Publicación Electrónica ISSN 2718-6636

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA
SECRETARIA DE CIENCIA Y TECNOLOGIA E INSTITUTOS DE INVESTIGACIONES

Resumen Ampliado

Jornadas Anuales

“Investigaciones en la Facultad” de Ciencias Económicas y Estadística

Bartolelli, Gino[1]; Cuesta Cristina[1]; Wojdyla Daniel[2]

[1]*Facultad de Ciencias Económicas y Estadística, UNR*
[2]*Duke Clinical Research Institute, Duke University*

MODELIZACIÓN DE DATOS LONGITUDINALES CON CLASES LATENTES[1]

Resumen

La modelización de datos longitudinales consiste en describir la evolución de observaciones medidas a través del tiempo teniendo en cuenta la estructura de correlación presente entre las mediciones. En situaciones donde se presume la existencia de factores no observables que generan una heterogeneidad que no puede ser explicada a partir de los datos, los modelos longitudinales con clases latentes surgen como una alternativa flexible para reconocer patrones e identificar subgrupos de observaciones con trayectorias similares. El proceso de estimación de estos modelos por el método de máxima verosimilitud se ve dificultado debido a la presencia de múltiples óptimos locales en la superficie de la verosimilitud. Como se muestra en este trabajo, falencias a la hora de tener en cuenta estas soluciones pueden conducir a resultados sesgados e interpretaciones incorrectas. Los resultados de este estudio de simulación muestran que las soluciones locales están presentes incluso en modelos parsimoniosos con datos completos y bien comportados. El uso de rutinas de búsqueda automática considerando valores iniciales distintos es de vital importancia para asegurar la validez de este tipo de análisis.

Palabras claves: Datos longitudinales, Modelos mixtos, Mezclas finitas, Clases latentes, Soluciones locales.

Introducción

Los datos longitudinales se presentan, generalmente, como un conjunto de observaciones medidas repetidamente en distintos momentos. El análisis de este tipo de datos consiste en describir la evolución de las observaciones a través del tiempo, contemplando la estructura de dependencia entre las mediciones (Fitzmaurice, Laird, Nan M., y Ware, James H. 2018). La trayectoria media puede ser modelada satisfactoriamente con las técnicas de regresión tradicionales incluyendo la covariable tiempo o alguna función de esta. Por otro lado, la estructura de correlación de los datos es más difícil de captar y puede modelarse de diversas maneras.

Modelos Mixtos

Una manera eficiente de tener en cuenta el carácter longitudinal de los datos es la inclusión de efectos aleatorios en el modelo, lo cual induce una dependencia entre las mediciones. Por ejemplo, la inclusión de un efecto aleatorio en el intercepto introduce una correlación positiva entre mediciones repetidas de una misma observación, mientras que la inclusión de un efecto aleatorio en la pendiente genera una estructura de correlación que decrece en el tiempo. Además, se pueden incluir efectos aleatorios de mayor jerarquía para captar asociaciones entre grupos de observaciones. Un modelo lineal mixto puede expresarse matricialmente de la siguiente manera:

[1] Trabajo elaborado en el marco del Proyecto (80020230200046UR), titulado: “Modelos Lineales Avanzados. Nuevas propuestas, usos y procedimientos.”, dirigido por Cristina Cuesta.

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_i + \mathcal{E}_i,\ i = 1, \ldots, n \\ \mathbf{b}_i &\sim \text{NMV}(\mathbf{0}, \mathbf{G}),\ \mathcal{E}_i \sim \text{NMV}(\mathbf{0}, \mathbf{R}) \end{aligned}$$

Modelos de Mezclas Finitas

Dado un vector observaciones $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, un modelo mezcla describe el comportamiento probabilístico de los datos a partir de una ponderación de (K) componentes (McLachlan y Peel 2000):

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \, f_k(\mathbf{y}_i)$$

En la ecuación del modelo, $\{f_k\}$ son las componentes que conforman la mezcla y $\{\pi_k\}$ son los pesos asociados a cada una. La mezcla $f(\mathbf{y}_i)$ constituye una función de probabilidad válida ya que cada componente es en sí misma una función de probabilidad. El número de componentes que conforman la mezcla se asume fijo, si bien en muchas aplicaciones este número es desconocido y se infiere a partir de los datos.

Habitualmente, los modelos mezcla son útiles cuando se presume que la muestra está conformada por observaciones heterogéneas que provienen de poblaciones diferentes pero desconocidas de antemano. La flexibilidad de estos modelos permite ajustar distribuciones que no pueden ser captadas por los métodos tradicionales y explicar heterogeneidad no asignable a fuentes conocidas, lo cual implica que deben tomarse muchos recaudos a la hora de utilizarlos.

Modelos mixtos con clases latentes

La inclusión de clases latentes en un modelo lineal parte de presumir que las observaciones que conforman la muestra provienen de un cierto número de poblaciones diferentes. Esta conjetura puede encontrar sustento en conocimientos de la problemática, o bien puede introducirse para modelar una fuente de heterogeneidad no atribuible a factores conocidos.

En presencia de (K) clases latentes, el modelo mixto más general postula una trayectoria media y una estructura de correlación específica para cada una de las clases:

$$\begin{aligned} \mathbf{y}_i^{(k)} &= \mathbf{X}\boldsymbol{\beta}^{(k)} + \mathbf{Z}\mathbf{b}_i^{(k)} + \mathcal{E}_i^{(k)},\ k = 1, \ldots, K \\ \mathbf{b}_i^{(k)} &\sim \text{NMV}\left(\mathbf{0}, \mathbf{G}^{(k)}\right),\ \mathcal{E}_i^{(k)} \sim \text{NMV}\left(\mathbf{0}, \mathbf{R}^{(k)}\right) \end{aligned}$$

La ecuación del modelo explicita la distribución condicional de los datos dada su clase, teniendo en cuenta que esta pertenencia es desconocida. Como las clases se suponen mutuamente excluyentes, por ley de probabilidad total podemos deducir la distribución marginal de la respuesta:

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} P\,(i \in k) f(\mathbf{y}_i | i \in k) = \sum_{k=1}^{K} \pi_k \, f_k(\mathbf{y}_i)$$

donde (π_k) es el tamaño relativo de la población (k) y $f_k(\mathbf{y}_i)$ está especificada por el modelo. Esto demuestra que el modelo mixto con clases latentes es una mezcla finita de normales multivariadas NMV$(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, cuyos vectores de esperanzas y matrices de covariancia están dados por:

$$\boldsymbol{\mu}_k = \mathbf{X}\boldsymbol{\beta}^{(k)},\ \boldsymbol{\Sigma}_k = \mathbf{R}^{(k)} + \mathbf{Z}\mathbf{G}^{(k)}\mathbf{Z}'$$

Estimación por Máxima Verosimilitud

El método de estimación más utilizado para ajustar modelos mezcla es el de Máxima Verosimilitud (MV), el cual consiste en maximizar la función de verosimilitud de la muestra suponiendo independencia entre las observaciones:

$$L(\mathbf{y}, \boldsymbol{\Psi}) = \prod_i f\,(y_i, \boldsymbol{\Psi})$$

donde ($\boldsymbol{\Psi}$) representa el vector completo de parámetros a ser estimados. Por lo general, en vez de emplear directamente la verosimilitud se trabaja con su logaritmo porque simplifica considerablemente los cálculos. La log-verosimilitud del modelo mezcla viene dada por:

$$\begin{aligned} \log L(\mathbf{y}, \boldsymbol{\Psi}) &= \sum_i \log\, f(y_i, \boldsymbol{\Psi}) \\ &= \sum_i \log \sum_k \pi_k\, f_k(y_i, \boldsymbol{\theta}_k) \end{aligned}$$

Bajo condiciones de regularidad, los estimadores MV son una solución del sistema de derivadas parciales respecto del vector de parámetros:

$$\frac{\partial \log L(\mathbf{y}, \boldsymbol{\Psi})}{\partial\, \boldsymbol{\Psi}} = \mathbf{0}$$

En la amplia mayoría de los casos, este sistema no tiene solución cerrada y se manipula algebraicamente para inducir un esquema iterativo. Este patrón recursivo puede ser identificado sin ningún artilugio reformulando el enunciado como un problema de datos faltantes y aplicando un algoritmo conocido por sus siglas (EM, del inglés *expectation-maximization*) (Dempster, N. M. Laird, y D. B. Rubin 1977).

Aplicación del algoritmo EM

En el marco de este algoritmo, el modelo mezcla se reformula conceptualmente presumiendo que cada observación ($\mathbf{y}_i$) surge de una de las componentes que conforman la mezcla. Es necesario introducir el vector no observado de pertenencia a las componentes (o vector de etiquetas):

$$\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$$

donde ($\mathbf{z}_i$) es un vector K-dimensional de variables indicadoras de pertenencia de la observación (i) a cada una de las componentes. Es decir:

$$z_{ik} = \begin{cases} 1, & \text{si } i \in k \\ 0, & \text{si } i \notin k \end{cases}$$

Bajo el supuesto de independencia entre observaciones, es válido asumir que las ($\mathbf{z}_i$) siguen una distribución multinomial de tamaño uno:

$$\mathbf{z}_i \underset{\text{ind}}{\sim} \text{Multinomial}(\pi_1, \ldots, \pi_{K-1})$$

Dada la ausencia de este vector de etiquetas, el vector de datos observados:

$$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$$

se considera incompleto, y la verosimilitud de una observación completa se expresa de la siguiente forma:

$$\begin{aligned} f(\mathbf{y}_i, \mathbf{z}_i) &= f(\mathbf{z}_i) f(\mathbf{y}_i|\mathbf{z}_i) \\ &= \prod_k \pi_k^{z_{ik}} f_k(\mathbf{y}_i)^{z_{ik}} \end{aligned}$$

La log-verosimilitud de la muestra completa resulta:

$$\log L(\mathbf{y}, \mathbf{z}, \boldsymbol{\Psi}) = \sum_i \sum_k z_{ik} \cdot \{\log \pi_k + \log f_k(\mathbf{y}_i, \boldsymbol{\theta}_k)\}$$

El algoritmo EM procede aplicando iterativamente dos pasos: E de esperanza y M de maximización.

Paso E

En la iteración $(m+1)$, el paso E consiste en calcular la esperanza condicional de la verosimilitud completa respecto del vector de datos observados, con el vector de parámetros obtenido en la iteración anterior:

$$E\{\log L(\mathbf{z}|\mathbf{y}, \boldsymbol{\Psi}^{(m)})\}$$

Dado que la verosimilitud es lineal en (z_{ij}), el paso E se reduce a calcular una sola esperanza condicional:

$$\begin{aligned} E(z_{ik}|\mathbf{y}_i, \boldsymbol{\Psi}^{(m)}) &= P(z_{ik} = 1|\mathbf{y}_i, \boldsymbol{\Psi}^{(m)}) \\ &= \frac{P(z_{ik} = 1) f(\mathbf{y}_i|z_{ik} = 1, \boldsymbol{\Psi}^{(m)})}{f(\mathbf{y}_i, \boldsymbol{\Psi}^{(m)})} \\ &= \frac{\pi_k f_k\left(\mathbf{y}_i, \boldsymbol{\theta}_k^{(m)}\right)}{\prod_k \pi_k f_k\left(\mathbf{y}_i, \boldsymbol{\theta}_k^{(m)}\right)} = \tau_{ik}^{(m+1)} \end{aligned}$$

Los $(\tau_{ik}^{(m+1)})$ pueden interpretarse como la probabilidad a posteriori que tiene la observación (i) de pertenecer a la componente (k) en la iteración $(m+1)$. De esta manera, el paso E radica en actualizar la probabilidad de pertenencia de las observaciones con el nuevo vector de parámetros obtenido en la iteración anterior.

Paso M

En la iteración $(m+1)$, el paso M consiste en maximizar la esperanza condicional obtenida en el paso E respecto del vector de parámetros para obtener $(\boldsymbol{\Psi}^{(m)})$:

$$\frac{\partial}{\partial \boldsymbol{\Psi}} \sum_i \sum_k \tau_{ik}^{(m+1)} \cdot \{\log \pi_k + \log f_k(\mathbf{y}_i, \boldsymbol{\theta}_k)\} = \mathbf{0}$$

El paso M reside en actualizar el valor de los parámetros con las nuevas probabilidades de pertenencia derivadas en el paso anterior.

Los pasos E y M se repiten de manera alternada hasta que el cambio en la verosimilitud reducida es despreciable (menor que una constante arbitrariamente pequeña ϵ):

$$L(\mathbf{y}, \boldsymbol{\Psi}^{(m+1)}) - L(\mathbf{y}, \boldsymbol{\Psi}^{(m)}) < \epsilon$$

Se puede demostrar que el algoritmo EM tiene convergencia monótona, es decir, la verosimilitud en cualquier iteración es siempre mayor respecto de la iteración anterior.

Elección de una solución consistente

El proceso de obtención de los estimadores MV en los modelos mezcla no es un problema trivial, porque presenta dificultades que los modelos clásicos no tienen:

- La función de verosimilitud puede presentar múltiples raíces asociadas a máximos locales, por lo que surge el problema de identificar una solución para definir el estimador MV del vector de parámetros.
- La verosimilitud de la muestra puede ser no acotada, lo que implica que bajo ese escenario el estimador MV es un máximo local y no global.
- Algunas raíces pueden ser soluciones espurias, las cuales responden a patrones localizados en vez de a una estructura subyacente que conforma los grupos.

En la práctica, la estrategia que se adopta para ajustar un modelo mezcla es partir de valores iniciales distintos con la intención de captar la mayor cantidad de raíces posibles. De ese conjunto de soluciones, se descartan las consideradas espurias y se elige la que maximiza la verosimilitud.

Si el algoritmo cae en una región donde la verosimilitud tiende a infinito, su detección es automática porque el proceso diverge. Las soluciones espurias son más difíciles de identificar, pero suelen tomar valores en las fronteras del espacio paramétrico y conformar clases pequeñas.

Objetivos

Este trabajo tiene por objetivo dar respuesta a algunos interrogantes en torno a las soluciones locales que se obtienen al modelar datos longitudinales con clases latentes. Se pretende analizar su frecuencia, los factores asociados con su aparición, y su impacto en el análisis de los resultados.

Metodología

Para dar respuesta a los objetivos planteados se simularon datos longitudinales de tres poblaciones distintas, cuya evolución a través del tiempo se encuentra completamente especificada por un modelo lineal mixto conocido. Se consideraron r = 100 muestras diferentes para tener en cuenta la aleatoriedad de la respuesta en la conformación del conjunto de datos y de esta manera robustecer el análisis.

La convergencia del proceso de estimación de los parámetros del modelo está directamente relacionada con el conjunto de valores iniciales que se toma como punto de partida del algoritmo. Se simularon v = 100 conjuntos de valores iniciales distintos distribuidos uniformemente en un rango pre especificado (Hipp, John R. y Bauer, Daniel J. 2006). Variando el punto de partida, el algoritmo puede explorar la superficie multidimensional de la verosimilitud captando soluciones locales.

Sobre cada muestra y para cada conjunto de valores iniciales se ajustó un modelo mixto correctamente especificado con K = 3 clases latentes, ignorando la pertenencia a cada una de las poblaciones.

La clasificación de las soluciones se hizo en base a la comparación de la verosimilitud al punto de convergencia entre los distintos valores de partida. Para cada conjunto de datos, las soluciones que alcanzaron la verosimilitud máxima se consideraron soluciones MV, y las que alcanzaron la convergencia con una verosimilitud inferior se consideraron soluciones locales.

Como primer análisis descriptivo, se tabuló la frecuencia de soluciones locales y MV

para cada muestra y para cada valor inicial por separado.

Para medir el efecto de las condiciones de partida del proceso de ajuste, se graficó la frecuencia de soluciones locales en función de los valores iniciales de parámetros específicos como el tamaño de las clases.

Con el propósito de medir el impacto de las soluciones locales en el análisis, se compararon las clases conformadas a partir de soluciones locales respecto de las conformadas a partir de soluciones MV. Además, se reajustaron los modelos con k = 2 y 4 clases para contrastar el proceso de recuento de clases en presencia y ausencia de soluciones locales.

Todos los análisis se llevaron a cabo con el software R. Los datos fueron simulados con la función *mvrnorm* del paquete *MASS*, y los modelos fueron ajustados con la función *hlme* del paquete *lcmm* (Proust-Lima, Cécile, Philipps, Viviane, y Liquet, Benoit 2017).

Resultados

La frecuencia de aparición de soluciones locales mostró una amplia variabilidad no sólo entre valores iniciales, sino también entre muestras (Figura 1). Esto último es señal de que existen configuraciones particulares de los datos que favorecen la aparición de esta clase de soluciones.

	solucion		
repeticion	div	local	mv
3	3	0	97
36	4	0	96
83	6	0	94
5	1	1	98
⋮	⋮	⋮	⋮
72	6	25	69
78	5	29	66
8	4	40	56
44	2	42	56

	solucion		
valor ini	div	local	mv
18	0	0	100
59	0	0	100
64	0	0	100
95	0	0	100
⋮	⋮	⋮	⋮
50	9	33	58
26	17	35	48
72	3	43	54
88	13	44	43

Figura 1: Frecuencia de las soluciones por muestra (izq.) y por valor inicial (der.)

Los valores iniciales que partieron de clases muy pequeñas tuvieron mayor convergencia en soluciones locales en comparación con los que partieron de clases balanceadas (Figura 2). Esto es un resultado interesante teniendo en cuenta que las clases verdaderas eran desbalanceadas.

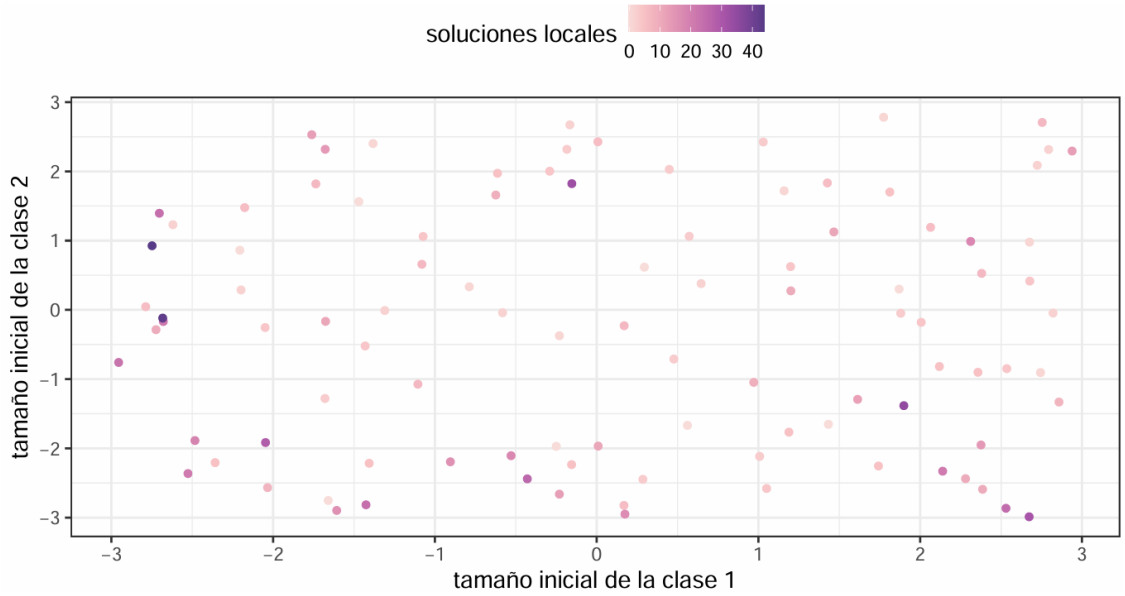

Figura 2: Frecuencia de las soluciones por tamaño inicial de las clases en escala logit

La gran mayoría de las soluciones locales subestimó el tamaño de la clase minoritaria y sobreestimó el tamaño de la clase predominante (Figura 3), lo cual aporta información relevante para la detección de esta clase de soluciones.

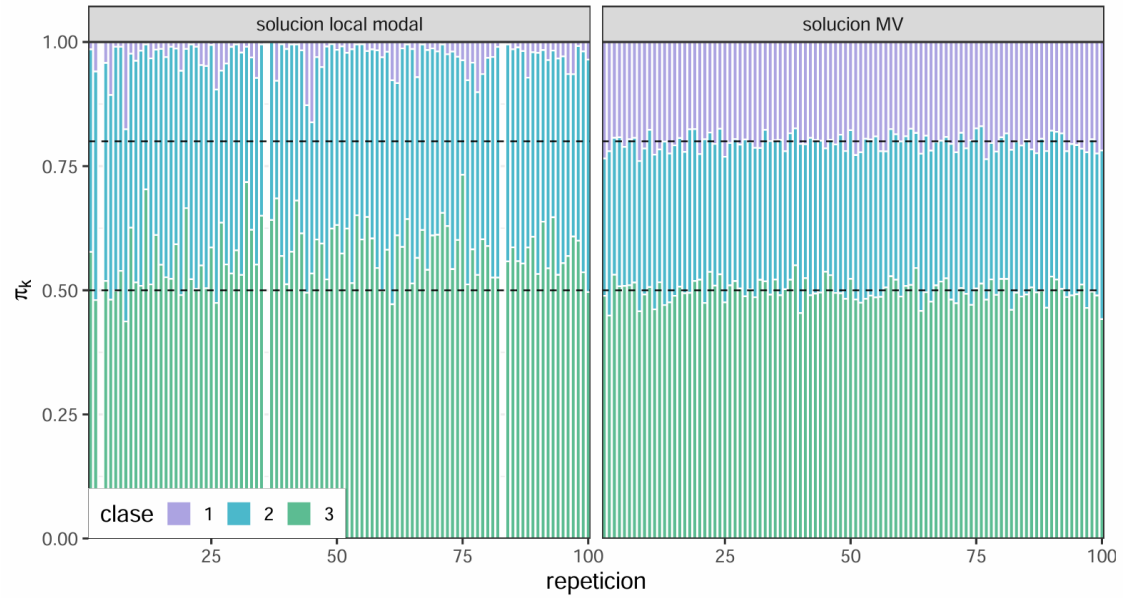

Figura 3: Efecto de las soluciones en el tamaño de las clases

El análisis de conformación de las clases de una solución local particular hace al entendimiento de la naturaleza de estas soluciones (Figura 4). En este agrupamiento particular, cuatro observaciones con mediciones muy similares formaron una clase espuria que distorsionó la clasificación de las observaciones ocultando una de las clases teóricas.

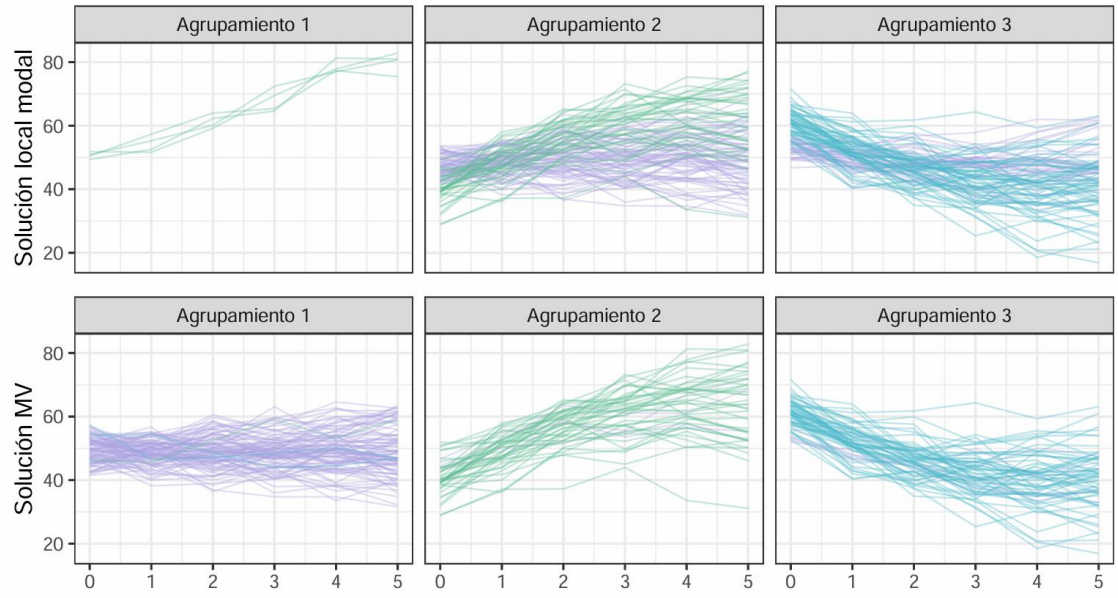

Figura 4: Efecto de las soluciones en la conformación de clases

Discusión final

Los resultados del análisis demuestran que las soluciones locales están presentes incluso en aplicaciones con conjuntos de datos reducidos, registros completos, modelos simples y bajo el cumplimiento de los supuestos distribucionales. Estas soluciones pueden alterar severamente la conformación de las clases y la interpretación de los resultados, frecuentemente sacrificando tendencias generales por patrones específicos y conductas aisladas no representativas.

El uso de rutinas automatizadas de búsqueda de soluciones es imperativo para la validez del análisis. Los agrupamientos conformados, el tamaño relativo de las clases y sus matrices de covariancia deben ser inspeccionados cuidadosamente para descartar la presencia de clases espurias y soluciones subóptimas.

Referencias Bibliográficas

Dempster, A. P., N. M. Laird, y D. B. Rubin. 1977. «Maximum Likelihood from Incomplete Data via the EM Algorithm», Journal of the Royal Statistical Society, 1-38.

Fitzmaurice, Garrett M., Laird, Nan M., y Ware, James H. 2018. *Applied Longitudinal Analysis*. John Wiley & Sons, Incorporated.

Hipp, John R., y Bauer, Daniel J. 2006. «Local Solutions in the Estimation of Growth Mixture Models», Psychological Methods, 36-53.

McLachlan, Geoffrey, y David Peel. 2000. *Finite Mixture Models*. Wiley Series en Probability & Statistics. John Wiley & Sons, Inc.

Proust-Lima, Cécile, Philipps, Viviane, y Liquet, Benoit. 2017. «Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm», Journal of Statistical Software.