Revista de Matemática: Teoría y Aplicaciones
ISSN: 1409-2433
mta.cimpa@ucr.ac.cr
Universidad de Costa Rica
Costa Rica

ALUJA, TOMÁS; GONZÁLEZ, VÍCTOR MANUEL

GNM - NIPALS: ESTIMACIÓN GENERAL NO MÉTRICA Y NO LINEAL POR MÍNIMOS CUADRADOS PARCIALES ITERATIVOS

Revista de Matemática: Teoría y Aplicaciones, vol. 21, núm. 1, enero, 2014, pp. 85-106
Universidad de Costa Rica
San José, Costa Rica

Disponible en: http://www.redalyc.org/articulo.oa?id=45331281006

Cómo citar el artículo
Número completo
Más información del artículo
Página de la revista en redalyc.org

Sistema de Información Científica
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

REVISTA DE MATEMÁTICA: TEORÍA Y APLICACIONES 2014 **21**(1) : 85–106
CIMPA – UCR ISSN: 1409-2433

GNM-NIPALS: ESTIMACIÓN GENERAL NO MÉTRICA Y NO LINEAL POR MÍNIMOS CUADRADOS PARCIALES ITERATIVOS

GNM-NIPALS: GENERAL NONMETRIC–NONLINEAR ESTIMATION BY ITERATIVE PARTIAL LEAST SQUARES

TOMÁS ALUJA* VÍCTOR MANUEL GONZÁLEZ†

Received: 7/May/2013; Revised: 14/Nov/2013;
Accepted: 15/Nov/2013

---

*Laboratori de Modelització i Anàlisi de la Informació (LIAM), Universitat Politècnica de Catalunya, Barcelona, España. E-Mail: tomas.aluja@upc.edu

†Docente, Universidad del Valle, Cali, Colombia. E-Mail: victor.m.gonzalez@correounivalle.edu.co, victor.manuel.gonzalez.rojas@upc.edu

Resumen

En este trabajo se desarrolla GNM-NIPALS 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 solo 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 reconstituida, 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 esta explicada, caso en el cual la función de cuantificación definitiva es de rango menor (k^*).

Palabras clave: NM-PLS, ACP, datos mixtos, cuantificación, k-dimensional, inercia saturada, maximal, razón correlación.

Abstract

This paper develops GNM-NIPALS as an extension of the NM-PLS methods, which allows to quantify the qualitative variables of mixed data, by means of the reconstitution function using the first k principal components, maximizing the inertia in the plane k subspace associated with the PCA of the quantified matrix. It generalizes the NM-NIPALS algorithm in the sense that the latter only uses the first principal component in the quantification of qualitative variables. From the maximization and positivity of the correlation ratio between each qualitative variable and the reconstituted function, we have that the accumulated inertia on the k-dimensional subspace associated to the quantification function of the same range is greater than or equal to the one generated on subspaces of equal dimension, but with quantification functions of different range. With the k principal components associated to the quantified matrix, a saturated inertia analysis is performed to evaluate if a dimension $k^* < k$ still exists, from which the accumulated inertia on the axes of equal or superior order is already explained, in which case the definitive quantification function is of lesser range (k^*).

Keywords: NM-PLS, PCA, mixed data, quantification, k-dimensional, saturated inertia, maximal, correlation ratio.

Mathematics Subject Classification: 62H25.

1 Introducción

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) [3], requieren en su desarrollo que las variables sean de tipo cuantitativo; el Análisis de Componentes Principales (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 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) [4] 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, como un *todo*, 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) [7], 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.

En este artículo, se presenta una generalización de NM-NIPALS, denominado GNM-NIPALS, el cual implementa la cuantificación a partir de una función lineal $\gamma = f(t_h)$ de h componentes principales vía reconstitución de la $q_{ésima}$ variable como en ACP, Aluja & Morineau (1999) [1], es decir de la forma $f(t_h) = p_{_{1q}}t_1 + p_{_{2q}}t_2 + \ldots + p_{_{hq}}t_h$, 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 $\mathbf{X}_p$, 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 $\eta_{\gamma|X_q}$ es máxima y positiva, Saporta (2011) [5], 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 $h \neq k$.

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 sería, $f(t_k) = p_{1q}t_1 + p_{2q}t_2 + \ldots + p_{kq}t_k$.

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: $f(t_{k^*}) = p_{1q}t_1^k + p_{2q}t_2^k + \ldots + p_{k^*q}t_{k^*}^k$ donde el superíndice k indica que son las componentes finales que se obtuvieron con la función $f(t_k)$.

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 *gustación* del ejemplo vinos, ver Escofier & Pagès (1992) [2]. 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, la sección dos iniciará explicando la conceptualización de este procedimiento y luego se expondrá lo relacionado con el NM-NIPALS; al final de la sección se presenta el procedimiento algorítmico objeto de este artículo denominado GNM-NIPALS y fundamentado en los métodos NM-PLS.

La sección tres contiene la interpretación y resultados del ejemplo de aplicación de GNM-NIPALS y finalmente en la sección cuatro se dan las conclusiones evidenciando la ganancia de inercia frente al NM-NIPALS.

2 Metodologías

2.1 El algoritmo NIPALS

NIPALS es la base de la regresión PLS, Tenenhaus (1998) [6]. Fundamentalmente realiza una descomposición singular de una matriz de datos, mediante secuencias iterativas de proyecciones ortogonales (concepto geométrico de regresión) obtenidas como productos escalares. Con bases de datos completas se tiene equivalencia con los resultados del ACP; además, 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.

Para la matriz de datos $\mathbf{X}_{n,p}$ de rango a cuyas columnas $X_1, \ldots, X_p$ se suponen centradas o estandarizadas, la descomposición derivada del ACP permite la reconstitución mediante el siguiente esquema:

$$\mathbf{X} = \sum_{h}^{a} t_h P_h'$$

(t_h es la h-ésima *componente principal* y representa los *scores*, y P_h es el vector propio o *loadings* en el eje h), por tanto,

$$[X_1 \ldots X_p] = t_1 P_1' + \ldots + t_a P_a'. \tag{1}$$

Así, la variable

$$X_j = \sum_{h}^{a} t_h p_{_{hj}}, \quad j = 1, \ldots, p \tag{1a}$$

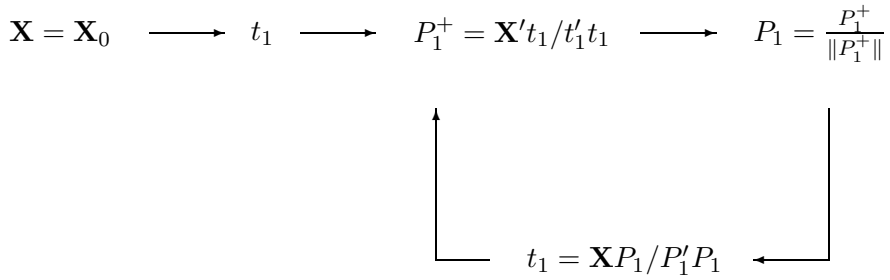
y la i-ésima fila (individuo)

$$x_i = \sum_{h}^{a} t_{hi} P_h, \quad i = 1, \ldots, n. \tag{1b}$$

.

Observe 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) en la regresión sin intercepto de X_j sobre t_h, con lo cual para todas las j-variables se obtiene el h-ésimo *vector propio* $P_h = (p_{h1}, \ldots, p_{hp})$. En el espacio 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. Así, el flujograma asociado al procedimiento iterativo 2.2 del algoritmo NIPALS es:



Se construirá una serie de tablas notadas $\mathbf{X}_h$ cuyas columnas se denotan como X_{hj} y la i-ésima fila se notará x'_{hi}. El algoritmo comienza tomando la matriz original como $\mathbf{X}_0$ e iniciando la primera *componente principal* t_1 con 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.

Con t_1 se calcula

$$\mathbf{X}' \underbrace{\mathbf{X}P_1}_{t_1} = \lambda P_1 \Rightarrow \mathbf{X}'t_1/\lambda = P_1,$$

que después de ser normado permite recalcular t_1 e iterar hasta la convergencia; además, $\lambda_1 = \frac{1}{n}t_1't_1$.

Luego se deflacta la matriz mediante $\mathbf{X}_1 = \mathbf{X}_0 - t_1P_1'$ 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. $X_0 = X_h$
Etapa 2. $h = 1, 2, \ldots, a$:
  Etapa 2.1. $t_h = 1^a$ columna de X_{h-1}
  Etapa 2.2. Repetir hasta la convergencia de P_h
    Etapa 2.2.1. $P_h = \frac{X'_{h-1}t_h}{t'_ht_h}$
    Etapa 2.2.2. Normar p_h a 1
    Etapa 2.2.3. $t_h = X_{h-1}P_h/P'_hP_h$
  Etapa 2.3. $X_h = X_{h-1} - t_hP'_h$ [garantiza la ortogonalidad]
Siguiente h.

NIPALS entrega las *componentes* t_h y los *vectores propios* P_h correspondientes a la matriz $\mathbf{X}$ excepto tal vez por signo, tal como si se hubiese aplicado la función `svd`($\mathbf{X}$) de R.

2.2 NM-NIPALS

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 es requerido.

Respecto a la propiedad grupal, la variable escalada $\hat{x}$ debe ser restringida tal que:

$$x_i^* \sim x_{i'}^* \Rightarrow \hat{x}_i = \hat{x}_{i'}$$

donde $\sim$ significa *pertenencia* a la misma categoría.

Si la variable a ser cuantificada es ordinal debe añadirse la restricción de *orden* $(\prec)$, con lo cual

$$x_i^* \sim x_{i'}^* \Rightarrow \hat{x}_i = \hat{x}_{i'} \text{ y } x_i^* \prec x_{i'}^* \Rightarrow \hat{x}_i < \hat{x}_{i'}.$$

Se define la función de cuantificación $q()$, Young (1981) [8], 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) γ *o* t más cercano:

$$\tilde{q}(x^*, \gamma) : \hat{x} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\gamma = P_{\tilde{X}}\gamma. \tag{2}$$

El coeficiente de determinación de esta regresión equivale al cuadrado de la razón de correlación de Pearson entre la variable categórica original y el LC. Por tanto la razón de correlación entre γ y x^* que siempre es *positiva* puede ser expresada en términos de la correlación lineal de Pearson:

$$\text{cor}(\gamma, \hat{x}) = \eta_{\gamma|x^*}$$

además

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

Como $R = \sup_{a_1,\ldots,a_k} r(\gamma; \sum_{j=1}^k a_j\tilde{x}_j)$, 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 NM-NIPALS obtiene en cada cuantificación

$$\max\{\text{cor}^2(\hat{x}, t_1)\}. \tag{3}$$

El pseudocódigo del algoritmo de NM-NIPALS es:

Input X^*
Output $P_h : [p_1, \ldots, p_h]$; $T_h : [t_1, \ldots, t_h]; \hat{\mathbf{X}}$.

1. *Inicializa* t_1

2. *Repeat*
   $\hat{x}_q = q(x_q^*, t_1)$ *# cuantificación mediante ecuación (2)*
   $\hat{\mathbf{X}} = [x_1 \ldots \hat{x}_q]$
   $\mathbf{p}_1 = \hat{\mathbf{X}}'\mathbf{t}_1/(\mathbf{t}_1'\mathbf{t}_1)$
   $\mathbf{p}_1 = \mathbf{p}_1/\|\mathbf{p}_1\|$
   $\mathbf{t}_1 = \hat{\mathbf{X}}\mathbf{p}_1$
   Until convergencia de $\mathbf{p}_1$.

3. $\mathbf{E}_1 = \hat{\mathbf{X}} - \mathbf{t}_1\mathbf{p}_1'$

4. *for* $h = 2, \ldots, p^*$
   Inicializa $\mathbf{t}_h$

5. *Repeat*
   $\mathbf{p}_h = \mathbf{E}_{h-1}'\mathbf{t}_h/(\mathbf{t}_h'\mathbf{t}_h)$
   $\mathbf{p}_h = \mathbf{p}_h/\|\mathbf{p}_h\|$
   $\mathbf{t}_h = \mathbf{E}_{h-1}\mathbf{p}_h/(\mathbf{p}_h'\mathbf{p}_h)$
   Until convergencia de $\mathbf{p}_h$.

6. $\mathbf{E}_h = \mathbf{E}_{h-1} - \mathbf{t}_h\mathbf{p}_h'$

7. *End.*

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 $t_1, \ldots, t_{p^*}$ si $\hat{X}$ es de rango completo p^*.

Para garantizar el orden se usa en vez de las indicadoras las matrices de orden $\bar{\bar{X}}$ donde para cada individuo se tiene una de las siguientes recodificaciones según la categoría de orden asumida $(a < b < \ldots < k)$:

a	1	0	$\ldots$	0
b	1	1	$\ldots$	0
$\vdots$				
k	1	1	$\ldots$	1

Luego, mediante regresión monótona de γ sobre $\bar{\bar{X}}$ 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 $\tilde{\tilde{X}}$. 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 $\hat{x}_q$ analizada a nivel de escala *ordinal*, la cuantificación está dada por:

$$\tilde{\tilde{q}}(x_q^*, t_1) : \hat{x}_q \propto \tilde{\tilde{X}}_q (\tilde{\tilde{X}}_q' \tilde{\tilde{X}}_q)^{-1} \tilde{\tilde{X}}_q' t_1 \tag{4}$$

Ahora, puesto que $\mathrm{cor}(\hat{x}_q, t_1) \propto p_{_{1q}}$, 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.

2.3 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 las variables cualitativas de una matriz de datos mixtos mediante una función lineal de h componentes contenidas en la matriz de datos cuantitativos, esto es:

$$f(t_h) = p_{_{1q}} t_1 + p_{_{2q}} t_2 + \ldots + p_{_{hq}} t_h. \tag{5}$$

La función $f(t_h)$ es una aplicación directa del concepto de *reconstitución* de una variable q derivada del ACP. Los pesos $p_{_{hq}}$ (estandarizados) corresponden a la q-ésima coordenada del *vector propio* P_h asociado a la componente t_h del mismo rango; ver (1a). Por tanto, con cada p_{hq} se puede obtener la correlación de la variable q cuantificada con el eje h, que en este caso equivale a la razón de correlación $\eta_{\gamma|x_q^*} = \mathrm{cor}(\hat{x}_q, \gamma)$, la cual es máxima y positiva, $\gamma = p_{_{hq}} t_h$. Esta correlación se puede calcular bajo t_h ya que

$$\hat{x}_{q\gamma} = P_{\tilde{X}_q} \gamma = p_{_{hq}} P_{\tilde{X}_q} t_h = p_{_{hq}} \hat{x}_{qt}.$$

Por tanto

$$\forall x_q, \mathrm{cor}(\hat{x}_{q\gamma}, \gamma) = \mathrm{cor}(p_{_{hq}} \hat{x}_{qt}, p_{_{hq}} t_h) = \mathrm{cor}(\hat{x}_{qt}, t_h). \tag{6}$$

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 $\mathrm{cor}(\hat{x}_{q\gamma}, t_h)$ toma el mismo valor excepto por el signo de $p_{_{hq}}$.

El método inicia tomando h *componentes* asociadas a la matriz $\mathbf{X}p$ (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 (5) se realiza la cuantificación $\hat{x}_q$ como en (2) y se obtiene la correlación (6), 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) de dimensión p^* .

Se toman así las q-ésimas coordenadas de los vectores propios P_h permitiendo formular la función de inicio $f(t_h) = p_{_{1q}}t_1 + p_{_{2q}}t_2 + \ldots + p_{_{hq}}t_h$ 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 cuantificamos con $f(t_1)$ se tiene la matriz cuantificada $\mathbf{X}C_1$ de cuya descomposición singular obtenemos 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 solo con la primera componente principal t_1 de la matriz $\mathbf{X}p$. Si cuantificamos con las dos primeras componentes $f(t_2)$, de la matriz así cuantificada $\mathbf{X}C_2$ obtendremos los vectores propios $P_1^b, P_2^b, \ldots, P_{p^*}^b$ y las componentes principales $t_1^b, t_2^b, \ldots, t_{p^*}^b$, donde el superíndice b indica que la cuantificación se realizó con dos componentes.

Observe entonces que $t_1^a \neq t_1^b, t_2^a \neq t_2^b, \ldots$; 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 $h = 1, 2, \ldots, p$.

De acuerdo con (2) haciendo $\gamma = f(t_h)$, al cuantificar sin orden simultáneamente cada variable cualitativa con h componentes, la $\mathrm{cor}^2(\hat{x}, f(t_h))$ 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.

Esta correlación maximal puede ser un índice de la dimensionalidad de $f(t)$, ya que valores relativamente grandes, por ejemplo $\mathrm{cor}^2(\hat{x}, f(t_h)) > 0.90$ indican que h componentes serán suficientes para la cuantificación vía reconstitución; observe que con $h = p^*$ entonces $\mathrm{cor}^2(\hat{x}, f(t_{p^*})) = 1$, 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: $f(t_h) = p_{1q}t_1 + p_{2q}t_2 + \ldots + p_{hq}t_h,\ h \leq p$. 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, $\lambda_1^a \geq \lambda_1^b$.

Sean $t_{h\cdot} = p_{hq}t_h$ y $q_{h\cdot} = PI_q t_{h\cdot}$ la cuantificación asociada, donde PI_q es el proyector ortogonal de las indicadoras de la variable q. En el caso, $f(t_2) = t_{12\cdot} = t_{1\cdot} + t_{2\cdot}$ la cuantificación asociada $q^b_{12\cdot}$ se consigue mediante $q^b_{12\cdot} = PI_q t^b_{12\cdot} = PI_q t^b_{1\cdot} + PI_q t^b_{2\cdot}$ gracias a la convergencia y ortogonalidad de $t^b_{1\cdot}, t^b_{2\cdot}$. De la razón de correlación se tiene que $r^2(q^b_{12\cdot},\ t^b_{12\cdot}) = r^2(q^b_{12\cdot},\ t^b_{1\cdot}) + r^2(q^b_{12\cdot}, t^b_{2\cdot})$ es máxima, induciendo máxima inercia en el plano generado por t^b_1, t^b_2.

Se muestra algorítmicamente que estos resultados se extienden a $h = 3, \ldots, p$. 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,

$$\lambda_1^c + \lambda_2^c + \lambda_3^c \geq \begin{cases} \lambda_1^a + \lambda_2^a + \lambda_3^a \\ \lambda_1^b + \lambda_2^b + \lambda_3^b. \end{cases}$$

Si denominamos $\mathbf{X}C_k$ la matriz conteniendo las p variables cuantitativas (estandarizadas) y las q variables (cualitativas) cuantificadas y estandarizadas, es evidente que al realizar la descomposición singular de $\mathbf{X}C_k$ se obtienen tantas *componentes* $t_1^k, t_2^k, \ldots, t_s^k$ como el rango s de $\mathbf{X}C_k$, las mismas obtenidas en la convergencia de la cuantificación sin orden.

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.

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 $f(t_k) = p_{1q}t_1 + p_{2q}t_2 + \ldots + p_{kq}t_k$, es mayor o igual que la explicada en planos de igual dimensión k, asociados a funciones de cuantificación con menor número de componentes finales, es decir con $f(t_{k^*}) = p_{1_q}t_1^k + p_{2_q}t_2^k + \ldots + p_{k^*_q}t_{k^*}^k$ se tiene:*

$$I_k^k \geq I_{k^*}^k,\ \textit{donde } k^* < k.$$

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, solo es necesario analizar las corre-

laciones de las variables cualitativas recuantificadas para obtener dichas inercias; ver Tabla 1.

$f(t_{1.})$	t_1	t_2	$\ldots$	t_{p+q}
X_1	r_{11}	r_{12}		r_{1p^*}
X_2	r_{21}	r_{22}		r_{2p^*}
$\vdots$				
q_1^a	$r^a_{q_1 1}$	$r^a_{q_1 2}$		
$\sum r_{jk}^2$	λ_1^a	λ_2^a	$\ldots$	$\lambda_{p^*}^a$

$f(t_{12.})$	t_1	t_2	$\ldots$	t_{p+q}
X_1	r_{11}	r_{12}		r_{1p^*}
X_2	r_{21}	r_{22}		r_{2p^*}
$\vdots$				
q_1^b	$r^b_{q_1 1}$	$r^b_{q_1 2}$		
$\sum r_{jk}^2$	λ_1^b	λ_2^b	$\ldots$	$\lambda_{p^*}^b$

Tabla 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.

Demostración. Omitiremos en esta demostración el superíndice k solo por comodidad, pero no debemos 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:

$$t_{123.} = t_{1.} + t_{2.} + t_{3.} = t_{12.} + t_{3.} \tag{7}$$

y su cuantificación asociada es:

$$q_{123.} = PI_q * t_{123.} = PI_q t_{1.} + PI_q t_{2.} + PI_q t_{3.}$$

$$q_{123.} = q_{1.} + q_{2.} + q_{3.} = q_{12.} + q_{3.}. \tag{8}$$

De la ortogonalidad de las componentes se tiene que las correlaciones al cuadrado son:

$$r^2(q_{123.}, t_{123.}) = r^2(q_{123.}, t_{1.}) + r^2(q_{123.}, t_{2.}) + r^2(q_{123.}, t_{3.}) = I_3^3 \tag{9}$$

Análogamente, de las recuantificaciones con $k^* = 2$, y $k^* = 1$ se tiene respectivamente:

$$r^2(q_{12.}, t_{12.}) = r^2(q_{12.}, t_{1.}) + r^2(q_{12.}, t_{2.}) = I_2^2 \tag{10}$$

$$r^2(q_{1.}, t_{1.}) = I_1^1. \tag{11}$$

Las expresiones (9), (10) y (11) 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) [5].

El extremo derecho de la igualdad en las ecuaciones (12), (13) y (14) demuestran la "inercia maximal intra" en los planos de igual dimensión. En el plano $k = 1$,

$$r^2(q_{1.}, t_{1.}) \geq \begin{cases} r^2(q_{12.}, t_{1.}) = r^2(q_{12.}, q_{1.}) r^2(q_{1.}, t_{1.}) = I_2^1 \\ r^2(q_{123.}, t_{1.}) = r^2(q_{123.}, q_{1.}) r^2(q_{1.}, t_{1.}) = I_3^1 \end{cases} \tag{12}$$

con lo cual $I_1^1 \geq \begin{cases} I_2^1 \\ I_3^1. \end{cases}$

En el plano de dimensión dos,

$$r^2(q_{12.}, t_{12.}) \geq \begin{cases} r^2(q_{1.}, t_{12.}) = r^2(q_{1.}, q_{12.}) r^2(q_{12.}, t_{12.}) = I_1^2 \\ r^2(q_{123.}, t_{12.}) = r^2(q_{123.}, q_{12.}) r^2(q_{12.}, t_{12.}) = I_3^2 \end{cases} \tag{13}$$

entonces $I_2^2 \geq \begin{cases} I_1^2 \\ I_3^2. \end{cases}$

Análogamente, ya que la expresión (9) es maximal

$$r^2(q_{123.}, t_{123.}) \geq \begin{cases} r^2(q_{1.}, t_{123.}) = r^2(q_{1.}, q_{123.}) r^2(q_{123.}, t_{123.}) = I_1^3 \\ r^2(q_{12.}, t_{123.}) = r^2(q_{12.}, q_{123.}) r^2(q_{123.}, t_{123.}) = I_2^3 \end{cases} \tag{14}$$

por tanto, $I_3^3 \geq \begin{cases} I_1^3 \\ I_2^3. \end{cases}$

De las expresiones (12), (13) y (14) se concluye que las inercias en los planos k-dimensionales son máximas cuando son generadas por funciones de cuantificación de igual dimensión. ■

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 $k^*, \ldots, k, k+1, k+2, \ldots$ de la matriz $\mathbf{X}C_k$.*

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 $f(t)$.

La generalización de estos resultados al caso en el cual $k > 3$ es evidente e inmediato. El pseudoalgoritmo asociado al procedimiento GNM-NIPALS se presenta a continuación:

Input $\mathbf{X}p$; *Output* $\mathbf{X}C, \mathbf{T}, \mathbf{P}$

Inicializa $\mathbf{T}$=($t_1, t_2, \ldots, t_H$) [H componentes en $\mathbf{X}p$ via NIPALS]

1. for h=1,2,...,H
   $\forall p, p_{hp} = r(X_p, t_h)$ [correlación p-v.cuantitativas con t_h]
   $\forall q, p_{hq} = r(\hat{X}_q, t_h)$; $\hat{X}_q = P_{\tilde{X}_q}.t_h$, [razón correlación v.cualit]
   $P_h \leftarrow (p_{h1}, \ldots, p_{hp}, p_{hq_1}, \ldots, p_{hq})$, normar P_h
   $P[, h] \leftarrow P_h$
   end h

2. repetir
   2.1. for $q = 1, 2, \ldots, Q$
   $f(t_H) = p_{1q}t_1 + p_{2q}t_2 + \ldots + p_{Hq}t_H = \gamma$ [función cuantificn]
   $\hat{X}_q = P_{\tilde{X}_q} \cdot \gamma$ [cuantificación estandarizada]
   $\mathbf{X}C[, p + q] \leftarrow \hat{X}_q$
   end q
   2.2. for $h = 1, 2, \ldots, H$ [actualizar $\mathbf{P}$, $\mathbf{T}$: función NIPALS]
   $P_h = \mathbf{X}C' \cdot \mathbf{T}[, h]$, normar P_h
   $t_h = \underset{\sim}{\mathrm{X}}C \cdot P_h$
   $P[, h] \leftarrow P_h$
   $T[, h] \leftarrow t_h$
   $\mathbf{X}C_h \leftarrow \mathbf{X}C$, $\mathbf{X}C \leftarrow \mathbf{X}C_h - t_h \cdot P_h'$ [deflactar]
   end h
   hasta convergencia de P_h

La fase 1 inicia obteniendo H componentes principales de $\mathbf{X}_k$ 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 las variables cualitativas (ver fase 2.1). En el caso que se requiera cuantificación con orden se utiliza la ecuación (4). En la fase 2.2 se recalcula las matrices de vectores propios $\mathbf{P}$ y de componentes $\mathbf{T}$, y se iteran están dos ultimas fases hasta la convergencia de los P_h.

3 Aplicación

La base de datos cuantitativa (vinos) utilizada como ejemplo de aplicación se encuentra descrita por Escofier y Pagès (1992) [2], fue complementada con las variables cualitativas denominación de origen *Appe* y tipo de suelo *Terr* que

contienen la siguiente codificación sin considerar orden:

$$\text{Appe:}(1,1,2,3,1,2,2,1,3,1,2,1,1,1,1,2,3,2,3,1,1)$$

$$\text{Terr:}(2,2,2,3,1,1,1,2,2,1,2,3,3,3,1,1,1,2,3,4,4)$$

.

Appe contiene tres categorías con los siguientes significados $1 =$ *saumur*, $2 =$ *bourgueil* y $3 =$ *chinon*; mientras *Terr* contiene $1 =$ *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 $\mathbf{X}C_k$; así,

$\mathbf{X}p:$ gustacion # contiene los datos cuantitativos $(n = 21, p = 9)$

$\mathbf{X}q:$ denom.f # tabla con los factores *Appe* y *Terr*.

De acuerdo con la ecuación (13), el proceso comienza con la descomposición singular vía NIPALS de la matriz $\mathbf{X}p$ 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 obtengo las matrices cuantificadas con $1, 2, \ldots,$ 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 2 y con ellos se obtiene la Figura 1 de inercia maximal acumulada de la cual se determina la dimensionalidad "optimal" de la función de cuantificación principal.

4 Resultados – datos *gustación*

Los 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.

De la Figura 1, 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 $\mathbf{X}C_5$ cuantificada bajo $f(t_5)$ se muestran en el extremo derecho de la Tabla 2. Observe que hasta el eje 5 se

	Cuantificación $f(t_{1.})$			Cuantificación $f(t_{2.})$			Cuantificación $f(t_{3.})$		
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	0.8528	1.32232	9.638	**0.8762**
4	0.69502	9.924	0.9022	062546	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	1.000	0.01933	11.000	1.0000	0.01726	11.000	1.0000

	Cuantificación $f(t_{4.})$			Cuantificación $f(t_{5.})$			Cuantificación $f(t_{9.})$		
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 2: Máxima Inercia acumulada $(0.5508, 0.7637, 0.8762, \ldots)$ de las matrices cuantificadas con $f(t_{1.}) = t_{1.}, f(t_2) = t_{1.} + t_{2.}, ft_3 = t_{1.} + t_{2.} + t_{3.} \ldots$ respectivamente.

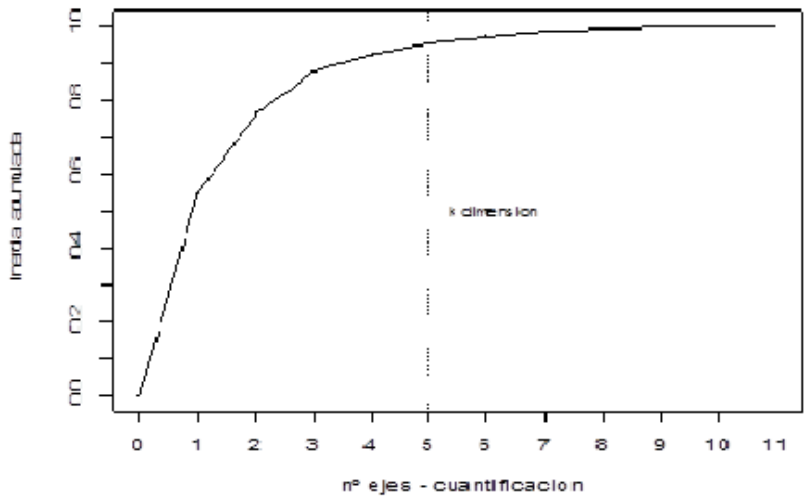


Figura 1: Inercia acumulada en los planos de dimensión $1, 2, 3, \ldots, 9$ base gustación.

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 ecuación (14).

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 $f(t^5_{4.})$ que contiene las primeras cuatro componentes derivadas de la descomposición espectral de $\mathbf{X}C_5$ y se obtienen los resultados de la Tabla 3.

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 3: Cuantificación $f(t^5_{4.})$.

En la fila `ratio` de la Tabla 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 $f(t^5_{4.})$, que toma las cuatro primeras componentes finales. Observe, que antes del quinto eje en el que la inercia es igual, 0.9547, la inercia acumulada eje por eje es prácticamente igual o superior, y esta característica se mantiene hasta el último eje; esto sugiere que cuatro componentes serán

Variable/Categoría	1	2	3	4
Appe	-0.654	-0.142	2.011	
Terr	-0.790	-0.255	0.346	2.791

Tabla 4: Valores de cuantificación asociados a las categorías de Appe y Terr.

suficientes en la cuantificación.

De hecho, al revisar la inercia obtenida con la cuantificación $f(t_{4.})$ en la Tabla 2, se tiene que ésta efectivamente ya contiene la inercia derivada de las cuantificaciones $f(t_5)$ y $f(t^5_{4.})$, con lo cual $f(t_{4.}) = t_{1.} + t_{2.} + t_{3.} + t_{4.}$ es la función de cuantificación definitiva, la cual genera inercia maximal por valor de 0.9228 en el plano de igual dimensión.

En la Tabla 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.

La matriz de correlaciones de las variables incluyendo las cuantificadas con los primeros cuatro ejes (Tabla 5) es muy importante, porque permite identificar con cuales de ellos existe mayor relación lineal y por tanto contribuyen más a su formación.

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 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 $\lambda_1 = 5.80041$, $\lambda_2 = 2.45020$, $\lambda_3 = 1.36351$, $\lambda_4 = 0.53656$ es maximal ya que ha sido generada por la función de cuantificación de igual dimensión $f(t_4)$. De la ecuación (9), 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

$$r(Appe, t1234.) = 0.9806 \text{ y } r(Terr, t12345.) = 0.9773,$$

entonces:

$$\begin{aligned} 0.9806^2 &= (-0.3222^2) + (-0.02425^2) + 0.849134^2 + 0.36884^2 \\ 0.9538^2 &= (-0.2877^2) + 0.86729^2 + (-0.311385^2) + (-0.15213)^2. \end{aligned}$$

Del círculo de correlaciones (Figura 2) y del análisis de las contribuciones en el primer plano factorial, las variables comprometidas en la formación de

	[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 5: Correlación entre las variables y los cuatro primeros ejes.

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".

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 esta 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, Figura 3, 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.

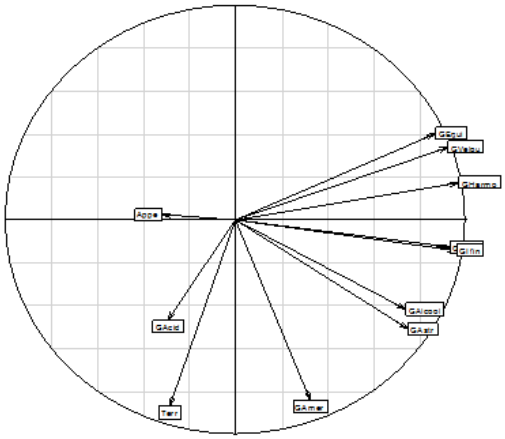


Figura 2: Círculo de correlaciones gustación; horizontal = eje1, vertical = eje2.

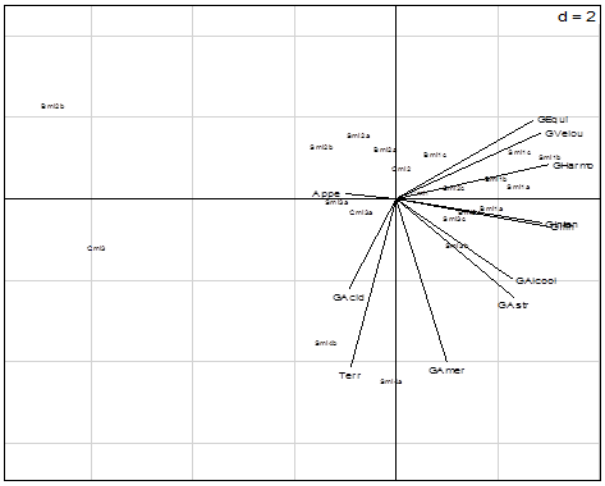


Figura 3: Representación simultánea en primer plano factorial de la matriz Gustacion.

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 Figura 3.

5 Conclusiones

a) Se desarrolla el método GNM-NIPALS para cuantificar óptimamente cada una de las variables cualitativas mediante una función lineal de k componentes, tal que formen parte activa del ACP en un conjunto de datos mixtos.

b) La función de cuantificación $f(t) = p_{1q}t_1 + p_{2q}t_2 + \ldots + p_{kq}t_k$ se basa en el concepto de reconstitución del ACP, e induce máxima inercia en el plano k-dimensional asociado a la matriz así cuantificada.

c) Si existe saturación de inercia, el análisis SIE permite cuantificar con una función de menor dimensión, es decir con $f(t) = p_{1q}t_1 + p_{2q}t_2 + \ldots + p_{k^*q}t_{k^*}$, $k^* < k$. La base *gustación* presenta SIE con lo cual solo se requieren cuatro y no cinco componentes, y la función de cuantificación definitiva es de la forma, $f(t) = t_{1.} + t_{2.} + t_{3.} + t_{4.}$.

d) El ACP de la base *gustación*, 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.

Referencias

[1] Aluja, T.; Morineau, A. (1999) *Aprender de los Datos: El Análisis de Componentes Principales*. EUB S.L, Barcelona.

[2] Escofier, B.; Pàges, J. (1992) *Análisis Factoriales Simples y Múltiples: Objetivos, Métodos e Interpretación*. Servicio Editorial Universidad del País Vasco, Bilbao.

[3] Lebart, L.; Morineau, A.; Piron, M. (2006) *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

[4] Russolillo, G. (2012) "Non-metric partial least squares", *Electronic Journal of Statistics* **6**: 1648–1655.

[5] Saporta, G. (2011) *Probabilités, Analyse des Données et Statistique*. Editions Technip, Paris.

[6] Tenenhaus, M. (1998) *La Régression PLS. Théorie et Pratique*. Editions Technip, Paris.

[7] Wold, H. (1975) “Path models with latent variables: The non-linear iterative partial least squares (NIPALS) approach”, Academic Press, New York: 307–357.

[8] Young, F. (1981) “Quantitative analysis of qualitative data”, *Psychometrika* **44**(4): 357–388.