Análisis comparativo de técnicas de Análisis Multivariado para tratamiento de datos mixtos

Guillermo Sabino[1], *Sillvia Boché*[1], *Sergio Bramardi*[1]

[1]agasabino@gmail.com, Departamento de Estadística, FAEA, Universidad del Comahue

Abstract

The development of the analysis of mixed multivariate databases is extensive. The knowledge about its performance under different conditions is extremely important. Three techniques for its treatment taking into account diverse correlation structures between quality and quantity variables were analyzed. The results show a differential performance in relation to the used technique.

Keywords: mixed multivariate databases.

1. Introducción

En el extenso bagaje de técnicas dentro del análisis multivariado se encuentra la riqueza que permite contemplar la naturaleza de las variables, aplicando un tratamiento diferenciado de acuerdo a que sean cualitativas o cuantitativas. Dado que la realidad estudiada se manifiesta de manera compleja en cuanto a la diversificación de las variables, se requiere de distintas técnicas capaces de transformar los datos recogidos a campo, en información útil para el investigador idóneo de la disciplina.

Estudiar en forma comparativa técnicas y estrategias en la caracterización y agrupamiento de individuos caracterizados por variables mixtas (cuantitativas y cualitativas simultáneamente) representa un trabajo útil para el conocimiento de la adecuación de cada una de ellas en diversas situaciones. En este trabajo, las alternativas propuestas para el abordaje multivariado de variables mixtas comprendieron tres estrategias: Análisis de Coordenadas Principales (ACoP) aplicado sobre el coeficiente de similaridad de Gower, Análisis de Procrustes Generalizados (APG) y Análisis Factorial Múltiple (AFM). Los interrogantes a responder son: las técnicas utilizadas, ¿tienen distinta calidad de representación de los individuos cuando la estructura de correlación entre las variables cualitativas y cuali-cuantitativas tiene distinta intensidad? En caso de obtener diferentes performances, ¿cuál recomendaría y por qué?

Con el fin de precisar el comportamiento de las tres estrategias multivariadas ante las diversas estructuras de datos mixtos, se realizaron simulaciones modificando paulatinamente el nivel de relación entre las variables. Para ello se creó un generador de datos multivariados mixtos, ya que no se encuentra disponible en los paquetes estadísticos. La rutina propuesta genera vectores normales multivariados con esperanza especificada para cada combinación de las variables categóricas y matriz de varianza-covarianza común. De esta forma pueden generarse conjuntos de datos que combinan variables categóricas y continuas, respetando las estructuras de correlación de las variables cuantitativas y modificando las relaciones entre las variables cualitativas y cuantitativas-cualitativas.

2. METODOLOGÍA

A continuación se detalla cada una de las estrategias empleadas:

1. Hallar la matriz de distancias a partir del coeficiente general de similaridad de Gower [1]. Luego, aplicar ACoP a la matriz hallada.

2. Para el grupo de variables cualitativas, calcular la matriz de similaridad a partir del coeficiente de similaridad Simple Matching y posteriormente aplicar Análisis de Coordenadas Principales. Análogamente, para el grupo de variables cuantitativas, realizar un Análisis de Componentes Principales a partir de la matriz de correlaciones. Aplicar APG [2] para hallar la configuración consenso, resultante de las dos configuraciones halladas previamente.

3. Realizar un Análisis Factorial de Correspondencias sobre las variables cualitativas y Análisis de Componentes Principales sobre las cuantitativas. Posteriormente aplicar AFM [3].

Se comenzó a trabajar a partir de una base de datos empírica que contaba con 6 variables cuantitativas, y se obtuvo su matriz de varianza-covarianza. Todas las matrices simuladas constaron de 120 individuos caracterizados por esas variables cuantitativas y 3 variables cualitativas, de las cuáles dos son binarias y la restante tiene tres categorías, generando 12 posibles combinaciones en las categorías de las variables cualitativas. Sin alterar esta estructura de correlación entre las variables cuantitativas y teniendo en cuenta que la intensidad en la estructura de correlación entre las variables cualitativas y entre cuali-cuanti podía afectar los resultados posteriores, se contemplaron inicialmente cuatro intensidades diferentes de correlación dentro de las variables cualitativas: baja (todas las variables cualitativas con una correlación no significativa $p > 0.10$); moderada ($0.05 < p < 0.10$), significativa ($0.01 < p < 0.05$) y altamente significativa ($p < 0,01$). Por otra parte, para cada estructura de correlación de las variables cualitativas, se estudiaron tres tipos de relaciones entre las variables cuanti y cualitativas: baja, moderada y alta; cuando una, dos o tres variables cualitativas se relacionaron significativamente con las variables cuantitativas, respectivamente. Para determinar la significancia en la asociación entre las variables cualitativas se utilizó la prueba chi-cuadrado de independencia, mientras que para las variables cualitativas con las cuantitativas, se realizó un análisis de la varianza. Para cuantificar el sentido y grado de asociación se utilizó la matriz de correlación para variables mixtas propuesta por Boché [4], cuyo análisis propone, para el estudio de la relación entre modalidades de diferentes variables categóricas multi-estado, descomponer cada variable cualitativa en tantas variables dicotómicas como modalidades contiene, lo que permite asociarle una variable Bernoulli a cada sub-variable que indica presencia-ausencia de la modalidad.

Para modificar la estructura de correlación interna entre las variables cualitativas, se asignó una cantidad diferenciada de individuos para las doce combinaciones posibles. Mientras que para alterar la estructura de correlación entre las variables cuali y cuantitativas, a cada una de esas combinaciones se modificó el valor promedio del vector normal multivariado base a partir del cual se realizaban las simulaciones.

Se llevó adelante un estudio que permitió responder el primer interrogante planteado anteriormente. Cada estructura de correlación de las variables cualitativas –baja, moderada, significativa y altamente significativa– se combinaron con distintas estructuras de correlación de las variables cuali-cuantitativas (baja, moderada, alta), obteniendo un total de doce

combinaciones de intensidades de relación. Para cada una de ellas, se simularon mil matrices, lo que arrojó un total de 12.000 matrices. A cada matriz posteriormente se aplicaron las tres estrategias de caracterización para variables mixtas. Con el objetivo de evaluar similaridad en la calidad de representación **entre estrategias**, se hallaron las distancias entre los individuos en la caracterización final en el plano principal de cada una de ellas, y se las correlacionó entre sí, logrando un total de 36000 correlaciones. De esta forma se intentó cuantificar si la representación entre las técnicas era similar o diferente.

Se realizó un análisis descriptivo de las correlaciones mediante diagramas de cajas. Todos los análisis multivariados se realizaron con el paquete FactoMineR del lenguaje R [5].

4. Resultados

En la Tabla 1 se muestran las doce estructuras de correlación. Se planea continuar ampliando otras estructuras de correlación, como por ejemplo modificar la correlación entre las variables cuantitativas.

	Cualitativas vs. Cuantitativas		
Cualitativas vs. Cualitativas	baja (una variable cualitativa vs. una cuantitativa)	moderada (dos variables cualitativas vs. cuantitativas)	alta (tres variables cualitativas vs. cuantitativas)
Baja ($p>0,10$)	B1	B2	B3
Moderada ($0,05<p<0,10$)	M1	M2	M3
Significativa ($0,01<p<0,05$)	S1	S2	S3
Altamente Significativa ($p<0,01$)	A1	A2	A3

Tabla 1: Codificación de las intensidades de asociación entre las variables cuali y cuantitativas

En términos generales se puede apreciar que en las tres estrategias de análisis se logra una caracterización de los individuos en el plano principal similar, ya que todas ellas tienen una relación no menor a 0,75. En las correlaciones obtenidas por simulación se pudo valuar que a medida que la relación dentro de las variables cualitativas va aumentando, también lo hace la asociación entre la representación de los individuos para las tres técnicas (Figura 1). También se puede observar que dentro de cada nivel de asociación de las variables cualitativas, a medida que aumenta la relación entre las variables cuali-cuanti, también aumenta la correlación entre las representaciones de las técnicas estudiadas. Cabe destacar la variabilidad en las correlaciones cuando la relación entre las variables cualitativas es alto.

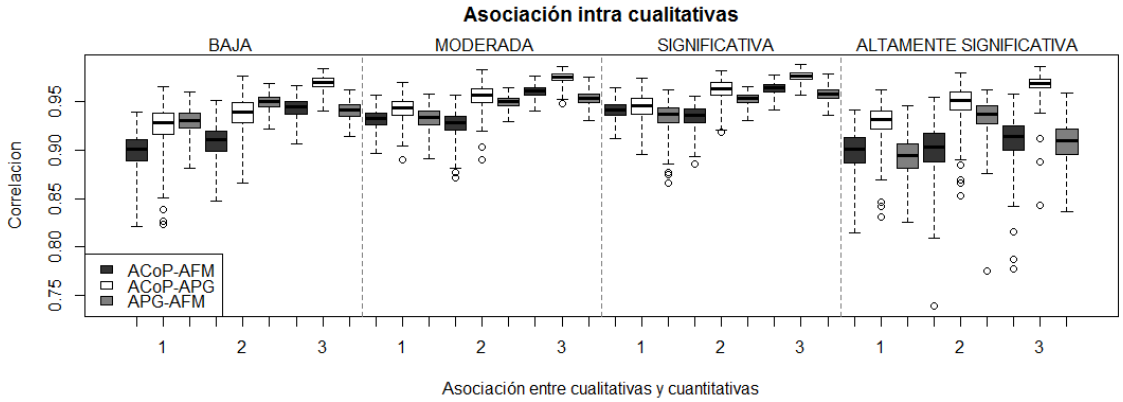

Figura 1: Semejanza en la calidad de representación según estrategias y tipo de asociación

Por otra parte, para determinar si existe un comportamiento similar o diferenciado de las tres alternativas metodológicas de análisis según las intensidades de asociación, la relación de las configuraciones ACoP y APG surgen como las dos técnicas que representan de forma similar a lo largo de todas las condiciones. También podemos afirmar que la relación entre AFM y las otras dos técnicas, depende del grado de asociación de las variables cuali y cuantitativas: cuando la relación es modearada, del AFM resultan configuraciones, en general, más parecida a APG; mientras que cuando la relación es baja o alta, las configuraciones obtenidas por AFM son similares tanto a APG como a ACoP.

La pregunta que surge cuando las correlaciones entre las configuraciones de las tres alternativas de análisis propuestas son bajas, es decir, cuando la calidad de representación de las estrategias no es similar, es, ¿a qué se debe? Seguramente una de las metodologías responde mejor que otra. Determinar cuál de ellas responde mejor a estas estructuras de datos estudiadas es un tema pendiente que se intentará abordar en continuidad con este trabajo. También se planea seguir trabajando modificando no sólo la estructura de las correlaciones de las variables cuantitativas, sino también ampliar a otras estrategias para representación de datos mixtos, en particular la discretización de Escofier.

5. Bibliografía

[1] Gower, J. C. (1966). *Some distance properties of latent root and vector methods in multivariate analysis*. Biometrika 53, 315-328

[2] Gower, J. C. (1975). *Generalized procrustes analysis.* Psychometrika 40, 33-51.

[3] Escofier, B., Pagès, J., 1990. *Analyses factorielles simples et multiples*. Dunod, Paris.

[4] Boché, S.; Bramardi, S.; Lavalle, A.; Reeb, P. (2014). *Variables mixtas, una medida de correlación*. XIX Reunión Científica del Grupo Argentino de Biometría.

[5] Husson, F.; Josse, J.; Le, S.; Mazet, J. (2015). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package version 1.29.
http://CRAN.R-project.org/package=FactoMineR