Análisis de Correspondencias Múltiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo NIPALS (ACMpdd)

Andrés Felipe Ochoa Muñoz
Director: PhD. Victor Manuel Gonzalez

Maestría en Estadística
UNIVERSIDAD DEL VALLE

SUSTENTACIÓN DE TRABAJO DE GRADO

Junio 2018

Tabla de Contenido

Tabla de Contenido

Tabla de Contenido

Tabla de Contenido

Tabla de Contenido

Tabla de Contenido

Tabla de Contenido

Tabla de Contenido

Introducción

En la actualidad, cuando se estudia algún fenómeno de las ciencias experimentales, se toman mediciones a distintas variables sobre muchas unidades de observación dando origen a grandes volúmenes de datos. Los métodos estadísticos multivariados son apropiados en estas situaciones, ya que, analizan simultáneamente toda la información [Aluja and Morineau, 1999]

En algunas circunstancias estas variables en su mayoría suelen ser de tipo cualitativo, de tal forma que si se quiere realizar análisis con este tipo de variables se debe contar con los métodos adecuados. Un método que se conoce para analizar este tipo variables es el Análisis de Correspondencias Múltiples (ACM), sin embargo este método clásico solo trabaja con información completa, es decir, no permite la presencia de datos faltantes

El Análisis de Correspondecias Múltiple en presencia de datos faltantes usualmente se trabaja eliminando los registros en donde exista el dato faltante o no disponible (NA), algunas veces se elimina toda la fila o toda la columna de la matriz de datos, lo cual no es adecuado ya que al realizarlo se pierde información relevante sobre algún individuo o variable del estudio.

Otras veces se puede asumir el dato faltante como una modalidad de la variable.

Algunas Soluciones: Una solución para esta situación puede ser la imputación del dato faltante o utilizar un algoritmo que permita trabajar con la presencia de éste tipo de datos (principio de datos disponibles).

Tabla disjuntiva completa para ACM con NA

Individuo	Género		Religión			Raza		
1	1	0	1	0	0	1	0	0
2	1	0	na	na	na	1	0	0
.	0	1	0	1	0	0	0	1
.	na	na	0	1	0	na	na	na
.	na	na	na	na	na	1	0	0
.	1	0	0	1	0	0	1	0
n	1	0	1	0	0	na	na	na

Requerimientos del ACM clásico

- La base de datos debe tener los registros completos
- El número de observaciones n debe superar el número de variables p

Propuestas para trabajar en presencia de datos faltantes

Una propuesta que se utiliza para trabajar el análisis multivariado en presencia de datos faltantes es el algoritmo NIPALS [Wold, 1975].

Este método se basa en el ACP y tiene la ventaja de que es un algoritmo iterativo, lo que hace que el investigador pueda intervenir dicho algoritmo, de acuerdo a los objetivos de su análisis (Datos faltantes, Multicolinealidad). NIPALS es la base de la regresión PLS, [Tenenhaus, 1998].

Para trabajar datos faltantes

Actualmente existen varios autores trabajando el algoritmo NIPALS en análisis multivariado [Russolillo, 2009], [Aluja and González, 2014], [Trinchera et al. [Sanchez, 2013] y otros trabajando bajo el enfoque de la imputación de datos con el algoritmo EM [Josse et al., 2012], [Audigier et al., 2015].

No se sabe exactamente que enfoque genere mejores resultados, sin embargo se han encontrado trabajos donde se realiza la comparación para el caso de ACP [Vitelleschi and Quaglino, 2009].

En el caso de Análisis de Correspondencias Múltiple los autores [Josse et al., 201 [Audigier et al., 2015] han trabajado con el enfoque del algoritmo EM y no se conoce de trabajos o ideas que intenten trabajar ACM bajo NIPALS.

Por esta razón esta propuesta investigación daría a entender de mejor manera cómo se puede trabajar los datos faltantes con el ACM, lo cual sería un gran aporte a los métodos multivariantes.

- ¿Cómo realizar el ACM en presencia de datos faltantes usando el principio de datos disponibles?
- ¿Que ventajas o desvantajas existen entre un ACM con el principio de datos disponibles y un ACM bajo el enfoque de imputación?

En el caso de Análisis de Correspondencias Múltiple los autores [Josse et al., 201 [Audigier et al., 2015] han trabajado con el enfoque del algoritmo EM y no se conoce de trabajos o ideas que intenten trabajar ACM bajo NIPALS.

Por esta razón esta propuesta investigación daría a entender de mejor manera cómo se puede trabajar los datos faltantes con el ACM, lo cual sería un gran aporte a los métodos multivariantes.

- ¿Cómo realizar el ACM en presencia de datos faltantes usando el principio de datos disponibles?
- ¿Que ventajas o desvantajas existen entre un ACM con el principio de datos disponibles y un ACM bajo el enfoque de imputación?

Objetivos

Objetivo General

Desarrollar el método ACMpdd mediante principio de datos disponibles del algoritmo NIPALS para matrices con datos faltantes

Objetivos especificos

- Identificar y realizar las expresiones exactas para encontrar las matrices simétricas en R^p y en R^n sobre las cuales se implementa el algoritmo NIPALS para el Análisis de Correspondencias Múltiples
- Desarrollar las funciones bajo el lenguaje del software R, para el método ACMpdd usando el principio de datos disponibles
- Analizar como influye el porcentaje de datos faltantes en los resultados del análisis factorial: inercia, representación en los planos factoriales, relaciones de transición, ortogonalidad en los ejes, entre otros.
- Evaluar si las propiedades de inercia total, por modalidad y pregunta se conservan o se pierden al trabajar con el principio de datos disponibles
- Comparar el método propuesto ACMpdd contra un método de imputación de datos bajo el algoritmo EM (ACM-EM)

Objetivos especificos

- Identificar y realizar las expresiones exactas para encontrar las matrices simétricas en R^p y en R^n sobre las cuales se implementa el algoritmo NIPALS para el Análisis de Correspondencias Múltiples
- Desarrollar las funciones bajo el lenguaje del software R, para el método ACMpdd usando el principio de datos disponibles
- Analizar como influye el porcentaje de datos faltantes en los resultados del análisis factorial: inercia, representación en los planos factoriales, relaciones de transición, ortogonalidad en los ejes, entre otros.
- Evaluar si las propiedades de inercia total, por modalidad y pregunta se conservan o se pierden al trabajar con el principio de datos disponibles
- Comparar el método propuesto ACMpdd contra un método de imputación de datos bajo el algoritmo EM (ACM-EM)

Objetivos especificos

- Identificar y realizar las expresiones exactas para encontrar las matrices simétricas en R^p y en R^n sobre las cuales se implementa el algoritmo NIPALS para el Análisis de Correspondencias Múltiples
- Desarrollar las funciones bajo el lenguaje del software R, para el método ACMpdd usando el principio de datos disponibles
- Analizar como influye el porcentaje de datos faltantes en los resultados del análisis factorial: inercia, representación en los planos factoriales, relaciones de transición, ortogonalidad en los ejes, entre otros.
- Evaluar si las propiedades de inercia total, por modalidad y pregunta se conservan o se pierden al trabajar con el principio de datos disponibles
- Comparar el método propuesto ACMpdd contra un método de imputación de datos bajo el algoritmo EM (ACM-EM)

Objetivos especificos

- Identificar y realizar las expresiones exactas para encontrar las matrices simétricas en R^p y en R^n sobre las cuales se implementa el algoritmo NIPALS para el Análisis de Correspondencias Múltiples
- Desarrollar las funciones bajo el lenguaje del software R, para el método ACMpdd usando el principio de datos disponibles
- Analizar como influye el porcentaje de datos faltantes en los resultados del análisis factorial: inercia, representación en los planos factoriales, relaciones de transición, ortogonalidad en los ejes, entre otros.
- Evaluar si las propiedades de inercia total, por modalidad y pregunta se conservan o se pierden al trabajar con el principio de datos disponibles
- Comparar el método propuesto ACMpdd contra un método de imputación de datos bajo el algoritmo EM (ACM-EM)

Objetivos especificos

- Identificar y realizar las expresiones exactas para encontrar las matrices simétricas en R^p y en R^n sobre las cuales se implementa el algoritmo NIPALS para el Análisis de Correspondencias Múltiples
- Desarrollar las funciones bajo el lenguaje del software R, para el método ACMpdd usando el principio de datos disponibles
- Analizar como influye el porcentaje de datos faltantes en los resultados del análisis factorial: inercia, representación en los planos factoriales, relaciones de transición, ortogonalidad en los ejes, entre otros.
- Evaluar si las propiedades de inercia total, por modalidad y pregunta se conservan o se pierden al trabajar con el principio de datos disponibles
- Comparar el método propuesto ACMpdd contra un método de imputación de datos bajo el algoritmo EM (ACM-EM)

Antecedentes

Antecendentes

En la literatura de datos faltantes en el contexto del Análisis Multivariado se encontraron diferentes enfoques:

- Métodos de tablas incompletas ("missing passive", "missing passive modified margin", etc)
- Métodos de imputación (ACP iterativo, ACP iterativo regularizado, ACM iterativo, etc)
- Métodos que usan el principio de datos disponibles de NIPALS (NM-NIPALS, GNM-NIPALS, etc)

Métodos de Tablas Incompletas

En el contexto de Análisis de Correspondencias con datos faltantes, se encontrarón algunos métodos, los cuales hacen referencia al manejo de los datos faltantes dentro de la Tabla Disjuntiva Completa, en particular, los datos faltantes se asumen cómo valores que no existen, por ende en la variable indicadora se asumen como 0 [Van der Heijden and Escofier, 2003]

El primer método que se menciona es "missing passive", que se puede denominar como un método de tablas incompletas, ya que la marginal por fila no es igual al número de variables s para cada individuo, es decir la marginal por fila ya no es constante. Este enfoque fue propuesto por [Benzécri et al., 1973] y las propiedades fueron estudiadas por [Meulman, 1982]

Métodos de Tablas Incompletas

En el artículo desarrollado por [Van der Heijden and Escofier, 2003], se muestran otros método cómo: el "missing passive modified margin", el cual cómo su nombre lo dice modifica la marginal para que sea constante [Escofier, 1981].

En el artículo se encuentrán más métodos, algunos utilizan los datos faltantes como si fueran una modalidad de la variable cualitativa ("missing insertion") [Nishisato, 1980], también se hace referencia a la teoría propuesta por [Rubin, 1976].

ACP iterativo: "Handling missing values in exploratory multivariate data analysis methods"

[Josse and Husson, 2012]: Presentan dos métodos para trabajar el Análisis de Componentes Principales, los métodos que se presentan tienen en cuentan los principios de imputación estudiados por [Rubin, 1976]. Se presenta el algoritmo ACP Iterativo y el ACP Iterativo Regularizado. Se muestrán los algoritmos y se da una idea de como trabajarlos usando imputación simple y múltiple bajo el algoritmo EM. Además se muestra un caso aplicado que se trabaja con el **paquete missMDA** del software R [Husson et al., 2013].

ACM iterativo: "Handling missing values with regularized iterative multiple correspondence analysis"

[Josse et al., 2012]: En este trabajo se propuso un algoritmo, llamado Análisis de Correspondencias Múltiple Iterativo, el cual es útil para manejar valores faltantes en el Análisis de Correspondencias Múltiples (ACM). Este algoritmo, se basa en un algoritmo ACP Iterativo, en el artículo se describen los algoritmos y se estudian sus propiedades. Los algoritmos son implementados en el **paquete missMDA** del software R husson2013package.

En este artículo se comparan los métodos: ACM iterativo, el ACM iterativo regularizado y algunos métodos que de tablas incompletas, como lo son: "missing passive", "missing passive modified margin" "missing fuzzy", donde se encontrarón mejores resultados para los métodos ACM iterativo y ACM iterativo regularizado.

Marco Teórico

Tabla Disjuntiva Completa Z_{ij} para ACM clásico

Individuo	Z_A		Z_b			Z_c			$Z_{i.}$
1	1	0	1	0	0	1	0	0	s
2	1	0	1	0	0	1	0	0	s
.	.	.	.	.	.	.	.	.	s
.	.	.	.	.	.	.	.	.	s
n	1	0	0	1	0	0	1	0	s
$Z_{.j}$	$Z_{,1}$	$Z_{,2}$	$Z_{,3}$	$Z_{,4}$	.	.	.	$Z_{.p}$	ns

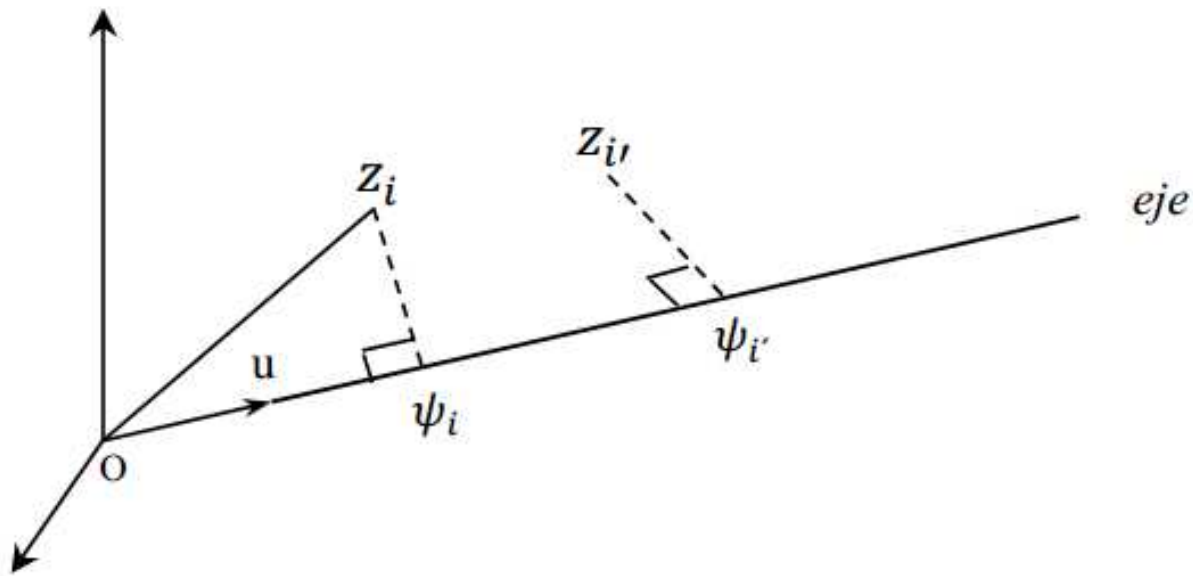
$M_p = [\ddots 1/f_{.j} \ddots]$; $f_{.j} = Z_{.j}/ns$

$M_n = [\ddots 1/f_{i.} \ddots] = nI$;

$F = \frac{1}{ns} Z$

Maximización y matriz a diagonalizar

El Objetivo geométrico del ACM es buscar un nuevo sistema de ejes ortogonales u_α, los que se proyecte la inercia de la nube de individuos, tal que los primeros ejes concentren la mayor parte de la misma y en forma decreciente.



Maximización y matriz a diagonalizar

$$I = \psi' M_n^{-1} \psi; \qquad \psi = M_n F M_p u$$

$$I = u' M_p F' M_n M_n^{-1} M_n F M_p u$$

$$I = u' M_p F' M_n F M_p u; \qquad S = F' M_n F M_p$$

$$I = u' M_p S u; \qquad u' M_p u = 1$$

Maximización y matriz a diagonalizar

Por tanto en el espacio de los individuos R^p, para encontrar los ejes directores u_α

se diagonaliza la matriz S

$$S = F' M_n F M_p \text{ tal que } u' M_p u = 1$$

S No necesariamente es simétrica y por ende no se garantiza que los vectores propios sean ortonormales.

Maximización y matriz a diagonalizar

Si en vez de diagonalizar S se diagonaliza la matriz:

$$S^* = M_p^{1/2} F' M_n F M_p^{1/2}; \qquad w = M_p^{1/2} u$$

$$S^* = S_o' S_o; \qquad S_o = M_n^{1/2} F M_p^{1/2}$$

$$S^* = S_o' S_o; \quad en \quad R_p; \quad w' w = 1$$

$$T^* = S_o S_o'; \quad en \quad R_n; \quad r' r = 1$$

ACM relación ACP

En ACS se vio la relación que hay con ACP, esa relación es importante y tambien se utiliza en ACM, de tal forma que

$$S_o = M_n^{1/2} F M_p^{1/2}$$

es lo mismo que tener la matriz Z del ACP, por lo cual se puede hacer un ACP (no centrado ni estandarizado) a dicha matriz, para así obtener los resultados del ACM.

Inercia Total

La inercia total I no tiene un significado estadístico interesante, por que depende del número de modalidades y variables y no de las relaciones entre variables:

$$I = \sum_{q}^{s} I_q = \frac{p}{s} - 1 = \sum_{j=1}^{p} I_j = \sum_{q=1}^{s} I_q$$

Inercia por modalidad

$$I_j = f_{.j} d^2(j, G) = \frac{Z_{.j}}{ns}\left(\frac{n}{Z_{.j}} - 1\right) = \frac{1}{s}\left(1 - \frac{Z_{.j}}{n}\right)$$

La inercia de una modalidad es más grande si la modalidad es rara; es decir si tiene frecuencia muy baja.

Inercia por pregunta

La inercia debida a una pregunta (subtabla) q es función creciente de su número de modalidades p_q

$$I_q = \sum_{j}^{p_q} I_j = \frac{1}{s}(p_q - 1)$$

Se debe equilibrar el número de modalidades por variable y evitar que sean artificialmente activas.

Nonlinear estimation by Iterative Partial Least Square (NIPALS)

NIPALS es la base de la regresión PLS, [Tenenhaus, 1998]. Fundamentalmente realiza una descomposición singular de la matriz de datos, mediante secuencias iterativas de proyecciones ortogonales (concepto geométrico de regresión) obtenidas como productos escalares.

Cuando la base de datos esta completa hay una equivalencia con los resultados del ACP, además se puede trabajar con datos faltantes y obtener estimaciones de la matriz de datos reconstituida.

El algoritmo NIPALS

Para la matrix de datos $X_{n,p}$ de rango a cuyas columnas $X_1, \ldots, X_p$ se suponen centradas o estandarizadas, la descomposición derivada del ACP permite la reconstitución mediante lo siguiente:

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

t_h es la h-ésima componente principal y representa los scores.
P_h es el vector propio en el eje h.

$$[X_1, ..., X_p] = t_1 P_1' + ... + t_a P_a'$$

$$X_j = \sum_{h}^{a} t_h phj, j = 1, ..., p$$

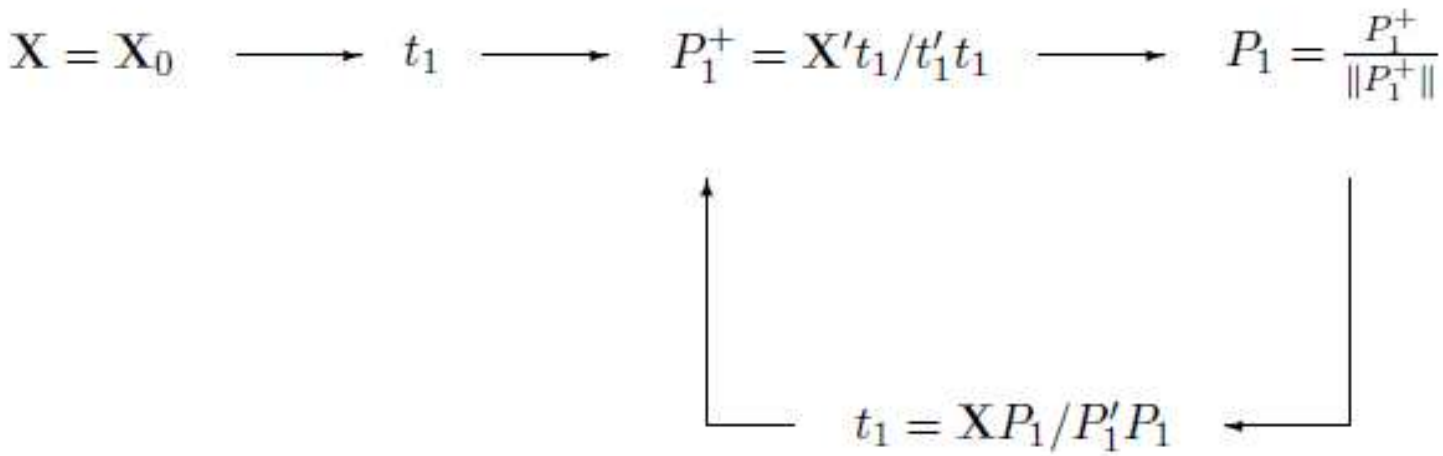
$$X_i = \sum_{h}^{a} t_h phi, i = 1, ..., n$$

El algoritmo inicia tomando la primera columna de X_0 como la 1^a componente principal t_1.

Luego se construirán una serie de tablas deflactadas notadas $X_h = X_0 - t_h P_h'$ las cuales permiten reiniciar el ciclo y obtener las componentes (ortogonales) restantes $t_2, .., t_h$ y sus respectivos vectores propios $P_1, ..., P_h$

Como se observa 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

El Algoritmo NIPALS - Esquema



El criterio de convergencia:

$$||t^{l+1} - t^{l}||^2 < \epsilon$$

El Algoritmo NIPALS

Etapa 1: $X_0 = X_h$
Etapa 2: $h = 1, 2, ..., a$

Etapa 2.1 $t_h = 1^a$ primera columna de X_{h-1}
Etapa 2.2 Repetir hasta la convergencia de P_h
Etapa 2.2.1 $P_h = \frac{X'_{h-1} t_h}{t'_h t_h}$
Etapa 2.2.2 Normar P_h a 1
Etapa 2.2.3 $t_h = X_{h-1} P_h / P'_h P_h$

Etapa 2.3 $X_h = X_{h-1} - t_h P'_h$ (garantiza la ortogonalidad)
Siguiente h

Principio de datos disponibles

Este principio hace referencia a que uno puede realizar algunas operaciones entre vectores omitiendo los NA´s y trabajando con los puntos emparajedos disponibles. Es decir que si se tiene un vector X y un vector Y (ambos con NA´s), el producto interno entre los vectores usando el principio de datos disponibles, sería de la siguiente forma:

$$X = \begin{pmatrix} x_1 \\ NA \\ x_3 \\ \vdots \\ x_n \end{pmatrix} \qquad Y = \begin{pmatrix} NA \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}$$

Entonces $\langle x, y \rangle = \sum_{x;y \ existent} x_i y_i = x_3 y_3 + x_4 y_4 + \ldots + x_n y_n$

El Algoritmo NIPALS - Datos Faltantes

Etapa 1: $X_0 = X_h$

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

Etapa 2.1 $t_h = 1^a$ primera columna de X_{h-1}

Etapa 2.2 Repetir hasta la convergencia de P_h

Etapa 2.2.1: Para $j = 1, 2, .., p$

$$p_{hj} = \frac{\sum_{i:x_{ji} ett_{hi} existent} x_{h-1,ji} t_{hi}}{\sum_{i:x_{ji} ett_{hi} existent} t_{hi}^2}$$ 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: $X_h = X_{h-1} - t_h p_h^{'}$

ACMpdd: ACM usando el principio de datos disponibles de NIPALS

ACM bajo el principio de datos disponibles (ACMpdd)

Para realizar un ACMpdd en primer lugar, se construye la Tabla Disjuntiva Completa $Z^*_{n.p}$ la cual contiene datos faltantes, como se observa:

$$Z^*_{ij} = \begin{cases} 1; \ \textit{Si la Modalidad existe} \\ 0; \ \textit{Si la Modalidad no existe} \\ \textit{NA}; \ \textit{Si la Modalidad tiene dato faltante} \end{cases}$$

ACMpdd pseudo-algoritmo

1. Se construye la Tabla disjuntiva con NA (Z_{ij}^*)
2. Se construye $F^* = \frac{Z^*}{k^*}$. ($k^* = ns^*$; disponible)
3. Se contruye la matriz S_0^* usando el principio de datos disponibles

$$S_0^* = M_n^{*1/2} F^* M_p^{*1/2}$$

$M_p^* = [\ddots 1/f_{.j} \ddots]$; $M_n^* = [\ddots 1/f_{i.} \ddots]$

4. Se aplica un NIPALS (no estandarizado) para la matriz S_0^*

ACMpdd

$$S^* = M_p^{*1/2} F^{*'} M_n^* F^* M_p^{*1/2}; \qquad w = M_p^{*1/2} u$$

$$S^* = S_o^{*'} S_o^*; \qquad S_o^* = M_n^{*1/2} F^* M_p^{*1/2}$$

$$S^* = S_o^{*'} S_o^*; \quad en \quad R_p; \quad w^{'} w = 1$$

$$T^* = S_o^* S_o^{*'}; \quad en \quad R_n; \quad r^{'} r = 1$$

Para que las relaciones de transición sean legitimas los λ_α deben ser equivalentes en ambos espacios.

Componentes en R_n y R_p

Para realizar el cálculo de las componentes $\psi_{n.p}$ en R_p, se realiza el producto punto de la matriz S_o con el vetor propio asociado al espacio de las variables $w = M_p^{1/2}u$. Y para calcular las componentes $\varphi_{p.p}$ en R_n, se realiza el producto punto de la matriz T_o con el vector propio asociado al espacio de los individuos $r = M^{1/2}v$. De tal forma que se tienen las siguientes expresiones:

$$\psi = S_o w$$

$$\varphi = T_o r; \qquad T_o = S_o^{'}$$

Expresiones de Inercias

Inercia Total

$$I = \frac{p}{s^*} - 1; dondes^* = \frac{\sum z_{i.}^*}{n}$$

Inercia por modalidad

$$I_j = \frac{1}{s^*}(1 - \frac{Z_{.j}^*}{n})$$

Inercia por pregunta

$$I_q = \frac{1}{s^*}(p_q^* - (\frac{Z_{.j}^*}{n}))$$

Metodología

Metodología

El método propuesto fue aplicado a la base de datos ***BredsDogs*** $(27 * 6)$ de la librería **FactoClass** del software R [Pardo and Del Campo, 2007], con el fin de que el método pueda ser desarrollado y verificado por cualquier usuario.

También se trabajo en una base de datos de mayor dimensión $(300 * 11)$, para analizar el comportamiento del ACMpdd. La base de datos se denomina ***tea*** y se encuentra en la librería **FactoMineR** [Husson et al., 2013].

En cada conjunto de datos se realizará un ACM con datos completos, ACMpdd y un ACM iterativo (ACM-EM). Además de eso se compararán los métodos diferentes porcentajes de datos faltantes generados aletoriamente (MCAR).

Conjunto de datos BredsDogs (perros)

La base de datos perros contiene 27 razas y 6 variables cualitativas, las cuales son: Tamaño (TAM), Peso (PES), Velocidad (VEL), Inteligencia (INT), Afectividad (AFE), Agresividad (AGR), dichas variables tienen entre 2 y 3 modalidades, tal como se ilustra en la siguiente tabla 1:

Razas	TAM	PES	VEL	INT	AFE	AGR
bass	peq	liv	len	baj	baj	alt
beau	gra	med	alt	med	alt	alt
boxe	med	med	med	med	alt	alt
buld	peq	liv	len	med	alt	baj
bulm	gra	pes	len	alt	baj	alt
cani	peq	liv	med	alt	alt	baj
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabla: Visualización de la base de datos perros

Escenarios de Simulación en BreedsDogs

Estructura NA	Cantidad NAs	% NAs	s^*
MCAR	1 NA por fila	16.7 %	5
MCAR	2 NA por fila	33.3 %	4
MCAR	3 NA por fila	50 %	3
MCAR	0:1 NA por fila	9.26 %	5.44
MCAR	0:2 NA por fila	13.58 %	5.18
MCAR	0:3 NA por fila	27.16 %	4.37

Análisis del Poder Descriptivo

Estructura NA	Métodos	% NAs	m
BreedsDogs tea	ACMpdd ACM-EM	5 %	1000
		10 %	1000
		15 %	1000
		20 %	1000
		25 %	1000
		30 %	1000
		50 %	1000

Tabla: Escenarios de simulación para el Análisis del Poder Descriptivo

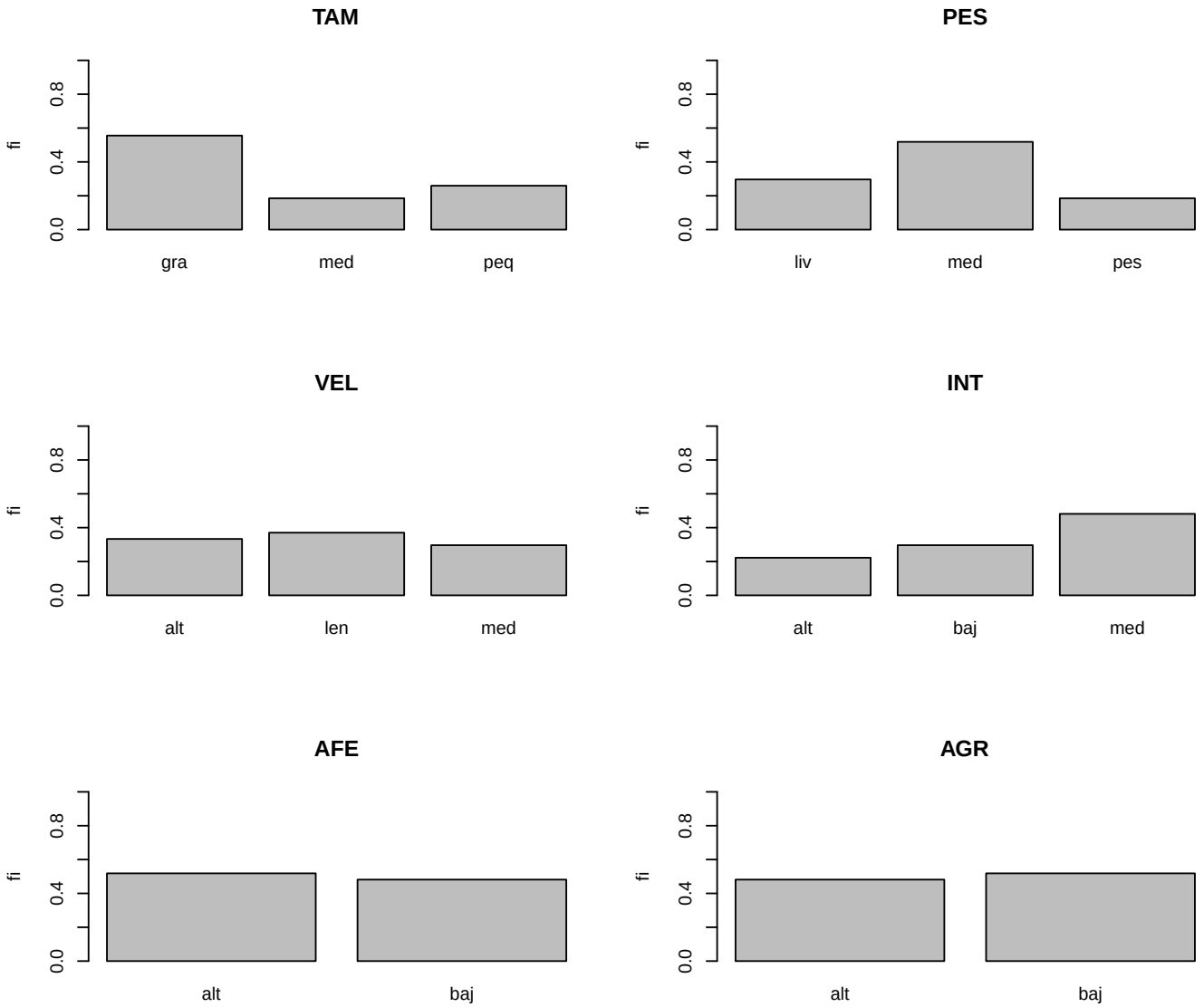
Análisis Estadístico de los escenarios de simulación

Se debe analizar la matriz con datos completos y las matrices asociadas a los escenarios de simulación. Analizando como se comportan cada uno de los siguientes indicadores:

- Valores λ y vectores propios u
- Componentes ψ, φ en R_n y R_p
- Inercia total, Inercia por modalidad y por pregunta.
- Poder descriptivo $(\lambda_1 + \lambda_2)/\sum \lambda$
- Planos factoriales
- Ortogonalidad en las componentes y Ortonormalidad en los vectores propios
- Correlación entre componentes con datos completos y faltantes.

Resultados

Descriptivas datos completos



$Z_{i.}$

ACMpdd: Caso 0:1 NA por fila

bass	beau	boxe	buld	bulm	cani	chih	cock	coll	dalm	dobe	dogo	foxh	foxt
5	5	5	5	5	5	5	6	5	5	5	6	5	5
galg	gasc	labr	masa	mast	peki	podb	podf	poin	sett	stbe	steck	tern	
6	6	6	6	6	5	6	6	6	5	6	6	5	

Datos completos

bass	beau	boxe	buld	bulm	cani	chih	cock	coll	dalm	dobe	dogo	foxh	foxt
6	6	6	6	6	6	6	6	6	6	6	6	6	6
galg	gasc	labr	masa	mast	peki	podb	podf	poin	sett	stbe	teck	tern	
6	6	6	6	6	6	6	6	6	6	6	6	6	

$Z_{.j}$

ACMpdd: Caso 0:1 NA por fila

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len	VEL.med
y	14	4	5	8	12	5	8	10	8
	AFE.alt	AFE.baj	AGR.alt	AGR.baj					
	12	12	13	14					

Datos completos

TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len
15	5	7	8	14	5	9	10
VEL.med	INT.alt	INT.baj	INT.med	AFE.alt	AFE.baj	AGR.alt	AGR.baj
8	6	8	13	14	13	13	14

Valores propios

ACMpdd: Caso 0:1 NA por fila

1	2	3	4	5	6	7	8
0.5087	0.4130	0.2562	0.1819	0.1600	0.1223	0.0979	0.0643
9	10	11	12	13	14	15	16
0.0464	0.0329	0.0213	0.0157	0.0101	0.0044	0.0012	

Datos completos

1	2	3	4	5	6	7	8
0.4816	0.3847	0.2109	0.1575	0.1501	0.1232	0.081	0.045
9	10	11	12	13	14	15	16
0.0235	0.0077						

Inercias (ACMpdd: Caso 0:1 NA por fila)

Inercia Total

$$I = 1{,}938776; \qquad (\sum \lambda = 1{,}936123)$$

Inercia por Pregunta

$$\sum I_q = [1{,}938776] = [0{,}3945 + 0{,}3809 + 0{,}3741 + 0{,}4013 + 0{,}2040 + 0{,}1836]$$

Inercia por Modalidad

$$\sum I_p = 1{,}938776$$

1	2	3	4	5	6	7	8
0.0884	0.1564	0.1496	0.1292	0.1020	0.1496	0.1292	0.1156
9	10	11	12	13	14	15	16
0.1292	0.1496	0.1360	0.1156	0.1020	0.1020	0.0952	0.0884

Inercias datos completos

Inercia Total

$$I = 1{,}6667$$

Inercia por Pregunta

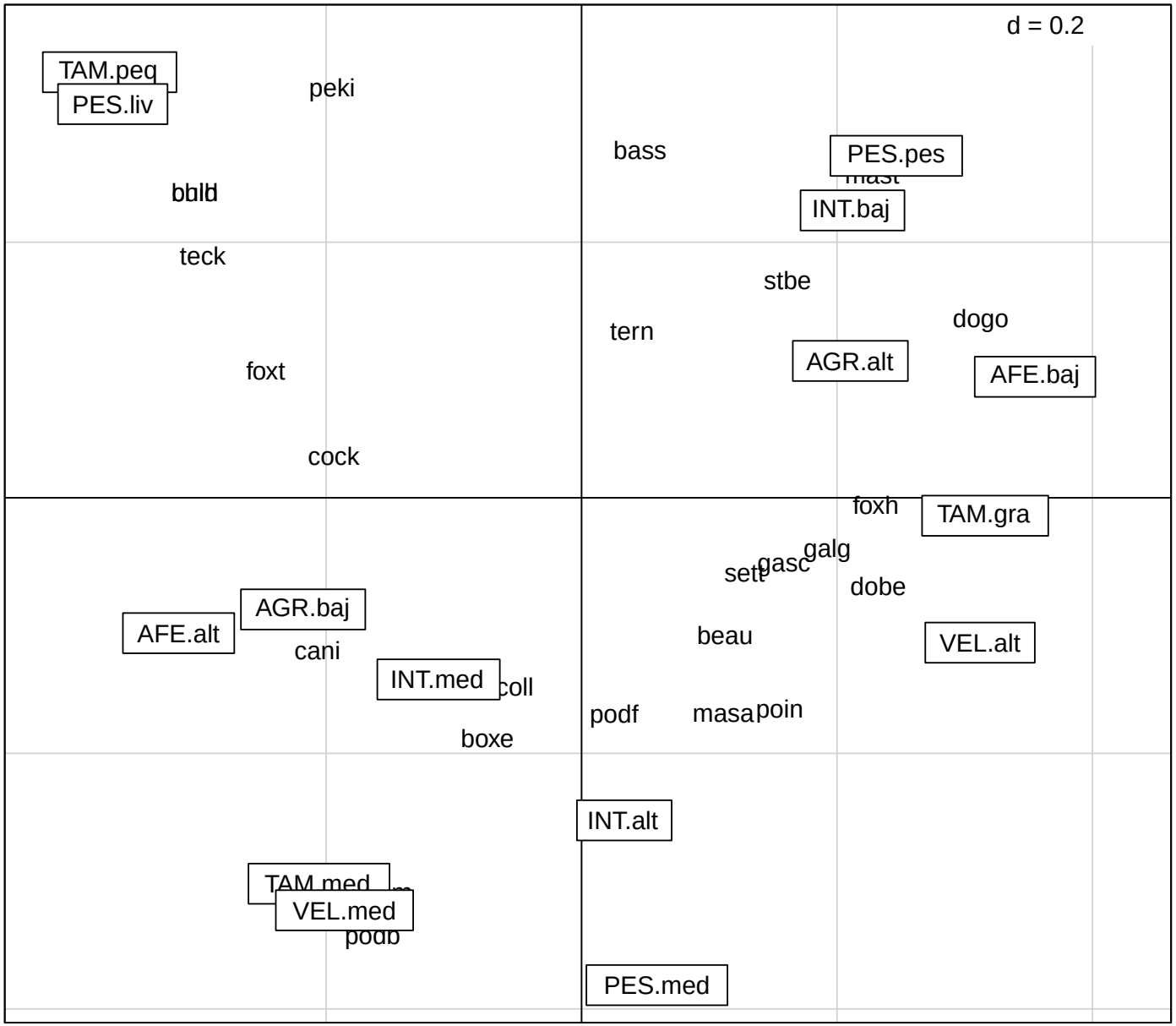
$$\sum I_q = [1{,}6667] = [0{,}333 + 0{,}333 + 0{,}333 + 0{,}333 + 0{,}16667 + 0{,}16667]$$

Inercia por Modalidad

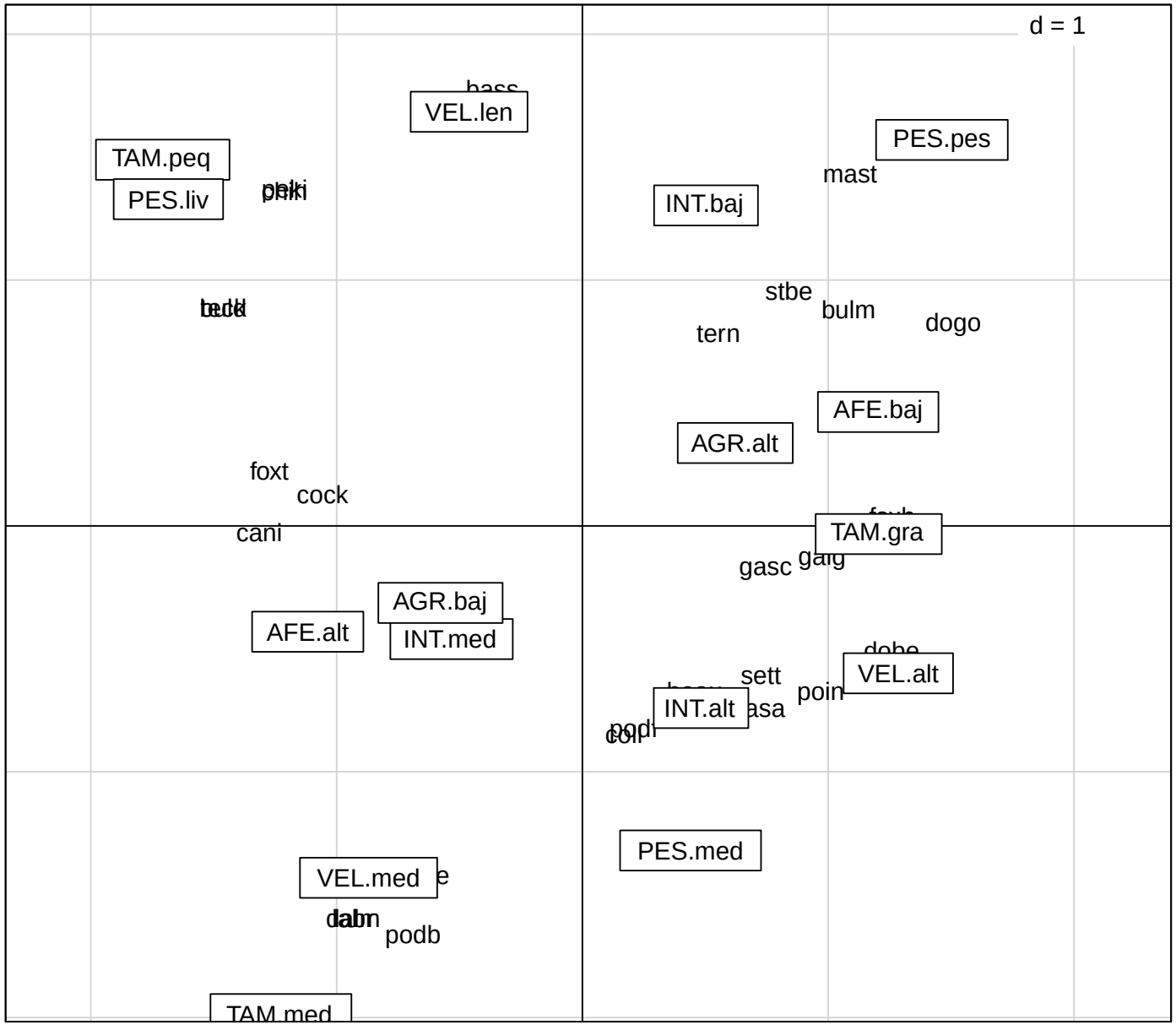
$$\sum I_p = [1{,}6667]$$

1	2	3	4	5	6	7	8
0.0740	0.1358	0.1234	0.1172	0.0802	0.1358	0.111	0.1049
9	10	11	12	13	14	15	16
0.1172	0.1296	0.1172	0.0864	0.0802	0.0864	0.0864	0.0802

Plano Factorial (Eje 1 y 2) ACMpdd (Caso 0:1)



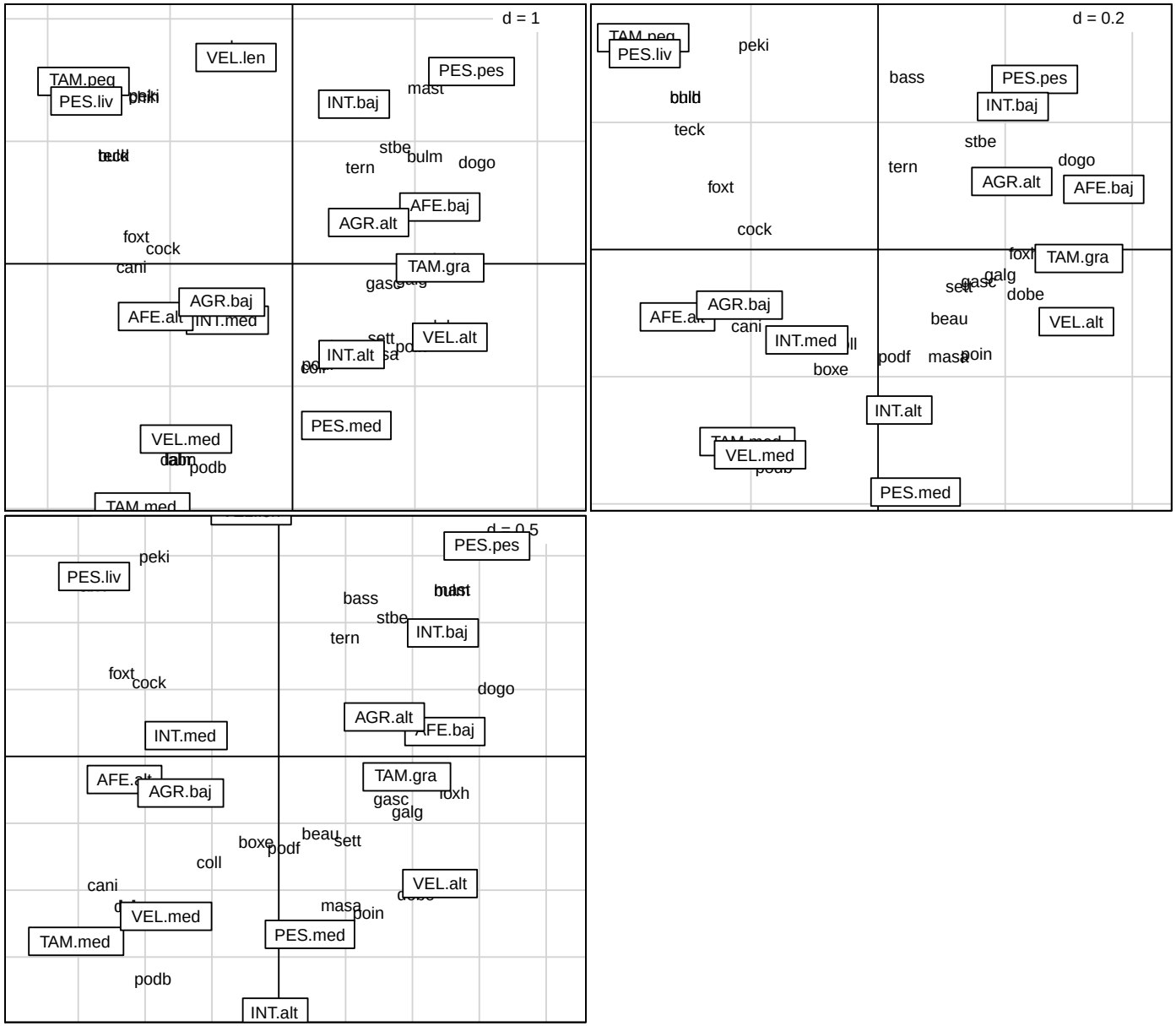
Plano Factorial (Eje 1 y 2) (Datos completos)



Resultados todos los escenarios

	Cantidad NAs	$\sum \lambda_\alpha$	Inercia Total	Poder descriptivo
Método	Datos completos	1.6667	1.6667	0.5198
ACMpdd	1 NA por fila	2.2	2.2	0.4094
	2 NA por fila	3	3	0.3466
	3 NA por fila	4.3333	4.3333	0.2864
	0:1 NA por fila	1.9361	1.9387	0.4760
	0:2 NA por fila	2.0650	2.0857	0.4375
	0:3 NA por fila	2.6700	2.6610	0.3756
ACM-EM	1 NA por fila	1.6667	1.6667	0.5381
	2 NA por fila	1.6667	1.6667	0.6077
	3 NA por fila	1.6667	1.6667	0.6683
	0:1 NA por fila	1.6667	1.6667	0.5410
	0:2 NA por fila	1.6667	1.6667	0.5001
	0:3 NA por fila	1.6667	1.6667	0.5394

Planos Factoriales (0:1 NA por fila)



Correlación entre componentes ψ en ACMpdd

$$cor(\psi_1, \psi_1 na)$$

	ψ_1 Completo		ψ_2 Completo		ψ_3 Completo
ψ_1 1 NA	0.9648	ψ_2 1 NA	0.9544	ψ_3 1 NA	0.5487
ψ_1 2 NA	0.8431	ψ_2 2 NA	0.8113	ψ_3 2 NA	0.1630
ψ_1 3 NA	0.8064	ψ_2 3 NA	0.3618	ψ_3 3 NA	0.2777
ψ_1 0:1 NA	0.9786	ψ_2 0:1 NA	0.9752	ψ_3 0:1 NA	0.8467
ψ_1 0:2 NA	0.9611	ψ_2 0:2 NA	0.9212	ψ_3 0:2 NA	0.7860
ψ_1 0:3 NA	0.9214	ψ_2 0:3 NA	0.8482	ψ_3 0:3 NA	0.7154

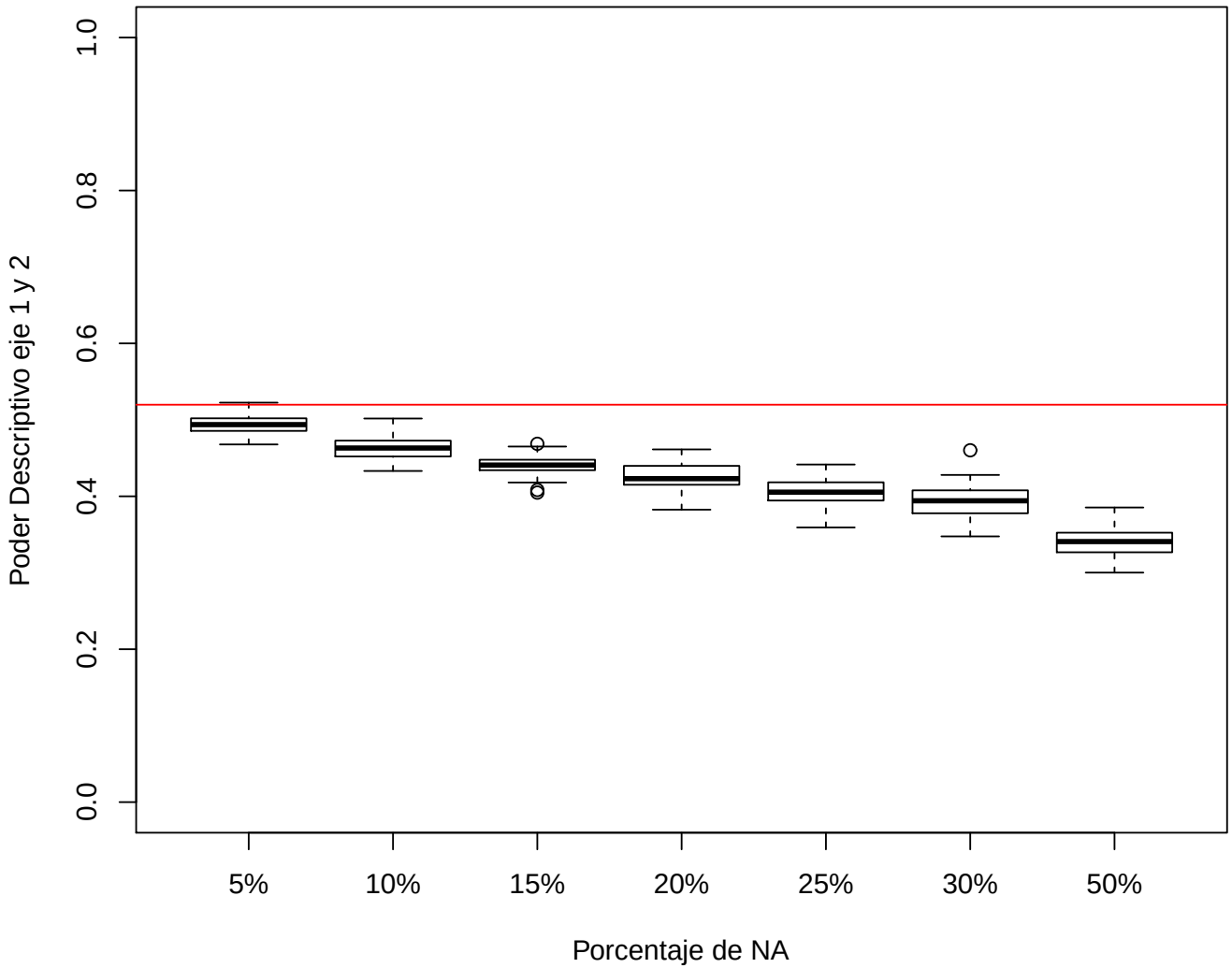
Tabla: Correlación entre coordenadas en R^p con datos completos y faltantes (ACMpdd)

Correlación entre componentes ψ en ACM-EM

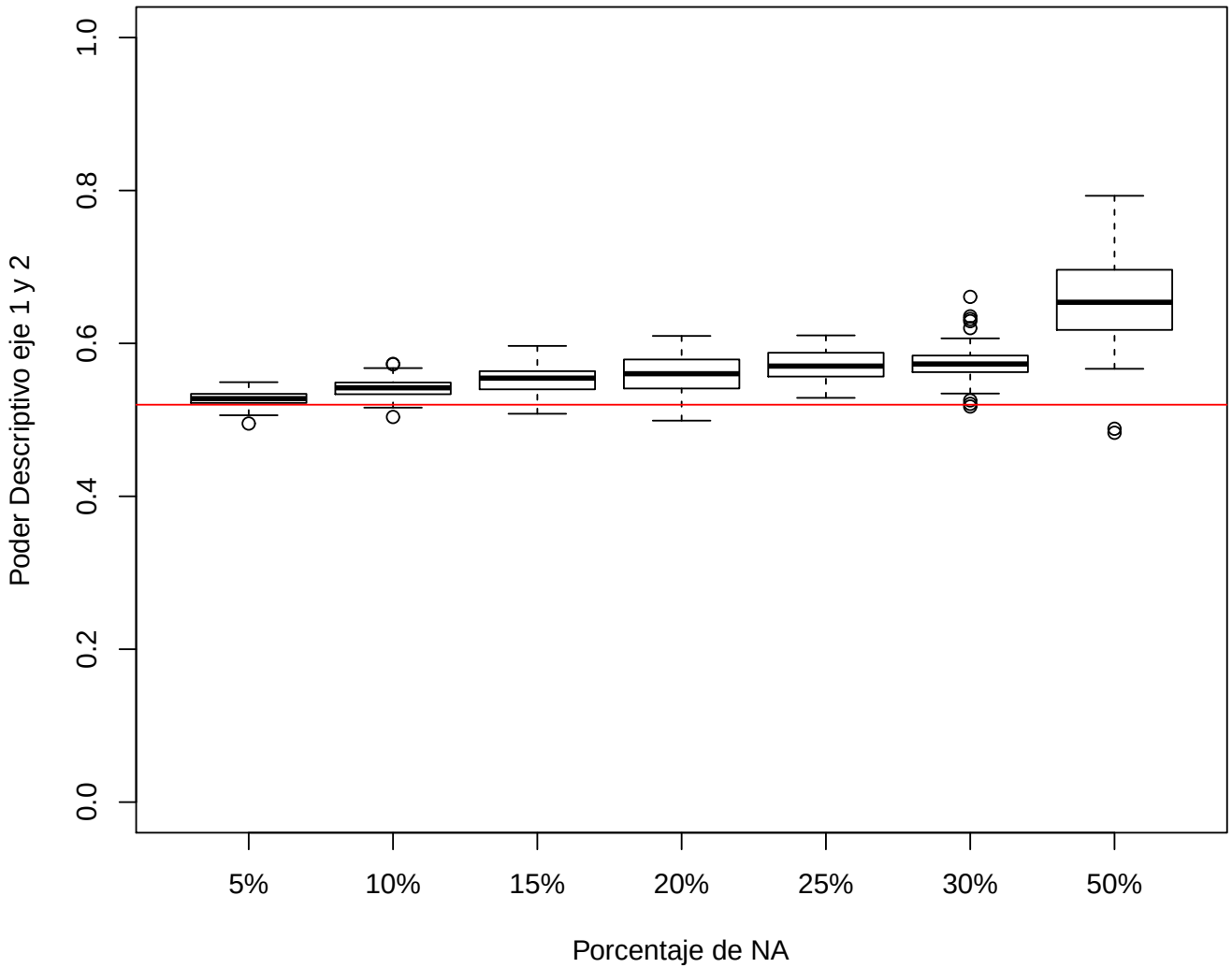
	ψ_1 Completo		ψ_2 Completo		ψ_3 Completo
ψ_1 1 NA	0.9805	ψ_1 1 NA	0.9263	ψ_1 1 NA	0.8559
ψ_1 2 NA	0.9027	ψ_1 2 NA	0.8627	ψ_1 2 NA	0.6258
ψ_1 3 NA	0.8295	ψ_1 3 NA	0.6375	ψ_1 3 NA	0.4717
ψ_1 0:1 NA	0.9568	ψ_1 0:1 NA	0.9232	ψ_1 0:1 NA	0.8422
ψ_1 0:2 NA	0.9737	ψ_1 0:2 NA	0.9398	ψ_1 0:2 NA	0.7869
ψ_1 0:3 NA	0.9388	ψ_1 0:3 NA	0.8349	ψ_1 0:3 NA	0.7856

Tabla: Correlación entre coordenadas en R^p con datos completos y faltantes (ACM-EM)

ACMpdd Simulación (Poder Descriptivo)



ACM-EM Simulación (Poder Descriptivo)



Conclusiones y Perpesctiva

Conclusiones

- Fue posible encontrar la implementación del algoritmo NIPALS en el caso de ACM con datos faltantes, situación que aclará mejor el panorama de cómo trabajar los datos faltantes en ACM, puesto que no se encontraron artículos formales donde se realizará dicha implementación.
- Los resultados de inercia total, modalidad y por pregunta cambiaron con respecto al caso completo, donde fue importante encontrar unas expresiones para los cálculos de inercia, los cuales involucraron la cantidad de datos disponibles en cada escenario de simulación. Respecto al número de ejes en ACM, se sabe que en caso completos es $p - s$ y con este procedimiento resultó $p - 1$ caso similar a ACS.

Conclusiones

- Las componentes del método ACMpdd conservan su ortogonalidad, los vectores propios son ortonormales, los valores propios son descrecientes $(\lambda_1 > \lambda_2 > ... > \lambda_{p-1})$. También, es importante mencionar que los valores propios en el espacio R^p y R^n coinciden, lo cuál indica que las relaciones de transición son legitimas $(\varphi = \frac{1}{\sqrt{\lambda}} M_p F' \psi)$
- Se considera que el método propuesto en este trabajo de investigación ACMpdd, presenta una solución práctica, puesto que su programación es sencilla y se encontrarón propiedades interesantes como: la ortogonalidad en las componentes, ortonormalidad en los vectores propios, equivalencia en los valores propios en R^p y en R^n, entre otras

Conclusiones

- También es importante mencionar que el método ACM-EM tuvó mejores aproximaciones, al realizar la comparación de las componentes con datos completos y datos faltantes ($cor(\psi_1, \psi_1 na)$). En esta comparación se usó el coeficiente correlación y se encontró correlaciones más altas. Algo a tener presente es que la correlación de las componentes con datos completos y faltantes fue mayor en las primeras dos componentes ψ_1 y ψ_2, de hecho la correlación fue mayor en matrices con menor porcentaje de datos faltantes.
- Para matrices con porcentajes de datos faltantes mayores al 30 % se recomienda el ACM iterativo regularizado [Josse et al., 2012]

Trabajos Futuros

- Para trabajos futuros sería interesante comparar el método ACMpdd, con el método de imputación múltiple [Audigier et al., 2015], teniendo en cuenta en los escenarios de simulación mécanismos de datos faltantes NMAR.
- Algunos métodos interesantes a desarrollar son el Análisis de Conglomerados y el Análisis Factorial Múltiple, ambos con datos faltantes vía NIPALS y adaptarlos a las librerías actuales del software R
- También sería interesante estudiar el problema de variables e individuos suplementarios cuando estos vienen con datos faltantes

Bibliografía

Bibliografía I


Aluja, T. and González, V. M. (2014).
Gnm-nipals: general nonmetric-nonlinear estimation by iterative partial least squares.
Revista de Matemática Teoría y Aplicaciones, 21(1):85–106.


Aluja, T. and Morineau (1999).
Aprender de los datos: el análisis de los componentes principales: una aproximación desde el data mining.
Barcelona :EUB.


Audigier, V., Husson, F., and Josse, J. (2015).
Mimca: Multiple imputation for categorical variables with multiple correspondence analysis.
arXiv preprint arXiv:1505.08116.


Benzécri, J.-P. et al. (1973).
L'analyse des données, volume 2.
Dunod Paris.

Bibliografía II


Burt, C. (1950).
The factorial analysis of qualitative data.
British Journal of Statistical Psychology, 3(3):166–185.


Escofier, B. (1981).
Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte.
PhD thesis, INRIA.


Guttman, L. (1941).
The quantification of a class of attributes: A theory and method of scale construction.
The prediction of personal adjustment, pages 319–348.


Hayashi, C. (1956).
Theory and examples of quantification.(ii).
In *Proc. of the Institute of Statist. Math*, volume 4, pages 19–30.

Bibliografía III


Husson, F., Josse, J., Husson, M. F., and FactoMineR, D. (2013).
Package "missmda".
methods, 153(2):79–99.


Josse, J., Chavent, M., Liquet, B., and Husson, F. (2012).
Handling missing values with regularized iterative multiple correspondence analysis.
Journal of classification, 29(1):91–116.


Josse, J. and Husson, F. (2012).
Handling missing values in exploratory multivariate data analysis methods.
Journal de la Société Française de Statistique, 153(2):79–99.


Lebart, L., Fénelon, J.-P., and Dunod, A. (1975).
Statistique et informatique appliquées.

Bibliografía IV


Meulman, J. (1982).
Homogeneity analysis of incomplete data, volume 1.
DSWO Press.


Nishisato, S. (1980).
Analysis of categorical data: Dual scaling and its applications.
Number 24. Univ of Toronto Pr.


Pardo, C. E. and Cabarcas, G. (2001).
Métodos estadísticos multivariados en investigación social.


Pardo, C. E. and Del Campo, P. C. (2007).
Combinación de métodos factoriales y de análisis de conglomerados en r: el paquete factoclass.
Revista colombiana de estadística, 30:231–245.

Bibliografía V


Rubin, D. B. (1976).
Inference and missing data.
Biometrika, 63(3):581–592.


Russolillo, G. (2009).
Partial Least Squares methods for non-metric data.
PhD thesis, Università degli Studi di Napoli Federico II.


Sanchez, G. (2013).
PLS path modeling with R.


Tenenhaus, M. (1998).
La régression PLS, théorie et pratique.
Editions technip.

Bibliografía VI


Trinchera, L., Squillacciotti, S., and Esposito Vinzi, V. (2006).
Pls typological path modeling: a model-based approach to classification.
Proceedings of KNEMO, page 87.


Van der Heijden, P. and Escofier, B. (2003).
Multiple correspondence analysis with missing data.
Analyse des correspondances. Recherches au cżur de l'analyse des donnees, pages 152–170.


Vitelleschi, M. and Quaglino, B. (2009).
Modelos pca a partir de conjuntos de datos con información faltante.
Master's thesis.

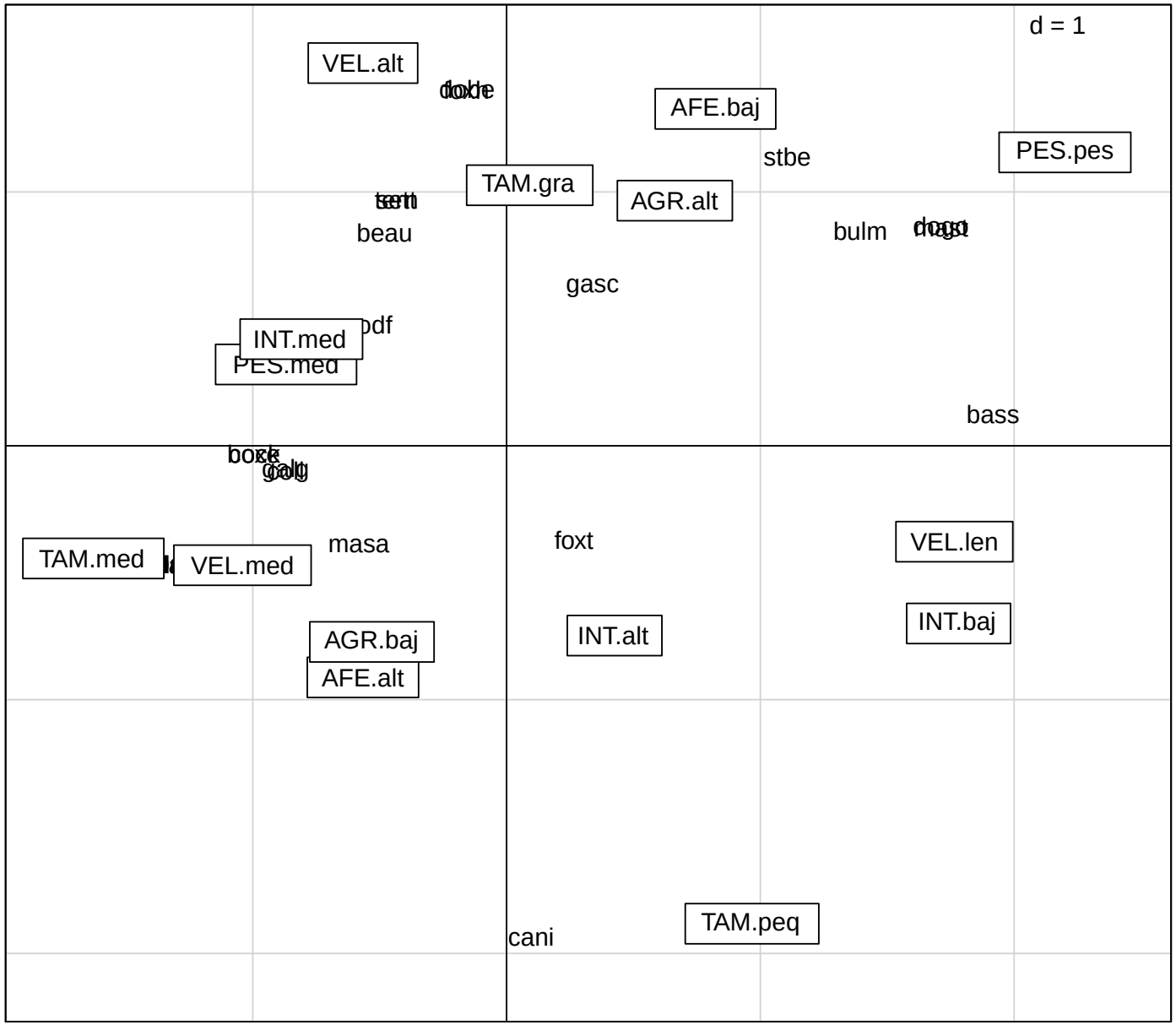

Wold, H. (1975).
Path models with latent variables: The NIPALS approach.
Acad. Press.

Bibliografía VII

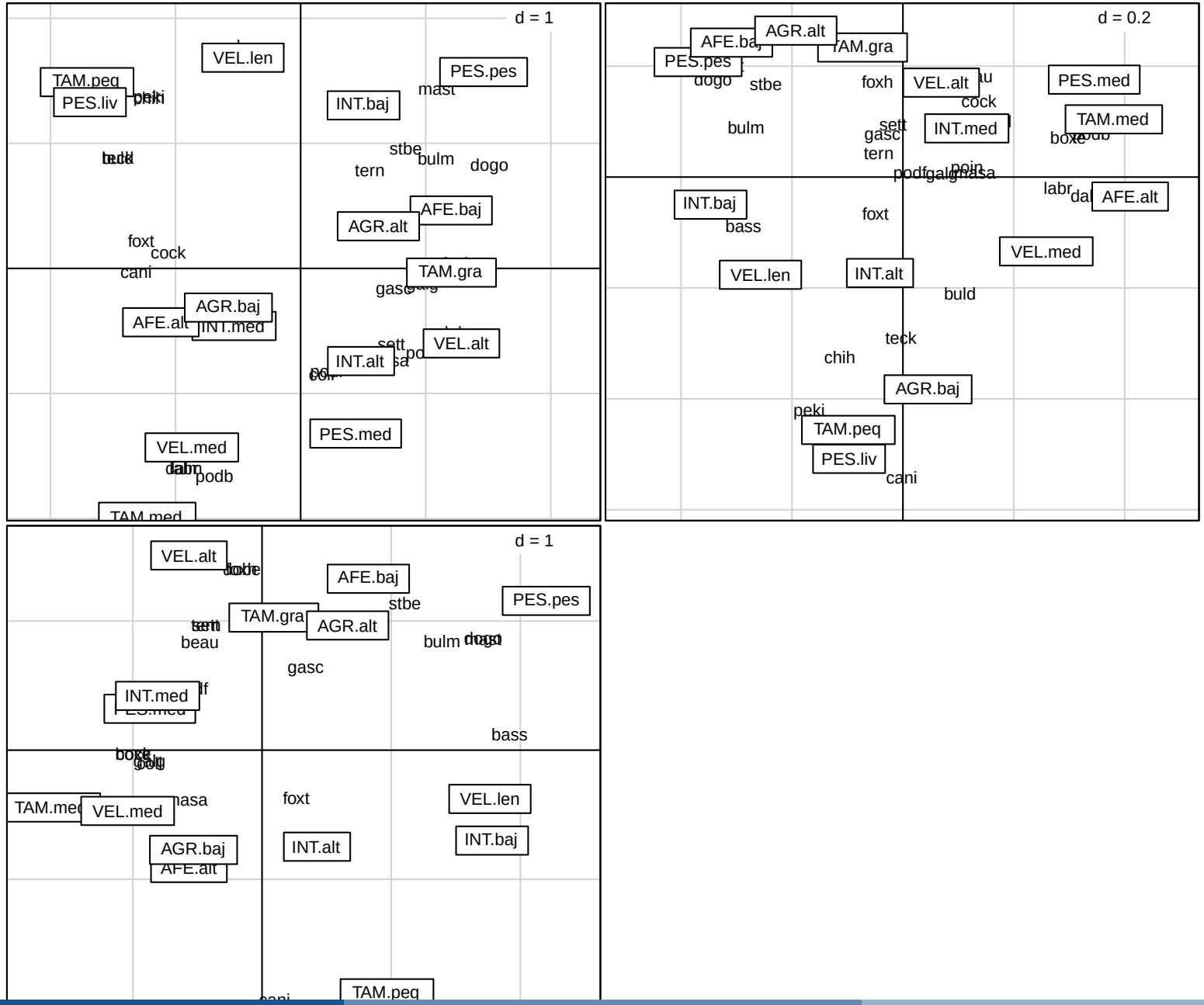

Wold, S., Martens, H., and Wold, H. (1983).
The multivariate calibration problem in chemistry solved by the pls method.
In *Matrix pencils*, pages 286–293. Springer.

¡¡GRACIAS!!

Planos Factoriales (Imputada)

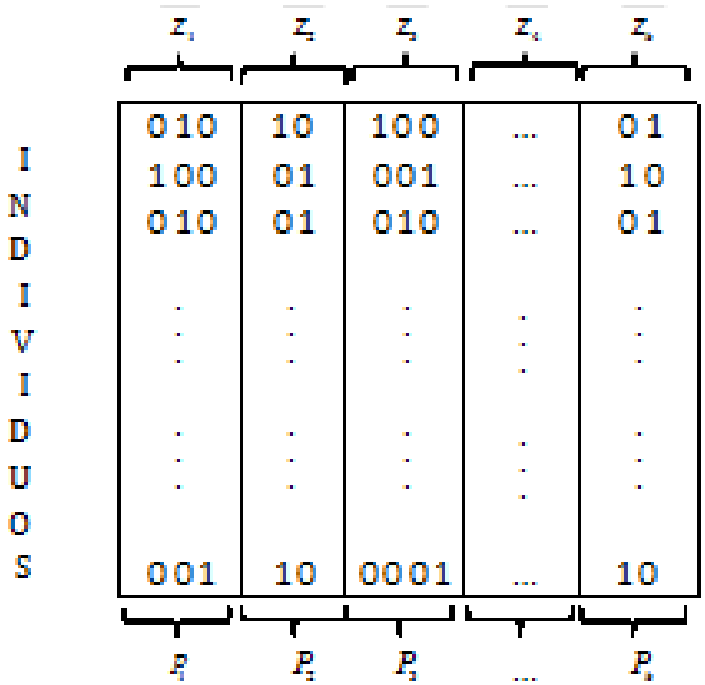


Comparación



Análisis de Correspondencia Múltiple (ACM)

- Se utiliza en el análisis de tablas de individuos descritos por variables categóricas y para estudiar las asociaciones entre diferentes modalidades de las variables en estudio. Este análisis parte de una tabla disyuntiva completa **Z** de n individuos y k variables categóricas [Pardo and Cabarcas, 2001].



Análisis de Correspondencia Múltiple (ACM)

- El ACM es considerado una representación gráfica de la asociación entre variables categóricas.
- Los individuos que aparecen cerca se parecen porque asumen las mismas modalidades de diferentes variables.
- Las modalidades de variables diferentes se consideran asociadas porque son asumidas por los mismos individuos.

Distancia ji-cuadrado

Las respectivas distancias ji-cuadrado, entre las modalidades y entre individuos, aplicada a una tabla disyuntiva completa están dadas por

$$z_{im} =$$

$$\begin{cases} d^2(j,j') = \Sigma_{i=1}^{n} n(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}})^2 \text{ distancia entre las modalidades } j \text{ y } j' \text{ (en } R^n \\ d^2(i,i') = \frac{i}{k}\Sigma_{j=1}^{p} \frac{n}{z_{.j}}(z_{ij} - z_{ij'})^2 \text{ distancia entre los individuos } i \text{ y } i' \text{ (en } R^p \end{cases}$$