ANÁLISIS DE CLASES LATENTES COMO TÉCNICA DE IDENTIFICACIÓN DE TIPOLOGÍAS

Daniel Ondé Pérez
Universidad Complutense de Madrid
Facultad de Psicología de la UCM
donde@ucm.es
Jesús María Alvarado Izquierdo
Universidad Complutense de Madrid
Facultad de Psicología de la UCM

Fecha de Recepción: 7 Febrero 2019
Fecha de Admisión: 30 Abril 2019

RESUMEN

En Psicología es frecuente encontrar situaciones en las que se necesita realizar algún tipo de clasificación de personas en subgrupos o clases. Existen técnicas de análisis multivariado como el Análisis Clúster Jerárquico (HCA) que se utilizan habitualmente para este fin. Actualmente, existe un interés creciente por la técnica de Análisis de Clases Latentes (LCA), si bien es una técnica relativamente poco conocida y utilizada. Varios autores han destacado que el LCA presenta importantes ventajas respecto al HCA, en especial que el LCA permite obtener medidas de bondad de ajuste. El objetivo de este trabajo es presentar varias aplicaciones del LCA tanto a partir de un estudio de simulación como a partir de datos reales, y comparar el desempeño de esta técnica frente al HCA. Los resultados a partir de la simulación indican que el LCA tiene una elevada capacidad para detectar estructuras de clase. Los resultados del estudio a partir de datos reales muestran que las distintas clases o mixturas presentes en los datos pueden estar solapadas, lo que dificulta la agrupación de clases al aplicar LCA. El HCA puede ser una buena herramienta de análisis para el investigador aplicado, ya que puede orientar sobre el mejor modelo de LCA que se debería interpretar. En contextos de investigación en los que el modelo teórico no es claro, se recomienda utilizar ambas técnicas con el fin de buscar convergencia de resultados.

Palabras clave análisis de clases latentes; análisis de clúster jerárquico; variable latente categórica; BIC; AIC

ABSTRACT

Latent Class Analysis as a typology identification technique. In Psychology, it is common to find situations in which some kind of classification of people in subgroups or classes is needed. There are multivariate analysis techniques such as Hierarchical Cluster Analysis (HCA) that are

commonly used for this purpose. Currently, there is a growing interest in the technique of Latent Class Analysis (LCA), although it is a relatively little known and used technique. Several authors have pointed out that the LCA has important advantages with respect to HCA, especially that the LCA allows for measures of goodness of fit. The aim of this paper is to present several applications of the LCA both from a simulation study and from real data and compare the performance of this technique against the HCA. The results from the simulation indicate that the LCA has a high performance to detect class structures. The results of the study from real data show that the different classes or mixtures present in the data may be overlapping, which makes grouping classes more difficult when applying LCA. The HCA can be a good analysis tool for the applied researcher since it can guide on the best model of LCA that should be interpreted. In research contexts in which the theoretical model is not clear, it is recommended to use both techniques in order to seek convergence of results.

Keywords: latent class analysis; hierarchical cluster analysis; categorical latent variable; BIC; AIC

INTRODUCCIÓN

En el campo de la Psicología existen numerosas situaciones y contextos de investigación en los que el objetivo es realizar algún tipo de clasificación de personas (selección de personal, diagnóstico y evaluación, tipologías de personalidad, rendimiento académico, etc.). Una característica común en muchas de estas situaciones es que la clasificación debe sustentarse sobre alguna variable latente o constructo no observable. El LCA es una familia de modelos cuyo fin es identificar una variable latente categórica a través de indicadores dicotómicos u ordinales, bajo la asunción de que la muestra completa es una agregación de subgrupos, tipos, mixturas o clases de sujetos. El objetivo es clasificar a las personas en clases utilizando los indicadores o ítems para identificar los patrones de respuesta que mejor definen a las clases (Magidson & Vermunt, 2003; Nylund, Asparouhov, & Muthén, 2007; Vermunt & Magidson, 2002; von Davier, Naemi, & Roberts, 2012). Estos modelos fueron desarrollados inicialmente por Lazarsfeld y Henry (1968) con el fin de identificar una variable latente de actitud medida mediante ítems dicotómicos a través de una encuesta. Actualmente, los avances en los algoritmos (y el software estadístico) utilizados para estimar los modelos LCA permiten analizar datos dicotómicos, ordinales, nominales, conteos y continuos, así como cualquier combinación de ellos.

El interés por los modelos LCA ha crecido en las dos últimas décadas. A pesar de este interés, sigue siendo una técnica de análisis poco utilizada en comparación con otras técnicas tradicionales de segmentación de la muestra como el HCA. El HCA funciona describiendo solamente las relaciones entre variables observables, mientras que el LCA parte de un modelo probabilístico en el que se asume la existencia de una variable latente categórica (la comparación del LCA con el análisis clúster de *K*-medias puede consultarse en Magidson & Vermunt, 2002). Entre las ventajas que tiene el LCA frente al análisis clúster tradicional destacan las siguientes: 1) el LCA se basa en la probabilidad de clasificación a partir de un modelo, en lugar de basarse en algoritmos "ciegos" de agrupación por distancias o similitudes; 2) del modelo de LCA se pueden obtener medidas de ajuste que orientan al investigador sobre cuál es la mejor agrupación, por lo que la elección del número de clases es menos arbitraria que en el HCA; 3) las variables pueden ser continuas, categóricas (nominales u ordinales) o cualquier combinación de estas; y 4) las variables demográficas y otras covariables se pueden utilizar para la descripción del clúster.

El objetivo de este trabajo es mostrar algunas de las ventajas de utilizar LCA respecto a HCA. Para ello, se han realizado dos estudios, el primero procedente de una simulación y el segundo de carácter aplicado. El estudio de simulación (Estudio 1) nos ha servido para mostrar cómo se elabo-

ran modelos LCA mediante el paquete poLCa del entorno gratuito R (Linzer and Lewis 2011; R Development Core Team 2010). Una cuestión importante cuando se elaboran estos modelos es que cada aplicación no parte de las mismas probabilidades condicionales de entrada, por lo que resulta recomendable aplicar el mismo análisis varias veces a partir de un mismo conjunto de datos. En este caso, se ha repetido el análisis 100 veces, y se ha evaluado cuál de los criterios de clasificación es el más fiable o estable a la hora de recuperar la estructura de clases simulada.

El estudio aplicado (Estudio 2) ha consistido en replicar un estudio realizado en el contexto de investigación del tabaquismo (Reyna y Brussino, 2011), el cual consistió en una aplicación del LCA a partir de datos que provienen de una encuesta. Reyna & Brussino (2011) han señalado que la metodología en la que se basa el LCA permite identificar tipologías de uso de sustancias más allá de la clasificación fundamentada exclusivamente en patrones patológicos, por lo que se pueden detectar otros problemas de consumo. En el presente trabajo se muestra el resultado de aplicar HCA sobre los mismos datos, y se compara con la solución obtenida mediante LCA.

BACKGROUND TEÓRICO

El HCA (también denominado como análisis Q, construcción de tipologías, análisis de clasificación y taxonomía numérica), es una técnica estadística multivariante que se engloba dentro de las técnicas de interdependencia. Tiene como objetivo principal agrupar sujetos (u objetos), de tal manera que se obtiene el mayor grado de homogeneidad interna de cada clúster o conglomerado al tiempo que se maximiza el grado de heterogeneidad entre clústers distintos (Hair, Black, Babin, & Anderson, 2010). Para elaborar clústers de sujetos mediante HCA, el investigador debe especificar algún criterio de selección. Existen numerosos criterios de selección disponibles, y varían según el nivel de medida de las variables introducidas en el análisis. Dichos criterios de selección descansan sobre el cálculo de algún tipo de medida de similitud entre sujetos, que puede definirse como una medida de correspondencia o de tendencia. En otras palabras, dos sujetos obtendrán una medida de similitud fuerte si se parecen entre sí (por ejemplo, si comparten un buen número de características sociodemográficas), o sí presentan una tendencia similar de respuesta ante determinados ítems de un test o de un cuestionario. Podemos encontrar medidas de similitud basadas en el cálculo de correlaciones y de distancias cuando las variables analizadas se definen a nivel de intervalo o de razón (como la distancia euclídea o la distancia de Mahalanobis), basadas en el cálculo de alguna medida de asociación entre variables de carácter nominal u ordinal (como Chi-cuadrado o Phi-cuadrado), y basadas en el cálculo de medidas para datos binarios o dicotómicos (como la diferencia de tamaño, la diferencia de configuración, de varianza, de dispersión, D de Anderberg, o Jaccard, entre otras). El cálculo de todas estas medidas de similitud puede consultarse en Sneath y Sokal (1973). Un problema de difícil solución se produce cuando se analizan variables que se encuentran definidas con distintos niveles de medida, situación más o menos frecuente en la práctica, dado que solamente se puede seleccionar una medida de asociación para analizar la información.

Una vez el investigador define el tipo de medida de similitud a utilizar, se calcula una matriz de distancias entre todos los sujetos. Esta matriz de distancias es siempre la misma; esto es, la matriz de distancias es independiente del número de clústers que se decide evaluar. El investigador procede generando nuevas variables con la asignación de los sujetos a los clústers, variando entre una cla-sificación de 2 clústers, 3, 4, etc., a partir de la matriz de distancias, que permanecerá invariante. Finalmente, el investigador tratará de describir cada clúster generado en cada variable nueva en fun-ción de su distribución respecto a otras variables, generalmente las variables introducidas en el HCA para elaborar la clasificación.

Por su parte, los modelos de LCA tienen como objetivo estratificar la información que proviene

de las tablas cruzadas de las variables analizadas (variables observadas o manifiestas), mediante una variable categórica latente o no observada (LV). En función de los valores de esta LV, se asume que las respuestas a todas las variables analizadas son estadísticamente independientes (Linzer & Lewis, 2011; Nylund et al., 2007). Por tanto, frente a los supuestos del HCA de representatividad de la muestra y de ausencia de multicolinealidad, en el LCA se debe asumir independencia condicional o local (esto es, para cada valor o nivel en la LV, los valores de los sujetos en las variables analizadas serán independientes). El modelo LCA agrupa probabilísticamente cada observación en la LV, calculando las probabilidades de respuesta en cada variable observable. Este procedimiento se suele realizar mediante el método de estimación Maxima Verosimilitud (*Maximum Likelihood*, ML) con algoritmo de optimización Esperanza-Maximización (*Expectation Maximization*, EM). A diferencia del HCA, este proceso se realiza para cada intento de clasificación evaluado; esto es, se generarán las probabilidades asociadas a una posible solución de 2 clústers, 3, 4, etc., siendo estas diferentes a priori en cada solución.

El modelo básico del LCA se puede formular mediante la ecuación 1 (ver Vermunt & Magidson, 2002) para un tratamiento en profundidad de los aspectos técnicas de este tipo de modelos).

$$f\left(\mathrm{y}_i|\,\theta\right)=\sum_{k=1}^{K}\pi_k f_k\left(\mathrm{y}_i|\theta_k\right) \qquad 1$$

En donde y_i se refiere a la puntuación de un sujeto en un conjunto de variables observables (las variables analizadas), *K* es el número de clústers o clases latentes, y π_k indica la probabilidad previa o a priori de pertenecer al clúster o clase latente *k*. La distribución de y_i dados los parámetros del modelo θ, $f(\mathrm{y}_i|\theta)$, se asume como una mixtura de distribuciones de densidad de clases específicas, $f(\mathrm{y}_i|\theta_k)$. Dado que la LV es nominal, el modelo de LCA se denomina también como modelo de mixtura finita. Los parámetros estimados por el modelo son la proporción de observaciones en cada clase latente y las probabilidades de que se produzca cada una de las respuestas en la LV, analizando patrones en lugar de respuestas individuales, como sucede en la Teoría de Respuesta al Ítem (TRI). Por tanto, las observaciones con conjuntos similares de respuestas en las variables analizadas tenderán a agruparse dentro de las mismas clases latentes (Linzer & Lewis, 2011).

Aunque el modelo LCA no determina automáticamente el número de clases latentes en un conjunto determinado de datos, ofrece algunas medidas de bondad de ajuste (como el criterio de información de Akaike – AIC y el criterio de información bayesiano – BIC), que el investigador puede usar para realizar una evaluación de carácter teórico y empírico. Aquí reside una de las principales ventajas respecto al HCA, en donde no se dispone de medidas de ajuste para evaluar cada clasificación.

MÉTODO

Para mostrar el funcionamiento de las clases latentes, en el Estudio 1 se han simulado las respuestas de 1.000 sujetos a 10 variables observadas ($Y_1 - Y_{10}$). Las variables observadas son de tipo politómico con tres categorías de respuesta (valores a, b y c). La estructura de clases simulada corresponde a una mixtura de 3 grupos, con una probabilidad asociada de 0,3, 0,4 y 0,3 (30%, 40% y 30% de la muestra, respectivamente). Para la primea clase (C1), las 10 variables tenían una probabilidad de respuesta de 0,6, 0,3, y 0,1 para las opciones de respuesta a, b y c, para la segunda clase (C2), de 0,2, 0,6 y 0,2, y para la tercera clase (C3), de 0,1, 0,3 y 0,6. El método de estimación fue ML con algoritmo de optimización EM, y se utilizaron como medidas de bondad de ajuste AIC y BIC.

Para evaluar el funcionamiento en datos reales del LCA y poder compararlo con HCA (Estudio 2), se tomaron los datos de un estudio de ámbito nacional sobre hábitos relacionados con el tabaco[1], que son los mismos que utilizaron Reyna y Brussino (2011) en su trabajo. El universo comprendía la población española de ambos sexos y mayor de 18 años, y el procedimiento de muestreo se realizó mediante selección aleatoria de teléfonos-hogares. El muestreo fue estratificado por Comunidad Autónoma y tamaño de hábitat (17x7 estratos). El número de entrevistas válidas registradas fue de 2.002 (error muestral ± 2,24% para un nivel de confianza del 95,5% y P = Q).

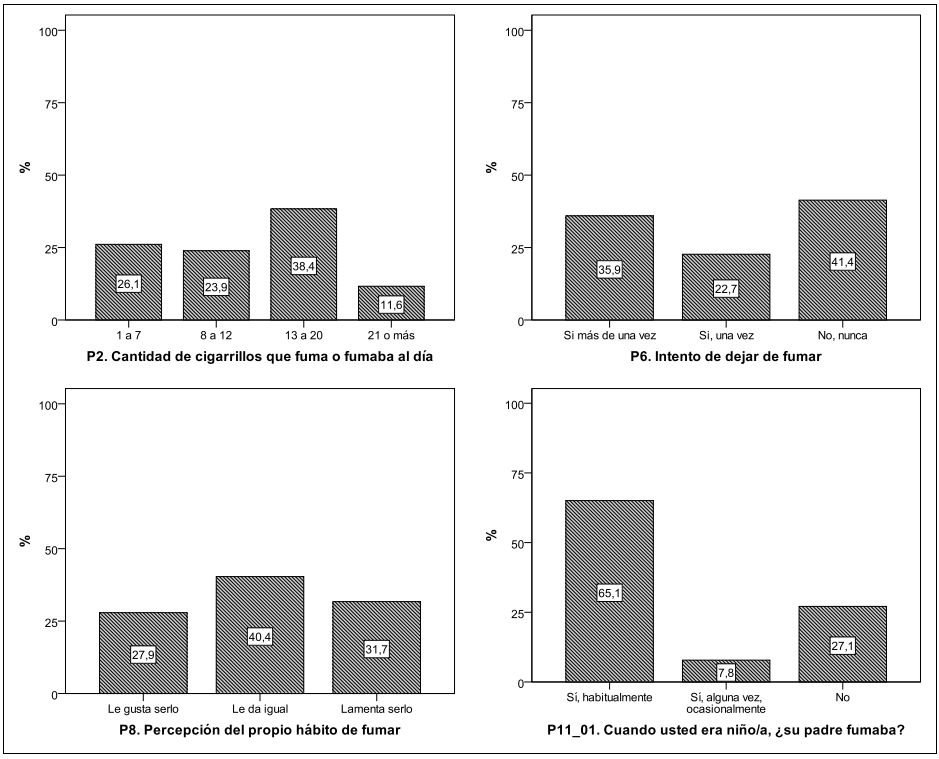
Para el análisis se utilizaron las preguntas P2 (recodificada en 4 categorías), P6, P8, P11_01 y P11_02 del cuestionario, consideradas por varios autores como variables relevantes para el estudio de los hábitos de consumo del tabaco (Chen et al., 2004; Furberg et al., 2005; Poletto, Pezzotto, Morini, & Andrade, 1991). Se seleccionaron previamente a todas aquellas personas entrevistadas que escogieron la categoría de respuesta "Sí, ahora fumo" de P1: "Para empezar, ¿podría decirme si fuma o ha fumado alguna vez en su vida de forma habitual?" (N = 513; 25,6%). No se tuvieron en cuenta las respuestas del tipo "No sabe", "No recuerda" o "No contesta" presentes en las variables analizadas. La muestra finalmente analizada estaba compuesta por 498 personas (N = 498; 24,9%). En la Figura 1 se muestra la correspondiente distribución de respuestas de las variables analizadas (base = 498).

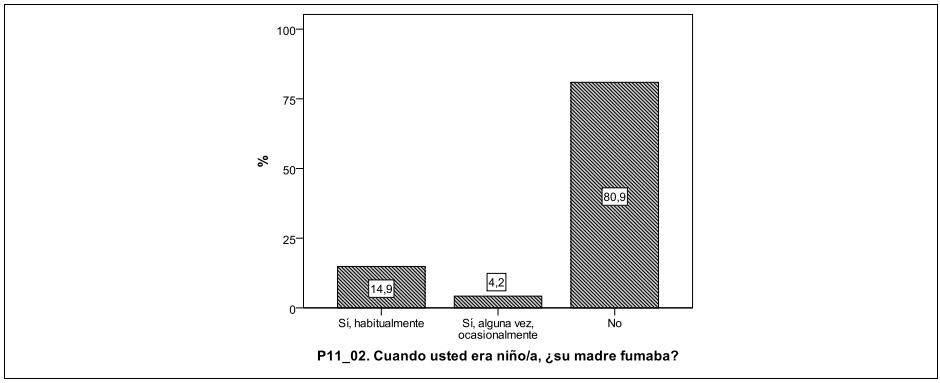
Para elaborar el HCA se ha utilizado el paquete estadístico SPSS (versión 22). Dado que las variables analizadas son de carácter categórico, se han utilizado como medidas de similitud-asociación el valor de Chi-cuadrado y el de Phi-cuadrado a partir de los recuentos. La medida Chi-cuadrado se basa en el test de Chi-cuadrado de igualdad para dos conjuntos o grupos de frecuencias, mientras que la medida Phi-cuadrado equivale a la medida Chi-cuadrado normalizada por la raíz cuadrada de la frecuencia combinada. Existen distintos métodos de conglomeración para establecer las agrupaciones o clústers. Para este trabajo hemos utilizado varios de estos métodos con el fin de comparar distintas soluciones del HCA con la clasificación obtenida mediante LCA. Más concretamente, hemos utilizado el método de agrupación vecino más cercano (VMC), el método vecino más lejano (VML), y el método de vínculos entre grupos (VEG).

El HCA comienza uniendo las dos observaciones más cercanas. Una vez formado el primer grupo, los valores iniciales se sustituyen por algún tipo de observación que caracterice al nuevo grupo. El siguiente paso consiste en recalcular la matriz de distancias para volver a unir las dos observaciones más cercanas. El proceso finaliza cuando todas las observaciones acaban unificadas o integradas en un solo clúster. La diferencia entre los distintos métodos de agrupación estriba en la forma en la que se establece la nueva observación de cada unificación antes de recalcular la matriz de distancias: distancia del miembro del grupo más cercano (VMC), distancia del miembro del grupo más alejado (VML), distancia promedio entre todos los pares de observaciones posibles entre miembros de distintos grupos (VEG). Dado que las variables analizadas son de carácter categórico no ha sido preciso realizar ningún tipo de estandarización. Se han guardado como nuevas variables de clasificación de los datos en 2, 3 y 4 clústers.

Sobre la elaboración de modelos LCA, el programa utilizado por Reyna y Brussino (2011) fue Latent GOLD 4.0 (Vermunt & Magidson, 2005). El método de estimación utilizado fue ML con algoritmo de optimización EM. Para evaluar el ajuste de los modelos, Latent GOLD ofrece el valor de los criterios de información BIC, AIC y AIC3 (este último es una corrección de los grados de libertad utilizados en la fórmula de AIC). En el Estudio 2 hemos replicado estos análisis, pero utilizando el paquete poLCa del entorno R. Se ha utilizado el mismo método de estimación y se ha evaluado el ajuste mediante AIC y BIC. El análisis mediante poLCa se ha repetido 10 veces con el fin de evaluar la estabilidad de la clasificación.

Figura 1 Distribución en porcentaje de las respuestas a P2 (recodificada en 4 categorías), P6, P8, P11_01 y P11_02 utilizadas en el estudio de Reyna y Brussino (2011). Información elaborada a partir de las personas que contestan "Sí, ahora fumo" en P1. Se excluyen respuestas del tipo "No sabe", "No recuerda" o "No contesta" (N = 498; 24,9%). Fuente: Estudio 2751 elaborado por el CIS; febrero de 2008. Elaboración propia.





RESULTADOS

Estudio 1

Se han elaborado 4 modelos LCA sobre el conjunto de datos simulados, asumiendo el modelo nulo como estructura (1 clase), y los modelos de 2 a 4 clases latentes (modelos 1 a 4, respectivamente). En la Tabla 1 se muestran las medidas de ajuste consideradas para evaluar cuál de estos modelos es el que recupera mejor la estructura de clases simulada. Más concretamente, se utilizan los valores de BIC y AIC para evaluar dicho ajuste, que indican un mejor ajuste cuanto menor es su valor. En la Tabla 1 se recoge el ajuste de la mejor solución entre las 100 analizadas en el proceso de simulación. Se observa que el valor de BIC más bajo corresponde al modelo LCA compuesto por una LV con tres clases (Modelo 3), mientras que el valor más bajo de AIC se produce en el Modelo 4, aunque seguido muy de cerca por el Modelo 3. El valor de AIC es sensible al tamaño de la muestra, por lo que en este escenario (N = 1.000) debe ser interpretado con cautela. Por tanto, para este caso preferimos quedarnos solamente con la información que proporciona BIC[2].

Tabla 1
Ajuste de los modelos LCA elaborados a partir de los datos simulados en el Estudio 1

Modelo LCA	Log-Likelihood	df	AIC	BIC
Modelo 1	-10.774,06	980	21.622,75	21.686,27
Modelo 2	-10.130,36	959	20.342,71	20.543,93
Modelo 3	-9.895,66	938	19.915,32	20.219,60
Modelo 4	-9.873,91	917	19.913,82	20.321,16

El Modelo 3 clasifica adecuadamente entre el 86% y el 90% de los grupos simulados, y asigna una probabilidad de pertenecía a cada clase casi idéntica a los valores poblacionales simulados de 0,3, 0,4 y 0,3 (0,29, 0,40, y 0,30, respectivamente).

Estudio 2

Respecto a los resultados obtenidos tras aplicar HCA, independientemente de si se utiliza la medida Chi-cuadrado o la medida Phi-cuadrado, e independientemente de si la clasificación se realiza para 2, 3 o 4 clústers, el método VMC clasifica a la inmensa mayoría de los sujetos en un solo grupo (> 98%). Algo similar ocurre con el método de vinculación VEG cuando se busca una solución de 4 clústers mediante la medida Chi cuadrado (el clúster 4 solamente está compuesto por un 0,8% de los sujetos). Además, cuando se utiliza VEG y la medida Phi cuadrado, existen varias agrupaciones compuestas por muy pocos sujetos, obteniéndose una agrupación del 84% o más en un solo grupo tanto si se realizan clasificaciones para 2 clústers como si se hacen para 3 o para 4. Con el método VML se obtienen agrupaciones más heterogéneas. No obstante, conviene señalar que utilizar la medida Chi cuadrado o la medida Phi cuadrado produce resultados marcadamente diferentes. Por ejemplo, cuando la clasificación se hace para dos clústers (situación en la que se produce una mayor aproximación entre ambas medidas de asociación), el grado de coincidencia es del 97,4% para uno de los clústers pero solo de un 76,4% para el otro. El grado de coincidencia mayor que se obtiene en la solución de tres clústers es del 78,9%, mientras que para una clasificación en base a cuatro clústers es del 86,8%.

Respecto a la aplicación de LCA mediante poLCa, en la Tabla 2 se recoge el ajuste obtenido por modelo siguiendo la misma lógica que la expuesta en el Estudio 1. En esta ocasión, el valor de AIC más bajo se sitúa en el Modelo 4, mientras que el valor de BIC más bajo se sitúa en el Modelo 1.

Tabla 2
Ajuste de los modelos LCA elaborados a partir de los datos simulados en el Estudio 2

Modelo LCA	Log-Likelihood	df	AIC	BIC
Modelo 1	-2.434,12	312	4.910,65	4.936,57
Modelo 2	-2.407,41	300	4.860,81	4.957,65
Modelo 3	-2.388,75	288	4.847,50	4.994,87
Modelo 4	-2.374,47	276	4.842,94	5.040,84

DISCUSIÓN

Reyna y Brussino (2011) concluyeron en su trabajo en base a cuestiones teóricas que la mejor solución era la del Modelo 4, avalada también por el valor de AIC. No obstante, estos autores no aplicaron HCA. Los resultados obtenidos tras aplica HCA indican que la ausencia de segmentación (Modelo 1) puede ser una solución plausible, avalada también por el valor más bajo de BIC obtenido tras aplicar LCA con poLCa. El grado de convergencia entre los resultados del HCA y del LCA elaborados en el presente trabajo para una agrupación basada en cuatro clústers no es muy elevado, lo que podría estar indicando que las distintas clases están demasiado entremezcladas, componiendo mixturas difíciles de detectar.

CONCLUSIONES

El objetivo de este trabajo ha sido mostrar el funcionamiento del LCA, en comparación con el HCA. En comparación con esta última técnica, el LCA permite realizar análisis desde una perspectiva más teórica, ya que el investigador puede poner a prueba distintos modelos y evaluar su bondad de ajuste. Además, el funcionamiento del LCA permite superar la limitación del HCA de utilizar una sola medida de asociación para todo el conjunto de variables analizadas. El LCA ha mostrado en el estudio de simulación elaborado una elevada capacidad para detectar estructuras de clases. No obstante, en contextos aplicados en donde las mixturas pueden estar más solapadas que en un estudio de simulación, el LCA puede mostrar resultados difíciles de interpretar al no coincidir la información que proveen los distintos índices de bondad de ajuste. Una estrategia que puede ser de utilidad para el investigador aplicado es aplicar LCA y HCA en busca de convergencia de resultados.

REFERENCIAS BIBLIOGRÁFICAS

Chen, X., Li, X., Stanton, B., et al. (2004). Patterns of cigarette smoking among students from 19 colleges and universities in Jiangsu Province, China: a latent class analysis. *Drug and Alcohol Dependency*, *76*, 53–163. https://doi.org/10.1016/j.drugalcdep.2004.04.013

Furberg, H., Sullivan, P.F., Bulik, C., Maes, H., Prescott, C.A., Kendler, K.S., & Lerman, C. (2005). The types of regular cigarette smokers: A latent class analysis. *Nicotine & Tobacco Research*, *7*(3), 351-360. https://doi.org/10.1080/14622200500124917

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis*. Pearson. Upper Saddle River, NJ.

Haughton, D., Legrand, P., & Woolford, S. (2009). Review of three latent class cluster analysis packages: Latent Gold, poLCA, and MCLUST. *The American Statistician*, *63*(1), 81-91. https://doi.org/10.1198/tast.2009.0016

Lazarsfeld, P., & Henry, N. (1968). *Latent structure analysis*. New York: Houghton-Mifflin.

Linzer, D.A., & Lewis, J.B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of statistical software*, *42*(10), 1-29.

Magidson, J., and Vermunt, J.K. (2002). Latent class models for clustering: A comparison with K-

means. *Canadian Journal of Marketing Research*, *20*, 37-44.
Magidson, J., and Vermunt J.K. (2003). A nontechnical introduction to latent class models. White Paper. Statistical Innovations.
McCutcheon, A.L. (2002). Basic concepts and procedures in single and multiplegroup latent class analysis. In J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis* (pp. 56–87). Cambridge University Press. Cambridge.
Nylund, K.L., Asparouhov, T., & Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*, *14*(4), 535-569. https://doi.org/10.1080/10705510701575396
Poletto, L., Pezzotto, S.M., Morini, J., & Andrade, J. (1991). Prevalencia del hábito de fumar en jóvenes y sus padres: asociaciones relevantes con educación y ocupación. *Revista de Saúde Pública*, *25*, 388-393.
R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
Reyna, C., & Brussino, S. (2011). Revisión de los fundamentos del análisis de clases latentes y ejemplo de aplicación en el área de las adicciones. *Trastornos Adictivos*, *13*(1), 11-19.
Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy*. Freeman. San Francisco.
Vermunt, J.K., Magidson, J. (2002). Latent Class Cluster Analysis. In J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis* (pp. 89–106). Cambridge University Press. Cambridge.
Vermunt, J.K., Magidson, J. (2005). Latent GOLD [computer program on disk]. Version 4.0. Belmont: Statistical Innovations. USA.
von Davier, M., Naemi, B., & Roberts, R.D. (2012). Factorial versus typological models: A comparison of methods for personality data. *Measurement: Interdisciplinary Research and Perspectives*, *10*(4), 185-208. https://doi.org/10.1080/15366367.2012.732798

1 Estudio encargado por Ministerio de Sanidad y Consumo de España y realizado por el Centro de Investigaciones Sociológicas – CIS (Estudio 2751; publicado en febrero de 2008).

2 Se repitió la simulación, pero esta vez a partir de N = 500. En este escenario tanto el valor de BIC como de AIC más bajos se situaron en el Modelo 3.