ***w*** entre los subespacios V_X e V_Y

2.3.1. Cálculo de las variables canónicas

Dos multiplicadores de lagrange λ , μ intervienen en el desarrollo, tal que:

$$L = \tfrac{1}{n}a'X'Yb - \lambda\,(\tfrac{1}{n}a'X'Xa - 1) - \mu(\tfrac{1}{n}b'Y'Yb - 1) \ .$$

Las derivadas del lagrangiano ∂L respecto a los vectores *a* y *b* conducen a

$$\tfrac{1}{n}X'Yb - 2\lambda\,\tfrac{1}{n}X'Xa = 0\,; \qquad \tfrac{1}{n}Y'Xa - 2\mu\tfrac{1}{n}Y'Yb = 0 \ . \qquad [2.5]$$

Premultiplicando los dos miembros de las relaciones anteriores por *a'* y *b'* respectivamente y teniendo en cuenta las restricciones en [2.4], se obtiene

$$\tfrac{1}{n}a'X'Yb = 2\lambda \ , \ \tfrac{1}{n}b'Y'Xa = 2\mu \ .$$

Por consiguiente $\lambda = \mu$. Note que $r = 2\,\lambda$ es el valor del coeficiente de correlación maximal buscado.

Así, el sistema [2.5] se escribe

$$X'Yb = rX'Xa \ , \ Y'Xa = rY'Yb \ . \qquad [2.6]$$

Si las matrices $X'X$ e $Y'Y$ son inversibles la solución es inmediata ya que

$$a = \frac{1}{r}(X'X)^{-1}X'Yb \qquad [2.7]$$

que reemplazando en [2.6] conduce a:

$$(Y'Y)^{-1}Y'X(X'X)^{-1}X'Yb \;=\; r^2 b \qquad [2.8]$$

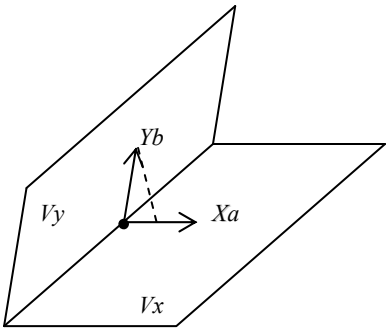
y entonces en [2.8] *b* es *vector propio* relativo al *valor propio* más grande r^2; de manera análoga *a* es *vector propio* del sistema

$$(X'X)^{-1}\,X'Y(Y'Y)^{-1}Y'Xa = r^2 a \qquad [2.9]$$

de [2.7] se obtiene $Xa = \frac{1}{r}\,X(X'X)^{-1}\,X'Yb = \frac{1}{r}P_x Yb$. [2.10]

Análogamente $Yb = \frac{1}{r}\,Y(Y'Y)^{-1}Y'Xa = \frac{1}{r}P_y Xa$. [2.11]

Así en [2.10], $Xa = t$ es colineal a la proyección ortogonal de $Yb = u$ sobre el espacio engendrado por las columnas de *X*. Análogamene para *Yb* en [2.11].



Premultiplicando [2.10] por $\frac{1}{n}a'X'$ se obtiene:

$$r\underbrace{\tfrac{1}{n}a'X'Xa}_{[2.4]} = \tfrac{1}{n}a'X'Yb \;= cor\,(Xa, Yb) = r$$

De las propiedades de la regresión múltiple, se define *R* como el *cos(w)* que es la correlación entre *t* e *u*. En la *h-ésima* etapa, $R^2(Yb_h; Xa_h) = \; cor^2(Yb_h, Xa_h) = r_h^2$

Así, *r* es el coeficiente de correlación canónica (ccc), y r_h el ccc asociado a las *h-ésimas* componentes.

Considérese ahora las *componentes canónicas* $t = Xa$ e $u = Yb$ reducidas $(t't = u'u = n)$; de [2.6], se tiene

$$\underbrace{\tfrac{1}{n}X'u}\; = \; \tfrac{1}{n}rX't \;\; ; \;\; \tfrac{1}{n}Y't \; = \; r\tfrac{1}{n}Y'u \quad \text{es decir}$$

$$\underbrace{cov(X_j\,, u)} = r\; cov(X_j\,, t) \;\; ; \;\; cov(Y_k\,, t) = r\; cov(Y_k\,, u) \quad k = 1,.,q \;\; . \qquad [2.12]$$

Así, el vector de correlaciones entre las variables X_j y la componente canónica del otro grupo *u* es colineal al vector de correlaciones entre los X_j y la componente canónica *t*. Análogamente en Y_k, ver [2.12].

Si se nota $R_{11} = \frac{1}{n} X'X$, $R_{22} = \frac{1}{n} Y'Y$ como las correlaciones intragrupos y $R_{12} = \frac{1}{n} X'Y$ con $R_{21} = R'_{12}$ las correlaciones intergrupos, entonces [2.8] y [2.9] se pueden escribir como:

$$R_{22}^{-1}\ R_{21}\ R_{11}^{-1}\ R_{12}\ b = r^2 b\ ,\ R_{11}^{-1}\ R_{12}\ R_{22}^{-1}\ R_{21}\ a = r^2 a\ . \qquad [2.13]$$

En general estas matrices no son simétricas. Su simetrización garantizará la ortogonalidad de los componentes canónicos t_h de orden diferente, para ello se premultiplica el lado derecho en [2.13] por $R_{11}^{1/2}$ y se obtiene:

$$R_{11}^{-1/2}\ R_{12}\ R_{22}^{-1}\ R_{21}\ R_{11}^{-1/2}\ \underbrace{R_{11}^{1/2} a}_{\tilde{a}} = r^2\ \underbrace{R_{11}^{1/2} a}\ . \qquad [2.14]$$

Así, los *vector propio* $R_{11}^{1/2} a$, asociados a los *valor propio* no nulos r^2 son ortogonales y se pueden escribir como $\tilde{a} = R_{11}^{1/2} a$ de donde $a = R_{11}^{-1/2} \tilde{a}$; la variable canónica *a* es el *vector propio* asociado en [2.9] y en [2.13].

Además, $\frac{1}{n} t'_h t'_l = \frac{1}{n} a'_h X'X a_l = a'_h\ R_{11} a_l = a'_h R_{11}^{1/2} R_{11}^{1/2} a_l = 0$ con lo cual se prueba la ortogonalidad de los t_h . Se muestra de la misma forma la ortogonalidad de los u_h .

En fin, también se tiene ortogonalidad entre los t_h y u_l

$$t'_h u_l = a'_h X' u_l = a'_h X' Y b_l = a'_h \underbrace{r_l X'X a_l}_{(7)} = a'_h r_l X' t_l$$

$$= r_l a'_h X' t_l = r_l t'_h t_l = 0 \quad ; \quad h \neq l\ .$$

Si se denomina la matriz simétrica dada en [2.14] por $Z'Z$ como en ACP, la descomposición singular de *Z* es:

$$\underbrace{R_{11}^{-1/2} R_{12} R_{22}^{-1/2}}_{Z} = \sum_h r_h\ \underbrace{R_{11}^{1/2} a_h}_{u}\ \underbrace{b'_h R_{22}^{1/2}}_{v'}$$

que premultiplicando por $R_{11}^{1/2}$ y pos multiplicando por $R_{22}^{1/2}$ conlleva a:

$$R_{12} = \sum_h r_h \, R_{11} a_h b_h' R_{22} \qquad [2.15]$$

con $R_{11}a_h = \frac{1}{n}X'Xa_h = \frac{1}{n}X't_h = \frac{1}{nr_h}X'u_h$; análogamente $R_{22}b_h = \frac{1}{nr_h}Y't_h$

por tanto de [2.15]

$$cor(x_j, y_k) = \sum_h \tfrac{1}{r_h} cor(x_j, u_h).cor(y_k, t_h) \underbrace{=}_{[2.12]} \sum_h cor(x_j, t_h).cor(y_k, t_h)$$

permitiendo la representación gráfica (biplot) de ambos conjuntos de variables.

2.3.2 Interpretación de resultados

Se describe en esta sección los conceptos básicos de *redundancia, comunalidad y componentes canónicas* útiles en la interpretación de los resultados.

2.3.2.1. Varianza explicada (Redundancia)

La proximidad entre un grupo de variables *X* y una componente *t* es medida en ACP por la proporción de inercia de *X* explicada por la variable *t,* denotada: $p(X,t) = \frac{1}{p}\sum_j^p cor^2\left(x_j, t\right) = \lambda_t/p$, por tanto

- Parte de la varianza de *X* explicada por su componente t_h y la componente u_h del otro grupo (*redundancia* de *X* con relación a u_h) es respectivamente:

$$Rd(X, t_h) = \tfrac{1}{p}\underbrace{\sum_j^p cor^2\left(x_j, t_h\right)}_{\lambda_{xt_h}}$$

$$Rd(X, u_h) = \tfrac{1}{p}\sum_j^p cor^2\left(x_j, u_h\right) = r_h^2 Rd(X, t_h) \ .$$

La generalización a *m* componentes u_h es:

$$Rd(X, u_1, \dots, u_m) = \tfrac{1}{p}\sum_j^p R^2(X_j, u_1, \dots, u_m) \underbrace{=}_{ort} \tfrac{1}{p}\sum_j^p \sum_h^m r^2(X_j, u_h) \ .$$

- Igualmente, parte de la varianza de *Y* explicada por su componente u_h e t_h del otro grupo (*redundancia* de *Y* respecto a t_h) es:

$$Rd(Y, u_h) = \frac{1}{q}\sum_k^q cor^2\,(y_k, u_h)$$

$$Rd(Y, t_h) = \frac{1}{q}\sum_k^q cor^2\,(y_k, t_h) = r_h^2 Rd(Y, u_h)\ .$$

Por tanto, y derivado de la generalización, parte de la varianza de *X e Y* explicadas por *m* componentes u_h y t_h:

$$\underbrace{\sum_h^m Rd(X, u_h)}_{\sum \lambda_{xu_h}} \quad \text{y} \quad \sum_h^m Rd(Y, t_h)\ .$$

2.3.2.2. Comunalidad Intra-grupo

Se puede medir la parte de varianza de cada variable explicada en sus *m* componentes canónicas a retener, y estos índices son llamados *comunalidades* como en análisis factorial. Así, de [2.12] y de $v(t) = v(u) = 1$ entonces,

$$R^2(x_j, t_1, \dots, t_m) = \sum_h^m cor^2(x_j, t_h) = \sum_h^m \frac{1}{r_h^2} cor^2(x_j, u_h)$$

$$R^2(y_k, u_1, \dots, u_m) = \sum_h^m cor^2(y_k, u_h) = \sum_h^m \frac{1}{r_h^2} cor^2(y_k, t_h)\ .$$

Las variables con comunalidad intra-grupo débil, participan poco en el estudio, puesto que ellas están poco relacionadas con las variables activas del otro grupo.

2.3.2.3. Comunalidad Inter-grupo

Está definida como la varianza *cruzada,* es decir, la varianza de cada variable explicada en las *m* componentes del otro grupo; esto es:

$$R^2\big(x_j, u_1, \dots, u_m\big) = \sum_h^m cor^2\big(x_j, u_h\big)$$

$$R^2(y_k, t_1, \dots, t_m) = \sum_h^m cor^2(y_k, t_h)\ .$$

Las variables con poca comunalidad intergrupos son específicas de su propio grupo, ellas están poco relacionadas con el otro grupo; se pueden suprimir estas variables sin perturbar el análisis.

Nota: Suponga que $s = \min(p, q)$ entonces las componentes $u_1, \dots, u_s$ engendran el mismo espacio que *Y* y por consiguiente *t* con *X*; entonces:

$$R^2(x_j, u_1, \dots, u_m) = R^2(x_j, y_1, \dots, y_q) \quad , \quad R^2(y_k, t_1, \dots, t_s) = R^2(y_k, x_1, \dots, x_p) \quad .$$

Además si *s=p, R_{12}* es de rango pleno y globalmente:

$$Rd(X, Y) = Rd(X, u_1, \dots, u_s) \quad , \quad Rd(Y, X) = Rd(Y, t_1, \dots, t_s) \quad .$$

2.3.2.4. Interpretación de las componentes canónicas

A nivel de las variables. En ACP hay proporcionalidad entre el *vector propio* y el vector de correlaciones de la variable y la *componente principal,* debido a la relación $\sqrt{\lambda}u = cor(X_j, \psi_\alpha) = \varphi_\alpha$. Así, las coordenadas de los vectores canónicos *a* o *b* miden el aporte marginal de cada variable a la construcción de las componentes canónicas *t* o *u* de su grupo.

Ellas pueden tener un signo y un nivel diferentes de la correlación entre las variables y sus componentes canónicas la cual puede reflejar adecuadamente las estructuras originales de correlación intragrupos. Si hay multicolinealidad intragrupo es muy difícil interpretar su vector canónico.

El AC de *X e Y* es equivalente al AC de las *componentes principales* de *X* y de las de *Y*, conduciendo a las mismas componentes canónicas. Pero la ortogonalidad de las componentes principales permite relacionar directamente los coeficientes de regresión a_j o b_k como las correlaciones entre las *componentes principales* y las componentes canónicas. Es pues muy ilustrativo interpretar las componentes canónicas proyectándolas sobre los círculos de correlación obtenidos del ACP de *X* y de *Y*.

A nivel de individuos. Las cartas $(t_1, u_1), \dots, (t_h, u_h)$ permiten visualizar la relación entre las CC y situar los individuos dentro de la relación.

2.4. ANÁLISIS FACTORIAL INTERBATERIAS

De acuerdo con Tucker (1958), se investigan las componentes $t_h = Xa_h$ e $u_h = Yb_h$ explicando su propio grupo y estando siempre tan correlacionadas como sea posible. Se impone a a_h y b_h ser *ortonormales*.

Se busca entonces *maximizar la covarianza* o simultáneamente el producto de sus varianzas con su correlación; esto es:

$$\max\ cov\left(Xa_{h,}\, Yb_h\right) = \max\{cor\left(Xa_{h,}\, Yb_h\right)\sqrt{var(Xa_h)}\sqrt{var(Yb_h)}\}\ .$$

Este método es en sí, es un compromiso entre el AC de *X e Y* que $[\max cor(Xa_h\, Yb_h)]$ y del ACP de *X* que $[\max var(Xa_h)]$ e *Y* que $[\max var(Yb_h)]$.

Tal como en AC, las variables se suponen centradas y reducidas, con lo cual la matriz de covarianzas[3] o correlaciones intra-*X* es $R_{11} = \frac{1}{n}X'X$ y la intergrupos corresponde a $R_{12} = \frac{1}{n}X'Y$. Observe que si *A* contiene todos los a_h entonces:

$$\sum_h^p var(t_h) = \tfrac{1}{n}\|XA\|^2 = \tfrac{1}{n}traza(XAA'X') = \tfrac{1}{n}traza(X'X) = p$$

igualmente, $\sum_h^q var(u_h) = q$.

2.4.1. Búsqueda de la solución óptima

De la covarianza

$$\begin{aligned}\gamma_h &= cov\left(Xa_{h,}\, Yb_h\right) = cov(t_h, u_h) = \tfrac{1}{n}t_h'u_h \\ &= a_h'R_{12}b_h = \cos(a_{h,}R_{12}b_h)\|R_{12}b_h\| \qquad [2.16]\end{aligned}$$

se deduce que un óptimo se da cuando el coseno es 1, esto es, cuando el vector a_h sea colineal con $R_{12}b_h$ y entonces $\gamma_h = \|R_{12}b_h\|$. De manera análoga, se tiene un óptimo cuando el vector b_h es colineal con $a_h'R_{12}$, con lo cual $\gamma_h = \|R_{21}a_h\|$.

[3] Bajo R, la cov() se calcula sobre n-1. Observe que las restricciones ahora actúan sobre a_h e b_h y no sobre *t* e *u*.

Aplicando el lagrangiano a $cov(Xa_{h,}Yb_h)$ $bajo$ $a_h'a_h = 1$ y $b_h'b_h = 1$ se obtiene el sistema:

$$\tfrac{1}{n}X'Yb = 2\lambda a \quad y \quad \tfrac{1}{n}Y'Xa = 2\mu b \ . \qquad [2.17]$$

Este sistema es relativamente diferente al encontrado en AC. Premultiplicando las dos ecuaciones por *a´* y *b´* se tiene que $2\lambda = 2\mu = \gamma$ y por tanto,

$$a = \tfrac{1}{\gamma}X'Yb \ \ ; \ \ b = \tfrac{1}{\gamma}Y'Xa$$

y las ecuaciones de estacionariedad son

$$X'YY'Xa = \gamma^2 a \ \ ; \ \ Y'XX'Yb = \gamma^2 b \ .$$

Es decir, a_h es *vector propio* de la matriz simétrica $R_{12}R_{21}$ de orden *p*, asociado al *valor propio* γ_h^2 más grande garantizando máxima covarianza; asi, los a_h conforman una base ortonormal en R^p. De forma análoga, b_h es *vector propio* de $R_{21}R_{12}$ asociado al mismo *valor propio* γ_h^2 más grande; y conforman una base ortonormal en R^q.

Se obtiene la misma solución investigando globalmente las componentes t_h *e* u_h maximizando el criterio

$$\sum_{h=1}^{s} cov^2(Xa_h, Yb_h)$$

bajo las restricciones $a_1, \dots, a_s$ ortonormales y $b_1, \dots, b_s$ ortonormales. Cuando R_{12} es de rango completo $s = \min(p, q)$.

2.4.2. Propiedades de las componentes t_h y u_h

- Las componentes t_h *o* u_l del mismo grupo no son ortogonales.

- Las componentes t_h *e* u_l de orden diferente son ortogonales, ya que

$$cor(t_{h,}u_l) = \tfrac{1}{n}t_h'u_l = \tfrac{1}{n}a_h'X'Yb_l = a_h'R_{12}b_l = a_h'a_l\gamma_l = 0 \ .$$

- La interpretación de las componentes a partir de [2.17] es:

$$a_h = \frac{1}{\gamma_h} X' u_h = \frac{1}{r_h \sigma_{t_h}} \frac{X' u_h}{\sigma_{u_h}} = \frac{1}{r_h \sigma_{t_h}} \{cor(x_j, u_h) \dots \forall j\}\,; \quad b_h = \frac{1}{\gamma_h} Y' t_h \quad [2.18]$$

con lo cual a_h es colineal al vector de correlaciones entre las X_j e u_h ; análogamente, b_h es colineal al vector de correlaciones entre las Y_k e t_h .

Hay coherencia entre los coeficientes de las variables y las correlaciones entre las variables de un grupo y las componentes del otro grupo.

Sin embargo las componentes del mismo grupo no son ortogonales pues

$$X'YY'Xa_h = \gamma_h^2 a_h \;\; ; \;\; \text{y} \;\; a_l' X'YY'Xa_h = t_l' YY' t_h = \gamma_h^2 a_l' a_h = 0$$

con lo cual $t_h' t_l \neq 0$ y $u_h' u_l \neq 0$; y esto ha de tenerse en cuenta en el cálculo de las varianzas explicadas o redundancias.

2.4.3. Descomposición de la matriz de correlaciones R_{12}

La descomposición en valores singulares de R_{12}está dada por: $R_{12} = \sum_h^s \gamma_h a_h b_h'$ la cual mediante [2.17] conduce a:

$$cor(x_j\,, y_k) = \sum_h^s \frac{1}{r_h} cor(x_j, u_h) cor(y_k\,, t_h) \;\; .$$

Se puede visualizar la matriz de correlaciones intergrupos, utilizando las correlaciones entre las variables de un grupo y las componentes del otro grupo.

Así, se obtiene la mejor aproximación en el sentido de los mínimos cuadrados de R_{12} , mediante su similar de dimensión *p, q* y rango *m*:

$$R_{12_m} = \sum_h^m \gamma_h a_h b_h' \quad ; \text{ y ya que } \; \|R_{12}\|^2 = \left\|R_{12_m}\right\|^2 + \left\|R_{12} - R_{12_m}\right\|^2 \;\; .$$

Se puede medir la calidad de la aproximación, definiendo el número de componentes a retener. Además, estas normas se calculan en términos de los *valores propios*:

$$\|R_{12}\|^2 \;=\; \text{traza}(R_{12}R_{21}) \;=\; \sum_h^s \gamma_h^2 \;;\; \left\|R_{12_m}\right\|^2 = \sum_{h=1}^s \gamma_h^2 \;;\; \left\|R_{12} - R_{12_m}\right\|^2$$

$$= \sum_{h=m+1}^s \gamma_h^2$$

2.4.4. Estadístico χ^2 para el número de componentes

Tucker (1958) propuso el test $\emptyset_m$ para determinar el número mínimo *m* de componentes a tomar para el cómputo:

$$\emptyset_m = \frac{n(p-m)(q-m)\sum_{m+1}^{s}\gamma_h^2}{(p-\sum_h^m var(t_h))(\sum_h^m var(u_h))} \ .$$

Para $m = 0$, $\emptyset_0 = n\sum_1^s \gamma_h^2$. $\emptyset_m$ se aproxima a una χ^2 con $(p-m)(q-m)\ g.l$ sobre las siguientes hipótesis:

a. Las *m* componentes está bien determinadas (con val-p grandes).
b. La matriz R_{12} es de rango *m*.
c. Los datos son de una población multinormal.
d. La muestra utilizada es aleatoria y suficientemente grande.

Así, la estadística $\emptyset_m$ permite determinar si el residuo $\left\|R_{12} - R_{12_m}\right\|^2 = \sum_{h=m+1}^{s}\gamma_h^2$ es significativo.

Una vez se obtienen los vectores *a* y *b* y por consiguiente las componentes *t* y *u,* se calculará la varianzas de estas componentes, las estadísticas $\emptyset_m$, las correlaciones entre las componentes, y tal como en AC las correlaciones de las x_j e v_k con las u_h y t_h .

También se obtendrán las partes de varianza explicada y las comunalidades, pero en este caso y debido a la correlación entre las componentes el cálculo ha de realizarse con la ayuda de la regresión.

Para la comunalidad intra-grupo se calcula la varianza de X_j explicada en t_1 $(t_1\ ,\ t_2\ )\ ;\ \ldots\ ;\ (t_1, t_2, \ldots, t_s)$. Tal como en ACP se tiene la reconstitución[4] $X = \sum_h^p t_h\, a_h'$ con lo cual la variable $X_j = t_1 a_{1j} + \cdots + t_p a_{pj}$.

Por tanto, en la regresión con *s* componentes se obtiene $X_j = \hat{X}_j + e = \hat{\beta}_1 t_1 + \cdots + \hat{\beta}_s t_s + e$; la cual corresponde con una regresión simple cuando $s = 1$ caso en el cual $\hat{\beta}_1 = X_j'.t_1$, y con una estimación exacta a la del ACP para $s = p$, pues los coeficientes $a_{1j} = \hat{\beta}_1, \ldots, a_{pj} = \hat{\beta}_p$ coinciden.

[4] $Xa_\alpha a_\alpha' = t_\alpha a_\alpha' \Rightarrow X\sum_\alpha a_\alpha a_\alpha' = X = \sum_\alpha t_\alpha a_\alpha'$

En cualquier regresión, se obtiene el coeficiente de determinación $R^2 = \frac{v(\hat{X}_j)}{v(X_j)}$ que mide el porcentaje de variabilidad de X_j explicado por la regresión $\hat{X}_j$; pero como $v(X_j) = 1$ entonces

$$v(\hat{X}_j) = \hat{\beta}_1^2\, v(t_1) + \cdots + \hat{\beta}_s^2 v(t_s) + 2\sum_{i,k>i}^{s} \beta_i\, \beta_k cov(t_i, t_k) = R^2$$

representa la comunalidad intragrupo de X_j en las *s* componentes. Así, es necesario realizar tantas regresiones progresivas, como componentes se tenga, esto es con t_1 ; (t_1, t_2); … ; $(t_1, t_2, \dots, t_s)$. Para $s = p$, $R^2 = 1$.

2.5. ANÁLISIS DE REDUNDANCIAS

Van den Wollemberg (1977), investiga para los grupos *X* e *Y* las componentes centradas reducidas y no correlacionadas $t_h = Xa_h$ maximizando el criterio

$$\sum_h^s \sum_k^q cor^2(y_k, t_h) = \sum_h \tfrac{1}{n} t_h' YY' t_h \tfrac{1}{n} = \sum_h a_h' R_{12}\, R_{21} a_h \quad . \qquad [2.19]$$

Se obtendrá así las componentes t_h que mejor explican el conjunto de variables y_k , bajo la restricción $\frac{1}{n} t_h' t_h = a_h' R_{11} a_h = 1$. Se evita de esta manera el defecto del AC que puede conducir a componentes no explicando alguna parte de las *Y*.

2.5.1. Solución óptima

Si se hace el cambio de variable $c_h = R_{11}^{1/2} a_h$, el criterio [2.19] a maximizar queda:

$$\sum_h c_h' (R_{11}^{-1/2} R_{12} R_{21} R_{11}^{-1/2}) c_h \quad . \qquad [2.20]$$

El número máximo de componentes que es posible extraer es igual al rango *s* de R_{12} . El máximo de [2.20] bajo esta restricción, es alcanzado por los vectores propios normados $c_1, \dots, c_s$ de la matriz simétrica $R_{11}^{-1/2} R_{12} R_{21} R_{11}^{-1/2}$ asociados a los valores propios más grandes $\lambda_1, \dots, \lambda_s$.

Se deducen luego los vectores $a_h = R_{11}^{-1/2} c_h$ y las componentes $t_h = X R_{11}^{-1/2} c_h$. Los vectores a_h son vectores propios de la matriz $R_{11}^{-1} R_{12} R_{21}$ asociados a los mismos valores propios λ_h .

2.5.2. Relaciones importantes

Se utilizan las relaciones habituales entre los vectores propios de la matriz $R_{11}^{-1/2} R_{12} R_{21} R_{11}^{-1/2} (B'B)$ y los de $R_{21} R_{11}^{-1} R_{12} (BB')$. Se puede resaltar que el vector

$$d_h = R_{21} R_{11}^{-1/2} c_h = \tfrac{1}{n} Y' X R_{11}^{-1/2} c_h = \tfrac{1}{n} Y' X a_h = \tfrac{1}{n} Y' t_h$$

es el vector propio de la matriz $R_{21} R_{11}^{-1} R_{12}$ asociado al valor propio λ_h , y *representa el vector de correlaciones de cada* y_k *con las componentes* t_h y tiene por norma $\sqrt{\lambda_h}$.

Además, la matriz $R_{21} R_{11}^{-1} R_{12}$ también se escribe $\frac{1}{n} \hat{Y}' \hat{Y}$ donde $\hat{Y} = X(X'X)^{-1} X'Y$ es la matriz formada de las columnas $\hat{y}_k$, como proyecciones de las y_k sobre el espacio engendrado por las columnas de *X*. Así, la matriz $R_{21} R_{11}^{-1} R_{12}$ representa la matriz de covarianzas entre las $\hat{y}_k$. El ACP de las proyecciones no reducidas conduce entonces a los vectores propios d_h de la matriz de covarianzas $\frac{1}{n} \hat{Y}' \hat{Y}$.

Los vectores propios c_h en f unción de los d_h son: $c_h = \frac{1}{\lambda_h} R_{11}^{-1/2} R_{12} d_h$.

Se deduce que t_h es la *h*-ésima componente principal reducida de las proyecciones $\hat{y}_k$:

$$t_h = X R_{11}^{-1/2} c_h = \tfrac{1}{\sqrt{\lambda_h}} P_x Y d_h \tfrac{1}{\sqrt{\lambda_h}} = \tfrac{1}{\sqrt{\lambda_h}} \hat{Y} d_h \ .$$

Se puede notar que la componente t_h se obtiene por regresión de $Y d_h$ sobre *X*, luego de la normalización. Las componentes $Y d_h$ pueden ser útiles, pero tienen el inconveniente de estar correlacionadas entre ellas. De

$$\tfrac{1}{n} \hat{Y}' \hat{Y} d_h = \lambda_h d_h \quad \text{y} \quad var(\hat{y}_k) = R^2(y_k; X_1, \dots, X_p)$$

se deduce:

$$cor(y_k, t_h) = cov(\hat{y}_k, t_h) = R(y_k; X_1, \dots, X_p).cor(\hat{y}_k, t_h) \ .$$

Esto muestra la posibilidad de visualizar las variables *X* e *Y* con la ayuda de sus correlaciones con las t_h. Así, como lo anotó Tyler (1982), el AR de *Y* respecto a *X* es un ACP de la proyección de *Y* sobre *X*.

2.6. REGRESIÓN MULTIVARIANTE

Considere el modelo definido por $Y = XB + U$ donde $Y_{n,p}$ es una matriz observada de *p* variables de respuesta sobre cada uno de *n* individuos, $X_{n,q}$ (independiente) es una matriz conocida, $B_{q,p}$ es la matriz con los parámetros de regresión desconocidos y *U* es una matriz no observada de perturbaciones aleatorias cuyas filas para *X* dado son incorrelacionadas, cada una con media *0* y matriz de covarianza común Σ. (Mardia et al. 1979)

Las columnas de *Y* representan las variables dependientes, las cuales serán explicadas por las columnas de *X*. Note que $E(y_{ij}) = X_i' B_j$. En muchas aplicaciones se supone que $U \sim N_p(0, \Sigma)$ donde *U* es independiente de *X*. Bajo el supuesto de errores normales, *la log-likelihood* para los datos *Y* en términos de los parámetros *B* y Σ es

$$\ell(B, \Sigma) = -1/2\, nlog|2\pi\Sigma| - 1/2\, tr(Y - XB)\Sigma^{-1}(Y - XB)' \ .$$

Se supone que el rango de *X* es q.

Sea $P_{nm} = I - X(X'X)^{-1}X'$ la matriz proyección sobre el subespacio de R^n ortogonal a V_x , simétrica e idempotente de rango *n-q*, tal que *PW = W sii W* es ortogonal a todas las columnas de *X* y *PW = 0 sii* es combinación lineal de las columnas de $X \Rightarrow PX = 0$.

Así, el estimador máximo verosímil de *B* y Σ es $\hat{B} = (X'X)^{-1}X'Y$ y $\hat{\Sigma} = n^{-1}Y'PY$ y además, $\hat{Y} = X\hat{B} = X(X'X)^{-1}X'Y$, $\hat{U} = Y - X\hat{B} = PY$.

2.7. ANÁLISIS CANÓNICO GENERALIZADO

La ampliación del AC con más de dos grupos de variables se topa con la dificultad siguiente: no existe una medida simple de la relación entre más de dos variables.

Así que habrá muchas maneras de obtener las variables canónicas como la forma de definir una correlación entre las *p* variables: se puede tomar por ejemplo como medida la suma de las correlaciones dos a dos, la suma de los cuadrados de las correlaciones, el determinante de la matriz de correlaciones, etc.

Todas estas generalizaciones son más o menos arbitrarias. El que se presenta aquí tiene la ventaja de ser probablemente la interpretación más simple y más rica, ya que se conecta fácilmente a todos los demás métodos de análisis de datos. (Ver Carroll 1968, Saporta 2011)

2.7.1. Una propiedad del análisis canónico ordinario

Sean los datos de dos conjuntos de variables centradas X_1 e X_2 ; los subespacios de R^n engendrados por las columnas de X_1 e X_2 se denotan: $W_1 = \{x|x = X_1 a\}$ e $W_2 = \{y|y = X_{2i} b\}$.

Los operadores A_1 e A_2 de proyección ortogonal sobre W_1 e W_2 son respectivamente:

$$A_1 = X_1\left(X_1' X_1\right)^{-1} X_1'$$

$$A_2 = X_2\left(X_2' X_2\right)^{-1} X_2' \ .$$

Las variables canónicas $\xi \epsilon\, W_1$ e $\eta \,\epsilon\, W_2$, vectores propios de $A_1 A_2$ e $A_2 A_1$ respectivamente, poseen la siguiente propiedad:

$$\xi + \eta \text{ es vector propio de } A_1 + A_2 \ .$$

En efecto, definiendo Z tal que $(A_1 + A_2)Z = \mu\, Z$ y premultiplicando por $A_1\, o\, A_2$ esta ecuación, se tiene:

$$A_1(A_1 + A_2)Z = \mu\, A_1 Z$$

o bien

$$A_1 A_2 Z = (\mu - 1)A_1 Z \quad \text{e} \quad A_2 A_1 Z = (\mu - 1)A_2 Z$$

lo que da:

$$A_1 A_2 A_1 Z = (\mu - 1)^2 A_1 Z$$

$$A_2 A_1 A_2 Z = (\mu - 1)^2 A_2 Z \ .$$

$A_1 Z$ e $A_2 Z$ no son otras que las variables canónicas ξ y η ; como $A_1 Z + A_2 Z = \mu Z$ se tiene $\mu Z = \xi + \eta$, lo que demuestra la propiedad enunciada.

La variable Z tiene la propiedad de ser la más correlacionada con los dos conjuntos X_1 e X_2 , en el sentido que ella tiene suma máxima de los cuadrados de los coeficientes de correlación múltiple con X_1 e X_2 .

El coeficiente de correlación múltiple de *Z* con X_i vale:

$$R_i^2 = \frac{Z' A_i Z}{Z' Z} = \frac{\|A_i Z\|^2}{\|Z\|^2} \ .$$

Dado que las variables están centradas, R_i es el coseno del ángulo formado por Z e W_i .

2.7.2. Generalización

De la propiedad anterior J. D. Carroll deduce la generalización del AC: en lugar de buscar directamente las variables canónicas de cada uno de los subespacios W_i asociados a las tablas de datos X_i , se busca una variable auxiliar *Z* que pertenece a la suma de los W_i tal que $\sum_i^p R^2(Z; X_i)$ sea máxima.

Z es entonces vector propio en: $(A_1 + A_2 + \cdots + A_p)Z = \mu Z$. Se obtiene a continuación, si es necesario, las variables canónicas ξ_i sobre los W_i ; $\xi_i = A_i Z$.

Si se pone la matriz $X = (X_1|X_2|\ldots|X_p)$ con *n* filas y $\sum_i^p m_i$ columnas, la variable *Z* es de la forma *Xb* y en vez de buscar *Z* como vector propio de una matriz $n_x n$ es mejor encontrar *b* que posee $\sum_i^p m_i$ componentes.

Siendo $A_i = X_i (X_i' X_i)^{-1} X_i'$, $V_{ii} = X_i' X_i$ la matriz de covarianzas del grupo *i* y

$$M = \begin{bmatrix} V_{11}^{-1} & & \\ & : & \\ & & V_{pp}^{-1} \end{bmatrix}$$

se encuentra fácilmente que $\sum_i^p A_i = \sum_i^p X_i V_{ii}^{-1} X_i'$ se escribe $\sum_i^p A_i = XMX'$.

Por lo tanto Z es vector propio de XMX' , y puesto que $Z = Xb$, si X es de rango pleno, b es vector propio de $MX'X$:

$$XMX'z = \mu z \quad , \quad MX'Xb = \mu b \ .$$

CAPÍTULO 3

MÉTODOS PLS

Son algoritmos basados en los conceptos del Análisis de Componentes Principales (ACP) y Regresión, que iteran hasta la convergencia conllevando el cálculo de componentes PLS y la estimación de parámetros del modelo que relaciona *J* bloques de datos ($X_1, \dots, X_J$) a través de sus LV. Los bloques X_J pueden presentar problemas de datos faltantes, multicolinealidad y además contener más variables que individuos ($p > n$).

Dentro de los métodos PLS (Wold 1985, Tenenhaus 1998) más importantes se encuentran NIPALS, PLS-R y PLS-PM para el tratamiento y análisis de una, dos y más matrices de datos respectivamente.

3.1. ALGORITMO NIPALS

El algoritmo NIPALS (Nonlinear estimation by Iterative Partial Least Square) es la base de la regresión PLS (Wold 1966). Fundamentalmente realiza la descomposición singular de una matriz de datos, mediante secuencias iterativas convergentes de proyecciones ortogonales [concepto geométrico de regresión simple]. Con bases de datos completas se tiene equivalencia con los resultados del ACP; sin embargo, y esta quizá es su mayor virtud, se puede realizar el ACP con datos faltantes (missing data) y obtener sus estimaciones a partir de la matriz de datos reconstituida.

Sea $X_{n,p}$ la matriz de datos de rango $a \leq p$ cuyas columnas $X_1, \dots, X_p$ se suponen centradas o estandarizadas (bajo S_n). La reconstitución derivada del ACP conlleva a $X = \sum_h^a t_h P_h'$ donde t es la *componente principal* (*scores*) y P_h' el *vector propio* (*loadings*) en el eje h.

$$[X_1 \dots X_p] = t_1 P_1' + \cdots + t_a P_a' \quad . \qquad [3.1]$$

Así, la columna $X_j = \sum_h^a P_{hj} t_h \quad j = 1, \dots p$ y la i-ésima fila $x_i = \sum_h^a t_{hi} P_h \quad i = 1, n$.

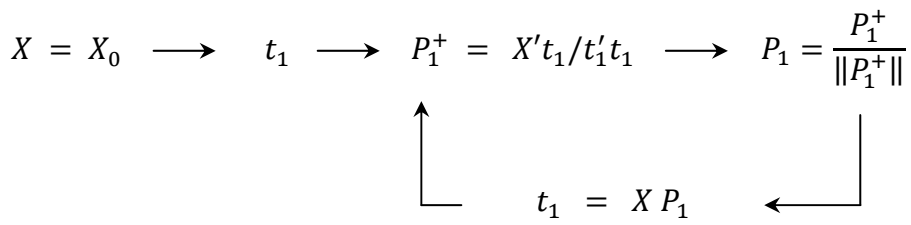
Observe entonces que si $h = 1$, la columna *j* se expresa como $X_j = P_{1j}t_1$ es decir $P_{hj} = X_j't_h$ es como el coeficiente (pendiente[5]) en la regresión de X_j sobre t_h. En el espacio de las filas, t_{hi} es el coeficiente de la regresión sin constante del individuo x_i sobre P_h.

Para $h > 1$, P_{hj} es el coeficiente de regresión de t_h en la regresión simple del vector deflactado $X_j - \sum_l^{h-1} P_{lj}\, t_l$ sobre t_h y t_{hi} el de P_h en la regresión de $x_i - \sum_l^{h-1} t_{li}\, P_l$ sobre P_h.

El objetivo de cualquier algoritmo PLS es el procedimiento iterativo para calcular los *parámetros* P_h del modelo. Para cada componente las cargas son computadas, una como función de la otra, a través del procedimiento iterativo.

3.1.1. Descripción pseudocódigo NIPALS

El flujograma asociado al procedimiento de convergencia en la etapa 2.2, es:



Luego se construirán una serie de tablas deflactadas notadas X_h cuyas columnas son $X_{h1}, \ldots, X_{hp}$; la *i*-ésima fila se notará $x'_{hi} = (x_{h1i}, \ldots, x_{hpi})$. El algoritmo inicia tomando X_{01} como la 1ª *componente principal* t_1.

3.1.1.1. Sin datos faltantes

Etapa 1. $X_0 = X_h$

Etapa 2. $h = 1, 2, \ldots, a$:

Etapa 2.1. t_h = 1ª columna de X_{h-1} [o $\bar{X}$]

[5] De la regresión simple $\hat{\beta}_1 = \hat{p}_{hj} = \frac{cov(t_h, X_j)}{s_{t_h}^2} = \frac{X_j't_h}{\|t\|^2} = X_j't_h = r$ si *x* e *t* estandarizadas.

Etapa 2.2. Repetir hasta la convergencia de P_h

Etapa 2.2.1 $P_h = \frac{X'_{h-1}t_h}{t'_h t_h}$ $\left[u = \frac{X'Xu}{\lambda}\ ,\ \lambda = \frac{1}{n}t't\right]$

Etapa 2.2.2 normar P_h a 1

Etapa 2.2.3 $t_h = X_{h-1}P_h$ [$t = Xu$]

Etapa 2.3. $X_h = X_{h-1} - t_h P'_h$ [garantiza la ortogonalidad]

Siguiente *h*.

Fin

Tal como se estudió inicialmente, en la etapa 2.2.1 P_{hj} representa, antes de la normalización, el coeficiente [pendiente] de la regresión de $X_{h-1,j}$ sobre la componente t_h.

Análogamente, en la etapa 2.2.3, t_{hi} representa el coeficiente de regresión (sin constante) de $x_{h-1,i}$ sobre P_h ; y ya que $P'_h P_h = 1$, t_{hi} también es el largo de la proyección ortogonal de $X_{h-1,i}$ sobre P_h .

Para $h = 1$ se obtiene el primer eje factorial P_1 y la primera *componente principal* t_1 de $X'X$. Ya que la matriz $X_1 = X - t_1 P'_1$ representa el *residuo* de la regresión de *X* sobre la primera *componente principal*, de [3.1], el vector propio P_2 de la matriz $X'_1 X_1 \,/\, n$ asociado al *valor propio* más grande, corresponde al *vector propio* de $X'X/n$ asociado al segundo *valor propio* más grande λ_2 .

Las relaciones cíclicas de la etapa 2.2 muestran que en el límite se verifican las ecuaciones:

$$\tfrac{1}{n}X'_{h-1}X_h P_h = \lambda_h P_h \ ; \quad \tfrac{1}{n}X_h X'_{h-1}t_h = \lambda_h t_h \ \Rightarrow \ \lambda_h = \tfrac{1}{n}t'_h t_h$$

Una vez se consigue la convergencia, en la etapa 2.3 se deflacta la matriz precedente para garantizar la ortogonalidad de las siguientes componentes. Con datos completos el divisor $t'_h t_h$ en la etapa 2.2.1 no es necesario.

La convergencia en NIPALS se puede conseguir con P_h tal como se ha expuesto, pero también con los t_h que representarán los vector propio en R^n y los resultados serán equivalentes.

Así, el problema del ACP bajo NIPALS es resolver una serie de regresiones simples locales hasta alcanzar la convergencia de los coeficientes de regresión

P_{hj} y t_{hi} que es el nuevo valor proporcionado de la regresión sin constante de $x_{h-1,i}$ sobre la 'nueva' variable P_h después de su normalización.

La principal característica del NIPALS es que trabaja respecto a una serie de productos escalares como suma de productos de los elementos emparejados. Esto permite manejar *missing data*, agregando en cada operación los pares disponibles. Geométricamente el procedimiento 'toma' los elementos omitidos como si ellos cayeran sobre la recta de regresión; no son puntos de apalancamiento.

Así, con datos faltantes se obtiene sin embargo las componentes t_h y los vectores P_h que permiten luego 'reconstituir' la matriz de datos mediante $\hat{X}$ y de ésta, estimar los datos faltantes utilizando la fórmula de reconstitución (1) derivada del ACP : $\hat{x}_{ji} = \sum_l^h t_{li} p_{lj}$.

3.1.1.2. Pseudocódigo NIPALS con datos faltantes

Etapa 1. $X_0 = X_h$

Etapa 2. $h = 1, 2, \ldots, a$:

Etapa 2.1. t_h = 1ª columna de X_{h-1}

Etapa 2.2. Repetir hasta la convergencia de P_h

Etapa 2.2.1 para *j=1,2,...,p*:

$$P_{hj} = \frac{\sum_{\{i: x_{ji} \; e \; t_{hi} \; existen\}} x_{h-1,ji} t_{hi}}{\{i: x_{ji} \; e \; t_{hi} \; existen\} t_{hi}^2} \qquad [cov(t_h, x_{h-1,j})/s^2_{th}]$$

Etapa 2.2.2 normar P_h a 1.

Etapa 2.2.3 para *i=1,2,...,n*: $t_{hi} = \frac{\sum_{\{j: x_{ji} \; existe\}} x_{h-1,ji} P_{hj}}{\sum_{\{j: x_{ji} \; existe\}} P_{hj}^2}$

Etapa 2.3. $X_h = X_{h-1} - t_h P_h'$

Fin

En las etapas 2.2.1 y 2.2.3 se calculan las pendientes de las rectas de mínimos cuadrados pasando por el origen de la nube de puntos sobre los *datos disponibles*. Los P_{hj} y los t_{hi} deben conservar en sus posiciones *j* e *i*, la característica de dato faltante dada por x_{ij} , la cual se puede expresar con 0.

En el anexo A se presentan las funciones bajo R que desarrollan los algoritmos NIPALS para matrices con y sin datos faltantes.

3.2. REGRESIÓN PLS

La regresión PLS fue creada por Wold (1975), como una técnica de regresión lineal que permite relacionar el conjunto de p variables predictoras X con una o varias variables de respuesta Y, implementado conceptos de ACP y de regresión simple (proyecciones ortogonales). Descompone la matriz predictor X con ayuda de Y, extrayendo secuencialmente H ($< p$) componentes PLS ortogonales, las cuales a su vez resumen las variables exógenas y nos permite modelar y predecir las variables de respuesta. Posteriormente es denominada Partial Least Squares Regression (PLS-R) por Wold et al. 1983, Tenenhaus 1998.

Es un método de análisis factorial que resume información redundante de una matriz predictor en pocas componentes ortogonales sin involucrar inversión de matrices. Esto hace del PLS-R una potente herramienta de visualización de información no redundante, por medio de proyecciones sobre el espacio generado por las componentes PLS.

PLS-R ha sido aplicada con gran eficiencia en campos tales como quimiometría, genética, sensometría etc, donde los datos son caracterizados por un lote de variables medidas sobre pocas observaciones ($p > n$). Este tipo de datos genera tres inconvenientes: inferencial, computacional y descriptivo.

El problema inferencial es debido al hecho que grandes conjuntos de variables siempre están fuertemente correlacionadas e inducen multicolinealidad, conllevando al aumento de variabilidad de los estimadores de los coeficientes de regresión, en detrimento a su significancia. El computacional es debido a que el rango de la matriz predictor, conlleva a una matriz de correlación singular; y el problema descriptivo se refiere a la dificultad analizando relaciones entre muchas variables al mismo tiempo.

Desde el punto de vista algorítmico PLS-R puede ser visto como una extensión del NIPALS para el análisis de una matriz de *covarianza cruzada*; además, puede ser considerado como una versión ligeramente modificada del algoritmo PLS-PM para dos bloques.

- **El Modelo**

Sea X un conjunto de p variables predictor y Y un conjunto de r variables de respuesta medidas sobre n observaciones. Se supone que todas las variables están centradas o estandarizadas.

El modelo PLS-R asume que hay una estructura común subyacente en los dos bloques de variables, y que esta estructura puede ser resumida por pocas componentes latentes $t_h \in T$ $(h = 1,2, \dots, H)$ calculadas como una combinación lineal de las variables predictor lo más relacionadas con *Y*.

Las matrices predictor y respuesta son descompuestas de la forma:

$$X = T_H P_H' + X_H$$
$$Y = T_H C_H' + Y_H$$

donde P_H y C_H son las matrices de cargas conteniendo los parámetros del modelo, y X_H e Y_H las matrices de residuos representando la parte de variabilidad de los datos debido al ruido. Los parámetros del modelo son calculados por medio de los algoritmos de regresión PLS denominados PLS2 en el caso de respuesta múltiple y PLS1 en el caso de respuesta simple. Ya que PLS1 es un caso particular de PLS2, esta distinción es puramente formal.

3.2.1. Regresión PLS1

Wold et al. (1983) modificaron ligeramente el NIPALS para obtener componentes regularizadas (estimadores de contracción) basadas en regresión, lo que se conoce como PLS Regressión (PLS-R). El algoritmo para PLS1 que contiene sólo una variable respuesta (caso particular del PLS2) es presentado por Tenenhaus (1998) y también es desarrollado con datos faltantes.

Se quiere estimar via PLS la regresión multiple

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \qquad [3.2]$$

de una sola variable a explicar *Y* sobre todas las variables explicativas $X_1, \dots, X_p$ que pueden estar altamente correlacionadas [multicolinealidad] y ser mayor en número que las observaciones, esto es $p > n$. Tanto las X_j como *Y* que $\in R^n$ se suponen *centradas – reducidas.*

Se busca en el espacio de las predictoras componentes ortogonales $t = Xw$ correlacionadas con *Y*, tal que se realice la regresión $Y = c_1 t_1 + \cdots + c_H t_H + Y_H$ para luego mediante desdoblamiento de las $t = f(X)$ que son combinación lineal de las X_j estimar el modelo expresado en [3.2]. Generalmente se tiene que $H < p$.

La idea es encontrar *w* tal que se maximice el cuadrado de la covarianza entre el componente $t = Xw$, y la variable respuesta Y bajo $w'w = 1$; (Vega y Guzmán 2011).

Sea V la matriz de orden $p\ x\ 1$, el vector de covarianzas de X e Y $(V = X'Y)$, entonces,

$$cov^2(Xw, Y) = [w'cov(X, Y)]^2 = [w'V]^2 = w'VV'w \ . \qquad [3.3]$$

La función lagrangiana $\emptyset$ que maximiza este cuadrado, $cov^2(t, Y)$, sujeta a la restricción de ortonormalidad de *w*, es:

$$\emptyset = w'VV'w - \lambda(w'w - 1)$$

y su derivada respecto a *w*, igualada a cero, $\frac{\partial \emptyset}{\partial w} = 2VV'w - 2\lambda w = 0$, por tanto

$$VV'w = \lambda w \qquad [3.4]$$

con lo cual $\lambda\ y\ w$ son el autovalor y autovector de la matriz VV' respectivamente. Observe que premultiplicando [3.4] por w', e igualmente, premultiplicando [3.4] por V' se deduce λ. Comparando las dos expresiones de λ, se tiene que el *w* que maximiza [3.3] es $w = \frac{V}{\|V\|} = \frac{X'Y}{\|X'Y\|}$ y corresponde al vector de covarianzas normalizado.

Observe que para cada etapa $h = 1, \dots, H$ se maximiza $cov^2(t_h, Y_{h-1})$ según [3.3]. Existen numerosas versiones de este algoritmo también basado en los principios del ACP que generalmente difieren en los niveles de normalización, pero todos convergen a la misma regresión. Se presenta aquí la versión que deriva naturalmente del NIPALS. En adelante por simplicidad se usa *y* en vez de *Y*.

3.2.1.1. El algoritmo de regresión PLS1

1. $X_0 = X,\ y_0 = y \equiv$ *componente de* R^n
2. *para* $h = 1, 2, \dots, a$. *(X es de rango* $a \leq p$ *)*
   - 2.1 $w_{hi} = X'_{h-1}y_{h-1}/\|y_{h-1}\|$ [coef reg x_j sobre y_{h-1}]
   - 2.2 $w_h = w_{hi}/\|w_{hi}\|$ [normar w_h a 1]
   - 2.3 $t_h = X_{h-1}w_h/w'_h w_h$ [*componente_h de X*]

2.4 $p_h = X'_{h-1}t_h / t'_h t_h$ [*coef reg* x_j *sobre* t_h]

2.5 $X_h = X_{h-1} - t_h p'_h$ [*residuo X* y *ortogonalidad*]

2.6 $c_h = y'_{h-1}t_h / t'_h t_h$ [*coef. de t sobre y*]

2.7 $u_h = y_{h-1}/c_h$ [*componente_h de Y*]

2.8 $y_h = y_{h-1} - t_h c_h$ [*residuos en y*]

Fin h.

Note que no hay iteraciones para la convergencia, debido a que sólo se tiene una *y*. Las coordenadas de los vectores w_h , t_h , p_h , e c_h representan las pendientes de las rectas de mínimos cuadrados pasando por el origen y pueden entonces ser calculadas con datos faltantes.

En este caso, los cálculos en *t* y *w* son de la forma: $t_{1i} = \frac{\sum_{\{j: x_{ji}\ existe\}} w''_{1j} x_{ji}}{\sum_{\{j: x_{ji}\ existe\}} (w''_{1j})^2}$ con

$$w''_{1j} = \frac{w'_{1j}}{\sqrt{\sum_j^p \left(w'_{1j}\right)^2}} \quad ; \quad w'_{1j} = \frac{\sum_{\{i: x_{ji}\ e\ y_i\ existen\}} x_{ji} y_i}{\sum_{\{i: x_{ji}\ e\ y_i\ existen\}} y_i^2} \quad .$$

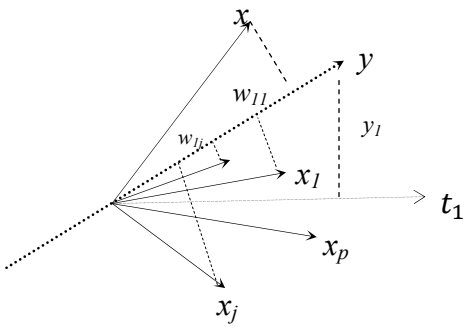
Cuando no hay datos faltantes se puede remplazar las etapas 2.1 y 2.2 por: $w_h = X'_{h-1}y_{h-1} / \| X'_{h-1}y_{h-1} \|$ y 2.3 por $t_h = X_{h-1}w_h$.

La idea es conseguir en el subespacio R_x engendrado por las variables $x_1, \ldots, x_p$ de R^n, una "*componente*" t_1 altamente correlacionada con *y,* es decir, cuya regresión con y minimice el residuo generado y_1 . (Ver gráfica 3.1).

Así, $t_1 = w_{11}x_1 + \cdots + w_{1p}x_p$ es combinación de las variables exógenas ponderadas por las coordenadas del "vector propio" w_1 , el cual se puede obtener buscando una relación con *y* siguiendo los lineamientos del ACP.

Se toma inicialmente *y* normalizado como el "*eigenvector"* de R^n sobre el cual se proyecta ortogonalmente cada una de las variables x_j, para obtener los coeficientes $w^*_{1j} = cov(x_j, y) = x'_j . y$ derivados de esta regresión, los cuales conformarán el vector w^*_1. Para garantizar norma 1, se toma $w_1 = w^*_1 / \| w^*_1 \|$.

Se calcula entonces la primera componente t_1 más relacionada con *y*, de la forma $t_1 = Xw_1 = w_{11}x_1 + \cdots + w_{1p}x_p$.



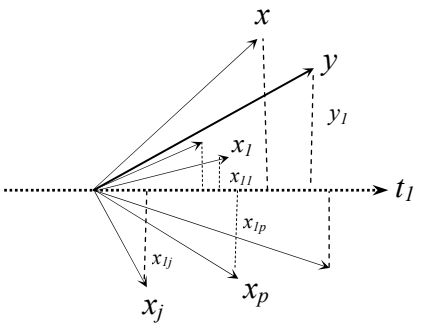
Gráfica 3.1. Obtención de las coordenadas de w_1 por proyección de x_j sobre y

Luego se realiza una *regresión* simple de y sobre t_1:

$$y = c_1 t_1 + y_1 = c_1 w_{11} x_1 + \cdots + c_1 w_{1p} x_p + y_1$$

donde $c_1 = y' t_1 / t_1' . t_1$ es el coeficiente de regresión, y y_1 el vector de *residuos* (parte de y no explicada por t_1); estos coeficientes de regresión son más 'fáciles' de interpretar por el investigador.

Es posible que aún falte explicar buena parte de los residuos. Si el poder explicativo de la regresión $y \sim t_1$ es muy débil, se construye una segunda componente t_2 en el espacio de los residuos de x ortogonales a t_1 y se realiza entonces la regresión $y \sim t_1 + t_2$.



Gráfica 3.2. Los residuos en x e y son ortogonales a t_1

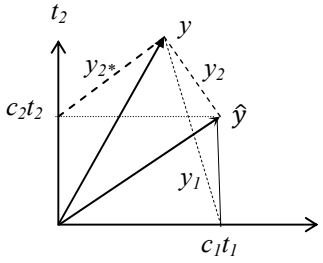
En la gráfica 3.2 los residuos x_{1j} (líneas punteadas) forman un subespacio ortogonal a R_x , y representan la parte de las x^s no explicada por t_1 ; se obtienen

de realizar regresiones simples de las x_j con t_1 , con lo cual[6] $x_{1j} = x_j - p_{1j}t_1$. La nueva componente t_2 es combinación lineal de estos residuos; $t_2 = w_{21}x_{11} + \ldots + w_{2p}x_{1p}$ y explicará el residuo y_1 también ortogonal a t_1 .

Se procede análogamente como antes, situados ahora en el subespacio de los residuos ortogonal a R_x , tomo como "*vector propio"* el residuo y_1 y proyecto ortogonalmente sobre este los residuos x_{1j} para obtener sus ponderadores w_{2j} y generar la combinación lineal denominada componente t_2 ortogonal a t_1 . Los coeficientes $w_{2j} = cov(x_{1j}, y_1)/\|w_2\|$ conforman el vector "propio" w_2 de norma 1.

Se realiza luego la regresión $y = c_1t_1 + c_2t_2 + y_2$, la cual es más precisa que la primera. (Ver gráfica 3.3).

Este procedimiento es iterativo, ahora se consiguen los residuos y_2 , $x_{21}, \ldots, x_{2p}$ de las regresiones[7] de y, $x_1, \ldots, x_p$ sobre el subespacio de dos dimensiones (t_1, t_2). Los residuos x_{2j} se proyectan sobre y_2 y estas coordenadas conformarán el vector w_3 que permitirá además, construir la componente t_3 ortogonal a (t_1, t_2). Se efectúa entonces la regresión $y = c_1t_1 + c_2t_2 + c_3t_3 + y_3$ etc.



Gráfica 3.3. Regresión de y sobre las componentes ortogonales t_1 , t_2

El número de componentes $t_1, \ldots, t_H$ a retener en la regresión es determinado por *validación cruzada.* Para cada valor *h* se calcula las predicciones $\hat{y}_{hi}$ e $\hat{y}_{h(-i)}$ de y_i con la ayuda del modelo con *h* componentes, calculadas utilizando todas las observaciones, y luego sin utilizar la observación *i*.

---

[6] $p_1 = X't_1/\|t_1\|$. El residuo será muy útil para explicar tanto t_h como y en función de las x_j.

[7] Como $x_j = c_{1j}t_1 + c_{2j}t_2 + x_{2j} \Rightarrow \underbrace{x_j - c_{1j}t_1}_{X1} = c_{2j}t_2 + x_{2j}$ es equivalente la regresión del residuo $X1$ sobre t_2.

Se calcula enseguida los criterios RSS_h (Residual Sum of Squares) y $PRESS_h$ (Prediction Error Sum of Squares) definidos por:

$$RSS_h = \sum(y_i - \hat{y}_{hi})^2 \quad , \quad PRESS_h = \sum(y_i - \hat{y}_{h(-i)})^2 \ .$$

Así, la componente t_h es retenida si $\sqrt{PRESS_h} \leq 0.95\sqrt{RSS_{h-1}}$.

En lugar de eliminar las observaciones una por una, se puede también eliminar bloque por bloque.

3.2.1.2. Las componentes PLS1

La regresión $y = c_1 t_1 + y_1$, se puede 'normalizar' haciendo $y/c_1 = u_1$ y correr de nuevo la regresión $u_1 = c_1^* t_1 + y_1^*$ con lo cual se obtendría como recta de mínimos cuadrados la primera bisectriz, pues $\hat{u}_1 = t_1$ ya que $\hat{c}_1^* = 1$; además, $y_1^* = y_1/c_1$.

Análogamente, de la regresión $y = c_1 t_1 + c_2 t_2 + y_2$ se tiene:

$$y - c_1 t_1 = y_1 = c_2 t_2 + y_2$$

con lo cual son equivalentes las regresiones

$$y \sim t_1 \, , t_2 \quad \text{o} \quad y_1 \sim t_2$$

pues obtienen el mismo coeficiente c_2 y los mismos residuales y_2 .

Aquí también se puede realizar una regresión bajo la transformación

$$y_1/c_2 = u_2 = t_2 + y_2^*$$

para obtener la primera bisectriz entre u_2 y t_2 que tiene 1 como coeficiente asociado. Sin embargo las variables $u_h = y_{h-1}/c_h$ no juegan un papel importante en el análisis PLS1, pues no son interpretables.

Las componentes t_h se han construido para simultáneamente y de la mejor manera posible, describir las x_j y explicar la variable *y*.

3.2.1.3. Puntos atípicos: T^2 de Hotelling

La T^2 de la observación *i*, calculada utilizando *H* componentes, está definida por:

$$T_i^2 = \frac{n}{n-1}\sum_h^H t_{hi}^2/s_h^2$$

donde s_h^2 es la varianza (sobre n-1) de la componente t_h .

Se detectan los puntos atípicos utilizando la estadística (Tracy, Young y Mason. 1992):

$$\frac{n(n-H)}{H(n^2-1)}T_i^2 \sim F_{H,n-H} \ .$$

Una observación se considera como atípica si

$$T_i^2 \geq \frac{H(n^2-1)}{n(n-H)}\, F_{0.95,H,n-H} = c^2.$$

Se puede construir en el plano $(t_1\,, t_2)$ la elipse al 95% de la ecuación:

$$\frac{t_{1i}^2}{s_{t_1}^2} + \frac{t_{2i}^2}{s_{t_2}^2} = \frac{n-1}{n}C^2$$

que para *H=2* el umbral $C^2 = 9.77$.

3.2.1.4. Interpretación de las componentes t_h

En regresión PLS clásica, se utiliza la notación siguiente:

$$t_h = \sum_j^p w_{hj}\, x_{hj} = \sum_j^p w_{hj}^*\, x_j \ .$$

Las componentes t_h son calculadas con la ayuda de los residuales x_{hj} , pero serán interpretadas con la ayuda de las variables x_j ; los coeficientes w_{hj}^* traducen la importancia de cada variable x_j en la construcción de t_h .

En ACP la interpretación de las componentes principales es simple porque la correlación entre las variables y la *componente principal s* son las coordenadas de las variables en la componente principal. Este no es el caso en regresión PLS.

Los coeficientes w_{hj}^{*} son el producto de los coeficientes aleatorios obtenidos en etapas anteriores y traducen la importancia de cada variable x_j en la construcción de t_h . Los c_h miden la relación de la variable *y* con las componentes t_h .

Es decir, la relación entre la variable *y* con las variables x_j se resumen a través de las variables t_h , en las ecuaciones siguientes:

$$T = XW^{*} \quad \text{(ver 3.2.1.8)} \quad ; \quad y \approx TC \ .$$

3.2.1.5. La previsión

El modelo de regresión de *y* sobre las *H* componentes PLS es:

$$y = c_1 t_1 + \cdots + c_H t_H + \varepsilon \ .$$

El cual y luego de estimar los c_H se puede expresar en función de las variables originales por:

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k \ .$$

El cual, para valores fijos de $x = (1, x_1, \ldots, x_k)$, $\hat{y} = \hat{y}_x$ se constituye en un estimador insesgado de $E(y|x) = \mu_x$. Por tanto, suponiendo $\varepsilon \sim N(0, \sigma)$, un intervalo de confianza al 95% para μ_x es:

$$\hat{y}_x \ \pm \mathrm{t}_{0.975,n-p}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{t_{1x}^2}{t_1' t_1} + \cdots + \frac{t_{Hx}^2}{t_H' t_H}} \ .$$

Análogamente, un intervalo de predicción para y_x esta dado por la fórmula

$$\hat{y}_x \ \pm \mathrm{t}_{0.975,n-p}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{t_{1x}^2}{t_1' t_1} + \cdots + \frac{t_{Hx}^2}{t_H' t_H}}$$

t_{Hx} es la coordenada de la componente t_H asociada al individuo x y $\hat{\sigma}^2$ el CME.

3.2.1.6. Propiedades matemáticas de la Regresión PLS1

La matriz X_h es el residuo de la regresión de X_{h-1} sobre t_h . La variable y_h es el vector de residuos en la regresión de y_{h-1} sobre t_h .

De estas regresiones se puede deducir las propiedades de *ortogonalidad*:

$$t_h' X_h = 0 \quad ; \quad t_h' y_h = 0 \ .$$

Se describen a continuación un conjunto de propiedades cuando no hay datos faltantes (validas también para PLS2):

a) $t_h' t_l = 0, \ \ l > h$.

b) $t_h' X_l = 0, \ \ l \geq h$.

c) $w_h' X_l' = 0, \ \ l \geq h$.

d) $w_h' w_l = 0, \ \ l > h$.

e) $w_h' p_l = 0, \ \ l > h$.

f) $w_h' p_h = 1$.

g) $X_h = X \prod_j^h (I - w_j p_j'), \ \ h \geq 1.$

3.2.1.7. Simplificación del algoritmo PLS1

La ortogonalidad de las componentes t_h permite simplificar el algoritmo PLS. Se evita el cálculo de los residuos y_h . En efecto ya que,

$$X_h = X - t_1 p_1' - \cdots - t_h p_h' = X_{h-1} - t_h p_h'$$

$$y_h = y - c_1 t_1 - \cdots - c_h t_h = y_{h-1} - c_h t_h \ .$$

Se tiene que

$$X_{h-1}' y_{h-1} = X_{h-1}' (y - c_1 t_1 - \cdots - c_{h-1} t_{h-1}) \underbrace{=}_{prop.\ f} X_{h-1}' y \ .$$

De manera análoga para c_h , p_h con lo cual en la etapa 2, se tienen los siguientes remplazamientos:

$$w_h = X_{h-1}' y / y' y \ ; \ \ p_h = X' t_h / t_h' t_h \ \ ; \ \ c_h = y' t_h / t_h' t_h \ .$$

3.2.1.8. Estudio de los vectores w_h^*

Los componentes t_h se definen a partir de los residuos X_{h-1}, $t_h = X_{h-1}w_h$, pero pueden también ser expresados en función de X (ver propiedad 1.g)

$$t_h = Xw_h^* \quad ; \quad T_h = XW_h^*$$

con $W_h^* = W_h(P_h'W_h)^{-1}$.

Se muestra que la matriz $P_h'W_h$ es triangular superior con todos sus elementos diagonales iguales a 1. Se puede también constatar que $P_h'W_h^* = I$ y que por consiguiente W_h^* es una inversa generalizada de P_h'.

3.2.1.9. Estudio de la ecuación de regresión PLS

Se puede escribir la fórmula de regresión de Y sobre las componentes $t_1, \ldots, t_h$ en función de las variables X, así:

$$\begin{aligned}\hat{y} &= T_hC_h = t_1c_1 + \ldots + t_hc_h \\ &= XW_h^*C_h \\ &= XW_h(P_h'W_h)^{-1}C_h = Xb_h\end{aligned}$$

donde $b_h = W_h(P_h'W_h)^{-1}C_h$ es el vector de los coeficientes de regresión PLS de Y sobre X utilizando h componentes.

3.2.2. Regresión Multivariante PLS2

Se retoman las presentaciones más importantes de este algoritmo en los libros de Martens 1989, Esbensen 1994, Höskuldsson 1996.

Sea Y la matriz de variables dependientes $y_1, \ldots, y_r$ e X la matriz de variables independientes $x_1, \ldots, x_p$ de rango a medidas sobre n individuos. Todas las variables están centradas y reducidas. Puede existir multicolinealidad al interior de cada bloque e inclusive r y p ser mayor que n, e igualmente existir datos faltantes.

Se tiene ahora, dos conjuntos de variables *Y* e *X* para los cuales se asume que existe una relación subyacente entre los dos bloques explicada por *H* componentes latentes ortogonales t_h $(h = 1,2, \dots, H)$ calculadas como combinación lineal de las variables del conjunto predictor *X*, y guardando alta relación con *Y* a través de su combinación lineal $u_h = Yc_h$.

Así, las matrices predictor y respuesta son descompuestas de la forma:

$$X = T_H P_H' + X_H$$

$$Y = T_H C_H' + Y_H$$

donde P_H y C_H son las matrices de cargas conteniendo los parámetros del modelo, y X_H y Y_H las matrices de residuos representando la parte de variabilidad de los datos no explicada por los modelos de parámetros.

Hay numerosas versiones del algoritmo PLS2, que difieren en el nivel de normalización escogido. Se describe el algoritmo de regresión PLS2 clásico, teniendo en cuenta la gestión de datos faltantes de acuerdo con los principios del NIPALS (Lindgren et at, 1993).

$X_0 = X$, $Y_0 = Y$
para $h = 1, 2, \dots, a$

1. Inicializo: u_h $(u_1 : 1^{\underline{a}}\ col\ de\ Y_{h-1}, \dots)$
2. Repetir hasta convergencia de w_h

$w_h = X_{h-1}' u_h / \|u_h\|$ # [coef regresión[8] simple x_j sobre u_h]
$w_h/\|w_h\|$: normar w_h a 1

$t_h = X_{h-1} w_h / (w_h' w_h)$ [*componente_h de X*]

$c_h = Y_{h-1}' t_h / (t_h' t_h)$ [‘*vector de* R^r]

$u_h = Y_{h-1} c_h / (c_h' c_h)$ [*componente_h de Y*]

3. $p_h = X_{h-1}' t_h / (t_h' t_h)$ [*vect-p de X en* R^p]
4. $X_h = X_{h-1} - t_h p_h'$ [*residuo de X, ortogonalidad*]
5. $Y_h = Y_{h-1} - t_h c_h'$

Fin h.

[8] No se realiza regresión múltiple, ya que bajo ACP $X'\psi = \lambda u$; $\frac{X'u}{\|u\|} = w$.

Cuando hay datos faltantes, se aplica los principios del algoritmo NIPALS: las coordenadas de los vectores w_h , t_h , c_h , u_h e p_h son calculadas como las pendientes de las rectas de mínimos cuadrados pasando por el origen, sólo sobre los datos disponibles.

3.2.2.1. Criterio de optimización

Se puede precisar la convergencia en la etapa 2. Las relaciones cíclicas de esta etapa muestran que en el *límite*, los vectores w_h , t_h , c_h e u_h mediante reemplazamientos sucesivos, verifican las ecuaciones siguientes:

$$\left(\frac{1}{n-1}X'_{h-1}Y_{h-1}\right)\left(\frac{1}{n-1}Y'_{h-1}X_{h-1}\right)w_h = \lambda_h w_h \quad [3.5]$$

$$\left(\frac{1}{n-1}X_{h-1}X'_{h-1}\right)\left(\frac{1}{n-1}Y_{h-1}Y'_{h-1}\right)t_h = \lambda_h t_h$$

$$\left(\frac{1}{n-1}Y'_{h-1}X_{h-1}\right)\left(\frac{1}{n-1}X'_{h-1}Y_{h-1}\right)c_h = \lambda_h c_h \quad [3.6]$$

$$\left(\frac{1}{n-1}Y_{h-1}Y'_{h-1}\right)\left(\frac{1}{n-1}X_{h-1}X'_{h-1}\right)u_h = \lambda_h u_h \ .$$

$\lambda_h = t'_h t_h . c'_h c_h . \|X'_{h-1}u_h\|$ es el valor propio más grande común a estas matrices, las cuales han sido divididas por *n-1* para recobrar los *valores propios*. La etapa *2* corresponde entonces a una aplicación de la potencia iterativa para el cálculo del vector propio de una matriz, asociado al valor propio más grande en cada *h*.

Se obtienen las componentes t_h e u_h a partir de la *primera* etapa del AIB de Tucker de las tablas X_{h-1} e Y_{h-1} . En cada etapa *h* se investiga entonces dos vectores normados w_h e c_h^* maximizando el criterio $cov(X_{h-1}w_h, Y_{h-1}c_h^*)$, o globalmente maximizando el criterio

$$\sum_{h=1}^{s} cov^2(X_{h-1}w_h, Y_{h-1}c_h^*) \quad [3.7]$$

el vector c_h es colineal al vector $c_h^* = c_h/\|c_h\|$ y $s \leq a$.

Se construye enseguida las tablas X_h e Y_h en las etapas 4 y 5 como residuos de las regresiones de X_{h-1} e Y_{h-1} sobre la componente t_h .

Se puede observar que la propiedad *a*) sigue siendo cierta; se deduce en particular la ortogonalidad de las componentes PLS t_h. Esta ortogonalidad se deriva directamente de la construcción del algoritmo de regresión PLS2 como se puede ver utilizando la siguiente proposición:

- ***Proposición***

Sean $t_1 = Xa_1, \dots, t_h = Xa_h$ un conjunto de componentes ortogonales y X_h el residuo de la regresión de *X* sobre $t_1, \dots, t_h$. Entonces el conjunto de combinaciones lineales de las columnas de *X* ortogonales a las componentes $t_1, \dots, t_h$ es igual al conjunto de las combinaciones lineales de las columnas de X_h .

La ecuación de regresión de *X* sobre las componentes ortogonales se escribe:

$$X = t_1 p_1' + \cdots + t_h p_h' + X_h \quad ; \quad p_k' = (t_k' t_k)^{-1} t_k' X \ .$$

Se define enseguida los conjuntos

$$A = \{x | x = Xa \ , \ x'(t_1, \dots, t_h) = 0\} \quad , \quad B = \{x | x = X_h c\} \ .$$

Se muestra que

$A \subset B$: $x = Xa = (t_1 p_1' + \cdots + t_h p_h' + X_h)a = X_h a$

$B \subset A$: $x = X_h c = (X - t_1 p_1' - \cdots - t_h p_h')c = X(I - a_1 p_1' + \cdots + a_h p_h')c = Xa$

y por tanto *A=B* ; de otra parte $x' t_k = c' X_h' t_k \underbrace{=}_{1.f} 0$.

3.2.2.2. Versión modificada del algoritmo de regresión PLS2

La ortogonalidad de las componentes t_h implica que las matrices X_h e Y_h son los residuos respectivos de las regresiones de *X* e *Y* sobre las componentes $t_1, \dots, t_h$.

Sea $\mathcal{P}_h$ el operador de proyección ortogonal sobre el espacio engendrado por las componentes $t_1, \dots, t_h$.

Se deduce entonces de $X_{h=}(I - \mathcal{P}_h)X \quad ; \quad Y_{h=}(I - \mathcal{P}_h)Y$

las relaciones $X_h'Y_h = X'(I - \mathcal{P}_h)(I - \mathcal{P}_h)Y = X_h'Y$

$$X_h'X_h = X'X_h \ .$$

de donde las modificaciones posibles en el algoritmo de regresión PLS2 son:

$$c_h = Y't_h/t_h't_h \ ; \quad \tilde{u}_h = Yc_h/(c_h'c_h) \ ; \quad p_h = X't_h/t_h't_h$$

por tanto $w_h = X_{h-1}'Yc_h/\|X_{h-1}'Yc_h\|$ y u_h ahora es $\tilde{u}_h$.

El cálculo de los residuos Y_h es del todo inútil para la determinación de las componentes PLS $t_1, \dots, t_h$. Sin embargo estos residuos son utilizados para escoger el número de componentes.

Entonces la versión modificada del algoritmo PLS2 conduce a los mismos vectores w_h, t_h, c_h e p_h que el algoritmo clásico cuando no hay datos faltantes. Sin embargo, utilizando los principios de NIPALS el algoritmo de regresión PLS2 modificado puede también funcionar con datos faltantes.

3.2.2.3. Algoritmo PLS2 modificado

$X_0 = X$, $Y_0 = Y$

para $h = 1, 2, \dots, a$.

1. Inicializo: $\tilde{u}_h$ (*1ª columna* de *Y,…*)
2. Repetir hasta convergencia de w_h

   $w_h = X_{h-1}'\tilde{u}_h/\|\tilde{u}_h\|$ # [w_{hj} coef de regresión [prod escalar] x_j sobre u_h]

   $w_h/\|w_h\|$: normar w_h a 1.

   $t_h = X_{h-1}w_h/(w_h'w_h)$ [*componente h de X, (w_h vector propio inic R^p)*]

   $c_h = Y't_h/(t_h't_h)$ [“*vector propio de* R^q”], *en realidad es c**

   $\tilde{u}_h = Yc_h/(c_h'c_h)$ [“*componente*”_*h de Y*]

3. $p_h = X't_h/(t_h't_h)$ [‘‘*vect-p’’ de X en* R^p]
4. $X_h = X_{h-1} - t_hp_h'$ [*residuo de X, garantiza ortogonalidad siguiente* p_h]
5. $Y_h = Y_{h-1} - t_hc_h'$

Fin h.

Las relaciones cíclicas de la etapa 2, de este algoritmo conducen en cada etapa *h* a las ecuaciones siguientes:

$$\left(\frac{1}{n-1}X'_{h-1}Y\right)\left(\frac{1}{n-1}Y'X_{h-1}\right)w_h = \lambda_h w_h$$

$$\left(\frac{1}{n-1}X_{h-1}X'_{h-1}\right)\left(\frac{1}{n-1}YY'\right)t_h = \lambda_h t_h$$

$$\left(\frac{1}{n-1}Y'X_{h-1}\right)\left(\frac{1}{n-1}X'_{h-1}Y\right)c_h^* = \lambda_h c_h^*$$

$$\left(\frac{1}{n-1}YY'\right)\left(\frac{1}{n-1}X_{h-1}X'_{h-1}\right)\tilde{u}_h = \lambda_h \tilde{u}_h$$

λ_h es el valor propio común a estas matrices.

Así cada etapa del algoritmo de regresión PLS2 modificado se reduce a la primera etapa del AIB de las tablas X_{h-1} e *Y*. En cada etapa se obtiene las componentes t_h e $\tilde{u}_h$ en búsqueda de los vectores *normados* w_h e c_h^* maximizando el criterio $cov(X_{h-1}w_h, Yc_h^*)$; como antes, c_h es colineal al vector c_h^* ; es decir $c_h^* = c_h/\|c_h\|$. Se pasa a la etapa siguiente construyendo el residuo X_h de la regresión de X_{h-1} sobre t_h .

3.2.2.4. Estudio de las componentes PLS2 t_h , u_h , $\tilde{u}_h$

Sean $T_h = [t_1, \dots, t_h]$, $W_h = [w_1, \dots, w_h]$, $P_h = [p_1, \dots, p_h]$, $W_h^* = [w_1^*, \dots, w_h^*]$. Se tiene en regresión PLS2 que $T_h = XW_h^*$ con $W_h^* = W_h(P_h'W_h)^{-1}$.

El algoritmo PLS2 clásico proporciona las componentes $u_h = Y_{h-1}c_h/(c_h'c_h)$ y el algoritmo modificado las componentes $\tilde{u}_h = Yc_h/(c_h'c_h)$. La componente u_h es difícil de interpretar puesto que, salvo para u_1 , ella no pertenece al espacio de las *Y*. Por el contrario $\tilde{u}_h$ es una combinación lineal de las *Y*, y por consiguiente de naturaleza más interpretable.

La división por $c_h'c_h$ proviene de la formulación del algoritmo de acuerdo al enfoque NIPALS con el fin de poder tomar en cuenta los datos faltantes. Permite también obtener la primera bisectriz como recta de mínimos cuadrados de la nube de puntos (t_h, u_h) y $(t_h, \tilde{u}_h)$.

Los coeficientes de regresión de u_h sobre t_h y de $\tilde{u}_h$ sobre t_h son:

$$\frac{u_h' t_h}{t_h' t_h} = \frac{c_h' Y_{h-1}' t_h}{t_h' t_h c_h' c_h} = 1 \quad , \quad \frac{\tilde{u}_h' t_h}{t_h' t_h} = \frac{c_h' Y' t_h}{t_h' t_h c_h' c_h} = 1$$

pues $Y_{h-1}' t_h = c_h t_h' t_h = Y' t_h$ tanto en el algoritmo clásico como en el modificado.

De la fórmula de regresión de *Y* sobre $t_1, \ldots, t_{h-1}$ [reconstitución de *Y* via ACP]:

$$Y = t_1 c_1' + \cdots + t_{h-1} c_{h-1}' + Y_{h-1} \ .$$

Postmultiplicando *Y* por $c_h / c_h' c_h$ se deduce también que u_h es el residuo de la regresión de $\tilde{u}_h$ sobre $t_1, \ldots, t_{h-1}$.

La ortogonalidad de las componentes t_h permite relacionar las correlaciones de u_h e $\tilde{u}_h$ con t_h . Se tiene en efecto

$$cor(\tilde{u}_h, t_h) = \frac{\tilde{u}_h' t_h}{\|\tilde{u}_h'\| \|t_h\|} = \frac{\|u_h\|}{\|\tilde{u}_h\|} = cor(u_h, t_h) \sqrt{1 - R^2(\tilde{u}_h; t_1, \ldots, t_{h-1})} \ .$$

Así, reemplazando $\tilde{u}_h'$ e Y' de la reconstitución, la correlación de la componente t_h con la componente $\tilde{u}_h$ es necesariamente inferior a su correlación con la componente u_h .

Las componentes $\tilde{u}_h = Y c_h / (c_h' c_h)$ están directamente relacionadas con las variables dependientes *Y* y pueden ser interpretadas. Las componentes $u_h = Y_{h-1} c_h / (c_h' c_h)$ construidas sobre los residuos Y_{h-1} están mejor correlacionadas con las variables t_h pero no se pueden interpretar.

3.2.2.5. Interpretación de las componentes t_h e $\tilde{u}_h$

Las componentes t_h e $\tilde{u}_h$ se escriben en función de las variables originales:

$$t_h = w_h^* x_1 + \cdots + w_{hp}^* x_p \quad , \quad \tilde{u}_h = (c_{h1} y_1 + \cdots + c_{hq} y_q) / c_h' c_h \ .$$

Las correlaciones de las variables x_j e y_k con las componentes t_h están relacionadas proporcionalmente con los vectores p_h e c_h respectivamente; en efecto:

$$cor(x_j, t_h) = \frac{x_j' t_h}{\|x_j\| \|t_h\|} = \frac{x_j' t_h}{\sqrt{n-1}\sqrt{t_h' t_h}} = s_{th} . p_{hj}$$

$$cor(y_k, t_h) = \frac{y_k' t_h}{\|y_k\| \|t_h\|} = s_{th} . c_{hk} \quad .$$

En la interpretación multivariada de las componentes t_h e $\tilde{u}_h$ se utiliza los vectores w_h^* e c_h .

La construcción de la carta superpuesta (w_1^*, w_2^*) e (c_1, c_2) permite interpretar visualmente las dos primeras componentes de la regresión PLS. Esta carta mezcla varios puntos de vista:

- Los puntos correspondientes a las variables x_j describen la construcción de las componentes t_h ;
- Los puntos correspondientes a las variables y_k expresan a la vez la construcción de las componentes $\tilde{u}_h$ y la correlación entre y_k e t_h .

La carta de correlaciones de las variables x_j e y_k con las componentes (t_1 , t_2) proporciona una interpretación univariada de esas componentes. A los factores de escala s_{t_1} e s_{t_2} cerca, esta carta corresponde a la superposición de (p_1 , p_2) e (c_1 , c_2) .

3.2.2.6. Estudio de la ecuación de regresión PLS2

Se expresa la regresión de *Y* sobre las componentes $t_1, \dots, t_H$ en función de las variables X. Se tiene

$$Y = t_1 c_1' + \cdots + t_H c_H' + Y_H \quad .$$

Haciendo $C_H = [c_1, \dots, c_H]$ se tiene

$$Y = T_H C_H' + Y_H = X W_H^* C_H' + Y_H$$

$$Y = X W_h (P_h' W_h)^{-1} C_H' + Y_H = XB + Y_H \quad .$$

Así, los coeficientes de regresión asociados a X luego de desdoblar el modelo en función de $t_1, \dots, t_H$ corresponden a $B = W_h (P_h' W_h)^{-1} C_H'$.

Para $H = 2$, la ecuación de regresión para cada variable respuesta es:

$$y_k \approx c_{1k}t_1 + c_{2k}t_2$$

$$y_k \approx (c_{1k}w_{11}^* + c_{2k}w_{21}^*)x_1 + \cdots + (c_{1k}w_{1p}^* + c_{2k}w_{2p}^*)x_p$$

$$y_k \approx b_{k1}x_1 + \cdots + b_{kp}x_p \ .$$

Por consiguiente la carta superpuesta (w_1^*, w_2^*) e (c_1, c_2) proporciona indicaciones sobre los coeficientes b_{kj} . Estos son positivos nulos o negativos según que los vectores x_j *e* y_k formen un ángulo agudo, recto u obtuso.

3.3. PLS – PM

El objetivo de PLS-PM es estimar las relaciones entre J bloques $X_1, \ldots, X_j, \ldots, X_J$ de variables 'manifiestas' (MV), las cuales son expresión de J constructos no observables $\xi_1, \ldots, \xi_j, \ldots, \xi_J$ respectivamente, usualmente llamados LV; (Wold 1975a, 1982)

PLS-PM estima la red de relaciones entre las MV y su propia LV, y entre las LV al interior del modelo, a través de un sistema de ecuaciones basadas en regresiones simples y múltiples.

Los métodos PLS-PM propuestos por Wold para el análisis de Modelos de Ecuaciones Estructurales (SEM) basados en componentes, operan sisn supuestos distribucionales y son una alternativa a los análisis de estructura de covarianzas basados en la multinormalidad, Joreskog (1970).

Al igual que los SEM cada modelo PLS-PM está compuesto de dos submodelos: el modelo *de medición*, y el modelo *estructural*. El primero tiene en cuenta las relaciones entre cada variable latente y sus correspondientes variables manifiestas, mientras el modelo *estructural* asume estar interconectado linealmente entre las LV de acuerdo a un modelo de relaciones causa-efecto.

El fenómeno de estudio es descrito por las relaciones *estructurales (inner)* entre las variables latentes de la forma

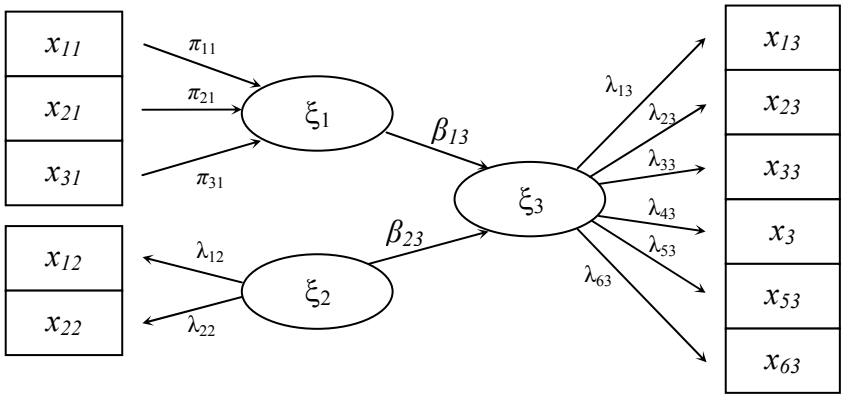
$$\xi_j = \beta_{j0} + \sum_k \beta_{jk}\xi_k + \zeta_j \ . \qquad [3.8]$$

Los coeficientes *'path'* β_{jk} expresan el impacto sobre la variable endogena ξ_j de las LV ξ_k conectadas, algunos pueden ser estructuralmente nulos y la variable correspondiente ξ_k no aparecer en la ecuación [3.8].

La 'hipótesis de especificación de predicción' es que el término de error aleatorio ζ_j tiene media 0 y no está correlacionado con las otras LV ξ_k de la ecuación; esto es, $E(\zeta_j) = 0$, $cov(\zeta_j,\ \xi_k) = 0$.

La formulación del modelo de *medición (outer)* depende de la dirección de la relación entre la LV y las correspondientes MV; así, el modelo puede ser *reflectivo, formativo* o ambos.

Se representan los datos y los modelos en un esquema de flechas, ver gráfica 3.4. Las variables *manifiestas* son representadas por rectángulos y las LV por elipses. Las relaciones de causalidad son simbolizadas por flechas, el origen de la flecha es la variable "causa" (explicativa) y la punta la variable "efecto" (a explicar) ó variable endógena.



Gráfica 3.4. Red de causalidad entre 3 grupos de variables; el primero *formativo* asociando pesos π y los otros *reflectivos* con pesos λ

Se define *reflectivo* si la variable latente (LV) es un factor común que refleja en si mismo cada MV; la LV es la causa de cada una de las MV. Las MV estarán entonces altamente correlacionadas; el bloque debe ser homogéneo[9]. Cada grupo de variables es esencialmente unidimensional:

[9] Las principales reglas para valorar homogeneidad son: a) λ_1 >>1 en el ACP b) ρ de Dillon-Goldstein's (Chin, 1998) > 0.7; $\rho = \left(\sum_p^{P_j} \lambda_{jp}\right)^2 / [\left(\sum_p^{P_q} \lambda_{jp}\right)^2 + \sum_p^{P_j}(1 - \lambda_{jp}^2)]$.

Cada variable es función del factor ξ_j centrado reducido. En un modelo *reflectivo* cada *p*-MV está relacionada a la correspondiente *j*-LV por un modelo de regresión simple, es decir:

$$x_{jp} = \lambda_{jp0} + \lambda_{jp}\,\xi_j + \varepsilon_{jp} \quad .$$

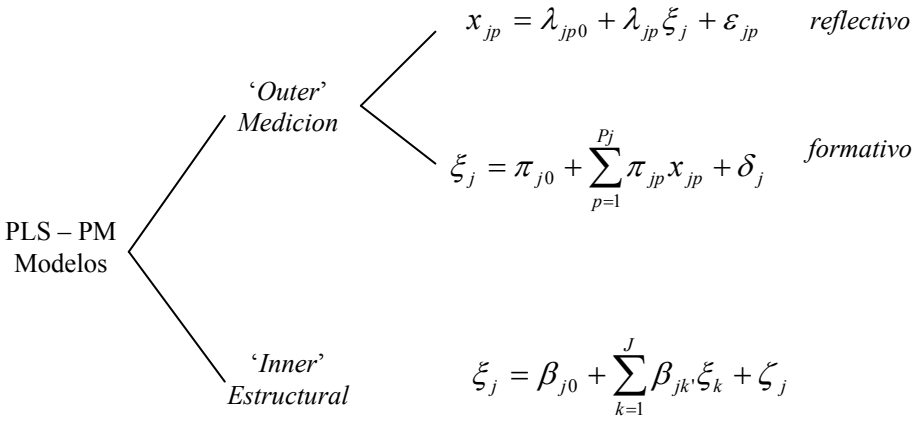
λ_{jp} es la carga asociada a la *p*-ésima MV en el bloque *j*, y la hipótesis es que el error ε_{jp} tiene media cero y no esta correlacionado con la variable latente ξ_j esto es: $E(\varepsilon_{jp}) = E(\xi_j \varepsilon_{jp}) = 0$.

En el caso *formativo* cada MV representa una dimensión diferente del concepto subyacente, y se consideran variables exógenas en el modelo de medición; cada sub-bloque de MV presenta una relación estrictamente causal con la LV y por tanto

$$\xi_j = \pi_{j0} + \sum_p^{Pq} \pi_{jp} x_{jp} + \delta_j \quad \text{con} \quad E(\delta_j) = E(x_{jp}, \delta_j) = 0 \quad .$$

donde π_{jp} es el coeficiente relacionando cada variable manifiesta a la correspondiente variable latente y el termino error δ_j la fracción de la correspondiente LV no tenida en cuenta por el bloque de variables manifiestas.

Una síntesis gráfica de los modelos *outer* e *inner* en PLS-PM es presentada en la gráfica 3.5:



Gráfica 3.5. Esquema de los modelos *outer* e *inner* en PLS-PM

3.3.1. Estimación de variables latentes

Independientemente del tipo de modelo (*outer*) de *medición*, las LV estandarizadas ξ_j son estimadas por las combinaciones lineales Y_j de las MV de su grupo *j*:

$$Y_j = \sum_p^{P_j} w_{jp} x_{jp} = X_j w_j$$

donde las variables x_{jp} están centradas y w_{jp} son los pesos *outer* escalados para dar varianza unitaria a Y_j . Es decir, $\hat{\xi}_j = Y_j = \pm a_j \sum_p^{P_j} \tilde{w}_{jp} x_{jp}$ donde a_j es un escalar que da varianza unitaria a Y_j , y la ambigüedad del $\pm$ es resuelta escogiendo el signo tal que la mayoría de las x_{jp} esta positivamente correlacionada con la LV estandarizada Y_j ; (Sánchez 2008).

También se define la aproximación *inner* Z_j de ξ_j como una combinación lineal normalizada de las estimaciones *outer* de las LV adyacentes,

$$Z_j = \sum_k c_{jk} e_{jk} Y_k \qquad [3.9]$$

donde $c_{jk} = 1$ si ξ_j esta conectada a ξ_k y $c_{jk} = 0$ en otro caso; e_{jk} son los pesos *inner*.

Relacionando estas dos aproximaciones, Wold obtiene las condiciones de *estacionaridad* que permiten determinar las variables Y_j. Se resuelven las ecuaciones de estacionaridad por un proceso iterativo hasta la convergencia.

Wold propone dos modos de relación entre las dos aproximaciones Y_j e Z_j de la variable latente ξ_j, que además permite actualizar los pesos $\underbrace{w_j}$ en el caso *outer:*

- En el *Modo A* (*reflectivo*) la variable Y_j está relacionada con la variable Z_j por la fórmula

$$Y_j \propto \sum_p \underbrace{cov(x_{jp} Z_j)}\, x_{jp} \;=\; X_j \underbrace{X_j' Z_j} \qquad [3.10]$$

  de [3.9] la condición de estacionaridad para Y_j en el modelo *A* es:

$$Y_j \propto X_j X_j' \sum_k c_{jk} e_{jk} Y_k \; . \qquad [3.11]$$

- En el *Modo B* (*formativo*) la variable Y_j es obtenida por regresión múltiple de Z_j sobre las columnas x_{jp} de la matriz X_j en la reducción

$$Y_j \propto X_j \underbrace{\left(X_j' X_j\right)^{-1} X_j' Z_j}_{} \quad . \qquad [3.12]$$

  Por lo tanto la condición de estacionariedad para una variable Y_j en el modo B es:

$$Y_j \propto X_j (X_j' X_j)^{-1} X_j' \sum_k c_{jk} e_{jk} Y_k \quad . \qquad [3.13]$$

Las estimaciones definitivas de las componentes latentes Y_j se tienen, eligiendo el modo *A* o *B* para su cálculo, según se consideren los grupos como *reflectivos o formativos*.

En el modo *A* la variable Y_j definida por la formula [3.10] corresponde igualmente a la primera componente PLS reducida de la regresión simple de Z_j sobre X_j. Además, permite tomar en cuenta sin dificultad los datos faltantes siguiendo los principios del algoritmo NIPALS.

En el modo *inner* hay tres vías *weighting* para calcular los pesos e_{jk}:

a) Esquema del '*centroide*' (Wold): $e_{jk} = sign[cor(Y_j, Y_k)]$ es el *signo* de la correlación entre Y_j *e* Y_k

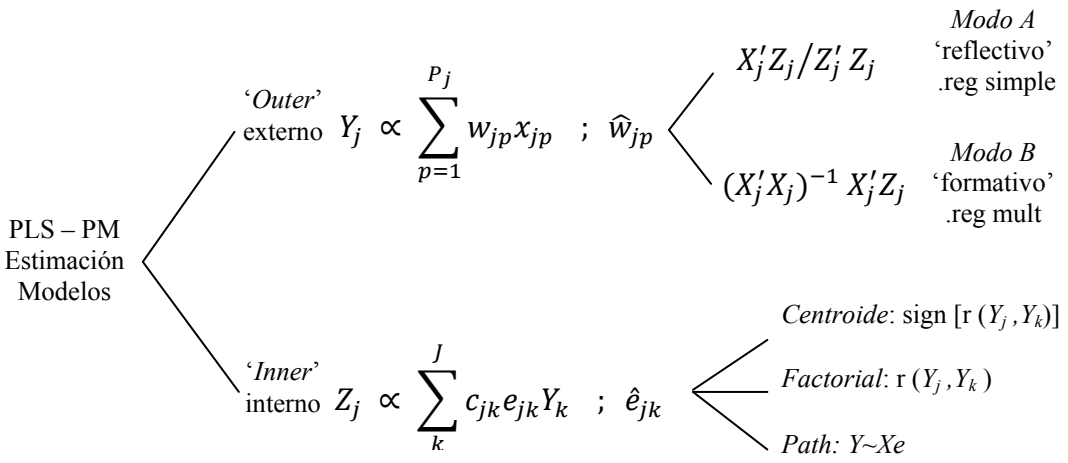
b) Esquema '*factorial*': $e_{jk} = cor(Y_j\ ,\ Y_k)$

c) Esquema '*path*'. Aquí $e_{jk} = cor\left(Y_j, Y_k\right)$ si ξ_j es explicada por ξ_k , ó

$$e_{jk} = \begin{cases} cor(Y_j, Y_k) \text{ si } \xi_j \text{ es explicada por } \xi_k \\ coeficientes\ en\ la\ regresion\ multiple\ Y_j = \sum_k e_{jk} Y_k \end{cases}$$

Una vez son obtenidos los pesos w_{jp}, los score normalizados de las LV son finalmente calculados mediante $\hat{\xi}_j \propto X_j w_j$.

El proceso de estimación de los modelos PLS-PM se puede esquematizar en la gráfica 3.6 como sigue:



Gráfica 3.6. Síntesis esquemática PLS-PM estimado

En el último paso del algoritmo PLS-PM los coeficientes *path* del modelo estructural son estimados a través de una regresión múltiple OLS entre los puntajes de las LV estimadas. Denotando Ξ_j la matriz de los correspondientes 'predictores' latentes, el vector de coeficientes path para cada ξ_j es $\hat{\beta}_j = (\hat{\Xi}_j{}'\hat{\Xi}_j)^{-1}\hat{\Xi}_j{}'\hat{\xi}_j$.

En caso de multicolinealidad entre las LV estimadas, los modelos PLS-R son usados en vez de la regresión convencional (Esposito Vinzi et al. 2010). Se observa que el algoritmo RGCCA propuesto por Tenenhaus and Tenenhaus (2011) para ganar propiedades de optimización, mantiene una estructura muy similar al algoritmo de Wold para PLS-PM, descrito a continuación.

El algoritmo del apartado 3.3.2 propuesto por Wold ahora es natural. En la etapa inicial se parte de variables Y_k arbitrariamente fijadas con alguna de las variables x_{kp} ligadas positivamente a su variable latente ξ_k .

Se obtiene con la ayuda de la ecuacion [3.9] nuevos valores de los pesos y componenters *inner*. Las estimaciones Y_j de las LV ξ_j son obtenidas iterando el proceso hasta la convergencia. Los vectores w_j se deducen enseguida de las ecuaciones [3.11] e [3.13].

3.3.2. Algoritmo PLS-PM de Wold

Entrada: $X = [\, X_1, \ldots, X_j, \ldots, X_J \,], C$; Salida: $\beta_j, w_j, \hat{\xi}_j$;

1. Inicio

   1.1. $w_j = w_j^{(0)}$

   1.2. $Y_j^{(0)} \propto \sum_{p=1}^{P_j} w_{jp}^{(0)} x_{jp} = X_j w_j^{(0)}$

   [*p*-ésima variable en el bloque *j*]

2. Repita

   2.1. Cálculo de los pesos *inner*

   $$\text{si } k < j \;;\quad e_{jk}^{(s)} = f\left(Y_j^{(s)}, X_k w_k^{(s+1)}\right)$$
   $$\text{si } k > j \;;\quad e_{jk}^{(s)} = f\left(Y_j^{(s)}, Y_k^{(s)}\right)$$

   2.2. Estimación de las LV *Inner*

   $$z_j \propto \left(\sum_{k<j} c_{jk} e_{jk}^{(s)} X_k w_k^{(s+1)} + \sum_{k>j} c_{jk} e_{jk}^{(s)} Y_k^{(s)}\right)$$

   2.3. Cálculo de los pesos *outer*

   $$w_j^{(s+1)} = (1/N) X_j^{'} z_j^{(s)} \quad (Mode\ A) \quad \text{ó}$$
   $$w_j^{(s+1)} = (X_j^{'} X_j)^{-1} X_j^{'} z_j^{(s)} \quad (Mode\ B)$$

   2.4. Estimación de las LV o*uter*

   $$Y_j^{(s+1)} \propto \sum_{p=1}^{P_j} w_{jp}^{(s+1)} x_{jp} = X_j w_j^{(s+1)}$$

   Hasta la convergencia de w_j

3. Computo de las LVs

   $$\hat{\xi}_j \propto X_j w_j$$

4. Cálculo de los coeficientes *Path*

   $$\beta_j = \left(\hat{\Xi}'_{\rightarrow \mathrm{j}} \hat{\Xi}_{\rightarrow \mathrm{j}}\right)^{-1} \hat{\Xi}'_{\rightarrow \mathrm{j}}\, \hat{\xi}_{\mathrm{j}}$$

Fin

3.3.3. Valoración y calidad del modelo

La calidad de un modelo PLS-PM depende de la calidad tanto del modelo de *medición* como del modelo *estructural*. En un buen modelo de *medición*, cada MV estará bien resumida por su propia LV. Así, para cada bloque un *índice de comunalidad* es

$$com_j = \frac{1}{P_j}\sum_p^{P_j} cor^2\left(x_{jp}, \hat{\xi}_j\right) = \frac{1}{P_j}\sum_p^{P_j} \hat{\lambda}_{jp}^2 \ .$$

Este es el promedio de las comunalidades entre cada MV y $\hat{\xi}_j$ del bloque *j*.

La bondad de ajuste de *todo* el modelo de *medición* se calcula con el índice de comunalidad promedio

$$\overline{com} = \frac{\sum_{p:P_j>1} P_j com_j}{\sum_{p:P_j>1}} \ .$$

Aunque la calidad de cada ecuación estructural es medida por el coeficiente de determinación R^2, este no es suficiente para evaluar todo el modelo estructural.

Para cada LV endógena, el siguiente índice de redundancia mide la porción de variabilidad de las MV relacionadas a la LV endógena ξ_j explicada por sus predictores latentes

$$red_j = com_j . R_j^2 \ .$$

El índice promedio de redundancia mide la calidad de todo el modelo estructural. Si J es el número de LV endógenas, entonces

$$\overline{red} = \frac{1}{J}\sum_j^J red_j \ .$$

La calidad global del modelo es evaluada por el índice de bondad de ajuste *GoF* (Tenenhaus et al. 2004). Este índice ha sido creado para medir el desempeño total de predicción del modelo conjunto, el de *medición* y el *estructural*. Así,

$$GoF = \sqrt{\overline{com}.\bar{R}^2} \ .$$

Este índice es normalizado (relativo) para que tome valores entre 0 y 1 y está determinado por

$$GoF_{rel} = \sqrt{\frac{1}{\sum_{j:P_j>1} P_j} \sum_{j:P_j>1} \frac{\sum_p^{P_j} \hat{\lambda}_{jp}^2}{\lambda_j^1} \cdot \frac{1}{J} \sum_j \frac{R_j^2}{\rho_j^2}}$$

λ_j es obtenido del ACP y ρ_j del ACC.

3.3.4. Criterios de optimización

Se presenta aquí como la utilización del algoritmo PLS general permite retomar la primera etapa de los principales métodos clásicos de análisis multivariante utilizados para relacionar dos o más grupos.

3.3.4.1. Estudio del caso de dos bloques (*J*=2)

La implementación del algoritmo PLS permite retomar la primera etapa de los principales métodos utilizados para relacionar dos grupos (X_1 e X_2) de variables. Se precisa en la tabla 3.1, los métodos correspondientes a las diferentes escogencias posibles de los modos de cálculo *A o B* para w_1 e w_2.

	AC (canónico)	*AIB* (interbaterias)	*AR*(redundancias, de X_2 respecto a X_1)
Modo de cálculo para w_1	B	A	B
Modo de cálculo para w_2	B	A	A

Tabla 3.1. Equivalencia entre el algoritmo PLS sobre dos bloques de variables X_1 e X_2 y las primeras etapas de los diferentes métodos multivariados.

Para verificar estos resultados es suficiente describir las condiciones de estacionaridad [3.12] e [3.13] para estas diferentes situaciones. Por simplicidad se supone que las LV ξ_1, ξ_2 están correlacionadas positivamente. Por consiguiente $Z_1 = Y_2$ *e* $Z_2 = Y_1$.

- **Análisis Canónico**

Las condiciones de estacionaridad se escriben utilizando el modo *B* para Y_1 e Y_2:

$$Y_1 \propto X_1(X_1'X_1)^{-1}X_1'Y_2 \quad , \quad Y_2 \propto X_2(X_2'X_2)^{-1}X_2'Y_1 \ .$$

El algoritmo PLS converge a las primeras componentes del AC de X_1 e X_2, que maximiza la $cor(Y_1, Y_2)$.

- **Análisis Interbaterias**

Las condiciones de estacionaridad se escriben utilizando el modo *A* para Y_1 e Y_2 :

$$Y_1 \propto X_1X_1'Y_2 \quad , \quad Y_2 \propto X_2X_2'Y_1 \ .$$

El algoritmo PLS converge entonces hacia las primeras componentes del AIB de las dos tablas X_1 e X_2, que maximiza la $cov(Y_1, Y_2)$. Se puede observar que el caso particular donde $X_1 = X_2$ permite obtener la primera componente principal reducida Y_1 de X_1 , puesto que Y_1 es *vect_p* de X_1X_1' asociado al más grande valor propio, es también *vect_p* de $(X_1X_1')^2$ asociado al más grande valor propio.

- **Análisis de Redundancias**

Las condiciones de estacionaridad utilizando el modo *B* para Y_1 y el *A* para Y_2 son:

$$Y_1 \propto X_1(X_1'X_1)^{-1}X_1'Y_2 \quad , \quad Y_2 \propto X_2X_2'Y_1 \ .$$

El algoritmo PLS-PM converge a la primera componente Y_1 del AR de X_2 respecto a X_1; la redundancia de X_2 sobre X_1 es maximizada.

3.3.4.2. Caso Multibloques ($j > 2$)

El PLS-PM puede ser visto como la generalización de un número de métodos multibloques tales como el modelo *jerárquico*, el modelo *confirmativo* y el modelo *general.*

Se observará que dependiendo del modo de estimación *outer* y del esquema de estimación *inner* elegido, las ecuaciones de estacionaridad de los algoritmos asociados a estos métodos PLS-PM a veces convergen a un criterio de optimización asociado con ciertos métodos de análisis multivariante.

Los PLS-PM no siempre tienen asegurado el criterio de convergencia, Tenenhaus (2011).

- **Modelo Jerárquico**

El modelo jerárquico (Tenenehaus et al. 2005) se caracteriza porque cada bloque X_j es conectado al *superbloque* X_{J+1} obtenido por yuxtaposición horizontal de los J bloques $|X_1|X_2|\ldots|X_J|$. Para los distintos tipos de estimación se tiene:

- Modo *A* y esquema *path* conllevan al AFM.
- Modo *B* según esquema *inner* converge a ecuaciones de Horst's y ACCG.

- **Modelo General**

- *Nuevo* Modo *A* (Tenenhaus, M. 2009), esquema del *centroide*. El algoritmo de Wold así ajustado monótonamente converge al criterio

$$\max_{\|w_j\| = 1} \sum_{j\#k} c_{jk} \left| cov(X_j w_j \, , \, X_k w_k) \right| \quad .$$

- *Nuevo* Modo *A*, esquema *factorial*. Análogamente, el algoritmo de Wold así ajustado monótonamente converge al criterio

$$\max_{\|w_j\| = 1} \sum_{j\#k} c_{jk} cov^2 (X_j w_j \, , \, X_k w_k) \quad .$$

- Modo *B*, esquema del *centroide*. Las ecuaciones de estacionaridad del algoritmo PLS-PM son equivalentes a las ecuaciones de Lagrange asociadas al criterio de optimización

$$\max_{\|X_j w_j\| = 1} \sum_{j\#k} c_{jk} \left| cor(X_j w_j \, , \, X_k w_k) \right| \quad .$$

- Modo *B*, esquema *factorial*. Las ecuaciones de estacionaridad del algoritmo PLS-PM son equivalentes a las ecuaciones de Lagrange asociadas al siguiente criterio de optimización:

$$\max_{\|X_j w_j\| = 1} \sum_{j \# k} c_{jk} cor^2(X_j w_j \, , \, X_k w_k) \quad .$$

- **Modelo Confirmativo**

En el modelo confirmativo (Tenenhaus and Hanafi, 2010), cada LV está relacionada a un bloque simple y conectada a todas las LV de los otros bloques, formando una especie de red de relaciones en la que cada LV está conectada sólo una vez a las otras LV bien sea de forma endógena o exógena. Este modelo lleva a las ecuaciones de estacionaridad del ACCG (Kettenring's 1971).

Sin embargo, para precisar el criterio a optimizar via PLS-PM, Tenenhaus et al. (2011), ha propuesto el método RGCCA que bajo el *nuevo modo A*, maximiza las funciones de covarianza (valor absoluto, cuadrado) de las componentes *outer*; y bajo el modo *B* se maximizan análogamente las funciones de correlación según los esquemas *inner* tipo *Centroide y Factorial* respectivamente.

3.4. ANÁLISIS DE CORRELACIÓN CANÓNICO REGULARIZADO GENERALIZADO (RGCCA)

El RGCCA es una generalización del RCCA para tres o más conjuntos de variables sobre los mismos individuos, constituye un marco general para muchos métodos de análisis de datos multibloques (Tenenhaus and Tenenhaus 2011).

RGCCA combina el poder de los métodos de análisis multibloques (maximización de criterios bien identificados) y la flexibilidad del PLS-PM. Buscando para un punto fijo de las ecuaciones de estacionariedad relacionadas al RGCCA, un nuevo algoritmo monótonamente convergente, muy similar al algoritmo PLS propuesto por Wold es obtenido.

Considerando una red de conexiones entre los bloques, el objetivo del RGCCA es encontrar combinaciones lineales del bloque de variables (componentes del bloque) que expliquen su propio bloque, y tal que los componentes que se asumen conectados, estén altamente correlacionados.

Muchos métodos cumplen este objetivo, algunos de ellos están basados sobre la maximización de una función de correlaciones (covarianza): SUMCOR, SSQCOR, SABSCOR, SUMCOV, SABSCOV, SSQCOV. Otros están basados sobre la maximización de una función de ambos, correlaciones y covarianzas.

RGCCA constituye un marco general para análisis de datos multibloques e incluye todos los métodos anteriores como casos particulares. También contiene los enfoques PLS de Wold, Lohmöller, Kramer, pero solamente cuando el método B para todos los bloques es seleccionado.

Sin embargo en bloques de alta dimensionalidad o presencia de multicolinealidad, los métodos basados en la correlación llevan a relaciones espurias entre bloques. RGCCA constituye una versión regularizada de varios métodos basados en correlación y hace un análisis estable de posibles bloques mal-condicionados.

No hay solución analítica para RGCCA, pero se propone un algoritmo monótonamente convergente basado en una modificación del algoritmo PLS de Wold.

RGCCA es obtenido siguiendo varios pasos:

1. Definición de GCCA a nivel poblacional.
2. Construcción de las ecuaciones de estacionariedad al nivel muestral.
3. Definición de RGCCA.
4. Encontrar las ecuaciones de estacionariedad a nivel muestral para un punto fijado (fijo): algoritmo PLS para RGCCA.

3.4.1. GCCA Poblacional

El análisis de correlación canónica de dos bloques puede ser extendido a tres o más conjuntos de variables en muchas vías.

Kettenring (1971) estudió cinco métodos: SUMCOR, MAXVAR, SSQCOR, MINVAR y GENVAR. Un sexto método el SABSCOR es también considerado y juega un rol central en el enfoque PLS de Wold (1985).

Se propone una modificación de los métodos SUMCOR, SSQCOR y SABSCOR la cual tiene en cuenta algunas hipótesis sobre las conexiones entre los conjuntos de variables.

Considérese *J* (bloques) vectores columna aleatorios p_j dimensionales con media cero $x_j = \left(x_{j1}\,, \ldots, x_{jp_j}\right)'$ definidads sobre la misma población y *J* vectores columna no aleatorios p_j dimensionales $\alpha_j = \left(\alpha_{j1}\,, \ldots, \alpha_{jp_j}\right)'$. También se debe tener en cuenta una red de conexiones entre los vectores aleatorios definiendo una matriz de diseño[10]

$$C = \{c_{jk}\} \;:\; c_{jk} = \begin{cases} 1 & \text{si } x_j \text{ y } x_k \text{ conectados} \\ 0 & \text{si no} \end{cases} .$$

Sean las dos componentes lineales $\eta_j = \alpha_j' x_j = \sum_h \alpha_{jh} x_{jh}$ e $\eta_k = \alpha_k' x_k$. La correlación entre estas dos variables aleatorias es

$$\rho\left(\eta_j\,, \eta_k\right) = \rho\left(\alpha_j' x_j\,, \alpha_k' x_k\right) = \frac{\alpha_j' \sum_{jk} \alpha_k}{(\alpha_j' \sum_{jj} \alpha_j)^{1/2} (\alpha_k' \sum_{kk} \alpha_k)^{1/2}} \qquad [3.14]$$

donde: $\sum_{jj} = E\left(x_j x_j'\right)$ matriz de covarianzas , $\sum_{jk} = E\left(x_j x_k'\right)$.

3.4.1.1. Definición GCCA poblacional

Se define como el siguiente problema de optimización:

$$\text{Maximizar}_{\alpha_1, \ldots \alpha_J} \sum_{j\,,k=1, j\neq k}^{J} c_{jk} g\left(\rho\left(\alpha_j' x_j\,, \alpha_k' x_k\right)\right) \qquad [3.15]$$

bajo la restricción $V(\underbrace{\alpha_j' x_j}_{\eta_j}) = 1 \;,\quad j = 1\,, \ldots, J$

donde $g(x) = \begin{cases} x & \text{identidad: esquema Horst (Kramer 2007)} \\ |x| & \text{valor absoluto: esquema Centroide (Wold 1985)} \\ x^2 & \text{función cuadrada: esquema Factorial (Lohmöller 1989).} \end{cases}$

El esquema de Horst puede ser muy útil cuando el investigador observa componentes correlacionados positivamente y puede producir soluciones muy diferentes de los otros dos esquemas.

[10] Define la conexión entre los bloques *j* e *k*.

Observe que en [3.15] lo que se quiere es maximizar la suma de correlaciones, del valor absoluto ó del cuadrado de las correlaciones de las componentes de los bloques *j, k* conectados y es equivalente de acuerdo con [3.14] a

$$\text{Maximizar}_{\alpha_1,\dots\alpha_J} \sum_{j,k=1,j\neq k}^{J} c_{jk} g\left(\alpha_j' \Sigma_{jk}\, \alpha_k\right) \qquad [3.16]$$

bajo $\alpha_j' \sum_{jj} \alpha_j = 1$, $j = 1, \dots, J$.

La función Lagrangiana del problema de optimización [3.16] es:

$$L\left(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_J, \lambda_1, \dots, \lambda_J\right) = \sum_{j,k=1,j\neq k}^{J} c_{jk}\, g(\boldsymbol{\alpha}_j' \Sigma_{jk}\, \boldsymbol{\alpha}_k) - \varphi \sum_{j=1}^{J} \frac{\lambda_j}{2}\left(\boldsymbol{\alpha}_j' \Sigma_{jj}\, \boldsymbol{\alpha}_j - 1\right) \qquad [3.17]$$

con $\varphi = \begin{cases} 1 \text{ si } g(x) = \begin{cases} x \\ |x| \end{cases} \\ 2 \text{ si } g(x) = x^2 \end{cases}$

Se supone que $\alpha_j' \sum_{jk} \alpha_k \neq 0 \quad sino \quad \Rightarrow \quad c_{jk} = 0,$ y además $\partial L / \partial \alpha_j = 0$ produce las ecuaciones de estacionariedad para cada j:

$$\frac{1}{\varphi} \Sigma_{jj}^{-1} \sum_{k*j}^{J} c_{jk}\, g'\left(\alpha_j' \Sigma_{jk}\, \alpha_k\right) \Sigma_{jk}\, \alpha_k = \lambda_j \alpha_j \, . \qquad [3.18]$$

Con la restricción de normalización

$$\alpha_j' \Sigma_{jj}\, \alpha_j = 1 \ . \qquad [3.19]$$

Estas ecuaciones de estacionariedad no tienen solución analítica, pero son usadas para construir un algoritmo convergente en la optimización del problema [3.16].

3.4.2. Algoritmo PLS para GCCA poblacional

Observe que se usará $Cov\left(\alpha_j' x_j \ , \ \alpha_k' x_k\right) = \alpha_j' \sum_{jk} \alpha_k$, $w(x) = \frac{1}{\varphi} g'(x)$ para facilitar tratamientos matemáticos.

- En terminología PLS α_j es el vector de pesos *outer* y $n_j = \alpha_j' x_j$ es llamada la componente *outer*, mientras la componente *inner* ξ_j es definida:

$$\xi_j = \sum_{k\neq j}^{J} c_{jk}\, w\left[Cov\left(\alpha_j' x_j\,, \alpha_k' x_k\right)\right]\alpha_k' x_k \qquad [3.20]$$

entonces, y de acuerdo con los esquemas:

$g(x)$

Horst: $x \Rightarrow g'(x) = 1,\ \varphi = 1$ y $w(x) = 1$:

$$\xi_j = \sum_{k\neq j} c_{jk}\ \alpha_k' x_k$$

Factorial: $x^2 \Rightarrow g'(x) = 2x,\ \varphi = 2$ y $w(x) = x$:

$$\xi_j = \sum_{j\neq k} c_{jk}\ Cov\left(\alpha_j' x_j\,, \alpha_k' x_k\right)\alpha_k' x_k$$

Centroide: $|x| \Rightarrow g'(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases},\ \varphi = 1$ y $w(x) = sig(x)$

$$\xi_j = \sum_{j\neq k} c_{jk}\ sig\left[Cov\left(\alpha_j' x_j\,, \alpha_k' x_k\right)\right]\alpha_k' x_k$$

- Estas componentes *inner* ξ_j son muy útiles para simplificar las ecuaciones de estacionariedad [3.18] para GCCA poblacional usando la siguiente expresión:

$$Cov\left(x_j\,, \xi_j\right) = E\left(x_j \xi_j\right) \dddot{=} \frac{1}{\varphi} \sum_{k\neq j} c_{jk} g'\left(Cov\left(\alpha_j' x_j\,, \alpha_k' x_k\right)\right) \Sigma_{jk}\alpha_k \ .$$

De acuerdo con [3.18]

$$\sum_{jj}^{-1} Cov\left(x_j, \xi_j\right) = \lambda_j \alpha_j \qquad [3.21]$$

de donde $\alpha_j = \frac{1}{\lambda_j} \sum_{jj}^{-1} cov(x_j, \xi_j)$ y por tanto de la restricción $\alpha_j' \sum_{jj} \alpha_j = 1$

se tiene $Cov\left(x_j\,, \xi_j\right)' \sum_{jj}^{-1} Cov\left(x_j\,, \xi_j\right) = \lambda_j^2$.

Así, las ecuaciones de estacionariedad son:

$$\alpha_j = \left[Cov\left(x_j\,, \xi_j\right)' \sum_{jj}^{-1} Cov\left(x_j\,, \xi_j\right)\right]^{-1/2} \sum_{jj}^{-1} Cov\left(x_j\,, \xi_j\right) . \qquad [3.22]$$

- También se abrevia el criterio a maximizar en [3.16] de acuerdo con la siguiente proposición:

Proposición 1: para toda $g(\ )$ igual a la identidad, valor absoluto, cuadrado:

$$\sum_{j\neq k}^{J} c_{jk} g\ \left[Cov\left(\alpha_j' x_j\ , \alpha_k' x_k\right)\right] = \sum_{j}^{J} Cov(\underbrace{\alpha_j' x_j}_{\eta_j}\ , \xi_j)\ . \qquad [3.23]$$

Las ecuaciones de estacionariedad [3.22] y la proposición 1, sugieren un algoritmo iterativo monótonamente convergente para la optimización del problema [3.16] a nivel poblacional mediante:

a) Comience con pesos *outer* anormalizados arbitrarios $\alpha_j\ ,\ j = 1, \dots, J$.
b) Compute los componentes *inner* ξ_j de acuerdo con [3.20], $j = 1, \dots, J$.
c) Compute nuevos pesos *outer* normalizados usando [3.22].
d) Itere este procedimiento.

Para obtener un algoritmo monótonamente convergente, se usa una secuencia de operaciones similares a las usadas por Wold (1985) y Hanafi (2007) para PLS-PM. El procedimiento inicia con el vector-peso *outer* normalizados $\alpha_1^0\ , \dots, \alpha_J^0$.

Luego empieza la convergencia con $s = 0, 1, 2, \dots$. Recorriendo el ciclo de las $j = 1$ calculando la componente *inner* ξ_1^0 según el tipo de esquema y actualizando los pesos *outer* α_1^1 con el cual para $j = 2$ calculo ξ_2^0 de acuerdo con la partición dada en ξ_j^s , inmediatamente actualizo α_2^1 . ... El procedimiento es iterado hasta la convergencia del criterio límite, debido a la siguiente proposición:

Proposición 2: la siguiente función se deriva de los pesos *outer* $\alpha_1, \dots, \alpha_j$ generados por el algoritmo GCCA poblacional:

$$f(\alpha_1\ , \dots\ , \alpha_J) = \sum_{j\neq k}^{J} c_{jk} g\left[\rho\left(\alpha_j' x_j\ , \alpha_k' x_k\right)\right]$$

entonces, se tiene la siguiente desigualdad:

$$\forall s \quad f\left(\alpha_1^s\ , \dots, \alpha_J^s\right) \leq\ f\left(\alpha_1^{s+1}\ , \dots, \alpha_J^{s+1}\right)\ .$$

La característica esencial de este algoritmo es que cada reemplazamiento es optimal y secuencial, es decir, α_j^s debe ser reemplazado por α_j^{s+1} antes de reemplazamiento α_{j+1}^s . Este enfoque secuencial lleva a la convergencia monótona de este algoritmo.

3.4.3. Ecuaciones de estacionariedad a nivel muestral

Se tienen J bloques $X_1, \dots, X_J$ de variables centradas medidas sobre un conjunto de n individuos. Una columna *h*, X_{jh} , de X_j es considerada como una variable observada sobre los n individuos. $C = \{c_{jk}\}$ es una matriz de diseño tal que $c_{jk} = 1$ para bloques j, k conectados, $c_{jk} = 0$ en otro caso.

La covarianza muestral es $S_{jj} = \frac{1}{n} X_j' X_j$ y $S_{jk} = \frac{1}{n} X_j' X_k$. En caso de alta multicolinealidad o cuando $n << p_j$, S_{jj} estima pobremente a Σ_{jj} ; sin embargo, se propone un mejor estimador de la clase restringida $\hat{S}_{jj} = \tau_j I + (1 - \tau_j) S_{jj}$ con $0 \leq \tau_j \leq 1$.

Las versiones muestrales $\underbrace{a_j, Y_j}_{outer}$, $\underbrace{Z_j}_{inner}$ respectivamente de $\underbrace{\underbrace{a_j, \eta_j}_{pesos,comp}}_{Outer}$, $\underbrace{\xi_j}_{inner}$ permiten calcular los componentes *outer* $Y_j = X_j a_j$ y los componentes *inner* $Z_j = \sum_{k \neq j}^{J} c_{jk}\, w[Cov(Y_j, Y_K)]\, Y_K$.

La versión muestral de las ecuaciones de estacionariedad [3.22] con Σ_{jj} reeemplazada por $\hat{S}_{jj}$ es:

$$a_j = \left[z_j' X_j \left(\tau_j I + (1 - \tau_j) \frac{1}{n} X_j' X_j \right)^{-1} X_j' z_j \right]^{-1/2} \left[\tau_j I + (1 - \tau_j) \frac{1}{n} X_j' X_j \right]^{-1} X_j' Z_j \quad [3.23]$$

bajo la restricción

$$a_j' \left[\tau_j I + (1 - \tau_j) \frac{1}{n} X_j' X_j \right] a_j = 1 \,; \quad j = 1, \dots, J \,. \qquad [3.24]$$

De acuerdo con la terminología PLS, con $\tau_j = 0$ se tiene el "modo B" con $\tau_j = 1$ se tiene "nuevo modo A" y con $0 \leq \tau_j \leq 1$ se tiene el "modo Ridge".

- **Modo B** $(\boldsymbol{\tau_j = 0})$

  Aqui la restricción de normalización $var(\underbrace{X_j a_j}_{Y_j}) = 1$ y la ecuación de estacionariedad

$$a_j = n^{1/2}\left[z_j' X_j \left(X_j' X_j\right)^{-1} X_j' z_j\right]^{-1/2} \left(X_j' X_j\right)^{-1} X_j' z_j \ . \qquad [3.25]$$

Este vector de pesos *outer* a_j es proporcional al vector de coeficientes de la regresión múltiple de Z_j sobre X_j . Esta via de cómputo de los pesos *outer* es el *modo B* usual del enfoque PLS.

- **Nuevo Modo A** $(\boldsymbol{\tau_j = 1})$

La restricción de normalización es $\|a_j\| = 1$ y la ecuación de estacionariedad es:

$$a_j = X_j' z_j \ / \ \|X_j' z_j\| \ . \qquad [3.26]$$

Note que la componente *outer* $Y_j = X_j a_j$ es la primera componente en la regresión PLS de la componente *inner* Z_j sobre el bloque X_j .

Esta vía de computo de los pesos *outer* es llamado "Nuevo Modo A". En el *modo A* original del enfoque PLS, los pesos *outer* son computados en la misma via de la fórmula [3.26], pero escalados para que la componente *outer* $Y_j = X_j a_j$ quede estandarizada.

- **Modo Ridge** $(\boldsymbol{0 < \tau_j < 1})$

El vector de pesos *outer* a_j es proporcional al vector de los coeficientes en la regresión ridge de Z_j sobre X_j con una constante ridge igual a $n\tau_j/(1-\tau_j)$. Este modo permite un encogimiento gradual de la matriz de covarianza intrabloque hacia la identidad.

Tres interesantes propiedades de los componentes *outer* $Y_j = X_j a_j$ son establecidos en Qannari and Hanafi (2005):

a) $Var\left(X_j a_j / \|a_j\|\right)$ es una función creciente de τ_j .

b) $Cov\left(X_j a_j / \|a_j\|, Z_j\right)$ es una función creciente de τ_j .

c) $Cor\left(X_j a_j \ , Z_j\right)$ es una función decreciente de τ_j .

Así, si el usuario está favoreciendo la estabilidad (alta varianza) comparada con la correlación, $\tau_j = 1$ es la escogencia natural. Si el usuario desea dar prioridad a la correlación entre Y_j y sus componentes vecinas $\tau_j = 0$ es la mejor escogencia.

3.4.4. Regularized Generalized Canonical Correlation Analysis (RGCCA)

Se define el RGCCA como el siguiente problema de optimización:

$$\text{Maximizar}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_J} \sum_{j\neq k}^{J} c_{jk}\, g\big(Cov(X_j a_j, X_k a_k)\big) \qquad [3.27]$$

sujeto a las restricciones $\tau_j \|a_j\|^2 + (1-\tau_j) var(x_j a_j) = 1 \;,\quad j = 1,\ldots,J.$

Las ecuaciones de estacionariedad asociadas a este problema de optimización son exactamente las descritas en [3.23]. No hay solución analítica a este problema, sin embargo escribiendo el algoritmo PLS para GCCA a nivel muestral, se produce en esta dirección un algoritmo monótonamente convergente. La gráfica 3.7 corresponde al algoritmo PLS para RGCCA, donde al nivel muestral los pesos *outer* a_j estiman α_j y las componentes *inner* Z_j estiman ξ_j.

El algoritmo descrito en la gráfica 3.7 empieza escogiendo arbitrariamente los vectores $\tilde{a}_1^0, \ldots \tilde{a}_j^0$ que han de ser normalizados de acuerdo con el ítem A. Suponga que en la iteración *s* se ha calculado los vectores de pesos *outer* $a_1^{s+1}, a_2^{s+2}, \ldots, a_{j-1}^{s+1}$ (ver ítem C); entonces para el cálculo de a_j^{s+1} es necesario considerar la componente *inner* Z_j separando la suma para los $k < j$ conteniendo los a_k^{s+1} y los $k > j$ conteniendo los a_k^s , tal como se ve en el ítem B. El procedimiento es iterado hasta la convergencia del criterio límite, descrito en la siguiente proposición:

<u>Proposición 3</u>: Sean $a_j^s \;,\; j = 1,\ldots,J \;;\; s = 0,1,2,\ldots,$ una secuencia de vectores de peso *outer* generados por el algoritmo RGCCA. Se define la siguiente función:

$$h(a_1, a_2, \ldots, a_j) = \sum_{j\neq k}^{J} c_{jk}\, g\big[Cov(X_j a_j, X_k a_k)\big]$$

para vectores $a_1 \,,\; \ldots, a_j$ verificando los controles de normalización en [3.24].

Entonces, se tiene la siguiente desigualdad para todo s :

$$h\left(\alpha_1^{\,s}, \ldots, \alpha_J^{\,s}\right) \leq h\left(\alpha_1^{\,s+1}, \ldots, \alpha_J^{\,s+1}\right) .$$

A. Inicio

A1. Escoja *J* vectores arbitrarios $a_1^{\sim 0}, a_2^{\sim 0}, \ldots, a_J^{\sim 0}$.
A2. Compute los vectores de pesos *outer* normalizados $a_1^0, a_2^0, \ldots, a_J^0$ como

$$a_j^0 = \left[\left(\tilde{a}_j^0\right)^t \left[\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j\right]^{-1} \tilde{a}_j^0\right]^{-1/2} \left[\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j\right]^{-1} \tilde{a}_j^0 .$$

Para $s = 0, 1, \ldots$ (hasta la convergencia)
Para $j = 1, 2, \ldots, J$

B. Computando la componente inner X_j

Compute la componente *inner* de acuerdo con el esquema seleccionado:

$$z_j^s = \sum_{k<j} c_{jk} w\left[Cov\left(X_j a_j^s, X_k a_k^{s+1}\right)\right] X_k a_k^{s+1} + \sum_{k>j} c_{jk} w\left[Cov\left(X_j a_j^s, X_k a_k^s\right)\right] X_k a_k^s$$

donde $w(x) = 1$ para el esquema Horst, x para el esquema factorial y $signo(x)$ para el esquema centroide.

C. Computando el vector de pesos outer para el bloque X_j

Compute el vector de pesos *outer*

$$a_j^{s+1} = \left[(z_j^s)^t X_j \left[\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j\right]^{-1} X_j^t z_j^s\right]^{-1/2} \times \left[\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j\right]^{-1} X_j^t z_j^s$$

Fin
Fin

Gráfica 3.7. Algoritmo PLS para RGCCA

No hay prueba de que el algoritmo converja siempre a un óptimo global; Kramer (2007) muestra ejemplos de convergencia a óptimo local.

El gran poder y flexibilidad del RGCCA permite un gran espectro de métodos a ser recuperados, especialmente el caso multibloques de acuerdo con la tabla 3.2.

Método	*Criterio a maximizar*	*Esquema*	*Normalización* $j = 1, \dots, J$	*constantes restringidas* $j = 1, \dots, J$
SUMCOR	$\sum_{j,k,j\neq k} Cor(X_j a_j, X_k a_k)$	Horst	$Var(X_j a_j) = 1$	$\tau_j = 0$
SSQCOR	$\sum_{j,k,j\neq k} Cor^2(X_j a_j, X_k a_k)$	Factorial	$Var(X_j a_j) = 1$	$\tau_j = 0$
SABSCOR	$\sum_{j,k,j\neq k} \lvert Cor(X_j a_j, X_k a_k)\rvert$	Centroid	$Var(X_j a_j) = 1$	$\tau_j = 0$
SUMCOV	$\sum_{j,k,j\neq k} Cov(X_j a_j, X_k a_k)$	Horst	$\lVert a_j \rVert = 1$	$\tau_j = 1$
SSQCOV	$\sum_{j,k,j\neq k} Cov^2(X_j a_j, X_k a_k)$	Factorial	$\lVert a_j \rVert = 1$	$\tau_j = 1$
SABSCOV	$\sum_{j,k,j\neq k} \lvert Cov(X_j a_j, X_k a_k)\rvert$	Centroid	$\lVert a_j \rVert = 1$	$\tau_j = 1$

Tabla 3.2. Casos especiales del RGCCA para analisis de datos multibloques.

En el análisis de datos multibloques, todos los bloques X_j, $j = 1, \dots, J$ están conectados, y muchos criterios existen con el objetivo de encontrar las componentes $Y_j = X_j a_j$ con propiedades muy útiles. Aquí, sólo se consideran los métodos relacionados con el problema de optimización [3.27], ver tabla 3.2.

El criterio SUMCOR ha sido propuesto por Horst (1961) y el SSQCOR por kettenring (1971). El criterio SUMCOV es un caso especial propuesto por Van de Geer (1984) y el SSQCOV ha sido propuesto por Hanafi y Kiers (2006). El SABSCOR ha sido estudiado por Hanafi (2007) y el criterio SABSCOV por Krämer (2007). Estos dos últimos criterios no son comunes en el análisis de datos multibloques, y simplemente son un caso especial de RGCCA.

Desde el punto de vista de optimización del problema [3.27], los primeros tres criterios en la tabla 3.2, corresponden al modo B de acuerdo con $\tau_j = 0$ para todos los bloques y los últimos tres criterios al nuevo modo A con $\tau_j = 1$.

También se puede considerar la situación donde las constantes contraídas son $\tau_j = 0$ para algunos bloques y $\tau_j = 1$ para otros. Entonces, el problema de optimización [3.27] con $c_{jk} = 1$ para todo $j \neq k$, es

$$\underset{a_1,\ldots,a_j}{\text{Maximizar}} \sum_{j\neq k}^{J} g\left[Cor(X_j a_j, X_k a_k)\left(var(X_j a_j)\right)^{\tau_j/2}\left(var(X_k a_k)\right)^{\tau_j/2}\right]$$

sujeto a $\tau_j \|a_j\|^2 + (1-\tau_j)var(X_j a_j) = 1$ con $\tau_j = 0$ ó $1 \quad j = 1,\ldots,J$.

Si $\tau_j = 0$ (modo B), el principal objetivo en la construcción de los componentes *outer* $Y_j = X_j a_j$ es maximizar su correlación con los componentes más cercanos.

Al contrario si $\tau_j = 1$ (nuevo Modo A) el objetivo es construir una componente $Y_j = X_j a_j$ el cual explicará su propio bloque como prioridad y al mismo tiempo estar bien correlacionada con sus componentes más cercanas.

3.5. TRANSFORMACIÓN DE DATOS NO MÉTRICOS VÍA PLS

El tratamiento de datos mixtos se lleva a cabo a partir de los métodos GNM-PLS (General Non Metric Partial Least Squares), los cuales son una extensión de los métodos algorítmicos convergentes NM-PLS propuestos por Russolillo (2009).

Los métodos NM-PLS toman para la cuantificación la 1ª componente proporcionada por el mismo método de análisis que se desarrolle y una vez los parámetros del modelo converjan; simultáneamente se verifica que la función de maximización alcance convergencia monótona. Los métodos así obtenidos se denominan NM-NIPALS, NM-PLSR o NM-PLSPM respectivamente.

Análogamente, GNM-PLS toma para la cuantificación de cada variable cualitativa una función agregada con las primeras *k = 1,2, ...* componentes suministradas por el propio método de análisis según NIPALS, PLS-R o PLS-PM. Los métodos así obtenidos se denominan GNM-NIPALS, GNM-PLSR o GNM-PLSPM respectivamente.

Tanto NM-PLS como GNM-PLS usan el mismo proceso de cuantificación de las variables no métricas propuesto por Young (1981), proporcionando propiedades de pertenencia grupal y orden si existe; en este último caso se implementa la regresión monótona (Aluja 1994).

3.5.1. Proceso de cuantificación

En el proceso de transformación, cada *categoría* observada en la variable no métrica cruda $x*$ es remplazada por un valor numérico en escala de intervalo. La variable escalada $\hat{x}$ debe preservar las propiedades grupales y de orden si existe.

Respecto a la propiedad grupal, la variable escalada $\hat{x}$ debe ser restringida tal que si: $x_i^* \sim x_{i'}^* \Rightarrow \hat{x}_i = \hat{x}_{i'}$ donde ~ significa *pertenencia* de los individuos *i* e *i'* a la misma categoría.

Se define la función de cuantificación $q()$, Young (1981), como una función real aplicada a $x*$ la cual genera un valor numérico optimo $\hat{x}$ para cada observación. Bajo los métodos NM-PLS, la cuantificación de las *k'* categorías de $x*$ satisfaciendo la pertenencia grupal, corresponden con el vector generado por el proyector ortogonal de su matriz indicadora $\tilde{X}$ sobre el criterio latente (LC) o función agregada γ más cercana:

$$\tilde{q}(x^*, \gamma): \ \hat{x} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\gamma = P_{\tilde{X}}\gamma \ . \qquad [3.28]$$

El coeficiente de determinación de esta regresión equivale al cuadrado de la razón de correlación de Pearson $\eta_{\gamma|x^*}$ entre la variable categórica original y el LC, así,

$$\eta^2_{\gamma|x^*} = R^2_{(\gamma,\tilde{x}_1,\ldots,\tilde{x}_m)} = \frac{\gamma' P_{\tilde{X}}\gamma}{\gamma'\gamma} = r^2(\gamma, \hat{x}) \ .$$

La razón de correlación además de ser siempre positiva, puede ser expresada en términos de la correlación lineal de Pearson:

$$\eta_{\gamma|x^*} = r(\gamma, \hat{x}).$$

Como $R = \underset{a_1, \ldots a_{k'}}{sup} r\left(\gamma; \sum_{j=1}^{k'} a_j \tilde{x}_j\right)$, de la razón de correlación se sabe que ese *máximo* se tiene para $a_j = \bar{\gamma}_j$, con lo cual cada categoría *j* queda cuantificada con la media de los valores de γ asociados a la *j-ésima* categoría, y por tanto con este procedimiento se obtiene en cada cuantificación

$$max\{cor^2(\hat{x}, \gamma)\} \ .$$

3.5.2. Cuantificación con orden

Si la variable a ser cuantificada es ordinal, además de la pertenencia grupal debe añadirse la restricción de *orden (≺)*, con lo cual

$$x_i^* \sim x_{i'}^* \;\Rightarrow\; \hat{x}_i = \hat{x}_{i'} \quad \text{y} \quad x_i^* \prec x_{i'}^* \;\Rightarrow\; \hat{x}_i < \hat{x}_{i'} \;\; .$$

Para garantizar el orden se usa en vez de las indicadoras las matrices de orden $\bar{\bar{X}}$ donde para cada individuo se tiene una de las siguientes recodificaciones según la categoría de orden asumida ($a < b < \ldots < k'$):

a	1	0	…	0
b	1	1	…	0
⋮				
k'	1	1	…	1

Luego, mediante regresión monótona (Aluja 1994), de γ sobre $\bar{\bar{X}}$ se seleccionan las categorías (ordenadas) con coeficientes positivos, excepto con la categoría *a* que puede tener cualquier signo, para conformar la matriz de regresión $\tilde{\bar{X}}$.

La exclusión de categorías con coeficientes de regresión negativos, induce empates con las categorías contiguas tomando por tanto el mismo valor de cuantificación. Así, para cada x_q analizada a nivel de escala *ordinal*, la cuantificación está dada por:

$$\tilde{\bar{q}}\left(x_q^*, \gamma\right): \;\; \hat{x}_q \propto \tilde{\bar{X}}_q \left(\tilde{\bar{X}}_q' \tilde{\bar{X}}_q\right)^{-1} \tilde{\bar{X}}_q' \gamma \qquad [3.29]$$

CAPÍTULO 4

ACP PARA DATOS MIXTOS VIA PLS

Se desarrolla GNM-NIPALS (General Non Metric – Nonlinear estimation by Iterative Partial Least Squares), Aluja y Gonzalez (2014), para formar parte de los métodos NM-PLS, el cual permite cuantificar las variables cualitativas de una matriz de datos mixtos mediante una función lineal de *k* componentes principales, tipo reconstitución, maximizando la inercia en el plano *k*-dimensional asociado al ACP de la matriz así cuantificada. Es entonces una generalización del algoritmo NM-NIPALS que usa sólo la primera componente principal en la cuantificación de variables cualitativas. De la maximización y positividad de la razón de correlación entre cada variable cualitativa y la función de reconstitucion, se tiene que la inercia acumulada en el plano *k-dimensional* asociado a la función de cuantificación del mismo rango, es mayor o igual que la generada en planos de igual dimensión pero con funciones de cuantificación de diferente rango.

Con las k componentes principales asociadas a la matriz así cuantificada, se desarrolla el análisis de inercia saturada para evaluar si aún existe una dimensión $k^* < k$, a partir de la cual la inercia acumulada en los ejes de orden igual o superior ya está explicada, caso en el cual la función de cuantificación definitiva es de rango menor (k^*).

Muchas de las bases de datos creadas para implementar análisis estadísticos suelen estar conformadas por datos mixtos, esto es, contienen tanto variables cuantitativas como cualitativas. La mayoría de los análisis clásicos multivariantes Lebart et al. (2006), requieren en su desarrollo que las variables sean de tipo cuantitativo; el ACP es muy útil para estudiar especialmente en el primer plano factorial las relaciones entre individuos y variables de tipo cuantitativo (métricas), sin embargo el tratamiento de datos mixtos propuesto en este trabajo, requiere que las variables cualitativas sean cuantificadas óptimamente, para ser incluidas como parte activa del análisis factorial junto a las demás variables cuantitativas.

Se sabe que remplazar cada variable cualitativa por su correspondiente matriz indicadora y luego desarrollar un ACP conlleva problemas de comparación de pesos entre las variables numéricas y las indicadoras, afectación (disminución proporcional) de la inercia en los primeros factores debido a la ortogonalidad de las indicadoras e incremento innecesario de la dimensionalidad (matrices esparcidas) dificultando la capacidad de síntesis en el análisis.

Russolillo (2012) presenta el método NM-NIPALS (Non Metric – Nonlinear estimation by Iterative PArtial Least Squares) y desarrolla algorítmicamente el ACP en una matriz de datos mixtos que contiene n individuos y $p^* = p + q$ variables con diferentes escalas de medida; q de ellas cualitativas.

Con el método NM-NIPALS se cuantifica bajo un criterio de optimización cada variable cualitativa, conservando las propiedades de pertenencia y orden (si existe) implícitas en las categorías correspondientes.

El NM-NIPALS aprovecha la flexibilidad del algoritmo NIPALS, Wold (1975), para en una primera fase del proceso cuantificar las variables cualitativas a partir de la primera componente principal t_1 que se obtiene iterando hasta la convergencia.

GNM-NIPALS presenta una generalización de NM-NIPALS, ya que implementa la cuantificación a partir de una función lineal $f(t_h)$ de h componentes principales vía reconstitución de la $q_{ésima}$ variable como en ACP, Aluja y Morineau (1999), es decir de la forma $f(t_h) = p_{1q}t_1 + p_{2q}t_2 + \cdots + p_{hq}t_h$, donde p_{hq} es la q-ésima coordenada del vector propio P_h. Las h componentes principales $t_1, t_2, \ldots, t_h$ que sirven de inicio en el algoritmo y que son proporcionadas por la matriz de datos cuantitativos Xp, indican la dimensión de la función de reconstitución.

El criterio de optimización asociado a la cuantificación se deriva del hecho de que la razón de correlación $\eta_{Y|X_q}$ es máxima y positiva, Saporta (2011), conllevando la generación de máxima inercia en el plano h-dimensional; con lo cual para $h = 1$, GNM-NIPALS es equivalente a NM-NIPALS y se tendrá máxima inercia en el primer eje factorial; ninguna otra cuantificación presenta mayor inercia en el primer eje. Para $h = 2$ se tendrá máxima inercia en el primer plano factorial, y ninguna otra cuantificación presenta mayor inercia en R^2; y así sucesivamente.

La matriz cuantificada presenta de hecho una estructura inercial decreciente eje por eje de acuerdo con la descomposición espectral; sin embargo, de las propiedades de la razón de correlación, cada función de cuantificación $f(t_k)$ hace que la inercia generada en el plano de igual dimensión (k) sea maximal, tal que la inercia de cualquier otro plano k-dimensional derivado de otra función $f(t_h)$ es inferior para todo $h\#k$.

La dimensión "ideal" de la función de cuantificación se puede determinar aplicando la regla de Cattell sobre la gráfica de inercia maximal acumulada eje por eje, identificando el punto k "*optimal*" a partir del cual la información de los ejes restantes no es relevante, y la función de cuantificación seria, $f(t_k) = p_{1q}t_1 + p_{2q}t_2 + \cdots + p_{kq}t_k$.

Con la matriz k cuantificada, se desarrolla el análisis de *Saturación de Inercia Explicada (SIE)* para evaluar si las primeras $k^* < k$ de las componentes finales involucradas en la cuantificación ya contienen la inercia explicada en los planos de dimensión $k^*, \ldots, k, \ldots, p^*$. Si es así, entonces la función de cuantificación definitiva es $f(t_{k^*}) = p_{1q}t_1 + p_{2q}t_2 + \cdots + p_{k^*q}t_{k^*}$ asociada a la q-ésima variable categórica.

Estas propiedades y algunas otras características también serán estudiadas aprovechando la ortogonalidad de las componentes principales $t_1, \ldots, t_k$ consideradas en $f(t_k)$. La aplicación se desarrolla tomando como base de datos el grupo g*ustacion* del ejemplo vinos, ver Escofier y Pagès (1992). El software utilizado es del entorno R, y las principales funciones desarrolladas proveen los resultados presentados.

Ya que la base fundamental de este trabajo reside en el método NIPALS, el capítulo dos iniciará explicando la conceptualización de este procedimiento y luego se expondrá lo relacionado con el NM-NIPALS; al final del capítulo se presenta el procedimiento algorítmico objeto de este articulo denominado GNM-NIPALS y fundamentado en los métodos NM-PLS. El capítulo tres contiene la interpretación y resultados del ejemplo de aplicación de GNM-NIPALS y finalmente en el capítulo cuatro se dan las conclusiones evidenciando la ganancia de inercia frente al NM-NIPALS.

4.1. EL ALGORITMO BASE: NIPALS

En la sección 3.1 se estudió el algoritmo NIPALS, el cual servirá de base fundamental para la construcción de los algoritmos NM-NIPALS y por consguiente de GNM-NIPALS.

Inicia tomando de la matriz original X_0 la primera *componente principal* t_1 como la primera columna, X_{01} ; en realidad la inicialización de t_1 puede ser el promedio de las variables o cualquier otra función lineal de las mismas.

Luego se deflacta la matriz original mediante $X_1 = X_0 - t_1 P_1'$ para garantizar la ortogonalidad de las siguientes componentes, e inicia nuevamente el proceso de iteración con $h = 2, 3, \ldots, a$. Para matrices con datos completos, el pseudo-algoritmo asociado a NIPALS es de la forma:

Etapa 1. $X_o = X_h$

Etapa 2. $h = 1,2, \ldots, a$:

Etapa 2.1. t_h = 1ª columna de X_{h-1} [prop $\bar{X}$]

Etapa 2.2. Repetir hasta la convergencia de P_h

Etapa 2.2.1. $P_h = \frac{X_{h-1}' t_h}{t_h' t_h}$ $\left[u = \frac{X'Xu}{\lambda}\ ,\ \lambda = \frac{1}{n-1} t't\right]$

Etapa 2.2.2. Normar p_h a 1

Etapa 2.2.3. $t_h = X_{h-1}P_h / P'_h P_h$ [$t = Xu$]

Etapa 2.3. $X_h = X_{h-1} - t_h P'_h$ [garantiza la ortogonalidad]

Siguiente *h.*

Fin

NIPALS entrega las *componentes* t_h y los *vectores propios* P_h correspondientes a la matriz X excepto tal vez por signo, tal como si se hubiese aplicado la función svd(X) de R.

4.2. NM-NIPALS

Bajo los métodos NM-PLS, la cuantificación de las k' categorías de x^* satisfaciendo la pertenencia grupal, corresponden con el vector generado por el proyector ortogonal de su matriz indicadora $\tilde{X}$ sobre el criterio latente (LC) γ o t más cercano:

$$\tilde{q}(x^*, \gamma): \ \hat{x} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\gamma = P_{\tilde{X}}\gamma \ . \qquad [4.1]$$

El coeficiente de determinación de esta regresión equivale al cuadrado de la razón de correlación de Pearson (*positiva*) entre la variable categórica original y el LC, y por tanto puede ser expresada en términos de la correlación lineal de Pearson:

$$cor(\gamma, \hat{x}) = \eta_{\gamma|x^*} \text{ además } \quad \eta^2_{\gamma|x^*} = R^2_{(\gamma, \tilde{x}_1, \ldots, \tilde{x}_k)} = \gamma' P_{\tilde{X}} \gamma / \gamma'\gamma \ .$$

Como $R = \underset{a_1, \ldots a_{k'}}{sup} r\left(\gamma; \sum_{j=1}^{k'} a_j \tilde{x}_j\right)$, de la razón de correlación se sabe que ese *máximo* se tiene para $a_j = \bar{\gamma}_j$, con lo cual cada categoría *j* queda cuantificada con la media de los valores de γ asociados a la *j-ésima* categoría. Por tanto con NM-NIPALS se obtiene en cada cuantificación con la primera componente $\gamma = t_1$

$$max\{cor^2(\hat{x}, t_1)\} . \qquad [4.2]$$

Es decir, la variable así cuantificada no solo conserva la propiedad derivada del ACP, $\sum_j^p cor^2(x_j, t_h) + \sum_r^q cor^2(\hat{x}_r, t_h) = \lambda_h$, sino que le imprime de acuerdo con [4.2] para las *q* variables cualitativas la característica maximal.

El pseudo algoritmo de NM-NIPALS es:

Entrada X^*

Salida $P_h : [p_1, \ldots, p_h]$; $T_h : [t_1 \ldots . t_h]$; $\hat{X}$.

1. *Inicializa t1*
2. *Repetir hasta convergencia de* p_h.

   $\hat{x}_q = q_\ell(x_q^*, t_1)$ # Cuantificación mediante ecuación [4.1]

   $\hat{X} = [x_1 \ldots \hat{x}_q]$

   $p_1 = \hat{X}' t_1 / (t_1' t_1)$

   $p_1 = p_1 / \|p_1\|$

   $t_1 = \hat{X} p_1$
3. $E_1 = \hat{X} - t_1 p_1'$
4. Para $h = 2, \ldots, p *$

   Inicializa t_h
5. *Repita*

   $p_h = E_{h-1}' t_h / (t_h' t_h)$

   $p_h = p_h / \|p_h\|$

   $t_h = E_{h-1} p_h / (p_h' p_h)$

   $E_h = E_{h-1} - t_h p_h'$
6. *Fin*

Note que en la fase 2, se cuantifica con el t_1 inicial, luego se calcula p_1 el cual permite a su vez recalcular t_1 iterando así hasta la convergencia de p_1.

En el paso 3 se deflacta, y a partir de la etapa 4, se obtienen las demás componentes $t_1, \dots, t_{p^*}$ si $\hat{X}$ es de rango completo p^*.

Para garantizar el orden se usa en vez de las indicadoras las matrices de orden $\tilde{\tilde{X}}$ descrita en el Proceso de Cuantificación (ítem 3.4.5). Luego, mediante regresión monótona de γ sobre $\tilde{\tilde{X}}$ se seleccionan las columnas (categorías-ordenadas) con coeficientes positivos, excepto con la categoría *a* que puede tener cualquier signo, para conformar la matriz de regresión $\hat{\tilde{X}}$. La exclusión de categorías con coeficientes de regresión negativos, induce empates con las categorías contiguas tomando por tanto el mismo valor de cuantificación.

Así, para cada x_q analizada a nivel de escala *ordinal*, la cuantificación está dada por:

$$\tilde{\tilde{q}}(x_q^*, t_1): \ \hat{x}_q \propto \tilde{\tilde{X}}_q(\tilde{\tilde{X}}_q'\tilde{\tilde{X}}_q)^{-1}\tilde{\tilde{X}}_q' t_1 \ . \qquad [4.3]$$

Ahora del ACP, puesto que $cor(\hat{x}_q, t_1) = \sqrt{\lambda_1}.p_{1q}$, cuando $p_{1q} > 0$ una regresión monótona creciente es implementada; y si $p_{1q} < 0$ la regresión monótona es decreciente. Cuando $p_{1q} \approx 0$ la relación es no monótona y la variable a cuantificar x_q^* en general no contendrá orden.

4.3. GNM-NIPALS

Si la matriz de datos para el análisis está constituida por múltiples dimensiones subyacentes significativas, es mucho más adecuado el uso de GNM-NIPALS que NM-NIPALS el cual se identifica más con sistemas de información unidimensionales que asocian factor tamaño.

GNM-NIPALS es también un método algorítmico, que busca bajo un criterio de optimización cuantificar cada una de las variables cualitativas de una matriz de datos mixtos mediante una función lineal de *h* componentes, esto es:

$$f(t_h) = p_{1q}t_1 + p_{2q}t_2 + \cdots + p_{hq}t_h \ . \qquad [4.4]$$

La función $\gamma = f(t_h)$ es una aplicación directa del concepto de *reconstitución* de una variable q derivada del ACP (ver apartado 2.2.4).

Los pesos p_{hq} corresponden a la q-ésima coordenada del *vector propio* P_h asociado a la componente t_h del mismo rango. Por tanto, cada p_{hq} se puede obtener como la correlación de la variable q cuantificada con el eje *h,* excepto por $\sqrt{\lambda_h}$, y en este caso equivale a la razón de correlación $\eta_{\gamma|x_q^*} = cor(\hat{x}_q, \gamma)$, la cual es máxima y positiva. Para $\gamma = p_{hq}t_h$ ésta correlación se puede calcular también bajo t_h ya que

$$\hat{x}_{q\gamma} = P_{\tilde{x}_q}\gamma = p_{hq}P_{\tilde{x}_q}t_h = p_{hq}\hat{x}_{qt} \quad \text{por tanto ,}$$

$$cor(\hat{x}_{q\gamma}, \gamma) = cor(p_{hq}\hat{x}_{qt}\,, p_{hq}t_h) = cor(\hat{x}_{qt}\,, t_h) \;\; \forall\, x_q \;. \qquad [4.5]$$

Observe que la correlación asociada a la cuantificación $\hat{x}_{q\gamma}$ con γ es equivalente a la correlación cuantificando $(\hat{x}_{qt})$ con el t asociado. Sin embargo la $cor(\hat{x}_{q\gamma}, t_h)$ toma el mismo valor excepto por el signo de p_{hq} .

El método inicia tomando *h componentes* asociadas a la matriz Xp (de rango completo) que contiene las p variables cuantitativas, lo cual garantiza la ortogonalidad al comenzar y una rápida convergencia. Con cada componente t_h de [4.4] se realiza la cuantificación $\hat{x}_q$ como en [4.1] y se obtiene la correlación [4.5], la cual permite estimar el p_{hq} correspondiente, que junto a las correlaciones de las variables cuantitativas con la misma componente, conducen a la conformación del *vector propio* P_h (normalizado para recuperar $\sqrt{\lambda_h}$) de dimensión p^* .

Se toman así, las q-ésimas coordenadas de los vectores propios P_h permitiendo formular la función de inicio $f(t_h) = p_{1q}t_1 + p_{2q}t_2 + \cdots + p_{hq}t_h$ con la cual se comienza el proceso GNM-NIPALS para cuantificar la q-ésima variable cualitativa, iterando hasta la convergencia de los t_h y P_h; note que $f(t_h)$ es un vector agregado de h componentes.

Se debe distinguir las cargas (pesos) y componentes finales (alcanzadas en la convergencia) asociadas a cada matriz cuantificada por su correspondiente $f(t_h)$. Así, si se cuantifica con $f(t_1)$ se tiene la matriz XC_1 de cuya descomposición singular se obtienen los vectores propios $P_1^a, P_2^a, \ldots, P_{p^*}^a$ y las componentes principales $t_1^a, t_2^a, \ldots, t_{p^*}^a$; el superíndice a indica que la cuantificación se realizó sólo con la primera componente principal t_1 de Xp.

Si se cuantifica con las dos primeras componentes $f(t_2)$, de la matriz así cuantificada XC_2 se obtienen los vectores propios $P_1^b, P_2^b, \dots, P_{p^*}^b$ y las componentes principales $t_1^b, t_2^b, \dots, t_{p^*}^b$ donde el superíndice b indica que la cuantificación se realizó con $(h = 2)$ las dos primeras componentes.

Observe entonces que $t_1^a \# t_1^b$, $t_2^a \# t_2^b, \dots$; las componentes del mismo orden asociadas a una y otra matriz son diferentes. En general estas diferencias se presentan para cada cuantificación realizada según la dimensión $h = 1,2, \dots p^*$.

De acuerdo con [4.1] haciendo $\gamma = f(t_h)$, al cuantificar sin orden simultáneamente cada variable cualitativa con h componentes, la $cor^2(\hat{x}, f(t_h))$ sigue siendo *máxima* y crece hasta la unidad de acuerdo con $f(t_{p^*})$, caso en el que se tiene rango completo (p^*) en la matriz cuantificada *XC* conteniendo las variables cuantitativas y las variables cualitativas cuantificadas.

Esta correlación maximal puede ser un índice de la dimensionalidad de $f(t_h)$, ya que valores relativamente grandes, por ejemplo $cor^2(\hat{x}, f(t_h)) > 0.90$ indican que h componentes serán suficientes para la cuantificación vía reconstitución; observe que con $h = p*$ entonces $cor^2\left(\hat{x}, f\left(t_{p^*}\right)\right) = 1$, lo cual es coherente con el hecho de que $\sum_{h=1}^{p^*} r^2_{(X_j, t_h)} = 1$.

La propiedad de máxima correlación expuesta anteriormente, conlleva a que fundamentalmente con GNM-NIPALS se consigue máxima inercia en el plano h-dimensional derivado de la matriz cuantificada con h componentes: $f(t_h) = p_{1q}t_1 + p_{2q}t_2 + \cdots + p_{hq}t_h$, $h \leq p$. Así, para $h = 1$, $f(t_1) = p_{1q}t_1$ induce máxima inercia proyectada en el primer eje y el proceso coincide con NM-NIPALS.

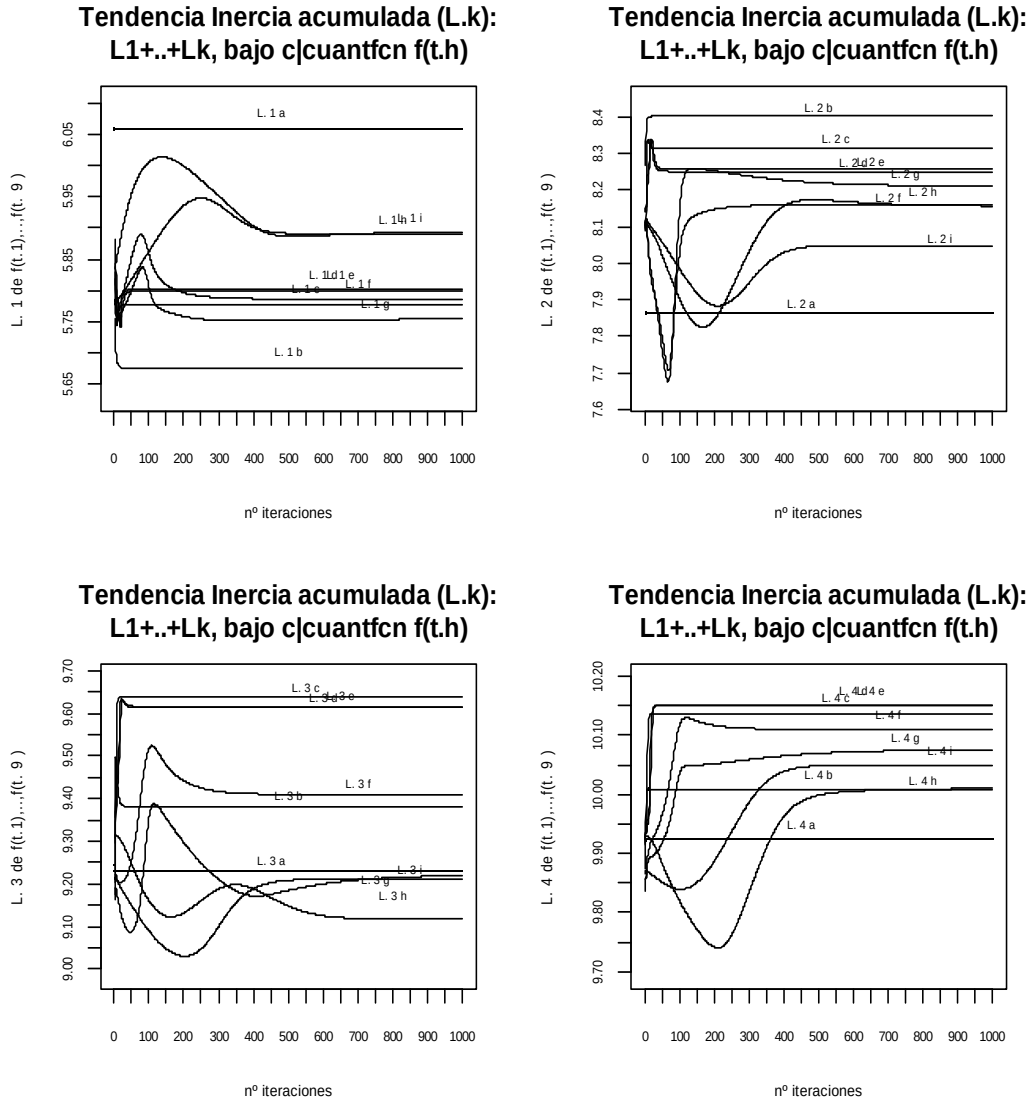
Cuantificando con $h = 2$ y por tanto con $f(t_2) = p_{1q}t_1 + p_{2q}t_2$, se consigue máxima inercia en el plano de los dos primeros ejes, tal que con ninguna otra cuantificación se consigue inercia superior en el primer plano factorial, esto es, se tiene que $\lambda_1^a + \lambda_2^a \leq \lambda_1^b + \lambda_2^b$; pero además, $\lambda_1^a \geq \lambda_1^b$.

Si $h = 3$ también se tiene que $\lambda_1^a \geq \lambda_1^c$ y que $\lambda_1^b + \lambda_2^b \geq \lambda_1^c + \lambda_2^c$. Así mismo,

$$\lambda_1^c+\lambda_2^c+\lambda_3^c \geq \begin{cases} \lambda_1^a + \lambda_2^a+\lambda_3^a \\ \lambda_1^b+\lambda_2^b+\lambda_3^b \end{cases}$$

Se muestra algorítmicamente que estos resultados se extienden a $h = 1,2, \ldots p^*$, y se pueden apreciar gráficamente en la convergencia de las inercias generadas en los planos de igual dimensión asociados a las funciones de cuantificación de orden *a, b, c, …*

De la ortogonalidad de las componentes finales y del concepto de inercia maximal descrito anteriormente, se presentan dos propiedades denominadas: Inercia Maximal Intra y Saturación de Inercia Explicada.



Gráfica 4.1. Tendencia de las primeras cuatro inercias acumuladas $\lambda_1, \ldots, \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$ derivadas de las matrices cuantificadas mediante $f(t_1), \ldots, f(t_9)$. *Datos Vinos : Gustacion.*

Propiedad 1: Inercia maximal intra. La inercia explicada I_k^k en el plano k dimensional derivada de la función de cuantificación del mismo orden $f(t_k) = p_{1q}t_1 + p_{2q}t_2 + \cdots + p_{kq}t_k$, es mayor o igual que la explicada en planos de igual dimensión k, asociados a funciones de cuantificación con diferente número de componentes finales, es decir

$$I_k^k \geq I_{k^*}^k \quad , \quad \text{para } k^* \# k.$$

Esta propiedad de inercia maximal I_k^k inducida por la función de cuantificación del mismo orden, se puede apreciar en las gráficas de tendencias de las inercias en el primer eje, en plano de los dos primeros ejes, en R^3, … , etc, generados por las funciones de cuantificación de todos los distintos órdenes.

Observe en la gráfica 4.1 que la inercia asociada al plano R^k de igual dimensión que la función de cuantificación $f(t_k)$, es máxima, es decir las curvas inerciales L_{1a}, L_{2b}, L_{3c} y L_{4d} son maximales.

Propiedad 2: Inercia Saturada. Se denomina SIE al caso en el que alguna de las matrices cuantificadas con las primeras $k^* < k$ de las componentes finales ya contiene la inercia acumulada explicada en los ejes $k^*, \ldots, k, k+1, k+2$ de la matriz XC_k.

La presencia de inercia saturada en el análisis, conlleva la disminución del orden k de dimensionalidad de la función de cuantificación generalmente a $k-1$; lo cual nos permite afinar la dimensión asociada a $f(t)$.

$f(t_{1.})$	t_1	t_2	…	t_{p+q}
X_1	r_{11}	r_{12}		r_{1p*}
X_2	r_{21}	r_{22}		r_{2p*}
⋮				
q_1^a	$r_{q_1 1}^a$	$r_{q_1 2}^a$		
$\sum r_{jk}^2$	λ_1^a	λ_2^a	…	λ_{p*}^a

$f(t_{12.})$	t_1	t_2	…	t_{p+q}
X_1	r_{11}	r_{12}		r_{1p*}
X_2	r_{21}	r_{22}		r_{2p*}
⋮				
q_1^b	$r_{q_1 1}^b$	$r_{q_1 2}^b$		
$\sum r_{jk}^2$	λ_1^b	λ_2^b	…	λ_{p*}^b

Tabla 4.1. Correlación de las variables e inercias con los ejes, derivados de funciones de cuantificación $f(t_{1.}) = t_{1.}$ y $f(t_{12.}) = t_{1.} + t_{2.}$ denotadas con superíndices [a] y [b] respectivamente.

Ya que la inercia asociada a la componente final h bajo la cuantificación de orden k se obtiene mediante $\sum_j^{p^*} r^2_{(X_j,t_h)} = \lambda_h^k$, y que estas correlaciones son invariantes para las variables cuantitativas, sólo es necesario analizar las correlaciones de las variables cualitativas recuantificadas para obtener dichas inercias; ver tabla 4.1.

Se omite en esta demostración el superíndice k sólo por comodidad, pero no se debe olvidar que las componentes finales usadas para la recuantificación provienen de funciones de dimensión k.

La función de recuantificación con tres de las componentes finales es:

$$t_{123.} = t_{1.} + t_{2.} + t_{3.} = t_{12.} + t_{3.} \qquad [4.6]$$

y su cuantificación asociada es :

$$q_{123.} = PI_q * t_{123.} = PI_q t_{1.} + PI_q t_{2.} + PI_q t_{3.}$$

$$q_{123.} = q_{1.} + q_{2.} + q_{3.} = q_{12.} + q_{3.}$$

De la ortogonalidad de las componentes se tiene de [4.6] que las correlaciones al cuadrado son:

$$r^2(q_{123.}, t_{123.}) = r^2(q_{123.}, t_{1.}) + r^2(q_{123.}, t_{2.}) + r^2(q_{123.}, t_{3.}) = I_3^3 \qquad [4.7]$$

Análogamente, de las recuantificaciones con $k^* = 2$, y $k^* = 1$ se tiene respectivamente:

$$r^2(q_{12.}, t_{12.}) = r^2(q_{12.}, t_{1.}) + r^2(q_{12.}, t_{2.}) = I_2^2 \qquad [4.8]$$

$$r^2(q_{1.}, t_{1.}) = I_1^1 \ . \qquad [4.9]$$

Las expresiones [4.7], [4.8] y [4.9] son maximales (en sus respectivas dimensiones), debido a que la razón de correlación $\eta^2_{Y|X} = r^2$ (correlación lineal) es máxima al aplicar el proyector ortogonal de las indicatrices PI_q a $t_{12\ldots k^*}$ (Saporta 2011).

El extremo derecho de la igualdad en las ecuaciones [4.10], [4.11] y [4.12] demuestran la "inercia maximal intra" en los planos de igual dimensión.

En el plano $k = 1$,

$$r^2(q_{1.}, t_{1.}) \geq \begin{cases} r^2(q_{12.}, t_{1.}) = r^2(q_{12.}, q_{1.}) r^2(q_{1.}, t_{1.}) = I_2^1 \\ r^2(q_{123.}, t_{1.}) = r^2(q_{123.}, q_{1.}) r^2(q_{1.}, t_{1.}) = I_3^1 \end{cases} \qquad [4.10]$$

con lo cual $I_1^1 \geq \begin{cases} I_2^1 \\ I_3^1 \end{cases}$.

En el plano de dimensión dos,

$$r^2(q_{12.}, t_{12.}) \geq \begin{cases} r^2(q_{1.}, t_{12.}) = r^2(q_{1.}, q_{12.}) r^2(q_{12.}, t_{12.}) = I_1^2 \\ r^2(q_{123.}, t_{12.}) = r^2(q_{123.}, q_{12.}) r^2(q_{12.}, t_{12.}) = I_3^2 \end{cases} \qquad [4.11]$$

entonces $I_2^2 \geq \begin{cases} I_1^2 \\ I_3^2 \end{cases}$

Análogamente, ya que la expresión [4.7] es maximal

$$r^2(q_{123.}, t_{123.}) \geq \begin{cases} r^2(q_{1.}, t_{123.}) = r^2(q_{1.}, q_{123.}) r^2(q_{123.}, t_{123.}) = I_1^3 \\ r^2(q_{12.}, t_{123.}) = r^2(q_{12.}, q_{123.}) r^2(q_{123.}, t_{123.}) = I_2^3 \end{cases} \qquad [4.12]$$

por tanto, $I_3^3 \geq \begin{cases} I_1^3 \\ I_2^3 \end{cases}$.

La generalización de estos resultados al caso en el cual $k > 3$ es evidente e inmediato. De las expresiones [4.10], [4.11] y [4.12] se concluye que las inercias en los planos k-dimensionales son máximas cuando son generadas por funciones de cuantificación de igual dimensión.

El pseudo-algoritmo asociado al procedimiento GNM-NIPALS se presenta a continuación:

Entrada Xp , Salida XC, *T, P*

1. *Inicializa* $T = (t_1, t_2, \dots, t_H)$ [*H componentes en Xp via NIPALS*]

$para\ h = 1,2, \dots, H$
$p_{hp} = r(X_p, t_h) \quad \forall\ p$ [*correlación p-v.cuantitativas con* t_h]
$p_{hq} = r(\hat{X}_q, t_h)$; $\hat{X}_q = P_{\tilde{\mathbf{X}}_q}.t_h \quad \forall\ q$ [*razón correlación v.cualit*]

$P_h \leftarrow (p_{h1}, \dots, p_{hp}, p_{hq_1} \dots, p_{hq})$, *normar* P_h

$P[, h] \leftarrow P_h$

fin h

2. *Repetir*

2.1 $para\ q = 1,2, \dots, Q$

$f(t_H) = p_{1q}t_1 + p_{2q}t_2 + \cdots + p_{Hq}t_H = \gamma$ [*función cuantificn*]

$\hat{X}_q = P_{\tilde{X}_q} \cdot \gamma$ [*cuantificación estandarizada*]

$XC[, p + q] \leftarrow \hat{X}_q$

fin q

2.2 $para\ h = 1,2, \dots, H$ [*actualizar P, T: función NIPALS*]

$P_h = XC' \cdot T[, h]$, *normar* P_h

$t_h = XC \cdot P_h$

$P[, h] \leftarrow P_h$

$T[, h] \leftarrow t_h$

$XC_h \leftarrow XC$, $XC \leftarrow XC_h - t_h \cdot P_h'$ [*deflactar*]

$fin\ h$

hasta convergencia de P_h

Fin

La fase 1 inicia obteniendo *H* componentes principales de X_k via NIPALS, luego, básicamente se constituyen las coordenadas p_{hq} de los vectores propios P_h mediante la razón de correlación, las cuales permiten formular la función $f(t_H)$ para la cuantificación de variables cualitativas (ver fase *2.1*).

En el caso que se requiera cuantificación con orden se utiliza la ecuación [4.3]. En la fase 2.2 se recalcula las matrices de vectores propios *P* y de componentes *T*, y se iteran están dos últimas fases hasta la convergencia de los P_h.

4.3.1. Aplicación

La base de datos cuantitativa (vinos) utilizada como ejemplo de aplicación se encuentra descrita por Escofier y Pagès (1992), fué complementada con las variables cualitativas denominación de origen *Appe* y tipo de suelo *Terr* que contienen la siguiente codificación sin considerar orden:

Appe : (1,1,2,3,1,2,2,1,3,1,2,1,1,1,1,2,3,2,3,1,1)

Terr : (2,2,2,3,1,1,1,2,2,1,2,3,3,3,1,1,1,2,3,4,4)

Appe contiene tres categorías con los siguientes significados:
1=*saumur*, 2=*bourgueil* y 3=*chinon*

Terr contiene las siguientes categorías:
1=*medio1* (*referencia*), 2=*medio2*, 3=*medio3* y 4=*medio4*

Se conforma la base de datos mixta denominada *vinos.k* del tipo *k-tablas,* que nos permite tomar el grupo *gustación* para el análisis conteniendo las variables cuantitativas y la base *denom.f* conteniendo los factores cualitativos.

Por comodidad en casi todo el desarrollo del procedimiento GNM- NIPALS bajo R se manejan los dos tipos de datos de forma separada; hasta conformar la matriz cuantificada XC_k ; así,

Xp : *gustacion* # contiene los datos cuantitativos (*n=21, p=9*)
Xq : *denom.f* # tabla con los factores *Appe* y *Terr*

De acuerdo con la ecuación [4.11], el proceso comienza con la descomposición singular vía NIPALS de la matriz Xp que presenta rango 9 y proporciona las componentes que han de conformar las funciones de cuantificación $f(t_1)$, $f(t_2)$, …, $f(t_9)$ con las cuales se obtiene las matrices cuantificadas con 1, 2, …, y 9 componentes respectivamente.

Los resultados inerciales maximales asociados a los planos de dimensión k en cada matriz k-cuantificada son presentados (resaltados) en la tabla 4.2, y con ellos se obtiene la gráfica 4.2 de inercia maximal acumulada de la cual se determina la dimensionalidad "optimal" de la función de cuantificación principal.

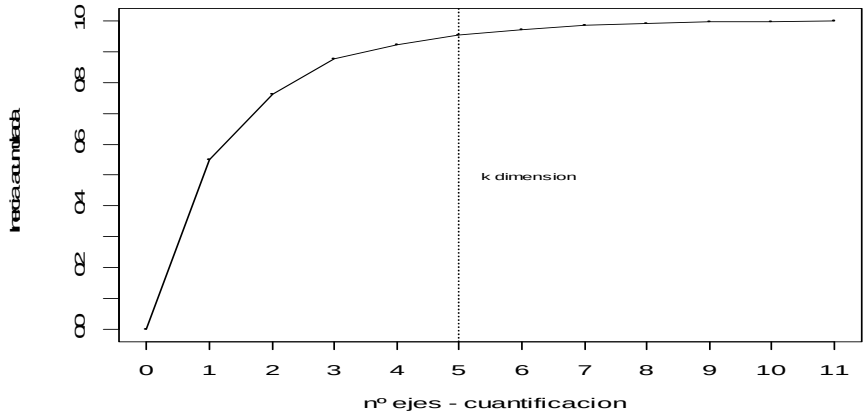
4.3.2. Resultados – datos *gustacion*

Loa valores propios optimales asociados a las cuantificaciones con $t_{1.}, t_{2.}, \ldots, t_{9.}$ corresponden a la columna denotada como inertia, mientras que ratio es la inercia porcentual acumulada eje por eje.

	Cuantificación		$f(t_{1.})$	Cuantificación		$f(t2_{.})$	Cuantificación		$f(t3_{.})$
k	inertia	cum	ratio	inertia	cum	ratio	inertia	cum	ratio
1	6.05913	6.059	0.5508	5.67415	5.674	0.5158	5.77589	5.776	0.5251
2	1.80870	7.868	0.7153	2.72696	8.401	0.7637	2.53983	8.316	0.7560
3	1.36104	9.229	0.8390	0.98033	9.381	0.8529	1.32232	9.638	0.8762
4	0.69502	9.924	0.9022	0.62546	10.007	0.9097	0.49907	10.137	0.9216
5	0.37148	10.295	0.9359	0.35013	10.357	0.9415	0.35206	10.489	0.9536
:	:	:		:	:		:	:	
11	0.01737	11.000	1.0000	0.01933	11.000	1.0000	0.01726	11.000	1.0000

	Cuantificación		$f(t4_{.})$	Cuantificación		$f(t5_{.})$	Cuantificación		$f(t9_{.})$
k	inertia	cum	ratio	inertia	cum	ratio	inertia	cum	ratio
1	5.80041	5.800	0.5273	5.79805	5.798	0.5271	5.95304	5.953	0.5412
2	2.45020	8.251	0.7501	2.45755	8.256	0.7505	1.93660	7.890	0.7172
3	1.36351	9.614	0.8740	1.36062	9.616	0.8742	1.25477	9.144	0.8313
4	0.53656	10.151	0.9228	0.53441	10.151	0.9228	0.83709	9.982	0.9074
5	0.35048	10.501	0.9547	0.35056	10.501	0.9547	0.36754	10.349	0.9408
6	0.20759	10.709	0.9735	0.20646	10.708	0.9734	0.30526	10.654	0.9686
7	0.12476	10.833	0.9849	0.12523	10.833	0.9848	0.15272	10.807	0.9825
8	0.06598	10.899	0.9909	0.06611	10.899	0.9908	0.10836	10.915	0.9923
9	0.05558	10.955	0.9959	0.05577	10.955	0.9959	0.05378	10.969	0.9972
10	0.02902	10.984	0.9986	0.02923	10.984	0.9985	0.01863	10.988	0.9989
11	0.01592	11.000	1.0000	0.01600	11.000	1.0000	0.01222	11.000	1.0000

Tabla 4.2. Máxima Inercia acumulada (0.5508, 0.7637,0.8762, …) de las matrices cuantificadas con $f(t_{1.}) = t_{1.}$, $f(t_2) = t_{1.} + t_{2.}$, $f(t_3) = t_{1.} + t_{2.} + t_{3.}$ ... respectivamente.



Gráfica 4.2. Inercia acumulada en los planos de dimension 1, 2, 3, …,9
base *gustacion*

De la gráfica 4.2 de distribución de inercia maximal acumulada se deduce bajo la regla de Cattell, que la función de cuantificación seleccionada es de dimensión 5, es como el punto de inflexión después del cual el aporte de inercia de cada uno de los ejes restantes no es relevante; por tanto $f(t_5) = t_{1.} + t_{2.} + t_{3.} + t_{4.} + t_{5.}$.

Los valores propios asociados a la matriz XC_5 cuantificada bajo $f(t_5)$ se muestran en la tabla 4.3. Observe que hasta el eje 5 se recoge el 95.47% de la inercia proyectada, y que ésta es máxima respecto a los otros planos de igual dimensión (cinco) de acuerdo con la generalización de la ecuación [4.12].

k	1	… 4	5	6	7	8	9	10	11
inert.	5.7985	0.5347	0.3505	0.2066	0.1251	0.0660	0.0557	0.0291	0.0159
cum	5.799	10.151	10.501	10.708	10.833	10.899	10.955	10.984	11.000
ratio	0.5271	0.9228	0.9547	0.9734	0.9848	0.9908	0.9959	0.9985	1.0000

Tabla 4.3. Cuantificación $f(t_{4.}^5)$

Es frecuente encontrar inercia saturada en los análisis, especialmente con funciones de un orden menor. Así, para iniciar el análisis de inercia saturada se recuantifica con una componente menos, es decir se recuantifica con $f(t_{4.}^5)$ que contiene las primeras cuatro componentes derivadas de la descomposición espectral de XC_5 y se obtienen los resultados de la tabla 4.3.

En la fila `ratio` de la tabla 4.3, se nota que existe SIE, es decir, la estructura inercial asociada con la cuantificación $f(t_5)$, ya está contenida en la matriz asociada con la recuantificación $f(t_{4.}^5)$, que toma las cuatro primeras componentes finales. Observe en este caso que la inercia acumulada eje por eje es prácticamente igual, y esta característica se mantiene hasta el último eje; esto sugiere que cuatro componentes serán suficientes en la cuantificación.

De hecho, al revisar la inercia obtenida con la cuantificación $f(t_{4.})$ en la tabla 4.2, se tiene que ésta efectivamente ya contiene la inercia derivada de las cuantificaciones $f(t_5)$ y $f(t_{4.}^5)$, con lo cual $f(t_{4.}) = t_{1.} + t_{2.} + t_{3.} + t_{4.}$ es la función de cuantificación definitiva generando inercia maximal por valor de 0.9228 en el plano de igual dimensión, y a partir de este la inercia que acumula es igual o superior.

En la tabla 4.4, se puede ver los valores cuantificados bajo $f(t_4)$ de las categorías de las variables *Appe* y *Terr* que parecen tener implícito un orden creciente natural.

Variable/categoría	1	2	3	4
Appe	-0.654	-0.142	2.011	
Terr	-0.790	-0.255	0.346	2.791

Tabla 4.4. Valores de cuantificación asociados a las categorías de Appe y Terr

La matriz de correlaciones de las variables incluyendo las cuantificadas con los primeros cuatro ejes (tabla 4.5) es muy importante, porque permite identificar con cuáles de ellos existe mayor relación lineal y por tanto contribuyen más a su formación.

	[t1]	[t2]	[t3]	[t4]
GInten	0,9285	0,1273	0,1015	0,1015
GAcid	-0,2961	0,4656	0,6732	0,4736
GAstr	0,7471	0,5079	-0,1151	0,1696
GAlcool	0,7368	0,4151	0,2114	-0,2640
GEqui	0,8664	-0,4044	0,0905	-0,0510
GVelou	0,9211	-0,3394	0,0383	-0,0429
Gamer	0,3243	0,8395	-0,0360	-0,1670
GIfin	0,9588	0,1420	0,1142	0,1084
GHarmo	0,9691	-0,1747	0,0081	-0,0152
Appe	-0,3222	-0,0243	0,8491	-0,3688
Terr	-0,2877	0,8673	-0,3114	-0,1521
$sum(r^2(,))$	λ_1	λ_2	λ_3	λ_4

Tabla 4.5. Correlación entre las variables y los cuatro primeros ejes

Las variables *Terr* y Gamer contribuyen en buena medida a la formación del eje 2, mientras que *Appe* y GAcid prácticamente definen el eje 3. En la misma tabla 4.5, se deducen los valores propios como la suma de los cuadrados de estas correlaciones con cada eje.

El área sombreada asociada a los cuatro primeros valores propios λ_1=5.80041, λ2=2.45020, λ_3=1.36351, λ_4=0.53656 es maximal ya que ha sido generada por la función de cuantificación de igual dimensión $f(t_4)$. De la ecuación [4.7], se evidencia en estos resultados que para cada variable cuantificada $\hat{x}_q$:

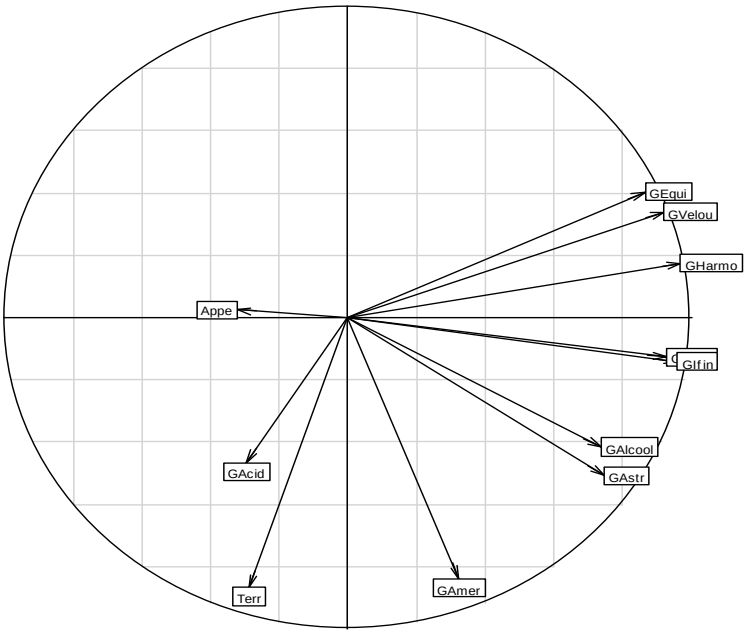
$r^2(\hat{x}_q, t1234.) = r^2(\hat{x}_q, t_1) + r^2(\hat{x}_q, t_2) + r^2(\hat{x}_q, t_3) + r^2(\hat{x}_q, t_4)$; ya que

`r(Appe,t1234.)= 0.9806` y `r(Terr,t12345.)= 0.9773`, entonces:

$0.9806^2 = (-0.3222^2) + (-0.02425^2) + 0.849134^2 + 0.36884^2$

$0.9538^2 = (-0.2877^2) + 0.86729^2 + (-0.311385^2) + (-0.15213)^2$.

Del circulo de correlaciones (gráfica 4.3) y del análisis de las contribuciones en el primer plano factorial, las variables comprometidas en la formación de la inercia recogida por el primer eje son `GInten` (14.86%), `GEqui` (12.94%), `GVelou` (14.63%), `GIfin` (15.85%) y `GHarmo` (16.19%) asociando altos cosenos cuadrados que oscilan entre 0.751 y 0.939. Este eje 1 representa la "calidad de los vinos".



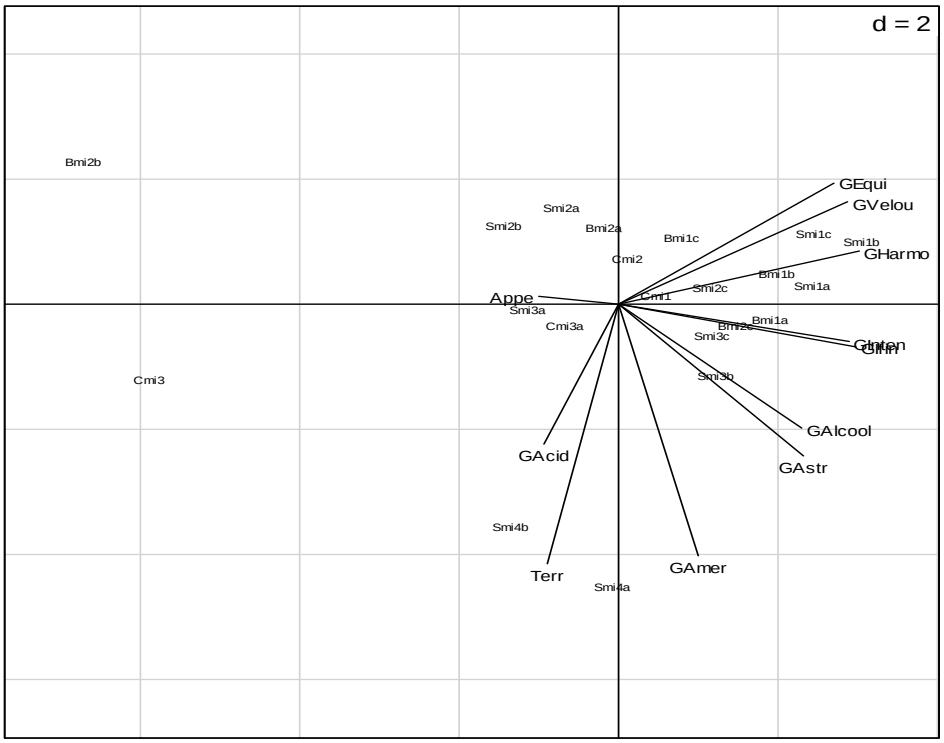
Gráfica 4.3. Círculo de correlaciones *gustacion*;
horizontal = eje1, vertical = eje2.

Análogamente, el segundo eje está caracterizado por las variables *Terr* y `GAmer` con contribuciones del 30.70% y 28.76% respectivamente. El origen de los vinos *Appe* no está bien representada en el primer plano principal, aunque si contribuye altamente en la formación del eje 3 junto con `GAcid`.

Las variables GAstr y GAlcool se pueden considerar medianamente influyentes en el primer eje, con contribuciones de 9.62 y 9.36 respectivamente, igualmente asocian cosenos cuadrados de aproximadamente 0.55. El tipo de suelo parece no tener relación con la intensidad de alcohol debido a la así "ortogonalidad" de las variables Terr y GAlcool en el círculo de correlaciones.

La representación simultánea, gráfica 4.4, de individuos y variables como vectores directores, permite el análisis de las interrelaciones entre individuos y variables. Los vinos más amargos *Smi4a, Smi4b* son de tipo de suelo *medio4*; mientras que el vino *Cmi3* presenta los más bajos índices de suavidad y armonía y *Bmi2b* el de menor intensidad.

Aunque los vinos de referencia *Smi1c* y *Smi1b* contribuyen medianamente en la formación del eje 1, presentan el mayor índice de suavidad (textura), armonía e intensidad, catalogándolos como los de mayor calidad. En este biplot los vinos *Bmi2b* y *Cmi3* claramente se diferencian por ser los de peor calidad.



Gráfica 4.4. Representación simultánea en primer plano factorial de la matriz *gustacion*

En general, los vinos asociados a las categorías (cuantificadas) de suelo *medio1*, *medio2*, y *medio3* también presentan índices medianos en las características que califican los ejes, ver su posicionamiento cerca al origen en la gráfica 4.4.

El ACP de la base *gustacion*, permite identificar el primer eje de calidad independientemente de las variables cuantificadas *Terr* y *Appe* asociadas a los ejes 2 y 3 ortogonales con el eje 1.

CAPÍTULO 5

REGRESIÓN PLS CON DATOS MIXTOS

GNM-PLSR está conformado por los métodos GNM-PLS1 y GNM-PLS2 cuyo objetivo es realizar una regresión PLS en dos conjuntos de datos mixtos $Y_{n.r}$ e $X_{n.p}$, conteniendo las submatrices Y_q^* e X_q^* de variables cualitativas en los conjuntos respectivos; es decir, $r_q$de las *r* variables en *Y* son cualitativas (La submatriz Y_k contiene las variables cuantitativas); análogamente, p_q de las *p* variables en *X* son cualitativas (X_k es la submatriz de variables exogenas cuantitativas).

Las variables cualitativas en uno y otro espacio, serán cuantificadas (estandarizadas), $\hat{Y}_q$, $\hat{X}_q$, en el desarrollo de la regresión PLS mediante un proceso de optimización, mientras que las variables cuantitativas mantienen su originalidad excepto por su estandarización. Cuando sólo se tiene una variable respuesta el método se denomina GNM-PLS1 y es un caso particular de GNM-PLS2.

Los métodos GNM-PLSR (General Non-Metric Partial Least Squares Regresion) son una extensión de los métodos NM-PLSR propuestos por Russolillo (2009), cuya cuantificación se lleva a cabo sólo con la primera componente $\gamma_t = t_1$ derivada de la regresión. GNM-PLSR se diferencia en que cuantifica óptimamente cada una de las variables cualitativas tomando una función agregada de las primeras $H \leq s$ componentes derivadas de la regresión; $s = rango(\frac{1}{n}X_k'Y_k)$.

En este capítulo se presentará primero la conceptualización del NM-PLSR ya que servirá de base en la estructuración del GNM-PLSR especialmente en la identificación de las funciones de cuantificación en uno y otro espacio, y de las funciones de maximización. Igualmente, se presentaran los casos de aplicación para los métodos GNM-PLS2 y GNM-PLS1 tomando como fuente de información la base de datos *cred.438* descrita en el apartado 5.2.3.

5.1. NM-PLSR

El algoritmo NM-PLSR propuesto por Russolillo (2009), esta basado en el algoritmo PLS-R descrito por Tenenhaus (1998).

En NM-PLSR las variables son cuantificadas óptimamente con funciones de la primera componente obtenidas en los espacios de respuesta y predictor, de acuerdo con las funciones [3.28] o [3.29] según se requiera orden o no. El algoritmo esta desarrollado basicamente en dos ciclos.

El primer ciclo inicia con la cuantificación de cada predictor cualitativo a partir de la componente u_1 (al inicio puede ser la 1ª columna de Y_k) del espacio de respuesta. $\hat{X}$ es ahora la matriz de predictores conteniendo las variables cuantitativas y las cuantificadas.

Inmediatamente, los parámetros del modelo w_1 son calculados y normalizados como función de $\hat{X}$ y u_1, para luego obtener $t_1 = \hat{X}w_1$ que a su vez permitirá cuantificar simultánea y optimamente las variables cualitativas de respuesta, para análogamente conformar $\hat{Y}$.

Seguidamente los pesos c_1 son computados como función de $\hat{Y}$ y t_1; y finalmente se recalcula la componente $u_1 = \hat{Y}c_1/c_1'c_1$ para reiniciar el ciclo hasta la convergencia de w_1.

Se obtienen por deflactacion las matrices residuales E_1 y F_1 de ambos espacios y son entradas al segundo ciclo PLS-R estándar, hasta obtener el resto de componentes ortogonales. El pseudo código del algoritmo NM-PLSR presenta la siguiente estructura:

Entrada: X_k, X^*, Y_k, Y^*
Salida: $W, C, T, U, \hat{X}, \hat{Y}$

1. Inicio u_1
Repita
    1.1. $\hat{X}_q = q_i(u_1, X_q^*)$
    1.2. $\hat{X} = X_k|\hat{X}_q$
    1.3. $w_1 = \hat{X}'u_1/\|\hat{X}'u_1\|$
    1.4. $t_1 = \hat{X}w_1$
    1.5. $\hat{Y}_q = q_i(t_1, Y_q^*)$
    1.6. $\hat{Y} = Y_k|\hat{Y}_q$
    1.7. $c_1 = \hat{Y}'t_1/t_1't_1$
    1.8. $u_1 = \hat{Y}c_1/c_1'c_1$
Hasta convergencia de w_1

1.9. $p_1 = \hat{X}t_1/t_1't_1$
1.10. $E_1 = \hat{X} - t_1p_1'$
1.11. $F_1 = \hat{Y} - t_1c_1'$

Para todo $h = 2, \dots, H$
2. Inicialice u_h
Repita
2.1. $w_h = E_{h-1}'u_h/\|E_{h-1}'u_h\|$
2.2. $t_h = E_{h-1}w_h$
2.3. $c_h = F_{h-1}'t_h/t_h't_h$
2.4. $u_h = F_{h-1}c_h/c_h'c_h$
Hasta convergencia de w_h

2.5. $p_h = E_{h-1}'t_h/t_h't_h$
2.6. $E_h = E_{h-1} - t_hp_h'$
2.7. $F_h = F_{h-1} - t_hc_h'$
Fin h

Dado que las cuantificaciones aquí se realizan con la primera componente obtenida en cada espacio, y de acuerdo con la ecuación [3.7], NM-PLSR obtiene matrices de datos óptimamente escaladas $\hat{X}$ e $\hat{Y}$ maximizando el criterio

$$cov^2(\hat{Y}c_1^*, \hat{X}w_1) = cov^2(u_1^*, t_1)\ . \qquad [5.1]$$

El criterio [5.1] depende de dos conjuntos de parámetros; el primero consiste de los parámetros del modelo restringidos a norma 1, esto es $\| w_1 \| = 1$ e $\| c_1^* \| = 1$. El otro conjunto consiste de los parámetros scaling con las restricciones propias debido al nivel de scaling escogido para cada variable ahora normalizada tal que $v(\hat{y}_r) = 1$ y $v(\hat{x}_p) = 1$.

El principal objetivo de NM-PLSR es maximizar el criterio [5.1] respecto a los parámetros scaling y del modelo. Este problema de la maximización será resuelto para los parámetros del modelo, manteniendo los parámetros scaling fijos. Después, se resolverá el mismo problema respecto a los parámetros scaling, fijando los parámetros del modelo.

La maximización de [5.1] para los parámetros scaling fijados, está dada según el AIB por los vectores normados w_1 y $c_1^* = c_1/\| c_1 \|$ correspondientes al valor propio común más grande λ verificando las ecuaciones de estacionariedad (ver ecuación 3.5).

$$\hat{X}'\hat{Y}\hat{Y}'\hat{X}w_1 = \lambda w_1 \qquad [5.2]$$
$$\hat{Y}'\hat{X}\hat{X}'\hat{Y}c_1^* = \lambda c_1^* \qquad [5.3]$$

Observe que a nivel de cálculo es equivalente tomar [5.3] o $\hat{Y}'\hat{X}\hat{X}'\hat{Y}c_1 = \lambda c_1$. Para $\hat{Y}c_1 = u_1$ fijos en [5.3], se maximiza λ así:

$$\lambda = c_1'\hat{Y}'\hat{X}\hat{X}'\hat{Y}c_1 = u_1'\hat{X}\hat{X}'u_1$$
$$\lambda = (\hat{X}'u_1)^2 = \sum_p cov^2\,(\hat{x}_p, u_1)$$
$$\lambda = \sum_p cor^2\,(\hat{x}_p, u_1)var(\hat{x}_p)var(u_1)$$
$$\lambda = \sum_p cor^2\,(\hat{x}_p, u_1)var(u_1)\ . \qquad [5.4]$$

Ya que $var(\hat{x}_p) = 1$. Como $var(u_1)$ es fija respecto a la suma, las soluciones óptimas para los parámetros scaling de las variables exógenas consisten en maximizar

$$\sum_p cor^2\,(\hat{x}_p, u_1)\ . \qquad [5.5]$$

La función [5.5] es separable respecto a cada variable $\hat{x}_p$ óptimamente escalada y puede ser descompuesta en sumandos los cuales son función solamente de los parámetros scaling de cada una de las *p* variables. Observe que $\hat{x}_p$ se cuantificó con la primera componente de *Y*.

Por tanto el problema puede ser resuelto maximizando separadamente la correlación al cuadrado de cada predictor cuantificado y la primera componente en el espacio de *Y*, teniendo en cuenta el nivel scaling de $\hat{x}_p$.

Si se desea cuantificar la variable predictor cualitativa x_p^* a nivel nominal, la siguiente función scaling ha de ser usada:

$$\tilde{q}(x_p^*, u_1):\ \hat{x}_p = \tilde{X}_p(\tilde{X}_p'\tilde{X}_p)^{-1}\tilde{X}_p'u_1 = P_{\tilde{X}_p}u_1$$

con lo cual se obtiene $max\{cor^2(\hat{x}_p, u_1)\}$ de acuerdo con [5.4].

Un razonamiento similar en [5.2] puede ser hecho para encontrar las cuantificaciones optimas asociadas a las variables respuesta, llevando a la maximización de cada

$$cor^2(\hat{y}_r, t_1)\ . \qquad [5.6]$$

Observe que $\hat{y}_r$ se cuantifica con la primera componente de X y que [5.6] es equivalente a la ecuación [3.3] para cada r. Por tanto y de acuerdo con [5.4], maximizar $cov^2(u_1^*, t_1)$ es equivalente a maximizar

$$\sum_p cor^2(\hat{x}_p, u_1) var(u_1) \quad \text{ó} \quad \sum_r cor^2(\hat{y}_r, t_1) var(t_1) .$$

5.2. GNM-PLSR: GNM-PLS2

Los métodos (algorítmicos) GNM-PLSR (General Non-Metric Partial Least Squares Regresion) propuestos en este trabajo de tesis, tienen por objetivo realizar una regresión PLS en dos conjuntos de datos mixtos $Y_{n.r}$ e $X_{n.p}$, conteniendo variables cualitativas en los conjuntos respectivos; es decir, r_q de las r variables en Y son cualitativas (r_k son cuantitativas), mientras que p_q de las p variables en X son cualitativas (p_k son cuantitativas).

Todas las variables cuantitativas mantendrán su originalidad excepto por estandarización, y sólo se cuantificarán las variables cualitativas que también han de ser estandarizadas. La yuxtaposición de los subconjuntos cuantitativos y cualitativos cuantificados, dan lugar a las matrices $\hat{X} = [X_{p_k} | \hat{X}_{p_q}]$ e $\hat{Y} = [Y_{r_k} | \hat{Y}_{r_q}]$.

Las variables cualitativas en Y e X serán cuantificadas respectivamente mediante las funciones agregadas γ_t e γ_u formadas generalmente con un número menor o igual de componentes asociadas al rango de la matriz $R_{12_k} = X_k Y_k$; es decir, el número H de componentes de la regresión-cuantificación PLS2 estará limitado de acuerdo con el AIB por $H \leq s = rango(R_{12_k})$, y será determinado mediante validación cruzada implementando previamente la función $fVC2(Y_k, X_k, s)$ que a su vez invoca la función $fPLS2(Y_k, X_k, H)$.

Por tanto, las funciones de cuantificación inicial de rango H respectivas son:

$$\gamma_t = c_{r1} t_1 + c_{r2} t_2 + \cdots + c_{rH} t_H \qquad [5.7]$$

$$\gamma_u = d_{p1} u_1 + d_{p2} u_2 + \cdots + d_{pH} u_H \ . \qquad [5.8]$$

Los pesos c_{rh} e d_{ph} se estiman como los coeficientes de las regresiones de las variables cuantificadas $\hat{y}_r \sim c_r \gamma_t$ e $\hat{x}_p \sim d_p \gamma_u$ respectivamente.

Así, el problema fundamental se resuelve consiguiendo una cuantificación (estandarizada) inicial $\hat{y}_r = \sum_{h=1}^{H} P_{\tilde{y}_r}\gamma_t$ bajo $c_{rh} = 1\ \forall\ h$, a partir de la cual se obtienen los coeficientes $\hat{c}_{rh} = r(\hat{y}_r, t_h)/s_{t_h}$ derivados de la regresión $\hat{y}_r \sim c_{rh}t_h$ con los cuales se reconstituye la función de cuantificación de orden H y se recuantifica la variable cualitativa, en un proceso iterativo convergente.

Un razonamiento análogo al anterior, permitirá inicialmente cuantificar las variables cualitativas x^* a partir de H componentes del espacio de las Y a través de $\hat{x}_p = \sum_{h=1}^{H} P_{\tilde{x}_q}\gamma_u$ bajo $d_{ph} = 1\ \forall\ h$, luego mediante regresiones sucesivas con cada componente $\hat{x}_q \sim d_{ph}u_h$ se estiman los coeficientes de regresión obteniendo $\hat{d}_{ph} = r(\hat{x}_p, u_h)/s_{u_h}$.

Las componentes u_h de la función γ_u son combinación lineal de las Y, y los pesos d_{ph} representan los coeficientes de regresión de $\hat{x}_p$ y la componente u_h y se pueden estimar a partir de la razón de correlación de la variable cualitativa x_p^* y la componente u_h.

Con estos resultados se conforman las matrices $\hat{X}$ e $\hat{Y}$ y se reconstituyen actualizar las funciones de cuantificación conllevando una nueva recuantificación. El proceso es iterativo hasta la convergencia de los pesos c_{rh} .

Observe que tanto las componentes t_h como la función de cuantificación γ_t que es una aplicación directa del concepto de regresión múltiple para cada variable respuesta $\hat{y}_r = c_{r1}t_1 + c_{r2}t_2 + \cdots + c_{rs}t_s$, se obtienen de manera natural con los mismos recursos generados en el proceso de regresión PLS2, permitiendo una fluidez limpia del método.

5.2.1. Maximización

Extendiendo el criterio [5.1] a H componentes, se maximiza globalmente

$$\sum_h^H cov^2(u_h, t_h)\ . \qquad [5.9]$$

De otro lado, sea $\gamma_t = c_{r1}t_1 + c_{r2}t_2$ la función de cuantificación tal que

$$\tilde{q}(y_r^*, \gamma_t):\ \hat{y}_r = \tilde{Y}_r(\tilde{Y}_r'\tilde{Y}_r)^{-1}\tilde{Y}_r'\gamma_t = P_{\tilde{y}_r}\gamma_t$$

con lo cual y debido a la razón de correlación se tiene

$$\text{máxima } cor^2(\hat{y}_r, \gamma_t) \; . \qquad [5.10]$$

La ecuación [3.5] obtenida de la convergencia en la cuantificación se obtiene recurrentemente de la etapa 2 del algoritmo PLS2, mediante

$$w_h = \hat{X}'_{h-1}u_h / \left\|\hat{X}'_{h-1}u_h\right\| = \hat{X}'_{h-1}\hat{Y}_{h-1}c_h / c'_h c_h \left\|\hat{X}'_{h-1}u_h\right\|$$

$$= \hat{X}'_{h-1}\hat{Y}_{h-1}\hat{Y}'_{h-1}t_h / t'_h t_h . c'_h c_h \left\|\hat{X}'_{h-1}u_h\right\|$$

$$= X'_{h-1}Y_{h-1}Y'_{h-1}X_{h-1}w_h / \lambda_h$$

donde $\lambda_h = t'_h t_h . c'_h c_h . \|X'_{h-1}u_h\|$

Con $h = 1, 2$ componentes se tiene los siguientes sistemas:

$$w'_1\hat{X}'\hat{Y}\hat{Y}'\hat{X}w_1 = \lambda_1 \;\; ; \;\; w'_2\,\hat{X}'_1\hat{Y}_1\hat{Y}'_1\hat{X}_1 w_2 = \lambda_2$$

$$t'_1\hat{Y}\hat{Y}'t_1 = \lambda_1 \;\; ; \;\; t'_2\hat{Y}_1\hat{Y}'_1 t_2 = \lambda_2$$

$$(\hat{Y}'t_1)^2 = \lambda_1 \;\; ; \;\; t'_2\left(\hat{Y} - t_1c'_1\right)\left(\hat{Y} - t_1c'_1\right)' t_2 = \lambda_2$$

$$t'_2\hat{Y}\hat{Y}'t_2 = \left(\hat{Y}'t_2\right)^2 = \lambda_2$$

de donde,

$$(\hat{Y}'t_1)^2 = \sum_r cov^2(\hat{y}_r, t_1) = \sum_r r^2\,(\hat{y}_r, t_1) var(\hat{y}_r) var(t_1) \; .$$

- maximizar $\sum_r cov^2(\hat{y}_r, t_1)$ es equivalente a maximizar $\sum_r r^2\,(\hat{y}_r, t_1)$ ya que $var(\hat{y}_r) = 1$ y $var(t_1)$ no depende de *r*; por tanto se maximiza

$$\sum_r r^2(\hat{y}_r, t_1) = \sum_r r^2(\hat{y}_r, c_{r1}t_1) \qquad [5.11]$$

análogamente,

$(\hat{Y}'t_2)^2 = \sum_r cov^2(\hat{y}_r, t_2) = \sum_r r^2\,(\hat{y}_r, t_2) var(\hat{y}_r) var(t_2)$ maximizando

- $\sum_r r^2(\hat{y}_r, t_2) = \sum_r r^2(\hat{y}_r, c_{r2}t_2) \; .$ [5.12]

Observe de [5.11] y [5.12] y de la ortogonalidad de t_1, t_2 que

$$\sum_r r^2(\hat{y}_r, c_{r1}t_1) + \sum_r r^2(\hat{y}_r, c_{r2}t_2) = \sum_r [r^2(\hat{y}_r, c_{r1}t_1) + r^2(\hat{y}_r, c_{r2}t_2)]$$

$$= \sum_r r^2(\hat{y}_r, \gamma_t) \ . \qquad [5.13]$$

Este resultado es equivalente al agregado de [5.10] para cada *r-ésima* variable y por tanto es maximal; además es igual a $\lambda_1 + \lambda_2$ que son los primeros valores propios asociados a la primera fase de las etapas $h = 1{,}2$ del AIB correspondiente.

Es equivalente entonces estudiar la maximización con $H \leq s$ componentes mediante [5.9] ó mediante [5.13] teniendo en cuenta que $\gamma_t = c_{r1}t_1 + \cdots + c_{rH}t_H$, esto es: $\sum_h^H cov^2(u_h, t_h) \equiv \sum_r r^2(\hat{y}_r, \gamma_t)$.

Se muestra gráficamente el proceso de convergencia monótona y creciente con los resultados derivados de [5.9] o [5.13] y la siguiente propiedad para $H \leq s$:

$$cor^2(u_1^a, t_1^a) \geq cor^2\left(u_1^b, t_1^b\right) \geq \cdots \geq cor^2(u_1^H, t_1^H)$$

$$\sum_{h=1}^H cor^2\,(u_h^H, t_h^H) \geq \cdots \geq \sum_{h=1}^2 cor^2\left(u_h^b, t_h^b\right) \geq cor^2(u_1^a, t_1^a) \ .$$

Nota: los superíndices indican el número *H* (rango) de componentes que conforman la función de cuantificación. Así, $t_1^a \neq t_1^b$ etc.

5.2.2. El algoritmo GNM-PLS2

- *Inicio: Cuantificación agregada*

El proceso inicia cuantificando las variables cualitativas exógenas X_{pq} a partir de una función agregada γ_u ($d_{ph} = 1 \ \forall \ h$) de $H \leq s$ componentes u_h.

El rango *s* de la matriz de correlaciones $R_{12} = X_k' Y_k / n$ proporcionará el número máximo de componentes requeridas en la cuantificación inicial, las cuales se obtendrán de la función de validación cruzada fVC2(Y_k ,X_k ,s) del entorno R (ver anexo), y corresponderán con el numero *H* de componentes en la regresión .

Se conforma entonces la matriz $\hat{X}$ que servirá de parámetro de entrada a la función f.PLS2(Y_k , $\hat{X}$, s) que a su vez suministrará las H componentes t_h para conformar γ_t ($c_{rh} = 1 \; \forall \; h$) y cuantificar las variables cualitativas endógenas obteniendo $\hat{Y}$. Observe que los scaling se realizan con los recursos propios derivados de la regresión PLS2.

- *Iteración – Convergencia*. Básicamente está conformada por tres fases:

  a) *Maximización*. Las matrices estimadas, sirven ahora de parámetros de entrada a la función $f.PLS2(\hat{Y}, \hat{X}, s)$ de la cual se obtienen las funciones de maximización [5.10] y [5.12] en cada iteración para su representación gráfica, y las componentes *U* (no ortogonales) para la siguiente fase.

  b) re-*Cuantificación de* X_{pq}. Se calcula la función agregada γ_u, con la cual se cuantifican (y estandarizan) la variables cualitativas exógenas, permitiendo obtener $\hat{X}$ que a su vez es parámetro de entrada en la siguiente fase.

  c) re-*Cuantificación de* Y_{rq}. Se corre nuevamente la función $f.PLS2(\hat{Y}, \hat{X}, s)$ y se obtienen las componentes t_h para conformar la función agregada γ_t, la cual permite cuantificar (estandarizar) las variables cualitativas de respuesta, para conformar $\hat{Y}$. Recuerde que $t_h = X_{h-1} w_h$.

Se repiten las fases a), b) y c) hasta la convergencia de los w_h que a su vez conlleva la convergencia en las variables cuantificadas y en las funciones de maximización.

El *pseudo-algoritmo* asociado a `fGNMPLS2()` tiene la forma siguiente:

Convenciones : $Y_{rq} = Y_q$; $X_{pq} = X_q$; $t_x = \gamma_t$, $u_y = \gamma_u$

Entrada Yq, Yk, Xk, Xq, H *(Componentes cuantificación, $H \leq s$)*
Salida $W, T, C, U, P, B, X, Y, S2tu, r2Yt$

Inicio

$X = X_k$; $Y = Y_k$

$f.PLS2(X, Y, H)$ => $U: u_1, u_2, \dots, u_H$
$u_y = U.du = u_1 + u_2 +, \dots, +u_H$; $du = (1,1, \dots, 1)$
$\hat{x}_q = q_i(x_q^*, u_y)$ cuantfcn $x_q \; \forall \; pq$; $X = \hat{X} = [X_k | \hat{X}_q]$

$f.PLS2(X,Y,H)$ => $T: t_1, t_2, \dots, t_H$

$t_x = T.ct = t_1 + t_2 +, \dots, +t_H$; $ct = (1,1,\dots,1)$

$\hat{y}_q = q_i(y_q^*, t_x)$ cuantfcn $y_q \forall rq$; $Y = \hat{Y} = [Y_k|\hat{Y}_q]$

Repita

- *maximización*

$f.PLS2(X,Y,H)$ => *T, U*

$S2tu = \sum_h^H cov^2(t_h, u_h)$; $r2Yt = r^2(Y, t_h)$

- *re-cuantificación* x_q

$sU = d.e(U)$; $du = r(U, \hat{X}_q)/sU$; $u_y = U.du$

$\hat{x}_q = q_i(x_q^*, u_y)$ cuantfcn $x_q \forall pq$; $X = \hat{X} = [X_k|\hat{X}_q]$

- *re-cuantificación* y_q

$f.PLS2(X,Y,H)$ => $T: t_1, t_2, \dots, t_H$; $[t_x = T.w_x]$

$sT = d.e(T)$; $ct = r(T, \hat{Y}_q)/sT$; $t_x = T.ct$

$\hat{y}_q = q_i(y_q^*, t_x)$ cuantfcn $y_q \forall rq$; $Y = \hat{Y} = [Y_k|\hat{Y}_q]$

Hasta convergencia w_x

$Retorne\ (W, T, C, U, P, B, X, Y, S2tu, r2Yt)$

Fin # Ver algoritmos `fGNMpls2()` bajo R en anexo C

5.2.3. Aplicación GNM-PLS2

Para los ejemplos de aplicación, se utiliza la base de datos "c*réd.438*" con *n* = 438 observaciones que corresponden a solicitudes de crédito por igual número de clientes de una institución financiera. Para cada método (GNM-PLS2, GNM-PLS1) se conforman los grupos de datos mixtos, los cuales sirven como casos prácticos para la implementación de los algoritmos correspondientes y la comprensión de los resultados.

- **Los datos**

La base c*réd.438* presenta la siguiente estructura para los primeros 15 individuos:

	Dictam	AnTrab	Vvienda	Plazo	Edad	EstCivil	Registros	TipTrab	Gastos
1917	1	1	1	36	27	2	1	1	55
2980	1	0	2	60	61	2	1	4	35
928	1	3	6	12	30	1	1	3	35
2112	1	1	5	36	24	1	1	1	35

```
3658      2      3       1    48   25        2         2       3     35
2617      1      7       5    60   30        1         1       1     35
388       1     20       5    60   36        2         1       1     60
3952      2      5       5    48   20        1         1       4     45
3841      1      0       5    24   20        1         1       1     35
1639      1     30       2    24   57        2         1       3     45
2326      1      3       1    48   32        2         1       1     57
2945      2      2       5    36   24        1         1       1     35
1038      1      2       2    36   38        2         1       1     45
999       1     13       2    60   42        2         1       1     60
2206      1     18       2    60   35        2         1       1     75
  :
```

```
     Ingresos Patrim CargPtrim ImprtSolic VrBfnciado

1917      100      0         0       1000       1300
2980      113   3000         0       1300       1726
928       150      0         0        920       1396
2112       60      0         0       1000       1602
3658        0      0         0       1300       1361
2617       95      0         0        900       1158
388       121      0         0        850        935
3952        0      0         0       2500       3542
3841      155   7000         0       1450       1688
1639       40  17000         0        400        500
2326      121      0         0       1200       1952
2945       65      0         0        895        895
1038      114   3600         0        650       1100
999       107   4000         0       1000       1359
2206      179   7000      3200       1000       1455
  :
```

Las variables *CargPtrim* (Carga Patrimonial) y *Plazo* han sido categorizadas como sigue:

CargPtrim	0	$0 < CargPtrim \leq 1500$	> 1500
	sin (carga)	baja	alta
Plazo	≤ 24	$24 < Plazo \leq 48$	> 48
	corto	medio	largo

De la base *cred.438* el bloque de las predictoras cuantitativas X_k está conformado por cinco variables cuantitativas (`Antrab, Edad, Gastos, Ingresos` y `Patrim`) y la sub-matriz X_q conteniendo las predictoras cualitativas (`Registros, CargPtrim`). El segundo bloque lo conforman las variables de respuesta *Y*, conteniendo a su vez la submatriz Y_k con dos variables endógenas cuantitativas (`ImprtSolic` y `VrBfnciado)` y la submatriz Y_q conformada por las endógenas cualitativas (`Dictam, Plazo`).

Se calcula el rango de la matriz de correlaciones entre las variables X_k, Y_k, para determinar el número *s* máximo de componentes agregadas en la cuantificación.

Luego mediante validación cruzada se determinará el número de componentes en la regresión-cuantificación evalúando los índices R^2, R^2vc (R^2 validación cruzada), RSS (suma de residuales cuadrados), PRESS (suma de cuadrados del error de predicción):

```
R2:
     ImprtSolic VrBfnciado
[1,]  0.1576505  0.3326368
[2,]  0.1793686  0.3790771

R2vc:
     ImprtSolic VrBfnciado
[1,]  0.1378973  0.2479709
[2,]  0.1554735  0.2849563

RSS:
     ImprtSolic VrBfnciado
[1,]   368.1067   291.6377
[2,]   358.6159   271.3433

PRESS:
     ImprtSolic VrBfnciado
[1,]   376.7389   328.6367
[2,]   369.0581   312.4741
```

Tomando en cuenta estos resultados, se observa que el número de componentes necesarios en la regresión-cuantificación es dos, ya que se tiene el mayor R^2vc (0.1555 y 0.2849) y el menor PRESS(369.05, 312.47) asociados respectivamente con las variables de respuesta ImprtSolic y VrBfnciado.

Se procede entonces a cuantificar las variables cualitativas tanto en Y_q como en X_q con $H = 1, 2$ componentes agregadas, iterando en cada etapa H el algoritmo `fGNMpls2()` hasta la convergencia.

En cada fase de cuantificación y en cada ciclo se calcula el cuadrado de la covarianza entre las componentes t_h *e* u_h, o equivalentemente de la correlación cuadrada entre las componentes t_h e *Y* de acuerdo con el proceso de maximización para estudiar su tendencia (ver tabla 5.1).

- Función de maximización : $\sum_h^H cov^2(t_h, u_{h.})$ de convergencia monótona creciente:

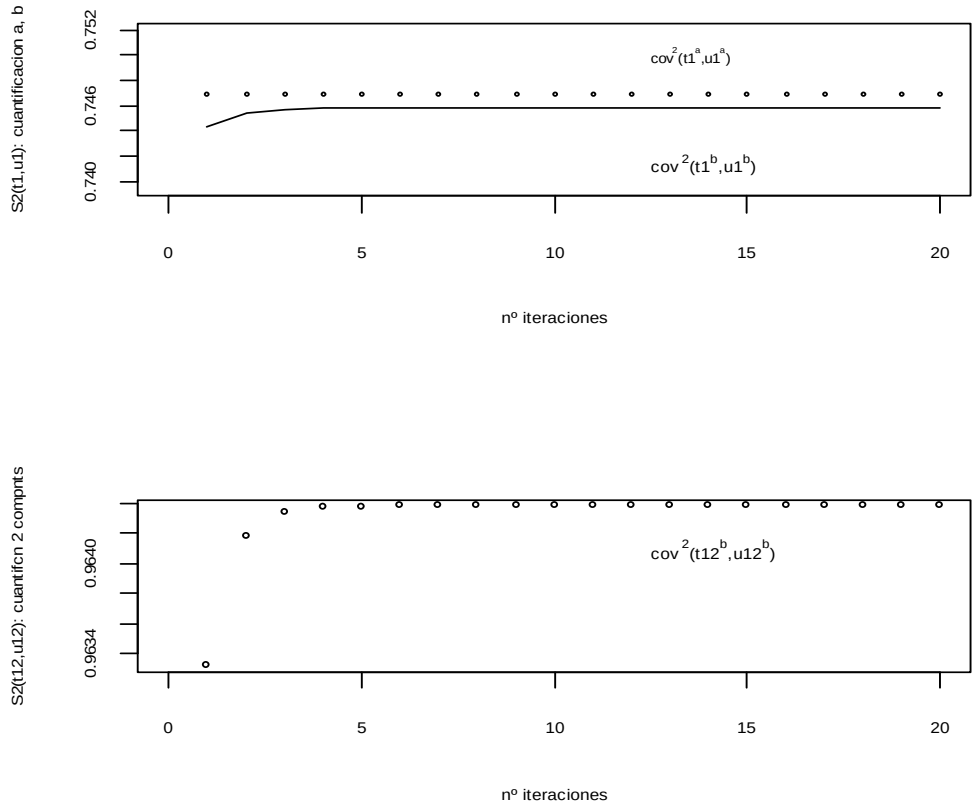
	S2tuH[,1^a,1]	S2tuH[,1^b,2]	S2tuH[,1^b,2]+ S2tuH[,2^b,2]
1	0.7469260	0.7443447	0.9633212
2	0.7469803	0.7454728	0.9641795
3	0.7469803	0.7457043	0.9643352
4	0.7469803	0.7457579	0.9643699
5	0.7469803	0.7457709	0.9643782
6	0.7469803	0.7457741	0.9643802
7	0.7469803	0.7457748	0.9643807
8	0.7469803	0.7457750	0.9643809
9	0.7469803	0.7457751	0.9643809
10	0.7469803	0.7457751	0.9643809
11	0.7469803	0.7457751	0.9643809
12	0.7469803	0.7457751	0.9643809
13	0.7469803	0.7457751	0.9643809
14	0.7469803	0.7457751	0.9643809
15	0.7469803	0.7457751	0.9643809
:			

	r2YtH[,1^a,1]	r2YtH[,1^b,2]	r2YtH...1...2...r2YtH...2...2.
1	0.7469260	0.7443447	0.9633212
2	0.7469803	0.7454728	0.9641795
3	0.7469803	0.7457043	0.9643352
4	0.7469803	0.7457579	0.9643699
5	0.7469803	0.7457709	0.9643782
6	0.7469803	0.7457741	0.9643802
7	0.7469803	0.7457748	0.9643807
8	0.7469803	0.7457750	0.9643809
9	0.7469803	0.7457751	0.9643809
10	0.7469803	0.7457751	0.9643809
11	0.7469803	0.7457751	0.9643809
12	0.7469803	0.7457751	0.9643809
13	0.7469803	0.7457751	0.9643809
14	0.7469803	0.7457751	0.9643809
15	0.7469803	0.7457751	0.9643809
:			

Tabla 5.1. Equivalencia (Tendencia[11])as funciones $cov^2(t_h, u_{h.}) \equiv r^2(y_r, t_h)$

Se compara entonces la tendencia de las series asociadas con t_1^a e t_1^b en la parte superior y se presenta la función asociada con $t_1^b + t_2^b$ en la parte inferior de la gráfica 5.1. Se aprecia que estas covarianzas son monótonas crecientes y convergentes. $cov^2(t_1^a, u_1^a)$ es máxima cuando se cuantifica con una (a) componentes, mientras que $cov^2(t_{12}^b, u_{12}^b)$ es máxima cuando se cuantifica con dos (b) componentes agregadas.

[11] El superíndice *a, b*, … indica el número 1, 2, … de componentes usadas en la cuantificación. Se distingue así, la primera componente t_1^a de t_1^b derivadas de las matrices cuantificadas.



Gráfica 5.1. Tendencia de $cov^2(t,u)$ bajo cuantificación con a y b componentes

Así, la matriz cuantitativa y cuantificada en el espacio de las Y utilizando dos componentes agregadas es:

```
      ImprtSolic  VrBfnciado     Dictam      Plazo
1917 -0.08002388 -0.20111041  0.6171159 -0.1193743
2980  0.55121264  0.48977571  0.6171159 -0.6734466
928  -0.24835362 -0.04541776  0.6171159  2.3772277
2112 -0.08002388  0.28867271  0.6171159 -0.1193743
3658  0.55121264 -0.10218070 -1.6167416 -0.1193743
2617 -0.29043606 -0.43140578  0.6171159 -0.6734466
388  -0.39564214 -0.79306682  0.6171159 -0.6734466
3952  3.07615875  3.43496157 -1.6167416 -0.1193743
3841  0.86683091  0.42814737  0.6171159  2.3772277
1639 -1.34249693 -1.49854911  0.6171159  2.3772277
2326  0.34080047  0.85630214  0.6171159 -0.1193743
2945 -0.30095667 -0.85793875 -1.6167416 -0.1193743
1038 -0.81646649 -0.52547008  0.6171159 -0.1193743
999  -0.08002388 -0.10542430  0.6171159 -0.6734466
2206 -0.08002388  0.05026834  0.6171159 -0.6734466
 :
```

Mientras que, la base de datos conteniendo tanto las variables cuantitativas como las cuantificadas con dos componentes en el espacio de las X es para las primeras 15 observaciones la siguiente:

	AnTrab	Edad	Gastos	Ingresos	Patrim	Registros
1917	-0.8472367	-0.8701936	-0.03248733	-0.3903932	-0.5048723	-0.4467028
2980	-0.9703996	2.2396992	-1.05618867	-0.2339062	-0.2097840	-0.4467028
928	-0.6009110	-0.5957913	-1.05618867	0.2114802	-0.5048723	-0.4467028
2112	-0.8472367	-1.1445959	-1.05618867	-0.8718920	-0.5048723	-0.4467028
3658	-0.6009110	-1.0531285	-1.05618867	-1.5941401	-0.5048723	2.2335139
2617	-0.1082596	-0.5957913	-1.05618867	-0.4505806	-0.5048723	-0.4467028
388	1.4928575	-0.0469867	0.22343801	-0.1376064	-0.5048723	-0.4467028
3952	-0.3545853	-1.5104656	-0.54433800	-1.5941401	-0.5048723	-0.4467028
3841	-0.9703996	-1.5104656	-1.05618867	0.2716675	0.1836671	-0.4467028
1639	2.7244860	1.8738294	-0.54433800	-1.1126414	1.1672949	-0.4467028
2326	-0.6009110	-0.4128564	0.06988281	-0.1376064	-0.5048723	-0.4467028
2945	-0.7240739	-1.1445959	-1.05618867	-0.8117047	-0.5048723	-0.4467028
1038	-0.7240739	0.1359482	-0.54433800	-0.2218687	-0.1507663	-0.4467028
999	0.6307175	0.5018179	0.22343801	-0.3061310	-0.1114212	-0.4467028
2206	1.2465318	-0.1384541	0.99121402	0.5605668	0.1836671	-0.4467028

:

	CargPtrim
1917	-0.3504225
2980	-0.3504225
928	-0.3504225
2112	-0.3504225
3658	-0.3504225
2617	-0.3504225
388	-0.3504225
3952	-0.3504225
3841	-0.3504225
1639	-0.3504225
2326	-0.3504225
2945	-0.3504225
1038	-0.3504225
999	-0.3504225
2206	2.8471828

:

Observe que para Carga Patrimonial *"CargPatrim"* el método cuantifica las categorías *sin* y *baja* con el mismo scoring -0.3504225 mientras que *"Plazo"* presenta cuantificación con orden inverso.

Los scaling asociados a cada una de las categorías de las variables cualitativas se presentan en la tabla 5.2 a seguir:

Xq : Categorías cuantificadas

Registros	si	no			
Cuantifcn	2.2335139	-0.4467028			

CargPtrim	sin	baja	alta		
Cuantifcn	-0.3504225	-0.3504225	2.8471828		

Yq : Categorias Cuantificadas

Dictam	positivo	negativo			
Cuantifcn	0.6171159	-1.6167416			

Plazo	corto	medio	largo		
Cuantifcn	2.4003373	-0.1639000	-0.6352217		

Tabla 5.2. Variables cualitativas X_q e Y_q cuantificadas

Una vez se tiene la matriz X e Y cuantificada óptimamente, se procede mediante *validación cruzada* a determinar el número de componentes en la regresión PLS2; se implementa el algoritmo fVC2 out leave one, y se evalúan los índices más importantes: R^2, R^2vc, RSS, PRESS

R^2

```
     ImprtSolic VrBfnciado     Dictam       Plazo

[1,]  0.1387317  0.2774524 0.07215094 0.007865552
[2,]  0.1797241  0.3688513 0.13186347 0.011901559
```

R^2vc

```
     ImprtSolic VrBfnciado     Dictam       Plazo
[1,]  0.1079348  0.1627483 0.05427656 -0.006872157
[2,]  0.1421934  0.2337684 0.10705998 -0.013132506
```

RSS

```
       ImprtSolic VrBfnciado   Dictam    Plazo
[1,]   376.3743   315.7533     405.4700  433.5628
[2,]   358.4606   275.8120     379.3757  431.7990
```

PRESS

```
       ImprtSolic VrBfnciado   Dictam    Plazo
[1,]   389.8325   365.8790     413.2811  440.0031
[2,]   374.8615   334.8432     390.2148  442.7389
```

Estos resultados especialmente los PRESS también pueden ser obtenidos a partir de la librería *pls* de R:

```
PRESSr
           1 comps  2 comps

ImprtSolic 391.9237 375.3475
VrBfnciado 371.8477 337.0535
Dictam     413.6782 390.5888
Plazo      439.9954 442.9060
```

Note que de la función fVC2, PRESS ~ PRESSr de la librería pls.

Si se tiene especial interés en la variable endógena Dictamen (Dictam), se evalúa el menor PRESS= 390,2148 el cual indica que se debe tomar *dos* componentes para la regresión PLS2 asociando a su vez el mayor R^2 = 0.1318635 y el mayor R^2vc = 0.10705998. Así, tomando dos componentes en la regresión se tiene los siguientes resultados:

```
WH.
                 [,1]        [,2]
AnTrab     0.13600030 -0.60443621
Edad       0.15062740 -0.33469190
Gastos     0.08971998 -0.16063954
Ingresos   0.50854993  0.02899684
Patrim     0.81001932  0.59194955
Registros -0.04374285  0.47034457
CargPtrim  0.18465607 -0.12742054

C.H
                [,1]        [,2]
ImprtSolic 0.3038173  0.19178036
VrBfnciado 0.4296541  0.28636669
Dictam     0.2191017 -0.23146473
Plazo      0.0723418  0.06017666

T.H[1:10,]
            [,1]        [,2]
 [1,] -0.9018726  0.33293371
 [2,] -0.2234255 -0.28981615
 [3,] -0.6128035  0.27410466
 [4,] -1.2799177  0.57525892
 [5,] -1.7171794  1.63543978
 [6,] -0.8824937 -0.04286937
 [7,] -0.3081052 -1.39080636
 [8,] -1.5894030  0.29676982
 [9,] -0.2124905  1.21289687
[10,]  0.9384731 -1.69322794
  :
```

Recuerde que $T = X * WH$. El análisis de redundancias muestra el porcentaje promedio de variabilidad explicada en *Y* a partir de las *X*; en este ejercicio es relativamente bajo, 0.17308; las siguientes regresiones sin intercepto de cada Y_r con las omponentes *T*, permiten obtener los R^2 a ser promediados:

$$Y_1 \sim T \ : R^2 = 0.1797$$

$$Y_2 \sim T \ : R^2 = 0.36885$$

$$Y_3 \sim T \ : R^2 = 0.13186$$

$$Y_4 \sim T \ : R^2 = 0.01190 \ .$$

Redundancia:

Rd(Yc;t1,t2)= (0.1797+...+0.0119)/4 = 0.1730851

El modelo de regresión gnmPLS2 de cada variable respuesta, se obtiene mediante los coeficientes B.H asociados a las variables explicativas en *X* estandarizadas; estos están dados a continuación:

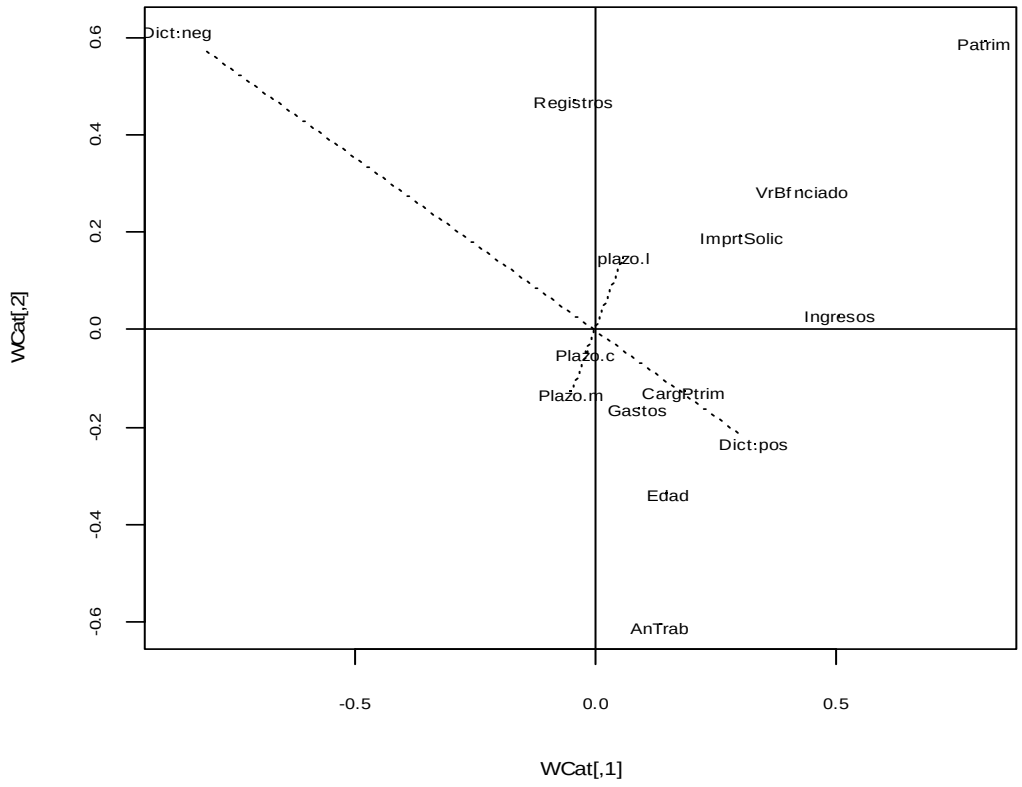
```
B.H
      ImprtSolic   VrBfnciado      Dictam        Plazo

[1,] -0.074599750 -0.114657305  0.16970356 -0.026534447
[2,] -0.018424124 -0.031126927  0.11047208 -0.009243984
[3,] -0.003549026 -0.007453252  0.05684019 -0.003176246
[4,]  0.160067294  0.226804310  0.10471239  0.038534350
[5,]  0.359622182  0.517542778  0.04046114  0.094219802
[6,]  0.076913017  0.115896721 -0.11845231  0.025139329
[7,]  0.031664951  0.042849245  0.06995181  0.005690609
```

Así, el modelo *gnmPLS2* estimado para Y_3 = *Dictam* asociando un $R^2 = 0.13186$ es de la forma:

$$\hat{Y}_3 = 0.1697 AnTrab + 0.1104 Edad + 0.0568 Gastos + 0.1047\ Ingresos + 0.0404\ Patrim - 0.1184\ Registros + 0.0699\ CargPtrim$$

En la gráfica 5.2 el dictamen positivo (*Dict.pos*) a la solicitud de crédito favoreció aquellos clientes con altos *Ingresos*, se prefiere en la aprobación clientes con mayor antigüedad laboral (*AnTrab*) y que no presenten *Registros*; mientras que se le niega a aquellos que presentan registros o poca antigüedad laboral.

Así mismo, la decisión de aprobación parece independiente del plazo (tiempo) de cancelación.



Gráfica 5.2. Carta de las variables *w1. , w2. y c1, c2*

- **Comparación usando indicadoras en Y e X**

Se procede ahora a desarrollar una regresión PLS2 tomando las indicadoras asociadas a las variables cualitativas tanto en *Y* como en *X*. Esto nos permitirá comparar con el análisis anterior (proceso de cuantificación) especialmente el Análisis de Redundancias y los gráficos de descripción en el primer plano factorial.

Se proporcionan las matrices *Y e X* así ampliadas y *H=2* como parámetros de entrada de la función fPLS2(), para obtener la matriz de componentes *T.h*. El promedio de los R^2 derivados de las regresiones múltiples sin intercepto de cada variable respuesta con las componentes *T.h*, permiten el cálculo de la redundancia; esto es:

$$Y_1 \sim T.h \ : \ R^2 = 0.1699495$$

$$Y_2 \sim T.h \ : \ R^2 = 0.3456674$$

$$Y_3 \sim T.h \ : \ R^2 = 0.1511729$$

$$Y_4 \sim T.h \ : \ R^2 = 0.1511729$$

$$Y_5 \sim T.h \ : \ R^2 = 0.01428954$$

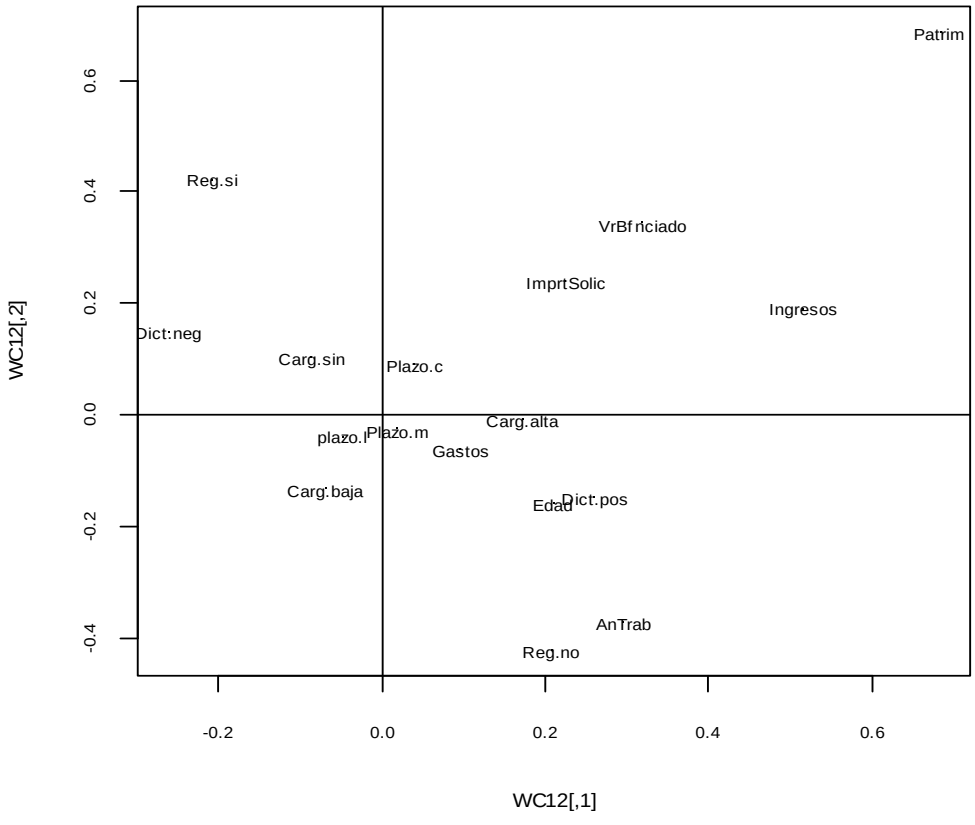
$$Y_6 \sim T.h \ : \ R^2 = 0.001491217$$

$$Y_7 \sim T.h \ : \ R^2 = 0.006113028.$$

Redundancia:

Rd(Y;t1,...,t7)=(0.1699495+...+0.006113028)/7 = 0.1199 = **0.12**

Se evidencia que el poder explicativo de las dos componentes disminuyó de 0.173 a 0.12 y será más severa en la medida que crezca el número de variables cualitativas y se desarrollen los análisis de regresión involucrando sus indicadoras.



Gráfica 5.3. Carta de las variables e indicadoras mediante (*w1., c1) ,(w2 , c2)*

La gráfica 5.3 con las variables dummy muestra que la aprobación a las solicitudes de crédito está muy influenciada por la Antigüedad en el trabajo y el no registro de los clientes. Sin embargo esta gráfica no provee suficiente información de las variables "Carga patrimonial" y "Plazo" como tal.

5.3. GNM-PLS1

Recuerde que GNM-PLS1 es un caso particular de GNM-PLS2 con *r* =1, por tanto los máximos obtenidos en la sección 5.2.1 son igualmente válidos, especialmente para el caso en el cual la variable *Y* es cuantitativa. En este caso, se tiene equivalencia con NM-PLS1 (Russolillo 2009).

Aquí ya no se requiere la función agregada t_x puesto que la única variable respuesta *y* es cuantitativa. Así, $u_y = y$ permite cuantificar en un único paso las variables cualitativas $\hat{x}_q$, maximizando de acuerdo con [5.4] $\sum_p cor^2(\hat{x}_p, y)$; mientras que globalmente en el cálculo de *H* componentes se maximiza el criterio $\sum_{h=1}^{H} cov^2(t_h, y)$.

Cuando *y* es cualitativa, el procedimiento que se podría denominar GNM-PLS1q es más interesante pero menos natural ya que las *H* componentes ortogonales de inicio para cuantificar *y* se obtienen del algoritmo NIPALS y no de la regresión PLS-R, aunque después de esta primera cuantificación si toma los recursos de la regresión invocando reiteradamente la función f.PLS1. Se presenta esta propuesta de análisis a continuación.

5.3.1. GNM-PLS1q

Se asume como caso de interés que la única variable *Y* es cualitativa mientras que en *X* se encuentran las submatrices X_q con q variables cualitativas y X_k con k variables cuantitativas. El algoritmo descansa en el PLSR y es denominado GNM-PLS1q(), y requiere para su inicio determinar las componentes ortogonales *t* para la cuantificación de *Y*, cuyo número está determinado por el rango *a* de la matriz cuantitativa X_k .

Una forma de encontrar la matriz *T* conteniendo las componentes ortogonales, consiste en obtener $H \leq a$ componentes t_h de la descomposición singular vía NIPALS de la matriz X_k , y proceder luego a conformar una cuantificación agregada inicial Y_{qc} bajo cargas arbitrarias $c_h = 1 \ \forall \ h$.

Para obtener Y_{qc} sin orden fcso() o con orden fcco(), se implementan adecuadamente en cada $h = 1,2,\dots,H$ las ecuaciónes [4.1] y [4.3] del capítulo NM-NIPALS, ver anexo B.

Se estandariza la cuantificacion y se recalculan las cargas c_h como los coeficientes de regresión de $YIt \sim T$. Una vez obtenidas las c_h como las correlaciones entre t_h *e* Y_{qt} se forma la función de cuantificación con $h\ (\leq a)$ componentes de la forma $t_x = c_1 t_1 + \cdots + c_h t_h$ y se entra al ciclo de la convergencia dentro del cual se recuantifica Y_q y con esta a su vez se cuantifican las variables cualitativas en X_q realizando secuencial y ciclicamente las siguientes actividades:

- Con t_x se cuantifica $Y : Y_{qc} = Y_{qt}$ normalizada
- Con Y_{qt} se cuantifica las $x_q : \hat{x}_q$ normalizadas
- Conformar la matriz ampliada $X = X_k | \hat{X}_q$
- Ejecutar fPLS1(*Y,X,a*) ; ouput: T_h, C_h tal que $t_x = T_h C_h$

El *pseudo-algoritmo* asociado a gnmPLS1q tiene la forma siguiente:

Entrada: Y_q, X_k, X_q, h (componentes *cuantificación*, $h \leq a$)
Salida: *Y,X,wH, cH, tH, bH, R2h, R2vch, Q2h*

Inicializa $t_1, t_2, \dots, t_h$ vía NIPALS(X_k)
con $c_h = 1\ \forall\ h$ se obtiene Y_{qt}
Conformar $t_x = c_1 t_1 + c_2 t_2 + \cdots + c_h t_h$, con $c_h = r(Y_{qt}, t_h)/St_h$
Repita
f.c(tx,y_q) = Yqt cuantificación estandarizada
f.c(Yqt, x_q) = $\hat{x}_q$ cuantificación etandarizada de las X_q
$X = X_k | \hat{X}_q$, $Y = Y_{qt}$ matrices ampliadas

f.PLS1(*Y,X,a*) : *Th, Ch ;* $t_h = X_{h-1} w_h$
maximización : $cov^2(t_h, Y_{h-1}) = S2ty$
$t_x = T_h C_h$
Hasta convergencia w_h

f.vc(Y,X,h) : *R2h, R2vch, PRESS* con estos índices de validación Cruzada
se obtienen H compnts para *regression-cuntificacion*
retorne: *Y,X, wH, tH, cH, bH, S2ty, R2h, R2vch, PRESS*
Fin.

Ver algoritmos f.PLS1, fgnmPLS1q, fvc1 bajo R en anexo B; donde $Y_{qt} = YI_t$.

En este proceso de cuantificación, todas las variables son estandarizadas (incluyendo las cualitativas luego de ser cuantificadas), lo cual induce la propiedad de *cuantificación invariante* para el caso de variables cualitativas con sólo dos categorías tanto en *Xq* como en *Yq*. En este caso, independientemente de la dimensión de la función de cuantificación $t_{xh} = c_1 t_1 + \cdots + c_a t_a$, el valor que toma cada categoría no cambia de una dimensión a otra.

Propiedad. Cuantificacion invariante: Si Y_q está conformada por sólo dos categorías, conteniendo *g* valores de la categoría "A" y *n-g* valores de la categoría "B", el valor de la cuantificación para "A" es

$$\sqrt{n-1}/\sqrt{\frac{(n-g)^2}{g} + \frac{n-g}{1}} = \sqrt{n-1}/\sqrt{(n-g)n/g}$$

mientras que para la categoría "B" es

$$-(n-g)\sqrt{n-1}/g\sqrt{\frac{(n-g)^2}{g} + \frac{n-g}{1}} \ .$$

5.3.2. Aplicación GNM-PLS1q

El bloque de las predictoras cuantitativas X_k está conformado por cinco variables cuantitativas: `Antrab` *(años de trabajo)*, `Edad`, `Gastos`, `Ingresos`, `Patrim` (*patrimonio*) y la sub-matriz X_q conteniendo las predictoras cualitativas `Registros`, `CargPtrim` (*carga patrimonial*). El segundo bloque lo constituye la matriz Y_q conteniendo una unica variable endógena cualitativa: *dictamen* notada `Dictam`.

Variables Cualitativas	*Descripción*	*Categorías*					
		1	2	3	4	5	6
Dictam : *Yq*	Dictamen	positivo	negativo				
EstCivil	Estado civil	soltero	casado	viudo	separado	divorciado	
Registros	Datacredito	no	si				
TipTrab	Tipo Trabajo	empleado	temporal	autónomo	otros		
Vvda	Tipo vivienda	alquiler	escritura	contrato	ignora	padres	otros
CargPtrim	Carga patrimonial	sin	baja	alta			

Tabla 5.3. Categorías asociadas a las variables cualitativas en GNM-PLS1q

La variable cualitativa de respuesta *Yq* en este caso *Dictamen*, y las demás variables exógenas cualitativas presentan las categorías dadas en la tabla 5.3.

El objetivo es cuantificar óptimamente la variable dictamen *Yq* y las dos variables cualitativas en *Xq* vía PLS1, encontrando el modelo de regresión PLS1 que mejor explique el *dictamen* debidamente cuantificado, en función de todas las 7 variables predictoras, incluyendo las cuantificadas.

El procedimiento *gnm PLS1* con variable respuesta cualitativa es realizado bajo R, a través de la función *fgnmPLS1q*(X_k, *Xq,Yq, a*) donde *a* es el rango de la matriz X_k y limita el número de componentes en la regresión-cuantificación; las variables exógenas cuantitativas se mantendrán en su forma original (excepto por su estandarización)

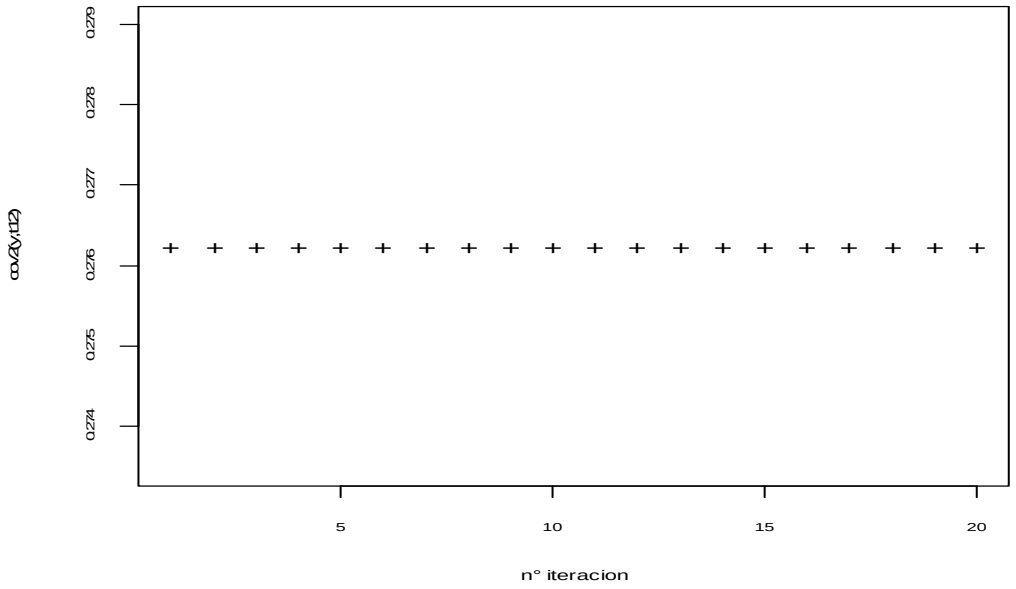
Respecto a la propiedad de cuantificación invariante, en este ejemplo de crédito la variable *Dictam* contiene *g=317* categorías "si" y *n-g* = 121 "no" fue cuantificada con los valores *Yq* : [0.6171, 0.6171, 0.6171, 0.6171, -1.6167, ...] excepto tal vez por signo debido al centramiento.

Este mismo razonamiento aplica para la variable explicativa *Registros* con dos categorías; sin embargo *Carga patrimonial* también se torna invariante ya que es cuantificada por *Dictam* que a su vez es invariante.

Los resultados derivados del proceso de cuantificación se resumen en la siguiente tabla:

Variables Cualitativas	*Descripción*	*Categorias - cuantificadas*		
		1	2	3
Dictam : *Yq*	Dictamen	Positivo 0.6171	Negativo -1.6167	
Registros	Datacredito	No 0.4467	Si -2.2335	
CargPtrim	Carga patrimonial	sin -0.3682	baja -0.1857	alta 2.8434

Observe que la variable CargPtrim contiene orden en la cuantificación. Simultáneamente al proceso de cuantificación se calcula la función de maximización $\sum_h cov^2(Y, t_h)$ con *h*=1,2 en cada iteración, conllevando como consecuencia de la cuantificación invariante a que esta función sea constante con valor 0.2762. Ver gráfica 5.4 para las primeras 20 iteraciones.



Gráfica 5.4. Función de maximización asociada a la cuantificación de *Dictam* via gnm PLS1q

El modelo PLS1 para *Dictam* cuantificada respecto a las 7 variables explicativas arroja los siguientes índices (*h* componentes en la regresión (fila)-cuantificación):

R^2h

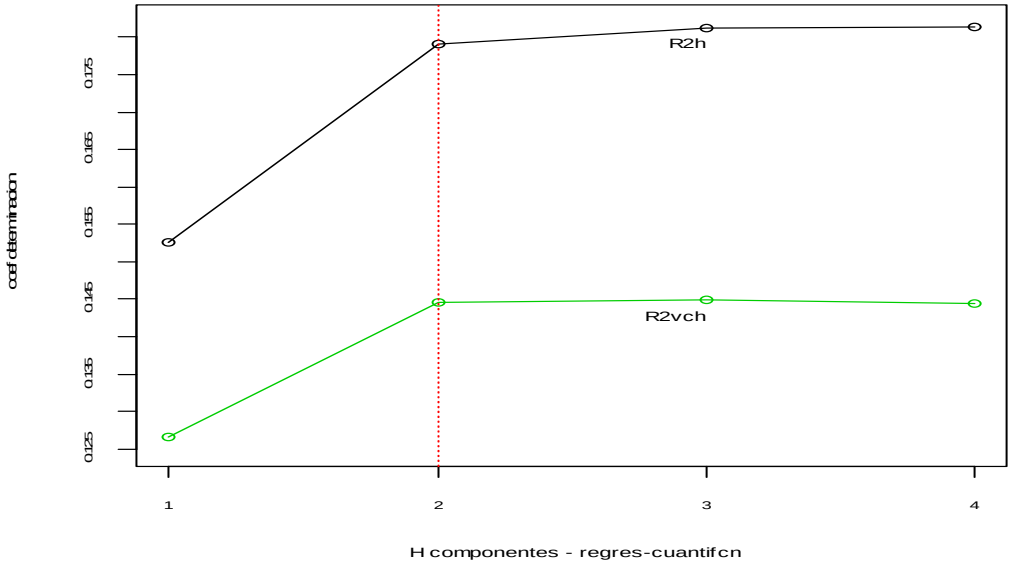
	[,1]	[,2]	[,3]	[,4]
[1,]	0.1526	0.1526	0.1526	0.1526
[2,]	0.1791	0.1791	0.1791	0.1791
[3,]	0.1812	0.1812	0.1812	0.1812
[4,]	0.1813	0.1813	0.1813	0.1813

R^2vch

	[,1]	[,2]	[,3]	[,4]
[1,]	0.1266	0.1266	0.1266	0.1266
[2,]	0.1445	0.1445	0.1445	0.1445
[3,]	0.1449	0.1449	0.1449	0.1449
[4,]	0.1444	0.1444	0.1444	0.1444

PRESSh

	[,1]	[,2]	[,3]	[,4]
[1,]	381.7	381.7	381.7	381.7
[2,]	373.9	373.9	373.9	373.9
[3,]	373.7	373.7	373.7	373.7
[4,]	373.9	373.9	373.9	373.9



Gráfica 5.5. Comportamiento coeficientes R^2h y R^2vch

Observe que tanto el coeficiente de determinación R^2h como el de cros-validación R^2vch son invariantes al número *a* de componentes en la cuantificación. También se aprecia que tomando $H=2$ componentes en la regresión PLS1, se tiene un valor R^2vch= 0.1445 a partir del cual la ganancia en explicación con más componentes no es importante; con dos componentes asocia un R^2h de 0.1791 y un error de predicción PRESSh = 373.9 a partir del cual la disminución no es significativa. (Ver gráfica 5.5).

Así, la matriz *X* cuantificada estandarizada bajo el modelo gnmPLS1 asociado a *Y=Dictam* con $H=2$ componentes presenta la siguiente estructura para las 15 primeras observaciones:

	Dictam	AnTrab	Edad	Gastos	Ingresos	Patrim	Registros	CargPtrim
1	0.6171	-0.8472	-0.87019	-0.03249	-0.3904	-0.5049	0.4467	-0.3682
2	0.6171	-0.9704	2.23970	-1.05619	-0.2339	-0.2098	0.4467	-0.3682
3	0.6171	-0.6009	-0.59579	-1.05619	0.2115	-0.5049	0.4467	-0.3682
4	0.6171	-0.8472	-1.14460	-1.05619	-0.8719	-0.5049	0.4467	-0.3682
5	-1.6167	-0.6009	-1.05313	-1.05619	-1.5941	-0.5049	-2.2335	-0.3682
6	0.6171	-0.1083	-0.59579	-1.05619	-0.4506	-0.5049	0.4467	-0.3682
7	0.6171	1.4929	-0.04699	0.22344	-0.1376	-0.5049	0.4467	-0.3682
8	-1.6167	-0.3546	-1.51047	-0.54434	-1.5941	-0.5049	0.4467	-0.3682
9	0.6171	-0.9704	-1.51047	-1.05619	0.2717	0.1837	0.4467	-0.3682
10	0.6171	2.7245	1.87383	-0.54434	-1.1126	1.1673	0.4467	-0.3682
11	0.6171	-0.6009	-0.41286	0.06988	-0.1376	-0.5049	0.4467	-0.3682
12	-1.6167	-0.7241	-1.14460	-1.05619	-0.8117	-0.5049	0.4467	-0.3682

```
13    0.6171 -0.7241  0.13595 -0.54434  -0.2219 -0.1508    0.4467   -0.3682
14    0.6171  0.6307  0.50182  0.22344  -0.3061 -0.1114    0.4467   -0.3682
15    0.6171  1.2465 -0.13845  0.99121   0.5606  0.1837    0.4467    2.8434
 :
```

De otro lado, el modelo de regresión gnmPLS1 retorna los coeficientes b_h asociados a las predictoras, tal que el modelo obtenido con $h = 2$ (col) componentes es:

```
bh
        [,1]      [,2]      [,3]     [,4]
[1,] 0.17569  0.217864  0.249369  0.25650
[2,] 0.08506  0.001632 -0.027228 -0.03194
[3,] 0.02647 -0.058583 -0.062195 -0.05728
[4,] 0.13581  0.174387  0.200613  0.19370
[5,] 0.09635  0.071008  0.066131  0.06663
[6,] 0.15744  0.261359  0.233200  0.23373
[7,] 0.04255  0.016812  0.009483  0.00692
```

$\widehat{Dictam}$ = 0.2179*AnTrab + 0.0016*Edad - 0.0586*Gastos + 0.1744*Ingresos + 0.0710*Patrim + 0.2614*Registros + 0.0168*CargPtrim

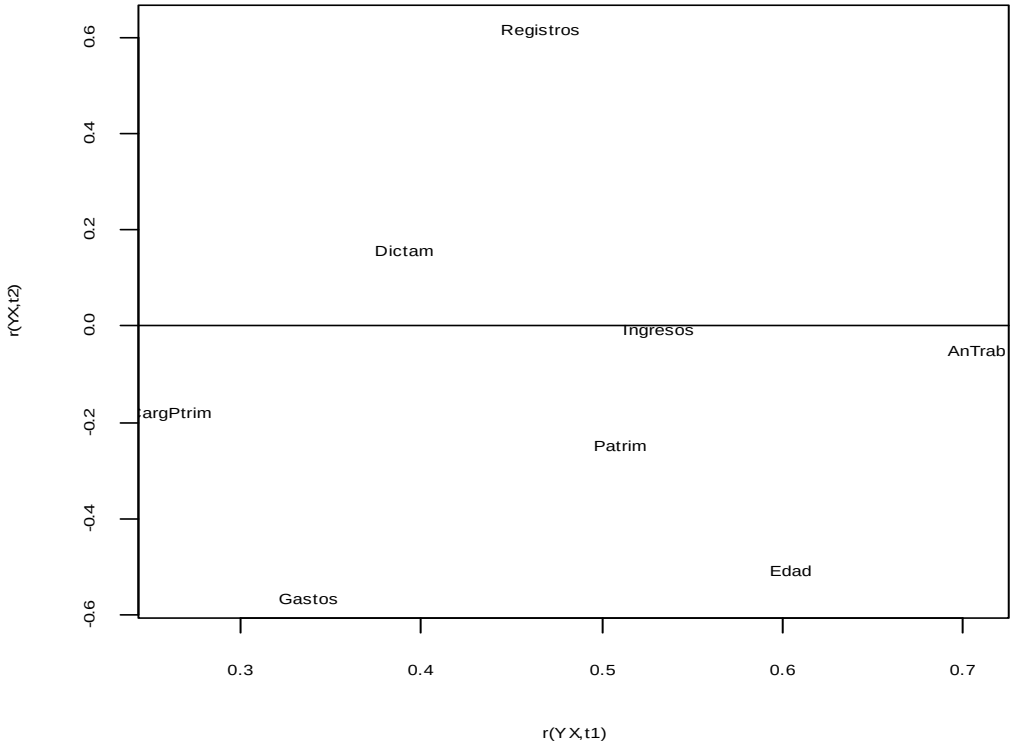
Este modelo estimado presenta un R^2= 0.179

- **Interpretación :**

Si un individuo presenta características similares al cliente 10, el modelo estima $\widehat{Dictam} = 0.62793$, y el crédito le será aprobado, ya que está más cerca de la categoría "positivo" cuantificada con 0.6171.

Las correlaciones de Y (dictamen) e X con las primeras dos componentes permitirán comprender mejor el sentido de las relaciones obtenidas en la cuantificación, asociando la gráfica 5.6:

```
cor(YX,th)
            [t1]      [t2]
Dictam    0.3907  0.162571
AnTrab    0.7073 -0.046380
Edad      0.6049 -0.503724
Gastos    0.3375 -0.558617
Ingresos  0.5317 -0.005332
Patrim    0.5104 -0.243504
Registros 0.4668  0.619856
CargPtrim 0.2621 -0.179132
```



Gráfica 5.6. Correlaciones de las variables *Y, X* con *t1, t2*

En la gráfica 5.6 y de acuerdo con el modelo gnmPLS1, la variable respuesta dictamen (*Dictam)* presenta correlación positiva con *Registros(no), Ingresos y AnTrab*, características que determinan prácticamente la aprobación. De otro lado, el crédito le es negado especialmente a aquellos clientes que presentan *Registros(si)* y altos *Gastos*.

CAPÍTULO 6

PLS-PM CON DATOS MIXTOS: NM- PLSPM

Los principios básicos del algoritmo NM-PLSR pueden ser extendidos al NM-PLSPM con dos bloques ya que el algoritmo es el mismo excepto por las restricciones de normalización. En PLS-PM Y_2 puede ser interpretada como la estimación *outer* del LC asociado con el bloque X_2, así como la estimación *inner* del LC subyacente a X_1 por medio del cual los pesos *outer* $a_1 = X_1'Y_2/\|X_1'Y_2\|$ son calculados. Simétricamente, Y_1 puede ser considerada simultáneamente como el estimador *outer* del LC subyacente a X_1 y el estimador *inner* del LC subyacente a X_2. Este es un paso oculto en el algoritmo PLS-R, el cual en términos del PLS-PM puede ser interpretado como: el estimador *outer* en un bloque es usado como el estimador *inner* en el otro bloque.

Esta doble función está justificada por las relaciónes *inner* $Z_2\ \alpha\ Y_1$, $Z_1\ \alpha\ Y_2$ que en terminos NM-PLSR nos permite obtener variables cuantificadas maximizando su correlación con el estimador *inner* de la correspondiente LV; esto es,

$$\max\{cor(x_1^*, Y_2)\} \equiv \max\{cor(x_1^*, Z_1)\}$$

$$\max\{cor(x_2^*, Y_1)\} \equiv \max\{cor(x_2^*, Z_2)\}$$

Así, se puede establecer que en el modelo NM-PLSR, interpretado como un NM-PLSPM de dos bloques, cualquier variable cuantificada es computada como una función del estimador *inner* del correspondiente LC. Por tanto, extendiendo esta idea a más de dos bloques, las componentes *inner* Z_j son las funciones de cuantificación agregadas de dimensión *J* en cada *j-ésimo* bloque, con lo cual estos procesos se podrían denominar JNM-PLSPM siendo un caso particular del hipotético GNM- PLSPM.

6.1. ALGORITMO NM-PLSPM

El ciclo del algoritmo NM-PLSPM difiere del ciclo estándar del PLS-PM en el hecho de que inicia con las estimaciones *inner* de cada LV, para obtener una primera cuantificación de las MV, de acuerdo con Russolillo (2009).

Las componentes *inner* Z_j pueden ser estimadas inicialmente aplicando la funcion PLS-PM a las J submatrices cuantitativas de los correspondientes bloques. Luego se obtienen los pesos *outer* w_j según el modo de estimación A o B, para calcular las componentes *outer* Y_j.

A su vez, las componentes *outer* permiten computar los pesos *inner* como función de las componentes *outer*, $e_{j,j'} = f(Y_j, Y_{j'})$, de acuerdo al esquema escogido. Finalmente se recalculan las componentes *inner* y el ciclo itera hasta la convergencia de los pesos *outer* w_j .

Pseudo algoritmo NM-PLSPM

Entradas: $X_1, \dots X_J; C, Esq, modo$
Salidas: $\hat{X}_1, \dots, \hat{X}_J; Y_j, Z_j, w_j, \beta_j$
Inicio $Z_j \; \forall \; j$

1. Iteración ($\forall \; j$)

1.1. Cuantificación
$\hat{X}_j = q_i(x_j^*, Z_j)$

1.2. Estimación pesos *outer*
$$w_j = \begin{cases} \hat{X}_j' Z_j & modo\ A \\ (\hat{X}_j' X_j)^{-1} \hat{X}_j' Z_j & modo\ B \end{cases}$$

1.3. Estimación componentes *outer*
$Y_j = \hat{X}_j w_j$

1.4. Computación pesos *inner*
$e_{j,j'} = f(Y_j, Y_{j'})$ según esquema

1.5. Componentes *inner*
$Z_j = \sum_{j'}^{J} c_{jj'} e_{jj'} Y_{j'}$

Hasta convergencia de w_j

2. Coeficientes path
$\beta_j = (\hat{\Xi}_j' \hat{\Xi}_j)^{-1} \hat{\Xi}_j' Y_j$

Fin

6.2. CRITERIO DE OPTIMIZACIÓN: RGCCA

Ya que el criterio para el cual el modo A del algoritmo PLS-PM ($J > 2$) converge es desconocido, el modo A del NM PLS-PM sufrirá los mismos inconvenientes. Sin embargo este problema desaparece si el *nuevo modo* A es usado, es decir, implementando un RGCCA (Tenenhaus et al. 2011) con parámetro de regularización $\tau_j = 1$ para todos los bloques. Con el *nuevo modo* A las estimaciones *outer* Y_j no son restringidas a varianza unitaria mientras que los pesos a_j *outer* si son restringidos a norma 1. (Ver fase C gráfica 6.1).

De acuerdo con [3.27], se demostró que el *nuevo modo A* PLS-PM, cuando los esquemas *factorial* y *centroide* son usados, converge monótonamente al criterio:

$$\underset{a_j}{arg\ max} \sum_{j \neq k}^{J} c_{jk}\, g\big(Cov(X_j a_j, X_k a_k)\big) \quad bajo \ \ \|a_j\| = 1$$

que puede ser escrito de acuerdo con [3.23] como

$$\underset{\forall a_j}{arg\ max} \sum_{j}^{J} Cov(X_j a_j, Z_j) \quad bajo \ \ \|a_j\| = 1 \ .$$

En la versión no métrica del *nuevo modo A* en PLS-PM (NM-RGCCA) la función de cuantificación $\hat{x}_{pj} \propto \mathcal{Q}(\tilde{X}_{pj}, Z_j^s)$ de las variables cualitativas del grupo j es agregada al final de la fase B en el paso B1 como se muestra en la gráfica 6.1, cuya consideración particular en la ecuación anterior conlleva el siguiente criterio:

$$\underset{\forall a_j}{arg\ max} \sum_{j}^{J} Cov\big(\hat{X}_j a_j, Z_j\big) = \sum_j \tfrac{1}{n} a_j' \hat{X}_j' Z_j$$

$$= \tfrac{1}{n} \sum_j Z_j' \hat{X}_j \hat{X}_j' Z_j / \big\|\hat{X}_j' Z_j\big\|$$

$$= \tfrac{1}{n} \sum_j \frac{\sum_p (n.cov(\hat{x}_{pj}, z_j))^2}{\sqrt{\sum_p (n.cov(\hat{x}_{pj}, z_j))^2}}$$

$$= \sum_j \sqrt{\sum_p cov^2(\hat{x}_{pj}, z_j)}$$

bajo las restricciones $\|a_j\| = 1$ y $var(\hat{x}_{pj}) = 1$.

Observe que $\hat{X}_j$ contiene tanto las variables cuantitativas 'originales' como las 'cualitativas' cuantificadas debidamente estandarizadas. Se maximizan p_j criterios, cada uno de los cuales es función de la variable escalada; por tanto este puede ser optimizado maximizando separadamente:

$$cov^2(\hat{x}_{pj}, z_j) \ .$$

Es decir, se maximiza la covarianza al cuadrado de cada variable del bloque *j* con su correspondiente componente *inner* $Z_j \ \forall j$, por medio de la función de cuantificación $Q(\tilde{X}_{pj}, Z_j)$. Esto implica que cuando el nuevo modo A es usado, optimiza el criterio del modelo con respecto a los parámetros de cuantificación para parámetros fijos del modelo (restringidos a norma 1).

A. Inicio

A1. $X_j = Xc_j$ cada bloque *j* contiene *inicialmente* sólo las variables cuantitativas

Ejecute la fucncion *fRGCCAu*(Xc_j, … ,)

A2. Determine los vectores de pesos *outer* normalizados $a_1^0 , a_2^0 , \dots , a_j^0$ como

$$a_j^0 = \left[\left(\tilde{a}_j^0\right)^t \left[\tau_j I + (1-\tau_j) \frac{1}{n} X_j^t X_j \right]^{-1} \tilde{a}_j^0 \right]^{-1/2} \left[\tau_j I + (1-\tau_j) \frac{1}{n} X_j^t X_j \right]^{-1} \tilde{a}_j^0 \ .$$

Para $s = 0, 1, \dots$ (hasta convergencia de todos los a_j)

Para $j = 1, 2, \dots , J$

B. Computo de la componente inner Z_j

Compute las componentes *inner* de acuerdo al esquema seleccionado:

$$z_j^s = \sum_{k<j} c_{jk} w \left[Cov\left(X_j a_j^s , X_k a_k^{s+1}\right) \right] X_k a_k^{s+1} + \sum_{k>j} c_{jk} w \left[Cov\left(X_j a_j^s , X_k a_k^s\right) \right] X_k a_k^s$$

donde $w(x) = 1$ para el esquema Horst, x para el esquema factorial y $signo(x)$ para el esquema centroide.

B1. Cuantificacion (*q-ésima variable cualitativa* ***)***

$$\hat{x}_{q_j} \propto Q\left(\tilde{X}q_j, z_j^s\right) \quad \forall q \;\; ; \;\; X_j = (Xc_j|\hat{X}q_j) \;\; .$$

C. Maximizacion

Implementación *fRGCCAu(X_j, …,)* y obtención componentes outer Y_j
Funciones a maximizar según esquema y modo:

aS_{ij} : valor absoluto de la covarianza entre Y_i, Y_j
aR_{ij} : valor absoluto de la correlación entre Y_i, Y_j
Sij^2 : cuadrado de la covarianza entre Y_i, Y_j
Rij^2 : cuadrado de la correlacion entre Y_i, Y_j

Recalcule el vector de pesos outer para cada bloque X_j

Actualizacion del vector de pesos *outer*

$$a_j^{s+1} = \left[(z_j^s)^t X_j \left[\tau_j I + (1-\tau_j)\frac{1}{n}X_j^t X_j\right]^{-1} X_j^t z_j^s\right]^{-1/2} \times \left[\tau_j I + (1-\tau_j)\frac{1}{n}X_j^t X_j\right]^{-1} X_j^t z_j^s$$

Fin

Fin

Gráfica 6.1. Algoritmo NM-RGCCA

6.3. ALGORITMO NM-RGCCA

Ya que se tiene datos mixtos en cada bloque *j*, se obtiene de los datos cuantitativos en $X_{j.}$ y mediante la función *fRGCCAu(X_j, …,)* las componentes *inner* Z_j y *outer* Y_j, que servirán de inicio al proceso de convergencia para la cuantificación y maximización respectivamente. Una vez cuantificadas (simultáneamente) y por primera vez las variables cualitativas, éstas se yuxtaponen en cada uno de los bloques, ahora denominados $\hat{X}_j$ con los cuales se obtienen nuevas componentes *inner* y *outer* de *fRGCCAu($\hat{X}_{j.}$, … ,)* para recuantificar una vez más; el proceso continua iterando hasta la convergencia de los pesos *outer* w_j.

El anexo D corresponde al programa *fnmRGCCA.R* elaborado bajo el entorno R, el cual contiene la funciones y procedimientos básicos invocando a su vez y en cada iteración el programa *fRGCCAu.R* que suministra las componentes *inner* y *outer* en este proceso de cuantificación.

Asocia además, una función de maximización de covarianzas o correlaciones de las Y_j conectadas (Tenenhaus and Tenenhaus, 2011) de acuerdo con el modo de relación A o B y el esquema (*centroide, factorial*) de estimación de los pesos, cuyas tendencias crecientes y convergentes pueden ser graficadas.

6.4. EJEMPLO DE APLICACIÓN

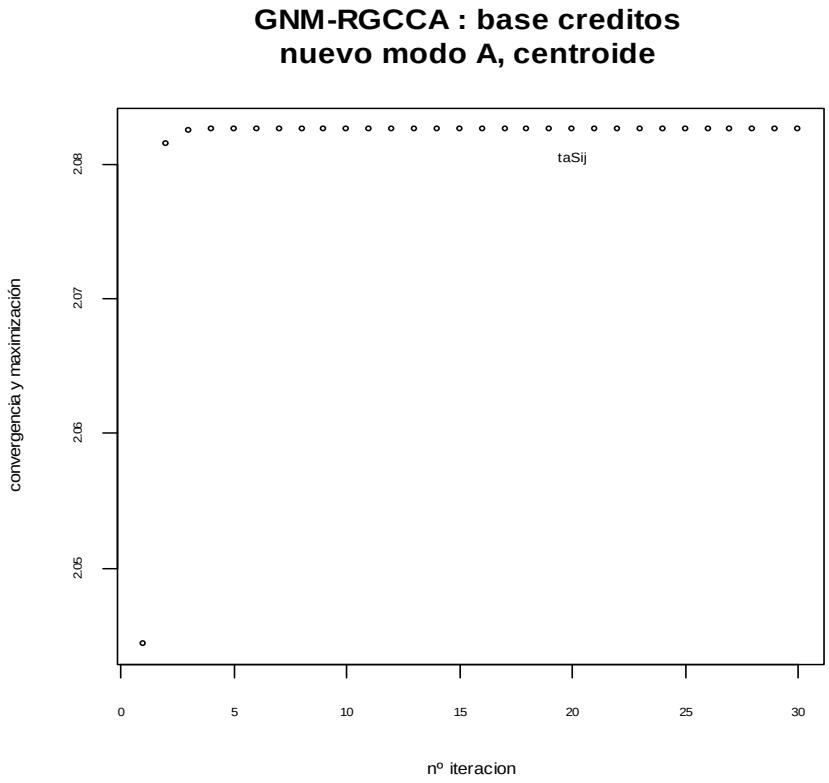
Se retoma la base de datos de *créd.438* del apartado 5.2.3. llevándola a la forma *k*-tablas ($X.k$) la cual contiene la parte cuantitativa (estandarizada) de cada uno de tres grupos. El número de variables cuantitativas en cada grupo se asocia en el vector `pX=c(2,3,2)`. Así, las cuantitativas del primer grupo son *AnTrab* y *Edad,* del segundo grupo *Gastos, Ingresos* y *Patrim;* y las del tercer grupo son *ImprtSolic*, *VrBfnciado*.

Las variables cualitativas asociadas con cada grupo se encuentran en el objeto *Q* tipo lista, y su posible orden en *oQ*. El primer grupo se refiere a la capacidad socio-económica "capSEC" asociando las variables *Vivienda, Estado civil* y *Tipo de trabajo*; el segundo describe la situación de las finanzas "sitFnza" mediante las variables *Registros* y C*arga patrimonial*; y el tercero a las características del crédito "carCred" conteniendo las variables *Dictamen* y *Plazo*; los bloques uno y dos esta conectados entre sí, y además cada uno conecta con el grupo 3 que hace las veces de grupo de respuesta o grupo Y.

Luego de conformar las bases para el análisis, se establecen las relaciones *inner* y se determinan los parámetros de entrada para implementar la función *fnmRGCCA()* de acuerdo con el anexo D.

La mejor respuesta al proceso de cuantificación y maximización se dio, bajo el *nuevo modo A* y esquema *centroide.*

Los resultados obtenidos se describen a continuación:



Gráfica 6.2. Maximización del valor absoluto de las covarianzas *outer*

De acuerdo con Tenenhaus and Tenenhaus (2011), la función a maximizar bajo estas características es $taSij = \sum_{j,k,j\neq k} Cov(X_j a_j, X_k a_k)$, la cual se obtiene en cada iteración formando la serie:

```
taSij

2.044401  2.081559  2.082557  2.082597  2.082599  2.082599  2.082599
2.082599  2.082599  2.082599  2.082599  2.082599  2.082599  2.082599
2.082599   ...
```

cuya gráfica 6.2 muestra la tendencia creciente y convergente de la misma.

La matriz (tipo *k* tablas) *Xc* conteniendo las variables cuantitativas y las cualitativas cuantificadas por grupos, se describe a continuación para los primeros 15 registros:

X1 grupo1 cuantificado(Vivienda, Estado Civil, Tipo de Trabajo)

	AnTrab	Edad	Vvienda	EstCivil	TipTrab
1917	-0.8472367	-0.8701936	-0.6194066	0.6144445	0.1105492
2980	-0.9703996	2.2396992	1.0197741	0.6144445	-1.0838542
928	-0.6009110	-0.5957913	-1.0748421	-1.6090718	1.1472060
2112	-0.8472367	-1.1445959	-1.3670905	-1.6090718	0.1105492
3658	-0.6009110	-1.0531285	-0.6194066	0.6144445	1.1472060
2617	-0.1082596	-0.5957913	-1.3670905	-1.6090718	0.1105492
388	1.4928575	-0.0469867	-1.3670905	0.6144445	0.1105492
3952	-0.3545853	-1.5104656	-1.3670905	-1.6090718	-1.0838542
3841	-0.9703996	-1.5104656	-1.3670905	-1.6090718	0.1105492
1639	2.7244860	1.8738294	1.0197741	0.6144445	1.1472060
2326	-0.6009110	-0.4128564	-0.6194066	0.6144445	0.1105492
2945	-0.7240739	-1.1445959	-1.3670905	-1.6090718	0.1105492
1038	-0.7240739	0.1359482	1.0197741	0.6144445	0.1105492
999	0.6307175	0.5018179	1.0197741	0.6144445	0.1105492
2206	1.2465318	-0.1384541	1.0197741	0.6144445	0.1105492

:

X2 grupo2 cuantificado (Registros,Carga Patrimonial)

	Gastos	Ingresos	Patrim	Registros	CargPtrim
1917	-0.03248733	-0.39039325	-0.5048723	0.4467028	-0.4571491
2980	-1.05618867	-0.23390615	-0.2097840	0.4467028	-0.4571491
928	-1.05618867	0.21148019	-0.5048723	0.4467028	-0.4571491
2112	-1.05618867	-0.87189199	-0.5048723	0.4467028	-0.4571491
3658	-1.05618867	-1.59414011	-0.5048723	-2.2335139	-0.4571491
2617	-1.05618867	-0.45058059	-0.5048723	0.4467028	-0.4571491
388	0.22343801	-0.13760641	-0.5048723	0.4467028	-0.4571491
3952	-0.54433800	-1.59414011	-0.5048723	0.4467028	-0.4571491
3841	-1.05618867	0.27166753	0.1836671	0.4467028	-0.4571491
1639	-0.54433800	-1.11264137	1.1672949	0.4467028	-0.4571491
2326	0.06988281	-0.13760641	-0.5048723	0.4467028	-0.4571491
2945	-1.05618867	-0.81170465	-0.5048723	0.4467028	-0.4571491
1038	-0.54433800	-0.22186869	-0.1507663	0.4467028	-0.4571491
999	0.22343801	-0.30613097	-0.1114212	0.4467028	-0.4571491
2206	0.99121402	0.56056678	0.1836671	0.4467028	2.6395132

:

X3 grupo3 cuantificado (Dictamen, Plazo)

	ImprtSolic	VrBfnciado	Dictam	Plazo
1917	-0.08002388	-0.20111041	0.6171159	0.6113082
2980	0.55121264	0.48977571	0.6171159	-1.1217099
928	-0.24835362	-0.04541776	0.6171159	1.4638141

```
2112 -0.08002388  0.28867271  0.6171159  0.6113082
3658  0.55121264 -0.10218070 -1.6167416  0.6113082
2617 -0.29043606 -0.43140578  0.6171159 -1.1217099
388  -0.39564214 -0.79306682  0.6171159 -1.1217099
3952  3.07615875  3.43496157 -1.6167416  0.6113082
3841  0.86683091  0.42814737  0.6171159  1.4638141
1639 -1.34249693 -1.49854911  0.6171159  1.4638141
2326  0.34080047  0.85630214  0.6171159  0.6113082
2945 -0.30095667 -0.85793875 -1.6167416  0.6113082
1038 -0.81646649 -0.52547008  0.6171159  0.6113082
999  -0.08002388 -0.10542430  0.6171159 -1.1217099
2206 -0.08002388  0.05026834  0.6171159 -1.1217099
 :
```

Q1: Categorías cuantificadas sin orden

Vvienda	alquiler	escritur	contrato	Ign.ctto	padres	otros
Cuantfcn	-0.61940	1.019774	0.04946	-0.84752	-1.3670	-1.074

EstCivil	soltero	casado	viudo	separado	divorciado
Cuantifcn	-1.609071	0.614444	-1.200889	-2.033247	-1.038589

TipTrab	empleado	e.temporal	autónomo	otros	
Cuantifcn	0.110549	-2.427553	1.1472060	-1.083854	

Q2 : Carga patrimonial (Cptrim) con orden

Registros	si	no			
Cuantifcn	-2.233513	0.4467028			

Cargptrim	sin	baja	alta		
Cuantifcn	-0.457149	0.865710	2.639513		

Q3:Yq , Plazo (orden - inverso)

Dictam	positivo	negativo			
Cuantifcn	0.6171159	-1.616741			

Plazo	corto	medio	largo		
Cuantifcn	1.463841	0.6113082	-1.121709		

Los pesos *wj* de las variables manifiestas asociados con su correspondiente componente *outer* Y_j son:

```
wj
            [[1]]
AnTrab    0.3610312
Edad      0.4012926
Vvienda   0.5440429
EstCivil  0.4651324
TipTrab   0.4430462

             [[2]]
Gastos    0.3985440
Ingresos  0.4793667
Patrim    0.7234320
Registros 0.1328105
CargPtrim 0.2652882

              [[3]]
ImprtSolic 0.4391587
VrBfnciado 0.6159320
Dictam     0.6437321
Plazo      0.1156567
```

De acuerdo con estos pesos y las cuantificaciones de los grupos se tiene las siguientes interpretaciones:

En el grupo 1, se tendrá mayor capacidad socioeconómica (Y_1) en la medida que se tenga más antigüedad en el trabajo, sea propietario de vivienda, trabaje como autónomo (o tenga un empleo fijo) y este casado.

Una "buena" situación financiera (Y_2) esta explicada en gran parte por un alto valor del patrimonio (Patrim) y de los ingresos, el potencial en gastos, un buen cubrimiento por carga patrimonial y el no presentar registro por incumplimiento crediticio.

Finalmente en el grupo 3 (Y_3) los "grandes" créditos solicitados (aprobados) se caracterizan por presentar un alto valor del bien financiado, un mayor importe solicitado, un plazo corto o mediano de cancelación y un dictamen positivo.

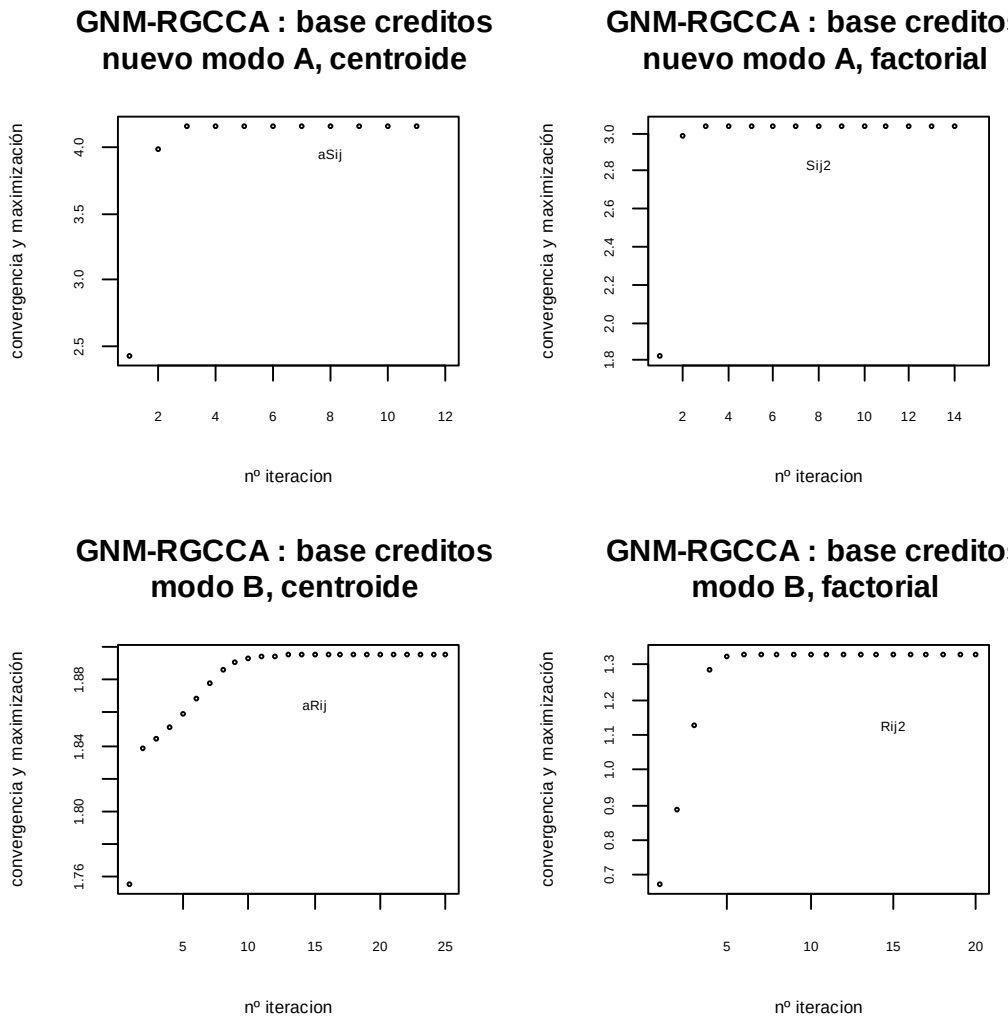
```
B               # coeficientes path

CSE               SitFnciera
0.03817647        0.48880868
```

El modelo estructural es de la forma $Y_3 = 0.0381Y_1 + 0.4888Y_2$
asociando un $R^2 = 0.28$

Así, una alta capacidad socioeconómica (Y_1) y una buena situación finaciera (Y_2) conllevan la aprobación de créditos mayor importe especialmente con financiación a corto y mediano plazo.

6.4.1. Funciones de maximización asociadas al procedimiento RGCCA

La base *Xc* tipo *k*-tablas debidamente cuantificada es aprovechada para verificar la maximización de las funciones de covarianza o correlación de las componentes *outer* según el modo y esquema previamente definidos. (Ver tabla 3.2).



Gráfica 6.3. Funciones de maximización derivadas del RGCCA

Los resultados asociados con estas funciones de maximizacion (gráfica 6.3) se obtienen del paquete rgcca() de R, de acuerdo con las siguientes salidas:

```
aSij        # valor absoluto covarianza

2.423384  3.992858  4.163707  4.165191  4.165197  4.165197  4.165197
4.165197  4.165197  4.165197 4.165197   ...


Sij2        # cuadrado de la covarianza

1.825536  2.985516  3.037593  3.039481  3.039539  3.039541  3.039541
3.039541  3.039541  3.039541 3.039541  3.039541  3.039541  ...


aRij        # valor absolute correlacion

1.755355  1.837977  1.844616  1.850793  1.858959  1.868679  1.878114
1.885458  1.890157  1.892749  1.894043  1.894649  1.894922  1.895043
1.895095  1.895118  1.895127  1.895131  1.895133  1.895134  1.895134
1.895134  1.895134  1.895134  ...


Rij2        # cuadrado correlacion

0.6723603   0.8874474   1.1248367   1.2862779   1.3220382   1.3271518
1.3279453   1.3280776   1.3281004   1.3281043   1.3281050   1.3281051
1.3281051   1.3281051   1.3281051  1.3281051    ...
```

CAPÍTULO 7

CONCLUSIONES

Después de los desarrollos NM-PLS presentados por Russolillo hacia 2009, se indagaba la posibilidad de cuantificar las variables cualitativas en conjuntos de datos mixtos mediante *k* componentes. A principios de 2014 se creó el proyecto denominado GNM-PLS, el cual conserva las variables cuantitativas originales (excepto por estandarización) y cuantifica óptimamente las variables cualitativas con una función agregada de las primeras *k* componentes derivadas del propio método de análisis.

Se creó GNM-NIPALS cuantificando cada *j*-ésima variable cualitativa mediante la función de reconstitución tipo ACP $\gamma_j = p_{1j}t_1 + \cdots + p_{kj}t_k$. Con este procedimiento se maximiza la inercia asociada al plano *k*-dimensional, permitiendo un mayor poder de descripción y conservando la variable cualitativa como tal. El ejemplo *vinos* permitió determinar que con *k*=4 componentes en la cuantificación se tiene una inercia maximal acumulada del 92.28%.

Con el método de regresión GNM-PLS2 se cuantifican las variables cualitativas en los dos conjuntos de datos mixtos Y *e* X. Las variables cualitativas en la matriz de respuesta Y se cuantifican mediante la función tipo regresión de las primeras H componentes de X, $\gamma_t = c_{q1}t_1 + \cdots + c_{qH}t_H$ y viceversa. Con la base *cred.438* se consiguió con *H*=2 un máximo de la función de covarianzas: $\sum_h^H cov^2\,(t_h, u_h) = 0.9643$. La *h*-ésima componente en el espacio de las Y es u_h.

En este mismo contexto de regresión GNM-PLS2 y como caso particular se creó GNM-PLS1q para cuantificar una única variable respuesta Y_q mediante una función γ_t como antes, obteniendo la cuantificación $\hat{Y}_q$ que a su vez cuantifica las variables cualitativas predictoras en X, iterando este procedimiento hasta la convergencia. En este caso se maximizó la función $\sum_h^H cov^2\left(t_h, \hat{Y}_{h-1}\right) = 0.276$ para *H*=2 y variable respuesta *Dictamen*.

Finalmente se desarrolló el método NM-RGCCA que modifica el modo A en NM PLS-PM para cuantificar las variables cualitativas de *J* bloques de datos mixtos. Esta función a su vez invoca el método RGCCA para obtener las componentes *inner* $Z_j = \sum_k e_k\, Y_k$ de dimensión *J* con la cual se cuantifican las cualitativas del bloque *j*.

El proceso asocia una función de maximización monótonamente convergente de covarianzas o correlaciones de las estimaciones *outer* Y_j conectadas, de acuerdo con el tipo de relación según el *nuevo modo A* o modo B y el esquema (*centroide o factorial*) de estimación de los pesos.

Luego de formar tres bloques de datos mixtos con la base *cred.438*, se obtuvo la ecuación estructural $Y_3 = 0.038Y_1 + 0.488Y_2$ indicando que las *Características del crédito* como *Dictamen, Plazo,* etc son explicadas con un R^2=0.28 por la *Capacidad socioeconómica* y en especial por la *Situación financiera.*

A pesar de que que la función de cuantificación *inner* en NM-RGCCA es de dimensión *J*, es importante resaltar el grado de complejidad que conlleva la obtención previa de las componentes *outer* multidimensionales en cada *j*-ésimo bloque para obtener GNM-RGCCA. El estudio de estas componentes *outer* limitadas por la dimensionalidad de cada bloque, pueden conllevar además a un nuevo marco teórico que se propone como investigación futura.

Para estudios futuros relacionados con GNM-PLS1 y GNM-PLS2 se deben analizar contrastes de significación para los coeficientes de regresión. Además será de mucho interés comparar con otra técnicas, por ejemplo la Regresión logística basada en distancias versus GNM-PLS1q.

Análogamente, en estudios posteriores relacionados con GNM-NIPALS, es muy importante considerar análisis de robustez con observaciones atípicas y datos faltantes, así como estabilidad en la representación factorial.

ANEXO A

GNM-NIPALS

```
fgnmNIPALS <- function(Xk,Xf,oXf)  # fcXn_p
{
   pXk <- ncol(Xk); pXf <- ncol(Xf); ph <- pXk+pXf ; n <- nrow(Xk)
   uhi <- matrix(0,ph,ph)     # matriz vect_p iniciales
   U <- matrix(0,ph,ph)

   XC <- cbind(Xk,Xf); MXfC <- matrix(0,n,(ph*pXf))
   XfC <- data.frame(MXfC); XfCh <- array(0,c(n,ph*pXf,rXk))

   ACPnipals <- fnipals(Xk)

   phi <- 1; eps <- 100
   while(phi <= rXk)   #cuantifica con 1; 1,2;...;1,2,...,phi
                        compals
   {
      ph1 <- phi
      T <- as.matrix(ACPnipals[[1]][,1:phi])  # inicio compals
                                                ortog.
      for(h in 1:phi)
      {    #* Conformacion inicial de U asociada a T
          for(j in 1:pXk){
             uhi[j,h] <- cor(Xk[,j],T[,h])
          }

          for(q in 1:pXf)
          {
             if(oXf[q]==1){YIto <- fcco(T[,h],q,Xf)} else
               {YIto <- fcso(T[,h],q,Xf)}
             YIt <- sqrt(n/(n-1))*scale(YIto)

             eh <- cor(YIt,T[,h])   # razon de correlacion
                                      invariante
             uhi[pXk+q,h]=eh
          }

          Nuhi <- sqrt(as.numeric(t(uhi[,h])%*%uhi[,h]))
               # norma uhi : recobra 1/Lj
          uh <- uhi[,h]*1/Nuhi      ;U[,h] <- uh
      }     #

      for(ej in 1:eps){ #** cuantificn simultanea q v.cualit

         for(q in 1:pXf){
            tnhi <- matrix(0,n)

            for(h in 1:phi){
                tnh <- as.matrix(U[pXk+q,h]*T[,h])+ tnhi
                tnhi <- tnh
            }
```

```
    if (oXf[q]==1){YIto <- fcco(tnh,q,Xf)} else
          {YIto <- fcso(tnh,q,Xf)}

    YIt <- sqrt(n/(n-1))*scale(YIto)    # estandarizo
    XC[,pXk+q] = YIt
 }

 Xh <- as.matrix(XC)

 for(h in 1:phi)        #*** recalculo Ph y th
 {
     p.hi <- t(Xh)%*%T[,h]
        Np.hi <- sqrt(as.numeric(t(p.hi)%*%p.hi))
     p.h <- p.hi*1/Np.hi ; U[,h] <- p.h

     t.h <- Xh%*%p.h ; T[,h] <- t.h
     Xo <- Xh; Xh <- Xo-t.h%*%t(p.h)
  }

} # end ej.

XfC[,((phi-1)*pXf+1):(phi*pXf)] <- XC[,(pXk+1):ph]
                                    # v.q cuantif.


# ................. INERCIA-AIE  .....................

XCh <- as.matrix(XC)  # matriz cuantificada 1.1, 2.2,...,
                              phi.phi

ACPnipals <- fnipals(XCh)
Th <- as.matrix(ACPnipals[[1]][,1:rXk])
Uh <- as.matrix(ACPnipals[[2]])

phc <- 1

while(phc <= phi)  # recuantifica 1.1;...;pXk.1,...,.pXk
{
   r.ytnh <- matrix(0,pXf,1)

   for(q in 1:pXf){
     tnhci <- matrix(0,n)

     for(h in 1:phc){
        tnhc <- as.matrix(Uh[pXk+q,h]*Th[,h])+ tnhci
        tnhci <- tnhc
     }

     if (oXf[q]==1){YIto <- fcco(tnhc,q,Xf)} else
          {YIto <- fcso(tnhc,q,Xf)}

     YIt <- sqrt(n/(n-1))*scale(YIto)
     XCh[,pXk+q] = YIt      # cuant intra.phc
     r.ytnh[q,1] <- cor(YIt,tnhc)
  }
```

```
        XfCh[,((phc-1)*pXf+1):(phc*pXf),phi]<- XCh[,(pXk+1):ph]

        acpXCh <- dudi.pca(XCh,scannf=FALSE,nf=ph)
        acpXChI <-inertia.dudi(acpXCh,row.inertia=TRUE,col.inertia=TRUE)

        cat("\n Componentes en la intra-cuantificn : ",phi,".",phc,"\n")
        cat("\n Valores propios asociados \n")




        print(acpXChI$TOT,2)

        plot(acpXCh$eig,type="h",xlab="ejes",ylab="val_p",xaxp=c(1,3,2))

        cat("\n Correlacion entre variables y rXk ejes ","\n")
        print(cor(XCh,Th[,1:phi]))

        cat("\n Correlacion entre v.cuantificadas yc.tnhc ","\n")
        print(r.ytnh)

        phc1 <- phc+1 ; phc <- phc1

      }     # end while phc.

        # ................. Fin AIE  ........................

       phi <- ph1+1

    } # end while phi

    s.fcXn_p <- list(XfC,XfCh,Th) # XfCh >> XfC: 1.1, 2.2,...,h.h
    return(s.fcXn_p)

} # end fgnmNIPALS().
```

```
fnipals <- function(Xii)
{
     p <- ncol(Xii); n <- nrow(Xii)
     X <- sqrt(n/(n-1))*scale(Xii)


     X0 <- X
     T <- matrix(1,n,p)
     P <- matrix(1,p,p)

     for(h in 1:p)
     {
        t1 <- as.matrix(X0[,1])

        for(e in 1:50)
        {
           P11 <- (t(X0)%*%t1)/(as.numeric(t(t1)%*%t1))
             nP11 <- as.numeric(t(P11)%*%P11)
           P1 <- 1/sqrt(nP11)*P11  # vect normalizado [matrix]
           t1 <- X0%*%P1
        }

        T[,h] <- t1
        P[,h] <- P1
        X1 <- X0 - t1%*%t(P1)      # deflacta
        X0 <- X1
     }

     resT.P <- list(T,P)
     return(resT.P)

 }   # End, Nipals datos completos.
```

ANEXO B

GNM-PLS1q

```
fgnmPLS1q <- function(Xk,Xq,Yq,oXq,oYq,H,a)
{

   n <- nrow(Xk); pXk <- ncol(Xk); pXq <- ncol(Xq)

   XC <- matrix(0,n,pXq); eps=100
   S2ty <- matrix(0,eps,a)

   Dk <- fnipals(Xk)       # descomposicion singular  NIPALS
   T <- as.matrix(Dk[[1]][,1:H]) # compnts ortogonales

   Yqo <- matrix(0,n,1); q <- 1

   for(h in 1:H)    # cuantifcn agregada inicial
   {
       if(oYq[q]==1){YIto <- fcco(T[,h],q,Yq)}else
         {YIto <- fcso(T[,h],q,Yq)}

       Yqc <- YIto + Yqo; Yqo <- Yqc
   }

   YIt <- scale(Yqc); sT <- sqrt(diag(var(T)))
   ct <- cor(T,YIt)/sT       # Conformacion C asociada a T
   tx <- T%*%ct


   for(ej in 1:eps)  # parametro estable ¡¡¡¡
   {
        q <-1
        if (oYq[q]==1){YIto <- fcco(tx,q,Yq)} else
             {YIto <- fcso(tx,q,Yq)}   # cuantif Yq

         Yqt <- scale(YIto)       # estandarizo
         colnames(Yqt) <- colnames(Yq)

      for(q in 1:pXq)
       {
         if (oXq[q]==1){XIto <- fcco(Yqt,q,Xq)} else
         {XIto <- fcso(Yqt,q,Xq)}# cuantif Xq

         XIt <- scale(XIto)       # estandarizo
         XC[,q] = XIt
       }

       colnames(XC) <- colnames(Xq)

       X <- as.matrix(cbind(scale(Xk),XC))     # matriz X ampliada
      Y <- Yqt
```

```
                    # .. Maximizacion ..

        f.PLS1 <- fPLS1(Y,X,a)
           TH <- f.PLS1[[2]]; CH <- f.PLS1[[3]]
           Yo <- Y

           for (h in 1:a)
           {
              th <- TH[,h]
              S2ty[ej,h] <- cov(th,Yo)^2

              Y1 <- Yo-th%*%t(CH[,h]); Yo <-Y1
           }

         T <- as.matrix(TH[,1:H]); C <- as.matrix(CH[,1:H])
         tx <- T%*%C

  } # end ej.

  f.vc1 <- fvc1(Y,X,a)

  R2h <- f.vc1[[1]]; R2cvh <- f.vc1[[2]]; Q2h <- f.vc1[[3]]
  PRESSh <- f.vc1[[5]]

       # print(data.frame(R2h,R2cvh,Q2h))
       # determine con estos indices, H compnts regresion

       wH <- f.PLS1[[1]]; tH <- T
       cH <- C; bH <- f.PLS1[[5]]

   r.gnmPLS1q <- list(Y,X,wH,tH,cH,bH,S2ty,R2h,R2cvh,Q2h,PRESSh)

   return(r.gnmPLS1q)

} # end fgnmPLS1q
```

```
fPLS1 <-function(Y,X,H)
{
   Xo <- scale(X)       # matriz n.p
    a <- qr(Xo)$rank   # rango de Xo, H <= a
   Yo <- scale(Y)   # matriz(col)n.1; es equiv sqrt(n/(n-1))

    pXo <- ncol(Xo); nXo <- nrow(Xo); pYo <- ncol(Yo)

    WH <- matrix(0,pXo,H); TH <- matrix(0,nXo,H)
    CH <- matrix(0,pYo,H); PH <- matrix(0,pXo,H)
    BH <- matrix(0,pXo,H)

    for(h in 1:H)# H : n° compon t en la regresion
    {
           whi <- t(Xo)%*%Yo/as.numeric(t(Yo)%*%Yo)
           nwhi <- as.numeric(sqrt(t(whi)%*%whi))

       wh <- whi/nwhi   # de norma 1

       th <- Xo%*%wh/as.numeric(t(wh)%*%wh)
           s.th2 <- as.numeric(t(th)%*%th)
       ch <- t(Yo)%*%th/s.th2
            s.ch2 <- as.numeric(t(ch)%*%ch)

       ph <- t(Xo)%*%th/s.th2 # Los ph no son ortonorm.

       X1 <- Xo -th%*%t(ph)
       Xo <- X1
       Y1 <- Yo - th%*%t(ch)
       Yo <- Y1

       WH[,h]<- wh; TH[,h]<- th; CH[,h]<- ch; PH[,h]<- ph

       BH[,h] <- WH[,1:h]%*%(solve(t(PH[,1:h])%*%WH[,1:h]))%*%CH[,1:h]
                 # coefs Xjz en Yo~c1t1 ; Yo~c1t1+c2t2 ; ...

    }  # end hH
       # CH == coef de los TH en la regres lm(Yo~TH-1)

    W.H <- WH%*%solve(t(PH)%*%WH)

    r.PLS1 <- list(WH,TH,CH,PH,BH,W.H)
    return(r.PLS1)

}   # end fPLS1
```

```
fvc1 <- function(Y,X,a)    #  Validacion Cruzada
{
  n <- nrow(X); pX <- ncol(X); Yz <- scale(Y)
  mY <- mean(Y); sY <- sqrt(var(Y))

  R2.h <- matrix(0,a,1); R2vc <- matrix(0,a,1)
  Q2h <- matrix(0,a,1);  PRESS <- matrix(0,a,1)
  RSS <- matrix(0,a,1)

  for (h in 1:a)
  {
    rg.PLS1 <- fPLS1(Y,X,h)
       T.h <- rg.PLS1[[2]]

        rgH <- lm(Yz~T.h -1)
          R2.h[h,1] <- summary(rgH)$r.squared
          Yze <- fitted(rgH); Ye <- mY+sY*Yze

          RSSh <- t(Y-Ye)%*%(Y-Ye)  # sum(residuo2)

    e_i2o <- 0
    for (i in 1:n)  # one leave out
    {

       Y_i <- Y[-i,]; X_i <- X[-i,]

        mXi <- colMeans(X_i); sXi <-sqrt(diag(var(X_i)))
              # mXi,sXi :vect numeric
        Xiz <- as.matrix((X[i,]-mXi)/sXi)  # matriz (col)
        mYi <- mean(Y_i); sYi <- sqrt(var(Y_i))

        PLS1_i <- fPLS1(Y_i,X_i,h)
           Wh_i <- PLS1_i[[1]]
           Ch_i <- PLS1_i[[3]]
           Ph_i <- PLS1_i[[4]]

           W.h_i <- Wh_i%*%solve(t(Ph_i)%*%Wh_i)       # Wh_i*
           thi <- t(Xiz)%*%W.h_i

           Yvc <- thi%*%t(Ch_i); Yivc <- mYi+sYi*Yvc
                    # estim Yz con V.C
           e_i2 <- (Y[i,]-Yivc)^2 + e_i2o ;  e_i2o <- e_i2
    }  # i

    PRESSh <- e_i2; PRESS[h,1] <- e_i2
    R2vc[h,1] <- as.numeric(1-PRESSh/((n-1)*var(Y)))

    if(h==1){Q2h[h,1]<- 1-PRESSh/(n-1)}else
       {Q2h[h,1]<- 1-PRESSh/RSSh_1}

    RSSh_1 <- RSSh ;  RSS[h,1] <- RSSh

  } # end h

  r.fvc1 <- list(R2.h,R2vc,Q2h,RSS,PRESS)     # Q2h :mal definida?
  return(r.fvc1)

} # end fvc1
```

ANEXO C

GNM-PLS2

```
fGNMpls2 <- function(Xk,Xq,Yk,Yq,oXq,oYq,H,a)
{
   n <- nrow(Xk); eps <- 100
   pXk <- ncol(Xk); pXq <- ncol(Xq)
   pYk <- ncol(Yk); pYq <- ncol(Yq)
       # H <= a; H compnts cuantif, a: compnts regresion

   XC <- matrix(0,n,pXq); colnames(XC) <- colnames(Xq)
   YC <- matrix(0,n,pYq); colnames(YC)<- colnames(Yq)
   r2Yt <-matrix(0,eps,a); L1e <-matrix(0,eps,a)
   S2tu <- matrix(0,eps,a)# 1ª etapa AIB

   Y <- scale(Yk); X <-scale(Xk)

   f.PLS2 <- fPLS2(Y,X,H)
       U <- f.PLS2[[4]]
       uy <- rowMeans(U)     # funCuant agregada inicial

       for(q in 1:pXq)     # Conformacion de d asociada a U
       {
         if(oXq[q]==1){XIto <- fcco(uy,q,Xq)} else
                {XIto <- fcso(uy,q,Xq)}

          XIt <- scale(XIto); sU <- sqrt(diag(var(U)))
          du <- cor(U,XIt)/sU
          uy. <- U%*%du

          if (oXq[q]==1){XIto <- fcco(uy.,q,Xq)} else
              {XIto <- fcso(uy.,q,Xq)}  # cuantif Xq
          XC[,q] = scale(XIto)

       }  # end q
       X <- cbind(scale(Xk),XC)

       f.PLS2 <- fPLS2(Y,X,H)
       T <- f.PLS2[[2]]

       for(q in 1:pYq)     # Conformacion de C asociada a T
       {
          Yqo <- matrix(0,n)
          for(h in 1:H)
          {
              if(oYq[q]==1){YIto <- fcco(T[,h],q,Yq)} else
                {YIto <- fcso(T[,h],q,Yq)}

              Yqc <- YIto + Yqo ; Yqo <- Yqc
          }

          YIt <- scale(Yqc); sT <- sqrt(diag(var(T)))
          ct <- cor(T,YIt)/sT
          tx <- T%*%ct
```

```
      if (oYq[q]==1){YIto <- fcco(tx,q,Yq)} else
         {YIto <- fcso(tx,q,Yq)}    # cuantif Yq


    YC[,q] = scale(YIto)

}  # end q

Y <- cbind(scale(Yk),YC)


for(ej in 1:eps)       # convergencia
{
   R12 <- t(X)%*%Y/(n-1)
   Rpls <- R12%*%t(R12)


         # ..  Maximizacion ..#

   f.PLS2 <- fPLS2(Y,X,a)

      T.h <- f.PLS2[[2]]; C.h <- f.PLS2[[3]]
      P.h <- f.PLS2[[5]]; Xo <- X; Yo <- Y

      for(h in 1:a)           # a : moderada
      {
          ch2 <- sum(C.h[,h]^2)
          uh. <- Yo%*%C.h[,h]/sqrt(ch2)    # ||c||=1
          th <- T.h[,h]; s2th <- var(th)

          S2tu[ej,h] <- cov(th,uh.)^2      # == r2Yt== L1e

         r2Yt[ej,h] <- sum(cor(Y,th)^2)*s2th  # equiv
          L1e[ej,h]<- eigen(Rpls)$values[1]  # 1ª etapa AIB

            X1 <- Xo-th%*%t(P.h[,h]); Y1 <- Yo-th%*%t(C.h[,h])
            R12 <- t(X1)%*%Y1/(n-1); Rpls <- R12%*%t(R12)
          Xo <- X1; Yo <- Y1
      }

   U <- f.PLS2[[4]]

   for(q in 1:pXq)
   {
      sU <- sqrt(diag(var(U))[1:H])
      du <- as.matrix(cor(U[,1:H],X[,pXk+q])/sU)
      uy <- U[,1:H]%*%du

      if (oXq[q]==1){XIto <- fcco(uy,q,Xq)} else
         {XIto <- fcso(uy,q,Xq)}          # re-cuantif Xq
      XC[,q] = scale(XIto)

   }  # end q

   X <- cbind(scale(Xk),XC)

   f.PLS2 <- fPLS2(Y,X,H); T <- f.PLS2[[2]]
```

```
      for(q in 1:pYq)     # Conformacion de C asociada a T
      {
         sT <- sqrt(diag(var(T)))
         ct <- as.matrix(cor(T,Y[,pYk+q])/sT)
         tx <- T%*%ct



         if (oYq[q]==1){YIto <- fcco(tx,q,Yq)} else
            {YIto <- fcso(tx,q,Yq)}          # re-cuantif Yq
         YC[,q] = scale(YIto)

      }  # end q

      Y <- cbind(scale(Yk),YC)

  }  # end ej

  W.h <- f.PLS2[[1]]; T.h <- f.PLS2[[2]]; C.h <- f.PLS2[[3]]
  U.h <- f.PLS2[[4]]; P.h <- f.PLS2[[5]]; B.h <- f.PLS2[[6]]
  Wh. <- f.PLS2[[7]]

  r.fGNMpls2 <-list(W.h,T.h,C.h,U.h,P.h,B.h,Wh.,X,Y,S2tu,r2Yt,L1e)
  return(r.fGNMpls2)

}  # end fGNMpls2
```

```
fPLS2 <- function(Y,X,H)   #  H <= a <- rango(X´Y/(n-1))
{
   Yo <- scale(Y) ; Xo <- scale(X)
   pXo <- ncol(Xo); n <- nrow(Xo); pYo <- ncol(Yo)

   WH <- matrix(0,pXo,H); TH <- matrix(0,n,H)
   CH <- matrix(0,pYo,H); PH <- matrix(0,pXo,H)
   BH <- array(0,dim=c(pXo,pYo,H)); UH <- matrix(0,n,H)

   for(h in 1:H)    # H : n° compon t en la regresion.
   {
       uh <- Yo[,1]

      for (ej in 1:200)
       {
            whi <- t(Xo)%*%uh/as.numeric(t(uh)%*%uh)
            nwhi <- as.numeric(sqrt(t(whi)%*%whi))

         wh <- whi/nwhi

         th <- Xo%*%wh/as.numeric(t(wh)%*%wh) # datos completos
             s.th2 <- as.numeric(t(th)%*%th)

         ch <- t(Yo)%*%th/s.th2
             s.ch2 <- as.numeric(t(ch)%*%ch)

         uh <- Yo%*%ch/s.ch2
       }

       ph <- t(Xo)%*%th/s.th2    # Los ph no son ortonorm.

       X1 <- Xo -th%*%t(ph); Xo <- X1
       Y1 <- Yo - th%*%t(ch); Yo <- Y1

       WH[,h]<- wh; TH[,h]<- th; CH[,h]<- ch; PH[,h]<- ph; UH[,h]<-uh

     BH[,,h] <- WH[,1:h]%*%solve(t(PH[,1:h])%*%WH[,1:h])%*%t(CH[,1:h])
                 # coefs Xjz

  }   # end hH
       # CH == coef de los TH en la regres lm(Y~TH-1)

  WH. <- WH%*%solve(t(PH)%*%WH)

  r.PLS2 <- list(WH,TH,CH,UH,PH,BH,WH.)
  return(r.PLS2)

} # end fPLS2 datos cuantitativos completos
```

```
fVC2 <- function(Y,X,a)    #  a: compnts regresn
{
   n <- nrow(X); r <- ncol(Y)
   R2 <- matrix(0,a,r); R2vc <- matrix(0,a,r)
   RSS <- matrix(0,a,r); PRESS <- matrix(0,a,r)
   ei2o <- matrix(0,a,r); ei2 <- matrix(0,a,r)

   for(h in 1:a)
   {
       rgPLS2 <- fPLS2(Y,X,h)  # estandariza Y,X
              T.h <- rgPLS2[[2]]

       for(k in 1:r)
       {
          Ykz <- scale(Y[,k]); mYk <- mean(Y[,k]); sYk <- sd(Y[,k])
          rgk.h <- lm(Ykz~T.h-1)
              R2[h,k] <- summary(rgk.h)$r.squared
              Ykze <- fitted(rgk.h)
              Yke <- mYk + sYk*Ykze
              RSS[h,k] <- sum((Y[,k]-Yke)^2)  # sum(residuo2)
                                                original
       }

       for(i in 1:n)
       {
          Y_i <-Y[-i,]; X_i <-X[-i,]
          mYi <-colMeans(Y_i); sYi <-sqrt(diag(var(Y_i)))
          mXi <-colMeans(X_i); sXi <-sqrt(diag(var(X_i)))
          Xi <- as.matrix(X[i,]); Xiz <- as.matrix((Xi-mXi)/sXi)
          Yi <- Y[i,]

          rgPLS2_i <- fPLS2(Y_i,X_i,h)

          Wh_i <- rgPLS2_i[[1]]; Th_i <- rgPLS2_i[[2]]
          Ch_i <- rgPLS2_i[[3]]; Ph_i <- rgPLS2_i[[5]]

          W.h_i <- Wh_i%*%solve(t(Ph_i)%*%Wh_i)
          thi <- t(Xiz)%*%W.h_i

          Yvc <- thi%*%t(Ch_i); Yivc <- mYi+sYi*Yvc
          ei2[h,] <-(Yi-Yivc)^2 + ei2o[h,]
             ei2o[h,] <-ei2[h,]

       }  # i

       R2vc[h,] <- as.numeric(1-ei2[h,]/((n-1)*diag(var(Y))))
       PRESS[h,] <- ei2[h,]

       colnames(PRESS) <- colnames(Y); colnames(RSS)<- colnames(Y)
       colnames(R2vc) <- colnames(Y); colnames(R2) <- colnames(Y)

   }  # h

   r.fvc2 <- list(R2,R2vc,RSS,PRESS,T.h)
   return(r.fvc2)

}  # end fVC2

library(pls)
Credt438.plsr <-plsr(Ykqc~Xkqc,scale=TRUE,validation="LOO",ncomp=2)
PRESSr <-Credt438.plsr$validation$PRESS       ==   fVC2[[4]]
```

ANEXO D

NM-RGCCA (NM-PLS PM nuevo modo A)

```
fgnmRGCCA <- function(q,Q,oQ,X.k,C,nameY,Esq,modo)
{
    n <- nrow(X.k[[1]]); ejl <- 40  #; pq <- rep(0,q)
    Qk <- Q; X. <- X.k
    taSij <- rep(0,ejl); taRij <- rep(0,ejl)
    tSij2 <- rep(0,ejl); tRij2 <- rep(0,ejl)

            #.......... k Compnts de inicio y ciclo ..........


    pls.RGCCAu <- fRGCCAu(q,X.,C,nameY,Esq,modo)
         Yk <- pls.RGCCAu[[1]]; Zk <- pls.RGCCAu[[5]]

    ej <- 1
    while(ej <= ejl)
    {
            #........ Pesos variables.q con las compnts........

       for(j in 1:q)          # bloques
       {
             Zkj <- Zk[,j] ; Xj <- X.k[[j]]
            qj <- ncol(Q[[j]])

            if(is.null(qj)){next} else
            {
                eq <- rep(0,qj)

              for(h in 1:qj)          # variabls
              {
                 if(oQ[[j]][h]==1)qhc <-fcco(Zkj,h,Q[[j]])else
                    qhc <- fcso(Zkj,h,Q[[j]])
                 eq[h]<- cor(qhc,Zkj)
              }
                ep <- cor(Zkj,Xj); epq <- cbind(ep,t(as.matrix(eq)))
                neq <- epq/sqrt(sum(epq^2))     # razon correl
            }

            #....... Cuantificn varbls.q  .......

            for(h in 1:qj)
             {
               if(oQ[[j]][h]==1)qhc <-fcco(Zkj,h,Q[[j]])else
                    qhc <- fcso(Zkj,h,Q[[j]])
               Qk[[j]][,h] <- scale(qhc)
            }
            colnames(Qk[[j]]) <- colnames(Q[[j]])
            X.[[j]] <- cbind(X.k[[j]],Qk[[j]])

       }  # fin bloques

       X.c <- X.
```

```
                    # ..... Componts Cuantifcn ........


          pls.RGCCAu <- fRGCCAu(q,X.c,C,nameY,Esq,modo)

          Yk <- pls.RGCCAu[[1]]
          Zk <- pls.RGCCAu[[5]]

                              #...... Maximizacion .............

          a.Sij <- 0 ; a.Rij <- 0; Sij.2 <- 0 ; Rij.2 <- 0
          covY <- cov(Yk); corY <- cor(Yk)

          for (i in 2:q)
          {
             for(j in 1:(i-1))
            {
                      aSij <-C[i,j]*abs(covY[i,j])+ a.Sij
                      aRij <- C[i,j]*abs(corY[i,j])+ a.Rij
                      Sij2 <- C[i,j]*covY[i,j]^2+ Sij.2
                      Rij2 <- C[i,j]*corY[i,j]^2+ Rij.2

                      a.Sij <- aSij; a.Rij <- aRij
                      Sij.2 <- Sij2 ; Rij.2 <- Rij2
            }
          }
          taSij[ej] <- aSij; taRij[ej] <- aRij
          tSij2[ej] <- Sij2; tRij2[ej] <- Rij2

         ej.u <- ej+1 ;  ej <- ej.u

         print(ej); print(covY)
          print(aSij); print(Sij2); print(aRij); print(Rij2)

     } #  convergencia

     wj <- pls.RGCCAu[[2]]
     B <- pls.RGCCAu[[3]]
     R2 <- pls.RGCCAu[[4]]

     rgnmRGCCA1 <- list(R2,B,wj,Yk,Zk,X.c,taSij,taRij,tSij2,tRij2)
     return(rgnmRGCCA1)

}    # end gnmRGCCA
```

```
fRGCCAu <- function(q,X.k,C,nameY,Esq,modo)  #
{
  n <- nrow(X.k[[1]]); ejl <- 200
  Y <- matrix(0,n,q); Z <-matrix(0,n,q); e <-matrix(0,q,q)
  a <- paste("a",1:q,sep=""); Wj <- list(seq(1:q))

  taSij <- rep(0,ejl); tSij2 <- rep(0,ejl)
  taRij <- rep(0,ejl); tRij2 <- rep(0,ejl)

  for(j in 1:q)# cualquier inicio conlleva result equiv :
  {
     pj <- ncol(as.matrix(X.k[[j]]))
     aj <- as.matrix(rep(1/sqrt(pj),pj))  #  RGCCA: tao=1

     Y[,j] <- as.matrix(X.k[[j]])%*%aj
  }
  colnames(Y) <- nameY

  for(j in 1:q)    # positividad
  {
     r.j <- cor(Y[,j],X.k[[j]])
     Sig.r <- sum(sign(r.j))
     if(Sig.r <= 0){sYj <- -Y[,j]; Y[,j] <- sYj}
  }

  ej <- 1
  while(ej <= ejl)
  {
    for(bj in 1:q)
    {
       for (i in 2:q){
          for(j in 1:(i-1))
          {
            if(C[i,j]==1){
               if(Esq==1) e[i,j]<- sign(cov(Y[,i],Y[,j])) #Cntroid
               if(Esq==2) e[i,j]<- cov(Y[,i],Y[,j])  # Factor
               if(Esq==3) e[i,j]<- C[i,j]# Horst
             }
            e[j,i] <- e[i,j]
          }
     }

     Zj <- matrix(0,n)
     for(i in 1:q){
       Z[,bj] <- e[i,bj]*Y[,i]+ Zj ; Zj <- Z[,bj]   # inner
     }

     Xj <- as.matrix(X.k[[bj]])

     if(modo[bj]==1){aj_i <- t(Xj)%*%Zj; naj<- sqrt(sum(aj_i^2))}else
        {aj_i <- solve(t(Xj)%*%Xj)%*%t(Xj)%*%Zj
         Yoj <- Xj%*%aj_i; naj <-sqrt(var(Yoj))}

         # naj<- sqrt(1/n*t(Zj)%*%Xj%*%aj_i)}

     aj <- aj_i*1/as.numeric(naj)

         # || aj||= 1 tipo RGCCA tao=1, v(Yj)=1 con tao=0

     assign(a[bj],aj)
     Wj[[bj]] <- get(a[bj])

     Y[,bj] <- as.matrix(X.k[[bj]])%*%aj  # outer: no stand
```

```
                    # positividad

       r.j <- cor(Y[,bj],X.k[[bj]])
       Sig.r <- sum(sign(r.j))
       if(Sig.r <= 0){sYj <- -Y[,bj]; Y[,bj] <- sYj}

    } # end bj

      # Maximizacion

      a.Sij <- 0; Sij.2 <- 0; a.Rij <- 0; Rij.2 <-0
      covY <- cov(Y); corY <- cor(Y)

      for (i in 2:q){
         for(j in 1:(i-1))
         {
            aSij <- C[i,j]*abs(covY[i,j])+ a.Sij
             Sij2 <- C[i,j]*covY[i,j]^2+ Sij.2
             aRij <- C[i,j]*abs(corY[i,j])+ a.Rij
             Rij2 <- C[i,j]*corY[i,j]^2+Rij.2

             a.Sij <- aSij; Sij.2 <- Sij2
             a.Rij <- aRij; Rij.2 <- Rij2
         }
      }
      taSij[ej] <- aSij; tSij2[ej] <- Sij2
      taRij[ej] <- aRij; tRij2[ej] <- Rij2

      ej.u <- ej+1 ;  ej <- ej.u

  } # end while

  rgY3 <- lm(Y[,3]~Y[,-3]-1, data=as.data.frame(Y))
  B <- coef(rgY3)
  R2 <- summary(rgY3)$r.squared

  rfRGCCAu <- list(Y,Wj,B,R2,Z,taSij,tSij2,taRij,tRij2)
  return(rfRGCCAu)

}
```

```
fPLSpm <- function(q,X.k,b,nameY,Esq,modo) # Modo A original
{
   n <- nrow(X.k[[1]]); ejl=1000; e <- matrix(0,q,q)
   Y <- matrix(0,n,q);  Z <- matrix(0,n,q)

   a <- paste("a",1:q,sep=""); wj <- list(seq(1:q))
   rXY <- paste("rXY",1:q,sep=""); rj <- list(seq(1:q))
   taSij <- rep(0,ejl); tSij2 <- rep(0,ejl)
   taRij <- rep(0,ejl); tRij2 <- rep(0,ejl)


   for(j in 1:q)
   {
      Y[,j] <- X.k[[j]][,1]
   }
   colnames(Y) <- nameY

   for(j in 1:q)
   {
      r.j <- cor(Y[,j],X.k[[j]])
      Sig.r <- sum(sign(r.j))
      if(Sig.r <0) {sYj <- -Y[,j]; Y[,j] <- sYj}

      if(Sig.r ==0){r.P <- sum(r.j[r.j > 0])
         r.N <- sum(r.j[r.j < 0])
         if(r.N > r.P){sYj <- -Y[,j]; Y[,j] <- sYj}
      }
   }

  ej <- 1
  while(ej <= ejl)
  {
    for(bj in 1:q)
    {

      for (k in 2:q){
         for(j in 1:(k-1))
         {
           if(b[k,j]==1){
              if(Esq==1) e[k,j]<- sign(cor(Y[,k],Y[,j]))
              if(Esq==2) e[k,j]<- cor(Y[,k],Y[,j])
              if(Esq==3) e[k,j]<- b[k,j]  # Horst
            }
           e[j,k] <- e[k,j]
         }
      }

      Zo <- matrix(0,n)

      for(k in 1:q)
      {
            Zj <- e[k,bj]*Y[,k]+ Zo ; Zo <- Zj
      }
```

```
      Z[,bj] <- Zj; sZ2 <- sum(Z[,bj]^2)    # ó scale(Zj)

      Xj <- as.matrix(X.k[[bj]])
      if(modo[bj]==1) aj_i <- t(Xj)%*%Z[,bj]*1/sZ2 else
         aj_i <- solve(t(Xj)%*%Xj)%*%t(Xj)%*%Z[,bj]

            # reg simple Xj~Zj ó aj_i <- cor(X.k[[bj]],Z[,bj])

      Yoj <- Xj%*%aj_i
      aj <- (1/as.numeric(sqrt(var(Yoj))))*aj_i
            # ajusto aj t.q v(Yj)=1
         assign(a[bj],aj); wj[[bj]] <- get(a[bj])

      Y[,bj] <- Xj%*%aj

      r.j <- cor(Y[,bj],Xj)
      Sig.r <- sum(sign(r.j))
      if(Sig.r <0){sYj <- -Y[,bj]; Y[,bj] <- sYj}

      if(Sig.r ==0){r.P <- sum(r.j[r.j > 0])
         r.N <- sum(r.j[r.j < 0])
         if(r.N > r.P){sYj <- -Y[,bj]; Y[,bj] <- sYj}
      }

   }  # end bj

      # Maximizacion

      a.Sij <- 0; Sij.2 <- 0; a.Rij <- 0; Rij.2 <-0
      covY <- cov(Y); corY <- cor(Y)

      for (i in 2:q){
         for(j in 1:(i-1))
         {
            aSij <- C[i,j]*abs(covY[i,j])+ a.Sij
            Sij2 <- C[i,j]*covY[i,j]^2+ Sij.2
            aRij <- C[i,j]*abs(corY[i,j])+ a.Rij
            Rij2 <- C[i,j]*corY[i,j]^2+Rij.2

            a.Sij <- aSij; Sij.2 <- Sij2
            a.Rij <- aRij; Rij.2 <- Rij2
         }
      }
      taSij[ej] <- aSij; tSij2[ej] <- Sij2
      taRij[ej] <- aRij; tRij2[ej] <- Rij2
      ej.u <- ej+1 ;  ej <- ej.u

} # end while

rownames(Y) <- rownames(X.k)
     # wj = aj <- list(a1,a2,a3)

rgY3 <- lm(Y[,3]~Y[,-3]-1)
  B3 <- rgY3$coeff
  R2 <- summary(rgY3)$r.squared
```

```
   for(j in 1:q)
   {
      assign(rXY[j],cor(X.k[[j]],Y[,j]))
      rj[[j]] <- get(rXY[j])
   }

   print(modo)
   rPLSpmt <- list(Y,wj,rj,B3,R2,Z,taSij,tSij2,taRij,tRij2)

} # end fPLSpm
```

BIBLIOGRAFÍA

ALUJA, T., Gonzalez, V. M. *GNM-NIPALS: General Nonmetric – Nonlinear Estimation by Iterative Partial Least Squares.* Revista de Matemática: Teoría y Aplicaciones, 21(1), 85-106. 2014.

ALUJA, T., Morineau, A. *Aprender de los datos: El análisis de Componentes Principales.* Ediciones Universitarias de Barcelona, Barcelona. 1999.

ALUJA, T. *Diseño del Producto Ideal*. Servei de Publicacions de la UPC, Barcelona. 1994.

BOJ, E., CLARAMUNT, M., GRANÉ, A., FORTIANA J. *Implementing PLS for distance –based regression: Computational issues*. Computational Statistics, 22, 237-248. 2007.

CARROLL, J. D. *Generalization of canonical analysis to three or more sets of variables*. Proc. Amer. Psychological Assoc. 227-228. 1968.

CUADRAS, C. M. *Distance analysis in discrimination and classification using both continuous and categorical variables*. In Y. Dodge (Ed.), North-Holland Publishing Co., 459-473. Amsterdam. 1989.

ESBENSEN, K., Schönkopf, S., Midtgaard, T. *Multivariate Analysis in Practice*. Olav Tryggvasons gt. 24, N-7011 Trondheim, Norway. 1994.

ESCOFIER, B., Pàges, J. *Análisis Factoriales Simples y Múltiples. Objetivos, Métodos e Interpretación.* Servicio Editorial Universidad del Pais Vasco. 1992.

ESPOSITO, V., Trinchera, L., Amato, S. *PLS Path Modeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement*. Handbook of Partial Least Squares, Springer, 47-82, Berlin. 2010.

GIFI, A. *Nonlinear multivariate analysis*. Wiley & Sons, New York. 1990.

HANAFI, M. '*PLS Path Modeling: computation of latent variables with the estimation mode B*'. Computational Statistics, 22, 275-292. 2007.

HANAFI, M., Kiers, H. A. L. *Analysis of K sets of data, with differential emphasis on agreement between and within sets*. Computational Statistics & Data Analysis, 51, 1491-1508. 2006.

HÖSKULDSSON, A. *Prediction Methods in Science and Technology*. 1, Basic Theory, Thor Publishing, Arnegaards Alle 7, Holte, Denmark. 1996.

HOTELLING, H. *Analysis of a complex of statistical variables into components*. Journal of Educational Psychology, 24. 1933.

HOTELLING, H. *Relation between two sets of variables*. Biometrika, 28, 129-149. 1936.

JÖRESKOG, K. *A General Method for Analysis of Covariance Structure*. Biometrika, 57, 239-251. 1970.

KETTENRING, J. R. *Canonical analysis of several sets of variables*. Biometrika 58, 433-451. 1971.

KRÄMER, N. *Analysis of high-dimensional data with partial least squares and boosting*. Doctoral dissertation, Technischen Universität Berlin. 2007.

LEBART, L., Morineau, A., y Piron, M. *Statistique Exploratoire Multidimensionnelle,* Dunod, Paris. 2006.

LINDGREN, F., Geladi, P., Wold, S. *The kernel algorithm for PLS.* Journal of Chemometrics, 7, 45-59. 1993.

LINTING, M., MEULMAN, J., GROENEN, P., VAN der KOOIJ, A. *Nonlinear Principal Cpomponents Analysis: Introduction and Application*. Psychological Methods, 12(3), 336-358. 2007.

LOHMÖLLER, J. *Latent variable path modeling with partial least squares*. Physica-Verlag, Heidelberg. 1989.

MARDIA, K., Kent, J., Bibby, J. *Multivariate Analysis*. Academic Press Inc, London. 1979.

MARTENS, H., Nars, T. *Multivariate calibration.* John Wiley & Sons, New York. 1989.

QANARI, E. M., Hanafi, M. *A simple continuum regression approach*. Journal of Chemometrics, 19, 387-392. 2005.

RUSSOLILLO, G. *Non Metric Partial Least Squares*. Electronic Journal of Statistics, 6, 1641-1669. 2012.

RUSSOLILLO, G. *Partial Least Squares Methods for Non-Metric Data*. PhD thesis, Università degli Studi di Napoli Federico II, Napoli, November 2009.

SANCHEZ, G. *PATHMOX Approach: Segmentation Trees in Partial Least Squares Path Modeling*. PhD thesis, Universidad Politecnica de Cataluña, Barcelona. 2008.

SAPORTA, G. *Probabilitiés Analyse des Données et Statistique*. Editions Technip, Paris. 2011.

TENENHAUS, M. *Analyse en composantes principales d'un ensemble de variables nominales ou numeric*. Revue de statistique appliquée 25, 39-56. 1977.

TENENHAUS, M. *La régression PLS théorie et pratique*. Editions Technip, Paris. 1998.

TENENHAUS, M., Amato, S., Esposito, Vinzi V. *A global goodness-of-fit index for PLS structural equation modelling*. XLII SIS scientific meeting, papers, CLEUP, Padova, 739-742. 2004

TENENHAUS, M., Esposito, V., Chatelin, Y. M., Lauro, C. *Path modeling*. Computational Statistics and data Analysis, 48, 159-205. 2005.

TENENHAUS, M. *An criterion based PLS approach to structural equation modeling*. In Programme and abstract of the 6th International Conference on Partial Least Squares Methods, 3. 2009.

TENENHAUS, M., Hanafi, M. *A bridge between PLS-PM and multi-block data analysis*. Springer Handbooks of Computational Statistics, Berlin. 2010.

TENENHAUS, A., Tenenhaus, M. *Regularized generalized canonical correlation analysis*. Psychometrika, 76, 257-284. 2011.

TRACY, N. D., Young, J. C., Mason, R. L. *Multivariate control charts for individual observations*. Journal of Quality Technology, 24, 88-95. 1992.

TUCKER, L. R. *An inter-battery method of factor analysis*. Psychometrika, 23(2), 111-136. 1958.

TYLER, D. E. *On the optimality of the simultaneous redundancy transformations*. Psychometrika, 47(1), 77-86. 1982.

VAN de GEER, J. P. *Linear relations among k sets of variables*. Psychometrika, 4(9), 70-94. 1984.

VAN de WOLLENBERG, A. L. *Redundancy analysis: an alternative for canonical correlation*. Psychometrika, 42, 207-219. 1977.

VEGA, J., Guzmán, J. *Regresión PLS y PCA como solución al problema de multicolinealidad en Regresión Múltiple.* Revista de Matemática CIMPA – UCR. Costa Rica. 2011.

WOLD, H. *Estimation of principal component and related models by Iterative Least Square*s. In P. R. Krishnaiah Ed., Multivariate Analysis, 391-420, Academic Press, New York. 1966.

WOLD, H. *Modelling in complex situations with soft information*. Third World Congress of Econometric Society, Toronto. 1975.

WOLD, H. *Path models with latent variables: The non-linear iterative partial least squares (NIPALS) approach*, Accademic Press, 307-357, New York. 1975a.

WOLD, H. *Soft modeling: the basic design and some extensions.* In K. G. jöreskog and H. Wold, eds, 'Systems under indirect observation', Part II, North Holland. 1982.

WOLD, S., Martens, H., Wold, H. *The multivariate calibration problem en chemistry solved by the PLS method*. In A. Ruhe & B. Kagstrom, eds, 'Proceedings of the Conference on Matrix Pencils. Lectures Notes in Mathematics', Springer, Heidelberg. 1983.

WOLD, H. *Partial least squares.* In: Encyclopedia of Statistical Sciences*,* 6, 581-591.Kotz, S., and Johnson, N.L. (Eds). New York. Wiley. 1985.

YOUNG, F., Takane, Y., de Leeuw, J. *The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features*. Psychometrika 43 (2), 279-281. 1978.

YOUNG, F. *Quantitative analysis of qualitative data*. Psychometrika, 44, 505-529. 1981.