ALGUNAS CONSIDERACIONES PRÁCTICAS ACERCA DE LA ESTIMACIÓN DE PARÁMETROS EN EL MODELO CLÁSICO DE CLASES LATENTES

Adalberto González Debén, Instituto de Cibernética, Matemática y Física, Cuba
Hannelore Liero y Henning Läuter, Instituto de Matemática, Universidad de Potsdam, Alemania
Jesús E. Sánchez García, Instituto de Cibernética, Matemática y Física, Cuba[1]

RESUMEN
En este trabajo se lleva a cabo un estudio de simulación para determinar el efecto de algunos factores sobre la estimación de los parámetros en el modelo clásico de clases latentes. Se utilizó el modelo clásico de dos clases latentes y se tomaron en cuenta cuatro factores: (1) tamaño de muestra, (2) fuerza de la relación entre las variables manifiestas y la variable latente, (3) tamaño relativo de las clases latentes, y (4) ruido. Las variables de respuesta estudiadas son la cantidad de resultados indeterminados, la cantidad de soluciones con valores en la frontera, el sesgo y el error cuadrático medio de dos de los parámetros del modelo. Los resultados sugieren que es posible encontrar estimaciones satisfactorias para los parámetros del modelo, aún cuando el tamaño de muestra es relativamente pequeño, cuando todas las variables están fuertemente relacionadas con la variable latente y cuando no hay una clase latente con tamaño relativo muy pequeño. Por último, se hacen algunas recomendaciones con respecto al tamaño de muestra necesario para lograr resultados satisfactorios en cada uno de los casos contemplados en el esquema de simulación.

ABSTRACT
In this paper the effect of some factors on the parameter estimation in the latent class model is studied by means of a simulation study, in which a classical two latent class model is used and four factors are taken into account: (1) sample size, (2) strength of relation between the manifest variables and the latent one, (3) the relative size of the latent classes, and (4) the level of noise. The response variables studied are the amount of indeterminate results, the amount of frontier solutions, the bias and the mean square error two of the model parameter estimation. The results suggest that it is possible to find satisfactory estimations of the model parameters of latent classes, even with the samples sizes being relatively small, when all the manifest variables are strongly related to the latent one, and when there is no relatively small latent class. In addition, for each simulated condition, some suggestions related with the necessary sample size are given.

Key words: Latent Class Analysis, Parameter Estimation, Parametric bootstrap.

MSC: 62H17

1 INTRODUCCIÓN

El análisis de clases latentes (ACL) (LAZARSFELD AND HENRY, 1968; GOODMAN, 1974) es una técnica multivariada para encontrar grupos de individuos a partir de variables categóricas. En el mismo se supone que la estructura de relaciones que se observa entre las variables manifiestas se debe a la existencia de una variable categórica latente. En el modelo de clases latentes clásico (MCLC) se supone que las variables manifiestas son condicionalmente independientes dada la variable latente. Por ejemplo, si se consideran tres variables manifiestas A, B y C, y se supone que existe una variable latente X con T categorías, el MCLC es el siguiente:

$$P(A=i, B=j, C=k) = \sum_{t=1}^{T} P(X=t)P(A=i/X=t)P(B=j/X=t)P(C=k/X=t)$$

Para una revisión de este método se recomienda CLOGG (1995). En ROST y LANGEHEINE (1997) se puede encontrar una gran variedad de aplicaciones.

Entre los usos más frecuentes del ACL está el descubrimiento de subtipos diagnósticos o subcategorías, así como la evaluación de test diagnósticos en ausencia de un “estandar de oro”. Las áreas de aplicación

[1]**E-mail**: grupoest@icmf.inf.cu

comprenden la investigación de mercados, el análisis de encuestas, la sociología, la psicología, la educación y la medicina. (UEBERSAX, 2001)

Una aplicación interesante del ACL es el uso de los parámetros del MCLC para determinar la confiabilidad de tests compuestos por indicadores cualitativos. (CLOGG y MANNING, 1996; GONZÁLEZ y MONTEAVARO, 2001)

Al igual que otros métodos de análisis de tablas de contingencia, el ACL requiere tamaños de muestra relativamente grandes para evitar los problemas relacionados con las tablas ralas. Una tabla de contingencia es rala cuando tiene muchas celdas con frecuencias bajas (AGRESTI, 2000). Una forma práctica de determinar el nivel de raleza en una tabla de contingencia es con el cociente entre el tamaño de muestra y el número de celdas. (AGRESTI y YANG, 1987)

Algunos autores han estudiado el efecto de la raleza sobre los test de bondad de ajuste y sobre la estimación de parámetros en el ACL. Con respecto a los tests de bondad de ajuste, COLLINS **et al**. (1993) propusieron el uso de estudios de simulación para estimar el valor esperado de los estadígrafos de prueba bajo la hipótesis nula de que el modelo se ajusta a los datos. FORMANN y KOHLMANN (1997), VAN DER HEIJDEN, HART y DESSENS (1997), VON DAVIER y ROST (1997), y GOETGHEBEUR **et al**. (2000) utilizan este procedimiento, también conocido como “bootstrap paramétrico”, para analizar la bondad de ajuste de modelos en tablas ralas. En síntesis, el método consiste en cinco pasos fundamentales: (1) estimar los parámetros del modelo a partir de la muestra de datos reales, (2) generar un número suficientemente grande de muestras a partir de los parámetros estimados, (3) estimar los parámetros del modelo para cada muestra simulada, (4) calcular el estadígrafo de prueba para cada muestra simulada, y (5) estudiar el comportamiento empírico del estadígrafo de prueba.

Hasta donde conocen los autores, existen dos estudios del efecto de la raleza sobre la estimación de los parámetros del MCL. GOETGHEBEUR **et al**. (2000) analizan la distribución de los parámetros estimados por medio del bootstrap paramétrico y señalan que en algunos casos se observa una “variación sustancial”.

COLLINS, FIDLER and WUGALTER (1996) llevaron a cabo un estudio de simulación con el objetivo de analizar este fenómeno en una forma más sistemática. Para ello utilizaron un modelo con dos clases latentes con probabilidades iguales ($P(X = 1) = P(X = 2)$) y tomaron en cuenta tres factores relacionados con la raleza: (1) el número de variables manifiestas, (2) la fuerza de la relación entre las variables manifiestas y la variable latente, y (3) el cociente entre el tamaño de muestra y el número de celdas (N/k.) Las variables dependientes que analizaron fueron: (1) la cantidad de soluciones indeterminadas, (2) el sesgo y el error cuadrático medio de las estimaciones y (3) la estimación del error estándar. La conclusión principal a la que arribaron es que es posible obtener buenas estimaciones de los parámetros aún cuando las tablas son ralas; sobre todo cuando la relación entre las variables manifiestas y la variable latente es muy fuerte.

Este trabajo puede ser considerado una continuación del llevado a cabo por COLLINS **et al**. (1996). En el mismo se tuvieron en cuenta cuatro factores:

(1) Fuerza de la relación entre las variables manifiestas y la variable latente,

(2) Tamaño relativo de las clases latentes,

(3) Ruido y

(4) Tamaño de muestra.

El objetivo es estudiar, además del efecto del tamaño de muestra, qué pasa cuando (1) la fuerza de la relación entre las variables manifestas y la variable latente no es la misma para todas las variables manifiestas, (2) las clases latentes tienen probabilidades diferentes, y (3) los datos están contaminados por error.

Las consecuencias prácticas de este estudio son directas. Es lógico esperar que en las investigaciones empíricas las magnitudes de la relación entre las variables manifiestas y la variable latente sean diferentes. COLLINS **et al**. (1996) recomiendan el estudio de este problema. Por ejemplo, si un test está compuesto por varios items categóricos, las confiabilidades específicas de cada uno de ellos no tienen por qué ser iguales. Una pregunta que interesa responder es si la determinación de la confiabilidad específica de un item depende de que los demás items que conforman el test sean o no igualmente confiables.

En general, en un modelo de dos clases latentes, éstas son interpretadas como “sano/enfermo”, “positivo/negativo”, “aprobado/desaprobado”, “con la habilidad/sin la habilidad”, etc. En la mayoría de las aplicaciones es lógico pensar que ambas clases latentes contienen proporciones diferentes de individuos. Esta es una condición favorable para la presencia de frecuencias bajas en algunas celdas en la tabla de contingencia formada por las variables manifiestas.

Por otro lado, los datos reales están usualmente contaminados por diferentes fuentes de error. Por ejemplo, algunos de los individuos de una muestra puede contestar al azar los items de un test.

Por último, el tamaño de la muestra está directamente relacionado con la raleza. Es muy común que los investigadores trabajen con tablas de contingencia que tienen una cantidad relativamente grande de celdas. Esto puede ser consecuencia de que se quieren considerar muchas variables, o variables con muchas categorías. Para ilustrar este problema GOETGHEBEUR **et al**. (2000) menciona cinco ejemplos de la literatura de análisis de clases latentes donde las tablas son ralas. El problema real que ellos analizan en su artículo también tiene esta característica. En el caso de la construcción de un test, usualmente se trata de incluir tantos items como sean necesarios con el objetivo de garantizar el control de la validez y la confiablidad. Se necesita un criterio que sirva de guía al investigador para saber qué tamaño de muestra se requiere en dependencia de la magnitud y de las características deseadas del test.

Desde el punto de vista del investigador, estos factores no son igualmente manipulables en la práctica. En particular, la fuerza de la relación entre las variables manifiestas y las variables latentes depende de las características del constructo que se quiere estudiar y de la calidad de los items. La proporción de individuos en cada clase latente es algo que depende del problema y no del investigador. La proporción de individuos que responden las preguntas al azar puede depender tanto de las características de la población como de la naturaleza de los items. Por último, el tamaño de muestra depende en última instancia de los recursos disponibles para llevar a cabo cada estudio concreto.

2. DISEÑO DEL ESTUDIO DE SIMULACIÓN

Se realizó un estudio de simulación con el objetivo de estudiar los factores que afectan la estimación de parámetros en el ACLC. El estudio de simulación consistió en la generación de conjuntos de datos a partir de modelos con parámetros conocidos, diferentes niveles de raleza y contaminación por errores aleatorios.

Para este estudio se escogió el modelo clásico de clases latentes:

$$P(A=i,B=j,C=k,D=l)=\sum_{t=1}^{2}P(X=t)P(A=i/X=t)P(B=j/X=t)P(C=k/X=t)P(D=l/X=t)$$

donde A, B, C, y D son cuatro variables manifiestas con dos categorías cada una, y X es una variable latente con dos clases.

Para seleccionar los parámetros del modelo se utilizaron dos factores:

- Estructura
- Asimetría

La estructura se refiere a la fuerza de la relación entre las variables manifiestas y la variable latente. La asimetría se refiere a la proporción de individuos que pertenece a cada clase.

Para el factor Estructura se escogieron dos niveles: “Fuerte” y “Débil”. En la Estructura Fuerte la probabilidad condicional de que las variables manifiestas tomen el valor “1” es igual a 0.9 para la primera clase latente y a 0.2 para la segunda clase latente. En la Estructura Débil la probabilidad condicional de que las variables A y B tomen el valor “1” es igual a 0.9 para la primera clase latente y a 0.2 para la segunda clase latente; mientras que la probabilidad condicional de que las variables C y D tomen el valor “1” es igual a 0.6 para la primera clase latente y a 0.4 para la segunda clase latente (ver Tabla 1). En la Estructura Fuerte todas las variables manifiestas están fuertemente relacionadas entre sí, pero esto no sucede en la Estructura Débil.

Tabla 1. Factor Estructura.

FACTOR	Estructura			
NIVELES	Fuerte		Débil	
CLASE LATENTE	t = 1	t = 2	t = 1	t = 2
P(A = 1/X = t)	0.9	0.2	0.9	0.2
P(B = 1/X = t)	0.9	0.2	0.9	0.2
P(C = 1/X = t)	0.9	0.2	0.6	0.4
P(D = 1/X = t)	0.9	0.2	0.6	0.4

Para el factor Asimetría se escogieron tres niveles: “Ninguna”, “Ligera” y “Mucha”. En el nivel Ninguna la probabilidad de pertenecer a cada clase latente es 0.5. En el nivel Ligera la probabilidad de pertenecer a la primera clase latente es 0.7, y la probabilidad de pertenecer a la segunda clase latente es 0.3. En el nivel Mucha la probabilidad de pertenecer a la primera clase latente es 0.9, y la probabilidad de pertenecer a la segunda clase latente es 0.1 (ver Tabla 2)

Tabla 2. Factor Asimetría.

FACTOR	Asimetría		
NIVELES	Ninguna	Poca	Mucha
P(X = 1)	0.5	0.7	0.9
P(X = 2)	0.5	0.3	0.1

Adicionalmente, se tuvieron en cuenta otros dos factores:

- Ruido
- Tamaño de muestra

Con respecto al factor Ruido, se refiere a la proporción de individuos de la muestra que no es generada por el modelo. Se seleccionaron tres niveles: 0, 1/8 y 1/4. Para las observaciones de estos individuos en cada una de las variables manifiestas, se generaron valores independientes de una distribución binomial con probabilidad de éxito igual a 0.5 (ver Tabla 3).

Tabla 3. Factor Ruido.

FACTOR	Ruido		
NIVELES	Ninguno	Poco	Mucho
SEGÚN EL MODELO	1	7/8	3/4
AL AZAR	0	1/8	1/4

Para el Tamaño de muestra n se seleccionaron cuatro valores: 32, 64, 128 y 256. El número de celdas de la tabla de contingencia formada por las cuatro variables manifiestas dicotómicas es constante (k = 16). Por lo tanto la razón n/k, que llamaremos nivel de raleza, toma los valores 2, 4, 8, y 16 respectivamente (ver Tabla 4).

Tabla 4. Factor Tamaño de muestra.

MUESTRA (n)	32	64	128	256
CELDAS (k)	16	16	16	16
(n/k)	2	4	8	16

En total, el estudio de simulación tiene 72 tratamientos, que resultan de combinar dos niveles de Estructura, tres niveles de Simetría, tres niveles de Ruido, y cuatro niveles de Tamaño de muestra.

Como la variable latente es nominal, es necesario definir una clase latente "principal" para evitar ambigüedad cuando se hace referencia a la clase latente X = 1 y X = 2. En el modelo de dos classe latentes $t^{(1)}$ y $t^{(2)}$), se define a $t^{(1)}$ como la clase latente principal si cumple la doble condición: $P(A = 1/X = t^{(1)}) > 0.5$ y $P(A = 2/X = t^{(2)}) > 0.5$.

Las soluciones donde no apareció una clase latente principal fueron consideradas indeterminadas, en el sentido de que no es posible determinar cuáles son los parámetros de interés para el estudio.

Teniendo en cuenta solo las soluciones donde apareció una clase latente principal, se estudió la calidad en la estimación de los dos parámetros siguientes:

- $P(X = t^{(1)})$
- $P(A = 1/X = t^{(1)})$

En resumen, las variables dependientes estudiadas fueron:

- Cantidad de soluciones indeterminadas
- Cantidad de soluciones con valor del parámetro $P(A = 1/X = t^{(1)})$ en la frontera (igual a 1)
- Sesgo para el parámetro $P(X = t^{(1)})$
- Error cuadrático medio (ECM) para el parámetro $P(X = t^{(1)})$
- Sesgo para el parámetro $P(A = 1/X = t^{(1)})$
- Error cuadrático medio (ECM para el parámetro $P(A = 1/X = t^{(1)})$

Para la generación de los datos se utilizó un programa escrito en MATLAB versión 5.3 y para el ajuste de los modelos de clases latentes se utilizó el paquete de programas LEM (VERMUNT, 1997).

Para cada tratamiento se generaron tantas réplicas como fue necesario para lograr un total de 1000 soluciones con clase latente principal.

3. RESULTADOS

Como se explicó en la introducción, desde el punto de vista del investigador los factores Tamaño de muestra y Estructura son más factibles de controlar, y por el contrario los factores Simetría y Ruido no dependen tanto de él sino de las características de la población y del constructo que se analiza. La discusión de los resultados se va a organizar teniendo en cuenta esto.

3.1. Cantidad de soluciones indeterminadas

La cantidad de soluciones indeterminadas para cada tratamiento se muestra en la Tabla 5.

Tabla 5. Muestras generadas hasta alcazar 1000 soluciones con clase maestra.

ASIMETRÍA	MUESTRA	ESTRUCTURA FUERTE			ESTRUCTURA DÉBIL		
		RUIDO			RUIDO		
		0	1/8	1/4	0	1/8	¼
P(X = 1) = 0.5	256	1000	1000	1000	1000	1000	1000
	128	1000	1000	1000	1000	1000	1003
	64	1002	1003	1002	1004	1011	1015
	32	1004	1014	1021	1023	1051	1075
P(X = 1) = 0.7	256	1000	1000	1000	1002	1002	1007
	128	1001	1003	1009	1006	1016	1035
	64	1005	1012	1031	1043	1086	1107
	32	1034	1060	1091	1144	1151	1175

P(X = 1) = 0.9	256	1008	1024	1060	1238	1326	1333
	128	1037	1097	1146	1410	1449	1409
	64	1131	1186	1225	1647	1540	1477
	32	1285	1333	1350	1869	1740	1659

Como se puede apreciar, para un tamaño de nuestra fijo, la cantidad de soluciones indeterminadas aumenta cuando:

- aumenta la asimetría,
- aumenta el ruido,
- la estructura es débil.

Este problema se hizo más frecuente con la disminución del tamaño de muestra. Considerando un 10% de soluciones indeterminadas (111 de 1111) como un límite razonable para afirmar que el problema no es grave, se puede decir cuál es el mínimo tamaño de muestra necesario para no exceder este límite. En el caso simétrico ($P(X = t^{(1)}) = 0.5$) el número de soluciones indeterminadas se mantuvo siempre por debajo de 75, alcanzándose este valor en las peores condiciones (estructura débil y mucho ruido). Para la asimetría ligera ($P(X = t^{(1)}) = 0.7$), no hubo problemas con la estructura fuerte, pero sí con la estructura débil y tamaño de muestra 32. En el caso de mucha asimetría ($P(X = t^{(1)}) = 0.9$), hubo problemas con la estructura fuerte y tamaños de muestra 32 y 64, y para todos los tamaños de muestra en el caso de la estructura débil.

3.2. Cantidad de soluciones límite $P(A = 1/X = t^{(1)}) = 1$

La cantidad de soluciones límite para cada tratamiento se muestra en la Tabla 6. Como se puede apreciar, este problema no parece estar relacionada con el ruido, y aumenta cuando disminuye el tamaño de muestra y cuando la estructura es débil. Analizando sólo el caso de ausencia de ruido, y considerando nuevamente el 10% (100 de 1000) como una medida de la gravedad del problema, se ve que este valor se supera en todos los niveles de asimetría y tamaños de muestra 32 y 64 para la estructura débil y en todos los niveles de asimetría y tamaño de muestra 32 para la estructura fuerte.

Tabla 6. Cantidad de soluciones límite P(A = 1/X = 1) = 1.

ASIMETRÍA	MUESTRA	ESTRUCTURA FUERTE			ESTRUCTURA DÉBIL		
		RUIDO			RUIDO		
		0	1/8	1/4	0	1/8	¼
P(X = 1) = 0.5	256	0	0	0	51	40	37
	128	4	5	8	79	95	81
	64	70	55	56	140	162	144
	32	246	219	201	231	200	208
P(X = 1) = 0.7	256	0	0	0	32	37	49
	128	1	1	2	97	79	96
	64	30	22	32	154	148	156
	32	144	164	154	280	238	231
P(X = 1) = 0.9	256	0	0	0	72	90	78
	128	0	2	4	151	157	145
	64	12	27	42	240	214	195
	32	119	137	136	428	298	275

3.3. Caso especial: estructura fuerte y tamaño de muestra grande

En los dos epígrafes anteriores se vio que la aparición de soluciones indeterminadas y de soluciones límite fue una característica dominante en los tratamientos con tamaño de muestra n = 32 y estructura débil. Podemos decir que las condiciones mínimas que debe garantizar el investigador son un tamaño de muestra suficientemente grande y unos items fuertemente relacionados con la variable latente. En el caso de este estudio de simulación esto se corresponde con los tratamientos que tienen un tamaño de muestra igual o mayor a 64 y una estructura fuerte.

Para facilitar la exposición, primero se analizará el efecto de los factores Asimetría y Ruido para el caso n = 256, y después los factores Asimetría y Tamaño de muestra en ausencia de ruido. Para el caso n = 256, no resultó importante el problema de las soluciones indeterminadas (ver Tabla 5), pues no hubo ninguna en los casos de simetría y asimetría ligera, y hubo pocas en el caso de mucha asimetría (0.8, 2.3 y 5.7% para ningún ruido, poco y mucho respectivamente).

Por otro lado, en ningún tratamiento hubo soluciones límite (ver Tabla 6).

En la Tabla 7 se muestran los resultados de la estimación de los parámetros.

Tabla 7. Estimación de los parámetros para el caso Estructura Fuerte y n = 256.

PARÁMETRO	$P(X = t^{(1)})$			$P(A = 1/X = t^{(1)})$		
RUIDO	0	1/8	2/4	0	1/8	2/4
P(X = 1) = 0.5	0.0017 (0.0014)	- 0.0173 (0.0022)	- 0.0414 (0.0043)	- 0.0005 (0.0011)	- 0.0145 (0.0016)	- 0.0328 (0.0031)
P(X = 1) = 0.7	0.0001 (0.0012)	- 0.0456 (0.0040)	- 0.0983 (0.0126)	0.0004 (0.0007)	- 0.0119 (0.0011)	- 0.0223 (0.0021)
P(X = 1) = 0.9	- 0.0017 (0.0006)	- 0.0750 (0.0076)	- 0.1664 (0.0315)	0.0015 (0.0005)	- 0.0081 (0.0009)	- 0.0136 (0.0014)

En todos los casos, la calidad de la estimación empeoró con el aumento del ruido.

La calidad de la estimación del parámetro $P(X = t^{(1)})$ mejoró con el aumento de la asimetría en ausencia de ruido, pero empeoró en presencia de ruido. Esta aparente contradicción se explica porque cuando no hay ruido la asimetría implica mayor cantidad de individuos en la clase $X= t^{(1)}$, y por lo tanto más información; pero cuando hay ruido, los individuos que responden al azar tienen una probabilidad igual a 0.5 de pertenecer a cada una de las clases latentes, y por lo tanto el efecto distorcionador de este ruido es mayor a medida que aumenta la asimetría.

La calidad de las estimación del parámetro $P(A = 1/X = t^{(1)})$ mejoró con el aumento de la asimetría, para todos los niveles de ruido. Esto se explica porque con el aumento de la asimetría, hay una mayor cantidad de individuos en la clase latente $X = t^{(1)}$.

Ya se vio que cuando el tamaño de muestra disminuye, aumenta la cantidad de soluciones indeterminadas y de soluciones límite. En la Tabla 8 se muestra el efecto sobre la calidad de la estimación de los parámetros de los factores asimetría y tamaño de muestra para el caso estructura fuerte y sin ruido.

Tabla 8. Estimación de los parámetros para el caso Estructura Fuerte y sin ruido.

PARÁMETRO	$P(X = t^{(1)})$			$P(A = 1/X = t^{(1)})$		
MUESTRA	256	128	64	256	128	64
P(X = 1) = 0.5	0.0017 (0.0014)	0.0014 (0.0030)	0.0059 (0.0061)	- 0.0005 (0.0011)	- 0.0004 (0.0021)	- 0.0014 (0.0043)
P(X = 1) = 0.7	0.0001 (0.0012)	- 0.0000 (0.0026)	0.0001 (0.0048)	0.0004 (0.0007)	0.0013 (0.0015)	0.0015 (0.0028)
P(X = 1) = 0.9	- 0.0017 (0.0006)	- 0.0032 (0.0014)	- 0.0097 (0.0025)	0.0015 (0.0005)	0.0005 (0.0010)	0.0039 (0.0019)

En resumen, la calidad de las estimaciones de los parámetros mejoró con el aumento de la asimetría y empeoró con la disminución del tamaño de muestra. Pese a esto último, se puede ver en todos los casos los valores del sesgo y el error cuadrático medio son pequeños. Por lo tanto se puede decir que para estos tamaños de muestra se obtuvieron soluciones aceptables en los tres niveles de asimetría.

4. CONCLUSIONES

Al igual que en todo estudio de simulación, los resultados de este trabajo sólo son válidos para las condiciones que se consideraron en el mismo. En este caso se utilizó el modelo clásico de dos clases latentes.

Un problema que se presenta cuando se trata de interpretar los parámetros en el modelo clásico de clases latentes es la existencia de indeterminación, esto es: que no hay posibilidad de distinguir entre las clases latentes. El aumento del número de soluciones indeterminadas está relacionado con el incremento de la asimetría y del ruido, y con la disminución del tamaño de muestra. También sucede cuando no todas las variables manifiestas tienen una relación fuerte con la variable latente. Con respecto a la magnitud del problema, este carece de importancia cuando las clases latentes tienen probabilidades similares, aún en presencia de ruido y en el caso de que el tamaño de muestra sea de sólo dos veces el número de celdas de la tabla de contingencia. Cuando la asimetría es ligera, no hay problemas siempre que la relación entre todas las variables manifiestas y la variable latente sea fuerte. Cuando la asimetría es grande, se necesita además que el tamaño de muestra sea de más de cuatro veces el número de celdas de la tabla de contingencia.

Otro problema que se puede presentar es que haya soluciones con valores límite o de frontera, esto es, probabilidades condicionales con valor 1 y 0. No se encontró relación entre este problema y el ruido, pero sí con la disminución del tamaño de muestra y la asimetría. Se necesita que el tamaño de muestra sea de al menos cuatro veces el número de celdas de la tabla de contingencia, y que la relación entre todas las variables manifiestas y la variable latente sea fuerte.

Con respecto a la estimación de los parámetros del modelo, el sesgo y el ECM son despreciables cuando las variables manifiestas están fuertemente relacionadas con la variable latente y el tamaño de muestra supera el doble del número de celdas de la tabla de contingencia, incluso en presencia de ruido y distintos niveles de asimetría.

Teniendo en cuenta lo anterior, se recomienda un tamaño de muestra de al menos cuatro veces el número de celdas de la tabla de contingencia para el mejor caso, que es cuando las clases latentes son de tamaños similares o ligeramente diferentes y la relación entre todas las variables manifiestas y la variable latente es fuerte.

Cuando las clases latentes son de tamaños muy diferentes, se necesita un tamaño de muestra de al menos ocho veces el número de celdas de la tabla de contingencia.

Cuando alguna variable manifesta no está fuertemente relacionada con la variable latente, se debe analizar la posibilidad de eliminarla, no sólo por su escasa contribución al modelo sino porque afecta la calidad de la estimación de los parámetros.

El incremento del número de variables manifiestas está directamente relacionado con la raleza. Aunque en este estudio el número de variables manifiestas se mantuvo constante, se pueden hacer algunos comentarios a partir de las diferencias encontradas entre los niveles de los factores “Estructura” y “Asimetría”. Suponiendo que el tamaño de muestra es suficientemente grande, el efecto de añadir una nueva variable depende de cuán grande es su relación con las variables manifiestas anteriores. Sólo cuando la nueva variable está fuertemente relacionada con las anteriores no habrá un efecto sobre la estimación de los parámetros del modelo. Por otro lado, el efecto de muchas celdas con frecuencias bajas cuando se añade una nueva variable es mayor cuando las clases latentes tienen probabilidades muy diferentes.

Todos estos criterios deben tenerse en cuenta cuando se utilice el análisis de classe latentes como metodología para la determinación de la confiabilidad de indicadores cualitativos.

AGRADECIMIENTOS

Los autores agradecen al Dr. Alexander von Eye haber sugerido este tema y la ayuda recibida para llevarlo a cabo.

Este trabajo se realizó durante estancias de los autores en la Universidad de Potsdam y el ICIMAF, respectivamente. Estas estancias fueron apoyadas en parte por el DAAD.

REFERENCIAS

AGRESTI, A. (1990): **Categorical Data Analysis**, Wiley, New York.

AGRESTI, A. and M.C. YANG (1987): “An Empirical Investigation of Some Effects of Sparseness in Contingency Tables”. **Computat. Statist. Data Anal**. 5, 9-21.

CLOGG, C.C. (1995): “Latent Class Models·. In: G. Arminger, C.C. Clogg, and M.E. Sobel (eds.): **Handbook of Statistical Modeling for the Social and Behavioral Sciences**. Academic Press, New York.

CLOGG, C.C. and W.D. MANNING (1996): “Assessing Reliability of Categorical Measurements Using Latent Class Models”. In A. von Eye and C.C. Clogg (eds.): **Categorical Variables in Developmental Research**. Academic Press, New York.

COLLINS, L.M.; P.L. FIDLER; S.E. WUGALTER and J.D. LONG (1993): “Goodness of Fit Testing for Latent Class Models”. **Multivariate Behavioral Research**, 28, 375-389.

COLLINS, L.M.; P.L. FIDLER and S.E. WUGALTER (1996): “Some Practical Issues Related to the Estimation of Latent Class and Latent Transition Parameters”. In: A. von Eye and C.C. Clogg (eds.): **Categorical Variables in Developmental Research**, Academic Press, New York.

DAVIER, M. von, and R. LANGEHEINE (1996): “Self monitoring – a Class Variable?” In: J. Rost and R. Langeheine (eds.): **Applications of Latent Trait and Latent Class Models in Social Sciences**, Plenum, New York.

FORMANN, A.K. and T. KOHLMANN (1996): “Latent Class Analysis in Medical Research”. **Statistical Methods in Medical Research** 5, 179-211.

GOETGHEBEUR, E.; J. LIINEV; M. BOELAERT and P. VAN DER STUYFT (2000): “Diagnostic Test Analyses in Search of their Gold Standard: Latent Class analysis with Random Effects”. **Statistical Methods in Medical Research**, 9, 231-248.

GOODMAN, L.A. (1974): “Exploratory Latent Structures analysis Using Both Identifiable and Unidentifiable Models”. **Biometrika**, 61, 215-231.

GONZÁLEZ, A. and M. MONTEAVARO (2001): **Sobre la aplicación de los modelos de clases latentes al estudio de la confiabilidad de indicadores cualitativos**. Reporte de Investigación del ICIMAF 2001, 131.

LAZARSFELD, P.F. and N.W. HENRY (1968): **Latent Structure Analysis**, Houghton Mifflin, Boston.

ROST, J. and R.A. LANGEHEINE (1997): Applications of Latent Trait and Latent Class Models in the Social Sciences, Waxmann, New York.

UEBERSAX, J. (2001): Latent Class Analysis Frequently Asked Questions (FAQ). Internet www page at URL: http://ourworld.compuserve.com/homepages/juebersax/faq.htm (version current as of October 5).

VAN der HEIJDEN, P.; H. HART and J. DESSENS (1996): “A Parametric Bootstrap Procedure to Perform Statistical Tests in a LCA of Anti-social Behaviour”. In: J. Rost and R. Langeheine (eds.): **Applications of Latent Trait and Latent Class Models in Social Sciences**, Plenum, New York.

VERMUNT, J.K. (1997): **LEM: A General Program for the Analysis of Categorical Data**. Internet www page at URL: http://cwis.kub.nl/~ fsw_1/mto.