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

Andrés Felipe Ochoa Muñoz

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Santiago de Cali, Colombia
2018

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

Andrés Felipe Ochoa Muñoz

Trabajo de investigación presentado como requisito para optar al título de:
Magíster en Estadística

Director:
Ph.D. Victor Manuel Gonzalez Rojas

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Santiago de Cali, Colombia
2018

Dedicatoria

Con cariño a todos mis familiares y amigos, en especial a mis padres Xiomara y Fabián; mi abuelita Nohora; a mis hermanos Kevin y Laura Sofia. A ellos muchas gracias por siempre acompañarme, por su amor y por creer en mis ideas.

También dedico este trabajo a mi compañera en el amor Yurany, gracias por el apoyo, por ser mi inspiración y traerme muchas bendiciones en la vida.

.

Agradecimientos

Agradezco a Dios quién me da salud y vida para seguir adelante con mis sueños, proyectos e ideas. Gracias a mi familia que siempre me da lo mejor y me apoyan en todos las cosas que emprendo.

Gracias al Profesor Víctor González por las enseñanzas, por el tiempo que me dedicó en las asesorías de la tesis, por ser un gran maestro y ser exigente con mi trabajo. Gracias a la Escuela de Estadística y el programa de Maestría Estadística de la Universidad del Valle, por brindarme la oportunidad de trabajar en este proyecto y de seguir aprendiendo de la Investigación y la Docencia.

Agradecimientos a los profesores Roberto Behar y Campo Elias Pardo por la lectura, evaluación y las sugerencias que se realizaron para la mejora de este documento.

También le agradezco mucho a el Profesor Carlos Osorio quién me motivo a seguir con mis estudios cuando me gradué de Estadístico, es un profe que considero un gran amigo y una gran persona. Igualmente agradezco a la profesores María Teresa Varela y Rafael Tovar por el apoyo y darme la oportunidad de trabajar con ellos en el programa de joven investigador del Grupo Salud y Calidad de Vida de la Pontificia Universidad Javeriana, ya que un fue trabajo donde aprendí mucho y fue de gran ayuda para mi Maestría.

Le agradezco a mis compañeros de la Maestría, les deseo lo mejor en la vida, en especial a quienes compartimos en el espacio de estudio: Cristian Garcia, Alejandro Delgado, Jennifer Portilla, Gustavo Gomez, Jose Luis Cabrera y todos los amigos que nos hacian compañia en este proceso de la Maestría; Alex Garcia, Jefferson Amado Peña, Orlando Joaqui, Lorena Ibarra, Jeison Mesa y los compas nuevos que fueron llegando (la cohorte 2016 y 2017).

Agradecimientos a mi primo Mayber Ramos Osorio quién me ayudo a escribir en inglés parte del documento para presentarlo en congresos de Estadística. Primo gracias por las clases y muchas gracias por la amistad.

Gracias nuevamente a mi compañera en el amor, por llenarme de inspiración para seguir teniendo nuevas ideas, para seguir trabajando y haciendo las cosas con mucho amor y dedicación, gracias por la buena vibra y por todos esos buenos momentos que seguimos creando. Je t´aime Yurany.

Resumen

El Análisis de Correspondencias Múltiples (ACM) en presencia de datos faltantes usualmente se trabaja eliminando los registros en donde exista el dato faltante, 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. En algunos otros casos, se asume que el dato faltante es una categoría de la variable cualitativa, trayendo como consecuencia mayor dispersión de varianza en los nuevos ejes. 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. Este trabajo se centra en realizar el método ACM en presencia de datos faltantes sin acudir a técnicas de imputación, para esto se utiliza el principio de datos disponibles del algoritmo NIPALS (Wold et al., 1966).

En el caso de ACM los autores Josse et al. (2012) y Audigier et al. (2015) han trabajado con el enfoque de imputación de datos y no se conocen trabajos o ideas que intenten trabajar ACM bajo el principio de datos disponibles; usando NIPALS. Por esta razón, este trabajo de investigación propone una forma de trabajar los datos faltantes con ACM usando el principio de datos disponibles. De tal manera, que se conforman las matrices simétricas en R^p y en R^n, obteniendo los valores y vectores propios; a su vez garantizando las relaciones de transición y las propiedades de ortogonalidad en los ejes factoriales.

En una primera instancia se analizó los resultados del ACM en una base de datos completa, luego se generarón aleatoriamente 1000 matrices con el 5 %, 10 %, 15 % y hasta un 50 % de datos faltantes. En cada una de las matrices se evaluó el ACM bajo el principio de datos disponibles (ACMpdd) y un método de ACM con el enfoque imputación. Se compararon los planos factoriales, la inercia total y el poder descriptivo con datos completos y faltantes. Se encontró que en ACMpdd a medida que aunmenta el porcentaje de datos faltantes el poder descriptivo disminuye. Por otro lado, con el método de imputación, se encontró que a medida que aunmenta el porcentaje de datos faltantes el poder descriptivo aumenta, situación que no es cohorente, por que se espera que a mayor porcentaje de datos faltantes se explique menos variabilidad en los ejes.
Palabras clave: ***Análisis de Correspondencias Múltiples; Datos Faltantes;Principio de datos disponibles; NIPALS*****)**

Abstract

Multiple Correspondence Analysis (MCA) in the presence of missing data is usually done by eliminating the records where missing data exists. Sometimes, the entire row or column is eliminated from the data matrix, and this is not appropriate because doing so, relevant information about some individual or variable of the study gets lost. In some other cases, it is assumed that the missing data is a category of the qualitative variable, resulting in a greater dispersion of variance in the new axes

A solution to this situation can be the imputation of the missing data, or using an algorithm that allows working with the presence of this type of data. This work focuses on performing the MCA method in the presence of missing data without using imputation techniques. For this, the principle of available data of the NIPALS algorithm is used (Wold et al., 1966).

In the case of MCA, the authors Josse et al. (2012) and Audigier et al. (2015) have worked with the data imputation approach, and there are not known works or ideas that attempt to work MCA under the available data principle; Using NIPALS. For this reason, this research explains in a better way how the missing data can be worked with MCA.Thus, the symmetric matrices are formed in R^p and in R^n, obtaining the eigenvalues and eigenvectors; in turn guaranteeing the transition relations and orthogonality properties in the factorial axes.

In the first instance, the results of the MCA were analyzed in a complete database, then 1000 matrices were randomly generated with 5 %, 10 %, 15 % and up to 50 % of missing data. In each of the matrices the MCA was evaluated under the principle of available data (MCAadp) and an MCA method with the imputation approach. The factorial planes, the total inertia and the descriptive power were compared with complete and missing data. It was found that in MCAadp, as the percentage of missing data continues to increase, the descriptive power decreases. On the other hand, with the imputation method, it was found that as the percentage of missing data increases, the descriptive power increases, a situation that is not cohort, because it is expected that a greater percentage of missing data will explain less variability in the axes

Keywords: ***Multiple Correspondence Analysis; Missing-Data;Principle of available data; NIPALS***

Contenido

Lista de Figuras

Lista de Tablas

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.

En este trabajo se plantea un método multivariado para una matriz de datos con variables cualitativas. El método a implementar en esta propuesta de investigación es el método de Análisis de Correspondencias Múltiples para el caso de datos faltantes o datos no disponibles (NA), evaluando como influyen este tipo de datos, en los ejes factoriales, en el poder descriptivo (porcentaje de varianza explicada), en la inercia que se genera en cada componente, entre otros. Para ello se tendrá en cuenta el principio de datos disponibles expuesto en el algoritmo NIPALS (Nonlinear estimation by Iterative Partial Least Square) y propuesto por Wold et al. (1966).

La idea es encontrar las matrices simétricas en R^p y en R^n usando el principio de datos disponibles del algoritmo NIPALS, para luego en cada matriz realizar una descomposición espectral, con lo cual se obtienen los valores y vectores propios en ambos espacios. Además es necesario verificar las relaciones de transición, propiedades de ortogonalidad en los ejes, ortonormalidad en los vectores propios, entre otros. Es importante mencionar que el método propuesto no realiza imputación de datos, ni se descartan individuos o variables con datos faltantes, fundamentalmente el método realiza los productos escalares con los datos emparejados disponibles. De esta manera, se encontró el método Análisis de Correspondencias Múltiples bajo el principio de datos disponibles (ACMpdd).

Este método se enmarca dentro de los métodos que utilizan el algoritmo NIPALS

(NM-NIPALS, GNM-NIPALS, etc). Existen otros métodos en la literatura, pero no se encontró un artículo formal donde se explicará como implementar NIPALS en el caso de ACM con datos faltantes. En el capitulo 4 se muestran las expresiones matemáticas para poder realizar el ACMpdd usando el principio de datos disponibles.

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).

A continuación, se muestra el planteamiento del problema, donde se analiza detalladamente aspectos teóricos que se tendrán presentes en esta investigación.

1. Planteamiento de la investigación

1.1. Planteamiento del problema

El análisis de datos describe sistemas de información a través de la matriz X, que en este caso está compuesta por variables cualitativas, donde en muchas ocasiones se tiene gran cantidad de variables y gran cantidad observaciones para analizar, con lo cual se necesitan técnicas para resumir dicha información encontrando lo más relevante de la base de datos para luego establecer indicadores o caracterizar la población. (Aluja and Morineau, 1999)

Entre las técnicas de análisis de datos se encuentran el Análisis de Componentes Principales (ACP) (Hotelling, 1933) para datos cuantitativos, el Análisis de Correspondencias Simples (ACS) para tablas de contingencia y el Análisis de Correspondencias Múltiples (ACM) (Lebart et al., 1997) para variables cualitativas.

Uno de los objetivos de estos análisis es extraer q variables latentes no correlacionadas, las cuales son combinación lineal de las p variables originales. Se desea encontrar las q variables, realizando una descomposición en vectores y valores propios de la matriz $W = X'X$ la cual contiene información de las variables originales. Al realizar dicha descomposición se requiere el supuesto de linealidad, ya que la matriz W por lo regular esta compuesta por las correlaciones entre las variables, por lo cual se necesitan relaciones lineales entre las mismas.

Para analizar solo datos cualitativos se utiliza el ACM, el cual parte de una tabla disyuntiva completa Z, la cual contiene s variables indicadoras asociadas a las variables cualitativas originales. En el ACM al igual que en ACS, la matriz Z se transforma en perfiles fila y columna ponderados, los cuales se basan en la distancia chi cuadrado. Al tener éstos perfiles fila y columna, se realiza al igual que en ACP una descomposición en vectores y valores propios, por lo cual el ACM utiliza conceptos del ACP en la matriz transformada para la representación de los individuos y las variables presentes en el análisis activo. (Lebart et al., 1997)

Un problema que se puede presentar en el análisis de datos cualitativos es la presencia de

datos faltantes, el cual se da generalmente en la recolección de la información, bien sea por errores de digitación, por fallos en los sistemas de medición, por destrucción o perdida del registro, por la no vigilancia del proceso, entre otros. Por ejemplo, en la investigación en salud la ausencia de datos puede surgir cuando las personas incluidas en el diseño del estudio no desean participar o no se pueden contactar con ellas mediante el mecanismo de selección establecido, también puede surgir cuando se tienen las preguntas del cuestionario en algunos sujetos, pero para ciertas preguntas no se tiene información del individuo. (Cañizares et al., 2004)

Uno de los métodos 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. 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 bajo el principio de datos disponibles según NIPALS.

El problema de datos faltantes se extiende al caso de ACM en cuyo análisis activo solo hay presencia de variables cualitativas en la forma de tabla disyuntiva completa. Existen autores como Josse et al. (2012) que han trabajado el ACP y ACM utilizando métodos de imputación, donde se utiliza el algoritmo EM propuesto por Rubin and Schenker (1991), estos autores han implementado el ACP y ACM en el paquete ***missMDA*** del software R (Husson et al., 2013). En esta propuesta de investigación se utilizará el principio de datos disponibles del algoritmo NIPALS para tratar los datos faltantes en el ACM. En síntesis, se conoce que existe el procedimiento para aplicar NIPALS en ACP, sin embargo, no existe en la literatura lo propio para ACM, es decir que nadie hasta el momento ha planteado la transformación adecuada adaptar NIPALS con ACM en el escenario de datos faltantes. Teniendo en cuenta todo lo anterior, las preguntas que se formula para esta investigación son:

- ¿Cómo realizar el ACM en presencia de datos faltantes usando el principio de datos disponibles?
- ¿Que tanto influyen los datos faltantes en los resultados de inercia, poder descriptivo, propiedades de ortogonalidad, entre otros?

- ¿Que ventajas o desvantajas existen entre un ACM con el principio de datos disponibles y un ACM bajo el enfoque de imputación?

1.2. Justificación

El Análisis de Correspondecias Múltiples en presencia de datos faltantes usualmente se trabaja eliminando registros donde exista el dato faltante, 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. 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 (Josse et al., 2012; Wold et al., 1966). Este trabajo se centra en realizar el método ACM en presencia de datos faltantes sin acudir a técnicas de imputación, para esto se utiliza el prinicipio de datos disponibles del algoritmo NIPALS.

El análisis multivariado de datos incluye matrices donde solo se tienen variables cualitativas, un método apropiado para describir este tipo de matrices es el ACM, sin embargo, algunas veces se tendrá el problema de datos faltantes, por lo que se deben buscar alternativas para desarrollar el método sin necesidad de eliminar filas o columnas de la matriz de datos. El hecho de eliminar filas o columnas de la matriz hace que se pierda información relevante de la matriz. Por tanto, sería un gran aporte implementar el ACM en presencia de datos faltantes bajo el principio de datos disponibles (ACMpdd); principio que se utiliza en el algoritmo NIPALS para trabajar en esta situación, igualmente se evaluará como influyen los datos faltantes en el poder descriptivo, la conformación de los ejes factoriales, las propiedades de ortogonalidad, entre otros.

Actualmente, existen varios autores trabajando el algoritmo NIPALS en análisis multivariado (Russolillo, 2009; Aluja and González, 2014; Trinchera et al., 2006; 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 intentan hacer la comparación para el caso de ACP (Vitelleschi and Quaglino, 2009).

En el caso de Análisis de Correspondencias Múltiples los autores Josse et al. (2012) han trabajado con el enfoque del algoritmo EM trayendo consigo dificultades en el proceso de imputación de la tabla disjuntiva completa; asigna 1 a las categorías de mayor frecuencia y también presenta problema de convergencia, si no se seleccionan las semillas adecuadamente.

Por otro lado, no se conoce de trabajos o ideas que intenten trabajar ACM bajo el principo de datos de NIPALS. Por esta razón esta propuesta investigación generará conocimientos sobre cómo se puede trabajar los datos faltantes con el ACM, lo cual sería un gran aporte para los métodos multivariados, además de una alternativa diferente a los métodos de imputación.

1.3. Objetivos

1.3.1. Objetivo General

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

1.3.2. Objetivo Específicos

- 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)

2. Antecedentes

Son números os los estudios realizados en el contexto de análisis multivariado para datos cualitativos, ya sea con el objetivo de medir variables latentes, identificar conglomerados en el conjunto de datos, encontrar las variables relevantes del estudio, entre otras. Particularmente en el análisis de datos se presenta el problema de datos faltantes, el cual ha sido abordado a través del algoritmo NIPALS y ha sido perfeccionado por Wold et al. (1983), éstos autores destacan el algoritmo por las características que presentaba para trabajar con datos multivariantes, como lo eran: la dimensionalidad, multicolinealidad y la resistencia hasta un 15 % de datos faltantes.

También se ha trabajado en diferentes áreas de Estadística el Algoritmo EM, para realizar el proceso de imputación múltiple (Rubin, 2004), teniendo en cuenta diferentes mécanismos generadores de datos faltantes (MAR,MCAR,MNAR) y éste algoritmo se ha trabajado en Análisis de Componentes Principales en presencia de datos faltantes (ACP iterativo ó ACP-EM) (Josse and Husson, 2012).

En el contexto de Análisis de Correspondencias con datos faltantes, también se econtrarón 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). 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).

Después de una revisión bibliográfica referente a estudios realizados en análisis multivariados sobre el algoritmo EM y NIPALS, se citan algunos trabajos relacionados con estas metodologías. En la primera sección de este capítulo se describen trabajos que hacen referencia a estudios relacionados con análisis multivariado, donde utilizan la teoría de imputación de datos faltantes (Algoritmo EM). En la segunda sección se tienen estudios donde utilizan el principio de datos disponibles (Algoritmo NIPALS).

2.1. Análisis multivariado utilizando métodos de imputación

- **(Vitelleschi et al., 2010):** Abordan la problemática de la construcción de modelos ACP a partir de conjuntos de datos con información faltante. Se trabaja sobre tres situaciones diferentes con relación a la matriz de datos originales. En cada situación se generaron pérdidas a través de mecanismos aleatorios y no aleatorios . A partir de cada conjunto de datos se construye el modelo ACP utilizando: Casos Completos, *Nonlinear Iterative Partial Least Squares* (NIPALS) y *Expectation Maximization* (EM). Se comparan los resultados con los obtenidos a través del conjunto de datos originales. Se definen una serie de medidas para estudiar cómo se afectan los resultados según la dimensión de la matriz de datos, el porcentaje y el mecanismo de pérdida, con relación a: bondad del ajuste, bondad de predicción, vectores cargas, ortonormalidad de la matriz de cargas y ortogonalidad de la matriz de scores. Una conclusión de este trabajo es que a partir de las situaciones estudiadas se observó que EM, muestra sus ventajas en los casos clásicos de información multivariada con pocas variables medidas sobre muchos individuos. Mientras que un método como NIPALS, es más adecuado en los casos extremos de información multivariada con pocos individuos y muchas variables. En ningún caso se muestra la conveniencia de desechar individuos por no contar con información completa para ellos.

- **(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 Regularizado Iterativo. 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 trabaja con el **paquete missMDA** del software R (Husson et al., 2013). En el marco teórico se explicará el ACP Iterativo.

- **(Josse et al., 2012):** Desarrollan un enfoque común para hacer frente a los datos

faltantes en el análisis multivariado, el cual consiste en minimizar la función de pérdida entre la matriz original y la matriz reconstuida con la imputación. Esto se puede lograr por medio de algoritmos tipo EM donde se realiza una imputación iterativa de los valores que faltan durante la estimación de los ejes y componentes. 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 (Husson et al., 2013). 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.

2.2. Análisis multivariado utilizando el principio de datos disponibles (NIPALS)

- **(Wold et al., 1966):** Desarrollan el algoritmo NIPALS el cual es la base de la regresión PLS, (Tenenhaus, 1998). Fundamentalmente el algoritmo 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.

- **(Tenenhaus, 1998):** Describe el algoritmo NIPALS, donde se ilustra un ejemplo de la aplicación de éste, utilizando la base de datos *carscomplete* del paquete ***plsdepot*** de R (Sanchez, 2012), la cual contiene las características de 24 modelos de autos teniendo en cuenta las variables cilindraje, potencia, velocidad, peso, largo y ancho. Se muestra como los valores de las componentes principales para el caso de datos completo y el caso de datos faltantes son similares en las componente 1, 2, 3.

- **(Russolillo, 2009):** Presentan tres métodos basados en el algoritmo NIPALS, para analizar las variables observadas en diferentes escalas de medición, ya que el método PLS nació para manejar solamente el conjunto de datos que forman espacios métricos, esto implica que todas las variables incluidas en el análisis se observan en escalas de intervalo o de razón. El método NM-PLS también tiene como objetivo investigar la no

linealidad de las variables observadas. Estos tres métodos: NM-NIPALS (non metric NIPALS), NM-PLS y el algoritmo de sendero NM-PLS, proporcionan el mismo tiempo para los parámetros específicos del modelo de PLS.

- **(Aluja and González, 2014):** Desarrollan el algoritmo GNM-NIPALS que forma 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 cuantificada. Este método es una generalización del algoritmo NM-NIPALS que usa solo la primera componente principal en la cuantificación de variables cualitativas. GNM-NIPALS será extendido al caso de datos faltantes

3. Marco Teórico

En este capitulo se presenta teóricamente el método de Análisis de Componentes Principales el cuál es la base para los métodos: Análisis de Correspondencias Simples y Análisis de Correspondencias Múltiples. También se describe el algoritmo NIPALS el cual se utiliza para trabajar en presencia de datos faltantes usando el principio de datos disponibles. En este capitulo, se hace referencia en cada método a los procesos de optimización, las matrices a diagonalizar, el concepto de inercia, los valores, vectores propios y demás conceptos relevantes del Análisis Multivariado. Además, se hace referencia a la relaciones que existen entre los métodos, especialmente en que los Análisis de Correspodencias Simples y Múltiples se pueden ver como un Análisis de Componentes Principales de una matriz transformada en perfiles ponderados. También se mostrarán algunos conceptos fundamentales de la teoría de datos faltantes y que aspectos se tienen en cuenta en el enfoque de imputación presentado por Rubin (1976). Además se mostrará como realizar un ACP iterativo y un ACM iterativo usando el algoritmo EM (Josse and Husson, 2012). A continuación, se presenta el Análisis de Componentes Principales:

3.1. Análisis de Componentes Principales (ACP)

El ACP es la técnica del análisis multivariado más importante, fue presentado por primera vez por Pearson (1901) e integrado a la estadística matemática por Hotelling (1933). El ACP es una técnica de representación de los datos, con un carácter óptimo según ciertos criterios algebráicos y geométricos, en donde se busca interpretar o comprender las relaciones entre variables e individuos en una matriz de datos $X_{n.p}$. Algunos ejemplos para trabajar con ACP son:

- Construcción de un índice para medir la capacidad economica de un individuo
- Análisis de la relación entre variables e individuos
- Segmentar una población en función de sus preferencias en consumo

Se puede decir que el ACP servira como una etapa intermedia a un análisis posterior, como por ejemplo: un análisis de regresión, clasificación, discriminación, entre otras. El ACP permite reducir la dimensionalidad de los datos, transformando el conjunto de p variables originales en otro conjunto de a variables incorrelacionadas ($a \leq p$) llamadas componentes principales.

$$\psi_1 = u_{11}X_1 + u_{12}X_2 + ... + u_{1p}X_p$$
$$\psi_2 = u_{21}X_1 + u_{22}X_2 + ... + u_{2p}X_p$$

...

$$\psi_a = u_{a1}X_1 + u_{a2}X_2 + ... + u_{ap}X_p$$

Las a componentes principales son obtenidas como combinaciones lineales de las variables originales. Las componentes se ordenan en función del porcentaje de varianza explicada. De tal manera, que el primer componente será el más importante por que explica mayor porcentaje de la variabilidad de los datos (λ_j valores propios $j = 1, .., a$).

El ACP tiene la opción de usar la matriz de correlaciones o la matriz de covarianzas.

- 1 Opción: Matriz de correlación. En esta opción se le esta dando la misma importancia a todas la variables y éstas tienen diferentes escalas de medida.
- 2 Opción: Matriz de covarianzas. Esta opción se utiliza cuando todas las variables tienen la misma escala de medida, y el investigador quiere analizar las variables en función del grado de variabilidad

Los únicos requerimientos previos para la aplicación del ACP son:

- a) Continuidad en las variables.
- b) Supuesto de linealidad en las variables
- c) El número n de individuos debe ser mayor que el número p.
- d) La base de datos es completa, es decir no permite trabajar con datos faltantes.

3.1.1. Maximización y Matriz a diagonalizar

El objetivo geométrico del Análisis Factorial es buscar un nuevo sistema de ejes ortogonales u_α , $\alpha = 1, 2, ..., p$, en 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. La proyección ortogonal del individuo i con respecto al eje u_α, se conoce como la α- ésima componente principal $\psi_\alpha = Xu_\alpha$ (X por lo regular está estandarizada Z para poder trabajar con diferentes escalas de medición), tal como se observa en la Figura **3-1**

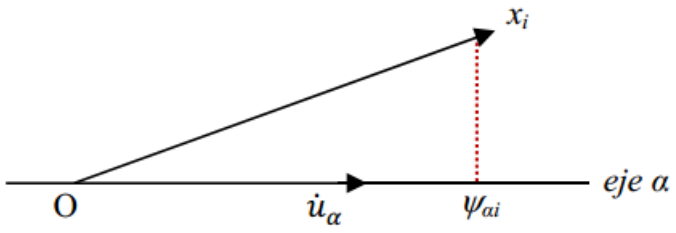

Figura 3-1.: Proyección ortogonal del individuo i sobre u

En ACP es de interés el maximizar la correlación que hay entre los nuevos ejes y las variables originales, es de interés maximizar la inercia o la variabilidad obtenida en cada eje α. La expresión de la inercia es:

$$\psi' M_n^{-1} \psi = u' Z' M_n Z u = \lambda \text{ bajo la restricción } u'u = 1.$$

$Mn = 1/n$ es una matriz diagonal que pondera cada individuo.

Entonces la matriz a diagonalizar es $Z' M_n Z = uDu'$, donde D es una matriz diagonal, que en su diagonal contiene los valores propios, siendo u los vectores propios. Es importante recordar que la matriz Z corresponde a la estandarización de la matriz X.

El anterior sistema de valores y vectores propios surge de maximizar la expresión de la inercia usando el Lagrangiano $L(u)$, esto con el fin de que los valores propios λ sean decrecientes. De esta manera se tiene lo siguiente:

$$L = u' Z' M_n Z u - \lambda(u'u - 1)$$

$$L = u^{'} Z^{'} M_n Z u - \lambda u^{'} u + \lambda$$

$$\frac{\partial L}{\partial u} = 2Z^{'} M_n Z u - 2\lambda u = 2(Z^{'} M_n Z u - \lambda u) = 0$$

$$Z^{'} M_n Z u = \lambda u$$

La anterior expresión es un sistema de valores y vectores propios, el cuál garantiza que las componentes principales $\psi = Zu$ sean ortogonales y sus valores propios λ sean decrecientes. Es importante mencionar que la matriz $Z^{'} M_n Z = cor(X_i, X_j)$ y es una matriz simétrica.

3.1.2. Indicadores para las interpretaciones del ACP

Contribuciones de los individuos Se define la contribución de un individuo en un eje de inercia como la parte de inercia total a lo largo del eje que se debe al individuo

$CTR(i,\alpha) = \frac{\psi_{i,\alpha}^2}{n\lambda_\alpha}$, donde $\sum_{i=1}^{n} CTR(i,\alpha) = 1$

La contribución de un individuo a la inercia de un eje mide la importancia de dicho individuo en el eje.

Contribuciones de las variables La contribución de una variable en un determinado eje de inercia es una medida de la importancia de la variable en dicho eje y, en concreto, determina que parte de la inercia total que existe a lo largo del eje se debe a la variable.

$CTR(j,\alpha) = u_{j,\alpha}^2$, igualmente $\sum_{i=1}^{n} CTR(j,\alpha) = 1$

3.1.3. Reconstitución de la matriz

En ACP es posible reconstituir la matriz de datos originales usando h componetes principales, es posible construir todos los valores de una variable, todo los valores de individuo o toda la matriz, este concepto es muy importante en el algoritmo NIPALS, que será tema en la siguiente sección. Entonces se recontituye la matriz de la siguiente forma:

$Z_{np} = \psi u^{'}$ donde ψ =componentes, u =vectores propios. La reconstitución de la j-ésima variable será entonces:

$$Z_j = \psi_1 u_{1j} + \psi_2 u_{2j} + ... + \psi_a u_{aj}$$

donde a es el rango Z, $(a \leqslant p)$

Este concepto de reconstitución tambien se utiliza para la cuantificación de las variables cualitativas, ya que:

$$u_{\alpha j} = \frac{cor(z_j, \psi_\alpha)}{\sqrt{\lambda_\alpha}}$$

3.1.4. Representación simultánea

En el ACP las dos nubes están definidas en espacio distintos, con origenes distintos y bases distintas. Para la nube de individuos, el origen es el centro de gravedad de los individuos, que es de dimensión p y designamos por u_α la correspondiente base hallada. Para la nube de variables, el origen de coordenadas es de dimensión n y los ejes factoriales los designamos por v_α. Los puntos fila y los puntos columna están pues en espacios distintos, por lo que es imposible poderlos visualizar en un solo espacio que respete las proximidades internas de las dos nubes.

Es posible representar las direcciones definidas por cada variable activa sobre la base de los ejes factoriales u_α. La posibilidad de una representación simultánea reside entonces en la proyección del antiguo eje canónico (en R^p) e_j sobre el nuevo eje u_α, cuya coordenada vale $e_j^{'} u_\alpha = u_\alpha$. Las relaciones entre los objetos (individuos) y las variables pueden apreciarse en gráficos *biplots*.

De las relaciones de transición, el análisis de la nube de variables se deduce del de la nube de individuos. Recordemos que en R^n la coordenada de la variable j sobre el eje α es $\varphi = \sqrt{\lambda} u$. Las dos nubes de variables se diferencian por una dilatación $\sqrt{\lambda}$; definida sobre cada eje. La interpretación de distancia entre dos variables solo se puede hacer en R^n.

Observaciones:

- No es posible hacer aparecer directamente las variables suplementarias en una representación simultanéa de variables e individuos en ACP. Las variables suplementarias no participan en la definición de la base original para la nube de los individuos.
- La representación simultánea en ACP puede aparecer en un sistema de representación más general introducido por Gabriel (1971), que consiste en descomponer la tabla de datos en el producto de dos matrices, una representando los elementos fila y otra los elementos columna *(Biplot)*

3.2. Nonlinear estimation by Iterative Partial Least Square (NIPALS)

(...)
Antes de empezar este capitulo hay que tener en cuenta esta notación de equivalencia con ACP: $\psi_\alpha = t_h$ (Componente α); $u_\alpha = p_h$ (Vector propio α); $Z = X$ (La matriz de partida).
(...)

El algritmo NIPALS fue propuesto por Wold et al. (1966) y es la base de la regresión PLS, (Tenenhaus, 1998). Se tiene la matriz de datos $X_{n,p}$ de rango a cuyas columnas X_1, ... , X_p se suponen centradas o estandarizadas. Se utiliza la descomposición derivada del ACP que permite realizar la reconstitución mediante lo siguiente:

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

t_h es la h-ésima componente principal
P_h es el vector propio en el eje h.

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

$$X_j = \sum_h^a t_h phj, j = 1, ..., p$$

$$X_i = \sum_h^a t_h phi, i = 1, ..., n$$

En la Figura **3-2** se observa el esquema de cómo funciona el Algoritmo Nipals para encontrar los cálculos correspondientes a t_h y p_h. 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 $P_1, ..., P_h$

Posteriormente se muestra el pseudocodigo del algoritmo cuando la matriz de datos esta completa. 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

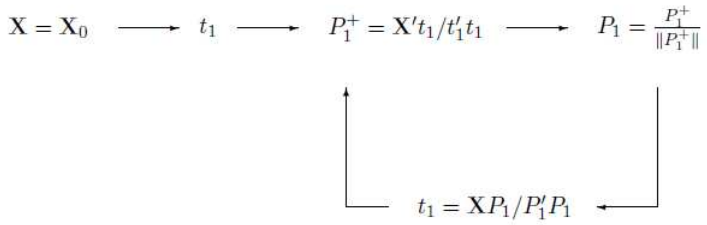

Figura 3-2.: Esquema del Algorimo NIPALS

3.2.1. Pseudocódigo 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

La principal característica de NIPALS es que trabaja respecto a una serie de productos escalares como suma de productos de los elementos emparejados. Esto permite trabajar con datos faltantes, realizando en cada operación los datos disponibles. Geométricamente el procedimiento toma los elementos omitidos como si ellos cayeran sobre la recta de regresión; no son puntos de apalancamiento (Tenenhaus, 1998)

En el pseudocódigo del algoritmo NIPALS con datos faltantes, se tiene las etapas 2.2.1 y 2.2.3 donde 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} (Aluja and González, 2014)

3.2.2. Principio de datos disponibles

Este principio hace referencia a que uno puede realizar algunas operaciones entre vectores omitiendo los datos no disponibles (NA) y trabajando con los puntos emparajedos disponibles. Es decir que si se tiene un vector X y un vector Y (ambos con NA), 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 $\langle x, y \rangle = \sum_{xyexistent} x_i y_i = x_3 y_3 + x_4 y_4 + \ldots + x_n y_n$

3.2.3. Pseudocódigo 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: $X_h = X_{h-1} - t_h p_h^{'}$

3.3. Análisis de Correspondencias Simples (ACS)

Presentado y desarrollado por Benzécri et al. (1973), tiene sus principios teóricos sobre las tablas de contingencia dentro del contexto de estadística inferencial clásica. La tabla de contingencia se obtiene mediante la clasificación de elementos de estudio según las coocurrencias asumidas entre modalidades fila y columna de las variables nominales.

Este análisis parte de una matriz K_{n*p}, dicha matriz es una tabla de contigencia de 2 variables cualitativas, en esta matriz se pretende obtener una tipología de filas, una tipología de columnas y relacionar las tipologías entre sí. De esta forma en la Figura **3-3** se observa la matriz k, donde k_{ij} = es el número de elementos presentando la característica i,j simultáneamente; $(k_{ij} \geqslant 0)$, El número total de elementos de la matriz es $k = \sum_i k_{i.} = \sum_j k_{.j} = \sum_{ij} = k_{ij}$. A su vez se puede obtener la matriz de frecuencias relativas F, la cual contiene los elementos: $f_{ij} = k_{ij}/k$; $f_{i.} = \sum_j f_{ij} = k_{i.}/k$; $f.j = \sum_i f_{ij} = k_{.j}/k$; $\sum_{i,j} = \sum_i f_{i.} = \sum_j f_{.j} = 1$

El ACS se basa en la prueba de independencia χ^2 de Karl Pearson, ya que es de interés analizar las relaciones existentes entre las modalidades de las variables. Se dice que hay independencia entre las variables i j, si para todo i,j se tiene que $f_{ij} = f_{i.} f_{.j}$, la prueba χ^2 permite apreciar la desviación entre f_{ij} y $f_{i.} f_{.j}$. Esta prueba tambien se expresa en términos de los perfiles fila, $\frac{f_{ij}}{f_{i.}} = f_{.j}$. Si todos los perfiles fila son idénticos al perfil fila esperado hay independencia entre variables. Igual sucede en los perfiles columna, lo cual indica que es lo mismo hacer el análisis por filas que por columnas.

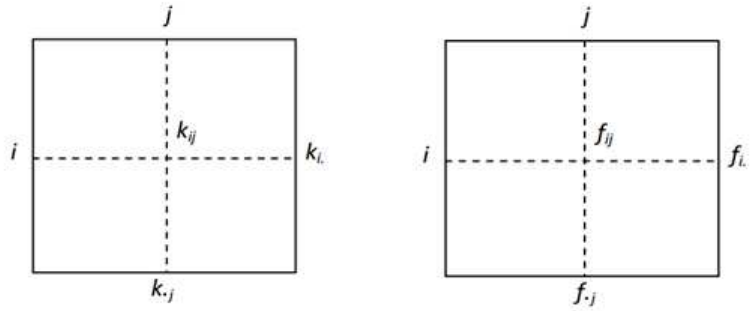

Figura 3-3.: Matriz K y F de partida en ACS

3.3.1. Construcción de la nube

Ya que $\sum_j f_{ij}/fi. = 1$, $\sum_i f_{ij}/f_{.j} = 1$ los n puntos de la nube fila y los p puntos de la nube columna están situados en subespacios de dimensión $p-1$ y $n-1$ respectivamente. El análisis consiste en buscar un conjunto de ejes ortogonales sobre los que será proyectada la nube, obteniendo una representación gráfica aproximada (Lebart et al., 1997). Es de interés representar geométricamente las similitudes entre las diferentes modalidades de una misma variable, lo que conduce a describir dichas similitudes por medio de distancias, esto en la nube N_I o en la nube N_J. Para construir la distancia entre dos modalidades i e i', se contruye la distancia χ^2 de la siguiente manera:

$$d^2(i,i') = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'}} \right)^2$$

En el ACP que $x_{ij} = f_{ij}/f_{i.}$ es el perfil fila y $1/f_{.j}$ para todo j que representa los elementos diagonales de la métrica, esto es $M_p = [\ddots 1/f_{.j} \ddots]$ análogamente $M_n = [\ddots 1/f_{i.} \ddots]$

Es posible transformar la distancia χ^2 en distancia euclidiana (perfil ponderado)

$$d^2(i,i') = \sum_j \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.}\sqrt{f_{.j}}} \right)$$

El perfil ponderado es: $\sum_i f_{i.} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} = \sqrt{f_{.j}}$, $\forall j$.

La inercia respecto al centro de gravedad para los perfiles ponderados (o simples), está definida por:

$$I_{G_I} = \sum_{i=1}^{I} p_i d^2(i, G_I) = \sum_{i=1}^{I} f_{i.} \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_i \sqrt{f_{.j}}} - \sqrt{f_{.j}} \right)^2 = \sum_{i,j} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = \frac{\chi^2}{k} = \phi^2$$

Como se observa el ACS es una descomposición de este estadístico y cada factor representa parte de la relación entre las variables. Es importante mencionar que de la misma forma se construye la inercia para los perfiles columna y se obtiene la equivalencia entre ellas $I_{G_I} = I_{G_J}$

3.3.2. Maximización y matriz a diagonalizar

Las coordenadas de los perfiles fila $X = M_n F$ proyectados, están dadas por $\psi = X M_p u$ bajo la métrica M_p y su inercia respecto al origen es $\psi' M_n^{-1} \psi = u' M_p F' M_n F M_p u = u' M_p S u = \lambda$ que es la cantidad a maximizar bajo $u' M_p u = 1$. La matriz no simétrica $S = F' M_n F M_p$ contiene vectores propios no necesariamente ortogonales asociados a los valores propios λ

Si en vez de S se diagonaliza la matriz simétrica $S^* = M_p^{1/2} F' M_n F M_p^{1/2}$ bajo la restricción $w^{*'} w^* = 1$, los vectores propios $w^* = M_p^{1/2} u$ son ortogonales y están asociados a los λ valores propios de S^*. Esto es equivalente a un ACP no centrado ni estandarizado de $Z^* = M_n^{1/2} F M_p^{1/2}$, ya que $S^* = Z^{*'} Z^*$. En este proceso se obtienen los valores y vetores propios en R^p. Un razonamiento análogo en R^n nos entregaría los valores y vectores propios ($T^* = Z^* Z^{*'}$)

3.4. Análisis de Correspondecias Múltiples (ACM)

Los principios de este método se deben a Guttman (1941), pero también a Burt (1950) y a Hayashi (1956). El 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 p variables categóricas (Pardo and Cabarcas, 2001). En la Figura

3-1 se observa una ilustración de la tabla disyuntiva completa, lo que se observa es que se asigna 1 a la presencia de la característica en la variable y 0 la ausencia de la característica, si se piensa en la relación con el ACS, el ACM es entonces un ACS donde se tiene una variable con n categorías y otra variable con p categorías.

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

Tabla 3-1.: Tabla Disjuntiva completa Z_{ij}

- 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.

Ya que es de interés analizar las similitudes entre individos i e $i^{'}$, se construyen las respectivas distancias ji-cuadrado, las cuales tambien se puenden construir entre modalidades, y se realizan para la tabla disyuntiva completa Z, dichas distancias están dadas por:

$$z_{im} =$$

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

3.4.1. Maximización y matriz a diagonalizar

Haciendo la analogía con ACS, se tiene que el objetivo geométrico del ACM es buscar un nuevo sistema de ejes ortogonales u_α, en los que se proyecte la inercia I_α de la nube de individuos, tal que los primeros ejes concentren la mayor parte de la misma y en forma decreciente.

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

$$I_\alpha = u_\alpha^{'} M_p F^{'} M_n M_n^{-1} M_n F M_p u_\alpha$$

$$I_\alpha = u_\alpha^{'} M_p F^{'} M_n F M_p u_\alpha; \qquad donde \;\; S = F^{'} M_n F M_p$$

$$I_\alpha = u_\alpha^{'} M_p S u_\alpha; \qquad u^{'} M_p u = 1$$

Por tanto en el espacio de los individuos R^p, para encontrar los ejes directores u_α se diagonaliza la matriz $S = F^{'} M_n F M_p$ tal que $u^{'} M_p u = 1$. S No necesariamente es simétrica y por ende no se garantiza que los vectores propios sean ortonormales. Si en vez de diagonalizar S se diagonaliza la matriz S^*:

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

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

$$S^* = S_o^{'} S_o; \quad en \quad R_p; \quad w^{'} w = M_p^{1/2} u$$

$$T^* = S_o S_o^{'}; \quad en \quad R_n; \quad r^{'} r = M_n^{1/2} v$$

Entonces en ACM se observa la relación que hay con Análisis de Componentes Principales (ACP), de tal forma que S_o 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 S_o, para así obtener los resultados del ACM. Este proceso es muy importante, debido a que se sigue el mismo esquema para la solución con datos faltantes usando el principio de datos disponibles.

3.4.2. 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:

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

3.4.3. Inercia por modalidad

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

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

3.4.4. 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.

3.5. Teoría de datos faltantes

La presencia de datos faltantes no solamente reducirá la precisión de las estimaciones y disminuirá la potencia de las pruebas estadísticas realizadas debido a la información incompleta, sino que bajo ciertas circunstancias puede generar sesgos y proporcionar conclusiones invalidas. Esta consecuencia esta relacionada con las razones que llevaron a la presencia valores faltantes, Rubin (1976) introduce los mecanismos de datos faltantes que describen la probabilidad de que una observación sea faltante y su relación con las variables del estudio.

Los tres tipos de mécanismos propuestos son:

- faltantes completamente aleatorios (Missing Completely at Random, MCAR)
- faltantes aleatorios (Missing at Random, MAR)
- faltantes no aleatorios (Missing Not at Random, MNAR)

Los mecanismos de datos faltantes tienen el objetivo de caracterizar las razones por las cuales hay observaciones faltantes. Se define R_{ij} que es variable aletoria que indica la presencia de datos faltantes, entonces $R_{ij} = 1$ si el dato x_{ij} es faltante, $R_{ij} = 0$ si el dato x_{ij} es observado. De esta manera, cada dato x_{ij} tiene su dato R_{ij} y se hace una partición de x_{ij}, tal que x^o los datos observados y x^m a los datos faltantes

3.6. Mécanismos de datos faltantes

MCAR Los datos son MCAR cuando la probabilidad de que la respuesta sea faltante no está relacionada con ningún valor de las variables que se planearon observar o de la misma variable respuesta.

$$P(R_i|x^o, x^m, X_i) = P(R_i)$$

Por lo tanto, los datos x_{ij} son MCAR cuando Ri es independiente de los componentes observados y no observados.

MCAR Se dice que los datos son MAR cuando la probabilidad de que una respuesta sea faltante depende de los valores observados de las variables, pero no depende que los valores no observados.

$$P(R_i|x^o, x^m, X_i) = P(R_{ij}|x^o, X_i)$$

Esto significa que la distribución condicionada de R_i dado x_i o es independiente de x^m

NMAR Se dice que los datos son NMAR cuando la probabilidad de que la respuesta sea faltante está relacionada no solamente con los valores observados, sino que también de los valores no observados de la variable respuesta.

$$P(R_i|x^o, x^m, X_i)$$

Lo que indica que la distribución condicionada de R_i dado x^o, depende de al menos un componente de x^m

3.7. Métodos de imputación

Imputación Simple La idea de reemplazar los valores faltantes con el promedio es antigua, que los metodólogos suelen atribuir a Wilks (1932) y Enders (2010). La sustitución de datos utilizando promedios es una vieja práctica entre investigadores de diversas disciplinas, sin tener en cuenta las limitaciones que este tiene. Es uno de los métodos mas sencillos para el caso de las técnicas de imputación simple.

Consiste en sustituir el valor faltante mediante el promedio de las unidades observadas de la misma variable. A pesar de que los promedios son estimadores consistentes de la media poblacional, la matriz de covarianza de la muestra sistemáticamente subestima el tamaño de las varianzas y covarianzas.

Imputación Múltiple La imputación múltiple, propuesta por Rubin (2004), se puede describir en tres pasos. el primer paso, llamado de imputación, consiste en imputar cada valor faltante por M valores imputados (creando M bases de datos) por medio de un modelo de imputación, el cual debe describir la relación distribucional entre los datos no observados y la información disponible ($f(y^m|y^o, x, \theta)$).

Posteriormente, se realiza el análisis de cada una de las M bases de datos imputada, produciendo igual número de estimaciones para cada parámetro de interés. En el último paso, los resultados de los M análisis son combinados, según las reglas desarrolladas por Rubin (2004), para producir una sola estimación de cada parámetro.

Función de perdida En general, en la estadística se busca representar la realidad por medio de un método que se plantea, de tal forma que con dicho método sea posible acercarse al conocimiento del fenomeno estudiando, teniendo en cuenta que las aproximaciones tengan mínimo error. Como pueden existir diferentes métodos para estudiar el fenomeno, se plantea como indicador de comparación de los métodos, la función de perdida, de la siguiente manera:

$$\ell = ||X - tp^{'}||^2$$

Donde $X_{n.p}$ es la matriz original de datos, $t_{n,1}$ son las componentes y $p_{p,1}$ los vectores propios

3.7.1. Algoritmo EM

El algoritmo EM propuesto por Dempster et al. (1977), es un algoritmo iterativo para calcular estimaciones máximo verosímiles en presencia de datos faltantes. Este es un algoritmo iterativo de dos pasos alternando entre completar los valores faltantes con su media condicional, dado las respuestas observadas y los parámetros estimados en la iteración previa

Paso E : $E[y^m|y^0, x, \theta]$

Luego, actualiza las estimaciones maximizando la verosimilitud de los resultados de los datos completos (Paso M).

Algo importante es que en cada iteración se cumple que $L(\theta^{t+1}|y^0) \geq L(\theta^t|y^0)$, lo cual garantiza la convergencia a un máximo (Schafer, 1997).

El algoritmo EM es uno de los algoritmos más usados para trabajar los datos faltantes bajo el enfoque de la imputación. A continuación, se presenta la propuesta del algoritmo EM en el contexto de Análisis de Componentes Principales y Análisis de Correspondencias Múltiple.

3.7.2. ACP-EM (ACP iterativo)

El método Análisis de Componentes Principales vía EM (ACP-EM), es un método propuesto por Josse and Husson (2012), él cual realiza un Análisis de Componentes principales en presencia de datos faltantes. Particularmente, los datos faltantes son estimados inicialmente por valores promedio y luego se desea minimizar la distancia entre los valores originales (Z) y los valores estimados ($\psi u^{'}$). De esta manera se propone minimizar la siguiente función:

$$\ell = w||(Z - \psi u^{'})||^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}(Z_{ij} - \psi_{i,\alpha}u^{'}_{\alpha,j})^2$$

donde:

Z_{n*p} es la matriz original de datos (estandarizada)
ψ_{n*q} las componentes

u_{p*q} los vectores propios

$\alpha = 1, 2, ..., q$. Donde $q < p$

w es una variable indicadora, de tal manera que $0 = NA$, $1 =$ Valor Observado.

Teniendo en cuenta el criterio anterior, se tiene que la función de perdida se cálcula solo para los valores completos (o sin NA). Entonces, el método ACP-EM sigue los siguientes pasos:

Pseudocódigo Algoritmo ACP-EM 1. Iniciación $L = 0$:

Z^0 Los datos faltantes son reemplazados por valores iniciales, como por ejemplo la media.

2. Paso L

2.1 Realize un ACP para calcular ψ^L, u^L, tomando q dimensiones

Observación: Al tomar $q < p$ se garantiza el proceso iterativo, pero es posible hacer un ACP utilizando p dimensiones, en este caso usamos toda la información de las componentes ψ. Usar q dimensiones garantiza un proceso iterativo que podría no tenerse en cuenta al usar las p dimensiones. Algo importante a mencionar es que la selección de q se encuentra por medio de validación cruzada generalizada.

2.2 Los valores faltantes son imputados via reconstitución de la matriz $Z^L = \psi^L u'^L$.

Los valores observados son los mismos y los faltantes se reemplazan por la imputación

3. El paso 2.1 y 2.2 se repiten hasta la convergencia.

Observaciones: Una de las observaciones que se puede hacer al método es que tiene problemas de robustes, puesto que los valores ψ y u dependen de la imputación inicial, la cuál no necesariamente debe ser el promedio, por ejemplo: puede ser la mediana, la moda, el mínimo o el máximo. Otra observación es la forma cómo se realiza reconstitución para imputar en Z puesto que la reconstitución se realiza usando q dimensiones, en vez de usar todas las p dimensiones ($p > q$). Para seleccionar el valor de q se utiliza la validación cruzada, pero la pregunta es si se utiliza solo en la primera iteración o si en todas la iteraciones se usa la misma validación cruzada, o si el q es el mismo para todas las iteraciones. Con lo anterior, se muestra que el algoritmo ACP iterativo (ACP-EM) tiene algunas dificultades.

3.7.3. ACM-EM (ACM iterativo)

El Análisis de Correspondencias Múltiples vía EM (ACM-EM) fue propuesto por Josse et al. (2012), éste método se basa en el ACP-EM, donde los datos faltantes se estiman por valores promedios y luego se minimizan las distancias entre los datos originales So_{ij} y los datos estimados $\psi u^{'}$. De esta manera, el ACM-EM utiliza la siguiente función de perdida:

$$\ell = ||w(So - \psi u^{'})||^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}(So_{ij} - \psi_{i\alpha}u^{'}_{\alpha j})^2$$

Donde:

$So = Mn^{1/2}FMp^{1/2}$

$Z_{n.p}$ es la tabla disjuntiva completa

$\psi_{n.q}$ son las componentes principales

$u_{p.q}$ es el vector propio en R^p

$\alpha = 1, 2, ..., q$. Donde $q < p - s$

w es una variable indicadora $0 = NA$ y $1 =$ Valor Observado.

Igualmente que en ACP-EM, el método minimiza la función de perdida asociada a los datos completos. A continuación se presenta el Pseudocódigo asociado al método ACM-EM

Pseudocódigo Algoritmo ACM-EM 1. Iniciación $L = 0$: Z^0

Los datos faltantes son reemplazados por la proporción de unos, en la tabla disjuntiva completa Z_{ij}. NOTA: El reemplazo de los datos faltantes debe sumar 1 por variable, lo cual hace que la marginal por fila sea igual a s como en datos completos

Ejemplo: $A = 0{,}4$, $B = 0{,}3$, $C = 0{,}3$

2. Paso L

2.1 Realize una descomposición singular de la matriz $So = Mn^{1/2}FMp^{1/2}$ (Aqui se obtienen: ψ y u)

2.2 Realizar la reconstitución de la matriz $\widehat{So} = \psi u'$, utilizando q dimensiones $(q < p - s)$.

2.3 Realizar la reconstitución hasta la tabla disjuntiva completa $\widehat{Z} = M_n^{-1/2}(\widehat{S}_o * ns)M_p^{-1/2}$. Aqui los valores faltantes se imputan con la reconstitución y los valores observados de Z son los mismos.

3. El paso 2.1, 2.2 y 2.3 se repiten hasta la convergencia.

Observaciones: Al igual que el ACP iterativo el ACM iterativo realiza la reconstitución usando q dimensiones, los cuales se eligen usando validación cruzada, pero ese número q puede ser el número total de ejes $(p - s)$, con lo cual se tendría

que la función de perdida es igual a cero y las iteraciones no serán necesarias. Por otro lado, en el articulo ACM iterativo se menciona que solo es posible encontrar una convergencia, si solo se imputan los valores como en el paso $L = 0$, si no se imputan de esa forma no es posible encontrar la convergencia en ψ y u.

Otra observación, es que al final (al obtener la convergencia) para indicar a que modalidad pertenece el individuo con NA, se hace una aproximación, en donde **los valores altos les asigno un 1 y a los valores pequeños les asigno un 0**, una situación parecida a al modelo logístico cuando se quiere analizar a partir de que valor θ asigno el éxito o el fracaso.

3.7.4. Selección del número de ejes q (Validación cruzada)

Una opción es dividir el conjunto de n observaciones en n sub-muestras de tamaño $n-1$ mediante el mecanismo de dejar por fuera una observación diferente cada vez. (Leave-one-out cross-validation)

Si denotamos $(\widehat{Z}_i)^q$ a la estimación de Z obtenida al suprimir de la muestra la observación i, entonces la observación Zi sería una observación adicional que podríamos utilizar para construir un estimador de MSE que denotaremos $CV(s)$ y que llamaremos criterio de validación cruzada

$$CV(s) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(Z_{ij} - (\widehat{Z}_{ij})^q)^2$$

MSE: Error cuadratico medio de la predicción

4. ACMpdd: Análisis de Correspodencias Múltiples bajo el principio de datos disponibles

En este capitulo se presenta el método ACMpdd en presencia de datos faltantes, se muestra como se obtienen los valores y vectores propios en el espacio R^n y en R^p. A partir de éstos resultados se tienen en cuenta las relaciones de transición para poder hallar las componentes en cada espacio. Además, se presentan las expresiones de Inercia total, Inercia por pregunta e Inercia por modalidad. Con este método se adapta la propuesta de Wold et al. (1966) y el Análisis de Correspondencias Múltiples, para trabajar con datos faltantes. A continuación se presenta el método ACMpdd:

4.1. Presentación del método 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 en la Tabla **4-1**.

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

Seguido a este paso se cálcula la matriz de frecuecias relativas $F^*_{n.p} = Z^*_{n.p}/k^*$. En este proceso es importante tener en cuenta que al construir F^* se divide Z^*/k^*, donde $k^* = ns^* = \sum Z^*_{i.} = \sum Z^*_{.j}$ se obtiene sumando por filas o por columnas para obtener las marginales correspondientes, esta vez con los datos disponibles.

Individuo	Género		Religión			Raza		
1	1	0	1	0	0	1	0	0
2	1	0	NA	NA	NA	0	1	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

Tabla 4-1.: Ejemplo: Tabla disyuntiva completa $Z^*_{n.p}$ con NA

4.2. ACMpdd en el espacio R^p

Usando la relación presente en ACM y ACP, se obtiene que la matriz a diagonalizar es S^*:

$$S^* = M_p^{*1/2} F' M_n^* F^* M_p^{*1/2} \text{ con la restricción } u' M_p^{'*1/2} M_p^{*1/2} u = 1$$

$$\text{Donde } w = M_p^{*1/2} u; \qquad w' w = 1$$

En el proceso de diagonalización de la matriz S^* se tiene el siguiente sistema de valores y vectores propios: $S^* w = \lambda w$.

Entonces, S^* contiene las submatrices $S_o^* = M_n^{*1/2} F^* M_p^{*1/2}$ tal que $S^* = S_o'^* S_o^*$. Es importante mencionar que M_n^* y M_p^* se obtiene con los datos disponibles y corresponden a las matrices de métricas para filas y columnas respectivamente, son matrices diagonales y en su diagonal contienen dichos pesos. En detalle, las matrices tienen la siguiente estructura:

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

De esta manera, se tiene que la matriz $S^*_{on.p} = Mn^{*1/2} F^* Mp^{*1/2}$, esta matriz no contiene datos faltantes, ya que al realizar los productos escalares se tiene en cuenta el principio de datos disponibles.

Finalmente, para la matriz $S^*_{on.p}$ se realiza una descomposición en valores singulares de manera iterativa, tal cual como lo hace el algoritmo NIPALS.

4.3. Pseudocódigo ACMpdd

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

$$S^*_0 = M^{*1/2}_n F^* M^{*1/2}_p$$

   $M^*_p = [\ddots 1/f_{.j} \ddots]$; $M^*_n = [\ddots 1/f_{i.} \ddots]$
4. Se aplica un NIPALS (no estandarizado) para la matriz S^*_0

4.4. ACMpdd en el espacio R^n

En la sección 4.1, se presentó el método en el espacio asociado a las variables R_p, éste mismo esquema se puede identificar en la nube de individuos de la siguiente manera:

Se construye la matriz $T^*_{n.n}$, la cuál es la matriz a diagonalizar en la nube de individuos. La construcción de la matriz T^* se realiza teniendo en cuenta el principio de datos faltante, ya que $F^*_{n.p}$ contiene registros NA.

$$T^* = F^* M^*_p F^{*'} M^*_n$$

Si se diagonaliza T. Entonces se tiene el siguiente sistema de valores λ y vectores propios v :

$$T^* v = \lambda v$$

Como T^* no necesariamente es simetrica y por ende no tiene vectores propios ortonormales. Se realiza la siguiente transformación $r = M^{*1/2}_n v$, tal que $r'r = 1$. Entonces:

$$M^{*1/2}_n T^* v = \lambda M^{*1/2}_n v$$

$$M_n^{*1/2} F^* M_p^* F^{*'} M_n^{*1/2} M_n^{*1/2} v = \lambda M_n^{*1/2} v$$

$$T^* r = \lambda r; \qquad r^{'} r = 1$$

Es importante mencionar que la matriz $T_{n.n}^*$ se construye teniendo en cuenta el principio de datos disponibles y que a partir de esta matriz se ecuentran los valores λ y vectores propios r.

Una situación más importante que se tiene en este procedimiento es que los valores propios λ en el espacio R^p y R^n son equivalentes, lo que hace que las relaciones de transición sean legitimas y pueda encontrar las coordenadas ψ y φ.

4.5. Relaciones de Transición

Como se mencionó anteriormente el método ACMpdd garantiza que los valores propios en los espacios R_n y R_p son equivalentes, de esta manera podemos relacionar las coordenadas de un espacio con las coordenadas de otro, teniendo en cuenta las siguientes expresiones:

$$\varphi_\alpha = \frac{1}{\sqrt{\lambda}} M_p S_o^{'} \psi_\alpha$$

$$\psi_\alpha = \frac{1}{s^* \sqrt{\lambda_\alpha}} S_o \varphi_\alpha$$

4.6. 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:

$$\psi = S_o^* w$$

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

4.7. Expresiones de Inercias para datos disponibles

En esta sección se tienen las nuevas expresiones de Inercia Total, Inercia por modalidad e Inercia por pregunta, de tal forma que estas expresiones tienen en cuenta el principio de datos faltantes, se trabaja s^* que es la marginal estimada por fila y se reemplaza en cada expresión de Inercia. A continuación se tienen las nuevas expresiones de Inercia:

Inercia Total: Esta Inercia depende del número de $NA's$ que existan puesto que si hay más, entonces s^* es más pequeño y por ende la Inercia Total aumenta.

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

Inercia por modalidad :

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

Inercia por pregunta

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

donde p_q es el número de modalidades en la pregunta q

Observación : En el capítulo de resultados se encontró que expresió de Inercia Total es equivalente a la sumatoria de los valores propios siempre y cuando la marginal por fila sea constante, ésto se observó en los casos 1, 2 y 3 NA por fila. Para los demás casos se encontró que las expresión es aproximada a la sumatoria de los valores propios. Para entender éstas inercias se puede observar el caso de datos completos en ACM 3.4.2

4.8. Propiedades del método ACMpdd

A continuación se listan algunas propiedades del método ACMpdd:

- Equivalencia de ACMpdd y ACM cuando la matriz de datos no tiene registros faltantes. Esta propiedad se deriva del algoritmo NIPALS particularmente al usar el principio de datos disponibles (Tenenhaus, 1998)
- Ortogonalidad de las coordenadas ψ y φ. Fue una propiedad que se verificó en el capitulo de Resultados
- Valores propios decrecientes (Se máxima la inercia disponible en los nuevos ejes ψ). También se verificó en el capitulo de Resultados

5. Metodología

En este capitulo, se presentan unos aspectos importantes para la elaboración de los análisis con el método ACMpdd. En primer lugar, es importante mencionar que el primer objetivo específico fue desarrollado en el Capitulo 4, donde se encontraron las expresiones exactas para implementar el principio de datos disponibles de NIPALS en el contexto de ACM con datos faltantes. Ahora bien, en este capitulo se describen las bases de datos **BreedsDogs** y **tea**, las cuales son bases de datos públicas y se encuentran en los paquetes **FactoClass** y **FactoMineR** del software R (Pardo and Del Campo, 2007) (Husson et al., 2017a). También, se explican los escenarios de simulación propuestos, los cuales dependen del número datos faltantes (NA) por fila. Posteriormente se menciona que indicadores se tendrán en cuenta en el Análisis Estadístico para cada uno de los escenarios de simulación.

5.1. Descripción de la matriz de datos BreedsDogs (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) y Agresividad (AGR), dichas variables tienen entre 2 y 3 modalidades, tal como se ilustra en la tabla **5-1**.

La variable TAM tiene 3 modalidades, las cuales son: Grande (gra), Mediano (med), Pequeño (peq). La variable PES también tiene 3 modalidades, que son: Pesado (pes), Mediano (med) y Liviano (liv). También se observa que la variable VEL tiene 3 modalidades: Alta (alt), Media (med), Lenta (len). Otra variable con 3 modalidades es INT, donde sus modalidades son: Alta (alt), Media (med) y Baja (baj). Por otro lado, las variables AFE y AGR tienen 2 modalidades, las cuales son Alta (alt) y Baja (baja). Es importante mencionar que estas variables cualitativas se trabajan como variables indicadoras. De esta manera, el estudio se realiza en una matriz de datos Z_{ij} la cuál contiene 1, 0 o NA dependiendo si la modalidad esta presente, ausente o si hay

dato faltante. Es importante mencionar que la matriz Z_{ij} es de dimensión $n * p$ (n : Individuos , p : de modalidades)

	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
chih	peq	liv	len	baj	alt	baj
cock	med	liv	len	med	alt	alt
coll	gra	med	alt	med	alt	baj
dalm	med	med	med	med	alt	baj
dobe	gra	med	alt	alt	baj	alt
dogo	gra	pes	alt	baj	baj	alt
foxh	gra	med	alt	baj	baj	alt
foxt	peq	liv	med	med	alt	alt
galg	gra	med	alt	baj	baj	baj
gasc	gra	med	med	baj	baj	alt
labr	med	med	med	med	alt	baj
masa	gra	med	alt	alt	alt	alt
mast	gra	pes	len	baj	baj	alt
peki	peq	liv	len	baj	alt	baj
podb	med	med	med	alt	alt	baj
podf	gra	med	med	med	baj	baj
poin	gra	med	alt	alt	baj	baj
sett	gra	med	alt	med	baj	baj
stbe	gra	pes	len	med	baj	alt
teck	peq	liv	len	med	alt	baj
tern	gra	pes	len	med	baj	baj

Tabla 5-1.: Visualizaión de la base de datos perros

5.2. Descripción de la matriz de datos tea

Los datos utilizados aquí se refieren a un cuestionario sobre el consumo de té. Se le preguntó a 300 personas cómo beben té (19 preguntas), cuál es la percepción de su producto (12 preguntas) y algunos detalles personales (4 preguntas). La base de datos se encuentra descrita en el libro ‘‘*Exploratory Multivariate Analysis by Example Using R*” (Husson et al., 2017b). La idea con ésta base de datos es analizar el comportamiento de ACMpdd en un escenario de mayor de dimensión. Se seleccionaron 11 preguntas para realizar el ACM con datos completos y faltantes. La descripción de las 11 preguntas del cuestionario se observa en la Tabla **5-2**.

Pregunta	**Descripción**	**Categoría**
P1	Sexo del encuestado	F:Femenino, M:Masculino
P2	Ocupación	O: Operario, MT: Medio tiempo AG: Alta gerencia, EM: Empleado OT: Otro trabajo, ES: Estudiante, NT: No trabaja
P3	¿Realiza deporte?	Si , No
P4	Edad del encuestado	”15-24”, ”25-34”, ”35-44”, ”45-59”, ”+60”
P5	¿Qué tipo de té bebes más?	N: té negro, V: té verde, S: té con sabor
P6	¿Cómo tomas tu té?	N: nada añadido, LI: con limón LE: con leche, O: otro
P7	¿Qué tipo de té compras?	B: bolsitas de té, TS: té suelto, A: ambos
P8	¿Agregas azúcar a tu té ?	1:Si, 2: No
P9	¿Dónde compras tu té ?	S: en el supermercado, TE: en tiendas especializadas A: ambos
P10	¿Qué tipo de té compras ?	EB: el más barato, DM: de marca EP: etiqueta privada, EX: exclusivo, V: variable, DS: desconocido
P11	¿Con qué frecuencia bebes té?	1D: una vez al día +2D: más de dos veces al día 1-2S: una o dos veces por semana 3-6S: 3 a 6 veces a la semana

Tabla 5-2.: Diccionario de datos para la base de datos tea

5.3. Generación de datos faltantes para BreedsDogs

Para la base de datos BreedsDogs se asignarán datos faltantes aleatoriamente, teniendo en cuenta que por fila pueden existir de 0 a 3 datos faltantes, considerando que se permita el 50 % de la información con NAs. A continuación se ilustra como se realiza la generación de datos faltantes.

En la siguiente estructura: j corresponde a la posición donde se ubica el dato faltante y w corresponde al número datos faltantes por fila, el cuál se acota como se meciono anteriormente (50 % de la información con NA´s)

```
j <-seq(1,ncol(dat.act)) # Posición del NA
w <-seq(0,ncol(dat.act)-ncol(dat.act)/2) # Número de NA´s

for(i in 1:nrow(dat.act)){

  dat.act[i,sample(j,sample(w,1))] <- NA

}
```

5.4. Escenarios de simulación para BreedsDogs

Como se observa en la Tabla **5-3** se tienen los escenarios de simulación propuestos, los cuales tienen presente un mécanismo de datos faltantes completamente aleatorio (MCAR). A su vez, se consideran escenarios cuando toda la matriz de datos tiene 1, 2 o 3 NA por fila, en estos 3 primeros escenarios, la marginal por fila $Z_{i.}$ es constante para todo i. Ahora bien, cuando se tiene de 0:1, 0:2 y 0:3 NA por fila, en estos últimos escenarios la marginal por fila $Z_{i.}$ ya no es constante para todo i. Importante mencionar que la marginal por columna no es constante para las j modalidades.

También es importante mencionar que el porcentaje de NAs se calcula con base al total de registros $(27 * 6)$. En este prooyecto se trabajó con un máximo de % 50 de NAs, el % de NAs en los últimos 3 escenarios se generó aleatoriamente.

Estructura NA	Cantidad NAs	% NAs
MCAR	1 NA por Fila	16.7 %
MCAR	2 NA por Fila	33.3 %
MCAR	3 NA por Fila	50 %
MCAR	0:1 NA por Fila	9.26 %
MCAR	0:2 NA por Fila	13.58 %
MCAR	0:3 NA por Fila	27.16 %

Tabla 5-3.: Escenarios de simulaión base de datos BreedsDogs

5.5. Análisis Estadístico de los escenarios de simulación

En primer lugar, se debe analizar la matriz con datos completos y ver 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

Colocando este punto de partida, se analiza los mimos indicadores para cada uno de los escenarios propuestos en la Tabla **5-3**. En cada uno de éstos análisis se identifican si las expresiones de inercia concuerdan con la teoría en datos completos. Posteriormente, en el capitulo 8 se hace una comparación con los métodos de imputación la cuál tiene presente el esquema de la Tabla **5-4**, donde m es igual al número de matrices a simular con dicha estructura. Esta comparación se realiza tanto para la base de datos BreedsDogs como para la base de datos tea.

Estructura NA	Métodos	% NAs	m
0:3 NA por Fila para BreedsDogs	ACMpdd ACM-EM	5 %	1000
		10 %	1000
		15 %	1000
		20 %	1000
		25 %	1000
		30 %	1000
		50 %	1000

Tabla 5-4.: Escenarios de simulaciǿn para el Análisis del Poder Descriptivo

5.6. Software

En la sección A.4 del Capitulo anexos, se presenta el código desarrollado en el software R, con el cuál se realizó el método **ACMpdd**, este código es un resultado de este proyecto de investigación. Para realizar los planos factoriales se uso la función **s.label** de la librería ade4 (Dray et al., 2007). Para el análisis con los métodos de imputación se utilizó la librería **missMDA** (Husson et al., 2013) también del software R en su versión 3.4.0.

6. Resultados

En este capitulo se presentan los resultados encontrados en cada escenario simulación usando el método ACMpdd, en cada escenario se muestra que es posible trabajar un ACM en presencia de datos disponibles, es importante mencionar que los indicadores que se análizan en cada escenario son los mencionados en la sección 5.5 de la metodología. En primer lugar, se muestra el caso con datos completos para la base de datos BreedsDogs, el cual se ilustra a continuación:

6.1. Datos completos para BreedsDogs

En esta sección se presentan los resultados de las Estadísticas descriptivas para la matriz con datos conmpletos. Además, se muestran los resultados relacionados con Análisis de Correspondencias Múltiples, donde es interés analizar la Inercia total, por modalidad y pregunta. Importante analizar las coordenadas ψ_α y φ_α. Igualmente se analiza el poder descriptivo en los dos primeros ejes factoriales.

En la figura **6-1** se observa que para la variable Tamaño la modalidad más frecuente es la modalidad grande, se observa que el 55.5 % de los perros son de tamaño grande, el 18.5 % son tamaño mediano y el 25.9 % de tamaño pequeño. Tambien se observa que para la variable Peso, la modalidad con mayor frecuencia es peso mediano que tiene una frecuencia del 51.8 %, la modalidad peso liviano tiene una frecuencia 29.6 % y 18.5 %. Tambien se observa que en la variable velocidad la modalidad con mayor frecuencia es lenta, la cual tiene una frecuencia de 37.03 %, seguido de la modalidad alta con 33.3 % y modalidad media con 29.6 %. Además se observa que la variable Inteligencia, tiene la modalidad media como la más frecuente con un porcentaje del 48.1 %, seguido de 29.6 % en modalidad baja y 22.2 % en modalidad alta. Las variables afectividad y agresividad tienen dos modalidades alta y baja, las cuales tienen porcentajes 51.8 % y 48.1 % en la variable afectividad y en agresividad los porcentajes están invertidos.

Ahora bien en la Tabla **6-1** se observa que la suma por fila $Z_{i.}$, la cuál indica que hay 6 variables cualitativas. Posteriormente se tiene la Tabla **6-2** la cuál es la suma por

columna en la tabla disjuntiva completa, ésta se denota por Z_j, que como se observa por cada variable la suma $Z_{.j}$ coincide con el número $n = 27$ razas de perros.

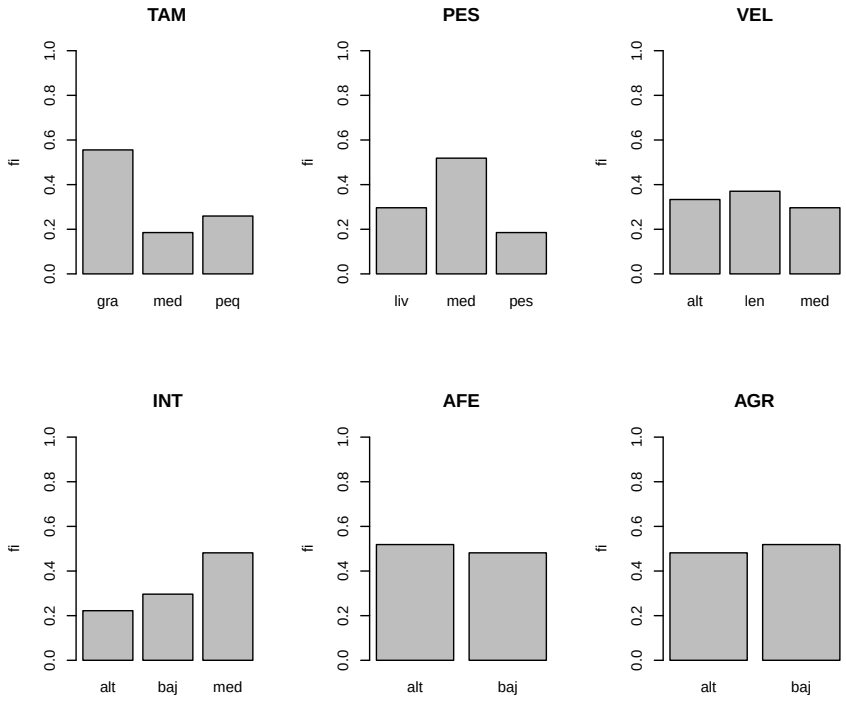

Figura 6-1.: Diagrama de Barras de la base de datos perros

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	steck	tern	
6	6	6	6	6	6	6	6	6	6	6	6	6	

Tabla 6-1.: $Zi.$ para datos completos

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

Tabla 6-2.: $Z.j$ con datos completos

En la Tabla **6-3** se observan los valores propios en el espacio R^p, los cuáles coinciden con los valores propios del espacio R^n que se encuentran en la Tabla **6-4**. Es importante recordar que esta equivalencia en los valores propios garantiza que se cumplan las relaciones de transición. Finalmente, se analizan los valores propios obtenidos, donde

	1	2	3	4	5	6	7	8	9	10
λ_α	0.4816	0.3847	0.2110	0.1576	0.1501	0.1233	0.0815	0.0457	0.0235	0.0077

Tabla 6-3.: Valores propios espacio R^p datos completos

1	2	3	4	5	6	7	8	9	10
0.4816	0.3847	0.2110	0.1576	0.1501	0.1233	0.0815	0.0457	0.0235	0.0077
11	12	13	14	15	16	17	18	19	20
0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	22	23	24	25	26	27			
0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000				

Tabla 6-4.: Valores propios espacio R^n datos completos

es importante mencionar que la sumatoria de los valores propios (1.6667) coincide con $(p/6) - 1 = 1{,}6667$, la cual se conoce como la Inercia Total.

$$\sum \lambda_i = 1{,}6667 = (p/s) - 1 = (16/6) - 1 = 1{,}6667$$

En la Tabla **6-5** se tienen las Inercias por Modalidad, las cuales al sumarlas coinciden con la Inercia Total. Además se observa que en la Tabla **6-6** se tienen las Inercias por pregunta, igualmente al sumarlas coinciden con la Inercia Total.

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len
Ij	0.0741	0.1358	0.1235	0.1173	0.0802	0.1358	0.1111	0.1049
	VEL.med	INT.alt	INT.baj	INT.med	AFE.alt	AFE.baj	AGR.alt	AGR.baj
	0.1173	0.1296	0.1173	0.0864	0.0802	0.0864	0.0864	0.0802

Tabla 6-5.: Inercia por Modalidad

En la Figura **6-2**, se tiene la representación simultánea donde se presentan las relaciones entre las razas de perros y las modalidades asociadas a sus variables. Por ejemplo, se observa que las razas de perros con Afectividad alta tienen Agresividad baja e Inteligencia media; tambien se observa que las razas con Agresividad alta tienen Afectividad baja y Tamaño grande; a su vez se observa que la modalidad Tamaño pequeño esta asociada con la modalidad Peso liviano. En éste plano factorial se observan algunas similitudes en las razas de perros, por ejemplo: foxt, cock y cani tienen son razas de perros con características similares.

	TAM	PES	INT	VEL	AGR	AFE
1	0.3333	0.3333	0.3333	0.3333	0.1667	0.1667

Tabla 6-6.: Inercia por Pregunta

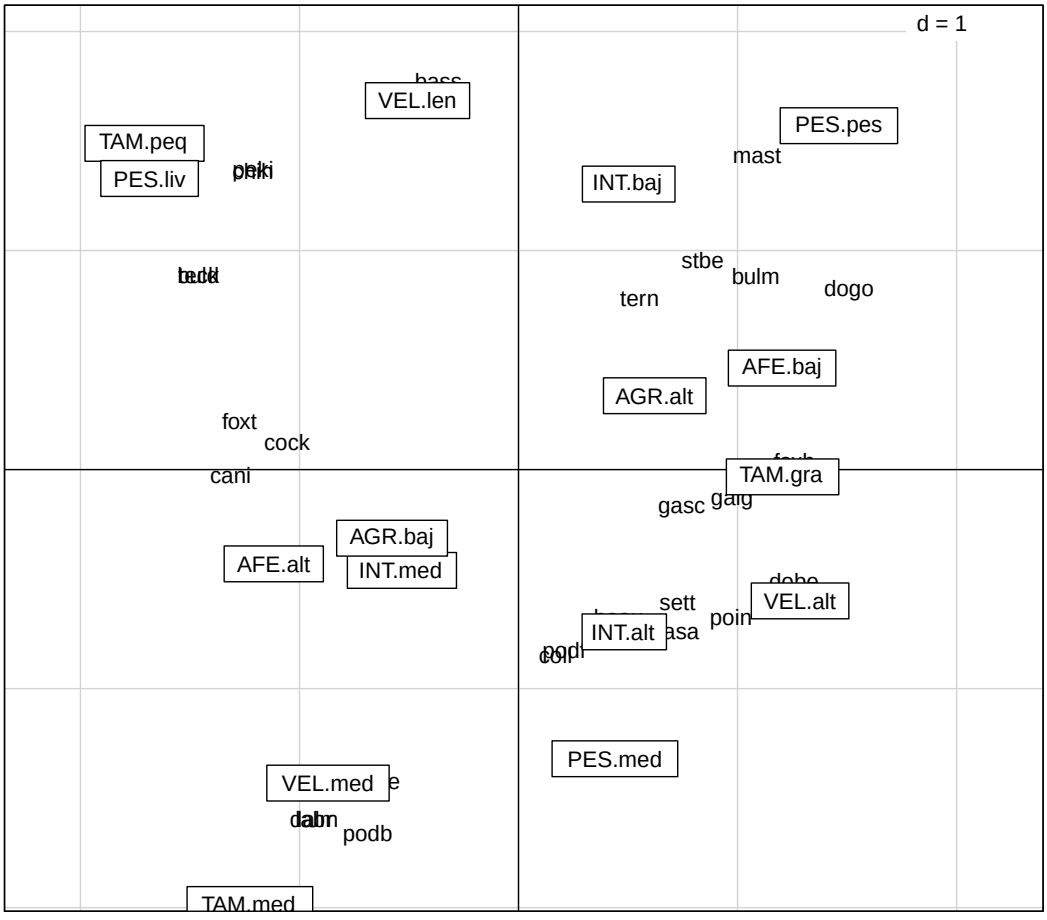

Figura 6-2.: Representación simultánea con datos completos

6.2. 1 NA por Fila BreedsDogs

A continuación se presenta el método ACMpdd cuando existe 1 NA aleatorio por cada individuo, para trabajar en este escenario, en la tabla **6-7** se ilustra la estructura de la base de datos.

Ahora bien en la Tabla **6-8** se observa que la suma por fila Z_i, la cuál indica que por cada raza de perro existe 1 NA. Posteriormente se tiene la Tabla **6-9** la cuál es la suma por columna en la tabla disjuntiva completa, ésta se denota por Z_j, que como se observa los resultados de cada suma por modalidad son variables.

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

Tabla 6-7.: Base de Datos con 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	5	5	5	5	5	5	5
galg	gasc	labr	masa	mast	peki	podb	podf	poin	sett	stbe	steck	tern	
5	5	5	5	5	5	5	5	5	5	5	5	5	

Tabla 6-8.: $Zi.$ con 1 NA

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len	VEL.med	INT.alt	INT.baj	INT.med
y	13	5	5	6	8	5	8	7	8	6	7	11
	AFE.alt	AFE.baj	AGR.alt	AGR.baj								
	13	10	12	11								

Tabla 6-9.: $Z.j$ con 1 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5180	0.3827	0.2884	0.2165	0.2072	0.1564	0.1422	0.0884	0.0586	0.0408
	11	12	13	14	15					
	0.0314	0.0295	0.0234	0.0109	0.0055					

Tabla 6-10.: Valores propios espacio R^p con 1 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5180	0.3827	0.2884	0.2165	0.2072	0.1564	0.1422	0.0884	0.0586	0.0408
	11	12	13	14	15	16	17	18	19	20
	0.0314	0.0295	0.0234	0.0109	0.0055	0.0000	0.0000	0.0000	0.0000	0.0000
	21	22	23	24	25	26	27			
	0.0000	0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000			

Tabla 6-11.: Valores propios espacio R^n con 1 NA

También se observan los valores propios en las Tablas: **6-11** y **6-10**. Dichos valores son equivalentes, en ambos espacios. Finalmente, se analizan los valores propios obtenidos, donde es importante mencionar que la sumatoria de los valores propios (2.2) coincide con $(p/5) - 1 = 2{,}2$

$$\sum \lambda_i = 2{,}2 = (p/s) - 1 = (16/5) - 1 = 2{,}2$$

En la Tabla **6-12** se tienen las Inercias por Modalidad, las cuales al sumarlas coinciden con la Inercia Total. Además se observa que en la Tabla **6-13** se tienen las Inercias por pregunta, igualmente al sumarlas coinciden con la Inercia Total (2.2).

En la Figura **6-3**, se tiene la representación simultánea para el caso en el que hay 1 NA por fila. Se observa que las razas de perros con Afectividad alta tienen Agresividad baja; tambien se observa que las razas con Agresividad alta tienen Afectividad baja; a su vez se observa que la modalidad Tamaño pequeño esta asociada con la modalidad Peso liviano. En éste plano factorial se observan algunas similitudes en las razas de perros, por ejemplo: foxt, teck y buld tienen son razas de perros con características similares. Comparando con el plano factorial con datos completos **6-2** se observó que algunas tipologías se mantienen.

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len
Ij	0.1037	0.1629	0.1629	0.1555	0.1407	0.1629	0.1407	0.1481
	VEL.med	INT.alt	INT.baj	INT.med	AFE.alt	AFE.baj	AGR.alt	AGR.baj
	0.1407	0.1555	0.1481	0.1185	0.1037	0.1259	0.1111	0.1185

Tabla 6-12.: Inercia por Modalidad con 1 NA por fila

	TAM	PES	INT	VEL	AGR	AFE
I_q	0.4296	0.4592	0.4296	0.4222	0.2296	0.2296

Tabla 6-13.: Inercia por Pregunta con 1 NA por fila

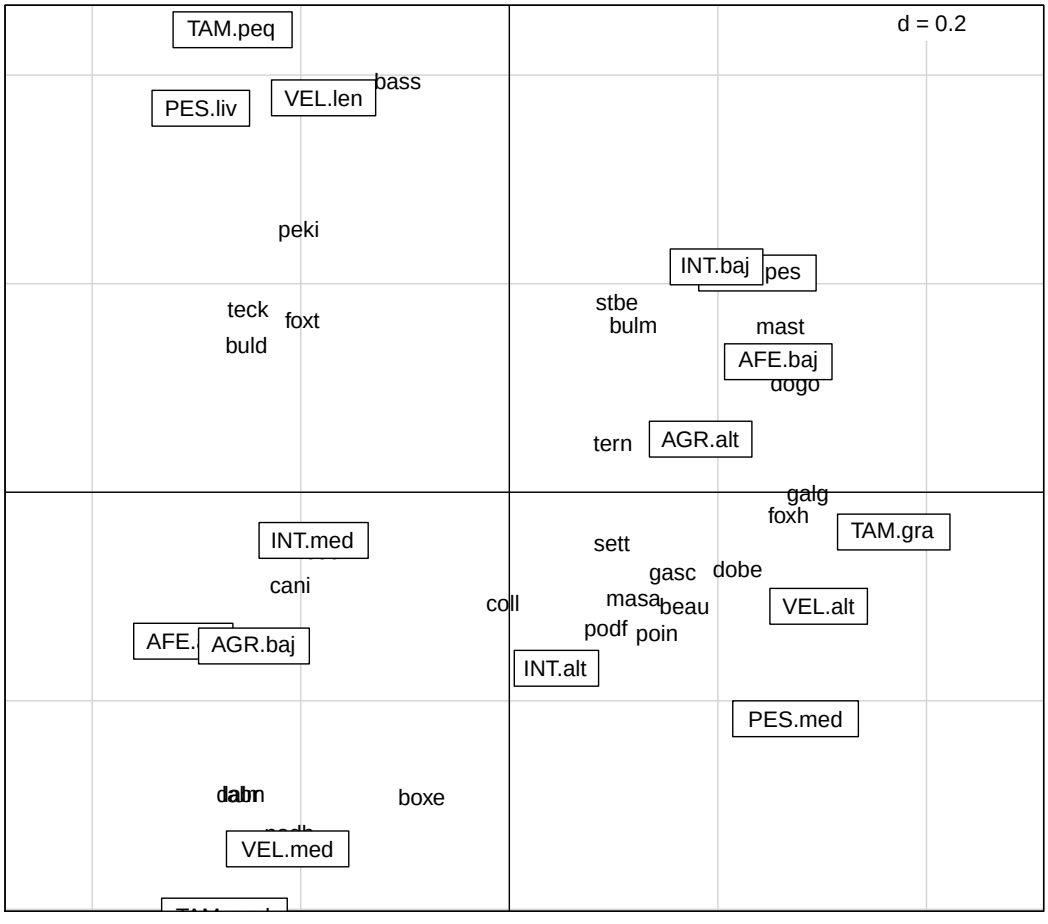

Figura 6-3.: Representación simultánea con 1 NA por fila

6.3. 2 NA por Fila BreedsDogs

En esta sección se presenta el método ACMpdd cuando existe 2 NAs aleatoriamente por cada individuo, para trabajar en este escenario, se ilustra a continuación la estructura de la base de datos, la cuál se observa en la tabla **6-14**.

Ahora bien en la Tabla **6-15** se observa que la suma por fila Z_i, la cuál indica que por cada raza de perro existe 2 NAs. Igualmente se tiene la Tabla **6-16** la cuál es la suma por columna en la tabla disjuntiva completa y al final de esta sección se analizaran los valores propios, de tal forma que se evalua si cumplen con las expresiones de inercia.

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

Tabla 6-14.: Base de Datos con 2 NA por fila

bass	beau	boxe	buld	bulm	cani	chih	cock	coll	dalm	dobe	dogo	foxh	foxt
4	4	4	4	4	4	4	4	4	4	4	4	4	4
galg	gasc	labr	masa	mast	peki	podb	podf	poin	sett	stbe	steck	tern	
4	4	4	4	4	4	4	4	4	4	4	4	4	

Tabla 6-15.: $Zi.$ con 2 NA

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len	VEL.med	INT.alt	INT.baj	INT.med
$Z.j$	12	4	5	4	11	5	6	7	6	3	3	10
	AFE.alt	AFE.baj	AGR.alt	AGR.baj								
	7	8	7	10								

Tabla 6-16.: $Z.j$ con 2 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5558	0.4842	0.3874	0.3154	0.2902	0.2331	0.1967	0.1291	0.1116	0.0943
	11	12	13	14	15					
	0.0659	0.0564	0.0339	0.0234	0.0226					

Tabla 6-17.: Valores propios espacio R^p con 2 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5558	0.4842	0.3874	0.3154	0.2902	0.2331	0.1967	0.1291	0.1116	0.0943
	11	12	13	14	15	16	17	18	19	20
	0.0659	0.0564	0.0339	0.0234	0.0226	0.0000	0.0000	0.0000	0.0000	-0.0000
	21	22	23	24	25	26	27			
	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000			

Tabla 6-18.: Valores propios espacio R^n con 2 NA

Es importante mencionar que la sumatoria de los valores propios (3) coincide con $(p/4) - 1 = 3$

$$\sum \lambda_i = 3 = (p/s) - 1 = (16/4) - 1 = 3$$

En la Tabla **6-19** se tienen las Inercias por Modalidad, las cuales al sumarlas coinciden con la Inercia Total. Además se observa que en la Tabla **6-20** se tienen las Inercias por pregunta, igualmente al sumarlas coinciden con la Inercia Total (3).

En la Figura **6-4**, se tiene la representación simultánea para el caso en el que hay 2 NA por fila. Comparando con el plano factorial con datos completos **6-2** se observó que algunas tipologías siguen igual, como lo son: que las razas de perros con Afectividad alta tienen Agresividad baja; tambien se observa que las razas con Agresividad alta tienen Afectividad baja. En éste plano factorial se observan algunas similitudes en las razas de perros, por ejemplo: boxe, dalm, podb y labr tienen son razas de perros con características similares. En general, se observa que hay cambios en la estructura del plano factorial, al comparar con datos completos, es importante recordar que con esta estructura se tiene un 33.3 % de NAs.

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len
Ij	0.1388	0.2129	0.2037	0.2129	0.1481	0.2037	0.1944	0.1851
	VEL.med	INT.alt	INT.baj	INT.med	AFE.alt	AFE.baj	AGR.alt	AGR.baj
	0.1944	0.2222	0.2222	0.1574	0.1851	0.1759	0.1851	0.1574

Tabla 6-19.: Inercia por Modalidad con 2 NA por fila

	TAM	PES	INT	VEL	AGR	AFE
I_q	0.5555	0.5648	0.5740	0.6018	0.3611	0.3425

Tabla 6-20.: Inercia por Pregunta con 2 NA por fila

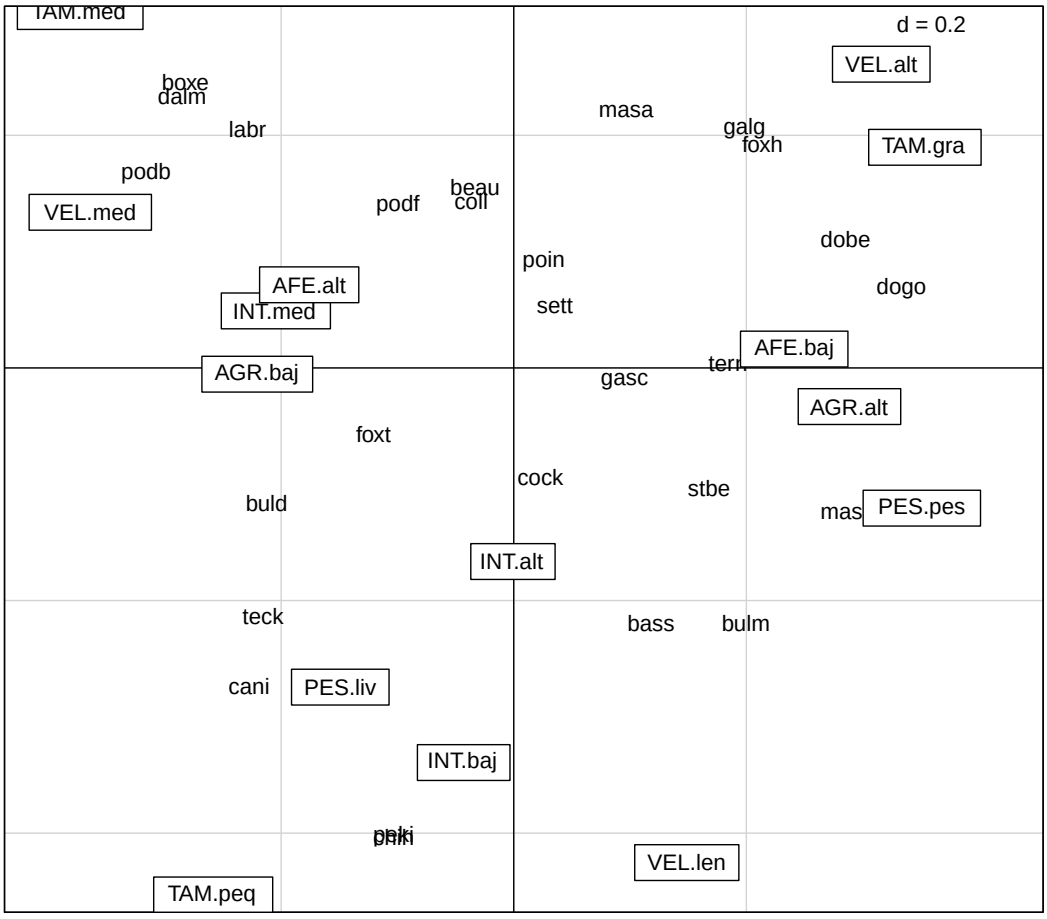

Figura 6-4.: Representación simultánea con 2 NA por fila

6.4. 3 NA por Fila BreedsDogs

En esta sección se presenta el método ACMpdd cuando existe 3 NAs aleatoriamente por cada individuo, para trabajar en este escenario, se ilustra a continuación la estructura de la base de datos, la cuál se observa en la tabla **6-21**. En la Tabla **6-22** se observa que la suma por fila Z_i, la cuál indica que por cada raza de perro existe 3 NAs. También se tiene la Tabla **6-23** la cuál es la suma por columna en la tabla disjuntiva completa y al final de esta sección se analizaran los valores propios, de tal forma que se evalua si cumplen con las expresiones de inercia.

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

Tabla 6-21.: Base de Datos con 3 NA por fila

bass	beau	boxe	buld	bulm	cani	chih	cock	coll	dalm	dobe	dogo	foxh	foxt
3	3	3	3	3	3	3	3	3	3	3	3	3	3
galg	gasc	labr	masa	mast	peki	podb	podf	poin	sett	stbe	steck	tern	
3	3	3	3	3	3	3	3	3	3	3	3	3	

Tabla 6-22.: $Zi.$ con 3 NA

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len	VEL.med	INT.alt	INT.baj	INT.med
y	8	3	3	4	6	2	3	2	3	5	5	7
	AFE.alt	AFE.baj	AGR.alt	AGR.baj								
	8	6	8	8								

Tabla 6-23.: $Z.j$ con 3 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.6858	0.5556	0.4818	0.4638	0.4084	0.4059	0.3385	0.2968	0.1750	0.1540
	11	12	13	14	15					
	0.1444	0.0822	0.0580	0.0481	0.0352					

Tabla 6-24.: Valores propios espacio R^p con 3 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.6858	0.5556	0.4818	0.4638	0.4084	0.4059	0.3385	0.2968	0.1750	0.1540
	11	12	13	14	15	16	17	18	19	20
	0.1444	0.0822	0.0580	0.0481	0.0352	0.0000	0.0000	0.0000	0.0000	-0.0000
	21	22	23	24	25	26	27			
	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000			

Tabla 6-25.: Valores propios espacio R^n con 3 NA

Es importante mencionar que la sumatoria de los valores propios (4.333) coincide con $(p/3) - 1 = 4{,}333$

$$\sum \lambda_i = 4{,}3333 = (p/s) - 1 = (16/3) - 1 = 4{,}33333$$

En la Tabla **6-26** se tienen las Inercias por Modalidad, las cuales al sumarlas coinciden con la Inercia Total. Además se observa que en la Tabla **6-27** se tienen las Inercias por pregunta, igualmente al sumarlas coinciden con la Inercia Total (4.333).

En la Figura **6-5**, se tiene la representación simultánea para el caso en el que hay 3 NA por fila. En este plano factorial se observa un cambio al comparar el plano factorial de datos completos **6-2**, ya que las tipologías mencionadas anteriormente no se observan tan evidentes como en los anteriores planos. Es importante recordar que este escenario de simulación se trabaja con el 50 % de NAs.

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len
Ij	0.2345	0.2962	0.2962	0.2839	0.2592	0.3086	0.2962	0.3086
	VEL.med	INT.alt	INT.baj	INT.med	AFE.alt	AFE.baj	AGR.alt	AGR.baj
	0.2962	0.2716	0.2716	0.2469	0.2345	0.2592	0.2345	0.2345

Tabla 6-26.: Inercia por Modalidad con 3 NA por fila

	TAM	PES	INT	VEL	AGR	AFE
I_q	0.8271	0.8518	0.9012	0.7901	0.4938	0.4691

Tabla 6-27.: Inercia por Pregunta con 3 NA por fila

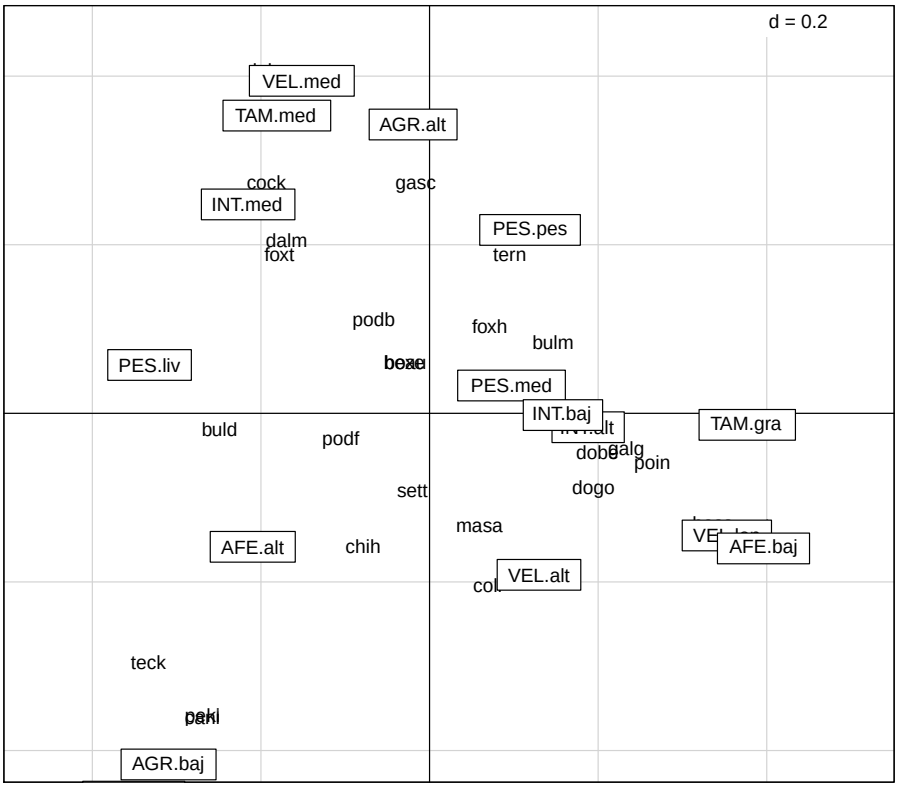

Figura 6-5.: Representación simultánea con 3 NA por fila

6.5. 0:1 NA por Fila BreedsDogs

En esta sección se presenta el método ACMpdd cuando existe de 0 a 1 NA aleatoriamente por cada individuo, para trabajar en este escenario, se ilustra a continuación la estructura de la base de datos, la cuál se observa en la tabla **6-21**. En la Tabla **6-22** se observa que la suma por fila Z_i, la cuál indica que por cada raza de perro existe de 0 a 1 NA. También se tiene la Tabla **6-23** la cuál es la suma por columna en la tabla disjuntiva completa y al final de esta sección se analizaran los valores propios, de tal forma que se evalua si cumplen con las expresiones de inercia.

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

Tabla 6-28.: Base de Datos con 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	

Tabla 6-29.: $Zi.$ con 0:1 NA

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len	VEL.med	INT.alt	INT.baj	INT.med
y	14	4	5	8	12	5	8	10	8	5	7	10
	AFE.alt	AFE.baj	AGR.alt	AGR.baj								
	12	12	13	14								

Tabla 6-30.: $Z.j$ con 0:1 NA

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

Tabla 6-31.: Valores propios espacio R^p con 0:1 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5087	0.4130	0.2562	0.1819	0.1600	0.1223	0.0979	0.0643	0.0464	0.0329
	11	12	13	14	15	16	17	18	19	20
	0.0213	0.0157	0.0101	0.0044	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000
	21	22	23	24	25	26	27			
	0.0000	0.0000	0.0000	0.0000	-0.0000	-0.0000				

Tabla 6-32.: Valores propios espacio R^n con 0:1 NA

Es importante mencionar que la sumatoria de los valores propios (1.936123) se aproxima a $(p/5{,}4444) - 1 = 1{,}938776$

$$\sum \lambda_i = 1{,}936123 \simeq (p/s^*) - 1 = (16/5{,}44444) - 1 = 1{,}938776$$

En la Tabla **6-33** se tienen las Inercias por Modalidad, las cuales al sumarlas coinciden con la Inercia Total. Además se observa que en la Tabla **6-34** se tienen las Inercias por pregunta, igualmente al sumarlas coinciden con la Inercia Total (1.938776).

En la Figura **6-6**, se tiene la representación simultánea para el caso en el que hay de 0 a 1 NA por fila. Este plano factorial es aquel que muestra más similitud al plano factorial con datos completos **6-2**. Se observa que las razas de perros con Afectividad alta tienen Agresividad baja; tambien se observa que las razas con Agresividad alta tienen Afectividad baja; a su vez se observa que la modalidad Tamaño pequeño esta asociada con la modalidad Peso liviano. En éste plano factorial se observan algunas similitudes en las razas de perros, por ejemplo: teck y buld tienen son razas de perros con características similares.

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len
Ij	0.0884	0.1564	0.1496	0.1292	0.1020	0.1496	0.1292	0.1156
	VEL.med	INT.alt	INT.baj	INT.med	AFE.alt	AFE.baj	AGR.alt	AGR.baj
	0.1292	0.1496	0.1360	0.1156	0.1020	0.1020	0.0952	0.0884

Tabla 6-33.: Inercia por Modalidad con 0:1 NA por fila

	TAM	PES	INT	VEL	AGR	AFE
I_q	0.3945	0.3809	0.3741	0.4013	0.2040	0.1836

Tabla 6-34.: Inercia por Pregunta con 0:1 NA por fila

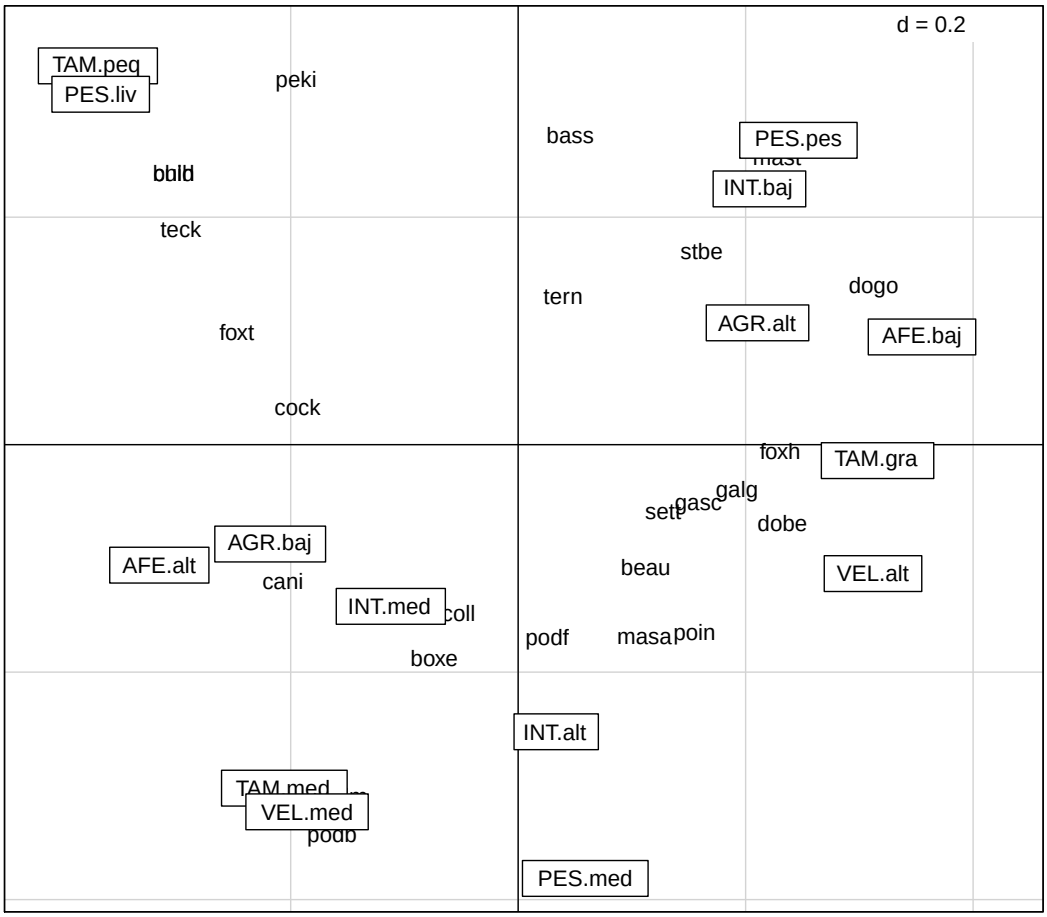

Figura 6-6.: Representación simultánea con 0:1 NA por fila

6.6. 0:2 NA por Fila BreedsDogs

En esta sección se presenta el método ACMpdd cuando existe de 0 a 2 NAs aleatoriamente por cada individuo, para trabajar en este escenario, se ilustra a continuación la estructura de la base de datos, la cuál se observa en la tabla **6-35**. En la Tabla **6-36** se observa que la suma por fila Z_i, la cuál indica que por cada raza de perro existe de 0 a 2 NAs. También se tiene la Tabla **6-37** la cuál es la suma por columna en la tabla disjuntiva completa y al final de esta sección se analizaran los valores propios, de tal forma que se evalua si cumplen con las expresiones de inercia.

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

Tabla 6-35.: Base de Datos con 0:2 NA por fila

	bass	beau	boxe	buld	bulm	cani	chih	cock	coll	dalm	dobe	dogo	foxh	foxt	galg	gasc	labr	masa
x	4	6	6	5	6	6	5	5	6	4	6	6	4	5	6	5	4	4
	mast	peki	podb	podf	poin	sett	stbe	teck	tern									
	6	5	5	4	6	5	6	4	6									

Tabla 6-36.: $Zi.$ con 0:2 NA

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len	VEL.med	INT.alt	INT.baj	INT.med
y	15	3	6	5	11	5	8	9	7	4	7	10
	AFE.alt	AFE.baj	AGR.alt	AGR.baj								
	14	11	12	13								

Tabla 6-37.: $Z.j$ con 0:2 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5132	0.3903	0.2512	0.2000	0.1815	0.1506	0.1316	0.0633	0.0616	0.0487
	11	12	13	14	15	16				
	0.0247	0.0213	0.0131	0.0087	0.0053					

Tabla 6-38.: Valores propios espacio R^p con 0:2 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5132	0.3903	0.2512	0.2000	0.1815	0.1506	0.1316	0.0633	0.0616	0.0487
	11	12	13	14	15	16	17	18	19	20
	0.0247	0.0213	0.0131	0.0087	0.0053	0.0000	0.0000	0.0000	0.0000	0.0000
	21	22	23	24	25	26	27			
	0.0000	0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000			

Tabla 6-39.: Valores propios espacio R^n con 0:2 NA

Es importante mencionar que la sumatoria de los valores propios (2.065059) se aproxima a $(p/5{,}185185) - 1 = 2{,}085714$

$$\sum \lambda_i = 2{,}065059 \simeq (p/s^*) - 1 = (16/5{,}185185) - 1 = 2{,}085714$$

En la Tabla **6-40** se tienen las Inercias por Modalidad, las cuales al sumarlas coinciden con la Inercia Total. Además se observa que en la Tabla **6-41** se tienen las Inercias por pregunta, igualmente al sumarlas coinciden con la Inercia Total (2.085714).

En la Figura **6-7**, se tiene la representación simultánea para el caso en el que hay un de 0 a 2 NAs por fila. Al comparar con el plano factorial con datos completos **6-2** se observó que algunas tipologías se mantienen, aunque no es tan evidente como en el caso de 0 a 1 NA por fila, donde las tipologías si se observan mejor. Se observa que las razas de perros con Afectividad alta tienen Agresividad baja; tambien se observa que las razas con Agresividad alta tienen Afectividad baja; a su vez se observa que la modalidad Tamaño pequeño esta asociada con la modalidad Peso liviano. En éste plano factorial se observan algunas similitudes en las razas de perros, por ejemplo: foxt y teck tienen son razas de perros con características similares.

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len
Ij	0.0857	0.1714	0.1500	0.1571	0.1142	0.1571	0.1357	0.1285
	VEL.med	INT.alt	INT.baj	INT.med	AFE.alt	AFE.baj	AGR.alt	AGR.baj
	0.1428	0.1642	0.1428	0.1214	0.0928	0.1142	0.1071	0.100

Tabla 6-40.: Inercia por Modalidad con 0:2 NA por fila

	TAM	PES	INT	VEL	AGR	AFE
I_q	0.4071	0.4285	0.4071	0.4285	0.2071	0.2071

Tabla 6-41.: Inercia por Pregunta con 0:2 NA por fila

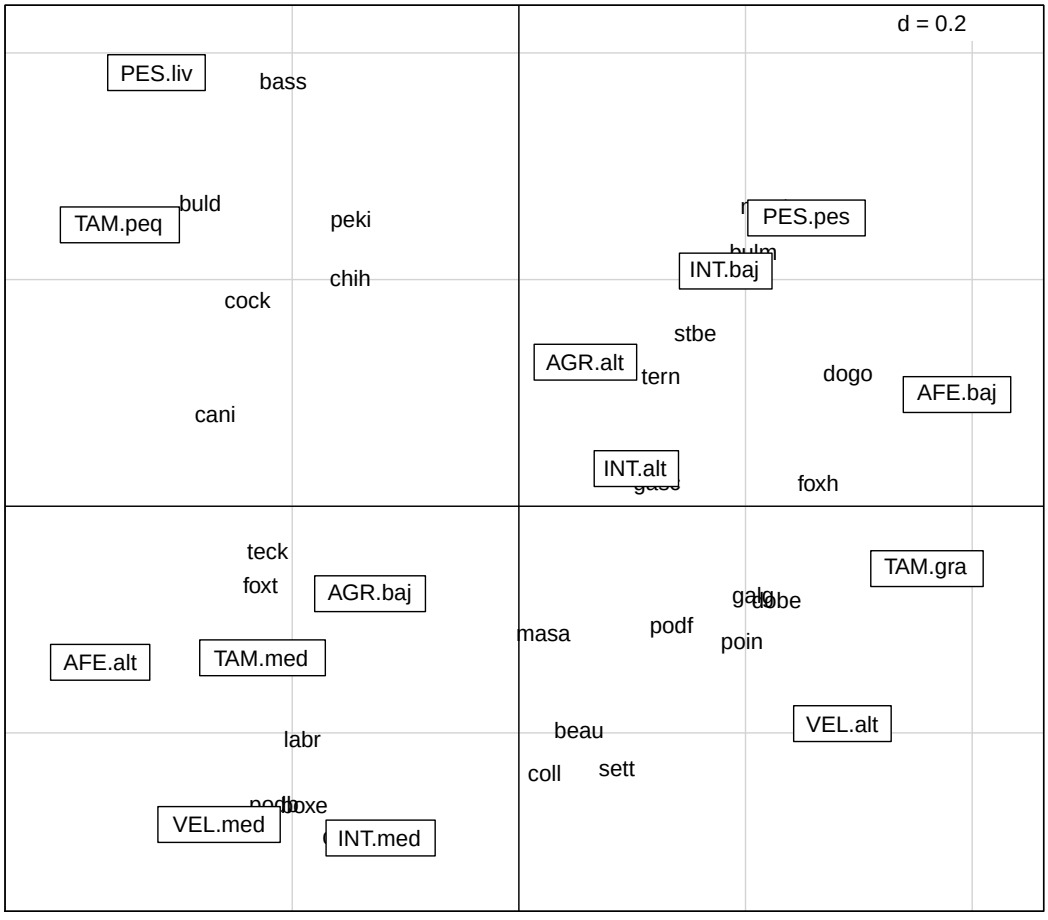

Figura 6-7.: Representación simultánea con 0:2 NA por fila

6.7. 0:3 NA por Fila BreedsDogs

En esta sección se presenta el método ACMpdd cuando existe de 0 a 3 NAs aleatoriamente por cada individuo, para trabajar en este escenario, se ilustra a continuación la estructura de la base de datos, la cuál se observa en la tabla **6-42**. En la Tabla **6-43** se observa que la suma por fila Z_i, la cuál indica que por cada raza de perro existe de 0 a 3 NAs. También se tiene la Tabla **6-44** la cuál es la suma por columna en la tabla disjuntiva completa y al final de esta sección se analizaran los valores propios, de tal forma que se evalua si cumplen con las expresiones de inercia.

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

Tabla 6-42.: Base de Datos con 0:3 NA por fila

	bass	beau	boxe	buld	bulm	cani	chih	cock	coll	dalm	dobe	dogo	foxh	foxt	galg	gasc	labr	masa
$Z_{i.}$	3	3	5	3	5	4	4	5	3	6	5	5	5	6	3	6	3	5
	mast	peki	podb	podf	poin	sett	stbe	teck	tern									
	5	3	3	5	3	5	6	4	5									

Tabla 6-43.: $Zi.$ con 0:3 NA

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len	VEL.med	INT.alt	INT.baj	INT.med
$Z_{.j}$	10	4	7	4	10	5	5	5	6	5	8	8
	AFE.alt	AFE.baj	AGR.alt	AGR.baj								
	9	12	10	10								

Tabla 6-44.: $Z.j$ con 0:3 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5609	0.4422	0.3461	0.2901	0.2597	0.2017	0.1582	0.1137	0.0777	0.0679
	11	12	13	14	15					
	0.0609	0.0457	0.0232	0.0136	0.0086					

Tabla 6-45.: Valores propios espacio R^p con 0:3 NA

	1	2	3	4	5	6	7	8	9	10
λ_α	0.5609	0.4422	0.3461	0.2901	0.2597	0.2017	0.1582	0.1137	0.0777	0.0679
	11	12	13	14	15	16	17	18	19	20
	0.0609	0.0457	0.0232	0.0136	0.0086	0.0000	0.0000	0.0000	0.0000	0.0000
	21	22	23	24	25	26	27			
	0.0000	0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000			

Tabla 6-46.: Valores propios espacio R^n con 0:3 NA

Es importante mencionar que la sumatoria de los valores propios (2.670013) se aproxima a $(p/4{,}37037) - 1 = 2{,}661017$

$$\sum \lambda_i = 2{,}670013 \simeq (p/s^*) - 1 = (16/4{,}37037) - 1 = 2{,}661017$$

En la Tabla **6-47** se tienen las Inercias por Modalidad, las cuales al sumarlas coinciden con la Inercia Total. Además se observa que en la Tabla **6-20** se tienen las Inercias por pregunta, igualmente al sumarlas coinciden con la Inercia Total (2.661017).

En la Figura **6-8**, se tiene la representación simultánea para el caso en el que hay de 0 a 3 NAs por fila. Ya no es tan clara la relación la similitud en las modalidades de las variables Afectividad y Agresividad. Una tipología que si se mantiene es Tamaño pequeño con la modalidad Peso liviano. En éste plano factorial se observan algunas similitudes en las razas de perros, por ejemplo: cani y foxt.

	TAM.gra	TAM.med	TAM.peq	PES.liv	PES.med	PES.pes	VEL.alt	VEL.len
Ij	0.1440	0.1949	0.1694	0.1949	0.1440	0.1864	0.1864	0.1864
	VEL.med	INT.alt	INT.baj	INT.med	AFE.alt	AFE.baj	AGR.alt	AGR.baj
	0.1779	0.1864	0.1610	0.1610	0.1525	0.1271	0.1440	0.1440

Tabla 6-47.: Inercia por Modalidad con 0:3 NA por fila

	TAM	PES	INT	VEL	AGR	AFE
I_q	0.5084	0.5254	0.5508	0.5084	0.2796	0.2881

Tabla 6-48.: Inercia por Pregunta con 0:3 NA por fila

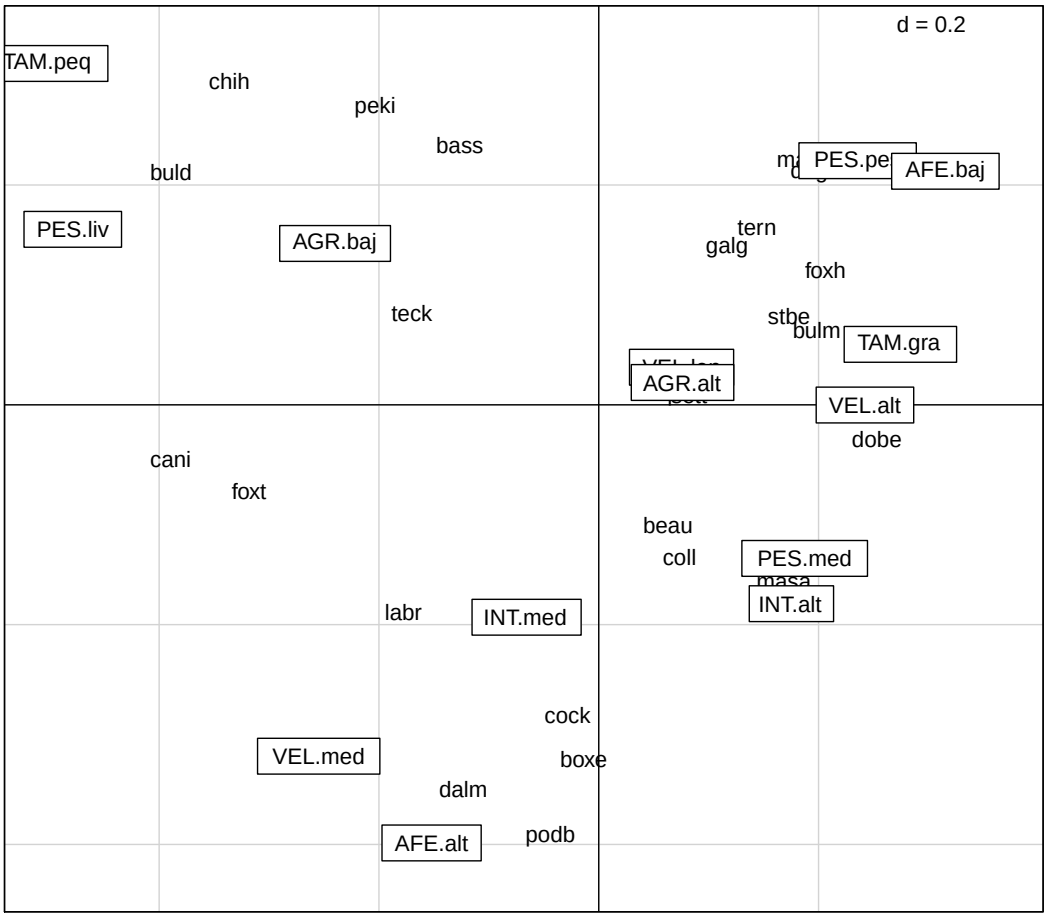

Figura 6-8.: Representación simultánea con 0:3 NA por fila

6.8. Comparación de ACMpdd y ACM-EM con la base BreedsDogs

En la Tabla **6-49** se tiene una comparación de los anteriores escenarios con el método de imputación propuesto por Josse et al. (2012). Se observa que la Inercia Total y el número de ejes con el método ACMpdd se incrementa al comparar con datos completos y con el método de imputación (ACM-EM) se observa que la Inercia Total es la misma que en datos completos, esto sucede por que la imputación lo que hace es recuperar las propiedades de datos completos, de tal forma, que las expresiones de Inercia son las mismas. Por otro lado, se observa que el poder descriptivo en el método ACMpdd disminuye mientras que con el método de imputación el poder descriptivo aunmenta, de hecho, se aunmenta en función del número de NAs. Entonces, se observa que esto puede ser una desventaja de los métodos de imputación, debido a que al aumentarse el poder descriptivo, es como si tener más NAs fuera una situación favorable, lo que se espera es que al tener más NAs en una matriz de datos la representación que se realice pierda poder descriptivo.

	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

Tabla 6-49.: Resultados de Inercia y Poder descriptivo

También se realizo una comparación de las componentes con datos completos y con datos faltantes, para ello se utilizo correlación de pearson ($cor(\psi_1, \psi_1 na)$), las

correlaciones se observan en las tablas **6-50**, **6-51**, **6-52** y **6-53**; las primeras dos tablas son la comparación con ACMpdd y las otras son con ACM-EM.

Para el caso de ACMpdd se observó que la correlación con la primera componente (ψ_1 y/o φ_1) es alta, principalmente en los casos donde hay menos faltantes. En la componente dos ψ_2 y φ_2, se observó que las correlaciones también son altas. Sin embargo, en el escenario 3 NA por fila, se observó que la correlación es muy baja. Para la tercer componentes las correlaciones ya son bajas, lo cuál indica que método ACMpdd generá unas componentes en el eje 3 diferentes a las componentes del eje 3 en datos completos, es decir, que las primeras componentes dos si son relacionadas con las componentes de datos completos.

Al analizar las correlaciones con ACM-EM, se tienen correlaciones más altas tanto para las 3 primeras componentes. Sin embargo, se observa al igual que en ACMpdd que las correlaciones con el eje 3 ya son más bajas al comparar con datos completos. Con base en éste criterio de correlación entre componentes faltantes y datos completos, el método ACM-EM presentó mejores resultados.

	φ_1 Completo		φ_2 Completo		φ_3 Completo
φ_1 1 NA	0.9662	φ_2 1 NA	0.9587	φ_3 1 NA	0.5658
φ_1 2 NA	0.8421	φ_2 2 NA	0.8665	φ_3 2 NA	0.2624
φ_1 3 NA	0.8190	φ_2 3 NA	0.3201	φ_3 3 NA	0.3701
φ_1 0:1 NA	0.9864	φ_2 0:1 NA	0.9887	φ_3 0:1 NA	0.8704
φ_1 0:2 NA	0.9808	φ_2 0:2 NA	0.9325	φ_3 0:2 NA	0.8582
φ_1 0:3 NA	0.9080	φ_2 0:3 NA	0.7829	φ_3 0:3 NA	0.7492

Tabla 6-50.: Correlación entre coordenadas en R^n con datos completos y faltantes (ACMpdd)

	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 6-51.: Correlación entre coordenadas en R^p con datos completos y faltantes (ACMpdd)

	φ_1 Completo		φ_2 Completo		φ_3 Completo
φ_1 1 NA	0.9860	φ_2 1 NA	0.9472	φ_3 1 NA	0.9021
φ_1 2 NA	0.9092	φ_2 2 NA	0.9090	φ_3 2 NA	0.5739
φ_1 3 NA	0.8904	φ_2 3 NA	0.6119	φ_3 3 NA	0.6864
φ_1 0:1 NA	0.9675	φ_2 0:1 NA	0.9455	φ_3 0:1 NA	0.9032
φ_1 0:2 NA	0.9922	φ_2 0:2 NA	0.9462	φ_3 0:2 NA	0.9122
φ_1 0:3 NA	0.9176	φ_2 0:3 NA	0.8596	φ_3 0:3 NA	0.8835

Tabla 6-52.: Correlación entre coordenadas en R^n con datos completos y faltantes (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 6-53.: Correlación entre coordenadas en R^p con datos completos y faltantes (ACM-EM)

Para estudiar de manera más general y aleatoria la presencia de NAs se realizo diferentes escenarios simulación tal como se muestra en la Tabla **5-4**, tal que los registros por individuo tendrán a lo más 3 Nas, para no exceder el 50 %. En la Figura **6-9** se tiene los planos factoriales en los primeros dos ejes, se observa como a medida que aumentan los datos faltantes se pierden algunas tipologías o características que se encontraban en datos completos; esto se analiza con el método ACMpdd. Ahora bien, en la Figura **6-10** se realiza la comparación de datos completos y matrices con datos faltantes donde se utilizó un ACM-EM, lo que se observa es que también se pierden algunas tipologías al aumentar los datos faltantes. Este tipo de análisis con los planos factoriales se dificulta, debido a que es un análisis visual y se debe tener un indicador de como son las características de variables e individuos en esos dos ejes, el indicador que se propone es el poder descriptivo $(\lambda_1 + \lambda_2)/\sum \lambda$, con dicho indicador se encuentra el porcentaje de varianza explicado en esos dos ejes.

Entonces, en la Figura **6-11**, se observa como a medida que se aumenta la cantidad de datos faltantes, el poder descriptivo disminuye; esto con el método ACMpdd. Se recomienda que máximo exista un porcentaje aleatorio de datos faltantes que no supere el 30 % del total de la base de datos. Es importante mencionar que la linea de referencia corresponde el poder descriptivo con datos completos, el cuál es del 0.5198. Por otro lado, en la Figura **6-12** se observa el poder descriptivo para un conjunto de matrices con datos faltantes, en particular, se observa que a medida que aumenta el porcentaje de datos faltantes, el poder descriptivo aumenta, es decir, que a mayor cantidad de datos faltantes la matriz imputada generará una matriz con pocas diferencias en las frecuencias de la modalidades (matriz homogenea), lo cuál hace que sea más facil explicar las relaciones entre las modalidades e individuos en el subespacio de 2 dimensiones. La anterior situación se considera que no es cohorente, debido al que tener mayor cantidad de datos faltantes, debería ser más dificil poder explicar las relaciones presentes en el conjunto de datos.

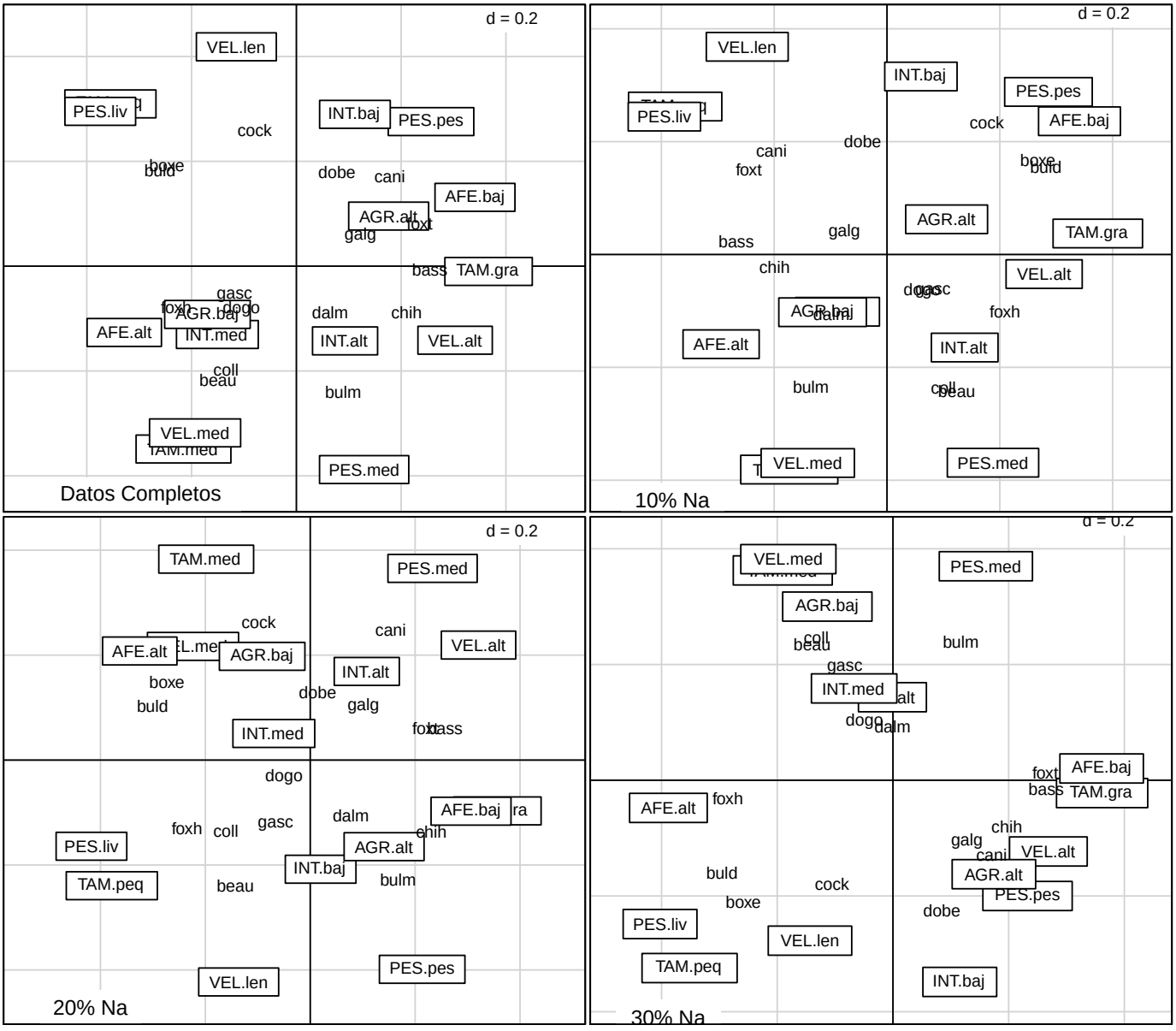

Figura 6-9.: Comparación Plano Factoriales: Datos Completos, ACMpdd 10 % , 20 % y 30 % de NAs

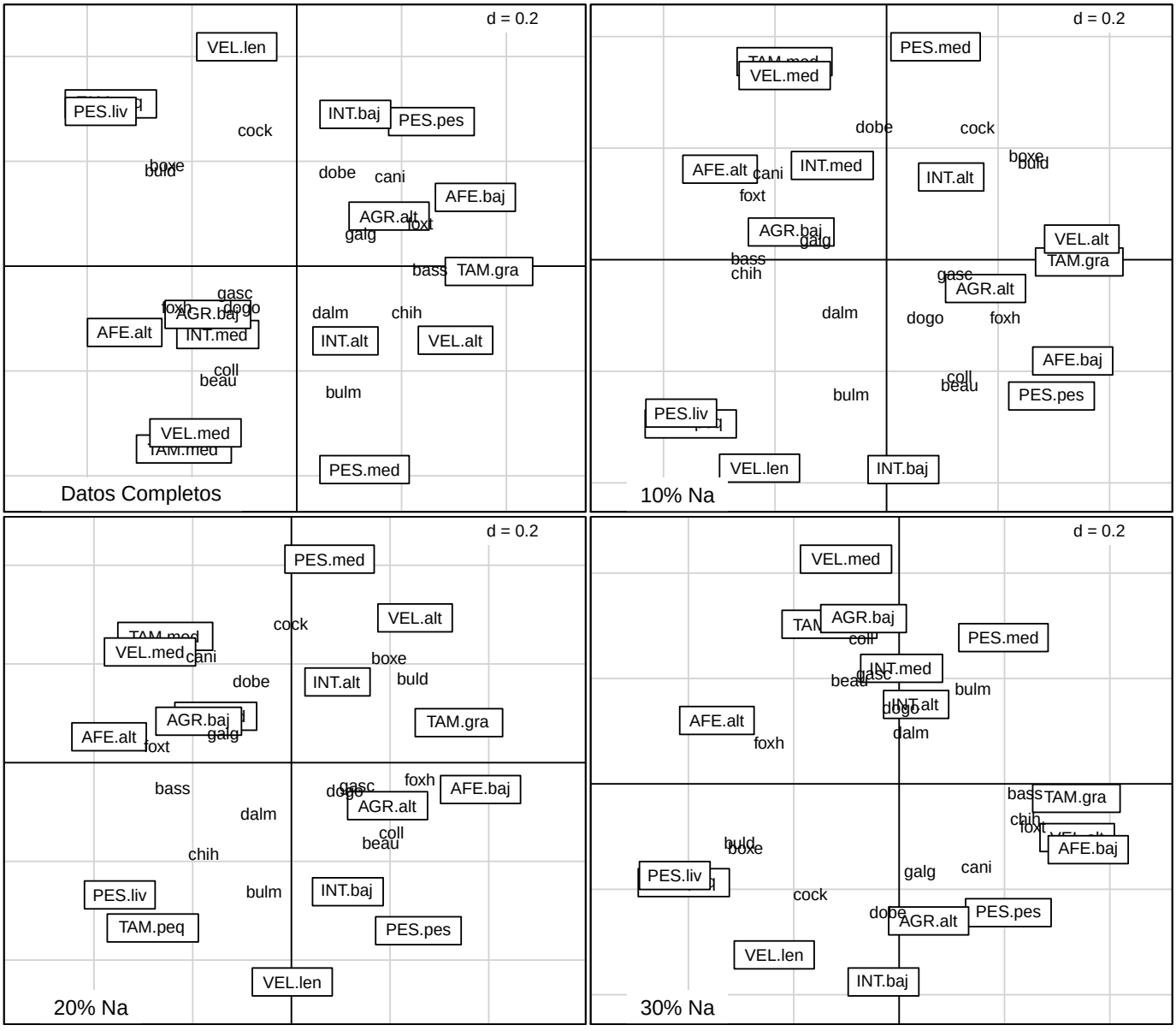

Figura 6-10.: Comparación Plano Factoriales: Datos Completos, ACM-EM 10 % , 20 % y 30 % de NAs

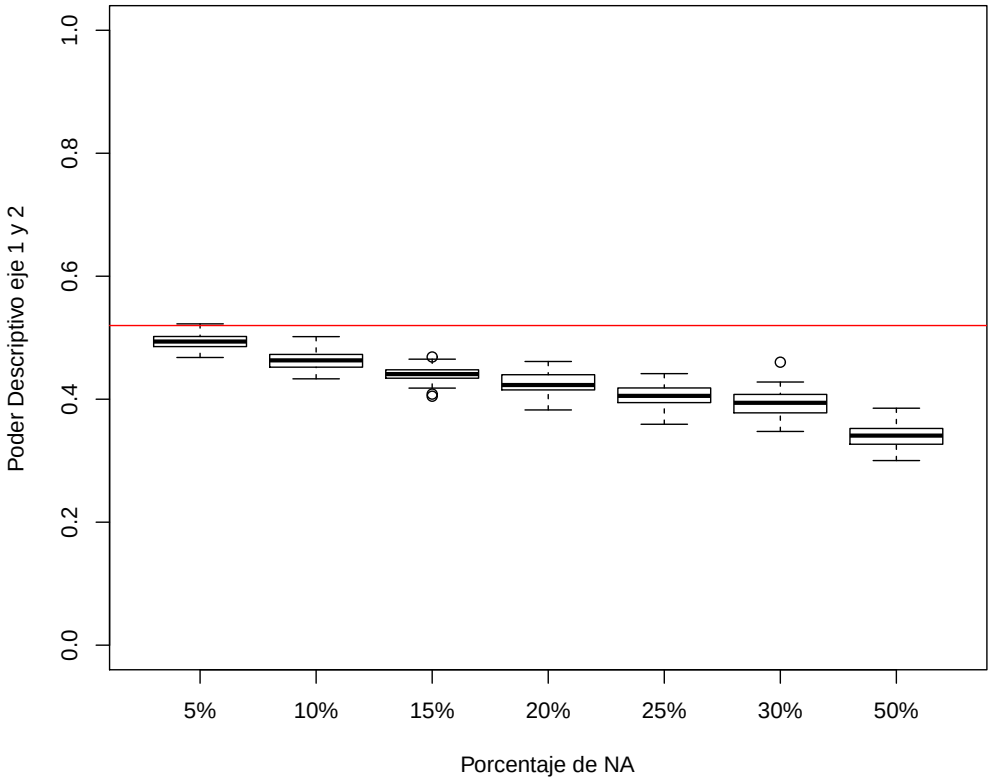

Figura 6-11.: Análisis del Poder Descriptivo según el Porcentaje de NA, ACMpdd (data BreedDogs)

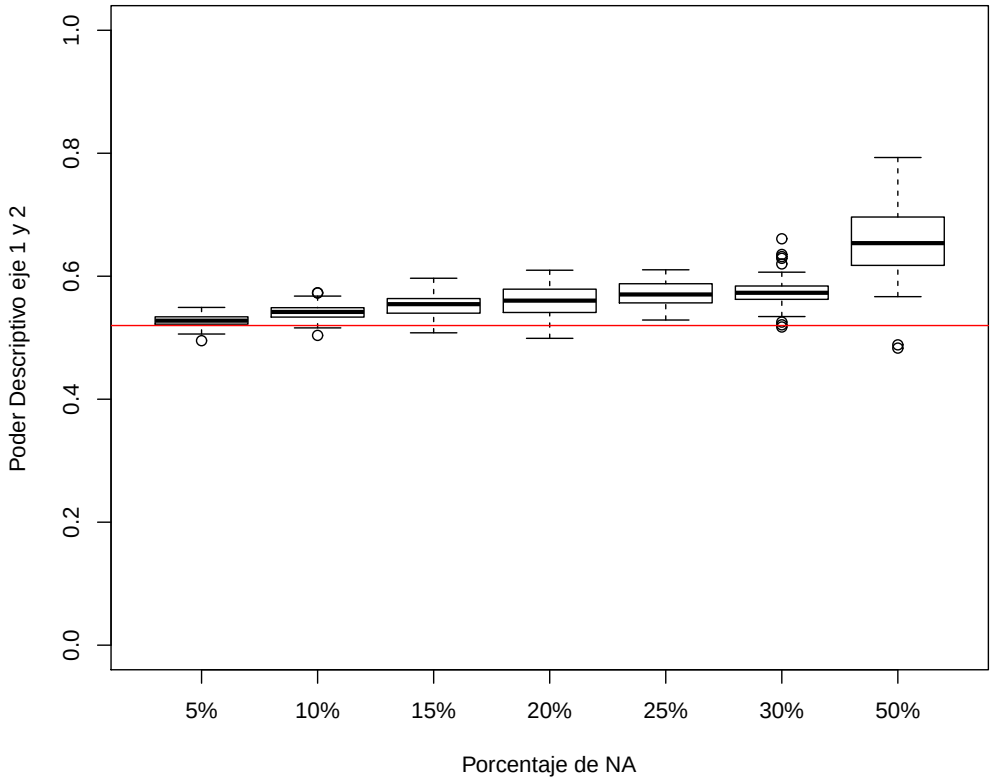

Figura 6-12.: Análisis del Poder Descriptivo según el Porcentaje de NA, ACM-EM (data BreedDogs)

6.9. Análisis de la base de datos tea

En esta sección se mostrará los resultados de la base de datos tea, en el caso con datos completos y datos faltantes. Con esta base de datos se quiere mostrar como el método ACMpdd trabaja para bases de datos con mayor dimensión. Dicha base de datos esta compuesta por 300 individuos a los cuales se les realizaron preguntas respecto al consumo de té, donde se realizaron 11 preguntas y se obtuvieron un total de 41 modalidades. La idea con el ACM es estudiar las relaciones con las modalidades y que individuos se caracterizan por medio de dichas relaciones. A continuación se muestran las estadísticas descriptivas de la base de datos.

En las Tablas **6-54** y **6-55**, se encuentran las frecuencias absolutas para cada una de las preguntas de la encuesta tea, donde se observa que hay mayor número de mujeres según la pregunta P1. También se observa que en la pregunta P2 (Ocupación), las modalidades con mayor frecuencia son EM (Empleado), ES (Estudiante), y NT (No trabaja). Además, en P2 se observa que las modalidades O (Operario) y OT (Otro trabajo) tienen baja frecuencia en la base de datos. Ahora bien, en la pregunta P3 se observa que la mayoría de personas si realizan deporte. Para la pregunta P4 se observa que hay más personas de edades entre los 15 y 24 años, y que la frecuencia más baja en la base de datos es para las personas mayores de 60 años. Por otro lado, en P5 se tiene los conteos respecto al tipo de té que bebén las personas, donde se observa que hay más personas que bebén té con sabor (S) y son pocas las personas que beben té verde (V).

Al analizar la pregunta P6 se observa que la mayoría de personas tomán el té sin nada añadido (N). En cuanto al tipo de té que compran las personas se observa que hay más personas que comprán té de bolsitas (B). Un dato muy importante es que muchas personas afirman no agregarle azucar al té. Además se encuentra que la mayoría de personas comprán el té en los supermercados (S). Para la pregunta P10 se tiene que muchas personas comprán té de marca (DM) y a veces su decisión es muy variable (V). Y en la pregunta P11 se observa que la frecuencia de consumo por lo regular es 1 vez al día (1D) y más de dos veces al día ($+2D$).

Las frecuencias absolutas en dichas tablas **6-54** y **6-55** son uno de los insumos más importantes en el ACM, debido a que cuando frecuencias bajas las modalidades serán excentricas con alta inercia y cuando las modalidades tengan alta frecuencia se arcercarán más al origen y tendrán poca inercia. A continuación se realizará el ACM, en este caso para la base de datos completa y posteriormente se realizará un ACMpdd con datos faltantes para comparar los resultados.

	P1	P2	P3	P4	P5	P6	P7
1	F:178	AG:35	No:121	15-24:92	N: 74	LE: 63	A : 94
2	M:122	EM:59	Si:179	25-34:69	S:193	LI: 33	B :170
3		ES:70		35-44:40	V: 33	N :195	TS: 36
4		MT:40		45-59:61		O : 9	
5		NT:64		+60 :38			
6		O :12					
7		OT:20					

Tabla 6-54.: Tabla de frecuencias absolutas P1 hasta p7

	P8	P9	P10	P11
1	No:155	A : 78	DM: 95	+2D :127
2	Si:145	S :192	DS: 12	1-2S: 44
3		TE: 30	EB: 7	1D : 95
4			EP: 21	3-6S: 34
5			EX: 53	
6			V :112	

Tabla 6-55.: Tabla de frecuencias absolutas P8 hasta P11

6.9.1. ACM con datos completos (tea)

En esta sección, se realizó un ACM para la base de datos tea completa. Es importante recordar que la base de datos cuenta con 300 individuos y 11 variables cualitativas 11. En primer lugar, se construyó la tabla disjuntiva completa en donde se encontrarón las marginales $Z_{i.}$ y $Z_{.j}$. Cómo la base de datos esta completa, entonces $Z_{i.}$ es igual a 11 para todos los individuos. En la tabla **6-56** se tienen las marginales $Z_{.j}$, donde se observa que hay 41 modalidades y que la suma de $Z_{.j}$ para las modalidades de una pregunta coincide con el número de individuos $n = 300$. Con éstas marginales se construyen las métricas y luego la matriz $S^* = S_0' S_0$, con el fin de realizar la descomposición en valores y vectores propios.

Al realizar el ACM se encontrarón los valores propios tal como se observa en la tabla **6-57**, en donde se muestran 30 valores propios ($p - s = 41 - 11$) y a su vez se observa que van decreciendo eje a eje. Una caracteristica importante del ACM es que la inercia total es la sumatoria de los valores propios, que en este caso es igual 2.727273.

P1.F	P1.M	P2.AG	P2.EM	P2.ES	P2.MT	P2.NT	P2.O
178	122	35	59	70	40	64	12
P2.OT	P3.No	P3.Si	P4.15-24	P4.25-34	P4.35-44	P4.45-59	P4.60
20	121	179	92	69	40	61	38
P5.N	P5.S	P5.V	P6.LE	P6.LI	P6.N	P6.O	P7.A
74	193	33	63	33	195	9	94
P7.B	P7.TS	P8.No	P8.Si	P9.A	P9.S	P9.TE	P10.DM
170	36	155	145	78	192	30	95
P10.DS	P10.EB	P10.EP	P10.EX	P10.V	P11.+2D	P11.1-2S	P11.1D
12	7	21	53	112	127	44	95
P11.3-6S							
34							

Tabla 6-56.: $Z.j$ datos completos (tea)

Además es importante mencionar que la sumatoria de los valores propios es igual $(p/s) - 1 = (41/11) - 1 = 2{,}727273$.

1	2	3	4	5	6	7
0.2384	0.1859	0.1709	0.1517	0.1404	0.1278	0.1136
8	9	10	11	12	13	14
0.1118	0.1062	0.1017	0.1004	0.0972	0.0956	0.0862
15	16	17	18	19	20	21
0.0847	0.0806	0.0787	0.0735	0.0714	0.0672	0.0653
22	23	24	25	26	27	28
0.0623	0.0547	0.0533	0.0495	0.0426	0.0372	0.0288
29	30					
0.0254	0.0242					

Tabla 6-57.: Valores propios en R^p para data tea

Con el ACM también se tienen los planos factoriales tanto para la nube de variables como de individuos. En la figura **6-13** se tiene la nube de variables en los dos primeros ejes, donde se tiene un poder descriptivo del 15.5 %. En el plano se observa que hay muchas modalidades en el origen y algunas en los extremos del plano. En este plano se puede interpretar que las modalidades P6.0, P4.60 y P2.NT presentan una

asociación, es decir que individuos que quedan cerca a esas modalidades tendrán dichas caracteristicas. En particular son, P6.0: personas que toman el té con otros añadidos, P4.60: personas mayores de 60 años y P2.NT: personas que no trabajan (Grupo 1).

También se encontró que las personas con la modalidad P9.TE, comparten la modalidad P7.TS y P10.EX; modalidades que indican que la persona comprá en tiendas especializadas, compra té suelto y té de marca exclusiva respectivamente (Grupo 2). Además se observó otro grupo de modalidades con asociación, las cuales fueron: P10.DS, P4.25-34, P2.0, que indican personas que compran té desconocido, con edades entre 24 y 35 años y con ocupación de operarios (Grupo 3). Otro grupo de modalidades son P4.15-24 y P2.ES, que hacen referencia a personas con edades entre 15 y 24 años, que son estudiantes (Grupo 4). Todos estos grupos de modalidades se analizarán en los planos con datos faltantes identificando si se siguén mantiendo dichas asociaciones.

Al analizar la nube de individuos se tiene presente los grupos de modalidades encontrados anteriormente. En la Figura **6-14** se observa que los individuos 44,22,66 y 148 son personas con modalidades P6.0, P4.60 y P2.NT. También se encontró que los individuos 74, 182, 100, y 195 son personas con modalidades P7.TS y P10.EX. Los individuos 145, 214, 290 y 90 comparten las modalidades P10.DS, P4.25-34, P2.0. Y el individuo 45 tiene las modalidades P4.15-24 y P2.ES. Igualmente se analizará esta información con la base de datos con información faltante.

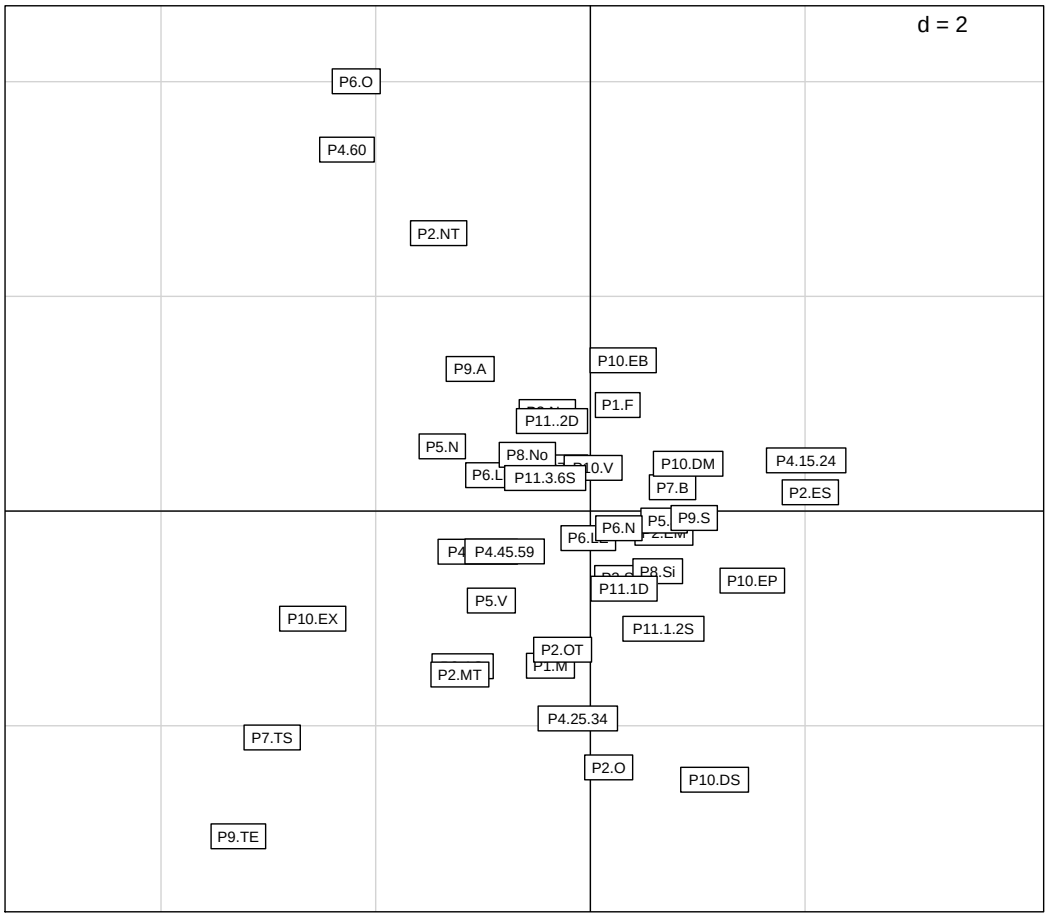

Figura 6-13.: Nube de variables con datos completos (tea)

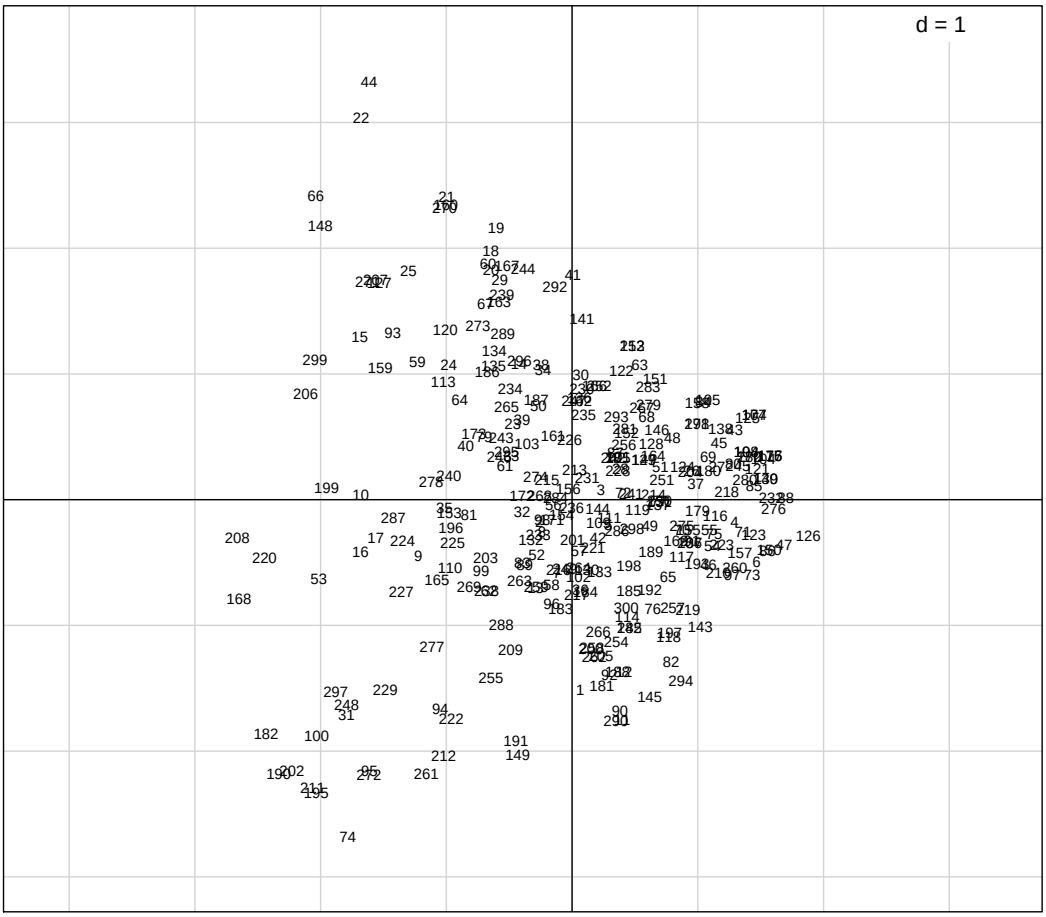

Figura 6-14.: Nube de individuos con datos completos (tea)

6.9.2. ACMpdd con el 10 % de datos faltantes (tea)

En esta sección se realizó un ACMpdd con un 10 % de datos faltantes con la base de datos tea, con el fin de comparar los resultados con la sección anterior, la cual fue un ACM con la base de datos completa. En las tablas **6-58** y **6-59** se tienen las frecuencias absolutas ésta vez con información faltante para cada pregunta de la encuesta, donde se el número de NA's y los conteos por cada una de sus modalidades.

En la tabla **6-60**, se tiene la marginal $z_{.j}$ ante la presencia de datos faltantes, una característica importante es que la suma de las marginales por cada una de las preguntas será diferente al tamaño de muestra 300, ésta característica se da por los datos faltantes.

Al realizar el ACMpdd se encontrarón los valores propios tal como se observa en la tabla **6-61**, en donde se muestran 40 valores propios ($p - 1 = 41 - 1 = 40$) y a su vez se observa que van decreciendo eje a eje. Una caracteristica importante del ACMpdd es que la inercia total es la sumatoria de los valores propios, que en este caso es igual 3.146693. Además es importante mencionar que la sumatoria de los valores propios es aproximado a $\sum \lambda = (p/s^*) - 1 = (40/9{,}896667) - 1 = 3{,}041765$.

	P1	P2	P3	P4	P5	P6	P7
1	F :162	ES :66	No :110	15-24:82	N : 67	LE : 61	A : 81
2	M :114	NT :56	Si :163	25-34:60	S :170	LI : 27	B :154
3	NA's: 24	EM :53	NA's: 27	35-44:36	V : 29	N :167	TS : 35
4		MT :34		45-59:56	NA's: 34	O : 9	NA's: 30
5		AG :33		60 :33		NA's: 36	
6		O:11		NA's :33			
7		OT:17					
8		NA's :30					

Tabla 6-58.: Tabla de frecuencias P1 a P7 con NAs

Con base en los planos factoriales encontrados en la sección 6.9.1, se realiza una comparación de como cambiaron los planos ahora con datos faltantes. De esta forma, en la Figura **6-15** se observa la nube de variables, donde al comparar con la Figura **6-13** con datos completos, se observa que las asociaciones descritas en los grupos

	P8	P9	P10	P11
1	No :136	A : 70	DM : 87	+2D :116
2	Si :131	S :169	DS : 11	1-2S: 40
3	NA's: 33	TE : 29	EB : 7	1D : 86
4		NA's: 32	EP : 18	3-6S: 34
5			EX : 42	NA's: 24
6			V :107	
7			NA's: 28	

Tabla 6-59.: Tabla de frecuencias P8 a P11 con NAs

1, 2, 3 y 4 se modifican un poco, por ejemplo: en el grupo 1 donde anteriormente estaban asociadas las modalidades P6.0, P4.60 y P2.NT, ahora se observa que ya no hay mucha relación con la modalidad P6.0, de hecho hay una modalidad, la cual es P9.A. Para el caso del grupo 2 se siguen manteniendo las mismas asociaciones. Ahora bien, en el grupo 3 se observa que las asociaciones se mantienen, solo que ahora la modalidad P1.M también presenta asociación. Por último, en el grupo 4 continuan las modalidades P4.15-24 y P2.ES, ahora con los datos faltantes aparece cerca la modalidad P9.S.

Analizando los planos para la nube de individuos con datos completos **6-14** y con datos faltantes **6-16** se observó similitudes con base a la conformación descrita por los grupos 1, 2 , 3 y 4. Sin embargo, parece que hay muchos individuos que ahora se fueron al origen del plano.

En términos generales, se observa que hay modalidades que mantiene las mismas asociaciones descritas en datos completos. Sin embargo, hay que prestar atención en las nuevas asociaciones que se encontrarón, pués hace que algunas interpretaciones cambien.

P1.F	P1.M	P2.AG	P2.EM	P2.ES	P2.MT	P2.NT	P2.O
162	114	33	53	66	34	56	11
P2.OT	P3.No	P3.Si	P4.15-24	P4.25-34	P4.35-44	P4.45-59	P4.60
17	110	163	82	60	36	56	33
P5.N	P5.S	P5.V	P6.LE	P6.LI	P6.N	P6.O	P7.A
67	170	29	61	27	167	9	81
P7.B	P7.TS	P8.No	P8.Si	P9.A	P9.S	P9.TE	P10.DM
154	35	136	131	70	169	29	87
P10.DS	P10.EB	P10.EP	P10.EX	P10.V	P11.+2D	P11.1-2S	P11.1D
11	7	18	42	107	116	40	86
P11.3-6S							
34							

Tabla 6-60.: $Z.j$ datos faltantes (tea)

1	2	3	4	5	6	7	8	9	10
0.2474	0.1931	0.1869	0.1694	0.1577	0.1398	0.1284	0.1238	0.1169	0.1147
11	12	13	14	15	16	17	18	19	20
0.1121	0.1058	0.1024	0.1007	0.0947	0.0914	0.0846	0.0834	0.0814	0.0774
21	22	23	24	25	26	27	28	29	30
0.0729	0.0697	0.0669	0.0625	0.0604	0.0506	0.0475	0.0429	0.0317	0.0296
31	32	33	34	35	36	37	38	39	40
0.0144	0.0128	0.0124	0.0108	0.0104	0.0093	0.0090	0.0084	0.0069	0.0056

Tabla 6-61.: Valores propios en R^p con datos faltantes (tea)

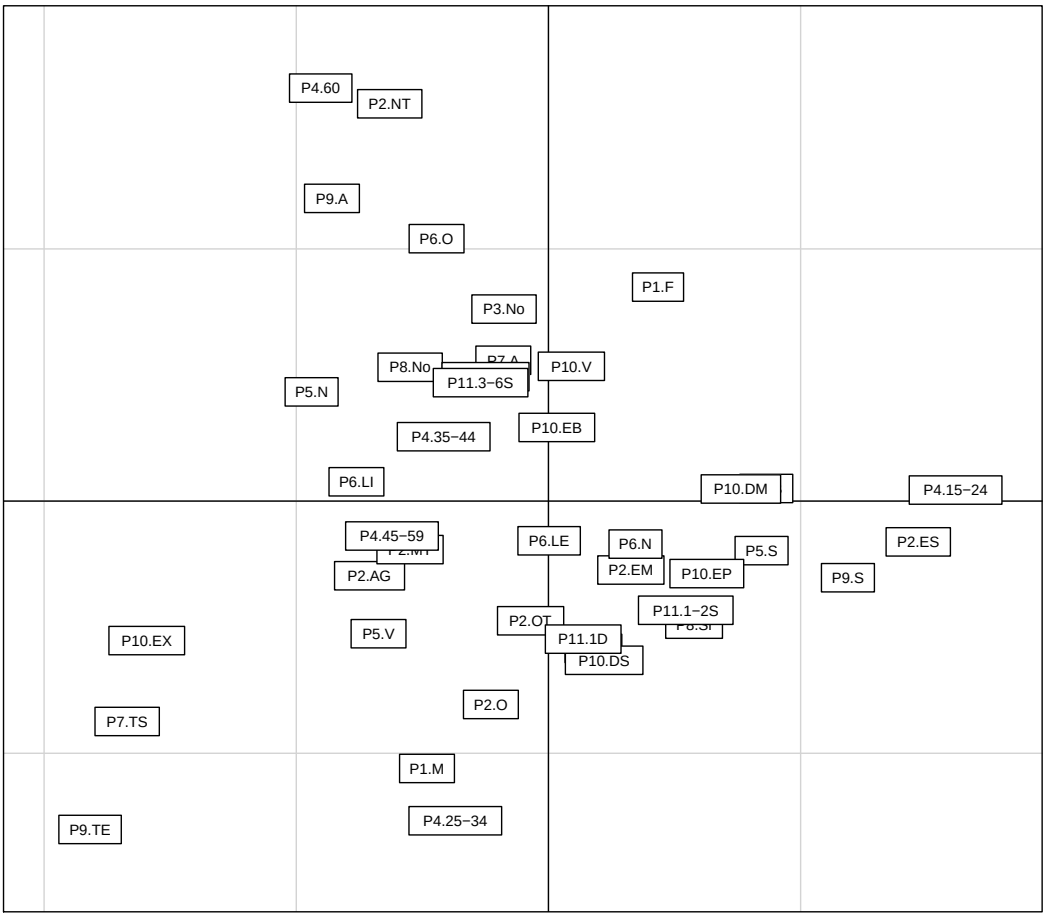

Figura 6-15.: Nube de variables con 10 % NAs (tea)

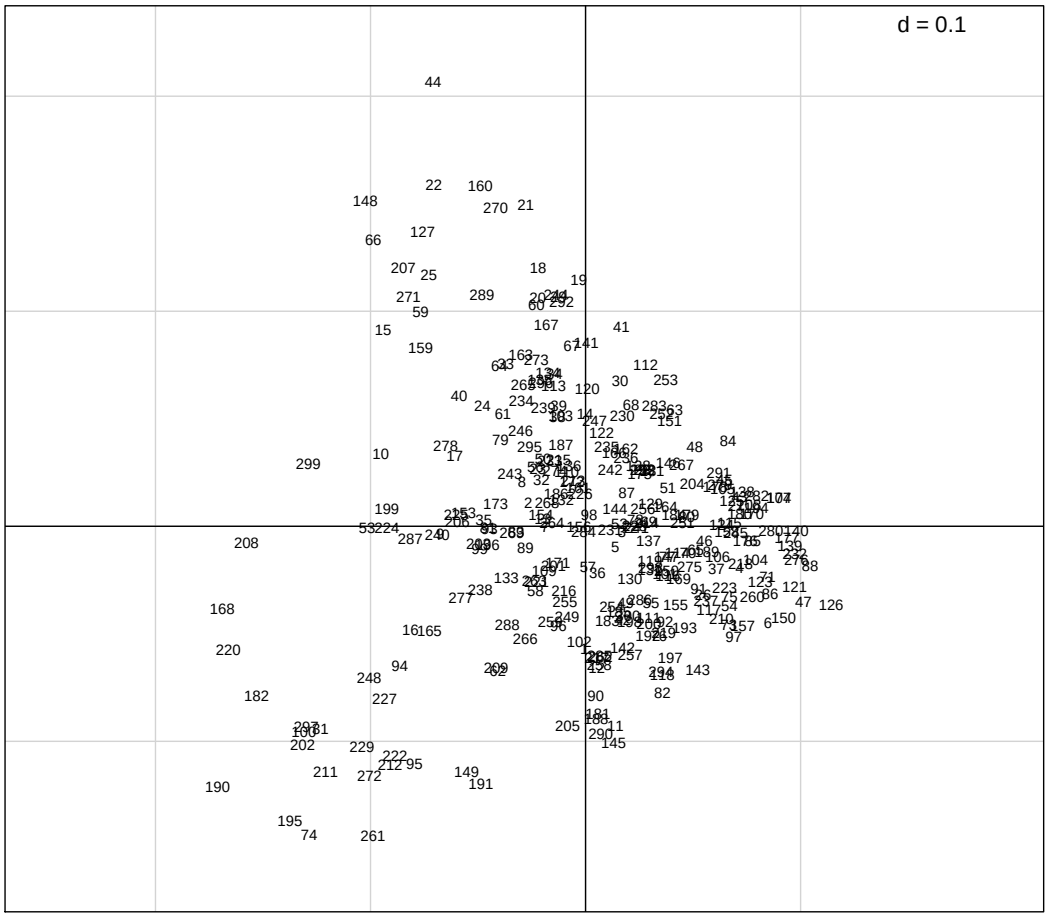

Figura 6-16.: Nube de individuos con 10 % NAs (tea)

6.10. Comparación de ACMpdd y ACM-EM con la base tea

En esta sección al igual que en la sección 6.8, se realizón una comparación del poder descriptivo en los dos primeros ejes, ésto con el fin de analizar si se sigue cumpliendo la característica de que a medida de más datos faltantes el poder descriptivo disminuye en ACMpdd y aumenta en ACM-EM. De esta manera, se trabajó con la base de datos tea y se generarón 1000 matrices con el 5 %, 10 %, 15 % , ..., 50 % y en cado una de esas matrices se realizó un ACMpdd y un ACM-EM. El reporte del poder descriptivo de esta simulación, se observa en las Figuras **6-17** y **6-18**, la primer Figura para ACMpdd y la segunda para ACM-EM.

En general, se mantiene que para ACMpdd el poder descriptivo disminuye cuando hay más datos faltantes y para ACM-EM el poder descriptivo aumenta cuando hay más datos faltantes. Se considera en éste caso un poder descriptivo 15.55 % para datos completos (linea roja).

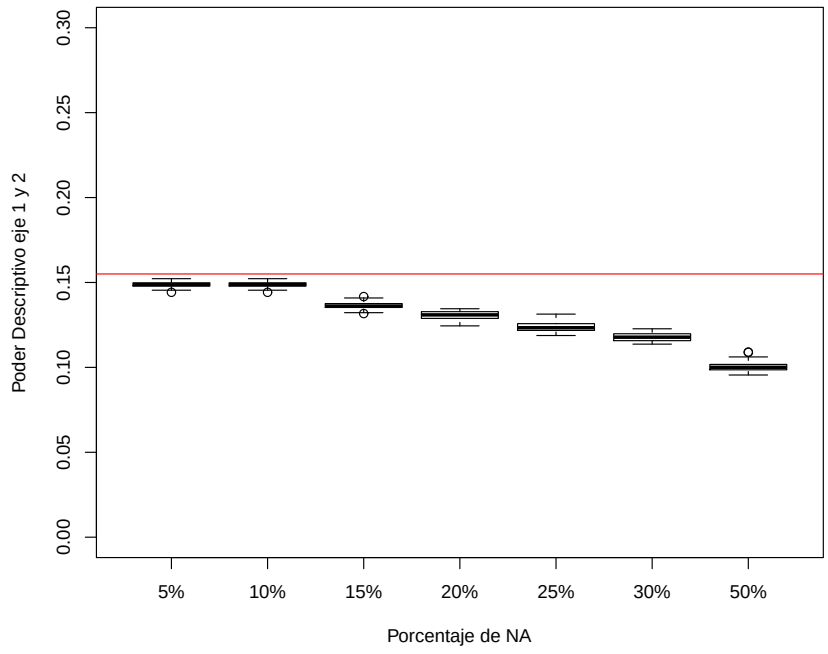

Figura 6-17.: Análisis del Poder Descriptivo según el Porcentaje de NA, ACMpdd (tea)

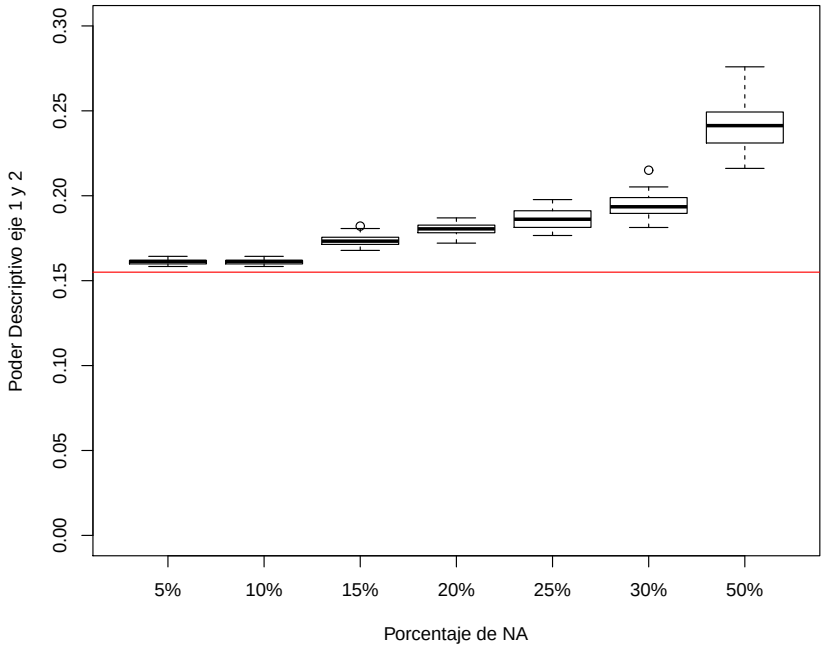

Figura 6-18.: Análisis del Poder Descriptivo según el Porcentaje de NA, ACM-EM (tea)

7. 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.

Se encontró que al usar el prinicipio de datos disponibles de NIPALS fue posible realizar un ACM en presencia de datos faltantes, al cuál se le asignó el nombre ACMpdd. Sin embargo, 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.

Una conclusión importante es que las componentes del método ACMpdd conservan su ortogonalidad (ver Anexos **A-19**), 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 ($\varphi = \frac{1}{\sqrt{\lambda}} M_p F' \psi$).

Con los anteriores resultados, 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. Es importante seguir trabajando en el método propuesto y seguir estudiando sus propiedades geométricas y matemáticas. El método ACMpdd es una solución alternativa al problema de datos faltantes, es un método que no acude a técnicas de imputación, pero el usario puede usar el método para realizar dicha imputación (vía reconstitución de la matriz). En comparación con el método ACM-EM el método ACMpdd presenta más cohorencia en términos del poder descriptivo, debido a que mayor cantidades esperamos que el poder descriptivo se

disminuya. Sin embargo, se considera importante pensar en otras comparaciones con otros métodos como los son el ACM-EM regularizado (Josse et al., 2012) o el ACM con imputación múltiple (Audigier et al., 2015), dichas comparaciones nos pueden dar a luz a cuál método presenta mejores propiedades en diferentes matrices cualitativas con datos faltantes.

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.

¿Que método se recomienda?: ACM-EM Regularizado.

A pesar que el método ACMpdd es facil en su implementación, se observa que presentó mejor aproximación el ACM-EM (en términos de la correlación de las componentes). La comparación con el poder descriptivo deja en duda la implementación del ACM-EM en bases de datos con gran cantidad de datos faltantes, puesto que su poder descriptivo se aumenta considerablemente. Sin embargo, se han encontrado estudios como el de los autores Josse et al. (2012), donde se comparan muchos métodos de datos faltantes y se utiliza el RV de escofier encontrando que el ACM-EM regularizado presenta mejores resultados cuando hay mayor porcentaje de datos faltantes ($> 30\,\%$); ésto por que las coordenadas se aproximan más a la realidad de datos completos. Algo importante a mencionar en el artículo es que no se realizaron comparaciones con NIPALS, por lo cuál surgió el interés en estudiar dicho algoritmo para ACM con datos faltantes.

¿ACMpdd vs los métodos de tablas incompletas estudiados por Van der Heijden and Escofier (2003)?

El ACMpdd presenta resultados equivalentes al método missing passive (ACMmp) propuesto por Benzécri et al. (1973) (ver (Van der Heijden and Escofier, 2003), (Meulman, 1982)), esto se puede observar en la sección A.2. Sin embargo, se debe resaltar que se partieron de enfoques diferentes, el interés de este trabajo fue inplementar NIPALS en el contexto de ACM con datos faltantes. De esta manera,

se utilizó el principio de datos disponibles para conformar las matrices S_o y T_o, cuya descomposición singular conllevan de acuerdo a las relaciones de transición a obtener los valores y vectores propios en la solución.

8. Trabajos Futuros

Un análisis interesante puede ser que pasa al realizar un clúster basado en ACMpdd. ¿La conformación de individuos dentro de los clúster cambia respecto a los clúster originales?. Esto es un tipo de análisis donde se trabaja con un clúster de la base de datos completa y con clúster con matrices de datos faltantes.

En cuanto, a procedimientos necesarios para mejorar el método ACMpdd es importante encontrar la generalización en la Inercia Total para cualquier estructura de datos faltantes. Igualmente es importante adaptar el algoritmo NIPALS con datos faltantes y dentro del algoritmo tener en cuenta las métricas M_n y M_p de tal forma que el algoritmo reconozca que debe hacer un ACM, aunque en esta situación garantizar la ortogonalidad de los factores podría ser complejo computacionalmente (Ver Pseudocódigo ACMpdd 4.3. paso 4 antes del 3).

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. Igualmente estudiar la relación entre el Modelo Multinomial y el ACM como se observa en el artículo (Groenen and Josse, 2016), identificando si es posible trabajar con esta relación y a su vez con datos faltantes.

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. Y como bien sabe la tendencia en estadística es el trabajo con grandes bases de datos, estudiar como se puede implementar el método propuesto usando las herramientas que provee el Big Data.

Bibliografía

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.

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

Cañizares, M., Barroso, I., and Alfonso, K. (2004). Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. *Gaceta Sanitaria*, 18(1):58–63.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dray, S., Dufour, A.-B., et al. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, 22(4):1–20.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Publications.

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

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467.

Groenen, P. J. and Josse, J. (2016). Multinomial multiple correspondence analysis. *arXiv preprint arXiv:1603.03174*.

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.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417.

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

Husson, F., Josse, J., Le, S., Mazet, J., and Husson, M. F. (2017a). Package "factominer". *Multivariate Exploratory Data Analysis and Data Mining*.

Husson, F., Lê, S., and Pagès, J. (2017b). *Exploratory multivariate analysis by example using R*. CRC press.

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., Morineau, A., and Piron, M. (1997). *Statistique exploratoire multidimensionnelle*. Dunod Paris.

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.

Pearson, K. (1901). Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559.

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

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Rubin, D. B. and Schenker, N. (1991). Multiple imputation in health-are databases: An overview and some applications. *Statistics in Medicine*, 10(4):585–598.

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

Sanchez, G. (2012). plsdepot: Partial least squares (pls) data analysis methods. r package version 0.1. 17.

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

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

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.

Vitelleschi, M. S. et al. (2010). Modelos pca a partir de conjuntos de datos con información faltante.¿ se afectan sus propiedades? *SaberEs*, (2).

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, pages 471–494.

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

Wold, H. et al. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, 1:391–420.

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.

A. Anexos

A.1. Bases de datos para la simulación

En el siguiente enlace se pueden encontrar las bases de datos para que los usuarios interesados pueden realizar las pruebas de simulación correspondientes.

https://drive.google.com/drive/folders/1a-aMDJeb4tU_wzGciAnAN4FE1gIZRJHj?usp=shar

A.2. Comparación de ACMpdd vs missing passive (ACMmp) vs missing passive modified margin (ACMmpmm)

En esta sección se realizó una comparación con los métodos missing passive y missing passive modified margin, encontrados en el artículo (Van der Heijden and Escofier, 2003). El método missing passive trabaja los datos faltantes como ceros en la tabla disjuntiva completa lo cuál hace que la marginal por fila no sea constante y sea diferente de s (número de preguntas de la encuesta), éste método fue propuesto por Benzécri et al. (1973) y Meulman (1982) estudió sus propiedades. Missing passive presenta algunas dificultades en sus propiedades, por ende, Escofier (1981) propone trabajar modificando las marginales, de tal forma que se trabaja con la métrica $Mn = 1/n$ para todos los individuos, cómo en el caso de datos completos. En la literatura se menciona que el método missing passive modified margin presenta mejores resultados en cuanto al cumplimiento de las propiedades del ACM.

La comparación propuesta en esta sección se realizó con la base de datos BreedsDogs en el caso 0:1 NA por fila. En primer lugar, se calcularon las inercias y el poder

descriptivo en cada uno de los 4 análisis; esto se puede observar en la Tabla **A-1**, donde se analiza que el método ACMpdd y ACMmm presentan los mismos resultados en cuanto la inercia total y el poder descriptivo, también se observa que el método ACMmpmm es el más cercano al caso de datos completos.

En las Figuras **A-1** y **A-2** se observan los planos factoriales para la nube de variables e individuos respectivamente. En general, se observa que hay algunas asociaciones entre modalidades que se conservan con datos faltantes y lo mismo se dice en el caso de la de individuos. Practicamente con los métodos ACMpdd, ACMmp y ACMmpmm se obtienen buenas aproximaciones a los planos factoriales con datos completos. Algo que se debe destacar es que el método ACMpdd y ACMmp presenta resultados iguales, esto se debe a la característica de marginales imcompletas o no constantes. Algo importante a mencionar es que los planos se realizaron con los dos primeros ejes.

Análisis	Inercia Total	Poder descritivo (Eje 1 y Eje 2)
ACM con datos completos	1.667	0.5198
ACMpdd con 0:2 NA por fila	2.065	0.4375
ACMmp con 0:2 NA por fila	2.065	0.4375
ACMmpmm con 0:2 NA por fila	1.776	0.4662

Tabla A-1.: Comparación ACMpdd vs ACMmp vs ACMmpmm

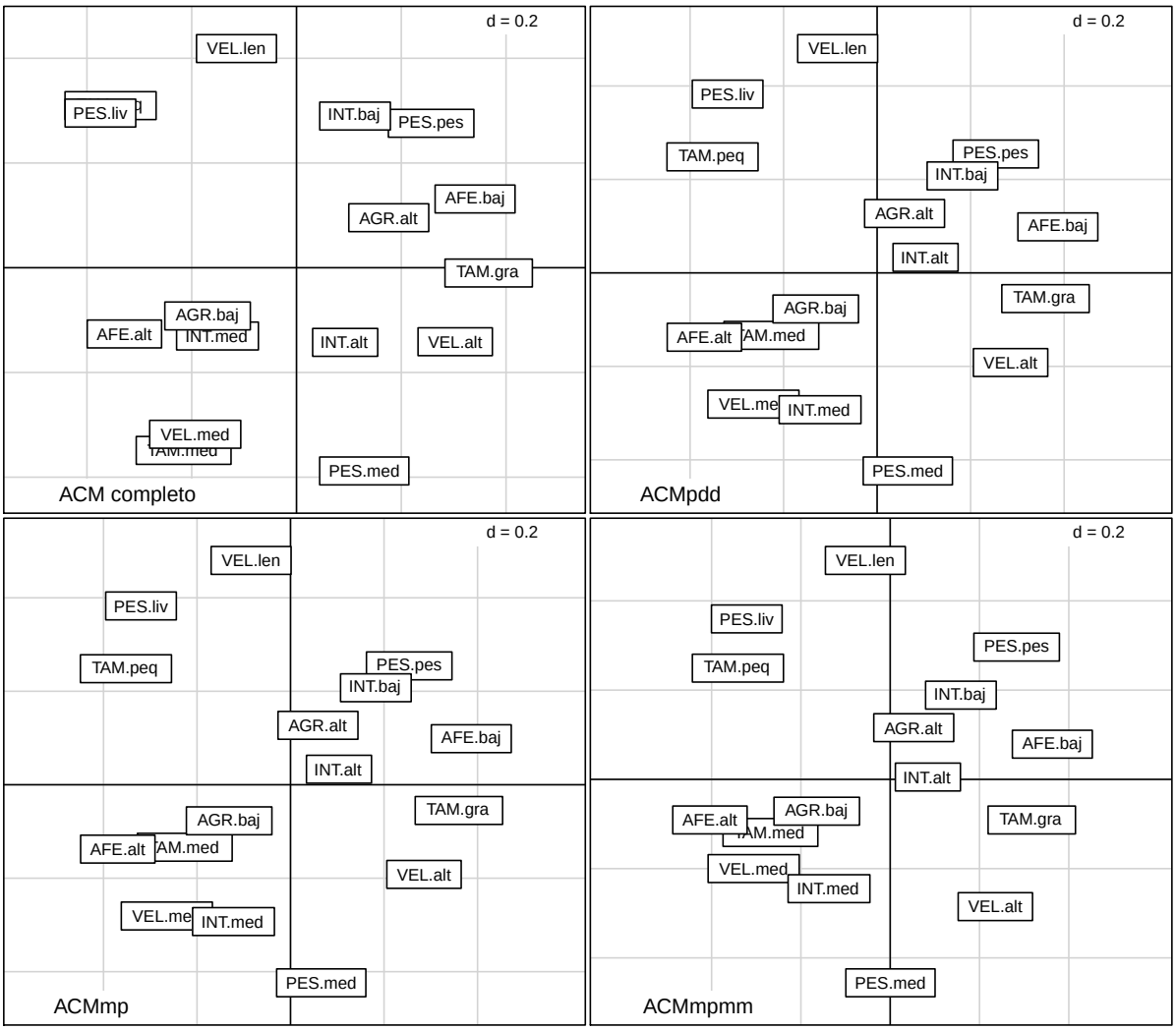

Figura A-1.: Nube de individuos comparación ACMpdd vs ACMmp vs ACMmpmm

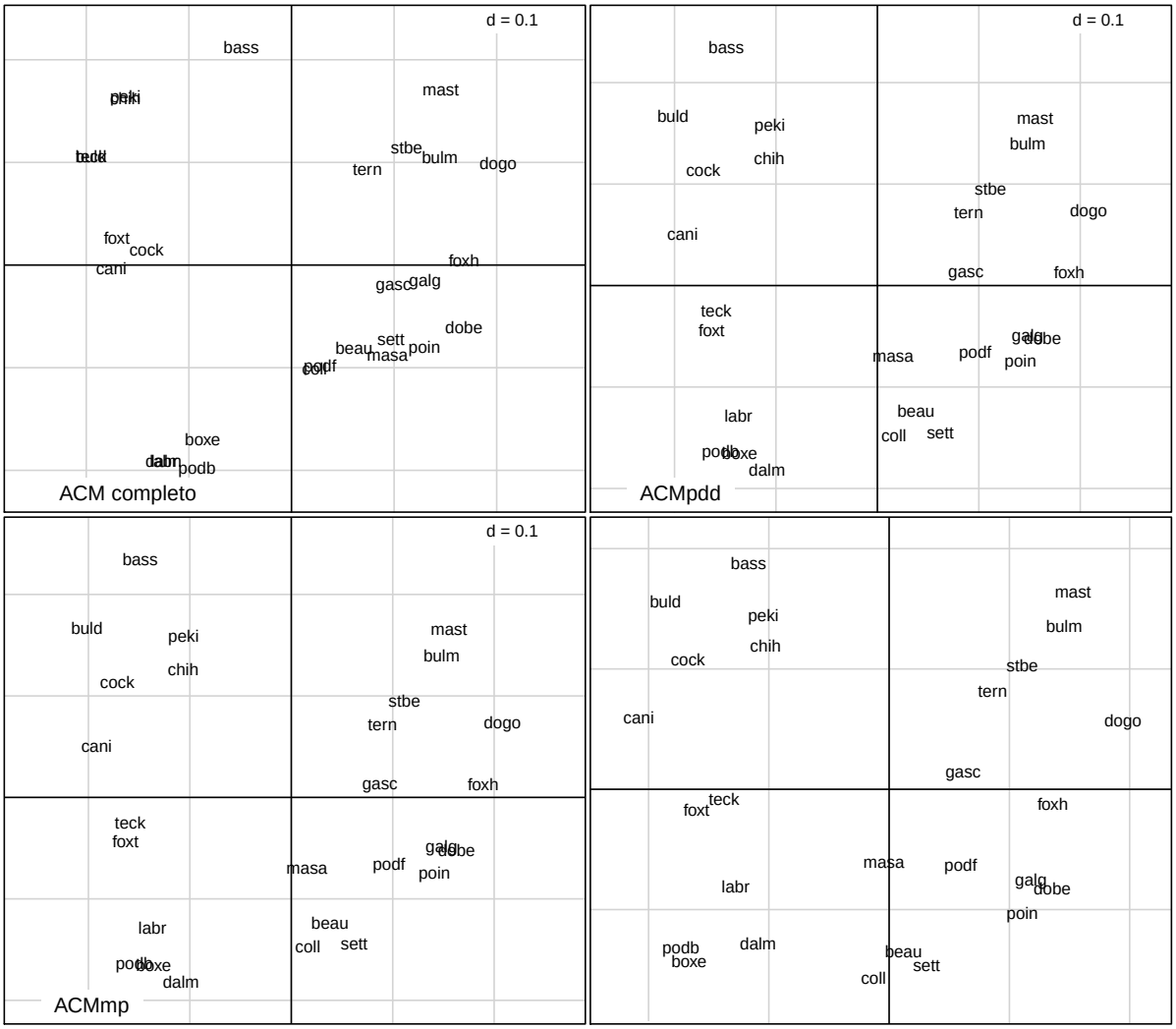

Figura A-2.: Nube de individuos comparación ACMpdd vs ACMmp vs ACMmpmm

A.3. Código ACMpdd en R, usando $NIPALS(S_o^*)$

```
ACMNipals <-function(dat.act) {

 n <-nrow(dat.act)
 s <-ncol(dat.act) # n° variables

## Funcion TDC_NA ## Tabla disjuntiva completa

 TDC_NA <-function (df)
 {
    acm.util.df <- function(i) {
        cl <- df[, i]
        cha <- names(df)[i]
        n <- length(cl)
        cl <- as.factor(cl)
        x <- matrix(0, n, length(levels(cl)))

  pos<-(1:n) + n * (unclass(cl) - 1)

  nas <- is.na(pos)
  for(cols in 1:length(nas)){
   if (nas[cols]){
x[cols,] <- NA
      }else{
x[pos] <- 1
}
  }
        dimnames(x) <- list(row.names(df), paste(cha, levels(cl),
            sep = "."))
        return(x)
    }
    G <- lapply(1:ncol(df), acm.util.df)
    G <- data.frame(G, check.names = FALSE)
    return(G)
```

```
}  ## end TDCNA

 TDCna <- TDC_NA(dat.act)  ## Funcn - tdc con NA

 x <-rowSums(TDCna,na.rm = TRUE) # Zi. suma por filas sin NA
 y <-colSums(TDCna,na.rm = TRUE) # Z.j                sin NA

 sumy <-sum(y); sum(x)
 s1 <-mean(x) ; k2 <- n*s1
 Fs <-as.matrix(TDCna/k2)  ## F con NAs

 fi. <-rowSums(Fs,na.rm = TRUE)
 f.j <-colSums(Fs,na.rm = TRUE)
 Mn <-diag(1/fi.)
 Mp <-diag(1/f.j)
 Mn1.2 <-diag(sqrt(1/fi.))
 Mp1.2 <-diag(sqrt(1/f.j))

## Multiplicación para obtener So con datos disponibles

 n <-nrow(Mn)
 p <-ncol(Fs)
 MnF <-matrix(0,n,p) # == Mn1.2F

 for(j in 1:p){
for(i in 1:n){

uu <-na.omit(cbind(Mn1.2[i,],Fs[,j]))
MnF[i,j] <-sum(uu[,1]*uu[,2])
}
 }

  Mn1.2F <- MnF

 So <- Mn1.2F%*%Mp1.2
 rownames(So) <-rownames(Fs)
 colnames(So) <-colnames(Fs)
```

```
### Función NIPALS

fnipalsACM <- function(Xi)
{
  a <- qr(Xi)$rank # rango de Xi
  p <- ncol(Xi); n <- nrow(Xi)
  Xo <- Xi # Xo <- sqrt(n/(n-1))*scale(Xi)

  T <- matrix(1,n,a) # a compals
  P <- matrix(1,p,a)

  for(h in 1:a)
  {
     t1 <- as.matrix(Xo[,1])
     for(e in 1:100)
     {
        P1i <- (t(Xo)%*%t1)/sum(t1^2)
        nP1i <- sqrt(sum(P1i^2))
        P1 <- P1i/nP1i # vect unitario
        t1 <- Xo%*%P1
     }
     T[,h] <- t1
     P[,h] <- P1
     X1 <- Xo - t1%*%t(P1) # deflacta
     Xo <- X1
  }

  L <- diag(t(T)%*%T)/n
  r.nip <- list(T,P,L)
  return(r.nip)

} # End, Nipals datos completos de rango a.

### Descomposición en valores singulares
```

```
NipalsSo <- fnipalsACM(So)  ## Rp

Lp <- (NipalsSo[[3]])*n ; Up <- NipalsSo[[2]]
Tp <- NipalsSo[[1]]

WP <- sum(Lp[2:3])/(sum(Lp)-1)  ## Poder descriptivo eje 1 y 2

results<-list(Lp,Up,TDCna,WP,Tp)
names(results)<-c("Lp","Up","Zij","WP","Tp")
return(results)


## Lp: Valores propios en Rp
## Up: Vectores propios en Rp
## Zij: Tabla disjuntiva completa con NA
## Tp: Coordenadas de los ejes en Rp

     } ## end  ACMNipals
```

A.4. Código ACMpdd en R, usando $eigen(S_o'^* S_o^*)$

```
ACMpdd <-function(dat.act) {

 n <-nrow(dat.act)
 s <-ncol(dat.act) # n° variables

## Funcion TDC_NA ## Tabla disjuntiva completa

 TDC_NA <-function (df)
 {
    acm.util.df <- function(i) {
        cl <- df[, i]
        cha <- names(df)[i]
        n <- length(cl)
        cl <- as.factor(cl)
        x <- matrix(0, n, length(levels(cl)))

  pos<-(1:n) + n * (unclass(cl) - 1)

  nas <- is.na(pos)
  for(cols in 1:length(nas)){
   if (nas[cols]){
x[cols,] <- NA
      }else{
x[pos] <- 1
}
  }
        dimnames(x) <- list(row.names(df), paste(cha, levels(cl),
            sep = "."))
        return(x)
    }
    G <- lapply(1:ncol(df), acm.util.df)
    G <- data.frame(G, check.names = FALSE)
```

```
    return(G)

}  ## end TDCNA

 TDCna <- TDC_NA(dat.act)  ## Funcn - tdc con NA

 x <-rowSums(TDCna,na.rm = TRUE) # Zi. suma por filas sin NA
 y <-colSums(TDCna,na.rm = TRUE) # Z.j                sin NA

 sumy <-sum(y); sum(x)
 s1 <-mean(x) ; k2 <- n*s1
 Fs <-as.matrix(TDCna/k2)  ## F con NAs

 fi. <-rowSums(Fs,na.rm = TRUE)
 f.j <-colSums(Fs,na.rm = TRUE)
 Mn <-diag(1/fi.)
 Mp <-diag(1/f.j)
 Mn1.2 <-diag(sqrt(1/fi.))
 Mp1.2 <-diag(sqrt(1/f.j))

## Multiplicación para obtener So con datos disponibles

 n <-nrow(Mn)
 p <-ncol(Fs)
 MnF <-matrix(0,n,p) # == Mn1.2F

 for(j in 1:p){
for(i in 1:n){

uu <-na.omit(cbind(Mn1.2[i,],Fs[,j]))
MnF[i,j] <-sum(uu[,1]*uu[,2])
}
 }

  Mn1.2F <- MnF

 So <- Mn1.2F%*%Mp1.2
```

```
 rownames(So) <-rownames(Fs)
 colnames(So) <-colnames(Fs)
 StS <-t(So)%*%So  ## en Rp
 SSt <-So%*%t(So)  ## en Rn

eigen(StS); Lp <- eigen(StS)$values; Up <- eigen(StS)$vectors
eigen(SSt); Ln <- eigen(SSt)$values; Un <- eigen(SSt)$vectors

WP <- sum(Lp[2:3])/(sum(Lp)-1)  ## Poder descriptivo eje 1 y 2

Tp<- So%*%Up ; Tn <- t(So)%*%Un  ## Componentes Principales

results<-list(Lp,Up,Ln,Un,TDCna,WP,Tp,Tn)
names(results)<-c("Lp","Up",
"Ln","Un","Zij","WP","Tp","Tn")
return(results)

## Lp: Valores propios en Rp
## Ln: Valores propios en Rn
## Up: Vectores propios en Rp
## Un: Vectores propios en Rn
## Zij: Tabla disjuntiva completa con NA
## Tp: Coordenadas de los ejes en Rp
## Tn: Coordenadas de los ejes en Rn



}  ### end ACMpdd usign eigen
```

A.5. Código missing passive en R

```
library(ade4) ## for zij
ACMmp <-function(dat.act) {
n <-nrow(dat.act)
s <-ncol(dat.act) # n? variables

#### Tabla disjuntiva completa  NA==0
library(ade4)
TDCna <- acm.disjonctif(dat.act) ## tdc con NA==0
x <-rowSums(TDCna) # Zi. suma por filas sin NA
y <-colSums(TDCna) # Z.j sin NA
sumy <-sum(y); sum(x)
s1 <-mean(x) ; k2 <- n*s1
Fs <-as.matrix(TDCna/k2) ## F con NAs
fi. <-rowSums(Fs)
f.j <-colSums(Fs)
Mn <-diag(1/fi.)
Mp <-diag(1/f.j)
Mn1.2 <-diag(sqrt(1/fi.))
Mp1.2 <-diag(sqrt(1/f.j))

## Multiplicación para obtener So con datos disponibles
n <-nrow(Mn)
p <-ncol(Fs)
MnF <- Mn1.2%*%Fs
Mn1.2F <- MnF
So <- Mn1.2F%*%Mp1.2
rownames(So) <-rownames(Fs)
colnames(So) <-colnames(Fs)
StS <-t(So)%*%So ## en Rp
SSt <-So%*%t(So) ## en Rn
eigen(StS); Lp <- eigen(StS)$values; Up <- eigen(StS)$vectors
eigen(SSt); Ln <- eigen(SSt)$values; Un <- eigen(SSt)$vectors
WP <- sum(Lp[2:3])/(sum(Lp)-1) ## Poder descriptivo eje 1 y 2
Tp<- So%*%Up ; Tn <- t(So)%*%Un ## Componentes Principales
```

```
results<-list(Lp,Up,Ln,Un,TDCna,WP,Tp,Tn)
names(results)<-c("Lp","Up",
"Ln","Un","Zij","WP","Tp","Tn")
return(results)

## Lp: Valores propios en Rp
## Ln: Valores propios en Rn
## Up: Vectores propios en Rp
## Un: Vectores propios en Rn
## Zij: Tabla disjuntiva completa con NA
## Tp: Coordenadas de los ejes en Rp
## Tn: Coordenadas de los ejes en Rn

} ### end ACMmp usingn eigen
```

A.6. Código missing passive modified margin en R

```
library(ade4)
ACMmpmm <-function(dat.act) {
n <-nrow(dat.act)
s <-ncol(dat.act) # n? variables

#### Tabla disjuntiva completa  NA==0

TDCna <- acm.disjonctif(dat.act) ## tdc con NA==0
x <-rowSums(TDCna) # Zi. suma por filas sin NA
y <-colSums(TDCna) # Z.j sin NA
p <- ncol(TDCna)
sumy <-sum(y); sum(x)
s1 <-mean(x) ; k2 <- n*s1
Fs <-as.matrix(TDCna/k2) ## F con NAs
fi. <-c(rep(s/k2,n))  ## margin constant
f.j <-colSums(Fs)
Mn <-diag(1/fi.)
Mp <-diag(1/f.j)
Mn1.2 <-diag(sqrt(1/fi.))
Mp1.2 <-diag(sqrt(1/f.j))

## Multiplicación para obtener So con datos disponibles
n <-nrow(Mn)

MnF <- as.matrix(Mn1.2)%*%as.matrix(Fs)
Mn1.2F <- MnF
So <- Mn1.2F%*%as.matrix(Mp1.2)
rownames(So) <-rownames(Fs)
colnames(So) <-colnames(Fs)
StS <-t(So)%*%So ## en Rp
SSt <-So%*%t(So) ## en Rn
eigen(StS); Lp <- eigen(StS)$values; Up <- eigen(StS)$vectors
eigen(SSt); Ln <- eigen(SSt)$values; Un <- eigen(SSt)$vectors
WP <- sum(Lp[2:3])/(sum(Lp)-1) ## Poder descriptivo eje 1 y 2
```

```
Tp<- So%*%Up ; Tn <- t(So)%*%Un ## Componentes Principales
results<-list(Lp,Up,Ln,Un,TDCna,WP,Tp,Tn)
names(results)<-c("Lp","Up",
"Ln","Un","Zij","WP","Tp","Tn")
return(results)


## Lp: Valores propios en Rp
## Ln: Valores propios en Rn
## Up: Vectores propios en Rp
## Un: Vectores propios en Rn
## Zij: Tabla disjuntiva completa con NA
## Tp: Coordenadas de los ejes en Rp
## Tn: Coordenadas de los ejes en Rn

} ### end ACMmpmm usingn eigen
```

A.7. Ejemplo: paquete missMDA (ACM-EM) en R

```
library(missMDA) ## Se carga la librería missMDA

nb = estim_ncpMCA(dat.act,ncp.max=10) ## Realiza una validación cruzada para
## encontrar el número de ejes q (dat.act la matriz perros con NAs)

imputeMCA(dat.act,ncp=nb$ncp)-> impute_perros  ## La función imputeMCA realiza
## la imputación de una matriz cualitativa vía ACM-EM usando q dimensiones

acm_impute<-dudi.acm(impute_perros$completeObs,scannf=FALSE,nf=16)
## Realiza un ACM de la matriz imputada
```

A.8. Componentes en R^p y R^n Casos Datos Completos

	ψ_1	ψ_2	ψ_3	ψ_4
bass	-0.2541	1.1012	-0.1907	0.2926
beau	0.3172	-0.4177	-0.1015	-0.2114
boxe	-0.4474	-0.8818	0.6920	0.2600
buld	-1.0134	0.5499	-0.1634	-0.3499
bulm	0.7526	0.5469	0.4976	0.6552
cani	-0.9123	-0.0162	-0.5766	0.6281
chih	-0.8408	0.8439	-0.4699	-0.0863
cock	-0.7333	0.0791	0.6622	0.1897
coll	0.1173	-0.5261	-0.3349	-0.6578
dalm	-0.6472	-0.9902	0.4586	-0.1863
dobe	0.8732	-0.3155	-0.4523	0.5101
dogo	1.0470	0.5070	0.1650	0.0629
foxh	0.8766	0.0252	-0.3622	-0.0152
foxt	-0.8816	0.1390	0.0535	0.2856
galg	0.6767	-0.0832	-0.5956	-0.4615
gasc	0.5173	-0.1134	0.0440	0.2410
labr	-0.6472	-0.9902	0.4586	-0.1863
masa	0.4864	-0.4644	-0.4981	0.5774
mast	0.7559	0.8876	0.5877	0.1299
peki	-0.8408	0.8439	-0.4699	-0.0863
podb	-0.4780	-1.0369	0.0619	0.6025
podf	0.1449	-0.5158	0.1171	-0.4689
poin	0.6733	-0.4239	-0.6857	0.0638
sett	0.5041	-0.3771	-0.2891	-0.7251
stbe	0.5834	0.5937	0.8942	-0.1337
teck	-1.0134	0.5499	-0.1634	-0.3499
tern	0.3835	0.4853	0.6608	-0.5800

Tabla A-2.: Componentes en R^p Casos Completos

	φ_1	φ_2	φ_3	φ_4
TAM.gra	0.8367	-0.0206	-0.0512	-0.1702
TAM.med	-0.8511	-1.2317	1.0161	0.3425
TAM.peq	-1.1850	0.9239	-0.6160	0.1201
PES.liv	-1.1689	0.8243	-0.3588	0.1649
PES.med	0.3054	-0.8189	-0.2313	-0.1184
PES.pes	1.0151	0.9739	1.2216	0.0676
VEL.alt	0.8921	-0.3718	-0.7631	-0.2398
VEL.len	-0.3199	1.0449	0.4017	-0.0803
VEL.med	-0.6037	-0.8878	0.3563	0.3702
INT.alt	0.3351	-0.4595	-0.5999	1.2752
INT.baj	0.3490	0.8086	-0.3515	0.0242
INT.med	-0.3694	-0.2855	0.4932	-0.6035
AFE.alt	-0.7755	-0.2669	-0.0608	0.0772
AFE.baj	0.8352	0.2875	0.0655	-0.0832
AGR.alt	0.4315	0.2092	0.3335	0.5512
AGR.baj	-0.4007	-0.1943	-0.3097	-0.5118

Tabla A-3.: Componentes en R^n Casos Completos

A.9. Componentes en R^p y R^n 1 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4
bass	-0.0770	-0.2441	0.0380	0.0936
beau	0.1209	0.0675	0.1045	-0.0465
boxe	-0.0600	0.1809	0.1034	0.1255
buld	-0.1812	-0.0873	-0.0578	-0.0665
bulm	0.0859	-0.0994	-0.1501	0.2193
cani	-0.1512	0.0553	0.0498	0.0442
chih	-0.1424	-0.2387	0.1326	-0.0104
cock	-0.1252	0.0347	-0.0594	0.0807
coll	-0.0045	0.0659	-0.0520	-0.1953
dalm	-0.1854	0.1792	-0.0531	-0.0232
dobe	0.1575	0.0456	0.0199	0.1232
dogo	0.1975	-0.0631	0.0155	-0.0230
foxh	0.1933	0.0133	0.1539	-0.0571
foxt	-0.1423	-0.1022	0.0917	0.0340
galg	0.2061	0.0021	0.0805	-0.0921
gasc	0.1128	0.0501	0.1559	0.0420
labr	-0.1854	0.1792	-0.0531	-0.0232
masa	0.0858	0.0640	0.0493	0.0194
mast	0.1870	-0.0981	-0.0507	0.0605
peki	-0.1453	-0.1549	0.1440	-0.0949
podb	-0.1518	0.2038	-0.0056	0.0905
podf	0.0667	0.0817	-0.0326	-0.0099
poin	0.1020	0.0846	-0.0075	-0.0179
sett	0.0708	0.0304	-0.1113	-0.1660
stbe	0.0744	-0.1127	-0.2311	0.0272
teck	-0.1803	-0.1088	-0.0551	-0.0766
tern	0.0714	-0.0289	-0.2197	-0.0573

Tabla A-4.: Componentes en R^p 1 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
ψ_1	0.5180	0	0	0	0	0
ψ_2	0	0.3827	0	0	0	0
ψ_3	0	0	0.2833	0	0	0
ψ_4	0	0	0	0.2164	0	0
ψ_5	0	0	0	0	0.2071	0
ψ_6	0	0	0	0	0	0.1564

Tabla A-5.: Ortogonalidad en las primeras 6 componentes ψ (1 NA por fila)

	φ_1	φ_2	φ_3	φ_4
TAM.gra	-0.2653	-0.0236	0.0272	0.0999
TAM.med	0.1966	-0.2514	0.0252	-0.1075
TAM.peq	0.1910	0.2744	-0.1308	0.0234
PES.liv	0.2130	0.2278	-0.1354	0.0000
PES.med	-0.1976	-0.1344	-0.1702	-0.0228
PES.pes	-0.1712	0.1300	0.2369	-0.0975
VEL.alt	-0.2135	-0.0677	-0.0686	0.1966
VEL.len	0.1282	0.2340	0.1205	-0.0971
VEL.med	0.1530	-0.2116	-0.0755	-0.0951
INT.alt	-0.0325	-0.1044	0.0150	-0.1879
INT.baj	-0.1430	0.1337	-0.1988	0.0636
INT.med	0.1351	-0.0286	0.2093	0.1380
AFE.alt	0.2250	-0.0884	-0.0572	0.0470
AFE.baj	-0.1856	0.0774	0.1750	-0.0549
AGR.alt	-0.1320	0.0315	-0.1135	-0.1863
AGR.baj	0.1763	-0.0912	0.1058	0.1699

Tabla A-6.: Componentes en R^n 1 NA por Fila usando ACMpdd

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
φ_1	0.5180	0	0	0	0	0
φ_2	0	0.3827	0	0	0	0
φ_3	0	0	0.2833	0	0	0
φ_4	0	0	0	0.2164	0	0
φ_5	0	0	0	0	0.2071	0
φ_6	0	0	0	0	0	0.1564

Tabla A-7.: Ortogonalidad en las primeras 6 componentes φ (1 NA por fila)

A.10. Componentes en R^p y R^n 2 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4
bass	0.0880	-0.1524	-0.1850	0.0151
beau	-0.0247	0.1082	0.0237	0.1573
boxe	-0.2100	0.1712	-0.0547	-0.0571
buld	-0.1582	-0.0809	0.0649	0.1527
bulm	0.1490	-0.1525	-0.0502	-0.2367
cani	-0.1696	-0.1907	-0.2168	-0.1972
chih	-0.0770	-0.2805	0.2394	-0.0064
cock	0.0171	-0.0654	-0.0761	0.1953
coll	-0.0273	0.0996	0.1089	0.0887
dalm	-0.2126	0.1626	0.0306	-0.1258
dobe	0.2127	0.0773	-0.0431	0.0076
dogo	0.2486	0.0475	-0.0569	0.0174
foxh	0.1599	0.1340	-0.0203	0.0332
foxt	-0.0894	-0.0398	-0.2846	0.0500
galg	0.1480	0.1431	0.0472	-0.0474
gasc	0.0710	-0.0081	0.1895	-0.0478
labr	-0.1706	0.1431	0.0666	-0.0195
masa	0.0721	0.1538	0.0509	0.0882
mast	0.2122	-0.0862	-0.0481	0.0389
peki	-0.0770	-0.2805	0.2394	-0.0064
podb	-0.2356	0.1166	0.0115	-0.0157
podf	-0.0742	0.0975	-0.0618	-0.0540
poin	0.0193	0.0640	0.0354	-0.2175
sett	0.0265	0.0372	0.0548	0.0210
stbe	0.1253	-0.0722	-0.0081	0.0519
teck	-0.1607	-0.1487	-0.1255	0.1768
tern	0.1373	0.0024	0.0685	-0.0627

Tabla A-8.: Componentes en R^p 2 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
ψ_1	0.5558	0	0	0	0	0
ψ_2	0	0.4842	0	0	0	0
ψ_3	0	0	0.3874	0	0	0
ψ_4	0	0	0	0.3154	0	0
ψ_5	0	0	0	0	0.2902	0
ψ_6	0	0	0	0	0	0.2331

Tabla A-9.: Ortogonalidad en las primeras 6 componentes ψ (2 NA por fila)

	φ_1	φ_2	φ_3	φ_4
TAM.gra	-0.2636	0.1320	-0.0851	0.0891
TAM.med	0.2779	0.2133	-0.0217	-0.0971
TAM.peq	0.1927	-0.3153	-0.0723	0.0476
PES.liv	0.1112	-0.1910	0.3261	0.0199
PES.med	0.0504	0.2749	-0.1008	-0.0541
PES.pes	-0.2616	-0.0839	0.0341	-0.0761
VEL.alt	-0.2356	0.1817	-0.0043	-0.0430
VEL.len	-0.1109	-0.2960	-0.0338	0.0174
VEL.med	0.2714	0.0931	0.1888	-0.1453
INT.alt	0.0005	-0.1158	0.1074	-0.3348
INT.baj	0.0321	-0.2361	-0.3099	-0.0312
INT.med	0.1525	0.0341	0.1018	0.1899
AFE.alt	0.1311	0.0498	-0.0177	0.2838
AFE.baj	-0.1798	0.0113	-0.0057	-0.1275
AGR.alt	-0.2152	-0.0231	0.2168	0.1203
AGR.baj	0.1644	-0.0037	-0.2337	-0.0540

Tabla A-10.: Componentes en R^n 2 NA por Fila usando ACMpdd

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
φ_1	0.5558	0	0	0	0	0
φ_2	0	0.4842	0	0	0	0
φ_3	0	0	0.3874	0	0	0
φ_4	0	0	0	0.3154	0	0
φ_5	0	0	0	0	0.2902	0
φ_6	0	0	0	0	0	0.2331

Tabla A-11.: Ortogonalidad en las primeras 6 componentes φ (2 NA por fila)

A.11. Componentes en R^p y R^n 3 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4
bass	0.2782	-0.0975	0.0663	-0.3404
beau	-0.0239	0.0454	-0.1629	0.0185
boxe	-0.0239	0.0454	-0.1629	0.0185
buld	-0.2060	-0.0144	0.1999	-0.0383
bulm	0.1215	0.0629	-0.0669	0.1452
cani	-0.2230	-0.2689	-0.0258	0.0165
chih	-0.0651	-0.1179	-0.0872	-0.0553
cock	-0.1602	0.2042	0.0062	-0.0283
coll	0.0570	-0.1517	0.0252	0.1895
dalm	-0.1403	0.1532	-0.0423	-0.0906
dobe	0.1649	-0.0343	-0.1179	-0.0357
dogo	0.1611	-0.0672	-0.0527	0.1236
foxh	0.0598	0.0772	-0.1187	-0.0631
foxt	-0.1468	0.1406	0.1391	-0.0180
galg	0.1932	-0.0328	0.0236	-0.0214
gasc	-0.0136	0.2009	-0.0990	-0.1351
labr	-0.1578	0.3033	-0.0128	-0.1335
masa	0.0490	-0.1004	-0.2385	0.1908
mast	0.3143	-0.0997	0.1556	-0.2023
peki	-0.2230	-0.2689	-0.0258	0.0165
podb	-0.0546	0.0812	-0.2322	0.0412
podf	-0.0872	-0.0232	0.0603	-0.0344
poin	0.2189	-0.0448	-0.0138	0.0488
sett	-0.0168	-0.0687	0.1453	-0.0980
stbe	0.1216	0.1560	0.1968	0.2691
teck	-0.2760	-0.2200	0.1432	-0.0155
tern	0.0786	0.1401	0.2979	0.2320

Tabla A-12.: Componentes en R^p 3 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
ψ_1	0.6858	0	0	0	0	0
ψ_2	0	0.5556	0	0	0	0
ψ_3	0	0	0.4818	0	0	0
ψ_4	0	0	0	0.4638	0	0
ψ_5	0	0	0	0	0.4084	0
ψ_6	0	0	0	0	0	0.4059

Tabla A-13.: Ortogonalidad en las primeras 6 componentes ψ (3 NA por fila)

	φ_1	φ_2	φ_3	φ_4
TAM.gra	0.3121	-0.0102	0.1664	-0.2351
TAM.med	-0.1500	0.2633	-0.1146	0.0590
TAM.peq	-0.2906	-0.3389	0.0439	-0.0085
PES.liv	-0.2751	0.0428	0.2031	0.0424
PES.med	0.0805	0.0246	-0.1625	0.0407
PES.pes	0.0987	0.1622	0.2910	-0.3004
VEL.alt	0.1075	-0.1428	-0.1277	-0.2466
VEL.len	0.2921	-0.1080	0.1305	0.3254
VEL.med	-0.1255	0.2940	-0.0740	0.1759
INT.alt	0.1558	-0.0123	-0.2489	-0.1480
INT.baj	0.1311	-0.0015	-0.1084	0.1784
INT.med	-0.1782	0.1847	0.2475	0.0579
AFE.alt	-0.1737	-0.1180	-0.2875	-0.0468
AFE.baj	0.3281	-0.1195	0.0880	0.2247
AGR.alt	-0.0161	0.2554	-0.0789	-0.0620
AGR.baj	-0.2564	-0.3105	0.1279	0.0058

Tabla A-14.: Componentes en R^n 3 NA por Fila usando ACMpdd

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
φ_1	0.6858	0	0	0	0	0
φ_2	0	0.5556	0	0	0	0
φ_3	0	0	0.4818	0	0	0
φ_4	0	0	0	0.4638	0	0
φ_5	0	0	0	0	0.4084	0
φ_6	0	0	0	0	0	0.4059

Tabla A-15.: Ortogonalidad en las primeras 6 componentes φ (3 NA por fila)

A.12. Componentes en R^p y R^n 0:1 Na por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4
bass	0.0327	0.1747	-0.0406	-0.1355
beau	0.0801	-0.0692	-0.0798	0.0055
boxe	-0.0522	-0.1208	0.0531	-0.0990
buld	-0.2158	0.1536	-0.0668	0.0599
bulm	0.1419	0.1398	0.1107	0.0147
cani	-0.1475	-0.0769	-0.1170	0.0288
chih	-0.2158	0.1536	-0.0668	0.0599
cock	-0.1383	0.0211	0.0893	-0.0731
coll	-0.0372	-0.0951	0.0618	0.0922
dalm	-0.1096	-0.1984	0.1455	-0.0589
dobe	0.1657	-0.0446	-0.1403	0.0669
dogo	0.2229	0.0889	0.0004	-0.0364
foxh	0.1647	-0.0036	-0.0955	-0.1161
foxt	-0.1759	0.0639	-0.0956	-0.1066
galg	0.1373	-0.0265	-0.0672	0.0148
gasc	0.1130	-0.0351	-0.0142	-0.1567
labr	-0.1372	-0.1936	0.1165	-0.0518
masa	0.0792	-0.1093	-0.1471	0.0541
mast	0.1623	0.1623	0.0798	-0.0559
peki	-0.1392	0.2053	-0.0723	-0.0182
podb	-0.1167	-0.2213	-0.0218	-0.0245
podf	0.0182	-0.1098	0.0833	0.0220
poin	0.1106	-0.1073	-0.1201	0.1333
sett	0.0910	-0.0379	0.0245	0.1329
stbe	0.1151	0.1092	0.1651	0.0352
teck	-0.2114	0.1219	0.0031	0.0765
tern	0.0281	0.0837	0.1958	0.1416

Tabla A-16.: Componentes en R^p 0:1 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
ψ_1	0.5087	0	0	0	0	0
ψ_2	0	0.4130	0	0	0	0
ψ_3	0	0	0.2562	0	0	0
ψ_4	0	0	0	0.1819	0	0
ψ_5	0	0	0	0	0.1600	0
ψ_6	0	0	0	0	0	0.1223

Tabla A-17.: Ortogonalidad en las primeras 6 componentes ψ (0:1 NA por fila)

	φ_1	φ_2	φ_3	φ_4
TAM.gra	0.2254	-0.0090	0.0365	-0.1299
TAM.med	-0.1466	-0.1941	0.1385	0.1024
TAM.peq	-0.2635	0.2140	-0.1180	-0.0304
PES.liv	-0.2618	0.1980	-0.1170	0.0402
PES.med	0.0421	-0.2451	-0.0180	0.0558
PES.pes	0.1757	0.1720	0.2096	-0.0489
VEL.alt	0.2225	-0.0729	-0.1862	-0.0892
VEL.len	-0.0861	0.2837	0.1028	-0.0354
VEL.med	-0.1324	-0.2074	0.0424	0.1587
INT.alt	0.0239	-0.1622	-0.2059	-0.1147
INT.baj	0.1513	0.1445	-0.0700	0.1917
INT.med	-0.0799	-0.0914	0.2510	-0.0720
AFE.alt	-0.2250	-0.0681	-0.0700	-0.0072
AFE.baj	0.2532	0.0608	-0.0064	0.0233
AGR.alt	0.1501	0.0688	-0.0318	0.1694
AGR.baj	-0.1555	-0.0562	0.0234	-0.1664

Tabla A-18.: Componentes en R^n 0:1 NA por Fila usando ACMpdd

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
φ_1	0.5087	0	0	0	0	0
φ_2	0	0.4130	0	0	0	0
φ_3	0	0	0.2562	0	0	0
φ_4	0	0	0	0.1819	0	0
φ_5	0	0	0	0	0.1600	0
φ_6	0	0	0	0	0	0.1233

Tabla A-19.: Ortogonalidad en las primeras 6 componentes φ (0:1 NA por fila)

A.13. Componentes en R^p y R^n 0:2 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4
bass	-0.1490	0.2346	-0.0049	0.0459
beau	0.0381	-0.1235	0.0115	-0.0097
boxe	-0.1352	-0.1649	0.2026	0.0448
buld	-0.2013	0.1674	-0.1193	0.0192
bulm	0.1485	0.1398	0.0921	0.0270
cani	-0.1919	0.0509	-0.1126	0.0851
chih	-0.1064	0.1259	-0.1150	-0.0682
cock	-0.1715	0.1135	0.1711	0.1757
coll	0.0163	-0.1472	-0.0929	-0.0235
dalm	-0.1087	-0.1817	0.0483	-0.0755
dobe	0.1629	-0.0518	-0.0210	0.1456
dogo	0.2082	0.0721	0.0496	-0.0371
foxh	0.1897	0.0131	-0.0690	0.0138
foxt	-0.1628	-0.0436	0.0529	-0.1271
galg	0.1481	-0.0502	-0.1139	0.0368
gasc	0.0874	0.0110	0.0819	-0.0280
labr	-0.1369	-0.1285	-0.0068	-0.1221
masa	0.0155	-0.0711	0.0530	0.0406
mast	0.1555	0.1651	0.1036	-0.0680
peki	-0.1061	0.1571	-0.0815	0.0371
podb	-0.1547	-0.1652	0.0980	0.1082
podf	0.0965	-0.0668	-0.0702	0.0314
poin	0.1411	-0.0755	-0.1254	0.1318
sett	0.0620	-0.1448	-0.1013	-0.0191
stbe	0.1116	0.0953	0.1285	-0.1220
teck	-0.1587	-0.0247	-0.1147	-0.1352
tern	0.0898	0.0716	0.0241	-0.1359

Tabla A-20.: Componentes en R^p 0:2 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
ψ_1	0.5132	0	0	0	0	0
ψ_2	0	0.3903	0	0	0	0
ψ_3	0	0	0.2512	0	0	0
ψ_4	0	0	0	0.2000	0	0
ψ_5	0	0	0	0	0.1815	0
ψ_6	0	0	0	0	0	0.1506

Tabla A-21.: Ortogonalidad en las primeras 6 componentes ψ (0:2 NA por fila)

	φ_1	φ_2	φ_3	φ_4
TAM.gra	-0.2580	-0.0344	-0.0148	-0.0004
TAM.med	0.1620	-0.0835	0.2339	0.1875
TAM.peq	0.2522	0.1551	-0.1522	-0.0809
PES.liv	0.2291	0.2391	-0.0550	0.1614
PES.med	-0.0471	-0.2648	-0.0257	0.1154
PES.pes	-0.1818	0.1589	0.1450	-0.1372
VEL.alt	-0.2045	-0.1198	-0.1404	0.0774
VEL.len	0.0605	0.2999	0.0499	-0.0193
VEL.med	0.1895	-0.1755	0.1219	-0.0564
INT.alt	-0.0743	0.0208	-0.0680	0.1778
INT.baj	-0.1308	0.1297	-0.0525	-0.0401
INT.med	0.0874	-0.1829	0.0337	-0.2350
AFE.alt	0.2647	-0.0860	-0.0021	-0.0250
AFE.baj	-0.2770	0.0616	0.0140	0.0008
AGR.alt	-0.0421	0.0794	0.2260	0.0287
AGR.baj	0.0941	-0.0482	-0.2271	-0.0247

Tabla A-22.: Componentes en R^n 0:2 NA por Fila usando ACMpdd

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
φ_1	0.5132	0	0	0	0	0
φ_2	0	0.3903	0	0	0	0
φ_3	0	0	0.2512	0	0	0
φ_4	0	0	0	0.2000	0	0
φ_5	0	0	0	0	0.1815	0
φ_6	0	0	0	0	0	0.1506

Tabla A-23.: Ortogonalidad en las primeras 6 componentes φ (0:2 NA por fila)

A.14. Componentes en R^p y R^n 0:3 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4
bass	-0.0946	-0.1568	0.0121	-0.1615
beau	0.0472	0.0731	0.0813	-0.2192
boxe	-0.0099	0.2145	-0.0446	-0.0820
buld	-0.2912	-0.1407	0.0254	0.0764
bulm	0.1487	-0.0454	-0.1988	-0.0252
cani	-0.2916	0.0331	0.0747	-0.0654
chih	-0.2516	-0.1957	0.0619	0.0234
cock	-0.0213	0.1879	-0.2099	-0.0984
coll	0.0550	0.0922	0.1152	0.0204
dalm	-0.0923	0.2328	-0.0045	0.0615
dobe	0.1895	0.0208	0.1596	0.0717
dogo	0.1476	-0.1408	0.0074	-0.1782
foxh	0.1550	-0.0815	0.1776	-0.0165
foxt	-0.2378	0.0523	-0.0086	-0.1081
galg	0.0875	-0.0953	0.1095	0.0154
gasc	0.0647	-0.0238	0.0868	-0.0410
labr	-0.1331	0.1253	0.0385	0.0440
masa	0.1261	0.1076	0.1797	0.0113
mast	0.1380	-0.1486	-0.0453	-0.0977
peki	-0.1524	-0.1799	0.0384	0.0375
podb	-0.0330	0.2611	-0.0007	0.0431
podf	0.0617	-0.0057	0.0133	0.1704
poin	0.0540	-0.0155	0.0106	0.1924
sett	0.0617	-0.0057	0.0133	0.1704
stbe	0.1297	-0.0537	-0.2224	-0.0280
teck	-0.1273	-0.0554	-0.1885	0.1074
tern	0.1078	-0.1072	-0.1760	0.1051

Tabla A-24.: Componentes en R^p 0:3 NA por Fila usando ACMpdd

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
ψ_1	0.5609	0	0	0	0	0
ψ_2	0	0.4442	0	0	0	0
ψ_3	0	0	0.3461	0	0	0
ψ_4	0	0	0	0.2901	0	0
ψ_5	0	0	0	0	0.2597	0
ψ_6	0	0	0	0	0	0.2017

Tabla A-25.: Ortogonalidad en las primeras 6 componentes ψ (0:3 NA por fila)

	φ_1	φ_2	φ_3	φ_4
TAM.gra	-0.1924	-0.0403	-0.0025	-0.0502
TAM.med	0.1890	-0.1540	0.2428	0.1456
TAM.peq	0.2946	0.2305	-0.2023	-0.0440
PES.liv	0.2316	0.1339	-0.1607	-0.0362
PES.med	-0.1154	-0.2030	-0.0435	0.0569
PES.pes	-0.1965	0.1714	0.1543	-0.1455
VEL.alt	-0.1746	-0.0631	0.0110	0.0673
VEL.len	0.0025	0.1230	0.1103	-0.0875
VEL.med	0.1934	-0.1513	0.0580	-0.0544
INT.alt	-0.0864	-0.0940	0.0354	0.2340
INT.baj	-0.0830	0.2131	-0.0720	-0.0136
INT.med	0.1041	-0.1036	0.0895	-0.0759
AFE.alt	0.1790	-0.1876	0.0831	-0.0097
AFE.baj	-0.3006	0.0275	-0.0280	0.0111
AGR.alt	-0.0188	0.1191	0.2449	-0.0202
AGR.baj	0.1063	-0.0084	-0.2229	-0.1058

Tabla A-26.: Componentes en R^n 0:3 NA por Fila usando ACMpdd

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
φ_1	0.5609	0	0	0	0	0
φ_2	0	0.4442	0	0	0	0
φ_3	0	0	0.3461	0	0	0
φ_4	0	0	0	0.2901	0	0
φ_5	0	0	0	0	0.2597	0
φ_6	0	0	0	0	0	0.2017

Tabla A-27.: Ortogonalidad en las primeras 6 componentes φ (0:3 NA por fila)