Vigilada Mineducación

Universidad Del Norte
División de Ciencias Básicas
Departamento de Matemáticas y Estadística
Programa de Matemáticas

Supuestos y Limitaciones en los Modelos de Clases Latentes

Proyecto de grado

Dylan Samuel Cantillo Arrieta
Barranquilla-Colombia
2024

Director: Dr.rer.nat Humberto Llinás Solano
División de Ciencias Básicas
Departamento de Matemáticas y Estadística
Universidad Del Norte

Abstract

Esta tesis realiza una revisión sistemática de la literatura sobre los Modelos de Clases Latentes (LCM), centrado en el análisis de sus supuestos y limitaciones. A partir de una búsqueda inicial de más de 2000 artículos, se seleccionaron aproximadamente 40 relevantes mediante criterios específicos, como calidad, relevancia y actualidad de la investigación. Se destacan y profundiza cuatro supuestos clave y cuatro limitaciones teórico-prácticas. A partir de lo abordado en la literatura, se habla sobre la importancia y consecuencias de estas, con propuestas estratégicas para mejorar la eficiencia computacional haciendo énfasis en la importancia de validar los supuestos para optimizar la aplicabilidad de los LCM en la toma de decisiones basada en datos.

Palabras Claves— modelo de clases latentes, supuestos, limitaciones, clase latente, variable latente

This thesis conducts a systematic literature review on Latent Class Models (LCM), focusing on the analysis of their assumptions and limitations. From an initial search of over 2,000 articles, approximately 40 relevant studies were selected using specific criteria, such as research quality, relevance, and timeliness. Four key assumptions and four theoretical-practical limitations are highlighted and examined in depth. Based on the literature review, the importance and consequences of these elements are discussed, with strategic proposals to improve computational efficiency, emphasizing the importance of validating assumptions to optimize the applicability of LCMs in data-driven decision-making.

Keywords— latent class model, assumptions, limitations, latent class, latent variable

CAPÍTULO 1

Introducción

En la actualidad, la capacidad para analizar y comprender grandes volúmenes de datos se ha convertido en una competencia esencial en diversas disciplinas. Los avances tecnológicos han facilitado la recolección masiva de información, pero también han generado un desafío significativo como extraer patrones útiles y significativos de los conjuntos de datos. Aquí es donde los modelos estadísticos, especialmente los de clases latentes, juegan un papel crucial. Estos modelos permiten descubrir estructuras ocultas dentro de los datos, revelando relaciones subyacentes que no son inmediatamente visibles.

La relevancia de los modelos de clases latentes trasciende distintos campos, desde el análisis de comportamiento del consumidor en el marketing, hasta la identificación de grupos de riesgo en estudios de salud pública. La capacidad para categorizar de manera eficiente a las personas o eventos con base en características comunes puede tener un impacto profundo en la toma de decisiones estratégicas. Sin embargo, la aplicación de estos modelos plantea preguntas sobre su precisión, la validez de los grupos identificados y los supuestos estadísticos que los sustentan.

Esta tesis tiene como objetivo explorar la teoría y la aplicación de los modelos de clases latentes, haciendo una revisión de su evolución y los retos asociados a su uso. A medida que la sociedad enfrenta la creciente complejidad de interpretar grandes cantidades de datos, este estudio busca contribuir al entendimiento de estas herramientas, analizando tanto su potencial como sus limitaciones. Además, no solo pretende esclarecer los fundamentos teóricos, sino también evaluar cómo estos modelos pueden ser aplicados de manera efectiva en contextos prácticos, aportando a la mejora en la toma de decisiones basada en datos.

El análisis y desarrollo de este proyecto es pertinente en un mundo que exige soluciones precisas y eficaces para enfrentar problemas cada vez más complejos. Posicionándose en la intersección entre la teoría matemática avanzada y la utilidad práctica, ofreciendo una reflexión profunda sobre cómo los modelos de clases latentes pueden ser una herramienta esencial en el análisis de datos y la toma de decisiones estratégicas en múltiples áreas.

CAPÍTULO 2

Antecedentes

Las clases latentes tienen su origen en los trabajos de Paul Lazarsfeld durante la década de 1950 [Lazarsfeld, 1950], quien introdujo el concepto de “variables latentes"para explicar las relaciones subyacentes entre variables observadas en análisis de datos categóricos. Esta innovación permitió abordar de manera más precisa la estructura de los datos complejos, especialmente en el contexto de encuestas y estudios sociales.

Durante los 60’s, Lazarsfeld, junto con Henry, desarrollaron formalmente el modelo de clases latentes para el análisis de datos de encuestas [Lazarsfeld, 1968], marcando un hito significativo en la metodología de investigación social. Este modelo permitió identificar grupos no observables dentro de una población, facilitando la comprensión de patrones subyacentes en los datos.

El desarrollo de los modelos de clases latentes continuó en la década de 1970 y 1980 con los trabajos de Goodman [Goodman, 1974], quien extendió la metodología a variables latentes continuas, ampliando así su aplicabilidad. Posteriormente, junto con Clogg, contribuyeron a popularizar el uso de estos modelos en las ciencias sociales [Clogg y Goodman, 1984], impactando la investigación empírica.

A partir de 1990, los modelos de clases latentes comenzaron a ser aplicados en una variedad de disciplinas, incluyendo psicología, educación, marketing y epidemiología [McCutcheon, 1987, Hagenaars y McCutcheon, 2002]. Estos modelos demostraron ser herramientas valiosas para identificar subpoblaciones ocultas y comprender mejor las dinámicas internas de los datos.

En la actualidad, los modelos de clases latentes se utilizan ampliamente en diversas disciplinas, con importantes extensiones a modelos longitudinales, multinivel y de mezclas finitas [Collins y Lanza, 2009, Vermunt, 2002]. La investigación metodológica en este campo sigue siendo activa, enfocándose en aspectos como la selección de modelos, el diagnóstico y las aplicaciones computacionales [Nylund-Gibson y Choi, 2018], consolidando así su relevancia en la estadística moderna y en el análisis de datos complejos.

Se han desarrollado guías para la práctica óptima del análisis de clases latentes, describiendo los elementos clave a considerar y proporcionando ejemplos de aplicación [Weller *et al.*, 2020]. Además, se han identificado consideraciones metodológicas y errores comunes en la aplicación de LCA, especialmente en campos como la medicina crítica y respiratoria [Sinha *et al.*, 2020].

Asimismo, los modelos de elección de clases latentes han integrado indicadores actitudinales mediante redes neuronales artificiales, mejorando la segmentación y la capacidad de capturar características comportamentales complejas [Lahoz *et al.*, 2023].

CAPÍTULO 3

Marco conceptual

3.1. Estadística fundamental

Para comprender los modelos de clases latentes, es fundamental entender algunos conceptos estadísticos (Tabla 3.1). Con estos conceptos se lograra avanzar en el análisis e interpretación de modelos complejos que facilitan la comprensión de las clases latentes dentro de un conjunto de datos, sus supuestos y limitaciones [Hagenaars y McCutcheon, 2002, Casella y Berger, 2024, Hastie *et al.*, 2009].

Concepto estadístico	**Descripción**
LCM	Modelo de clases latente (Latent Class Model).
LCA	Análisis de clases lantente (Latent Class Analysis).
Categórico	Valor discreto que representan categorías mutuamente excluyentes.
Ordinal	Dato categórico donde existe un orden inherente, como las escalas de opinión.
Continuo	Valor dentro de un rango determinado, incluyendo decimales o fracciones.
Longitudinal	Observación de los mismos individuos a través del tiempo.
Multinivel	Estructura jerárquica que permite el análisis de datos anidados, como estudiantes dentro de escuelas.
Mezcla finita	Modelo en los que la población se representa como una mezcla de varias subpoblaciones.
Covariable	Variable adicional a los datos que se agrega a un modelo estadístico para explicar o ajustar la relación entre algunas variables de interés.

Cuadro 3.1: Conceptos estadísticos y sus respectivas descripciones

3.1.1. Supuesto de normalidad

La normalidad en estadística es el supuesto donde los datos se distribuyen siguiendo una distribución normal, también conocida como distribución de Gauss o campana de Gauss (Figura 3.1) [Casella y Berger, 2024]. Esta distribución es simétrica alrededor de su media, con una forma de campana. La normalidad es un requisito

importante en muchos análisis estadísticos, como en pruebas t, regresión y ANOVA, ya que garantiza que los resultados sean válidos y generalizables.

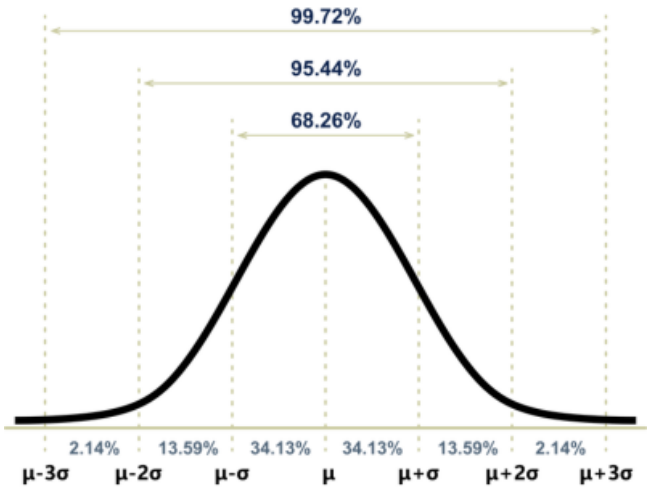

Figura 3.1: Distribución normal

Si los datos no cumplen con este el supuesto de normalidad, se pueden aplicar transformaciones para aproximar la normalidad o utilizar métodos no paramétricos que no dependen de este supuesto.

Cuando hablamos de desviaciones, en la normalidad, se refieren a las diferencias entre los valores individuales y la media del conjunto de datos. En una distribución normal, aproximadamente el 68 % de los datos se encuentran dentro de una desviación estándar de la media, el 95 % dentro de dos desviaciones estándar, y el 99.7 % dentro de tres desviaciones estándar. Cuando hay desviaciones significativas de la normalidad (como asimetría o curtosis), el supuesto de normalidad puede no cumplirse, lo que afecta la validez de ciertos análisis y sugiere que los datos podrían estar distribuidos de otra manera [Maity y Saha, 2023].

3.2. Introducción a los modelos de clases latentes

Se presentan los principios matemáticos fundamentales y las herramientas necesarias para describir la teoría relacionada con los modelos de clases latentes [Gelman *et al.*, 1995, Hastie *et al.*, 2009, Klemens, 2008].

3.2.1. Independencia condicional

Una de las bases de los modelos de clases latentes es el supuesto de independencia condicional, que permite simplificar la representación de la relación entre las variables observadas dado el conocimiento de la clase latente.

Definition 3.1 (Independencia condicional)**.** *Sea C una variable de clase latente y X_i, X_j variables observadas, para cualquier par X_i y X_j, y para $C = k$, tenemos [Chow y Teicher, 2012]*

$$P(X_i, X_j \mid C = k) = P(X_i \mid C = k)P(X_j \mid C = k)$$

Permite que la distribución conjunta de las variables observadas sea calculada de forma eficiente, especialmente cuando la dimensionalidad de los datos es alta, conceptos que se explicarán a continuación.

3.2.2. Medidas de evaluación y ajuste de modelos

Para evaluar la efectividad de un modelo de clases latentes, se utilizan métricas de clasificación como la entropía, la precisión. Y para evaluar y mejorar la calidad del ajuste del modelo se utilizan herramientas como residuos bivariados e índice de modificación.

Definition 3.2 (Entropía)**.** *La entropía mide la incertidumbre en la asignación de una observación a una clase específica. Para la observación i, su entropía se define como*

$$H_i = -\sum_{k=1}^{K} P(C = k \mid \mathbf{x}_i) \log P(C = k \mid \mathbf{x}_i),$$

donde $P(C = k \mid \mathbf{x}_i)$ es la probabilidad posterior de que la observación i pertenezca a la clase k.

Definition 3.3 (Precisión)**.** *La precisión representa la proporción de observaciones correctamente asignadas a su clase latente o al grupo más probable. Dada una muestra con asignaciones observadas $\{c_i\}$ y predichas $\{\hat{c}_i\}$*

$$\textit{Precisión} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(c_i = \hat{c}_i),$$

donde $\mathbb{I}(c_i = \hat{c}_i)$ indica 1 si la asignación predicha coincide con la real y 0 en caso contrario.

Definition 3.4 (Residuos bivariados)**.** *Para un par de variables observadas (X_i, X_j), el residuo bivariado se define como*

$$\begin{aligned} r_{ij}(x_i, x_j) = {} & P(X_i = x_i, X_j = x_j) \\ & - \sum_{y=1}^{k} P(X_i = x_i \mid Y = y) P(X_j = x_j \mid Y = y) P(Y = y) \end{aligned}$$

Un residuo bivariado significativo indicaría una dependencia no explicada por el modelo entre X_i y X_j, sugiriendo que el supuesto de independencia condicional entre X_i y X_j dado Y podría ser incorrecto.

Para cuantificar la significancia de los residuos bivariados, se emplea la siguiente estadística de prueba

$$Z_{ij} = \frac{r_{ij}(x_i, x_j)}{\sigma_{ij}(x_i, x_j)}$$

donde $\sigma_{ij}(x_i, x_j)$ es la desviación estándar de $r_{ij}(x_i, x_j)$.

Definition 3.5 (Índice de modificación)**.** *El índice de modificación indica cuánto mejoraría el ajuste de un modelo si se permitiera una nueva relación (p. ej., una covarianza entre errores o un parámetro adicional). Se calcula como el cambio esperado en el valor de chi-cuadrado de ajuste al añadir el parámetro sugerido. Matemáticamente, si q es el nuevo parámetro, el índice de modificación M_q es*

$$M_q = \frac{\partial \chi^2}{\partial q}$$

Un valor alto de M_q sugiere que el parámetro adicional podría mejorar sustancialmente el ajuste del modelo.

3.2.3. Método de imputación múltiple

La imputación múltiple permite manejar datos faltantes creando múltiples conjuntos de datos completos mediante simulación. Cada conjunto es analizado independientemente y los resultados se combinan para obtener estimaciones finales robustas [Lee *et al.*, 2020].

1. *Markov Monte Carlo (MCMC):* este método utiliza simulaciones de cadenas de Markov para generar imputaciones basadas en la distribución posterior de los datos observados, especialmente útil en situaciones con distribuciones complejas.
2. *Imputación múltiple basada en regresión cuantílica bayesiana (BQR):* la imputación basada en BQR ajusta distribuciones condicionales para cada cuantil, proporcionando imputaciones robustas en presencia de valores extremos o datos asimétricos, utilizando un enfoque bayesiano para cada cuantil específico.

3.3. Métodos de estimación de parámetros

La verosimilitud y la estimación de máxima verosimilitud (MLE) son esenciales para estimar los parámetros del modelo de clases latentes [Hagenaars y McCutcheon, 2002].

Definition 3.6 (Función de verosimilitud)**.** *La función de verosimilitud de un modelo de clases latentes para datos observados* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ *fijos y parámetros* θ *se define como*

$$L(\theta; \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N} \sum_{k=1}^{K} P(\mathbf{x}_i \mid C = k; \theta) P(C = k; \theta).$$

Debido a la complejidad computacional de esta función, el algoritmo de Expectation-Maximization es comúnmente utilizado para optimizarla.

Theorem 3.1 (Máxima verosimilitud)**.** *Los parámetros* θ *que maximizan la verosimilitud,* $\hat{\theta}_{ML}$ *[Hagenaars y McCutcheon, 2002], satisfacen*

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta; \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N).$$

El algoritmo de Expectación-Maximización (EM) se emplea para encontrar estos estimadores en presencia de variables latentes.

Definition 3.7 (Algoritmo Expectation-Maximization)**.** *El algoritmo Expectation-Maximization (EM) es un método iterativo para optimizar la verosimilitud en modelos con variables latentes. Los pasos para implementarlo son los siguientes*

1. *Paso de expectación (E-step): calcula la probabilidad posterior de cada clase latente* $C = k$ *para cada observación* $\mathbf{x}_i$*.*
2. *Paso de maximización (M-step): maximiza la verosimilitud completa actualizando los parámetros* θ *con las probabilidades calculadas en el E-step.*

Definition 3.8 (Razón de verosimilitud bootstrapped)**.** *La razón de verosimilitud bootstrapped (BLRT) es una prueba estadística utilizada para comparar la bondad de ajuste de dos modelos anidados, es decir, un modelo más complejo y uno más simple. La BLRT evalúa si el aumento en la verosimilitud al pasar del modelo más simple al modelo más complejo es significativo. Formalmente, la estadística de razón de verosimilitud se define como*

$$LR = 2\left(L(\hat{\theta}_1; \mathbf{x}) - L(\hat{\theta}_0; \mathbf{x})\right)$$

donde $L(\hat{\theta}_1; \mathbf{x})$ *es la verosimilitud máxima para el modelo más complejo y* $L(\hat{\theta}_0; \mathbf{x})$ *es la verosimilitud máxima para el modelo más simple. Dado que la distribución de la LR bajo la hipótesis nula puede ser desconocida en modelos de clases latentes, se utiliza el método bootstrap para aproximar su distribución, generando múltiples muestras bootstrap a partir de los datos observados y calculando la razón de verosimilitud en cada muestra. Este proceso permite construir un intervalo de confianza para la LR y realizar inferencias sobre la significancia del modelo complejo.*

3.4. Criterios de selección de modelos

Los criterios de selección, como el AIC y el BIC, ayudan a determinar el número óptimo de clases.

Definition 3.9 (Akaike Information Criterion)**.** *El Akaike Information Criterion (AIC) equilibra la bondad de ajuste del modelo con su simplicidad y se define como*

$$AIC = -2\log L(\hat{\theta}) + 2p$$

Definition 3.10 (Bayesian Information Criterion)**.** *El Bayesian Information Criterion (BIC) introduce una penalización dependiente del tamaño de muestra, calculado como*

$$BIC = -2\log L(\hat{\theta}) + p\log(N)$$

Definition 3.11 (Sample-Size Adjusted BIC)**.** *El Criterio de Información Bayesiano ajustado por el tamaño de la muestra (saBIC) es una versión modificada del BIC que introduce una penalización ajustada específicamente para tamaños de muestra más pequeños.*

$$saBIC = -2\log L(\hat{\theta}) + p\log\left(\frac{N+2}{24}\right)$$

donde $L(\hat{\theta})$ es la verosimilitud máxima del modelo, p es el número de parámetros, y N es el tamaño de la muestra. En comparación con el BIC tradicional, el saBIC disminuye la penalización en función del tamaño de la muestra, lo cual es particularmente útil cuando N es pequeño, ya que el BIC tiende a penalizar de manera más estricta en tales casos.

3.5. Introducción a los modelos de clases latentes

Se abordan varios conceptos desde una perspectiva matemática, describiendo su relevancia y conexión con los modelos de clases latentes (LCM). Este enfoque permitirá profundizar en los supuestos y limitaciones, propiedades y conceptos que fundamentan el análisis de clases latentes y, por ende, la formulación de estos modelos.

Definition 3.12 (Variables observables)**.** *Son aquellas que se pueden medir o registrar directamente en un conjunto de datos, siendo la fuentes principal de información empírica de los modelos de clases latentes.*

Definition 3.13 (Variables latentes)**.** *Directamente relacionado con las variables observables, las variables latentes son aquellas que no se observan directamente, pero que explican la estructura de correlación o patrón subyacente entre las variables observables. En los LCM, las variables latentes son categóricas y representan clases o grupos subyacentes.*

Si denotamos la variable latente como C, que puede tomar valores en un conjunto discreto de clases $\{C_1, C_2, \ldots, C_k\}$, el objetivo es modelar $P(\mathbf{X}, C)$ para inferir la probabilidad de pertenencia de cada observación a cada clase latente.

Definition 3.14 (Perfil)**.** *Un perfil representa una combinación específica de valores de las variables observables asociadas a una clase latente. El perfil describe una configuración probable de respuestas o valores de las variables observables condicionadas a una clase latente específica.*

Si una clase latente $C = c$ implica un perfil particular $\mathbf{x}_c$ de las variables observables, el modelo puede describir la probabilidad de que una observación siga ese perfil

$$P(\mathbf{X} = \mathbf{x}_c | C = c)$$

Este perfil facilita la interpretación de las clases latentes, pero también supone que cada clase corresponde a una estructura homogénea de respuestas o características, lo cual puede ser una limitación en sistemas complejos o dinámicos.

3.6. Clase latente

Definition 3.15 (Clase latente)**.** *Una clase latente lo podemos definir como un concepto estadístico utilizado para representar grupos o categorías no observables dentro de una población, en función de las observaciones. A diferencia de las clases observables, que se miden y documentan directamente, las clases latentes son categorías inferidas que no se representan directamente en los datos. En los LCM, estas clases actúan como una variable discreta subyacente que permite describir la estructura oculta de las observaciones.*

Definition 3.16 (Definición formal de una clase latente)**.** *Sea C una variable discreta no observable que toma valores en el conjunto $\{C_1, C_2, \ldots, C_k\}$, donde cada C_i representa una clase latente específica. Esta variable latente tiene una distribución de probabilidad que describe la proporción de la población que se espera pertenezca a cada clase. De este modo, la probabilidad de que una observación pertenezca a la clase latente C_i se denota como*

$$P(C = C_i) = \pi_i, \quad \text{con } i = 1, 2, \ldots, k$$

donde π_i representa el peso o proporción de la clase C_i en la población. Ya que los π_i forman una distribución de probabilidad, estos pesos satisfacen la condición de normalización

$$\sum_{i=1}^{k} \pi_i = 1$$

3.6.1. Interpretación probabilística

Desde una perspectiva probabilística, una clase latente permite modelar la heterogeneidad de la población dividiéndola en clases donde cada subconjunto tiene características distintivas que representan la variabilidad en las variables observables. Al clasificar los individuos en clases latentes, los LCM pueden asignar una probabilidad de pertenencia para cada observación respecto a cada clase latente. Esta probabilidad, denominada probabilidad posterior, se expresa como

$$P(C = C_i | \mathbf{X} = \mathbf{x})$$

donde $\mathbf{x}$ representa un vector de valores observados para $\mathbf{X}$. Este cálculo permite asignar cada observación a la clase latente más probable, aunque en algunos casos, la pertenencia puede estar distribuida entre varias clases si las probabilidades posteriores son similares.

3.6.2. Ejemplo aplicado

Consideremos un modelo con dos variables observables binarias, X_1 y X_2, y una variable de clase latente C con dos clases posibles, C_1 y C_2. Las probabilidades condicionales para X_1 y X_2 dada cada clase pueden definirse como

$$P(X_1 = 1 | C = C_i) = \theta_{1,i} \quad \text{y} \quad P(X_2 = 1 | C = C_i) = \theta_{2,i}, \quad \text{para } i = 1, 2$$

Aquí, $\theta_{1,1}$, $\theta_{1,2}$, $\theta_{2,1}$ y $\theta_{2,2}$ representan los parámetros que definen las probabilidades condicionales de cada variable observable dentro de cada clase latente. La probabilidad conjunta de observar un par (X_1, X_2) en función de la clase latente se expresa como:

$$P(X_1 = x_1, X_2 = x_2) = \sum_{i=1}^{2} P(X_1 = x_1 | C = C_i) \cdot P(X_2 = x_2 | C = C_i) \cdot P(C = C_i)$$

Este ejemplo muestra cómo las clases latentes introducen una estructura de mezcla que explica las observaciones en términos de componentes subyacentes.

Definition 3.17 (Estructura latente)**.** *La estructura latente hace referencia a la relación y dependencia entre las clases latentes y las variables observables. Formalmente, el conjunto de distribuciones condicionales que relacionan las variables observables con las clases latentes.*

3.7. Modelo de clases latentes

El modelo de clases latentes (LCM) es una técnica estadística que se utiliza para clasificar observaciones en grupos no observables o clases latentes, basándose en un conjunto de variables observadas. Este enfoque resulta particularmente útil en situaciones en las que se desea identificar patrones o estructuras subyacentes en los datos, especialmente cuando las clases a las que pertenecen las observaciones no son directamente observables, sino inferidas a partir de su comportamiento o características observables.

Definition 3.18 (Definición formal del modelo de clases latentes)**.** *Sea* $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ *un vector de variables observadas, donde cada variable* X_i *puede ser discreta o continua. Se asume que existe una variable latente* C *que toma valores en un conjunto finito de clases* $\{1, 2, \ldots, K\}$*. La variable* C *representa la clase latente a la cual pertenece cada observación. Formalmente, el modelo de clases latentes plantea que la distribución conjunta de las variables observadas se puede expresar como una combinación de distribuciones condicionadas a la variable latente* C

$$P(\mathbf{X} = \mathbf{x}) = \sum_{k=1}^{K} P(\mathbf{X} = \mathbf{x} \mid C = k) P(C = k)$$

La ecuación anterior representa una mezcla de distribuciones, donde cada componente $P(\mathbf{X} = \mathbf{x} \mid C = k)$ *describe la distribución de* $\mathbf{X}$ *dada una clase específica* k*, y* $P(C = k)$ *es la probabilidad a priori de que una observación pertenezca a la clase* k*.*

3.7.1. Algunos modelo de clases latentes

Existen diversos tipos de modelos de clases latentes que amplían la estructura básica del LCM para abordar diferentes tipos de relaciones entre las variables y los datos observados. Entre estos, destacan los siguientes

- **Modelos jerárquicos**: estructuran las clases en niveles jerárquicos, lo cual permite representar subgrupos dentro de subgrupos. Este enfoque es útil en situaciones donde se espera una organización multinivel de los datos. Por ejemplo, en estudios de comportamiento, los individuos pueden clasificarse en grupos según características generales, que luego se subdividen en grupos más específicos.

  Formalmente, este modelo jerárquico implica la existencia de múltiples niveles de clases latentes, donde las clases de un nivel dependen de las clases del nivel superior. Esta estructura puede representarse mediante una cadena de probabilidades condicionadas en la forma:

  $$P(Y|C_1, C_2, \ldots, C_L) = \prod_{l=1}^{L} P(Y|C_l, C_{l-1}, \ldots, C_1)$$

- **Modelos de clases latentes restringidos (RLCM)**: son una variante del LCM en la que se imponen restricciones adicionales sobre las relaciones entre variables observables y latentes. Estas restricciones suelen establecerse para simplificar el modelo o para hacerlo más interpretable. Por ejemplo, se pueden fijar ciertos parámetros a cero para indicar que no todas las variables observables dependen de todas las clases latentes.

  Una ventaja clave de los RLCM es que pueden reducir la complejidad del modelo, lo cual facilita su interpretación y ajuste en conjuntos de datos grandes o con múltiples variables.

3.7.2. Parámetros en los modelos de clases latentes

En el contexto de los modelos de clases latentes, un *parámetro* se refiere a una cantidad desconocida que debe estimarse a partir de los datos para caracterizar el modelo. Estos parámetros incluyen

- *Probabilidades de pertenencia a la clase*: representan la probabilidad de que un individuo pertenezca a cada una de las clases latentes.

- *Distribuciones de las variables observables*: dadas las clases latentes, describen la probabilidad condicional de cada respuesta en las variables observables.

La estimación precisa de estos parámetros es esencial para que el modelo represente de forma adecuada la estructura latente del conjunto de datos.

3.7.3. Ajuste de un modelo de clases latentes

El *ajuste de un modelo de clases latentes* consiste en estimar los parámetros del modelo para que los datos observados se ajusten de la mejor manera posible a la estructura postulada. Esto se realiza mediante métodos de estimación como la *máxima verosimilitud* o el *Método de Esperanza-Maximización* (EM). La calidad del ajuste del modelo puede evaluarse mediante criterios como el *Criterio de Información de Akaike* (AIC) o el *Criterio de Información Bayesiano* (BIC), que penalizan modelos más complejos y ayudan a seleccionar el modelo con un balance óptimo entre ajuste y parsimonia.

Un ajuste adecuado implica que el modelo seleccionado es capaz de capturar las relaciones relevantes entre las variables observables y las clases latentes sin sobreajustarse a los datos, lo cual es fundamental para asegurar que los resultados sean generalizables y útiles para análisis posteriores.

Definition 3.19 (Análisis de clases latentes)**.** *El análisis de clases latentes (LCA) es una técnica estadística específica para ajustar y evaluar modelos de clases latentes. A diferencia del LCM, que es el modelo teórico, el LCA se centra en el proceso empírico de ajuste del modelo y en la interpretación de los resultados obtenidos. A través del LCA, los investigadores buscan identificar el número óptimo de clases latentes y asignar probabilidades de pertenencia a cada clase para cada individuo en el conjunto de datos.*

La diferencia fundamental entre el LCM y el LCA radica en que el primero se refiere al modelo teórico, mientras que el segundo implica los procedimientos y herramientas estadísticos utilizados para estimar los parámetros del modelo y evaluar su adecuación.

3.8. Ejemplo ilustrativo

En esta sección, se presenta un ejemplo adaptado del material publicado por GeeksforGeeks [GeeksforGeeks, 2024], que describe detalladamente la implementación de un modelo de clases latentes (LCA, por sus siglas en inglés) utilizando el lenguaje de programación R. Este recurso, ampliamente reconocido por su enfoque pedagógico en herramientas de análisis de datos, ofrece una introducción práctica para identificar subgrupos no observados en una población a partir de datos categóricos. El análisis aquí expuesto se reformula y extiende para resaltar los fundamentos matemáticos del modelo, proporcionando una base conceptual sólida que facilita su integración en el marco teórico de esta tesis.

Planteamiento del problema

Consideremos una encuesta aplicada a una población de 1000 individuos. Cada participante responde a cinco preguntas categóricas, denotadas por Q_1, Q_2, Q_3, Q_4, Q_5. Cada pregunta ofrece un conjunto de opciones de respuesta. Por ejemplo, Q_1 y Q_3 tienen tres opciones posibles $\{1, 2, 3\}$, mientras que Q_2 y Q_4 tienen cuatro opciones posibles $\{1, 2, 3, 4\}$. El objetivo es determinar si existen patrones subyacentes que permitan agrupar a los individuos en clases latentes, lo cual aporta información clave sobre la heterogeneidad de la población.

Modelo matemático

El modelo de clases latentes considera las respuestas de un individuo i a las preguntas j representadas como X_{ij}, donde $i = 1, \ldots, n$ y $j = 1, \ldots, p$. Se asume la existencia de una variable latente C_i que representa la clase

a la que pertenece el individuo i, con $C_i \in \{1, 2, \ldots, k\}$, siendo k el número de clases latentes. Bajo esta formulación, las respuestas X_{ij} son condicionalmente independientes dado C_i, lo cual se expresa matemáticamente como:

$$P(X_{ij} \mid C_i = c) = \prod_{j=1}^{p} P(X_{ij} \mid C_i = c),$$

donde $P(X_{ij} \mid C_i = c)$ es la probabilidad de que el individuo elija la respuesta X_{ij} dado que pertenece a la clase c.

Adicionalmente, se asume que:

$$P(C_i = c) = \pi_c, \quad \text{con } \sum_{c=1}^{k} \pi_c = 1,$$

donde π_c representa la proporción de la población asignada a la clase c.

La probabilidad conjunta de las respuestas de un individuo i se obtiene sumando sobre todas las clases:

$$P(X_i) = \sum_{c=1}^{k} \pi_c \prod_{j=1}^{p} P(X_{ij} \mid C_i = c).$$

Resultados del caso de estudio

Se ajustó un modelo con $k = 3$ clases latentes utilizando datos simulados. Los resultados clave incluyen:

Probabilidades condicionales: Para la pregunta Q_1, las probabilidades de respuesta en cada clase son:

Clase	$P(X_{i1} = 1 \mid C_i = c)$	$P(X_{i1} = 2 \mid C_i = c)$	$P(X_{i1} = 3 \mid C_i = c)$
1	0,2480	0,4188	0,3332
2	0,3746	0,3205	0,3050
3	0,3310	0,2679	0,4011

Distribución de las clases: Las proporciones estimadas de individuos en cada clase son:

$$\pi_1 = 0{,}2754, \quad \pi_2 = 0{,}4101, \quad \pi_3 = 0{,}3144.$$

Evaluación del ajuste: El modelo fue evaluado mediante los siguientes criterios:

- **Criterio de Información de Akaike (AIC):** 12164.5
- **Criterio de Información Bayesiano (BIC):** 12350.99

Estos valores indican que el modelo con $k = 3$ clases ofrece un ajuste razonable.

Interpretación

El análisis sugiere la existencia de tres grupos principales en la población:

- **Clase 1**: Individuos con alta probabilidad de seleccionar la opción 2.
- **Clase 2**: Respuestas más dispersas, sin una preferencia dominante.
- **Clase 3**: Predilección por opciones menos comunes.

CAPÍTULO 4

Metodología

En el vasto universo de la estadística matemática, las clases latentes emergen como un paradigma fascinante, ofreciendo una lente a través de la cual podemos desentrañar estructuras ocultas en datos multivariados. Este apartado delinea la metodología empleada en nuestra revisión sistemática de la literatura, un proceso riguroso diseñado para explorar la estructura formal de las clases latentes, con un énfasis particular en sus fundamentos matemáticos, limitaciones intrínsecas y otros horizontes.

Esta investigación se centra en una revisión meticulosa que combina una perspectiva histórica con un análisis actual en profundidad. Este enfoque dual nos permite no solo entender la evolución del campo a lo largo de más de siete décadas, sino también estudiar con precisión los avances más recientes y las fronteras actuales de este conocimiento.

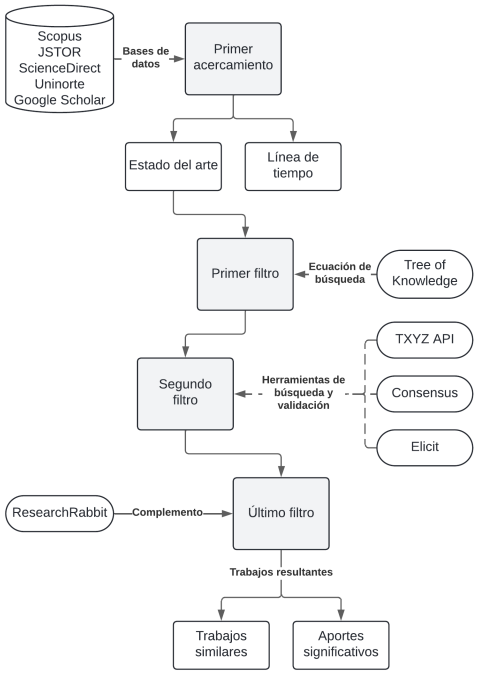

Figura 4.1: Flujo de la metodología

4.1. Criterios de selección para la revisión sistemática de la literatura

La selección de la literatura se rigió por la metodología PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [Liberati *et al.*, 2009], un enfoque sistemático y riguroso que garantiza la transparencia y reproducibilidad de nuestra revisión. El proceso de selección se estructuró en fases secuenciales, cada una diseñada para refinar y focalizar nuestro estudio.

Primer acercamiento

Se realizó una revisión de del estado del arte relacionado con las clases latentes en el campo de las matemáticas, estadística y posteriormente sobre la informática, analizando trabajos desde 1950 hasta la actualidad. Luego, se identificaron más de 2000 trabajos potencialmente relevantes a través de una búsqueda en bases de datos (Scopus, JSTOR, ScienceDirect, bases de datos Uninorte, Google Scholar, etc.), principales en idiomas inglés y español, y restringiendo el rango temporal a los últimos cinco años (2020 - 2024) para capturar investigaciones recientes en el campo.

Ecuación de búsqueda y enfoque teórico

Posteriormente, se aplicaron filtros específicos a esta investigación inicial. Se planteo una ecuación de búsqueda enfocada en la amplitud y precisión

> `"latent class"` AND (`"limitation"` OR `review"` OR `"guide"` OR `"literature"` OR `"suggestion"` OR `recommendation"` OR `"theory"`)

El objetivo fue incluir tanto estudios que abordaran los fundamentos teóricos de las clases latentes como aquellos trabajos que presenten desafíos, críticas constructivas o propuestas de mejora con respecto a las limitaciones/supuestos. Como resultado de este proceso de refinamiento, el número de artículos seleccionados se redujo a un total aproximado de 300.

Lectura inicial

Se llevó a cabo una revisión detallada de los títulos y resúmenes de los trabajos identificados. En esta etapa, se utilizaron criterios de inclusión que requerían que los artículos estuvieran centrados en el desarrollo teórico de los modelos de clases latentes, en especial aquellos que abordaran de manera explícita las limitaciones o supuestos subyacentes, o que propuestas de superación de los desafíos (limitaciones/supuestos). A través de este análisis, el estudio se redujo a 80 trabajos.

Selección final

Finalmente, los artículos preseleccionados fueron sometidos a un análisis exhaustivo del texto, con el fin de determinar su relevancia teórica y rigor matemático. Tras esta revisión integral, los artículos seleccionados se clasificaron en dos categorías: aquellos cuyos enfoques metodológicos o teóricos resultaban similares al de esta investigación, constituyendo un grupo de 15 trabajos, y aquellos que realizaban aportes significativos a la literatura o desarrollo sobre clases latentes, con un total de 41 trabajos.

4.2. Herramientas de búsqueda

Se utilizaron varias herramientas tecnológicas avanzadas con el objetivo de optimizar y complementar el proceso de búsqueda y filtrado de los trabajos a tratar. Estas herramientas no solo aceleraron el proceso de recolección de datos, sino que también mejoraron la calidad y profundidad del análisis. A continuación, se describe el rol específico de cada una de ellas y los beneficios profesionales que aportaron a la revisión sistemática

- **Tree-of-Knowledge**: herramienta especializada en la visualización de relaciones conceptuales entre artículos académicos. Utilizando gráficos y redes, esta aplicación permitió mapear las interconexiones entre investigaciones previas y actuales, ayudando a identificar los estudios seminales y aquellos que han generado más impacto en el campo de las clases latentes.
- **TXYZ API**: plataforma que emplea inteligencia artificial para la búsqueda automatizada de literatura científica. Una de sus ventajas clave es su capacidad para filtrar artículos de manera más precisa mediante el análisis semántico del contenido. Esta herramienta mejoró el proceso de búsqueda al reducir significativamente el ruido (artículos irrelevantes o poco pertinentes) y maximizar la relevancia de los resultados.
- **Consensus**: motor de búsqueda académica impulsado por IA con el objetivo de hacer que el mejor conocimiento del mundo sea más accesible. Facilitando la evaluación comparativa de los estudios seleccionados, destacando los puntos en común entre ellos y señalando discrepancias o vacíos en la literatura sobre las limitaciones y supuestos de las clases latentes.
- **Elicit**: herramienta que se enfoca en ayudar a investigadores a formular preguntas de investigación de manera estructurada y luego encontrar respuestas basadas en la literatura disponible. Proporcionó sugerencias sobre artículos que respondían directamente a las interrogantes teóricas planteadas sobre las clases latente.
- **ResearchRabbit**: plataforma de investigación que le permite descubrir y visualizar literatura y académicos relevantes. Utilizado como un asistente de búsqueda inteligente que permitió descubrir publicaciones adicionales que no habían sido capturadas en las búsquedas iniciales.

Es importante aclarar que los resultados arrojados por las diversas herramientas luego fueron validados analizando los trabajos ofrecidos, uno por uno.

CAPÍTULO 5

Supuestos inherentes

Los LCM se basan en una serie de supuestos fundamentales que, si bien permiten un mejor entendimiento y procesamiento de los datos, también presentan desafíos cuando no se cumplen en la práctica. A continuación, se detallan algunos de los supuestos mayormente encontrados en la literatura y las formas en que estos han sido debatidos, analizados, contrarrestados y, en algunos casos, modificados para mejorar la aplicabilidad de los LCM.

5.1. Independencia condicional local

Una suposición importante en el LCM es la de independencia condicional local (ICL), también conocida como independencia local. Este principio establece que las variables observadas son independientes entre sí dentro de una misma clase latente. En otras palabras, una vez que se conoce la clase a la que pertenece una variable, la respuesta a una variable observada no debe estar influenciada por las respuestas de otras [He y Fan, 2020].

El supuesto ICL permite simplificar el cálculo de la probabilidad de un patrón de respuesta dado. Sin embargo, la violación de esta suposición puede tener graves consecuencias en la calidad del modelo de medición de la siguientes maneras [Visser y Depaoli, 2022]

- Aumento de la probabilidad de asignar un elemento a la clase incorrecta.
- Deterioro del ajuste del modelo afectando la selección del número de clases. Esto se debe a que los índices de ajuste del modelo, como el AIC y el BIC, tienden a favorecer modelos con más clases cuando hay dependencia condicional no modelada.
- Dificultad en la interpretación de las clases, conduciendo a clases que no cuentan con un fundamento empírico y difíciles de interpretar.

Dado su importancia, investigadores buscan justificar la estructura de dependencia condicional. Esto se puede lograr mediante el uso de algunas técnicas estadísticas para detectar la violación de la ICL. Entre ellos [Visser y Depaoli, 2022]

- **Residuos bivariados**: esta técnica examina las asociaciones residuales entre pares de indicadores para identificar dependencias condicionales.
- **Índice de modificación**: esta medida estadística indica cuánto mejoraría el ajuste del modelo si se relajara la restricción de ICL para un par específico de indicadores.
- **Estrategia de priori restrictivo**: en la estimación bayesiana, se pueden usar información a priori para flexibilizar la ICL y detectar la dependencia condicional.

Una vez que se detecta la dependencia condicional, se puede ajustar el modelo para obtener estimaciones de parámetros más precisas. Esto suele implicar establecer correlaciones adicionales entre los indicadores que muestran esta dependencia.

5.2. Homogeneidad dentro de las clases latentes

Este supuesto establece que los elementos dentro de una misma clase latente son homogéneos en términos de sus respuestas a las variables observadas [Hess, 2024]. En otras palabras, asume que todos los elementos de una clase comparten la misma probabilidad de responder a una variable de una manera particular.

La homogeneidad dentro de las clases latentes tiene varias implicaciones para su aplicación e interpretación, entre ellos: permite simplificar el modelo y reducir la cantidad de parámetros a estimar, las clases latentes pueden interpretarse como grupos con perfiles de respuesta distintos y bien definidos, la clasificación de los elementos en las clases latentes tiende a ser más precisa.

En la práctica, es común encontrar heterogeneidad dentro de las clases latentes debido a diversos factores

- **Variables no incluidas en el modelo**: la omisión de variables observables importantes puede distorsionar la formación de las clases y afectar la precisión de la clasificación.
- **Naturaleza continua de las variables latentes**: las variables latentes subyacentes pueden ser continuas, lo que introduce una variación gradual dentro de las clases. En estos casos, el supuesto de homogeneidad puede no ser completamente válido.
- **Interacciones entre variables**: el modelo puede no capturar las interacciones complejas entre las variables. Las relaciones no lineales entre variables pueden contribuir a la variación dentro de una misma clase.

Si no se cumple este supuesto, la interpretación de los resultados del LCM puede ser erradas o inconsistentes. Las clases pueden no representar grupos realmente distintos, sino simplemente variaciones aleatorias dentro de una población más homogénea. Además, la violación del supuesto puede afectar la precisión de la clasificación de los elementos en las clases [Sinha *et al.*, 2021].

Para garantizar la validez, es fundamental detectar y abordar la heterogeneidad dentro de las clases latentes. Existen diversas estrategias para este fin

- **Análisis exploratorio de los datos**: un análisis exploratorio de los datos puede revelar patrones de heterogeneidad dentro de las clases. La inspección visual de los elementos dentro de cada clase puede proporcionar indicios sobre la presencia de subgrupos [Van Lissa *et al.*, 2024].
- **Inclusión de covariables**: incluir covariables en el modelo puede ayudar a explicar parte de la heterogeneidad dentro de las clases. Las covariables pueden ser variables observadas que se cree que están relacionadas con las variables latentes o con la pertenencia a una clase [**?**, Porcu y Giambona, 2017].
- **Aumento del número de clases**: una solución simple para capturar la variación dentro de las clases es aumentar el número de clases latentes en el modelo [Sinha *et al.*, 2021]. Esto permite una mayor flexibilidad para modelar la heterogeneidad, pero también aumenta la complejidad del modelo.
- **Modelos jerárquicos de clases latentes**: los modelos jerárquicos de clases latentes permiten modelar la heterogeneidad mediante la inclusión de niveles adicionales de clases latentes. Estos modelos son más complejos, pero pueden ser más realistas en situaciones donde existe una estructura jerárquica en los datos [Van Lissa *et al.*, 2024].

Finalmente, el supuesto de independencia condicional es un supuesto relacionado con la homogeneidad dentro de las clases, por lo que su flexibilidad implica una flexibilidad en la homogeneidad.

5.3. Número fijo de clases latentes

El número de clases latentes, denotado como K, es conocido y fijo. Este supuesto implica que el investigador, basándose en conocimiento a priori, ha determinado, con anterioridad, la cantidad de subgrupos o clases latentes que existen en los datos.

Si bien la suposición simplifica el modelado, también presenta restricciones importantes [Bakk, 2024]. En muchos casos, determinar el número óptimo de clases latentes no es trivial, especialmente cuando se exploran fenómenos complejos o poco estudiados. También, la elección de un K incorrecto puede afectar significativamente la estimación de otros parámetros del modelo, la clasificación de los elementos e incluso los análisis de la investigación. Además, si se asume un K mayor al real, se corre el riesgo de dividir artificialmente la población en subgrupos sin significado fundamentado. Por otro lado, un K menor al real puede llevar a una subestimación de la heterogeneidad de la población, agrupando elementos con perfiles distintos en una misma clase.

A pesar de estas restricciones, la suposición de un K fijo es común en la práctica, especialmente en estudios confirmatorios donde se busca validar una hipótesis específica sobre la estructura latente de la población [Chen *et al.*, 2023]. Para mitigarlo, se han propuesto diversas estrategias en la literatura [Van Lissa *et al.*, 2024, Sinha *et al.*, 2021]

- **Criterios de información**: AIC, BIC y saBIC son criterios de información que se utilizan para comparar modelos con diferente número de clases, buscando un equilibrio entre el ajuste a los datos y la complejidad del modelo.
- **Pruebas estadísticas**: la prueba de razón de verosimilitud bootstrapped (BLRT) se usa para comparar modelos anidados y evaluar la significancia estadística de la adición de una clase adicional.
- **Medidas de clasificación**: la entropía y la precisión de la clasificación se pueden utilizar para evaluar la separación y la fiabilidad de las clases identificadas.
- **Interpretación sustantiva**: la interpretabilidad y la relevancia teórica de las clases obtenidas son cruciales para justificar la elección de un K particular.

En la práctica, se recomienda combinar diferentes estrategias para determinar el número de clases latentes y verificar la elección final. La transparencia en el proceso de selección del modelo y la presentación de los resultados son esenciales para que los lectores puedan evaluar la validez de las conclusiones de la investigación.

5.4. Supuestos de distribución con datos longitudinales

En el contexto de los modelos de clases latentes (LCM), el análisis de datos longitudinales introduce complejidades adicionales en lo que respecta a los supuestos de distribución. A diferencia de los datos transversales, donde se asume que las observaciones son independientes, los datos longitudinales presentan una dependencia inherente entre las observaciones a lo largo del tiempo.

Varios de otros supuestos o limitaciones funcionan diferentes con los datos longitudinales [Lee *et al.*, 2020, Hess, 2024], en este caso mencionaremos algunos de estos

- El supuesto tradicional de normalidad para variables continuas puede no ser apropiado para datos longitudinales. Esto se debe a que a menudo presentan patrones no normales, como asimetría, curtosis o heterocedasticidad, que evolucionan con el tiempo. Ignorar estas desviaciones de la normalidad puede llevar a estimaciones sesgadas de los parámetros del modelo y a inferencias erróneas sobre la estructura de la clase latente.
- La presencia de datos faltantes, un problema común en estudios longitudinales, complica aún más los supuestos de distribución. Los métodos de imputación de datos faltantes, como la imputación múltiple basada en cadenas de Markov Monte Carlo (MCMC), a menudo asumen la normalidad de los datos, lo que puede no ser válido para datos longitudinales no normales. Se han desarrollado métodos de imputación más

sofisticados, como la imputación múltiple basada en regresión cuantílica bayesiana (BQR), para abordar esta limitación y proporcionar estimaciones más robustas en presencia de datos faltantes no normales.

- La heterogeneidad no observada puede manifestarse de manera diferente en datos longitudinales. Además de la heterogeneidad entre clases, también puede haber heterogeneidad dentro de la clase en las trayectorias de los elementos a lo largo del tiempo.

Al elegir y evaluar los supuestos de distribución para LCM con datos longitudinales, es esencial considerar la naturaleza específica del estudio, las características de los datos y los objetivos de la investigación. Las comprobaciones de diagnóstico, como los gráficos de residuos, las pruebas de bondad de ajuste y los análisis de sensibilidad, deben emplearse para evaluar la idoneidad de los supuestos y la robustez de los resultados del modelo.

El análisis de datos longitudinales utilizando LCM exige una consideración exhaustiva de los supuestos de distribución. Las complejidades introducidas por la dependencia temporal, la no normalidad, los datos faltantes y la heterogeneidad no observada deben abordarse cuidadosamente para garantizar una estimación precisa del modelo, una clasificación válida y conclusiones significativas.

CAPÍTULO 6

Limitaciones teórico-practicas

A pesar de su utilidad en la modelización de datos complejos y multivariados, los LCM no están exentos de limitaciones. Estas limitaciones pueden afectar tanto la validez de las conclusiones como la robustez de los resultados. En este apartado recaen algunas de las principales limitaciones de los LCM que la literatura ha identificado, basándose en estudios claves y revisiones sistemáticas que examinan la aplicación de los LCM en diferentes contextos.

6.1. Identificabilidad del modelo

La identificabilidad es un concepto fundamental en el amplio aspecto del modelado estadístico. Se refiere a la capacidad de estimar de forma única los parámetros del modelo a partir de los datos observados. Un modelo es identificable si existe una correspondencia única entre los valores de los parámetros y la distribución de probabilidad de los datos. La falta de identificabilidad, implica que diferentes conjuntos de parámetros podrían generar la misma distribución de datos, lo que imposibilita la estimación precisa de los parámetros y lleva a conclusiones ambiguas [Gu y Xu, 2020].

La no identificabilidad puede surgir en los LCM debido a diversas razones [Culpepper, 2023], incluyendo: redundancia en la parametrización, restricciones insuficientes en el espacio de parámetros o estructura compleja de las relaciones entre las variables latentes y observadas.

Las consecuencias a la que podría llevar la presencia de esta limitación varían con respecto al objetivo base. En general, las estimaciones de los parámetros pueden variar ampliamente entre diferentes ejecuciones del modelo, incluso con los mismos datos. También, las pruebas de hipótesis y los intervalos de confianza basados en un modelo no identificable pueden ser engañosos. Así como, la interpretación de los resultados puede ser ambigua, ya que diferentes conjuntos de parámetros podrían explicar los datos igualmente bien.

Es crucial verificar la identificabilidad de un LCM antes de realizar inferencias. Esto implica examinar la estructura del modelo y las restricciones impuestas en los parámetros, así como considerar la naturaleza de los datos. Existen diversos enfoques para abordar la limitación de la no identificabilidad [Culpepper, 2023, Wang *et al.*, 2020], entre ellos

- **Imposición de restricciones adicionales**: se pueden considerar restricciones adicionales en los parámetros para reducir el número de parámetros libres y mejorar la identificabilidad.
- **Uso de información previa**: se puede incorporar información a priori sobre los parámetros, a través de distribuciones previas, para mejorar la identificabilidad en un contexto bayesiano.
- **Simplificación del modelo**: en algunos casos, simplificar la estructura del modelo puede mejorar la identificabilidad.

Sobre la literatura se plantea el concepto de la identificabilidad genérica [Gu y Xu, 2020], el cual es más débil que la identificabilidad estricta. Un modelo es genéricamente identificable si es identificable en casi todos los puntos del espacio de parámetros, excepto en un conjunto de medida cero. La identificabilidad genérica implica que la no identificabilidad es un problema "improbable" en la práctica. Sin embargo, es importante tener en cuenta que incluso en modelos genéricamente identificables, pueden existir conjuntos específicos de parámetros no identificables.

La investigación sobre la identificabilidad de los LCM ha avanzado significativamente en las últimas décadas. Se han establecido condiciones para la identificabilidad de diversos tipos de LCM, incluyendo modelos con variables binarias, politómicas y nominales [Ouyang y Xu, 2022, Liu y Culpepper, 2024]. Estas condiciones suelen involucrar requisitos sobre la estructura del modelo, como la presencia de ciertas configuraciones de elementos o la relación entre las variables latentes y observadas.

6.2. Interpretación de las clases latentes

Aunque los algoritmos estadísticos pueden identificar agrupaciones distintas dentro de los datos, la interpretación de estos subgrupos en patrones significativos y válidos presenta un desafío [Sinha *et al.*, 2021].

Existe el peligro de sobrestimar el número de clases latentes. Esto puede ocurrir cuando los investigadores priorizan un mejor ajuste estadístico sobre la interpretabilidad sustantiva, lo que lleva a la identificación de clases que tienen poca relevancia práctica o que son difíciles de distinguir conceptualmente.

Se pueden enfrentar ante la trampa del "efecto Salsa" [Sinha *et al.*, 2021], donde, a veces, las clases latentes identificadas representan simplemente grados de severidad en las variables observadas en lugar de grupos cualitativamente distintos. Este fenómeno, se conoce como el "efecto Salsa", donde las clases se alinean en un continuo sin diferencias significativas en sus perfiles.

La asignación de nombres excesivamente simplistas a las clases latentes puede distorsionar su verdadera naturaleza. Si bien los nombres cortos son útiles para la comunicación, no deben afectar la complejidad de los perfiles de clase. Es esencial proporcionar una descripción detallada de las características de cada clase, incluyendo los parámetros del modelo, para evitar interpretaciones erróneas [Van Lissa *et al.*, 2024].

La composición y la calidad de las variables observadas utilizadas para construir el modelo de clases latentes influyen, en gran medida, en la interpretación de las clases [Porcu y Giambona, 2017].

- La elección de variables debe estar relacionada por la pregunta de investigación y estar encaminada el objetivo latente bajo estudio.
- Clases latentes mal seleccionados o con poca relación con el objetivo pueden resultar en clases latentes que son difíciles de interpretar o sin sentido [Culpepper, 2023].

La interpretación de las clases latentes requiere una evaluación cuidadosa que vaya más allá de los resultados estadísticos. Los investigadores deben considerar la relevancia teórica, la separación significativa entre clases y la validez [Sinha *et al.*, 2021, Van Lissa *et al.*, 2024]. Una interpretación sólida combina el rigor estadístico con la comprensión sustantiva para derivar conclusiones significativas a partir de los modelos de clases latentes.

6.3. Sensibilidad a la calidad de los datos y tamaño de la muestra

Los LCM son particulamente sensibles tanto a la calidad de los datos como al tamaño de la muestra. Esta sensibilidad puede afectar significativamente la precisión y la validez de los resultados del modelo [Aflaki *et al.*, 2023, Alagöz y Vermunt, 2022].

La presencia de valores atípicos pueden distorsionar la formación de clases. Es fundamental realizar un análisis exploratorio de los datos para detectar y manejar estos valores. Un análisis de sensibilidad con diferentes valores

de una constante sumada a las frecuencias observadas puede ayudar a medir el impacto de las observaciones con frecuencias bajas o nulas en las estimaciones de los parámetros y los estadísticos de bondad de ajuste [Alpízar, 2015].

En cuanto a las variables continuas, aquellas que no presentan una distribución normal pueden requerir transformaciones matemáticas para ser utilizadas en LCM. De igual forma, cuando se trabaja con variables ordinales o continuas que contienen categorías de baja frecuencia, estas pueden agruparse para facilitar la interpretación. Sin embargo, este procedimiento puede afectar la confiabilidad de la escala y la precisión de las mediciones.

Los valores faltantes en los valores observados de clase pueden sesgar las estimaciones de los efectos de las covariables. Se pueden utilizar métodos como la imputación múltiple o el análisis de clases latentes por pasos para manejar los valores faltantes [Vermunt y Magidson, 2021, Stahlmann *et al.*, 2023].

Por otro lado, el tamaño de la muestra tiene una influencia directa en la capacidad del modelo para identificar y estimar correctamente las clases latentes. Uno de los primeros desafíos es la detección de clases pequeñas; muestras limitadas pueden dificultar la identificación de subgrupos relevantes en los datos, aunque estos sean significativos para el análisis. Asimismo, la precisión en la estimación de parámetros depende del número de observaciones: un tamaño de muestra insuficiente puede resultar en estimaciones inexactas y en interpretaciones equivocadas de los resultados obtenido [Nylund-Gibson y Choi, 2018].

Otro aspecto importante es la separación entre clases. Cuando los modelos presentan una clara distinción entre clases, el tamaño de la muestra requerido es menor para obtener estimaciones precisas. Sin embargo, si la separación entre clases es baja, será necesario un tamaño de muestra más amplio para evitar problemas de precisión en el modelo [Bakk, 2024]. Además, la detección de efectos de covariables en las clases latentes también depende del tamaño de la muestra: un mayor tamaño aumenta la potencia estadística y, con ello, la posibilidad de detectar dichos efectos.

Para mejorar la precisión y confiabilidad de los modelos de clases latentes, es esencial seguir ciertas recomendaciones que aborden aspectos clave como la calidad de los datos y el tamaño de la muestra.

- Examinar la calidad de los datos cuidadosamente y abordar los valores atípicos, la no normalidad y los valores faltantes de forma adecuada.
- Utilizar un tamaño de muestra lo suficientemente grande para garantizar la precisión de la estimación del modelo y la detección de clases significativas.
- Considerar la separación de clases esperada al planificar el tamaño de la muestra.
- Utilizar métodos de validación, como el análisis de sensibilidad o la validación cruzada, para evaluar la solidez de los resultados del modelo.

En general, la sensibilidad de los LCM a la calidad de los datos y al tamaño de la muestra radica en la importancia de un diseño de estudio cuidadoso y una consideración exhaustiva de las limitaciones del modelo.

6.4. Complejidad computacional

Como hemos observado, los LCM son una herramienta poderosa para analizar datos, pero su aplicación práctica puede verse limitada por la complejidad computacional [Naldi y Cazzaniga, 2020, Figuera y García Bringas, 2024].

- **Algoritmo EM y su convergencia**: la estimación de parámetros en LCM generalmente se basa en el algoritmo Expectation-Maximization (EM). Si bien EM es conceptualmente elegante, su convergencia puede ser lenta, especialmente en modelos con muchas clases latentes.
- **Número de clases latentes (K)**: la complejidad computacional de los LCM aumenta exponencialmente con el número de clases latentes (K). Esto se debe a que el número de parámetros del modelo crece con

K, requiriendo más iteraciones de algoritmosm como EMm para converger. La elección óptima de K a menudo implica comparar modelos con diferentes números de clases, aumentando aún más la carga computacional.

- **Dimensionalidad de los datos**: la dimensionalidad de los datos, tanto en términos de número de individuos como de número de clases latentes, también influye en la complejidad computacional. Conjuntos de datos más grandes requieren más memoria y tiempo de procesamiento.
- **Restricciones del modelo**: los modelos de clases latentes restringidos (RLCM), los cuales imponen restricciones adicionales a los parámetros del modelo. Estas restricciones pueden requerir algoritmos de optimización más sofisticados y aumentar el tiempo de cálculo [Culpepper, 2023].

Se han desarrollado varias estrategias para mitigar la complejidad computacional de los LCM. Las variantes del algoritmo EM, como el tempered EM, buscan mejorar la velocidad de convergencia. También, reducir la dimensionalidad de los datos antes del análisis puede aliviar la carga computacional. Así como, la paralelización del algoritmo EM, utilizando GPUs o clusters de computación, puede acelerar significativamente la estimación del modelo.

Al abordar la complejidad computacional en LCM, es crucial considerar elegir modelos y algoritmos adecuados a los recursos computacionales disponibles. En algunos casos, puede ser necesario sacrificar cierta precisión para obtener resultados en un tiempo razonable. Es importante que la complejidad computacional no debe comprometer la validación del modelo, incluyendo la evaluación del ajuste y la interpretación de los resultados.

CAPÍTULO 7

Conclusión

Por medio de una revisión sistemática de la literatura, utilizando la metodología PRISMA, esta tesis ha permitido identificar y analizar los aspectos críticos en la implementación de los modelos de clases latentes (LCM), así como proponer estrategias para abordar sus principales desafíos metodológicos.

El análisis de los supuestos subyacentes a los LCM, como la independencia condicional local y la homogeneidad dentro de las clases latentes, reveló que la violación de estos supuestos puede afectar significativamente la validez y la interpretabilidad de los resultados. Es fundamental abordar la posible dependencia condicional entre variables, así como la heterogeneidad no observada, utilizando estrategias como el análisis de residuos bivariados, la inclusión de covariables o la consideración de modelos jerárquicos de clases latentes. La determinación del número óptimo de clases latentes (K) también es crucial y se puede abordar mediante criterios de información como AIC, BIC y saBIC, pruebas estadísticas como la BLRT y medidas de clasificación como la entropía y la precisión.

La identificabilidad del modelo en los LCM es una limitación clave, ya que su falta puede llevar a estimaciones imprecisas y conclusiones ambiguas. Para enfrentar esto, se pueden añadir restricciones, incorporar información previa o simplificar el modelo. La interpretación de las clases latentes también requiere cuidado, pues sobreestimar el número de clases o asignar nombres simplistas puede distorsionar su verdadero significado. Además, la sensibilidad de los LCM a la calidad de los datos y al tamaño de la muestra hace esencial un diseño cuidadoso del estudio, así como una gestión adecuada de valores atípicos, normalidad y valores faltantes. Finalmente, la complejidad computacional en modelos con muchas clases y datos de alta dimensión exige estrategias como el uso de variantes del algoritmo EM, reducción de dimensionalidad o paralelización.

En conclusión, la aplicación efectiva de los LCM depende de un diseño de estudio riguroso, un análisis cuidadoso de los supuestos y una gestión adecuada de las limitaciones. Al abordar estos aspectos, los investigadores pueden aprovechar el poder de los LCM para obtener información significativa de datos complejos.

AGRADECIMIENTOS

A lo largo del desarrollo de esta tesis, muchas personas han sido fundamentales en el apoyo y guía para alcanzar esta meta. A todos ellos, les expreso mi más sincero agradecimiento.

En primer lugar, quiero agradecer profundamente a mis padres, quienes siempre han creído en mí y han sido mi mayor fuente de motivación. Su apoyo sacrificio me han permitido llegar hasta aquí.

Mi agradecimiento al patrocinador de la Beca Konder, que me permitió realizar mis estudios y alcanzar este logro académico. Su apoyo fue fundamental en mi camino y sin esa oportunidad, esta tesis no habría sido posible. La confianza que depositaron en mí ha sido una fuente constante de motivación para esforzarme y dar lo mejor de mí en cada etapa de este ciclo educativo.

A mis compañeros y amigos, quienes han sido un pilar importante a lo largo de esta travesía. A Mariana Oquendo, gracias por su apoyo y amistad incondicional en el ciclo educativo.

A mi asesor de tesis, Dr.rer.nat Humberto Llinás, por su guía y sus valiosos comentarios. Y no puedo dejar de agradecer a los profesores y a la institución que me han formado y brindado las herramientas necesarias para alcanzar mis objetivos. En especial, al profesor Berdarno Uribe, quien fue la primera cara de las matemáticas que encontré y fue motivador.

Finalmente, a todos aquellos que, de una u otra forma, han contribuido a este proyecto con sus palabras de ánimo, su ayuda o su compañía. A ustedes, mi más sincero agradecimiento y reconocimiento.

Gracias a todos por ser parte de este camino.

Bibliografía

[Aflaki *et al.*, 2023] Aflaki, K., Vigod, S., y Ray, J. G. (2023). Part ii: A step-by-step guide to latent class analysis. *Journal of Clinical Epidemiology*, 159:348–351.

[Alagöz y Vermunt, 2022] Alagöz, Ö. E. C. y Vermunt, J. K. (2022). Stepwise latent class analysis in the presence of missing values on the class indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(5):784–790.

[Alpízar, 2015] Alpízar, C. A. (2015). Diagnóstico de modelos de clases latentes en tablas poco ocupadas. *Pensamiento Actual*, 15(25):77–84.

[Bakk, 2024] Bakk, Z. (2024). Latent class analysis with measurement invariance testing: Simulation study to compare overall likelihood ratio vs residual fit statistics based model selection. *Structural Equation Modeling: A Multidisciplinary Journal*, 31(2):253–264.

[Casella y Berger, 2024] Casella, G. y Berger, R. (2024). *Statistical inference*. CRC Press.

[Chen *et al.*, 2023] Chen, Y., Culpepper, S. A., y Chen, Y. (2023). Bayesian inference for an unknown number of attributes in restricted latent class models. *psychometrika*, 88(2):613–635.

[Chow y Teicher, 2012] Chow, Y. S. y Teicher, H. (2012). *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media.

[Clogg y Goodman, 1984] Clogg, C. C. y Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, 79(388):762–771.

[Collins y Lanza, 2009] Collins, L. M. y Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, volumen 718. John Wiley & Sons.

[Culpepper, 2023] Culpepper, S. A. (2023). A note on weaker conditions for identifying restricted latent class models for binary responses. *psychometrika*, 88(1):158–174.

[Figuera y García Bringas, 2024] Figuera, P. y García Bringas, P. (2024). Revisiting probabilistic latent semantic analysis: Extensions, challenges and insights. *Technologies*, 12(1):5.

[GeeksforGeeks, 2024] GeeksforGeeks (2024). Latent class analysis in r. `https://www.geeksforgeeks.org/latent-class-analysis-in-r/`. [Accessed 22-11-2024].

[Gelman *et al.*, 1995] Gelman, A., Carlin, J. B., Stern, H. S., y Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

[Goodman, 1974] Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

[Gu y Xu, 2020] Gu, Y. y Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics*, 48(4):2082–2107.

[Hagenaars y McCutcheon, 2002] Hagenaars, J. A. y McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

[Hastie *et al.*, 2009] Hastie, T., Tibshirani, R., Friedman, J. H., y Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volumen 2. Springer.

[He y Fan, 2020] He, J. y Fan, X. (2020). Latent class analysis. En *Encyclopedia of personality and individual differences*, pp. 2566–2570. Springer.

[Hess, 2024] Hess, S. (2024). Latent class structures: taste heterogeneity and beyond. En *Handbook of choice modelling*, pp. 372–391. Edward Elgar Publishing.

[Klemens, 2008] Klemens, B. (2008). *Modeling with data: tools and techniques for scientific computing*. Princeton University Press.

[Lahoz *et al.*, 2023] Lahoz, L. T., Pereira, F., Sfeir, G., Arkoudi, I., Monteiro, M. M., y Azevedo, C. L. (2023). Attitudes and latent class choice models using machine learning. *ArXiv*, abs/2302.09871.

[Lazarsfeld, 1950] Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. *Studies in social psychology in world war II Vol. IV: Measurement and prediction*, pp. 362–412.

[Lazarsfeld, 1968] Lazarsfeld, P. F. (1968). Latent structure analysis. *(No Title)*.

[Lee *et al.*, 2020] Lee, M., Rahbar, M. H., Gensler, L. S., Brown, M., Weisman, M., y Reveille, J. D. (2020). A latent class based imputation method under bayesian quantile regression framework using asymmetric laplace distribution for longitudinal medication usage data with intermittent missing values. *Journal of biopharmaceutical statistics*, 30(1):160–177.

[Liberati *et al.*, 2009] Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., y Moher, D. (2009). The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of internal medicine*, 151(4):W–65.

[Liu y Culpepper, 2024] Liu, Y. y Culpepper, S. A. (2024). Restricted latent class models for nominal response data: Identifiability and estimation. *Psychometrika*, 89(2):592–625.

[Maity y Saha, 2023] Maity, M. y Saha, P. (2023). Normal distribution. *International journal of science and research*.

[McCutcheon, 1987] McCutcheon, L. (1987). *Latent class analysis*. Sage.

[Naldi y Cazzaniga, 2020] Naldi, L. y Cazzaniga, S. (2020). Research techniques made simple: latent class analysis. *Journal of Investigative Dermatology*, 140(9):1676–1680.

[Nylund-Gibson y Choi, 2018] Nylund-Gibson, K. y Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, 4(4):440.

[Ouyang y Xu, 2022] Ouyang, J. y Xu, G. (2022). Identifiability of latent class models with covariates. *psychometrika*, 87(4):1343–1360.

[Porcu y Giambona, 2017] Porcu, M. y Giambona, F. (2017). Introduction to latent class analysis with applications. *The Journal of Early Adolescence*, 37(1):129–158.

[Sinha *et al.*, 2020] Sinha, P., Calfee, C., y Delucchi, K. (2020). Practitioner's guide to latent class analysis: Methodological considerations and common pitfalls. *Critical Care Medicine*.

[Sinha *et al.*, 2021] Sinha, P., Calfee, C. S., y Delucchi, K. L. (2021). Practitioner's guide to latent class analysis: methodological considerations and common pitfalls. *Critical care medicine*, 49(1):e63–e79.

[Stahlmann *et al.*, 2023] Stahlmann, K., Reitsma, J. B., y Zapf, A. (2023). Missing values and inconclusive results in diagnostic studies–a scoping review of methods. *Statistical Methods in Medical Research*, 32(9):1842–1855.

[Van Lissa *et al.*, 2024] Van Lissa, C. J., Garnier-Villarreal, M., y Anadria, D. (2024). Recommended practices in latent class analysis using the open-source r-package tidysem. *Structural Equation Modeling: A Multidisciplinary Journal*, 31(3):526–534.

[Vermunt, 2002] Vermunt, J. K. (2002). Latent class cluster analysis. *Applied latent class analysis/Cambridge UniversityPress*.

[Vermunt y Magidson, 2021] Vermunt, J. K. y Magidson, J. (2021). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(3):356–364.

[Visser y Depaoli, 2022] Visser, M. y Depaoli, S. (2022). A guide to detecting and modeling local dependence in latent class analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(6):971–982.

[Wang *et al.*, 2020] Wang, C., Lin, X., y Nelson, K. P. (2020). Bayesian hierarchical latent class models for estimating diagnostic accuracy. *Statistical methods in medical research*, 29(4):1112–1128.

[Weller *et al.*, 2020] Weller, B. E., Bowen, N. K., y Faubert, S. J. (2020). Latent class analysis: a guide to best practice. *Journal of black psychology*, 46(4):287–311.