Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Simulación Estadística
Cali, junio de 2018

TRABAJO 3
RESEÑA

Kevin Steven García[a], Cesar Andres Saavedra[b]

1. Análisis de correspondencias múltiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo NIPALS (ACMpdd)

El objetivo principal de esta tesis de maestría es proponer un método de análisis de correspondencias múltiples para matrices con datos faltantes. Para ello, se propone un método denominado ACMpdd, el cuál trabaja bajo el principio de datos disponibles del algoritmo NIPALS.

Primero, debemos saber que es el análisis de correspondencias múltiples y que es el algoritmo NIPALS, para tener una idea general sobre que tipo de problemas se pueden abordar bajo este método propuesto.
El análisis de correspondencias múltiples ACM, es un método estadístico multivariado (estudia tres o más variables al mismo tiempo) que sirve para analizar variables cualitativas (variables que expresan distintas características, cualidades o modalidades), el problema fundamental que posee este método es que para realizarlo se necesitan las bases de datos totalmente completas, es decir, no admite datos faltantes, lo cuál ocurre con demasiada frecuencia en la práctica, sobre todo en las encuestas; esto hace que este método solo se pueda usar en las condiciones ideales (datos completos). Aunque se han propuesto soluciones a este problema como eliminar los registros donde se produzca el dato faltante o la imputación del dato faltante (básicamente consiste en estimar los datos faltantes a partir de los datos disponibles), estos tienen muchas criticas y muchos problemas, el primero (eliminar registros), ha sido criticado porque significaría eliminar toda una fila o una columna de datos, lo cuál no solo generaría perdida de información sino también perdida de recursos; y el segundo (imputación del dato faltante), al hacer las estimaciones de los datos faltantes a partir de los datos disponibles, 'hala hacia el montón', es decir, no son buenas estimaciones, ya que estas se ven muy afectadas por las categorías con mayor frecuencia (más repeticiones).

La otra solución es la que se trabaja en esta tesis que es utilizar un algoritmo que permita trabajar con la presencia de éste tipo de datos (principio de datos disponibles), ese algoritmo es el algoritmo NIPALS, el cuál realiza una descomposición singular de la matriz de datos, mediante secuencias iterativas de proyecciones ortogonales. Esta solución bajo nuestra perspectiva es la mejor, ya que se trata de 'exprimir' al máximo los datos, dando uso a cada uno de ellos. Al aplicar este algoritmo a matrices de datos completas, el resultado es equivalente al ACM, pero, su mayor virtud es que se puede utilizar con datos faltantes y obtener sus estimaciones a partir de una reconstrucción de la matriz de datos. Además de esto, el algoritmo NIPALS es de vital importancia en la regresión PLS (minimos cuadrados parciales), ya que este sirve para descomponer la matriz y así obtener las componentes ortogonales sobre las cuales se realizara la regresión. Entonces, indirectamente el algoritmo NIPALS es de vital importancia en el proceso de estimar modelos de regresión con presencia de multicolinealidad, matrices sobredefinidas (con más variables que datos), y matrices con datos faltantes.

El método ACMpdd de forma general consta de los siguientes pasos:

1. Construir una tabla disyuntiva completa Z_{ij}^* en la que estén contenidos los datos faltantes, la cuál cual codifica las variables cualitativas como 0 si la modalidad no existe y 1 si la modalidad existe, además los datos faltantes son codificados como NA.
2. A partir de la tabla disyuntiva construir $F^* = \frac{Z^*}{k^*}$ donde $k^* = ns^*$ (s^* disponibles)

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1628466. E-mail: cesar.saavedra@correounivalle.edu.co

3. Se construye una matriz S_0^* utilizando el principio de datos disponibles
4. Se aplica el algoritmo NIPALS a la matriz construida en el item anterior.

El método propuesto se aplico a una base de datos de la librería **FactoClass** del software R denominada **BredsDogs**, que consta de 27 razas de perros y 6 variables cualitativas que son, el tamaño, el peso, la velocidad, inteligencia, afectividad y agresividad, las cuales tienen entre dos y tres modalidades o categorías. Además, también se realizó el procedimiento con la base de datos **Tea** de la libreria **FactoMineR**, la cuál es de mayor dimensión (300 * 11) y se refiere a un cuestionario sobre el té. Donde se le pidió a 300 individuos cómo beben té, cuál es la percepción de su producto y algunos detalles personales. Las dos bases de datos trabajadas estan completas y corresponden a variables cualitativas o categoricas. La idea de trabajar con bases de datos completas y de las librerias de R es que cualquier usuario pueda desarrollar y comprobar el método y que se puedan simular datos faltantes de distintas formas y distintas cantidades para ver la eficiencia del método con respecto a la cantidad de datos faltantes.

En general, el método propuesto ACMpdd se puede usar en cualquier base de datos donde sea de interés hacer un análisis multivariado de variables cualitativas, las bases de datos pueden ser completas o con datos faltantes, ya que se comprobó que los resultados con matrices completas son equivalentes al ACM común, y con datos faltantestambién se obtienen buenos resultados. En particular, este método puede tener gran utilidad en bases de datos correspondientes a encuestas donde se presentan muchos datos faltantes por preguntas no contestadas por los encuestados. También se usa mucho en bases de datos relacionadas al riesgo laboral y a la caracterización de empleados con respecto a sus cualidades y características.

En esta tesis, fue de fundamental y de vital importancia el uso de la simulación, ya que esta permite evaluar el método en distintos escenarios. En particular, se evaluó el método para la base de datos completas y además, se evaluó bajo los escenarios de 1,2 y 3 datos faltantes (NA) fijos por filas y también de forma aleatoria 0 a 1 NA por fila, 0 a 2 NA por fila y 0 a 3 NA por fila. Sin el usó de la simulación, no se podrían sacar conclusiones tan generales del método ya que sería imposible evaluarlo como fue evaluado en esta tesis.

Finalmente, se llegó a la conclusión que el método propuesto para el análisis de correspondencias múltiples con presencia de datos faltantes ACMpdd, es una solución practica para este tipo de problemas, ya que su programación y uso no es complejo. Además el método presento buenas propiedades en cuanto a la construcción de las componentes y de los vectores propios de la matriz, sin embargo, el metodo ACM-EM (bajo el enfoque de la imputación de datos) tuvo mejores aproximaciones, al realizar la comparación de las componentes con datos completos y datos faltantes.

2. Un contraste de hipótesis para datos funcionales