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

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

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)
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

Las variables funcionales se caracterizan por la evolución de una variable a lo largo del tiempo (proceso estocástico), de modo que los valores que toman son funciones en lugar de vectores como en análisis

multivariante clásico. La imposibilidad de medir la mayoría de estas variables continuamente en el tiempo, unida a la complejidad teórica de muchos de los métodos estadísticos disponibles para su análisis, conllevan a que se maneje resúmenes periódicos, los cuales permiten constituir un modelo de series temporales contenidas que describen fenómenos.

Estos problemas han planteado la necesidad de desarrollar metodologías potentes que permitan, en la práctica, la modelación y predicción a partir de datos funcionales, dando lugar a una especialidad estadística muy reciente conocida con el nombre de Análisis de Datos Funcionales (FDA), que en la actualidad está siendo objeto de un estudio intensivo por parte de los investigadores. A partir de lo anterior se llega al objetivo principal de la tesis de maestría en la que se basa la presente reseña, el cual es modelar el comportamiento de datos de dos estaciones de vigilancia de la calidad del aire para contaminación por partículas con un diámetro menor a 2.5 ţm, mediante el uso de datos funcionales y la simulación de estos para realizar un análisis estadístico que permita responder a la pregunta de si son estadísticamente iguales las curvas generadas.

La solución que se plantea en el trabajo de tesis es el de tomar todos los datos obtenidos por medio de las estaciones, partir de datos escalares y generar curvas por medio de análisis numérico, dichas curvas son llamadas datos funcionales. Una particularidad del estudio es que se evalúan las curvas mas no la estadística escalar, lo que permite a su vez obtener mediciones discretas de un fenómeno continuo a partir de técnicas de suavización y de esta manera extender conceptos de la estadística escalar a la funcional. Los datos funcionales sirven para comparar medias de medias funcionales de poblaciones de datos, para el estudio en particular corresponde a dos poblaciones de datos pareadas, generando así mediciones discretas de un fenómeno continuo a partir de técnicas de suavización. La razón por la cual para la modelación se simuló partiendo de escalares para llegar a datos funcionales fue porque no se conocía la distribución de las curvas que generaban los escalares, lo cual dificultaba el análisis estadístico.

Para facilitar el estudio, el autor hizo uso del análisis por medio de curvas de varianzas, puesto que es un método útil para constatar la utilidad de los datos obtenidos, en el caso específico se evaluó la aplicabilidad de los datos obtenidos a partir de dos estaciones meteorológicas de medición de calidad del aire en la ciudad de Cali, Valle del Cauca; a partir de las mediciones dadas, reportadas como medias, se calcula una curva y se tienen los datos funcionales de medición donde cada conjunto de curva tiene una media funcional, mediante la cual se busca responder si las curvas son estadísticamente iguales.

Es a partir de esto nace el problema estadístico en el cual se busca comparar las medias funcionales, donde el objetivo es construir un contraste de hipótesis para dichas pruebas a partir de dos enfoques:

1- Se realiza la integral de la diferencia de las parejas de curvas, se promedian y se incluye el coeficiente longitudinal.

2- Se toma de manera parecida el enfoque 1 pero primero se promedia y luego se integra, contrario al enfoque 1.

Es importante aclarar que el coeficiente longitudinal se encarga de mostrar discrepancias entre las curvas, el cual es el estadístico que interesa en el estudio.

En el enfoque 1 se tienen medias y la integral de las medias, al mirar la distribución de las medias se puede aplicar el Teorema del límite central, por lo cual todos tienen distribución normal y esto facilita el hacer contrastes. Ya que el objetivo del investigador es realizar contrastes de hipótesis; para realizar dicha hipótesis se realizan simulaciones de poblaciones bajo una hipótesis nula, hipótesis que plantea que las medias son iguales, es decir dos poblaciones que tienen medias iguales.

Es importante la simulación puesto que por medio de esta se quiere saber si la prueba es suficientemente buena para detectar diferencias entre las curvas que están simuladas por parejas pero que entregan dos poblaciones diferentes; poblaciones que a medida que se aumenta el tamaño de muestra, converge puesto

que son medias, por lo cual es un método de simulación convergente. Mediante esta simulación y el tamaño de muestra escogido se obtuvo que para el enfoque 1 todo es normal pese a algunas diferencias; sin embargo, en el enfoque 2 la diferencia entre las curvas es mucho mayor. También resaltar que la simulación nos permite conocer la potencia de la prueba para así saber cuándo rechazar la hipótesis planteada al inicio.

En general, el método propuesto se puede aplicar a cualquier conjunto de datos que tenga como estructura datos pareados para comparación de medias o problemas propios, un ejemplo claro es la variación meteorológica, donde se puede evidenciar cómo la calidad de aire se ve afectada por esta, de forma tal que si se contrastan se puedan obtener datos más viables que permitan al investigador obtener resultados más concluyentes para el estudio que se lleve a cabo y entregar estos a otros especialistas del tema para que tomen acciones y decisiones en base a estos.

3. Referencias

1. Análisis de Correspondencias Múltiples en presencia de datos faltantes:el principio de datos disponibles del algoritmo NIPALS (ACMpdd). Andrés Felipe Ochoa. Sustentación tesis de maestría. Universidad del Valle. Junio 2018.

2. Un Contraste de Hipótesis para datos funcionales. Cristhian Leonardo Urbano Leon. Exposición tesis de maestría.Universidad del Valle. Junio 2018.

3. METODOS PLS EN ANALISIS MULTIBLOQUES:TRATAMIENTO DE DATOS FALTANTES Y MULTICOLINEALIDAD. PhD. VICTOR MANUEL GONZALEZ ROJAS. Universidad del Valle. Abril 2016.

4. Principal Component Analysis (PCA) & NIPALS algorithm. Henning Risvik. May 10, 2007.

5. ANÁLISIS DE DATOS FUNCIONALES. IMPLEMENTACIÓN Y APLICACIONES. Valentín Navarro Pérez. Licenciatura en Ciencias y Técnicas Estadísticas. Universitat Politécnica de Catalunya. Julio 2004.