

PONTIFICIA UNIVERSIDAD
CATÓLICA DEL PERÚ

Escuela de Posgrado



Clusterización basada en una mixtura con distribuciones normales
contaminadas multivariadas con datos incompletos: Una aplicación a la
evaluación de habilidades socioemocionales

Tesis para obtener el grado académico de Magíster en Estadística
que presenta:

Ángel Christopher Zegarra López

Asesor:

Dr. Luis Enrique Benites Sánchez


Lima, 2023

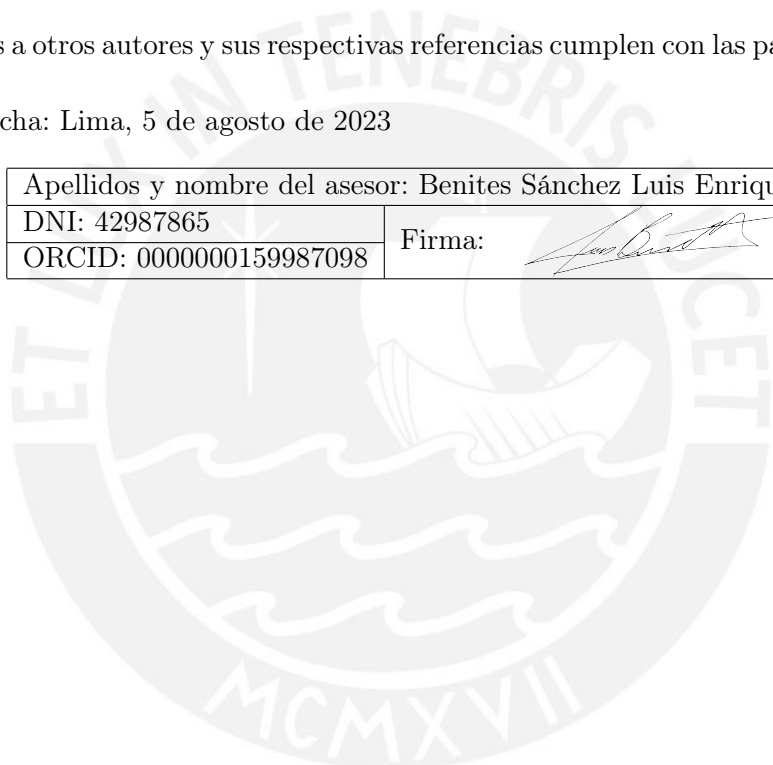
Informe de Similitud

Yo, Luis Enrique Benites Sánchez docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis titulada Clusterización basada en una mixtura con distribuciones normales contaminadas multivariadas con datos incompletos: Una aplicación a la evaluación de habilidades socioemocionales del autor Angel Christopher Zegarra López, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 14%, lo que está dentro del límite establecido. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 18/07/2023.
- He revisado con detalle dicho reporte y la Tesis, y no se advierte indicios de plagio..
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 5 de agosto de 2023

| | |
|---|---|
| Apellidos y nombre del asesor: Benites Sánchez Luis Enrique | |
| DNI: 42987865 | Firma:  |
| ORCID: 0000000159987098 | |



Resumen

Aunque la distribución normal es útil en una variedad de contextos, enfrenta ciertas limitaciones al modelar datos que contienen valores extremos. Estos valores pueden generar “colas” más pesadas en la distribución, en contraste con las colas más ligeras de la distribución normal. Por lo tanto, en tales circunstancias, la distribución normal contaminada se presenta como una alternativa efectiva. Este ajuste es especialmente significativo en aplicaciones como la agrupación basada en modelos. En este método, es habitual emplear distribuciones normales multivariadas como fundamento para la agrupación. No obstante, la estimación de parámetros puede verse afectada por la presencia de valores extremos. En este estudio, implementamos la distribución normal contaminada multivariada como base para la agrupación basada en modelos, tal como propone Tong y Tortora (2022). Explicamos las características del modelo y llevamos a cabo un estudio de simulación para contrastar su desempeño con la distribución normal multivariada y la distribución t multivariada. Finalmente, aplicamos un proceso de agrupación basado en una mezcla de distribuciones normales contaminadas multivariadas a un conjunto de datos reales. Estos datos se derivan de los resultados de la Evaluación de Habilidades Socioemocionales, una iniciativa implementada por el Ministerio de Educación de Perú en 2021.

Abstract

The normal distribution has limitations when modeling data with outliers. The presence of outliers implies heavier tails in the distribution; whereas, the normal distribution has very light tails. For this reason, the contaminated normal distribution is used as a better alternative to model in these cases. One of the applications where this change is pertinent is in model-based clustering. In this approach, using multivariate normal distributions as the basis for clustering is common practice; however, the parameter estimates may be biased due to the presence of outliers. In this thesis, the multivariate contaminated normal distribution is used as the basis for model-based clustering. The characteristics of the model were presented, as well as a simulation study that compares the performance of the model with respect to the multivariate normal distribution and the multivariate t-distribution. Finally, a clustering process was carried out based on a mixture of multivariate contaminated normal distributions to a data set of the results of the Socio-emotional Skills assessment, an operation implemented by the Ministry of Education of Peru in 2021.

Índice general

| | |
|--|-----------|
| Índice de figuras | VII |
| Índice de cuadros | VIII |
| 1. Introducción | 1 |
| 1.1. Consideraciones preliminares | 1 |
| 1.2. Objetivos | 2 |
| 1.3. Organización del Trabajo | 3 |
| 2. Clusterización basada en modelos | 4 |
| 2.1. Clusterización | 4 |
| 2.2. Clusterización basada en modelos | 5 |
| 2.2.1. Modelos de Mixtura Finita | 6 |
| 2.2.2. Estimación de parámetros vía MLE | 7 |
| 2.2.3. Estimación de parámetros vía EM | 8 |
| 2.2.4. Algoritmo EM en una mixtura de distribuciones normales univariadas | 10 |
| 2.2.5. Estimación de parámetros vía ECM | 12 |
| 2.2.6. Métodos de selección del modelo | 13 |
| 3. Distribución normal contaminada | 14 |
| 3.1. Distribución normal contaminada univariada | 14 |
| 3.2. Distribución normal contaminada multivariada | 17 |
| 3.3. Mixtura de distribuciones normales contaminadas multivariadas | 18 |
| 4. Distribución t | 20 |
| 4.1. Distribución t univariada | 20 |
| 4.2. Distribución t multivariada | 21 |
| 5. Estudio de simulación | 22 |
| 5.1. Características de la simulación | 22 |
| 5.1.1. Condiciones para la mixtura de distribuciones normales multivariadas | 23 |
| 5.1.2. Condiciones para la mixtura de distribuciones t multivariadas | 23 |
| 5.1.3. Condiciones para la mixtura de distribuciones normales contaminadas multivariadas | 23 |

| | |
|--|-----------|
| 6. Aplicación en las Habilidades Socioemocionales | 27 |
| 6.1. Habilidades Socioemocionales | 27 |
| 6.2. La Evaluación de las Habilidades Socioemocionales en Perú | 29 |
| 6.2.1. Características de la muestra | 29 |
| 6.2.2. Características del instrumento de recolección de datos | 30 |
| 6.2.3. Características de los datos | 31 |
| 6.2.4. Análisis exploratorio de datos | 32 |
| 6.3. Clusterización basada en una mixtura de distribuciones normales contaminadas multivariadas | 35 |
| 6.3.1. Modelo $G=2$ | 35 |
| 6.3.2. Modelo $G=4$ | 37 |
| 7. Conclusiones y sugerencias | 40 |
| 7.1. Conclusiones | 40 |
| 7.2. Sugerencias | 40 |
| 8. Código en R | 41 |
| 8.1. Código en R para el estudio de simulación | 41 |
| 8.2. Código en R para el análisis del estudio de simulación | 43 |
| 8.3. Código en R para la aplicación práctica | 50 |
| Bibliografía | 52 |
| Anexo 1 | 57 |
| A. Modelo Rasch | 57 |
| B. Modelo Rasch para ítems dicotómicos | 58 |
| C. Modelo de Escala de Valoración | 60 |
| D. Métodos de estimación de parámetros | 62 |
| E. Una nota sobre las medidas Rasch | 64 |

Índice de figuras

| | | |
|------|--|----|
| 3.1. | Dos distribuciones normales con mismo parámetro de localización $\mu = 0$, pero distinto parámetro de escala $\sigma^2 = 1$ y $\sigma^2 = 3$, respectivamente | 15 |
| 3.2. | Una distribución normal contaminada (en rojo) como resultado de dos componentes normales, con parámetros $\mu = 0$, $\sigma^2 = 1$, $\eta = 3$ y $\alpha = 0,20$ | 16 |
| 6.1. | Histograma de las variables Autoeficacia, Autonomía y Relación | 32 |
| 6.2. | Histograma de las variables Autorregulación conductual, Empatía cognitiva, Disposición empática, Resiliencia y Responsabilidad | 33 |
| 6.3. | Histograma de la variable Toma de decisiones y sus subdimensiones | 34 |
| 6.4. | Medias aritméticas de los conglomerados identificados en el primer modelo . . | 36 |
| 6.5. | Medias aritméticas de los conglomerados identificados en el segundo modelo . | 38 |
| B.1. | Curvas Características para tres ítems hipotéticos con parámetros de dificultad definidos como $b_1=-1,00$, $b_2=0,00$ y $b_3=1,00$, bajo el modelo Rasch para ítems dicotómicos | 59 |
| C.1. | Umbrales en un ítem tipo Likert de 5 alternativas de respuesta | 60 |
| C.2. | Curvas Características para para cinco categorías de respuesta de un ítem hipotético con parámetro de localización definido como $\delta=0$ y parámetros de umbrales equidistantes $\tau_1 = -2,00$, $\tau_2 = -1,00$, $\tau_3 = 1,00$, $\tau_4 = 2,00$ | 61 |

Índice de cuadros

| | |
|---|----|
| 2.1. Nomenclatura de los métodos de análisis de conglomerados | 5 |
| 3.1. Versiones Parsimoniosas del Modelo General de Mixtura de Distribuciones Normales Contaminadas Multivariadas | 19 |
| 5.1. Resultados de la simulación sobre la clusterización basados en modelos normales multivariados, modelos t multivariados y modelos normales contaminados multivariados, utilizando k-means para obtener valores iniciales | 24 |
| 5.2. Resultados de la simulación sobre la clusterización basada en modelos normales multivariados, modelos t multivariados y modelos normales contaminados multivariados, utilizando k-medoids para obtener valores iniciales | 25 |
| 5.3. Resultados de la simulación sobre la clusterización basada en modelos normales multivariados, modelos t multivariados y modelos normales contaminados multivariados, utilizando el método hierarchial para obtener valores iniciales | 26 |
| 6.1. Libro de códigos de la base de datos empleada | 31 |
| 6.2. Estadísticos de ajuste para la cantidad de conglomerados | 35 |
| D.1. Métodos de Estimación de Parámetros en Modelos Rasch | 62 |
| E.1. Requerimientos de la Medición Conjunta | 64 |

Capítulo 1

Introducción

1.1. Consideraciones preliminares

La distribución normal multivariada (MN), caracterizada por un vector de medias aritméticas $\boldsymbol{\mu}$ y una matriz de covarianzas $\boldsymbol{\Sigma}$, es una herramienta popular en inferencia estadística para datos continuos multivariados, gracias a sus propiedades teóricas y ventajas computacionales (Punzo y Tortora, 2021). De hecho, estas ventajas también han hecho de la distribución normal multivariada una de las aproximaciones más frecuentemente utilizadas en la implementación de algoritmos de agrupamiento basados en modelos (Akogul y Erisoglu, 2016; Bouveyron y Brunet-Saumard, 2014; Wolfe, 1967; McNicholas, 2016). Sin embargo, la distribución normal multivariada tiene limitaciones notables en situaciones prácticas que involucran *outliers* o valores extremos. Los valores extremos pueden inducir colas pesadas en la distribución, haciendo inadecuada a la distribución normal multivariada para modelar estas situaciones, ya que esta última presenta colas ligeras, lo cual podría afectar la estimación de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$. Como solución a este problema, la distribución normal contaminada multivariada (MCN) se presenta como una alternativa viable para el modelamiento en presencia de valores extremos (Punzo y Tortora, 2021).

La distribución normal contaminada multivariada se puede ver como una generalización de la distribución normal multivariada, pero con la capacidad de modelar colas más pesadas. Según se define en Tong y Tortora (2022), un vector aleatorio de d dimensiones, $\mathbf{X} = (X_1, \dots, X_d)^\top$, sigue esta distribución si posee un vector de medias $\boldsymbol{\mu}$, una matriz de escala $\boldsymbol{\Sigma}$, una proporción $\alpha \in (0, 5, 1)$ que representa la cantidad de observaciones 'buenas' (no outliers), y un grado de contaminación $\eta > 1$. En tal caso, la función de densidad de probabilidad conjunta se describe de la siguiente manera:

$$f_{MCN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \eta) = \alpha f_{MN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1 - \alpha) f_{MN}(\mathbf{x}; \boldsymbol{\mu}, \eta \boldsymbol{\Sigma}),$$

donde $f_{MN}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denota la función de probabilidad conjunta de un vector aleatorio con d dimensiones que sigue una distribución normal multivariada con un vector de medias $\boldsymbol{\mu}$ y una matriz de covarianzas $\boldsymbol{\Sigma}$.

La distribución normal contaminada multivariada representa una mixtura de dos componentes en donde un componente, con probabilidad α representa las buenas observaciones y el otro componente, con una probabilidad $1 - \alpha$ representa las "malas" observaciones u outliers. Ambas componentes comparten el mismo vector de medias aritméticas $\boldsymbol{\mu}$, pero el compo-

nente que representa a las malas observaciones tiene una matriz de covarianzas considerada como inflada $\eta\Sigma$. De esta manera, la MN puede ser considerada como un caso especial de la MCN cuando $\alpha = 1$ y $\eta = 1$. La ventaja principal de la distribución normal contaminada multivariada radica en que proporciona estimaciones más robustas de μ y Σ (Wilcox, 2017). Debido a sus beneficios, esta distribución ha sido utilizada como fundamento para diversos algoritmos, incluyendo la agrupación basada en modelos. Como ejemplo, Punzo y McNicholas (2016) llevaron a cabo una comparación del rendimiento de un algoritmo de agrupación basado en la distribución normal contaminada multivariada con otras mixturas ya reconocidas en la literatura, tales como mixturas de distribuciones normales, mixturas de distribuciones t , mixturas de una distribución normal con una distribución uniforme, y mixturas de distribuciones normales con un componente uniforme. Los autores encontraron que el método propuesto ofrece un rendimiento equivalente al de propuestas previas, con la ventaja adicional de poder identificar directamente a outliers sin la necesidad de establecer un punto de corte a priori.

Como limitación de esta aproximación, Tong y Tortora (2022) sostienen que emplear la distribución normal contaminada multivariada para la clusterización de datos y detectar outliers requiere de datos completos, es decir, requieren que todos los casos presenten observaciones en todas las variables, lo cual en muchos contextos es difícil de lograr. Por estos motivos, los autores proponen una estimación de parámetros basada en una variante del algoritmo esperanza-maximización (EM), el algoritmo EM condicional (ECM) para poder estimar parámetros y detectar outliers cuando se trabaja con datos que incluyen valores perdidos.

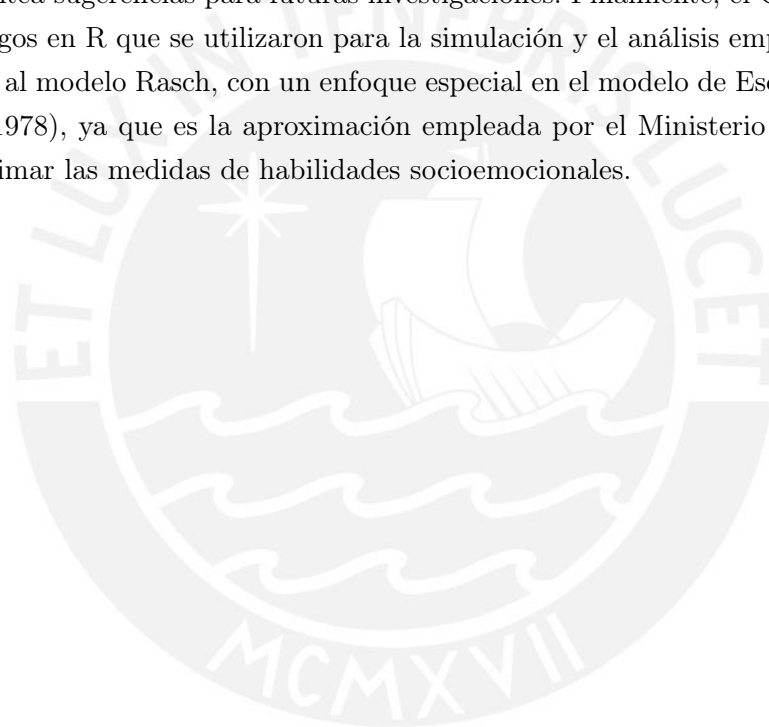
1.2. Objetivos

El objetivo principal de esta tesis es explorar y aplicar un modelo basado en la distribución normal contaminada multivariada como fundamento para un proceso de agrupación basado en modelos y detección de outliers, todo esto en el contexto de un conjunto de datos reales que presenta valores perdidos. Estos datos provienen de una evaluación de habilidades socioemocionales llevada a cabo por el Ministerio de Educación del Perú en 2021. Además, se incluye un estudio de simulación para contrastar el rendimiento del modelo propuesto con otros métodos convencionales, donde el porcentaje de valores perdidos en los datos se controla, constituyendo una extensión del estudio original de Tong y Tortora (2022). De manera más específica:

- Analizar el modelo de la distribución normal contaminada, sus propiedades y la estimación de sus parámetros.
- Desarrollar estudios de simulación que permitan comparar el desempeño de la metodología utilizada con respecto a otras aproximaciones de clusterización basada en modelos.
- Implementar un proceso de clusterización y detección de outliers a los resultados de la evaluación de habilidades socioemocionales implementada por el Ministerio de Educación del Perú en el año 2021.

1.3. Organización del Trabajo

El Capítulo 2 aborda conceptos esenciales relacionados con la agrupación basada en modelos, incluyendo la idea de un modelo de mezcla finita y los procesos de estimación de parámetros asociados con el algoritmo EM y sus variantes. El Capítulo 3 revisa las características de la distribución normal contaminada, considerando tanto el caso univariado como el multivariado y para una mezcla de distribuciones normales contaminadas multivariadas. El Capítulo 4 introduce la distribución t como una competidora de la distribución normal contaminada, ambas consideradas como alternativas a la distribución normal en situaciones donde se encuentran valores extremos. El Capítulo 5 presenta un estudio de simulación para explorar las propiedades del modelo; en tanto que el Capítulo 6 muestra una aplicación práctica del modelo a los resultados de la evaluación de habilidades socioemocionales implementada por el Ministerio de Educación del Perú en 2021. El Capítulo 7 recoge las conclusiones de este trabajo y plantea sugerencias para futuras investigaciones. Finalmente, el Capítulo 8 proporciona los códigos en R que se utilizaron para la simulación y el análisis empírico. El Anexo 1 está dedicado al modelo Rasch, con un enfoque especial en el modelo de Escala de Valoración de Andrich (1978), ya que es la aproximación empleada por el Ministerio de Educación del Perú para estimar las medidas de habilidades socioemocionales.



Capítulo 2

Clusterización basada en modelos

En este capítulo se introduce la noción general de la clusterización, con un énfasis en la clusterización basada en modelos al introducir la definición formal de un modelo de mixtura finita. De esta manera, cada conglomerado es conceptualizado como un componente de la mixtura, reduciendo el proceso de análisis de conglomerados a la estimación de parámetros del modelo de mixtura. Por estos motivos, se enfatiza el proceso de estimación de parámetros desde la perspectiva tradicional basada en máxima verosimilitud, hasta el algoritmo EM (Esperanza-Maximización) y sus extensiones.

2.1. Clusterización

El análisis de conglomerados o clusterización es un conjunto amplio de métodos cuyo propósito es identificar un número reducido de grupos (o clusters) a partir de los datos. De forma más precisa, dados n casos y p variables observadas en un conjunto de datos, el objetivo de estos métodos es discernir $k < n$ grupos de tal manera que cada grupo albergue al menos una unidad y cada unidad pertenezca exclusivamente a un grupo, es decir, los grupos son disjuntos. Esta disposición se conoce como partición de n unidades en k conglomerados. Dada la gran cantidad de particiones posibles, es crucial identificar aquellos grupos que permitan reconocer fácilmente las características que hacen similares a los casos dentro de cada grupo y los diferencian de los casos en otros grupos.

Giordani et al. (2020) propone una clasificación para las técnicas de clusterización que incluye los métodos estándar, fuzzy y basados en modelos. Los métodos estándar se dividen en dos enfoques principales. En primer lugar, tenemos las técnicas jerárquicas, las cuales generan una serie de particiones que permiten la fusión de dos conglomerados (es decir, la clusterización jerárquica aglomerativa) o la división de un conglomerado en dos (es decir, la clusterización jerárquica divisiva) basándose en un criterio determinado. En segundo lugar, las técnicas no jerárquicas buscan la mejor partición de n unidades en k conglomerados minimizando una función de coste. Ambos enfoques estándar también son conocidos como “hard clustering”, ya que cada caso solo puede ser clasificado en un único conglomerado. En contraposición, los métodos fuzzy asignan un grado de pertenencia a cada caso para los conglomerados y también buscan la minimización de una función de coste. Los métodos basados en modelos, por otro lado, establecen supuestos de distribución probabilística sobre los datos y utilizan probabilidades a posteriori como indicadores de pertenencia a cada conglomerado. Tanto los métodos fuzzy como los basados en modelos se consideran “soft clustering”, ya que permiten

una asignación más flexible de casos a cada conglomerado, permitiendo que cada individuo presente un cierto grado o probabilidad de pertenencia a distintos conglomerados.

Cuadro 2.1: Nomenclatura de los métodos de análisis de conglomerados

| Característica | Método estándar | Método fuzzy | Método basado en modelos |
|---------------------------|-----------------|--------------|--------------------------|
| Clusterización hard | Sí | No | No |
| Clusterización soft | No | Sí | Sí |
| Supuestos probabilísticos | No | No | Sí |

2.2. Clusterización basada en modelos

La clusterización basada en modelos consiste en emplear como base para la partición de conglomerados a un modelo probabilístico y los métodos estándares de la inferencia estadística. Básicamente, se asume que cada conglomerado tiene una función de densidad de probabilidad propia y los datos observados corresponden a una mixtura de dichas densidades en lo que formalmente se conoce como un modelo de mixtura finita (Bouveyron et al., 2019).

De este modo, la clusterización basada en modelos consiste en aproximar la verosimilitud de los datos a partir de un modelo de mixtura finita, en donde usualmente la cantidad de componentes de la mixtura representan la cantidad de conglomerados y que los datos observados son particionados utilizando una regla pre-especificada. En este sentido, existe una diferencia clara entre el modelamiento de mixturas finitas y la clusterización basada en modelos. El objetivo principal del modelamiento de mixturas finitas tiende a ser la inferencia sobre los parámetros del modelo; mientras que, la clusterización basada en modelos busca particionar datos en grupos con unidades homogéneas entre sí y diferentes con respecto a unidades de otros conglomerados. Una ventaja de emplear modelos como base para el análisis de conglomerados es que se emplean criterios formales más rigurosos basados en la función de log-verosimilitud penalizada por alguna función del número de parámetros en el modelo (Demg y Han, 2014).

2.2.1. Modelos de Mixtura Finita

Los modelos de mixtura finita constituyen una familia versátil de modelos capaces de modelar poblaciones heterogéneas. Aunque originalmente se emplearon para el modelado y estimación de densidades, su uso se ha extendido recientemente al campo de la agrupación y la clasificación (Giordani et al., 2020). En términos prácticos, estos modelos se han aplicado en una amplia variedad de disciplinas, incluyendo la astronomía (Kuhn y Feigelson, 2018), la biología (Christensen et al., 2021), la genética (Gianola et al., 2007), la medicina (Schlattmann, 2009), la psiquiatría (Miettunen et al., 2016), la economía (Compiani y Kitamura, 2016) y el marketing (Tuma y Decker, 2013), entre otros campos tanto de las ciencias físicas como sociales (McLachlan y Peel, 2000).

En la definición básica de un modelo de mixtura finita, $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ denotan una muestra aleatoria de tamaño n , en donde \mathbf{Y}_j es un vector aleatorio de d dimensiones con una función de densidad de probabilidad $f(\mathbf{y}_j)$ en \mathbb{R}^d . En la práctica, \mathbf{Y}_j contiene las variables aleatorias correspondientes a las d medidas obtenidas en los j registros de las características del fenómeno bajo estudio. Así, $\mathbf{Y} = (\mathbf{Y}_1^\top, \dots, \mathbf{Y}_n^\top)^\top$. En este caso, \mathbf{Y} es la muestra completa, es decir, \mathbf{Y} corresponde a n puntos en \mathbb{R}^p . Una realización del vector aleatorio \mathbf{Y}_j es representado por $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$ como una muestra aleatoria e \mathbf{y}_j es el valor observado de \mathbf{Y}_j . La densidad $f(\mathbf{y}_j)$ de \mathbf{Y}_j puede expresarse como:

$$f(\mathbf{y}_j) = \sum_{i=1}^g \pi_i f_i(\mathbf{y}_j), \quad (2.1)$$

en donde $0 \leq \pi_i \leq 1 (i = 1, \dots, g)$, y $\sum_{i=1}^g \pi_i = 1$. Las cantidades no negativas cuya suma es uno, π_1, \dots, π_g , se conocen como proporciones o pesos de mixtura, mientras que las funciones $f_i(\mathbf{y}_j)$ se denominan densidades de los componentes de la mixtura. En este planteamiento, el modelo de mixtura finita tiene un número fijo de componentes g ; no obstante, en la mayoría de las aplicaciones, el valor de g es desconocido y debe inferirse a partir de los datos, al igual que las proporciones de mixtura y los parámetros de las formas específicas de las densidades de cada componente.

McLachlan y Peel (2000) denotan una manera sencilla de interpretar a los modelos de mixtura finita en donde un vector aleatorio \mathbf{Y}_j con una densidad $f(\mathbf{y}_j)$ de la mixtura de g componentes, como se define en (2.2), puede establecerse tras definir Z_j como una variable aleatoria categórica que puede tomar valores $1, \dots, g$ con probabilidades π_1, \dots, π_g , respectivamente, y la densidad condicional de \mathbf{Y}_j dado $Z_j = i (i = 1, \dots, g)$ es $f_i(\mathbf{y}_j)$. Así, la densidad marginal de \mathbf{Y}_j está dada por $f(\mathbf{y}_j)$. Por consiguiente, la variable Z_j puede interpretarse como un componente que etiqueta al vector \mathbf{Y}_j . En este escenario, resulta más práctico trabajar con un vector \mathbf{Z}_j de g dimensiones en lugar de la variable categórica Z_j . En este caso, cada elemento i de \mathbf{Z}_j se define como 1 ó 0, dependiendo de si el componente original de \mathbf{Y}_j en la mixtura coincide con i o no ($i = 1, \dots, g$).

De este modo, \mathbf{Z}_j sigue una distribución multinomial que consiste en una extracción entre g categorías con probabilidades π_1, \dots, π_g . Esto se puede expresar como:

$$P(\mathbf{Z}_j = \mathbf{z}_j) = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \dots \pi_g^{z_{gj}},$$

en donde $\mathbf{Z}_j \sim \text{Multinomial}_g(1, \boldsymbol{\pi})$ y $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)^\top$. En el ámbito de la agrupación, Giordani et al. (2020) proporcionan una interpretación directa de los componentes de la mixtura como los conglomerados correspondientes. Cada conglomerado posee su propia densidad, por lo que la asignación de cada caso a un conglomerado específico puede efectuarse mediante el criterio del Máximo A Posteriori (MAP), que ubica cada unidad en el conglomerado donde tiene la probabilidad a posteriori más alta. Si más de una probabilidad a posteriori resulta ser máxima, se pueden utilizar métodos de aleatorización (McLachlan y Peel, 2000).

2.2.2. Estimación de parámetros vía MLE

Tradicionalmente, el método más comúnmente utilizado para la estimación de parámetros en un modelo de mixtura finita ha sido el de máxima verosimilitud (ML), maximizando generalmente la log-verosimilitud para una mayor eficiencia computacional. Según la terminología de Giordani et al. (2020), el vector $\boldsymbol{\Psi}$ incluye todos los parámetros de la mixtura, donde $\boldsymbol{\Psi} = (\pi_1, \dots, \pi_{g-1}, \boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_g)$ y $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_q)$ representa un vector de q parámetros del componente g . De este modo, la función de log-verosimilitud puede expresarse como:

$$\ell(\boldsymbol{\Psi}) = \sum_{j=1}^n \log \left[\sum_{i=1}^g \pi_i f(\mathbf{x}_p | \boldsymbol{\zeta}_i) \right]. \quad (2.2)$$

El estimador de máxima verosimilitud para $\boldsymbol{\Psi}$ puede obtenerse al considerar los ceros de la primera derivada de la función de log-verosimilitud, con la restricción de que $\sum_{i=1}^g \pi_i = 1$. Sin embargo, en los modelos de mixtura, es raro que dicha log-verosimilitud conduzca a soluciones cerradas para las estimaciones de máxima verosimilitud, ya que generalmente se obtendrá un conjunto de ecuaciones no lineales que deberán resolverse de manera iterativa. Debido a estas dificultades, en la práctica se prefiere utilizar el algoritmo de Esperanza-Maximización (EM) como un enfoque más apropiado para la estimación de los parámetros del modelo. Este algoritmo fue introducido por Dempster et al. (1977), quienes propusieron dicho método para la estimación de parámetros en datos incompletos.

2.2.3. Estimación de parámetros vía EM

McLachlan y Krishnan (2008) denotan la formulación del algoritmo EM y diversas extensiones. Al tener un vector aleatorio \mathbf{Y} que corresponde a los datos observados \mathbf{y} , con una función de densidad de probabilidad propuesta como $g(\mathbf{y}; \Psi)$, en donde $\Psi = (\Psi_1, \dots, \Psi_d)^\top$ es un vector de parámetros desconocidos en un espacio de parámetros Ω .

El algoritmo EM es un proceso iterativo con diversas aplicaciones para la estimación de parámetros en contextos de estimación por máxima verosimilitud, especialmente en situaciones en donde los datos presentan valores perdidos que dificultan la estimación directa. El vector de datos observados \mathbf{y} se considera como incompleto y como una función observable de datos completos. En la mayoría de casos, los datos incompletos se refieren a la presencia de valores perdidos. En este sentido, \mathbf{x} representa el vector de datos completos y \mathbf{z} denota el vector de datos adicionales, referidos como los datos no observados o perdidos. Es importante mencionar que, incluso en ausencia de valores perdidos, emplear el algoritmo EM puede resultar en una estimación más sencilla de los parámetros de máxima verosimilitud.

Sea $g_c(\mathbf{x}; \Psi)$ la distribución de densidad de probabilidad de un vector aleatorio \mathbf{X} correspondiente al vector de datos completos \mathbf{x} . De esta manera, la función de log-verosimilitud de los datos completos puede ser formada por Ψ si \mathbf{x} fuera completamente observado dado

$$\log \ell_c(\Psi) = \log g_c(\mathbf{x}; \Psi).$$

Formalmente, se tienen dos espacios muestrales \mathcal{X} y \mathcal{Y} . En lugar de observar el vector de datos completos \mathbf{x} en \mathcal{X} , se observa el vector de datos incompletos $\mathbf{y} = \mathbf{y}(\mathbf{x})$ en \mathcal{Y} . De esa manera se sigue que

$$g(\mathbf{y}; \Psi) = \int_{\mathcal{X}} g_c(\mathbf{y}; \Psi) dx,$$

en donde $\mathcal{X}(\mathbf{y})$ es el subconjunto de \mathcal{X} determinado por la ecuación $\mathbf{y} = \mathbf{y}(\mathbf{x})$.

El algoritmo EM aborda este problema al resolver la verosimilitud de los datos incompletos de manera indirecta al proceder de manera iterativa en términos de la función de verosimilitud ℓ de los datos completos $\log \ell_c(\Psi)$. Como no es observable, se reemplaza por su esperanza condicional dado \mathbf{y} , utilizando el ajuste actual para Ψ . De manera específica, si $\Psi^{(0)}$ representa un valor inicial para Ψ , entonces en la primera iteración, el paso E requiere la estimación de

$$Q(\Psi; \Psi^{(0)}) = E_{\Psi^{(0)}}(\log \ell_c(\Psi) | \mathbf{y}). \quad (2.3)$$

Por su parte, el paso M requiere de la maximización de $Q(\Psi; \Psi^{(0)})$ con respecto a Ψ sobre el espacio de parámetros Ω . Esto implica escoger un valor para $\Psi^{(1)}$ de modo que, para todo $\Psi \in \Omega$, se cumple lo siguiente:

$$Q(\Psi^{(1)}; \Psi^{(0)}) \geq Q(\Psi; \Psi^{(0)}). \quad (2.4)$$

Posteriormente, los pasos E y M se vuelven a ejecutar, pero en esta ocasión, $\Psi^{(0)}$ reemplaza el valor actual de $\Psi^{(1)}$. Así, en la iteración $(k + 1)$, los pasos E (2.3) y M (2.4) son definidos de la siguiente manera:

Paso E. Calcular $Q(\Psi; \Psi^{(k)})$, en donde:

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}}(\log \ell_c(\Psi) | \mathbf{y}).$$

Paso M. Escoger $\Psi^{(k+1)}$ para que sea cualquier valor de $\Psi \in \Omega$ que maximice $Q(\Psi; \Psi^{(k)})$, es decir, para todo $\Psi \in \Omega$:

$$Q(\Psi^{(k+1)}; \Psi^{(k)}) \geq Q(\Psi; \Psi^{(k)}).$$

También se puede expresar como si $\Psi^{(k+1)}$ pertenece al conjunto de puntos que maximiza $Q(\Psi; \Psi^{(k)})$:

$$\mathcal{M}(\Psi^{(k)}) = \operatorname{argmax}_{\Psi} Q(\Psi; \Psi^{(k)}).$$

Los pasos E y M son alternados de manera iterativa hasta que se alcance un nivel de convergencia. Por ejemplo, la diferencia

$$\ell(\Psi^{(k+1)}) - \ell(\Psi^{(k)})$$

cambie por una cantidad pequeña arbitraria en el caso de la convergencia de valores de la secuencia de valores de verosimilitud $\ell(\Psi^{(k)})$. Dempster et al. (1977) demostraron que la función de verosimilitud de los datos incompletos $\ell(\Psi)$ no decrece después de una iteración EM, es decir,

$$\ell(\Psi^{(k+1)}) \geq \ell(\Psi^{(k)}),$$

para $k = 0, 1, 2, \dots$

Como resultado, no es necesario especificar el mapeo completo de \mathcal{X} a \mathcal{Y} , ni la representación correspondiente de la densidad de los datos incompletos g a los datos completos g_c . Lo único que se requiere es de la especificación de un vector de datos completos \mathbf{x} y la densidad condicional de \mathbf{X} dado el vector de datos observados \mathbf{y} . La especificación de la densidad condicional es requerida para llevar a cabo el paso E. Como la elección de un vector de datos completos \mathbf{x} no es única, puede escogerse por conveniencia computacional con respecto a la implementación de los pasos E y M.

2.2.4. Algoritmo EM en una mezcla de distribuciones normales univariadas

En este ejemplo proporcionado por McLachlan y Peel (2000) se considera que la densidad de los g componentes son distribuciones normales univariadas con medias desconocidas μ_1, \dots, μ_g y varianza común desconocida σ^2 . De esta manera $f_i(x)$ puede ser expresado como $f_i(x; \zeta_i)$ en donde $\zeta_i = (\mu_i, \sigma^2)^\top$ y $\zeta = (\mu_1, \dots, \mu_g, \sigma^2)^\top$ contiene los parámetros desconocidos en las g densidades de los componentes normales. Así, el vector Ψ que contiene todos los parámetros desconocidos puede definirse como:

$$\Psi = (\zeta^\top, \pi_1, \dots, \pi_{g-1})^\top.$$

En este sentido, el modelo de mezcla con densidades normales corresponde a

$$f(w; \Psi) = \sum_{i=1}^g \pi_i f_i(x; \zeta_i).$$

Para este ejemplo particular se considera que

$$\begin{aligned} f_i(x; \zeta_i) &= \phi(x; \mu_i, \sigma^2), \\ f_i(x; \zeta_i) &= (2\pi\sigma^2)^{-1/2} \exp(-1/2(x - \mu_i)^2/\sigma^2). \end{aligned}$$

Además, el vector de datos completos \mathbf{x} se define como $\mathbf{x} = (\mathbf{y}^\top, \mathbf{z}^\top)^\top$, en donde \mathbf{z} corresponde a un vector de datos no observables $\mathbf{z} = (\mathbf{z}_1^\top, \dots, \mathbf{z}_n^\top)$ de valores 0 y 1 que indica si x_j pertenece o no al componente i de la mezcla ($i = 1, \dots, g; j = 1, \dots, n$). De la misma manera, la log-verosimilitud, como se expresó en (2.1), para Ψ tiene la forma multinomial

$$\ell(\Psi) = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log \pi_i + C.$$

De modo que C no depende de Ψ y se expresa como:

$$C = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log f_i(\mathbf{x}_j).$$

La función de verosimilitud para Ψ en el presente ejemplo está dada por

$$\begin{aligned} \ell(\Psi) &= \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log f_i(x_j; \zeta_i). \\ &= -\frac{1}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^g \sum_{j=1}^n z_{ij} (\log \sigma^2 + (x_j - \mu_i)^2/\sigma^2). \end{aligned}$$

Es por esta razón que el paso E, como se define en (2.3), para la iteración $k + 1$ requiere del cálculo del valor condicional esperado actual de Z_{ij} dado los datos observados \mathbf{y} , en donde Z_{ij} es una variable aleatoria correspondiente a z_{ij} .

$$E_{\Psi^{(k)}} = p_{\Psi^{(k)}}(Z_{ji} = 1 | \mathbf{y}) = z_{ij}^{(k)}.$$

Que por Teorema de Bayes implica obtener

$$z_{ij}^{(k)} = \tau_i(\mathbf{x}_j; \Psi^{(k)}),$$

en donde $\tau_i(\mathbf{x}_j; \Psi^{(k)}) = \pi_i^{(k)} f_i(\mathbf{x}_j) / f(\mathbf{x}_j; \Psi^{(k)})$, para $i = 1, \dots, g, j = 1, \dots, n$. La cantidad $\tau_i(\mathbf{x}_j; \Psi^{(k)})$ es la probabilidad a posteriori de que el miembro j de la muestra que presenta un valor observado \mathbf{x}_j pertenezca al componente i de la mixtura.

El Paso M (2.4) en la iteración $k + 1$ simplemente requiere reemplazar cada z_{ij} por $z_{ij}^{(k)}$, lo que da como resultado que

$$\pi_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} / n.$$

Asimismo, en este ejemplo particular deben estimarse los valores de $\mu_1^{(k+1)}, \dots, \mu_g^{(k+1)}$ y $\sigma^{(k+1)^2}$, junto con $\pi_1^{(k+1)}, \dots, \pi_{g-1}^{(k+1)}$, y maximizar $Q(\Psi; \Psi^{(k)})$

Ahora, los estimadores de máxima verosimilitud para μ_i y σ^2 si z_{ij} es observable corresponden respectivamente a

$$\sum_{j=1}^n z_{ij} x_j / \sum_{j=1}^n z_{ij},$$

y

$$\sum_{i=1}^g \sum_{j=1}^n z_{ij} (x_j - \mu)^2 / n.$$

Mientras $\ell(\Psi)$ sea lineal en z_{ij} , se sigue que estas funciones pueden ser reemplazadas por sus condicionales esperados actuales $z_{ij}^{(k)}$ en donde las estimaciones actuales $\tau_i(\mathbf{x}_j; \Psi^{(k)})$ de las probabilidades a posteriori de la membresía a un componente i de la mixtura están dados por

$$\tau_i(\mathbf{x}_j; \Psi^{(k)}) = \pi_i^{(k)} f_i(x_j; \zeta_i^{(k)}) / f_i(x_j; \Psi^{(k)}), (i = 1, \dots, g).$$

Como resultado se obtiene:

$$\mu_i^{(k+1)} = \sum_{j=1}^n z_{ij}^{(k)} x_j / \sum_{j=1}^n z_{ij}^{(k)}, (i = 1, \dots, g),$$

y

$$\sigma^{(k+1)^2} = \sum_{i=1}^g \sum_{j=1}^n z_{ij}^{(k)} (x_j - \mu_i^{(k+1)})^2 / n.$$

2.2.5. Estimación de parámetros vía ECM

Tras la popularidad del algoritmo EM y sus diversas aplicaciones, varios autores propusieron extensiones del método como el EM estocástico (SEM) (Nielsen, 2000), Clasificación EM (CEM) (Hunter y Lange, 2012), EM generalizado (GEM) (Tanner y Wong, 1987), etc. En este capítulo se introduce el algoritmo maximización de esperanza condicional (ECM) (Tanner y Wong, 1987). En términos sencillos, el algoritmo ECM reemplaza el paso de maximización M del algoritmo EM por una serie de pasos de maximización condicional CM que tienen mayor sencillez computacional. Como resultado, la convergencia en ECM puede requerir más iteraciones que en EM, pero computacionalmente el proceso es más rápido. Este es uno de los motivos por los cuales el algoritmo ECM es una extensión recomendada, especialmente en donde la estimación de máxima verosimilitud de los datos completos es complicada.

Meng y Rubin (1993) introdujeron esta extensión del algoritmo EM. El paso M es reemplazado por $S > 1$ pasos. Así, $\Psi^{(k+s/S)}$ denota el valor de Ψ en el paso s CM de la iteración $(k + 1)$, en donde $\Psi^{(k+s/S)}$ es escogido para maximizar

$$Q(\Psi; \Psi^{(k)}),$$

sujeto a la restricción:

$$g_s(\Psi) = g_s(\Psi^{(k+(s-1)/S)}).$$

De este modo, el conjunto $C = g_s(\Psi), s = 1, \dots, S$ es un conjunto de S preseleccionadas funciones de vectores. Así, $\Psi^{(k+s/S)}$ satisface:

$$Q(\Psi^{(k+s/S)}; \Psi^{(k)}) \geq Q(\Psi; \Psi^{(k)}) \forall \Psi \in \Omega_s(\Psi^{(k+(s-1)/S)}),$$

donde

$$\Omega_s(\Psi^{(k+(s-1)/S)}) \equiv \Psi \in \Omega : g_s(\Psi) = g_s(\Psi^{(k+(s-1)/S)}).$$

El valor de Ψ en el paso CM final $\Psi^{(k+S/S)} = \Psi^{(k+1)}$ se toma como input en la iteración $(k + 2)$.

Un ejemplo de la aplicación del algoritmo ECM en mixturas de distribuciones puede ser consultados en los trabajos de McLachlan y Peel (2000), y Tong y Tortora (2022).

2.2.6. Métodos de selección del modelo

Uno de los grandes desafíos de la clusterización consiste en determinar la cantidad de particiones que deben realizarse. No obstante, al emplear una clusterización basada en modelos, la selección de la cantidad de particiones puede ser abordada de manera formal. Para ello, es importante hacer una similitud con la cuestión de seleccionar la cantidad de componentes de un modelo de mixtura. En estos casos, una mixtura con muchos componentes puede sobreajustarse a los datos; mientras que, muy pocos componentes pueden no ser lo suficiente flexibles para aproximarse al modelo de verdadero. Por estos motivos, existen procedimientos estadísticos para inferir el número de componentes (Demg y Han, 2014).

Sea M_k la clase de todas las posibles mixturas de k componentes construidas para un tipo particular de distribución. El criterio de máxima verosimilitud no puede ser empleado para estimar directamente la cantidad de componentes porque estos se encuentran anidados ($M_k \subseteq M_{k+1}$) y la maximización de la log-verosimilitud $-\log p(\mathbf{X}|\zeta_{ML})$ es una función no decreciente de k .

En su lugar, las aproximaciones propuestas para determinar la cantidad de componentes pueden ser clasificados en métodos determinísticos y probabilísticos (Demg y Han, 2014). Los métodos determinísticos comienzan al obtener un conjunto de modelos candidatos para un rango de valores de k (desde k_{min} hasta k_{max} en donde se asume que se encuentra la cantidad óptima. Posteriormente, el número de componentes es seleccionado sobre la base de:

$$k^* = \operatorname{argmin}_k(C(\zeta_k, k), k = k_{min}, \dots, k_{max},$$

donde ζ_k es una estimación de los parámetros de la mixtura con k componentes y $C(\zeta_k, k)$ es un criterio de selección de modelos. Un criterio común puede ser expresado de la forma:

$$C(\zeta_k, k) = -\log p(\mathbf{X}|\zeta_k) + \mathcal{P}(k),$$

donde $\mathcal{P}(k)$ es una función que penaliza los valores altos de k . Existen diversos criterios empleados en modelos de mixtura finita como el criterio empírico de Laplace (LEC), criterio de inferencia bayesiano de Schwarz (BIC), longitud mínima de descripción (MDL) de Rissanen, longitud mínima del mensaje (MML), Criterio de Información de Akaike (AIC), criterio de probabilidad completa integrada (ICL), información de Kullback criterio (KIC).

En contraste a los métodos determinísticos, los métodos estocásticos se basan en simulaciones Monte Carlo con cadenas de Markov (MCMC) en donde pueden aplicarse los criterios de selección de modelos, o seguir el modelamiento bayesiano tras muestrear repetidamente de la distribución a posteriori.

Capítulo 3

Distribución normal contaminada

Este capítulo presenta la distribución normal contaminada, así como el caso de la familia de mixturas normales contaminadas multivariadas, detallando sus correspondientes funciones de densidad.

3.1. Distribución normal contaminada univariada

El origen de la distribución normal contaminada se acredita a los trabajos de Newcomb Simon, un astrónomo de origen canadiense reconocido por sus aportes sobre el tratamiento de outliers en estadística, aplicación de teoría de probabilidad para la interpretación de datos y al desarrollo de lo que hoy se denomina como estimación robusta en estadística (Dodge, 2008). Uno de los trabajos más destacados de Newcomb es la colección *Notes on the Theory of Probabilities* (Mathematics Monthly, 1859-1861) en donde se presentan aportes que hasta el día de hoy se consideran modernos. Entre las distintas aproximaciones planteadas en estas notas, Newcomb plantea un método novedoso para el tratamiento de outliers en datos obtenidos de la disciplina de astronomía. En este procedimiento, el astrónomo reconoció que la distribución normal no se ajustaba apropiadamente a los datos de su disciplina en donde era frecuente encontrar outliers; por ello, propuso la distribución normal contaminada como una alternativa para su modelamiento.

Con el objetivo de ilustrar las características de la distribución normal contaminada, primero se repasa la función de densidad de una distribución normal, definida como:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), x \in \mathbb{R}, \quad (3.1)$$

en donde μ representa la media aritmética, como un parámetro de localización y σ^2 representa la varianza, como un parámetro de escala que determina la forma de la distribución (Mavrakakis y Penzer, 2021).

En el caso univariado, una distribución normal contaminada corresponde a una mixtura finita de dos distribuciones normales con el mismo parámetro de localización, pero distintas varianzas (Everitt y Skrondal, 2010). Por ejemplo, en la Figura 3.1 se presentan dos distribuciones normales con la misma media aritmética μ y distintas varianzas σ^2 .

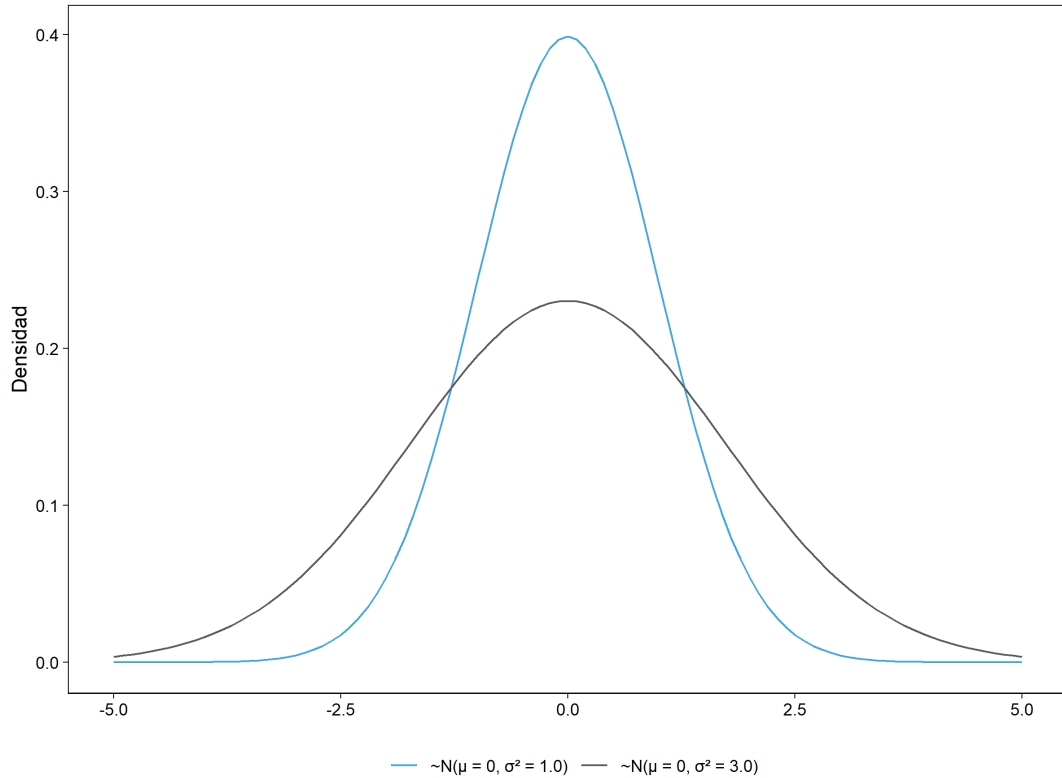


Figura 3.1: Dos distribuciones normales con mismo parámetro de localización $\mu = 0$, pero distinto parámetro de escala $\sigma^2 = 1$ y $\sigma^2 = 3$, respectivamente

En este sentido, la distribución normal con menor varianza se considera como la distribución de referencia o el primer componente de la mixtura finita, y el segundo componente, la distribución normal con mayor varianza representa aquella que permitirá transformar las colas de la distribución de referencia de ligeras a más pesadas. Tukey (1960) establece la función de densidad de la distribución normal contaminada como una mixtura de dos distribuciones normales, cada una con densidades expresadas como en (3.1), en donde se considera dos parámetros adicionales α y η . Aquí α representa la proporción de *contaminación* de los datos, o proporción de datos con mayor varianza y η representa un parámetro de escala para la varianza del segundo componente. Como se espera que la varianza del segundo componente sea mayor a la del primero, entonces se espera que $\eta > 1$ (McLachlan y Peel, 2000). De esta manera, la función de densidad de la distribución normal contaminada para el caso univariado puede expresarse como:

$$f_{CN}(x) = (1 - \alpha)\phi(x, \mu, \sigma^2) + \alpha\phi(x, \mu, \eta\sigma^2), x \in \mathbb{R}.$$

Así, la distribución normal puede considerarse como un caso especial de la distribución normal contaminada cuando $\alpha = 1$ y $\eta = 1$. En el ejemplo presentado en la Figura 3.1, la varianza del primer componente es $\sigma^2 = 1$; mientras que la varianza del segundo componente es $\sigma^2 = 3$; por lo tanto, si se deseará generar una mixtura de ambas considerando la densidad de la distribución normal contaminada, el parámetro de escala es $\eta = 3$. En la Figura 3.2 se presenta la distribución normal contaminada con ambos componentes y un $\alpha = 0,20$.

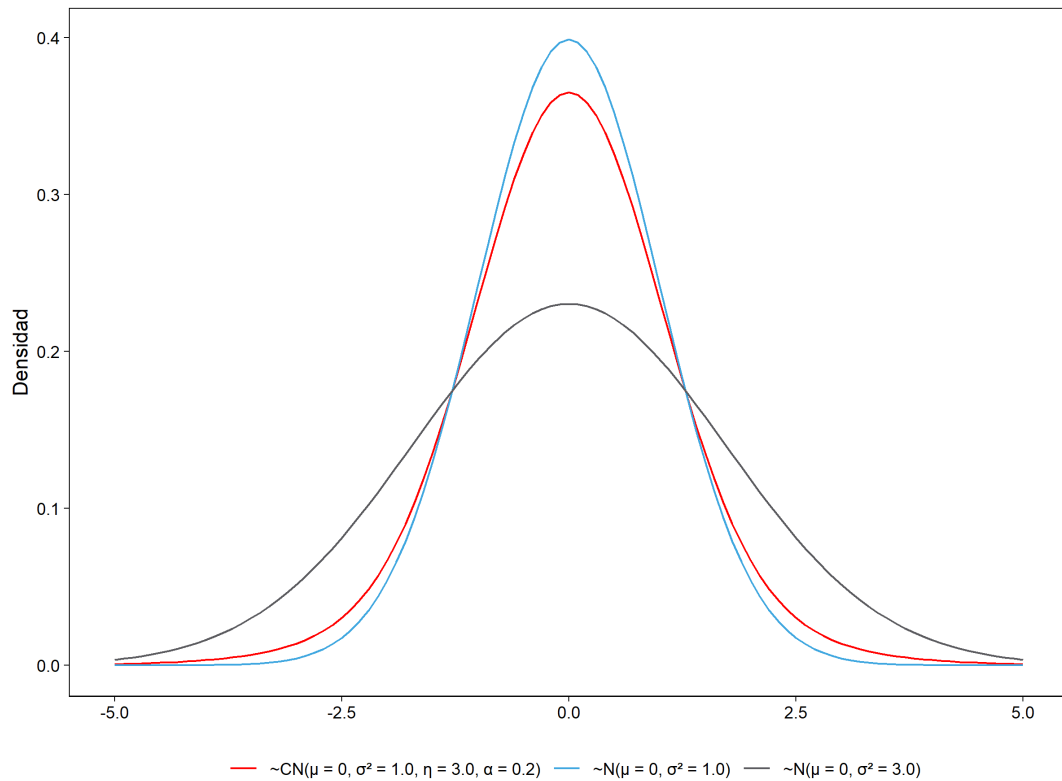


Figura 3.2: Una distribución normal contaminada (en rojo) como resultado de dos componentes normales, con parámetros $\mu = 0$, $\sigma^2 = 1$, $\eta = 3$ y $\alpha = 0,20$

Como se aprecia en la Figura 3.2, el primer componente de la mezcla es representado por la densidad en celeste y corresponde a la distribución de referencia. El segundo componente es representado por la densidad en negro que mantiene la misma media, pero tiene mayor varianza. Finalmente, la distribución normal contaminada representada por la densidad en rojo indica la mezcla finita de los dos componentes previamente mencionados. En este ejemplo es posible observar que la distribución normal contaminada tiene colas más pesadas que la distribución normal del primer componente de la mezcla, lo que permite un mejor modelamiento de los outliers.

3.2. Distribución normal contaminada multivariada

Como se mencionó anteriormente, la motivación original de la distribución fue el modelamiento de datos con presencia de outliers que pudieran dificultar el ajuste de un modelo normal convencional (Dodge, 2008). No obstante, aplicaciones modernas de la distribución normal contaminada ocurren en datos multivariados, en donde la distribución normal multivariada tiende a ser una de las alternativas tradicionales más empleadas en múltiples algoritmos. La distribución normal multivariada es susceptible ante la presencia de outliers; por ello, la aproximación contaminada representa una mejor alternativa para una estimación robusta de parámetros.

Un vector aleatorio de d dimensiones $\mathbf{X} = (X_1, \dots, X_d)^\top$ sigue una distribución normal multivariada con un vector de medias $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$, si su función de densidad de probabilidad conjunta es dada por:

$$f_{MN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp(-1/2(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})), x \in \mathbb{R}^d,$$

donde $\boldsymbol{\Sigma}$ es una matriz de covarianzas simétrica y no negativa definida. Esto implica que $\mathbf{a}' \boldsymbol{\Sigma} \mathbf{a} \geq \mathbf{0}$ para todo vector \mathbf{a} , Además, se asume que $\boldsymbol{\Sigma}$ es una matriz positiva definida, es decir que $\mathbf{a}' \boldsymbol{\Sigma} \mathbf{a} > \mathbf{0}$ para todo $\mathbf{a} \neq \mathbf{0}$. De la misma manera, esta matriz es no singular y su determinante es positivo $|\boldsymbol{\Sigma}| > 0$.

Al igual que en el caso univariado, una de las limitaciones de la distribución normal multivariada corresponde a sus colas ligeras que dificultan el modelamiento de datos con outliers. Es para estos mismos casos que la distribución normal contaminada multivariada ha sido propuesta como alternativa, En resumen, una distribución normal contaminada multivariada representa una simple generalización de la distribución normal multivariada que permite modelar colas más pesadas. Como se define en Tong y Tortora (2022), un vector aleatorio con d dimensiones $\mathbf{X} = (X_1, \dots, X_d)^\top$ sigue una distribución normal contaminada multivariada con un vector de medias $\boldsymbol{\mu}$, matriz de escala $\boldsymbol{\Sigma}$, proporción de observaciones buenas (no outliers) $\alpha \in (0, 1)$ y grado de contaminación $\eta > 1$, si su función de densidad de probabilidad conjunta es dada por:

$$f_{MCN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \eta) = \alpha f_{MN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1 - \alpha) f_{MN}(\mathbf{x}; \boldsymbol{\mu}, \eta \boldsymbol{\Sigma}),$$

donde $f_{MN}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denota la función de probabilidad conjunta de un vector aleatorio con d dimensiones que sigue una distribución normal multivariada (MN) con un vector de medias $\boldsymbol{\mu}$ y una matriz de covarianzas $\boldsymbol{\Sigma}$.

Es importante notar que la notación de Tong y Tortora (2022) difieren de la notación de Tukey (1960) en el establecimiento del parámetro α . En esta nomenclatura, Tong y Tortora (2022) consideran a α como la proporción de casos no contaminados; mientras que Tukey (1960) la definía como la proporción de casos contaminados. Esta diferenciación no tiene severas implicancias, pues $\alpha \in (0, 1)$, y por lo general se espera que $\alpha \in (0,5, 1)$ en la nomenclatura de Tong y Tortora (2022), lo que indica que hay más de la mitad de casos con buenas observaciones no contaminadas.

Al igual que en el caso univariado, la distribución normal contaminada multivariada re-

presenta una mezcla de dos componentes en donde un componente, con probabilidad α representa las "buenas" observaciones y el otro componente, con una probabilidad $1 - \alpha$ representa las "malas" observaciones u outliers. Ambos componentes comparten el mismo vector de medias aritméticas $\boldsymbol{\mu}$, pero el componente que representa a las malas observaciones tiene una matriz de covarianzas inflada $\eta\boldsymbol{\Sigma}$. De esta manera, la MN es un caso especial de la MCN cuando $\alpha = 1$ y $\eta = 1$ (Punzo y Tortora, 2021).

En el conjunto de parámetros $\vartheta = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \alpha, \eta)$, una vez estimados $\hat{\vartheta} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\alpha}, \hat{\eta})$, es posible establecer si un punto genérico \mathbf{x}^* constituye una buena observación (no outlier) considerando la probabilidad a posteriori (Punzo y Tortora, 2021):

$$P(x^* \text{ sea bueno} \mid \hat{\vartheta}) = \hat{\alpha} f_{MN}(\mathbf{x}^*; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) / f_{MNC}(\mathbf{x}^*; \hat{\vartheta}).$$

Así, x^* puede ser considerado como bueno si $P(x^* \text{ sea bueno} \mid \hat{\vartheta}) > 1/2$, de lo contrario será considerado como un punto malo, o outlier.

3.3. Mezcla de distribuciones normales contaminadas multivariadas

Las técnicas de clusterización propuestas sobre la base de distribuciones normales contaminadas multivariadas emplean como metodología una mezcla de dichas distribuciones y asumen que cada cluster corresponde a uno de los componentes de la mezcla (Punzo y Tortora, 2021). Punzo y McNicholas (2016) presentan el caso de una mezcla de G distribuciones normales contaminadas (MCNM). Este modelo puede ser definido como:

$$f_{MCNM}(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{g=1}^G \pi_g [\alpha_g \phi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + (1 - \alpha_g) \phi(\mathbf{x}; \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g)],$$

donde, para cada componente g , π_g representa la proporción combinada que cumple $\pi_g > 0$ y que $\sum_{g=1}^G \pi_g = 1$, $\alpha_g \in (0, 1)$ corresponde a la proporción de buenas observaciones y $\eta_g > 1$ denota su grado de contaminación. En esta ecuación, $\boldsymbol{\Psi}$ contiene a todos los parámetros de la mezcla, mientras que $\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ representa la distribución de un vector de d variables aleatorias normales con un vector de medias $\boldsymbol{\mu}$ y una matriz de covarianza $\boldsymbol{\Sigma}$ (Punzo et al., 2018). Una nomenclatura equivalente presentada por Tong y Tortora (2022) para esta distribución es:

$$f_{MCNM}(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{g=1}^G \pi_g f_{MCN}(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \alpha_g, \eta_g).$$

Este es un modelo generalizado para mezclas de distribuciones normales contaminadas multivariadas, en donde $\alpha \rightarrow 1^-$ y $\eta_g \rightarrow 1^+$, para cada componente $g = 1, \dots, G$, se obtendrá mezclas de distribuciones normales multivariadas clásicas (Punzo et al., 2018). Adicionalmente, Punzo y McNicholas (2016) propusieron una familia de catorce distribuciones de mezcla para distribuciones normales contaminadas multivariadas más parsimoniosas sobre la base de la descomposición espectral:

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g^\top,$$

donde $\lambda_g = |\Sigma_g|^{1/p}$, Δ_g es la diagonal escalada ($|\Delta_g| = 1$) de la matriz de autovalores de Σ_g ordenada en orden decreciente, y Γ_g es una matriz $p \times p$ ortogonal cuyas columnas son los autovectores normalizados de Σ_g , ordenados de acuerdo a sus respectivos autovalores. En este sentido, λ_g determina el tamaño o volumen de un cluster, Δ_g determina su forma, y Γ_g determina su orientación.

Cuadro 3.1: Versiones Parsimoniosas del Modelo General de Mixtura de Distribuciones Normales Contaminadas Multivariadas

| Familia | Modelo | Volumen | Forma | Orientación | Σ_g | Parámetros libres en Σ_g |
|----------|--------|----------|----------|-----------------|---|---------------------------------|
| Esférica | EII | Igual | Esférica | - | $\lambda \mathbf{I}$ | 1 |
| | VII | Variable | Esférica | - | $\lambda_g \mathbf{I}$ | G |
| Diagonal | E EI | Igual | Igual | Alineada al eje | $\lambda \Delta$ | p |
| | VEI | Variable | Igual | Alineada al eje | $\lambda_g \Delta$ | $G + p - 1$ |
| | EVI | Igual | Variable | Alineada al eje | $\lambda \Delta_g$ | $1 + G(p - 1)$ |
| | VVI | Variable | Variable | Alineada al eje | $\lambda_g \Delta_g$ | Gp |
| General | EEE | Igual | Igual | Igual | $\lambda \Gamma \Delta \Gamma^\top$ | $p(p + 1)/2$ |
| | VEE | Variable | Igual | Igual | $\lambda_g \Gamma \Delta \Gamma^\top$ | $G + p - 1 + p(p - 1)/2$ |
| | EVE | Igual | Igual | Igual | $\lambda \Gamma \Delta_g \Gamma^\top$ | $1 + G(p - 1) + p(p - 1)/2$ |
| | EEV | Igual | Variable | Variable | $\lambda \Gamma_g \Delta \Gamma_g^\top$ | $p + Gp(p - 1)/2$ |
| | VVE | Variable | Igual | Igual | $\lambda_g \Gamma \Delta_g \Gamma^\top$ | $Gp + p(p - 1)/2$ |
| | VEV | Variable | Variable | Variable | $\lambda_g \Gamma_g \Delta \Gamma_g^\top$ | $G + p - 1 + Gp(p - 1)/2$ |
| | EVV | Igual | Variable | Variable | $\lambda \Gamma_g \Delta_g \Gamma_g^\top$ | $1 + G(p - 1) + Gp(p - 1)/2$ |
| | VVV | Variable | Variable | Variable | $\lambda_g \Gamma_g \Delta_g \Gamma_g^\top$ | $Gp(p + 1)/2$ |

La lógica detrás de los catorce modelos que constituyen la familia de modelos de mixtura de distribuciones normales contaminadas multivariadas es ajustar todos los modelos para distintos valores de G y utilizar algún criterio para seleccionar al que mejor se ajusta a los datos McNicholas, 2016. En este caso, el criterio más popular es el criterio de información bayesiano (BIC) (Schwarz, 1978):

$$BIC = 2 \log(\hat{\vartheta}) - \rho \log(n),$$

donde $\hat{\vartheta}$ es la estimación de máxima verosimilitud de ϑ , $l(\hat{\vartheta})$ es la log-verosimilitud maximizada y ρ es la cantidad de parámetros libres. No obstante, en la práctica, Punzo y McNicholas (2016) trabajan con el modelo VVV (ver Tabla 3.1), que representa la versión menos restringida para demostrar las propiedades del modelo en comparación con la distribución normal multivariada y la distribución t multivariada como aproximaciones para la clusterización basada en modelos. En este sentido, una descripción más detallada de cada uno de los 14 modelos y sus propiedades puede consultarse en Punzo y McNicholas (2016).

Capítulo 4

Distribución t

En este capítulo se introduce la distribución t univariada y multivariada empleadas en la clusterización basada en modelos. La familia de modelos es presentada como una aproximación ante la distribución normal para el caso en donde se presentan valores extremos.

4.1. Distribución t univariada

La distribución t se conoce también como la distribución t de Student. Si Y y Z denotan variables aleatorias independientes, en donde Y es una variable con una distribución normal estándar y Z corresponde a una distribución chi-cuadrado con n grados de libertad, la variable aleatoria $X = \frac{Y}{\sqrt{\frac{Z}{n}}}$ tiene una distribución t. En este sentido, la distribución de densidad de una variable con una distribución t con n grados de libertad se caracteriza por:

$$f_t(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}}, \quad -\infty < x < \infty.$$

Si n corresponde a un valor grande (i.e., $n \geq 30$), la función de densidad de la distribución t tiene una forma de campana simétrica centrada en cero que se aproxima a una distribución normal estándar, pero con colas ligeramente más pesadas. En efecto, cuando una distribución normal representa a una población, la distribución t describe muestras extraídas de dicha población. Asimismo, la distribución t será distinta dependiendo del tamaño de muestra determinado y conforme el tamaño de muestra se hace más grande, la distribución t se aproxima a la distribución normal. En efecto, la distribución normal estándar corresponde al valor límite de la distribución t cuando $n \rightarrow \infty$ (Grami, 2020). En cuestiones prácticas, la distribución t se emplea como alternativa ante la distribución normal para estimar la media aritmética de una población cuando se desconoce la variabilidad poblacional y el tamaño de muestra es pequeño. Asimismo, tras tener colas más pesadas que una distribución normal estándar, es preferible para el modelamiento de datos con forma normal, pero con presencia de valores extremos (Punzo y Tortora, 2021).

4.2. Distribución t multivariada

La distribución t multivariada es una generalización de la distribución t para el caso de dos o más variables. Dado un vector aleatorio de d variables $\mathbf{X} = (X_1, \dots, X_d)^\top$, este tendrá una distribución t multivariada con un vector de medias $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^\top$ y una matriz de covarianzas $\nu(\nu - 2)^{-1}\Sigma$, $\nu > 2$, denotada por $T_\nu(\boldsymbol{\mu}, \Sigma, d)$, si tiene una función de densidad de probabilidad dada por:

$$f_{Mt}(x) = \frac{\Gamma(\frac{1}{2}(\nu + d))}{(\pi\nu)^{\frac{1}{2}\nu} \Gamma(\frac{1}{2}\nu) |\Sigma|^{\frac{1}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\nu} \right)^{-\frac{1}{2}(\nu + d)}, \nu > 2.$$

En este sentido, cuando $T_\nu(0, 1, 1) = t_\nu$ es la distribución t de Student univariada con ν grados de libertad (Lin, 1972). La distribución t multivariada es usualmente empleada como un sustituto de la distribución normal multivariada en situaciones en donde se sabe que la distribuciones marginales de las variables individuales tienen colas más pesadas que las que se pueden modelar con una distribución normal (Golam Kibria y Joarder, 2006).



Capítulo 5

Estudio de simulación

En este capítulo se introduce un estudio de simulación con el objetivo de demostrar las características del modelo normal contaminado multivariado como aproximación para la clusterización basada en modelos. Este estudio consiste en la adaptación de la metodología inicialmente planteada por Punzo y McNicholas (2016) en donde se comparan distintas aproximaciones para implementar una clusterización basada en modelos, pero empleando como modificación la inclusión de valores perdidos, tal y como se desarrolló en Punzo y Tortora (2021).

5.1. Características de la simulación

En este experimento Monte Carlo se simulan conjuntos de datos de 200 casos y dos dimensiones ($d = 2$) con un número fijo de dos conglomerados ($G = 2$). Así, se contrastarán los métodos de clusterización considerando tres mixturas: una mixtura de modelos normales multivariados, una mixtura de modelos t multivariados y una mixtura de distribuciones normales multivariadas.

Para la simulación de estas tres mixturas se considerará tanto condiciones en común como condiciones específicas. En primer lugar, las condiciones en común para las tres mixturas propuestas fueron los parámetros $\pi_1 = 0,3$ y $\pi_2 = 0,7$ que representan la ponderación del primer y segundo componente de cada mixtura, respectivamente. De la misma manera, se definieron los vectores de medias aritméticas de cada componente de la mixtura de dos dimensiones como $\boldsymbol{\mu}_1 = (0, -3)^\top$ y $\boldsymbol{\mu}_2 = (0, 3)^\top$. El último aspecto en común para estas tres mixturas corresponde a la matriz de varianzas y covarianzas del primer y segundo componente, establecidos como $\begin{pmatrix} 1 & -0,5 \\ -0,5 & 1 \end{pmatrix}$ y $\Sigma_1 = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$, respectivamente.

Del mismo modo, en los tres escenarios se consideró reemplazar el 10 % de las observaciones por valores perdidos a través de la función `ampute()` del paquete `mice`. Dichos valores perdidos son generados a través de patrones aleatorios, que Rubin (1976) denomina como Missing at Random (MAR). Asimismo, se consideraron 1000 réplicas en la simulación y tres condiciones que varían en función al método empleado para estimar los valores iniciales del algoritmo (i.e., k-means, k-medoids y hierarchical).

Dichos parámetros fueron la base para la simulación de mezclas correspondiente a tres modelos multivariados, cuyas características se describen a continuación.

5.1.1. Condiciones para la mezcla de distribuciones normales multivariadas

La función `rmvnorm()` del paquete `mvtnorm` fue utilizada para la simulación de dos distribuciones normales multivariadas, con la ponderación indicada en los parámetros comunes. Asimismo, se reemplazó el 1% de los casos por $(0, x^*)$, en donde x^* representa un valor aleatorio de una distribución uniforme con parámetros $a = 10$ y $b = 15$.

5.1.2. Condiciones para la mezcla de distribuciones t multivariadas

La función `rmvt()` del paquete `mvtnorm` fue utilizada para la simulación de dos distribuciones t multivariadas, con la ponderación indicada en los parámetros comunes. Adicionalmente, se consideraron como parámetros $\nu_1 = 4$ y $\nu_2 = 10$ grados de libertad para cada distribución, respectivamente.

5.1.3. Condiciones para la mezcla de distribuciones normales contaminadas multivariadas

La función `rCN()` del paquete `ContaminatedMixt` fue utilizada para la simulación de dos distribuciones normales contaminadas multivariadas, con la ponderación indicada en los parámetros comunes. Adicionalmente, se consideraron como parámetros $\alpha_1 = 0,9$ y $\alpha_2 = 0,8$, que representan la ponderación del primer componente de la mezcla de distribuciones normales, o la proporción de observaciones que son consideradas no outliers; $\eta_1 = 20$ y $\eta_2 = 30$ como parámetros de escala entre las matrices de varianzas y covarianzas de los dos componentes normales. Los resultados de la primera simulación se presentan en el cuadro 5.1 en la cual se exhiben los escenarios considerados en la simulación que emplea el método k-means como aproximación para obtener los valores iniciales, junto con estimaciones del sesgo, desviación estándar y error cuadrático medio para los parámetros comunes y específicos en donde es pertinente.

En el cuadro 5.1 es posible apreciar cómo el sesgo de los parámetros de ponderación del primer componente de la mezcla π_1 y los parámetros de medias aritméticas de ambos componentes μ_1 y μ_2 son mayores cuando se emplea la distribución normal estándar multivariada como la estrategia para desarrollar una clusterización basada en una mezcla de dicho modelo. Esto ocurre en principio por la presencia de un 1% de valores extremos que fueron insertados en el conjunto de datos. En contraste, al utilizar una mezcla de distribuciones t multivariadas o distribuciones normales contaminadas multivariadas se identifica un menor grado de sesgo en los parámetros mencionados. Asimismo, es importante notar que las estimaciones al emplear una mezcla de distribuciones t multivariadas son ligeramente más precisas y menos variables que las estimaciones obtenidas la distribución normal contaminada multivariada como base para la clusterización. En efecto, el sesgo y la variabilidad de las estimaciones son ligeramente altos en los parámetros η_1 y η_2 al emplear una mezcla de distribuciones nor-

Cuadro 5.1: Resultados de la simulación sobre la clusterización basados en modelos normales multivariados, modelos t multivariados y modelos normales contaminados multivariados, utilizando k-means para obtener valores iniciales

| Escenario | Parámetro | Sesgo | DE | RMSE |
|---|------------------|----------|----------|----------|
| Mixtura de distribuciones normales multivariadas | $\pi_1 = 0,3$ | -0,11312 | 0,02287 | 0,11518 |
| | $\mu_{11} = 0$ | 0,18999 | 0,10869 | 0,21616 |
| | $\mu_{12} = -3$ | -0,23490 | 0,13263 | 0,26648 |
| | $\mu_{21} = 0$ | -0,04342 | 0,07003 | 0,15782 |
| | $\mu_{22} = 3$ | 0,64611 | 0,15782 | 0,66323 |
| Mixtura de distribuciones t multivariadas | $\pi_1 = 0,3$ | -0,00193 | 0,01237 | 0,00773 |
| | $\mu_{11} = 0$ | 0,00762 | 0,16288 | 0,10065 |
| | $\mu_{12} = -3$ | -0,00794 | 0,16106 | 0,09954 |
| | $\mu_{21} = 0$ | -0,00896 | 0,09514 | 0,05899 |
| | $\mu_{22} = 3$ | -0,00541 | 0,09711 | 0,06003 |
| | $v_1 = 4$ | 2,52731 | 3,02626 | 2,43501 |
| | $v_2 = 10$ | 1,36193 | 3,19719 | 2,14549 |
| Mixtura de distribuciones normales contaminadas multivariadas | $\pi_1 = 0,3$ | -0,00250 | 0,05276 | 0,05279 |
| | $\mu_{11} = 0$ | -0,12734 | 1,00930 | 1,00297 |
| | $\mu_{12} = -3$ | 0,01555 | 0,24369 | 0,24074 |
| | $\mu_{21} = 0$ | -0,01263 | 0,82188 | 0,83260 |
| | $\mu_{22} = 3$ | -0,08583 | 0,55940 | 0,57326 |
| | $\alpha_1 = 0,9$ | -0,02341 | 0,06260 | 0,06681 |
| | $\alpha_2 = 0,8$ | -0,02763 | 0,06481 | 0,07042 |
| | $\eta_1 = 20$ | -4,09064 | 10,02118 | 10,81929 |
| $\eta_2 = 30$ | 5,56338 | 20,09052 | 21,83691 | |

Nota. DE = Desviación estándar. RMSE = Error cuadrático medio.

males contaminadas multivariadas. Finalmente, se debe mencionar que pueden encontrarse problemas de convergencia en la estimación de parámetros cuando se emplea el algoritmo k-means como método para obtener los valores iniciales; por ello, es recomendable considerar otra aproximación.

Los resultados de la segunda simulación se presentan en el cuadro 5.2 en la cual se exhiben los escenarios considerados en la simulación que emplea el método k-medoids como aproximación para obtener los valores iniciales, junto con estimaciones del sesgo, desviación estándar y error cuadrático medio para los parámetros comunes y específicos en donde es pertinente.

Cuadro 5.2: Resultados de la simulación sobre la clusterización basada en modelos normales multivariados, modelos t multivariados y modelos normales contaminados multivariados, utilizando k-medoids para obtener valores iniciales

| Escenario | Parámetro | Sesgo | DE | RMSE |
|--|------------------|----------|----------|----------|
| Mezcla de distribuciones normales multivariadas | $\pi_1 = 0,3$ | -0,10694 | 0,02006 | 0,10880 |
| | $\mu_{11} = 0$ | 0,15615 | 0,20283 | 0,25589 |
| | $\mu_{12} = -3$ | -0,18892 | 0,17512 | 0,25754 |
| | $\mu_{21} = 0$ | -0,03088 | 0,08272 | 0,08825 |
| | $\mu_{22} = 3$ | 0,68170 | 0,15683 | 0,69949 |
| Mezcla de distribuciones t multivariadas | $\pi_1 = 0,3$ | -0,00169 | 0,01199 | 0,01210 |
| | $\mu_{11} = 0$ | 0,00247 | 0,15736 | 0,15730 |
| | $\mu_{12} = -3$ | -0,00391 | 0,15815 | 0,15812 |
| | $\mu_{21} = 0$ | -0,00141 | 0,09262 | 0,09258 |
| | $\mu_{22} = 3$ | -0,00234 | 0,09520 | 0,09518 |
| | $v_1 = 4$ | 2,41717 | 2,93371 | 3,80010 |
| | $v_2 = 10$ | 1,49516 | 3,21547 | 3,54463 |
| Mezcla de distribuciones normales contaminadas multivariadas | $\pi_1 = 0,3$ | 0,00674 | 0,02615 | 0,02699 |
| | $\mu_{11} = 0$ | -0,00889 | 0,14883 | 0,14902 |
| | $\mu_{12} = -3$ | 0,00370 | 0,15113 | 0,15110 |
| | $\mu_{21} = 0$ | -0,00185 | 0,09991 | 0,09988 |
| | $\mu_{22} = 3$ | -0,00237 | 0,10078 | 0,10076 |
| | $\alpha_1 = 0,9$ | -0,02447 | 0,06315 | 0,06769 |
| | $\alpha_2 = 0,8$ | -0,02113 | 0,05320 | 0,05722 |
| | $\eta_1 = 20$ | -3,61257 | 9,96994 | 10,59958 |
| | $\eta_2 = 30$ | 6,98826 | 20,78132 | 21,91500 |

Nota. DE = Desviación estándar. RMSE = Error cuadrático medio.

En el cuadro 5.2 es posible apreciar cómo el sesgo de los parámetros de ponderación del primer componente de la mezcla π_1 y los parámetros de medias aritméticas de ambos componentes μ_1 y μ_2 son mayores cuando se emplea la distribución normal estándar multivariada como la estrategia para desarrollar una clusterización basada en una mezcla de dicho modelo. Esto ocurre en principio por la presencia de un 1% de valores extremos que fueron insertados en el conjunto de datos. En contraste, al utilizar una mezcla de distribuciones t multivariadas o distribuciones normales contaminadas multivariadas se identifica un menor grado de sesgo en los parámetros mencionados. De esta manera, es posible afirmar que utilizar una mezcla de distribuciones normales multivariadas en presencia de valores extremos brinda un menor error en la estimación de parámetros y que el funcionamiento de la mezcla de distribuciones normales contaminadas multivariadas tiene un desempeño similar a la mezcla de distribuciones t multivariadas como base para la clusterización de datos.

Los resultados de la tercera simulación se presentan en el cuadro 5.3 en la cual se exhiben los escenarios considerados en la simulación que emplea el método hierarchical como aproximación para obtener los valores iniciales, junto con estimaciones del sesgo, desviación estándar y error cuadrático medio para los parámetros comunes y específicos en donde es pertinente.

Cuadro 5.3: Resultados de la simulación sobre la clusterización basada en modelos normales multivariados, modelos t multivariados y modelos normales contaminados multivariados, utilizando el método hierarchical para obtener valores iniciales

| Escenario | Parámetro | Sesgo | DE | RMSE |
|---|------------------|----------|----------|----------|
| Mixtura de distribuciones normales multivariadas | $\pi_1 = 0,3$ | -0,10740 | 0,01864 | 0,05737 |
| | $\mu_{11} = 0$ | 0,14300 | 0,20085 | 0,12961 |
| | $\mu_{12} = -3$ | -0,18625 | 0,16993 | 0,13258 |
| | $\mu_{21} = 0$ | -0,03058 | 0,08342 | 0,04668 |
| | $\mu_{22} = 3$ | 0,68011 | 0,14588 | 0,36606 |
| Mixtura de distribuciones t multivariadas | $\pi_1 = 0,3$ | -0,00174 | 0,01202 | 0,01214 |
| | $\mu_{11} = 0$ | 0,00249 | 0,15740 | 0,15734 |
| | $\mu_{12} = -3$ | -0,00398 | 0,15839 | 0,15836 |
| | $\mu_{21} = 0$ | -0,00144 | 0,09262 | 0,09259 |
| | $\mu_{22} = 3$ | -0,00236 | 0,09524 | 0,09522 |
| | $v_1 = 4$ | 2,42497 | 2,94067 | 3,81043 |
| | $v_2 = 10$ | 1,51760 | 3,21652 | 3,55510 |
| Mixtura de distribuciones normales contaminadas multivariadas | $\pi_1 = 0,3$ | 0,00902 | 0,03161 | 0,03285 |
| | $\mu_{11} = 0$ | -0,00955 | 0,14888 | 0,14911 |
| | $\mu_{12} = -3$ | 0,00471 | 0,15160 | 0,15160 |
| | $\mu_{21} = 0$ | -0,00161 | 0,09985 | 0,09981 |
| | $\mu_{22} = 3$ | -0,00237 | 0,10077 | 0,10074 |
| | $\alpha_1 = 0,9$ | -0,03370 | 0,06953 | 0,07723 |
| | $\alpha_2 = 0,8$ | -0,03335 | 0,06479 | 0,07284 |
| | $\eta_1 = 20$ | -4,62148 | 10,30748 | 11,29141 |
| | $\eta_2 = 30$ | 7,38614 | 21,32351 | 22,55643 |

Nota. DE = Desviación estándar. RMSE = Error cuadrático medio.

En el cuadro 5.3 es posible apreciar cómo el sesgo de los parámetros de ponderación del primer componente de la mixtura π_1 y los parámetros de medias aritméticas de ambos componentes μ_1 y μ_2 son mayores cuando se emplea la distribución normal estándar multivariada como la estrategia para desarrollar una clusterización basada en una mixtura de dicho modelo. En síntesis, los resultados son similares al caso en donde se emplea el algoritmo de k-medoids para obtener los valores iniciales del algoritmo. A pesar de ello, es posible observar un menor sesgo y variabilidad al emplear k-medoids en comparación con el empleo del método hierarchical para obtener los valores iniciales. Por estos motivos, se concluye que el método más idóneo para asegurar mejor precisión en la estimación de parámetros es k-medoids, aunque las diferencias no se encuentran muy marcadas en comparación con hierarchical.

Capítulo 6

Aplicación en las Habilidades Socioemocionales

En el presente capítulo se expone una aplicación directa de la mixtura de distribuciones normales contaminadas multivariadas como base para una clusterización. Esto se realiza en una base datos secundaria que corresponde al estudio Evaluación Virtual de Aprendizajes (EVA), un operativo de evaluación desarrollado por la Oficina de Medición de la Calidad de los Aprendizajes (UMC) en el año 2021. En dicho operativo, además de evaluar competencias cognitivas como lectura y matemática, por primera vez se incluyó la evaluación de habilidades socioemocionales (HSE), cuyos resultados son el enfoque principal de esta aplicación práctica.

Este estudio tiene como objetivo principal formar conglomerados de sujetos en relación a los distintos constructos de HSE que fueron evaluados por el Ministerio de Educación del Perú. La Oficina de Medición de la Calidad de los Aprendizajes (UMC) provee de manera abierta los datos de la evaluación, pero no brinda las respuestas a cada uno de los cuestionarios, sino la medida estimada para cada estudiante en todas las HSE a través del modelo de Escala de Valoración de Andrich (Andrich, 1978). En otras palabras, el presente trabajo no aborda el modelamiento del rasgo latente en cada HSE, pues es un procedimiento previamente desarrollado por la UMC.

6.1. Habilidades Socioemocionales

Durante los últimos años ha surgido una creciente preocupación sobre si las habilidades cognitivas son adecuadas para predecir el éxito en diferentes ámbitos de la vida, incluyendo la escuela y el trabajo. En este contexto, Vila et al. (2021) consideran que existe otro conjunto de habilidades como la empatía, habilidades sociales, organización, autorregulación, entre otras que tienen tanta importancia como las capacidades cognitivas y conocimientos para el logro de los propios objetivos y el éxito futuro. Dichos constructos se denominan habilidades socioemocionales y recientemente han sido objeto de constante atención y estudio sobre cómo comprender y medir este conjunto de habilidades.

En la literatura, dichas habilidades son reconocidas bajo distintas nomenclaturas; por ejemplo, habilidades no cognitivas, habilidades del siglo 21, carácter, competencias socioemocionales, cualidades personales o habilidades blandas (Duckworth y Yeager, 2021; García-Cabrero, 2018). En muchas ocasiones, el término empleado para referirse a dichas habilidades depende de la disciplina científica y la elección de los investigadores, lo que deriva en definiciones distintas que pueden tener limitaciones en su amplitud o adecuación. Por ejemplo, el término “no cognitivo” sugiere que hay aspectos de la conducta humana que no implican el procesamiento cognitivo, pero como señalan Duckworth y Yeager (2021), todos los aspectos psicológicos de una persona involucran procesamiento de información y, por lo tanto, tienen un componente cognitivo. El término “habilidades blandas” también podría ser discutible ya que sugiere que dichas habilidades son menos importantes que las cognitivas (García-Cabrero, 2018).

Dado que no hay una terminología ideal para describir estas habilidades, Duckworth y Yeager (2021) han identificado una serie de características comunes que todas estas palabras comparten:

- a. son distintas de las habilidades cognitivas
- b. se consideran útiles para el estudiante y la sociedad en general
- c. tienden a ser estables a lo largo del tiempo sin interferencias externas
- d. son susceptibles a ser mejoradas mediante una intervención
- e. dependen de factores situacionales para ser manifestadas

Como resultado, García-Cabrero (2018) sugiere emplear el término habilidades no cognitivas, al cual define como patrones de pensamiento, emociones y comportamientos que pueden seguir evolucionando a lo largo de la vida y que desempeñan una función en el proceso educativo. De acuerdo con dicho autor, estas habilidades no son diferentes de los rasgos representados por las habilidades cognitivas, sino que se refieren a características socioemocionales y de comportamiento que no constituyen rasgos de personalidad fijos. Asimismo, García-Cabrero (2018) señala que las habilidades socioemocionales estarían relacionadas con el proceso educativo, ya sea porque se han adquirido durante la escolarización o porque contribuyen al desarrollo de las habilidades cognitivas durante esta etapa.

6.2. La Evaluación de las Habilidades Socioemocionales en Perú

La Oficina de Medición de la Calidad de los Aprendizajes (UMC), instancia técnica del Ministerio de Educación del Perú (MINEDU) desarrolló un operativo de evaluación denominado EVA 2021 cuyo objetivo principal fue brindar un diagnóstico actualizado sobre la situación de los aprendizajes de los estudiantes, sus habilidades socioemocionales y los factores contextuales relacionados con sus aprendizajes en el contexto de la pandemia por COVID-19. En este sentido, el operativo incluyó la evaluación de constructos cognitivos como Lectura y Matemática, y; por primera vez, la evaluación de las habilidades socioemocionales como constructos no cognitivos. Dicho operativo fue desarrollado en varios grados del ciclo estudiantil; no obstante, las HSE solo fueron evaluadas en segundo grado de secundaria.

6.2.1. Características de la muestra

La población objetivo fueron estudiantes que vivieran en hogares que cuenten con dispositivos electrónicos (Computadora, tablet o smartphone) e internet. Esta población corresponde al 48.0% del total de estudiantes matriculados durante el año 2021. A pesar de que el operativo alcanzó una muestra de 11227 estudiantes, solo una submuestra de 3363 estudiantes rindieron la evaluación de habilidades socioemocionales. Es importante mencionar que los resultados de la EVA 2021 no son representativos de toda la población escolar, pues tras ser un estudio de carácter virtual, solo aquellos estudiantes que contaban con un dispositivo electrónico y de conexión a internet pudieron tener acceso a la evaluación.

6.2.2. Características del instrumento de recolección de datos

Los instrumentos de recolección de datos empleados en el operativo EVA 2021 fueron entrevistas y cuestionarios, principalmente. Para el caso de las habilidades socioemocionales, se consideraron pruebas psicométricas sobre los siguientes constructos:

- Autoeficacia emocional: creencias sobre las capacidades propias para afrontar diversas situaciones. Estas se dividen en:
 - Autoeficacia emocional: creencias sobre las capacidades propias para experimentar y expresar emociones
 - Autoeficacia social: creencias sobre las capacidades propias para comportarse de manera efectiva en situaciones sociales
 - Autoeficacia académica: creencia sobre las capacidades propias para responder a las demandas del entorno académico
- Autorregulación conductual: capacidad para planificar, guiar y monitorear el propio comportamiento de manera flexible frente a circunstancias cambiantes
- Autonomía: grado de funcionamiento autónomo
- Relación: grado de interacción o distanciamiento con otras personas
- Resiliencia: capacidad para utilizar los recursos disponibles para hacer frente a situaciones adversas
- Toma de decisiones: Implica características vinculadas a la manera en la cual se aborda la toma de decisiones efectiva. Estas son:
 - Vigilancia: implica considerar diversas alternativas para alcanzar los objetivos
 - Hipervigilancia: búsqueda frenética de alternativas para salir del dilema y aliviar la tensión rápidamente
 - Procrastinación: aplazar la toma de decisiones al darle menor prioridad
 - Transferencia: atribución de la responsabilidad de la decisión propia
- Empatía cognitiva: capacidad para identificar el estado emocional propio y de otros individuos
- Disposición empática: disposición para apoyar a personas que atraviesan una dificultad
- Responsabilidad: rasgos orientados al orden, precisión, cumplimiento de compromisos y adhesión a normas

6.2.3. Características de los datos

Todos los constructos presentados en la sección anterior fueron modelados considerando el modelo de Escala de Valoración de Andrich (Andrich, 1978). La calibración de los instrumentos se desarrolló empleando el método de máxima verosimilitud marginal (MML), que es una aproximación bayesiana en donde una distribución normal es asumida como distribución a priori (Birnbaum, 1968); mientras que, las medidas de los estudiantes fueron estimadas a partir de la metodología de Valores Plausibles, la cual consiste en extraer valores aleatorios de la distribución posterior para cada estudiante. En particular se extrajeron 5 valores plausibles para cada constructo latente y por estudiante. Para esta aplicación solo se consideró el primer valor plausible. Una descripción del modelo Rasch puede ser consultada con más detalle en el Anexo 1.

Asimismo, la base de datos contiene variables sociodemográficas del estudiante, como el sexo, área y gestión de la institución educativa en la que se encuentra matriculado. Es importante mencionar que la base de datos contiene una variable adicional que se denomina *Deseabilidad social*, este es el resultado de una escala psicométrica que tiene como objetivo identificar un potencial sesgo en las respuestas de los estudiantes. Como esta no se considera una habilidad socioemocional en sí no es considerada en el estudio. En el cuadro 6.1 se presenta el libro de códigos de la base de datos.

Cuadro 6.1: Libro de códigos de la base de datos empleada

| Variable | Descripción | Tipo | Valores |
|-------------|---|----------|---------------------------|
| sexo | Sexo del estudiante | Numérica | 0=Hombre 1=Mujer |
| area | Área en la que reside el estudiante | Numérica | 1=Urbana 2=Rural |
| gestion2 | Gestión de la Institución Educativa | Numérica | 1=Estatal 2=No estatal |
| PV1.AFC_ACA | Autoeficacia académica Primer Valor plausible | Numérica | - |
| PV1.AFC_SOC | Autoeficacia social Primer Valor plausible | Numérica | - |
| PV1.AFC_EMO | Autoeficacia emocional Primer Valor plausible | Numérica | - |
| PV1.AYR_AUT | Autonomía Primer Valor plausible | Numérica | - |
| PV1.AYR_REL | Relación Primer Valor plausible | Numérica | - |
| PV1.ACC_1 | Autorregulación conductual Primer Valor plausible | Numérica | - |
| PV1.DSC_1 | Deseabilidad social Primer Valor plausible | Numérica | - |
| PV1.EMC_COG | Empatía cognitiva Primer Valor plausible | Numérica | - |
| PV1.EMC_DIS | Disposición empática Primer Valor plausible | Numérica | - |
| PV1.RSC_1 | Resiliencia Primer Valor plausible | Numérica | - |
| PV1.REC_1 | Responsabilidad Primer Valor plausible | Numérica | - |
| PV1.TDD_VIG | Toma de decisiones Vigilancia Primer Valor plausible | Numérica | - |
| PV1.TDD_PRO | Toma de decisiones Procrastinación Primer Valor plausible | Numérica | - |
| PV1.TDD_HIP | Toma de decisiones Transferencia Primer Valor plausible | Numérica | - |
| PV1.TDD_TRA | Toma de decisiones Hipervigilancia Primer Valor plausible | Numérica | - |

6.2.4. Análisis exploratorio de datos

Con respecto a las variables sociodemográficas, se identifica que la muestra está compuesta por un total de 54.42 % (n=1829) de estudiantes varones y 45.52 % (n=1532) de estudiantes mujeres. Asimismo, el 82.86 % (n=2785) de los estudiantes se encontraba habitando en una zona urbana; mientras que, el restante 17.14 % (n=576) habitaba en la zona rural. Finalmente, el 68.64 % (n=2307) de estudiantes se encontraba matriculado en instituciones educativas de gestión estatal y 31.56 % (n=1054) en instituciones educativas de gestión no estatal. La distribución de las variables correspondientes a las habilidades socioemocionales se presenta en las figuras 6.1, 6.2 y 6.3.

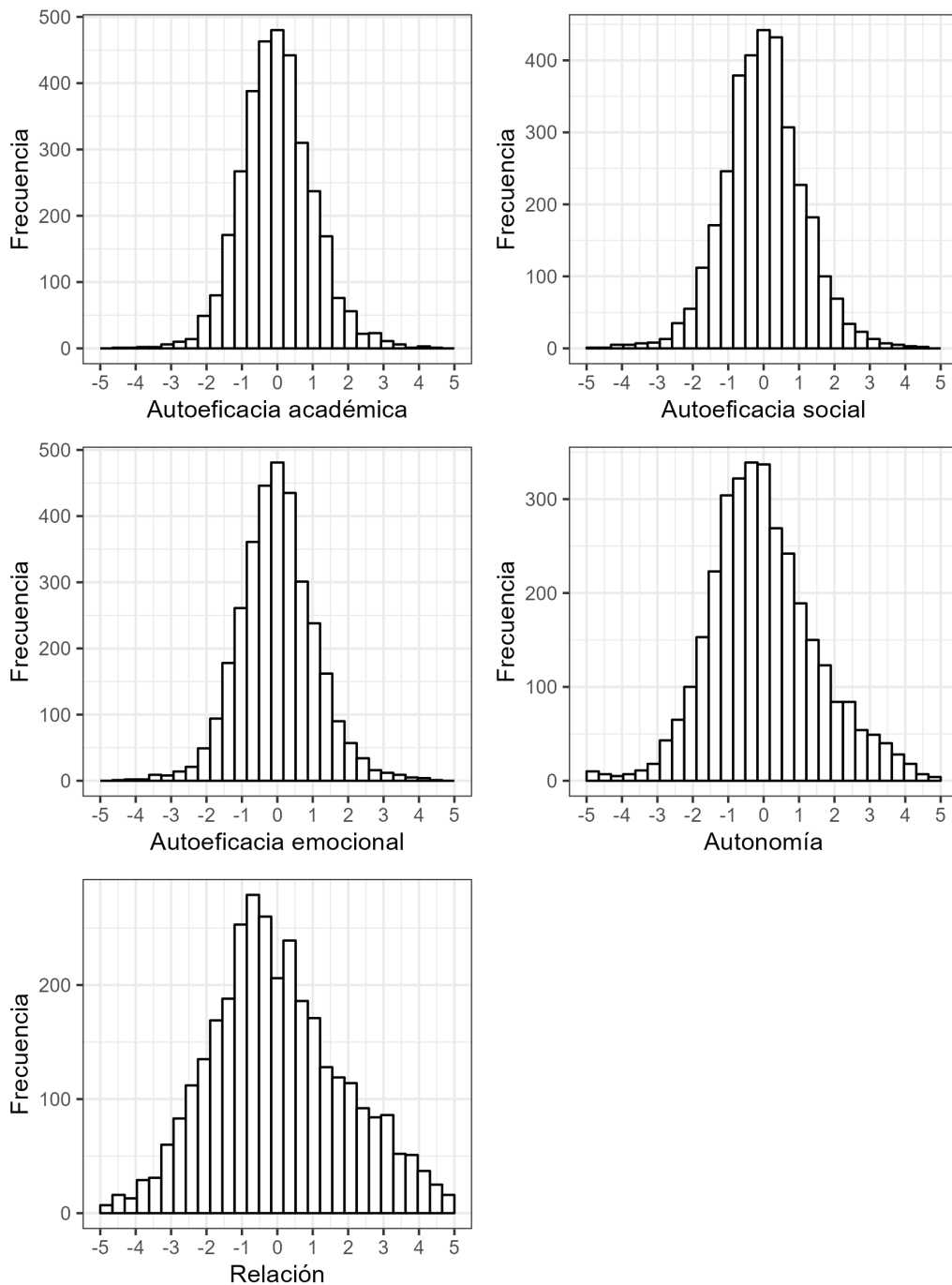


Figura 6.1: Histograma de las variables Autoeficacia, Autonomía y Relación

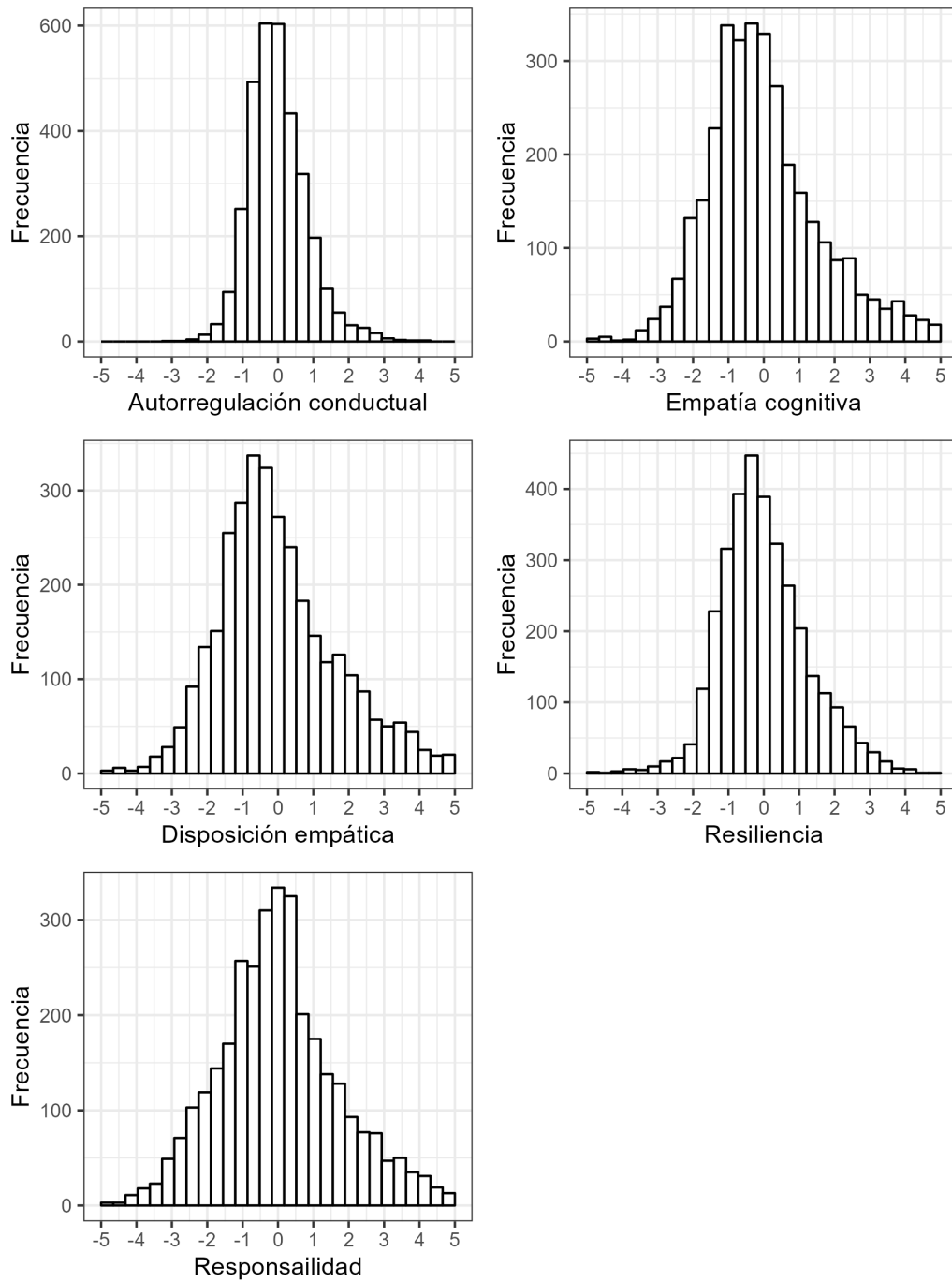


Figura 6.2: Histograma de las variables Autorregulación conductual, Empatía cognitiva, Disposición empática, Resiliencia y Responsabilidad

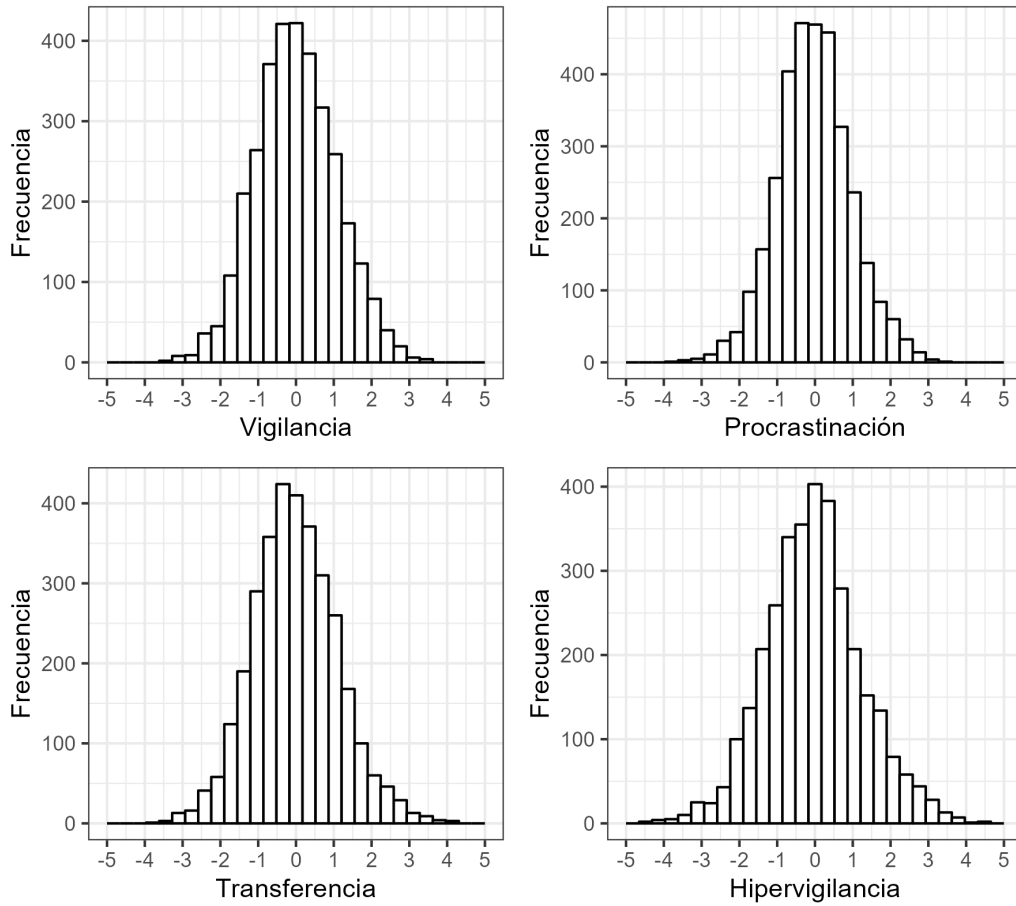


Figura 6.3: Histograma de la variable Toma de decisiones y sus subdimensiones

Como se puede apreciar, la distribución de cada una de las variables se observa como próxima a una distribución normal, aunque en algunos casos con unos potenciales valores extremos. En este sentido, emplear una distribución normal contaminada multivariada como base para el proceso de clusterización tiene un sustento empírico en los datos.

6.3. Clusterización basada en una mezcla de distribuciones normales contaminadas multivariadas

Con el objetivo de desarrollar una clusterización basada en una mezcla de distribuciones normales contaminadas multivariadas, se realizó un ajuste de diferentes modelos en donde se consideran desde $G = 2$ hasta $G = 6$ conglomerados potenciales, pues a partir de $G = 7$ conglomerados se observó problemas de convergencia. Los resultados de estos análisis se presentan en el cuadro 6.2.

Cuadro 6.2: Estadísticos de ajuste para la cantidad de conglomerados

| G | log-lik | AIC | BIC | KIC | KICc | AIC3 | CAIC | AICc | ICL |
|---|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 2 | -59807,46 | 120044,92 | 121357,50 | 120262,92 | 120300,59 | 120259,92 | 121572,50 | 120074,93 | 121715,95 |
| 3 | -59624,18 | 119894,37 | 121866,29 | 120220,37 | 120307,87 | 120217,37 | 122189,29 | 119964,44 | 122414,09 |
| 4 | -59457,88 | 119777,76 | 122409,01 | 120211,76 | 120372,65 | 120208,76 | 122840,01 | 119907,10 | 123000,50 |
| 5 | -59326,61 | 119731,23 | 123021,83 | 120273,23 | 120533,66 | 120270,23 | 123560,83 | 119941,30 | 123675,35 |
| 6 | -59265,38 | 119824,76 | 123774,70 | 120474,76 | 120863,85 | 120471,76 | 124421,70 | 120139,63 | 124476,19 |

Como se puede apreciar, los resultados indican un mejor ajuste del modelo que considera únicamente $G = 2$ mezclas de distribuciones normales contaminadas multivariadas, seguido por el modelo que considera $G = 4$ conglomerados. A continuación se presentan los resultados de la estimación de los parámetros en ambos modelos, junto con la estimación de la proporción de casos que son considerados outliers.

6.3.1. Modelo $G=2$

En el modelo que considera $G = 2$ conglomerados se asume que existe una mezcla de 2 distribuciones normales contaminadas multivariadas, ambas con una ponderación de $\pi_1 = 0,61$ y $\pi_2 = 0,39$. Los vectores de parámetros de localización μ_1 y μ_2 corresponden a las medias aritméticas de las medidas de habilidades socioemocionales obtenidas en cada una de las 14 variables para el primer y segundo componente de la mezcla, y son representadas en la Figura 6.4 con los colores celeste y naranja, respectivamente.

Específicamente, el primer componente es a su vez una mezcla de dos distribuciones normales multivariadas, con el mismo vector de parámetros de localización μ_1 , pero con distintas matrices de varianzas y covarianzas que se encuentran relacionadas por un factor de escala estimado de $\eta_1 = 1,76$, LA primera distribución normal multivariada corresponde a los casos no atípicos y; la segunda, a los casos atípicos, ambos ponderados por el parámetro $\alpha_1 = 0,71$. Es así que de los 2134 estudiantes que pertenecen al primer componente, 400 (18,74%) son identificados como outliers según el modelo.

Con respecto al segundo componente, este también es una mezcla de dos distribuciones normales multivariadas, con el mismo vector de parámetros de localización μ_2 , pero con distintas matrices de varianzas y covarianzas que se encuentran relacionadas por un factor de escala estimado de $\eta_1 = 1,88$, lo que indica una mayor variabilidad del segundo componente de la mezcla. De este modo, se identifica un parámetro de proporción de observaciones no atípicos de $\alpha_2 = 0,73$, que sugiere que, de los 1177 estudiantes que pertenecen a este componente, 253 (21,50%) son considerados como outliers.

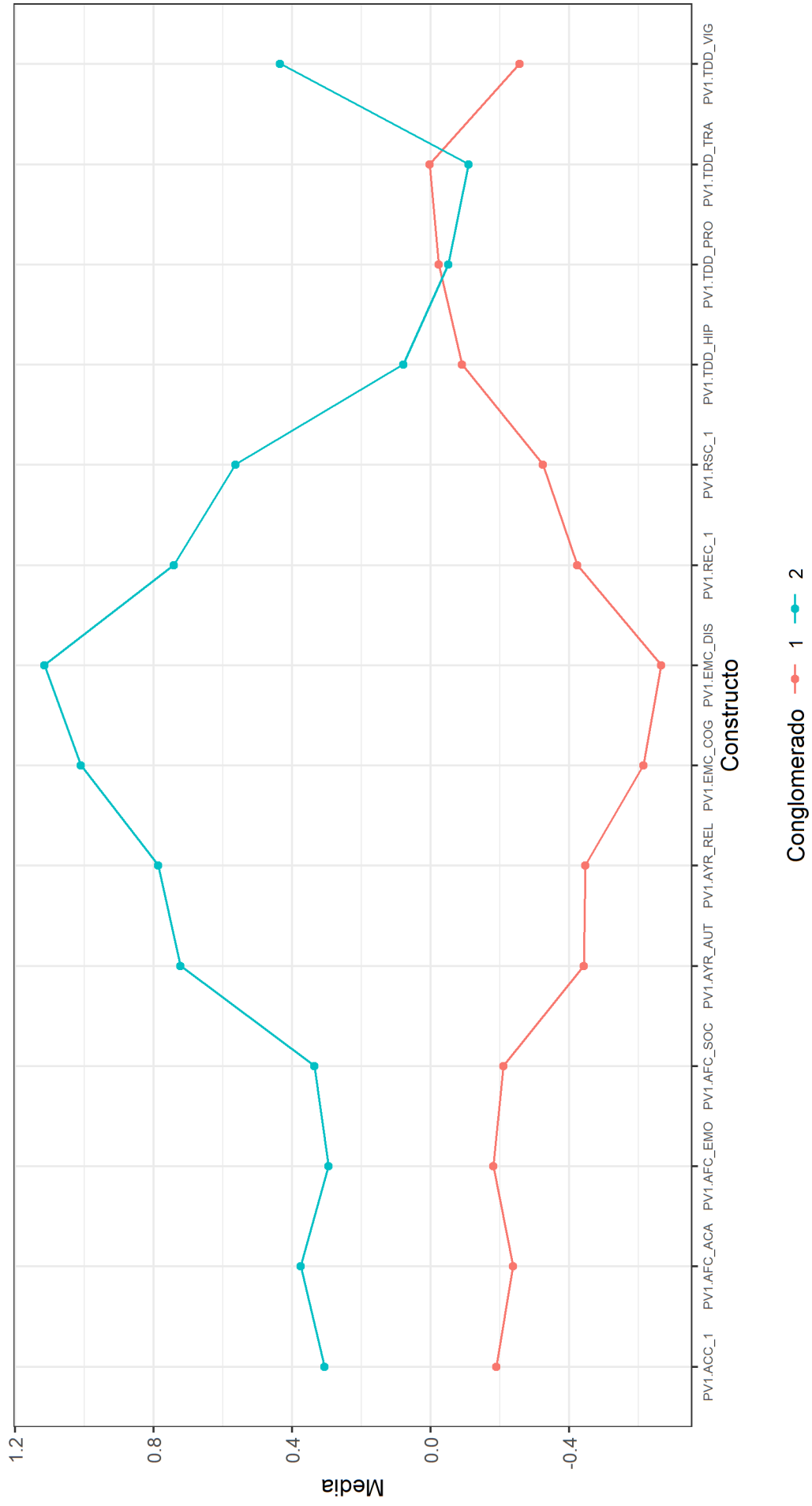


Figura 6.4: Medias aritméticas de los conglomerados identificados en el primer modelo

El resultado de la comparación entre medias aritméticas de las variables para cada conglomerado permite identificar claramente un patrón de estudiantes con un alto nivel alcanzado en todas las medidas de habilidades socioemocionales consideradas en el estudio, con excepción de las primeras tres dimensiones de la Toma de decisiones: hipervigilancia, procrastinación y transferencia. Sin considerar dichas variables, el modelo permitiría clasificar a estudiantes en estos dos grupos diferenciados.

6.3.2. Modelo $G=4$

En el modelo que considera $G = 4$ conglomerados se asume que existe una mixtura de 4 distribuciones normales contaminadas multivariadas, las que mantienen una ponderación de $\pi_1 = 0,42$, $\pi_2 = 0,26$, $\pi_3 = 0,22$ y $\pi_4 = 0,09$. Al igual que en el escenario anterior, cada uno de los cuatro componentes es a su vez una mixtura de dos distribuciones normales multivariadas con los mismos vectores de parámetros de localización μ_1 , μ_2 , μ_3 y μ_4 , representados den la Figura 6.5 con los colores naranja, verde, celeste y morado, respectivamente,

El primer componente es a su vez una mixtura de dos distribuciones normales multivariadas en donde la primera representa a los casos no atípicos y la segundo a los casos atípicos. Tras estimar un parámetro de escala entre matrices de varianzas y covarianzas de $\eta_1 = 1,48$, se concluye una mayor dispersión de la segunda distribución normal multivariada. Asimismo, la ponderación de ambas distribuciones de la mixtura está dada por un parámetro de casos no outliers de $\alpha_1 = 0,69$. Así, el componente se caracteriza por 1536 estudiantes, de los cuales 229 (13,43 %) son identificados como outliers según el modelo.

Con respecto al segundo componente, este corresponde a un conglomerado de 839 estudiantes, de los cuales 139 (16,57 %) son considerados como outliers; así, se establece un parámetro de proporción de observaciones no outliers de $\alpha_2 = 0,76$; mientras que, los parámetros de escala entre matrices de varianzas y covarianzas es de $\eta_2 = 1,87$, que exhibe una mayor dispersión en la distribución normal multivariada que representa a los valores atípicos con respecto a la que representa a los valores no atípicos.

El tercer componente se encuentra compuesto por un total de 674 estudiantes, de los cuales 177 (26,26 %) son considerados como outliers; así, se establece un parámetro de proporción de observaciones no outliers de $\alpha_3 = 0,68$; mientras que, el parámetro de escala entre matrices de varianzas y covarianzas es de $\eta_3 = 2,20$, que indica una mayor dispersión en la distribución normal multivariada que representa a los valores atípicos con respecto a la que representa a los valores no atípicos.

Con respecto al cuarto componente, este corresponde a un conglomerado de 262 estudiantes, de los cuales 87 (33,21 %) son considerados como outliers; así, se establece un parámetro de proporción de observaciones no outliers de $\alpha_4 = 0,59$; mientras que, los parámetros de escala entre matrices de varianzas y covarianzas es de $\eta_4 = 1,52$, que indica una mayor dispersión en la distribución normal multivariada que representa a los valores atípicos con respecto a la que representa a los valores no atípicos.

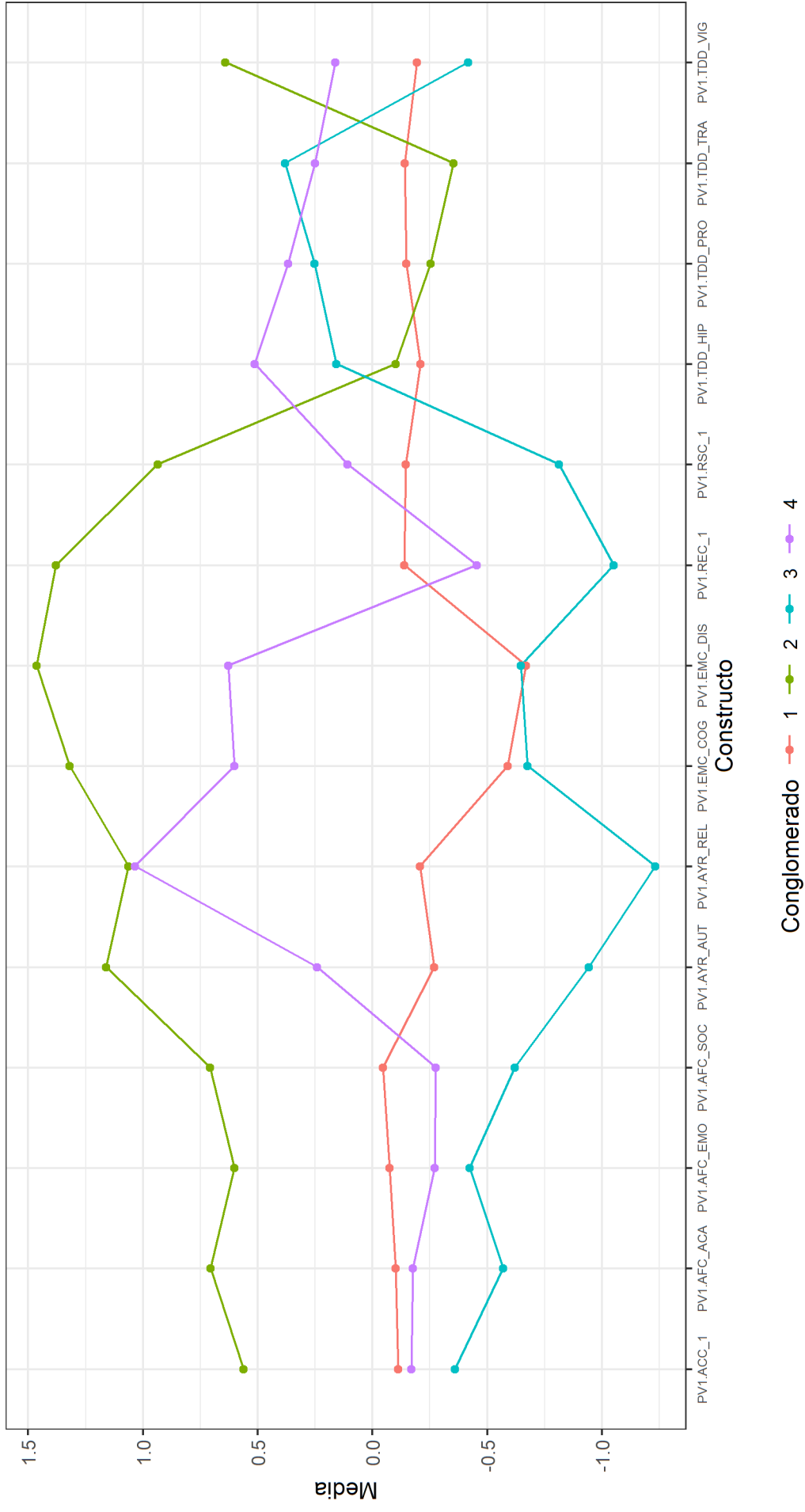
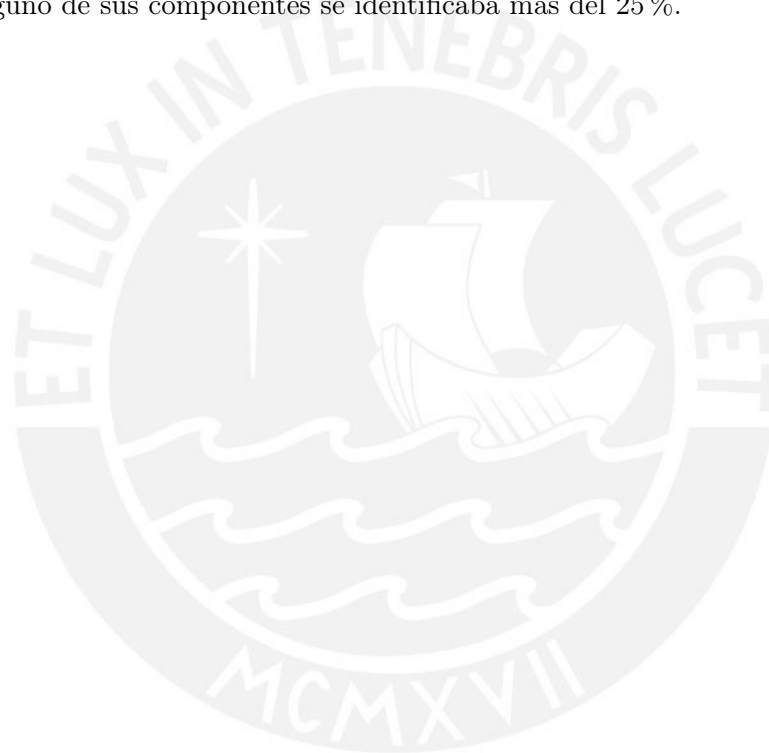


Figura 6.5: Medias aritméticas de los conglomerados identificados en el segundo modelo

A diferencia del primer modelo, al considerar $G = 4$ conglomerados la interpretación de los resultados se vuelve más compleja. Puede observarse que el conglomerado 2 se caracteriza por tener niveles más altos en todas las variables socioemocionales, con excepción de las tres subescalas principales de la Toma de decisiones, lo que demuestra cierta congruencia con lo observado en el modelo 1. El tercer conglomerado alcanza los niveles más bajos en todas las habilidades socioemocionales con excepción de las mismas tres subdimensiones de Toma de decisiones. Estos patrones podrían indicar que retirar dichas variables permitirían una identificación más parsimoniosa de las características de los conglomerados. No se presentan muchas diferencias entre los conglomerados 1 y 4. Adicionalmente, la proporción de casos atípicos identificados en los conglomerados 3 y 4 sobrepasa el 25 % del total de casos clasificados en los respectivos conglomerados, lo que indica que más de un cuarto de las observaciones están siendo detectadas como outliers, a diferencia del modelo de $G = 2$ conglomerados en donde en ninguno de sus componentes se identificaba más del 25 %.



Capítulo 7

Conclusiones y sugerencias

7.1. Conclusiones

- Se llevó a cabo un análisis del modelo de la distribución normal contaminada, tanto en su versión univariada como multivariada, así como de una mixtura de distribuciones normales contaminadas multivariadas..
- Se llevaron a cabo estudios de simulación que proporcionaron un marco comparativo entre la metodología utilizada y otras aproximaciones de agrupamiento basadas en modelos, en particular, aquellas que emplean la distribución normal multivariada y la distribución t multivariada. Los resultados de este análisis comparativo revelan que el modelo de distribución normal contaminada multivariada opera con una eficiencia comparable a la que se obtiene al emplear la distribución t multivariada. Sin embargo, se observó que el modelo de distribución normal contaminada multivariada supera a la distribución normal multivariada, especialmente en presencia de valores atípicos (outliers) y datos faltantes.
- Se implementó un proceso de clusterización y detección de outliers a los resultados de la evaluación de habilidades socioemocionales implementada por el Ministerio de Educación del Perú en el año 2021 y se optó por un modelo con $G = 2$. conglomerados como una buena aproximación para la interpretación de los resultados.

7.2. Sugerencias

- Futuras investigaciones pueden contemplar más escenarios para las simulaciones desarrolladas en el presente estudio, considerando evaluar el desempeño de las mixturas propuestas al modificar la proporción de valores extremos, cantidad de componentes de la mixtura y dimensionalidad.
- Futuras investigaciones pueden plantear estudios predictivos en donde se evalúe qué variables se encuentran asociadas al hecho de que un estudiante pertenezca al conglomerado de habilidades socioemocionales más desarrolladas y a aquel de habilidades socioemocionales menos desarrolladas.

Capítulo 8

Código en R

Este capítulo presenta los códigos implementados durante la simulación que se llevó a cabo para estudiar las propiedades del modelo, así como su aplicación a los datos empíricos.

8.1. Código en R para el estudio de simulación

```
1 #-----#
2 # Simulacion 1 en Punzo y Tortora
3 #-----#
4 setwd(dirname(rstudioapi::getSourceEditorContext()$path))
5 library(tidyverse)
6 library(mvtnorm) #Student t y MN
7 library(ContaminatedMixt) #MCN
8 library(mice) #Missing values
9 library(clusterGeneration) #genRandomClust una mezcla de MN
10 library(MixGHD) #ARI
11 library(haven)
12 set.seed(420)
13
14 #-----#
15 # Parametros generales
16 #-----#
17 # Numero de variables
18 nvariables <- 2
19
20 # Numero de clusters
21 clusters <- 2
22
23 # Tamanos de muestra
24 n <- c(200)
25
26 # Replicas
27 reps <- 1000
28
29 #Vectores de medias
30 mu <- list(c(0,-3),
31           c(0,3))
32
33 #Matrices de covarianzas
34 sigma <- list(
35   matrix(c(1,-0.5,
36           -0.5,1),byrow=T,ncol=2),
37   matrix(c(1,0.5,
38           0.5,1),byrow=T,ncol=2))
39
40 #Ponderacion de cada mezcla
41 pi <- c(0.3,0.7)
42
43 #-----#
44 # Simulacion
45 #-----#
46 tmulti <- list()
47 contam <- list()
48 mnorma <- list()
49
50 for (m in 1:reps) {
51
52 #-----#
53 # Mezcla de distribucion t multivariada
```

```

54 #-----#
55 # Grados de libertad
56 df <- c(4,10)
57
58 # Simulacion
59 dist_mt <- list()
60 for (i in 1:clusters) { dist_mt[[i]] <- rmvt(n = n[1]*pi[i],mu[[i]],sigma=sigma[[i]],df=df[i])
61   dist_mt[[i]] <- data.frame(dist_mt[[i]])
62   names(dist_mt[[i]]) <- paste0("R",m,"V",1:nvariables)
63 }
64
65 # Data final
66 dist_mt <- bind_rows(dist_mt)
67 dist_mt <- ampute(dist_mt,prop=.1)$amp
68 tmulti[[m]] <- dist_mt
69 #-----#
70 # Mixtura de distribucion normales contaminadas multivariadas
71 #-----#
72 # Grado de descontaminacion (porcentaje de casos no outliers)
73 alpha <- c(0.9,0.8)
74
75 # Escala entre matrices de covarianzas
76 eta <- c(20, 30)
77
78 #Simulacion
79 dist_cn <- list()
80 for (i in 1:clusters) { dist_cn[[i]] <- rCN(n=n[1]*pi[i],mu=mu[[i]],Sigma=sigma[[i]],alpha=alpha[i],eta = eta[i])
81   dist_cn[[i]] <- data.frame(dist_cn[[i]])
82   names(dist_cn[[i]]) <- paste0("R",m,"V",1:nvariables)
83 }
84
85 # Data final
86 dist_cn <- bind_rows(dist_cn)
87 dist_cn <- ampute(dist_cn,prop=.1)$amp
88 contam[[m]] <- dist_cn
89 #-----#
90 # Mixtura de distribucion normales multivariadas
91 #-----#
92 dist_nm <- list()
93
94 #Simulacion
95 dist_nm <- list()
96 for (i in 1:clusters) { dist_nm[[i]] <- rmvnorm(n=n[1]*pi[i],mean=mu[[i]],sigma=sigma[[i]])
97   dist_nm[[i]] <- data.frame(dist_nm[[i]])
98   names(dist_nm[[i]]) <- paste0("R",m,"V",1:nvariables)
99 }
100
101 # Data final
102 dist_nm <- bind_rows(dist_nm)
103
104 #Anadir 1% de valores extremos:
105 #Identificar casos que seran sustituidos
106 cases <- sample(1:nrow(dist_nm),nrow(dist_nm)*.1,replace = F)
107
108 #Replacements
109 repla <- cbind(rep(0,length(cases)),runif(length(cases),10,15))
110
111 #Reemplazar
112 dist_nm[cases,] <- repla
113 dist_nm <- ampute(dist_nm,prop=.1)$amp
114 mnorma[[m]] <- dist_nm
115 } #End of simulation
116
117 #-----#
118 # Resultados de la simulacion
119 #-----#
120 tmulti <- bind_cols(tmulti)
121 contam <- bind_cols(contam)
122 mnorma <- bind_cols(mnorma)
123
124 #-----#
125 # Exportacion
126 #-----#
127 write_sav(tmulti,"simula_t_multivariada.sav")
128 write_sav(contam,"simula_n_contaminada_multivariada.sav")
129 write_sav(mnorma,"simula_n_multivariada.sav")

```

8.2. Código en R para el análisis del estudio de simulación

```
1 #-----#
2 # Analisis de simulacion
3 #-----#
4 setwd(dirname(rstudioapi::getSourceEditorContext()$path))
5 library(tidyverse)
6 library(mvtnorm) #Student t y MN
7 library(ContaminatedMixt) #MCN
8 library(mice) #Missing values
9 library(clusterGeneration) #genRandomClust una mezcla de MN
10 library(MixGHD) #ARI
11 library(haven)
12 library(cluster)
13 library(MixtureMissing)
14 library(openxlsx)
15 set.seed(420)
16
17 #-----#
18 # Analisis de simulacion
19 #-----#
20 data_tm <- read_sav("simula_t_multivariada.sav")
21 data_nc <- read_sav("simula_n_contaminada_multivariada.sav")
22 data_nm <- read_sav("simula_n_multivariada.sav")
23
24 #-----#
25 # Parametros a recuperar
26 #-----#
27 #Vectores de medias
28 mu <- list(c(0,-3),
29            c(0,3))
30
31 #Matrices de covarianzas
32 sigma <- list(
33   matrix(c(1,-0.5,
34            -0.5,1),byrow=T,ncol=2),
35   matrix(c(1,0.5,
36            0.5,1),byrow=T,ncol=2))
37
38 #Ponderacion de cada mezcla
39 pi <- c(0.3,0.7)
40
41 #replications
42 reps <- 1000
43
44 #df para t multivariada
45 df <- c(4,10)
46
47 # Grado de descontaminacion (porcentaje de casos no outliers) en NC
48 alpha <- c(0.9,0.8)
49
50 # Escala entre matrices de covarianzas en NC
51 eta <- c(20, 30)
52
53 #x <- MCMN(data_nc[,1:2], G=2)
54 #y <- MNM(data_nm[,1:2], G=2)
55 #z <- MtM(data_tm[,1:2], G=2)
56
57 #-----#
58 # distribucion t multivariada
59 #-----#
60 par_pi <- list()
61 par_mu_v1_1 <- list()
62 par_mu_v2_1 <- list()
63 par_mu_v1_2 <- list()
64 par_mu_v2_2 <- list()
65 #par_sigma_1 <- list()
66 #par_sigma_2 <- list()
67 par_df_1 <- list()
68 par_df_2 <- list()
69
70 id <- seq(1,ncol(data_tm), by=2)
71 for (i in 1:(ncol(data_tm)/2)) {
72   #i=2
73   estim <- MtM(data_tm[,id[i]:(id[i]+1)], G=2, show_progress = F)
74   par_pi[[i]] <- min(estim$pi) #Pi
75   main_mu <- data.frame(estim$mu)
76   names(main_mu) <- c("V1","V2")
77   g1 <- main_mu %>% filter(V2<0)
78   g2 <- main_mu %>% filter(V2>0)
79   par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
```

```

80 par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
81 par_mu_v2_1[[i]] <- g2[1] #Mu de V2
82 par_mu_v2_2[[i]] <- g2[2] #Mu de V2
83 #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
84 #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
85 par_df_1[[i]] <- min(estim$df)
86 par_df_2[[i]] <- max(estim$df)
87 }
88
89 #Results t multivariate
90 result_tm <-
91 data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
92                   std = sd(unlist(par_pi)),
93                   rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/reps)),
94           mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
95                     std = sd(unlist(par_mu_v1_1)),
96                     rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/reps)),
97           mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
98                     std = sd(unlist(par_mu_v1_2)),
99                     rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/reps)),
100          mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
101                    std = sd(unlist(par_mu_v2_1)),
102                    rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/reps)),
103          mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],
104                    std = sd(unlist(par_mu_v2_2)),
105                    rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/reps)),
106          df_1 = c(bias = mean(unlist(par_df_1))-df[1],
107                 std = sd(unlist(par_df_1)),
108                 rmse = sqrt(sum((unlist(par_df_1)-df[1])^2)/reps)),
109          df_2 = c(bias = mean(unlist(par_df_2))-df[2],
110                 std = sd(unlist(par_df_2)),
111                 rmse = sqrt(sum((unlist(par_df_2)-df[2])^2)/reps))
112 write.xlsx(result_tm,"resultados_tm_kmedoids.xlsx",overwrite = T)
113
114 #-----#
115 # distribucion normal multivariada
116 #-----#
117 par_pi <- list()
118 par_mu_v1_1 <- list()
119 par_mu_v2_1 <- list()
120 par_mu_v1_2 <- list()
121 par_mu_v2_2 <- list()
122 #par_sigma_1 <- list()
123 #par_sigma_2 <- list()
124
125 id <- seq(1,ncol(data_nm), by=2)
126 for (i in 1:(ncol(data_nm)/2)) {
127   #i=2
128   estim <- MNM(data_nm[,id[i]:(id[i]+1)], G=2, show_progress = F)
129   par_pi[[i]] <- min(estim$pi) #Pi
130   main_mu <- data.frame(estim$mu)
131   names(main_mu) <- c("V1","V2")
132   g1 <- main_mu %>% filter(V2<0)
133   g2 <- main_mu %>% filter(V2>0)
134   par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
135   par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
136   par_mu_v2_1[[i]] <- g2[1] #Mu de V2
137   par_mu_v2_2[[i]] <- g2[2] #Mu de V2
138   #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
139   #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
140 }
141
142 #Results normal multivariante
143 result_nm <-
144 data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
145                   std = sd(unlist(par_pi)),
146                   rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/reps)),
147           mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
148                     std = sd(unlist(par_mu_v1_1)),
149                     rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/reps)),
150           mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
151                     std = sd(unlist(par_mu_v1_2)),
152                     rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/reps)),
153           mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
154                     std = sd(unlist(par_mu_v2_1)),
155                     rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/reps)),
156           mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],
157                     std = sd(unlist(par_mu_v2_2)),
158                     rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/reps))
159 write.xlsx(result_nm,"resultados_nm_kmedoids.xlsx",overwrite = T)
160
161 #-----#

```

```

162 # distribucion normal contaminada multivariada (kmedoids)
163 #-----#
164 par_pi <- list()
165 par_mu_v1_1 <- list()
166 par_mu_v2_1 <- list()
167 par_mu_v1_2 <- list()
168 par_mu_v2_2 <- list()
169 par_alpha_1 <- list()
170 par_alpha_2 <- list()
171 par_eta_1 <- list()
172 par_eta_2 <- list()
173
174 id <- seq(1,ncol(data_nc), by=2)
175 for (i in 1:(ncol(data_nc)/2)) {
176   #i=2
177   estim <- MCMC(data_nc[,id[i]:(id[i]+1)], G=2, show_progress = F)
178   par_pi[[i]] <- min(estim$pi) #Pi
179   main_mu <- data.frame(estim$mu)
180   names(main_mu) <- c("V1","V2")
181   g1 <- main_mu %>% filter(V2<0)
182   g2 <- main_mu %>% filter(V2>0)
183   par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
184   par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
185   par_mu_v2_1[[i]] <- g2[1] #Mu de V2
186   par_mu_v2_2[[i]] <- g2[2] #Mu de V2
187   par_alpha_1[[i]] <- max(estim$alpha)
188   par_alpha_2[[i]] <- min(estim$alpha)
189   par_eta_1[[i]] <- min(estim$eta)
190   par_eta_2[[i]] <- max(estim$eta)
191   #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
192   #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
193 }
194
195 #Results contaminated normal
196 result_nc <-
197   data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
198                     std = sd(unlist(par_pi)),
199                     rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/reps)),
200             mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
201                       std = sd(unlist(par_mu_v1_1)),
202                       rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/reps)),
203             mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
204                       std = sd(unlist(par_mu_v1_2)),
205                       rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/reps)),
206             mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
207                       std = sd(unlist(par_mu_v2_1)),
208                       rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/reps)),
209             mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],
210                       std = sd(unlist(par_mu_v2_2)),
211                       rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/reps)),
212             alpha_1 = c(bias = mean(unlist(par_alpha_1))-alpha[1],
213                       std = sd(unlist(par_alpha_1)),
214                       rmse = sqrt(sum((unlist(par_alpha_1)-alpha[1])^2)/reps)),
215             alpha_2 = c(bias = mean(unlist(par_alpha_2))-alpha[2],
216                       std = sd(unlist(par_alpha_2)),
217                       rmse = sqrt(sum((unlist(par_alpha_2)-alpha[2])^2)/reps)),
218             eta_1 = c(bias = mean(unlist(par_eta_1))-eta[1],
219                    std = sd(unlist(par_eta_1)),
220                    rmse = sqrt(sum((unlist(par_eta_1)-eta[1])^2)/reps)),
221             eta_2 = c(bias = mean(unlist(par_eta_2))-eta[2],
222                    std = sd(unlist(par_eta_2)),
223                    rmse = sqrt(sum((unlist(par_eta_2)-eta[2])^2)/reps)))
224 write.xlsx(result_nc,"resultados_nc_kmedoids.xlsx",overwrite = T)
225
226 #-----#
227 # distribucion t multivariada (kmeans)
228 #-----#
229 par_pi <- list()
230 par_mu_v1_1 <- list()
231 par_mu_v2_1 <- list()
232 par_mu_v1_2 <- list()
233 par_mu_v2_2 <- list()
234 #par_sigma_1 <- list()
235 #par_sigma_2 <- list()
236 par_df_1 <- list()
237 par_df_2 <- list()
238
239 id <- seq(1,ncol(data_tm), by=2)
240 for (i in 1:(ncol(data_tm)/2)) {
241   #i=2
242   estim <- MtM(data_tm[,id[i]:(id[i]+1)], G=2, show_progress = F,init_method = "kmeans")
243   par_pi[[i]] <- min(estim$pi) #Pi

```

```

244 main_mu <- data.frame(estim$mu)
245 names(main_mu) <- c("V1", "V2")
246 g1 <- main_mu %>% filter(V2<0)
247 g2 <- main_mu %>% filter(V2>0)
248 par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
249 par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
250 par_mu_v2_1[[i]] <- g2[1] #Mu de V2
251 par_mu_v2_2[[i]] <- g2[2] #Mu de V2
252 #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
253 #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
254 par_df_1[[i]] <- min(estim$df)
255 par_df_2[[i]] <- max(estim$df)
256 }
257
258 #Results t multivariate
259 result_tm <-
260 data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
261                   std = sd(unlist(par_pi)),
262                   rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/reps)),
263           mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
264                   std = sd(unlist(par_mu_v1_1)),
265                   rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/reps)),
266           mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
267                   std = sd(unlist(par_mu_v1_2)),
268                   rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/reps)),
269           mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
270                   std = sd(unlist(par_mu_v2_1)),
271                   rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/reps)),
272           mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],
273                   std = sd(unlist(par_mu_v2_2)),
274                   rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/reps)),
275           df_1 = c(bias = mean(unlist(par_df_1))-df[1],
276                   std = sd(unlist(par_df_1)),
277                   rmse = sqrt(sum((unlist(par_df_1)-df[1])^2)/reps)),
278           df_2 = c(bias = mean(unlist(par_df_2))-df[2],
279                   std = sd(unlist(par_df_2)),
280                   rmse = sqrt(sum((unlist(par_df_2)-df[2])^2)/reps)))
281 write.xlsx(result_tm, "resultados_tm_kmeans.xlsx", overwrite = T)
282
283 #-----#
284 # distribucion normal multivariada (kmeans)
285 #-----#
286 par_pi <- list()
287 par_mu_v1_1 <- list()
288 par_mu_v2_1 <- list()
289 par_mu_v1_2 <- list()
290 par_mu_v2_2 <- list()
291 #par_sigma_1 <- list()
292 #par_sigma_2 <- list()
293
294 id <- seq(1, ncol(data_nm), by=2)
295 for (i in 1:(ncol(data_nm)/2)) {
296   #i=2
297   estim <- MNM(data_nm[,id[i]:(id[i]+1)], G=2, show_progress = F, init_method = "kmeans")
298   par_pi[[i]] <- min(estim$pi) #Pi
299   main_mu <- data.frame(estim$mu)
300   names(main_mu) <- c("V1", "V2")
301   g1 <- main_mu %>% filter(V2<0)
302   g2 <- main_mu %>% filter(V2>0)
303   par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
304   par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
305   par_mu_v2_1[[i]] <- g2[1] #Mu de V2
306   par_mu_v2_2[[i]] <- g2[2] #Mu de V2
307   #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
308   #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
309 }
310
311 #Results normal multivariante
312 result_nm <-
313 data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
314                   std = sd(unlist(par_pi)),
315                   rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/length(par_pi))),
316           mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
317                   std = sd(unlist(par_mu_v1_1)),
318                   rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/length(par_pi))),
319           mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
320                   std = sd(unlist(par_mu_v1_2)),
321                   rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/length(par_pi))),
322           mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
323                   std = sd(unlist(par_mu_v2_1)),
324                   rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/length(par_pi))),
325           mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],

```

```

326         std = sd(unlist(par_mu_v2_2)),
327         rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/length(par_pi)))
328 write.xlsx(result_nm,"resultados_nm_kmeans.xlsx",overwrite = T)
329
330 #-----#
331 # distribucion normal contaminada multivariada (kmeans)
332 #-----#
333 par_pi <- list()
334 par_mu_v1_1 <- list()
335 par_mu_v2_1 <- list()
336 par_mu_v1_2 <- list()
337 par_mu_v2_2 <- list()
338 par_alpha_1 <- list()
339 par_alpha_2 <- list()
340 par_eta_1 <- list()
341 par_eta_2 <- list()
342
343 id <- seq(1,ncol(data_nc), by=2)
344 for (i in 1:(ncol(data_nc)/2)) {
345     #i=2
346     estim <- MCMC(data_nc[,id[i):(id[i]+1)], G=2, show_progress = F, init_method = "kmeans")
347     par_pi[[i]] <- min(estim$pi) #Pi
348     main_mu <- data.frame(estim$mu)
349     names(main_mu) <- c("V1","V2")
350     g1 <- main_mu %>% filter(V2<0)
351     g2 <- main_mu %>% filter(V2>0)
352     par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
353     par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
354     par_mu_v2_1[[i]] <- g2[1] #Mu de V2
355     par_mu_v2_2[[i]] <- g2[2] #Mu de V2
356     par_alpha_1[[i]] <- max(estim$alpha)
357     par_alpha_2[[i]] <- min(estim$alpha)
358     par_eta_1[[i]] <- min(estim$eta)
359     par_eta_2[[i]] <- max(estim$eta)
360     #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
361     #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
362 }
363
364 #Results contaminated normal
365 result_nc <-
366     data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
367                       std = sd(unlist(par_pi)),
368                       rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/reps)),
369               mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
370                           std = sd(unlist(par_mu_v1_1)),
371                           rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/reps)),
372               mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
373                           std = sd(unlist(par_mu_v1_2)),
374                           rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/reps)),
375               mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
376                           std = sd(unlist(par_mu_v2_1)),
377                           rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/reps)),
378               mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],
379                           std = sd(unlist(par_mu_v2_2)),
380                           rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/reps)),
381               alpha_1 = c(bias = mean(unlist(par_alpha_1))-alpha[1],
382                           std = sd(unlist(par_alpha_1)),
383                           rmse = sqrt(sum((unlist(par_alpha_1)-alpha[1])^2)/reps)),
384               alpha_2 = c(bias = mean(unlist(par_alpha_2))-alpha[2],
385                           std = sd(unlist(par_alpha_2)),
386                           rmse = sqrt(sum((unlist(par_alpha_2)-alpha[2])^2)/reps)),
387               eta_1 = c(bias = mean(unlist(par_eta_1))-eta[1],
388                        std = sd(unlist(par_eta_1)),
389                        rmse = sqrt(sum((unlist(par_eta_1)-eta[1])^2)/reps)),
390               eta_2 = c(bias = mean(unlist(par_eta_2))-eta[2],
391                        std = sd(unlist(par_eta_2)),
392                        rmse = sqrt(sum((unlist(par_eta_2)-eta[2])^2)/reps))
393 write.xlsx(result_nc,"resultados_nc_kmeans.xlsx",overwrite = T)
394
395 #-----#
396 # distribucion t multivariada (hierarchical)
397 #-----#
398 par_pi <- list()
399 par_mu_v1_1 <- list()
400 par_mu_v2_1 <- list()
401 par_mu_v1_2 <- list()
402 par_mu_v2_2 <- list()
403 #par_sigma_1 <- list()
404 #par_sigma_2 <- list()
405 par_df_1 <- list()
406 par_df_2 <- list()
407

```

```

408 id <- seq(1,ncol(data_tm), by=2)
409 for (i in 1:(ncol(data_tm)/2)) {
410   #i=2
411   estim <- MtM(data_tm[,id[i]:(id[i]+1)], G=2, show_progress = F,init_method = "hierarchical")
412   par_pi[[i]] <- min(estim$pi) #Pi
413   main_mu <- data.frame(estim$mu)
414   names(main_mu) <- c("V1","V2")
415   g1 <- main_mu %>% filter(V2<0)
416   g2 <- main_mu %>% filter(V2>0)
417   par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
418   par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
419   par_mu_v2_1[[i]] <- g2[1] #Mu de V2
420   par_mu_v2_2[[i]] <- g2[2] #Mu de V2
421   #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
422   #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
423   par_df_1[[i]] <- min(estim$df)
424   par_df_2[[i]] <- max(estim$df)
425 }
426
427 #Results t multivariate
428 result_tm <-
429   data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
430                     std = sd(unlist(par_pi)),
431                     rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/reps)),
432             mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
433                       std = sd(unlist(par_mu_v1_1)),
434                       rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/reps)),
435             mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
436                       std = sd(unlist(par_mu_v1_2)),
437                       rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/reps)),
438             mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
439                       std = sd(unlist(par_mu_v2_1)),
440                       rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/reps)),
441             mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],
442                       std = sd(unlist(par_mu_v2_2)),
443                       rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/reps)),
444             df_1 = c(bias = mean(unlist(par_df_1))-df[1],
445                   std = sd(unlist(par_df_1)),
446                   rmse = sqrt(sum((unlist(par_df_1)-df[1])^2)/reps)),
447             df_2 = c(bias = mean(unlist(par_df_2))-df[2],
448                   std = sd(unlist(par_df_2)),
449                   rmse = sqrt(sum((unlist(par_df_2)-df[2])^2)/reps)))
450 write.xlsx(result_tm,"resultados_tm_hierarchical.xlsx",overwrite = T)
451
452 #-----#
453 # distribucion normal multivariada (hierarchical)
454 #-----#
455 par_pi <- list()
456 par_mu_v1_1 <- list()
457 par_mu_v2_1 <- list()
458 par_mu_v1_2 <- list()
459 par_mu_v2_2 <- list()
460 #par_sigma_1 <- list()
461 #par_sigma_2 <- list()
462
463 id <- seq(1,ncol(data_nm), by=2)
464 for (i in 1:(ncol(data_nm)/2)) {
465   #i=2
466   estim <- MNM(data_nm[,id[i]:(id[i]+1)], G=2, show_progress = F,init_method = "hierarchical")
467   par_pi[[i]] <- min(estim$pi) #Pi
468   main_mu <- data.frame(estim$mu)
469   names(main_mu) <- c("V1","V2")
470   g1 <- main_mu %>% filter(V2<0)
471   g2 <- main_mu %>% filter(V2>0)
472   par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
473   par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
474   par_mu_v2_1[[i]] <- g2[1] #Mu de V2
475   par_mu_v2_2[[i]] <- g2[2] #Mu de V2
476   #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
477   #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
478 }
479
480 #Results normal multivariante
481 result_nm <-
482   data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
483                     std = sd(unlist(par_pi)),
484                     rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/reps)),
485             mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
486                       std = sd(unlist(par_mu_v1_1)),
487                       rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/reps)),
488             mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
489                       std = sd(unlist(par_mu_v1_2)),

```



```

490         rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/reps)),
491     mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
492               std = sd(unlist(par_mu_v2_1)),
493               rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/reps)),
494     mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],
495               std = sd(unlist(par_mu_v2_2)),
496               rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/reps))
497
498 write.xlsx(result_nm,"resultados_nm_hierarchical.xlsx",overwrite = T)
499
500 #-----#
501 # distribucion normal contaminada multivariada (hierarchical)
502 #-----#
503 par_pi <- list()
504 par_mu_v1_1 <- list()
505 par_mu_v2_1 <- list()
506 par_mu_v1_2 <- list()
507 par_mu_v2_2 <- list()
508 par_alpha_1 <- list()
509 par_alpha_2 <- list()
510 par_eta_1 <- list()
511 par_eta_2 <- list()
512
513 id <- seq(1,ncol(data_nc), by=2)
514 for (i in 1:(ncol(data_nc)/2)) {
515     #i=2
516     estim <- MCMC(data_nc[,id[i]:(id[i]+1)], G=2, show_progress = F, init_method = "hierarchical")
517     par_pi[[i]] <- min(estim$pi) #Pi
518     main_mu <- data.frame(estim$mu)
519     names(main_mu) <- c("V1","V2")
520     g1 <- main_mu %>% filter(V2<0)
521     g2 <- main_mu %>% filter(V2>0)
522     par_mu_v1_1[[i]] <- g1[1] #Mu de V1 3
523     par_mu_v1_2[[i]] <- g1[2] #Mu de V2 -3
524     par_mu_v2_1[[i]] <- g2[1] #Mu de V2
525     par_mu_v2_2[[i]] <- g2[2] #Mu de V2
526     par_alpha_1[[i]] <- max(estim$alpha)
527     par_alpha_2[[i]] <- min(estim$alpha)
528     par_eta_1[[i]] <- min(estim$eta)
529     par_eta_2[[i]] <- max(estim$eta)
530     #par_sigma_1[[i]] <- data.frame(estim$sigma)[c(1:2),c(1:2)] #Sigma 1
531     #par_sigma_2[[i]] <- data.frame(estim$sigma)[c(1:2),c(3:4)] #Sigma 2
532 }
533
534 #Results contaminated normal
535 result_nc <-
536     data.frame(pi_1 = c(bias = mean(unlist(par_pi))-pi[1],
537                       std = sd(unlist(par_pi)),
538                       rmse = sqrt(sum((unlist(par_pi)-pi[1])^2)/reps)),
539     mu_v1_1 = c(bias = mean(unlist(par_mu_v1_1))-mu[[1]][1],
540               std = sd(unlist(par_mu_v1_1)),
541               rmse = sqrt(sum((unlist(par_mu_v1_1)-mu[[1]][1])^2)/reps)),
542     mu_v1_2 = c(bias = mean(unlist(par_mu_v1_2))-mu[[1]][2],
543               std = sd(unlist(par_mu_v1_2)),
544               rmse = sqrt(sum((unlist(par_mu_v1_2)-mu[[1]][2])^2)/reps)),
545     mu_v2_1 = c(bias = mean(unlist(par_mu_v2_1))-mu[[2]][1],
546               std = sd(unlist(par_mu_v2_1)),
547               rmse = sqrt(sum((unlist(par_mu_v2_1)-mu[[2]][1])^2)/reps)),
548     mu_v2_2 = c(bias = mean(unlist(par_mu_v2_2))-mu[[2]][2],
549               std = sd(unlist(par_mu_v2_2)),
550               rmse = sqrt(sum((unlist(par_mu_v2_2)-mu[[2]][2])^2)/reps)),
551     alpha_1 = c(bias = mean(unlist(par_alpha_1))-alpha[1],
552               std = sd(unlist(par_alpha_1)),
553               rmse = sqrt(sum((unlist(par_alpha_1)-alpha[1])^2)/reps)),
554     alpha_2 = c(bias = mean(unlist(par_alpha_2))-alpha[2],
555               std = sd(unlist(par_alpha_2)),
556               rmse = sqrt(sum((unlist(par_alpha_2)-alpha[2])^2)/reps)),
557     eta_1 = c(bias = mean(unlist(par_eta_1))-eta[1],
558              std = sd(unlist(par_eta_1)),
559              rmse = sqrt(sum((unlist(par_eta_1)-eta[1])^2)/reps)),
560     eta_2 = c(bias = mean(unlist(par_eta_2))-eta[2],
561              std = sd(unlist(par_eta_2)),
562              rmse = sqrt(sum((unlist(par_eta_2)-eta[2])^2)/reps))
563 write.xlsx(result_nc,"resultados_nc_hierarchical.xlsx",overwrite = T)

```

8.3. Código en R para la aplicación práctica

```
1 #-----#
2 # Analisis de aplicacion
3 #-----#
4 setwd(dirname(rstudioapi::getSourceEditorContext()$path))
5 library(tidyverse)
6 library(mvtnorm) #Student t y MN
7 library(ContaminatedMixt) #MCN
8 library(mice) #Missing values
9 library(clusterGeneration) #genRandomClust una mixtura de MN
10 library(MixGHD) #ARI
11 library(haven)
12 library(cluster)
13 library(MixtureMissing)
14 library(openxlsx)
15 library(gridExtra)
16 set.seed(420)
17
18 #-----#
19 # Data
20 #-----#
21 data <- read_sav("EVA_2021_2S_HSE_VP.sav")
22 PV1 <- data %>% dplyr::select(starts_with("PV1"))
23 PV1$nas <- rowSums(is.na(PV1))
24 PV1 <- PV1 %>% filter(nas<5) %>% dplyr::select(-nas)
25 PV1 <- bind_cols(lapply(PV1, function(x) {as.numeric(as.character(x))}))
26 PV1 <- PV1 %>% dplyr::select(-PV1.DSC_1)
27
28 #-----#
29 # Descripcion de los datos
30 #-----#
31 prop.table(table(data$sexo))*100
32 prop.table(table(data$area))*100
33 prop.table(table(data$gestion2))*100
34
35 # Histogramas de las variables
36 vars <- names(PV1)
37 graphs <- list()
38 for (i in 1:ncol(PV1)) {
39   tgraph <- PV1[,i] %>% rename(value = 1)
40   graphs[[i]] <-
41     ggplot(tgraph, aes(x=value)) +
42       geom_histogram(color="black", fill="white")+theme_bw()+ylab("Frecuencia")+xlab(vars[i])+
43       scale_x_continuous(limits=c(-5,5), breaks = seq(-5,5))
44 }
45
46 gr1 <-
47 grid.arrange(graphs[[1]], graphs[[2]],
48             graphs[[3]], graphs[[4]],
49             graphs[[5]], nrow = 3)
50 gr2 <-
51 grid.arrange(graphs[[6]], graphs[[7]],
52             graphs[[8]], graphs[[9]],
53             graphs[[10]], nrow = 3)
54 gr3 <-
55 grid.arrange(graphs[[11]], graphs[[12]],
56             graphs[[13]], graphs[[14]], nrow = 2)
57
58 ggsave(paste0("histograma_1.png"),gr1,dpi=300,height = 8,width = 6)
59 ggsave(paste0("histograma_2.png"),gr2,dpi=300,height = 8,width = 6)
60 ggsave(paste0("histograma_3.png"),gr3,dpi=300,height = 5.333333,width = 6)
61
62 #-----#
63 # distribucion normal contaminada multivariada: Ajuste
64 #-----#
65 fit2 <- MCNM(PV1,G=2,max_iter=20)
66 fit3 <- MCNM(PV1,G=3,max_iter=20)
67 fit4 <- MCNM(PV1,G=4,max_iter=20)
68 fit5 <- MCNM(PV1,G=5,max_iter=20)
69 fit6 <- MCNM(PV1,G=6,max_iter=20)
70 #fit7 <- MCNM(PV1,G=7,max_iter=20)
71 #fit8 <- MCNM(PV1,G=8,max_iter=20)
72 #fit9 <- MCNM(PV1,G=9,max_iter=20)
73 #fit10 <- MCNM(PV1,G=10,max_iter=20)
74 #fit11 <- MCNM(PV1,G=11,max_iter=20)
75 #fit12 <- MCNM(PV1,G=12,max_iter=20)
76 #fit13 <- MCNM(PV1,G=13,max_iter=20)
77 #fit14 <- MCNM(PV1,G=14,max_iter=20)
78
79 fitstats <- list()
```

```

80 results <- list(fit2,fit3,fit4,fit5,fit6)
81 for (i in 1:5) {
82   fitstats[[i]] <- data.frame(
83     loglike=results[[i]]$final_loglik,
84     AIC=results[[i]]$AIC,
85     BIC=results[[i]]$BIC,
86     KIC=results[[i]]$KIC,
87     KICc=results[[i]]$KICc,
88     AIC3=results[[i]]$AIC3,
89     CAIC=results[[i]]$CAIC,
90     AICc=results[[i]]$AICc,
91     ICL=results[[i]]$ICL,
92     AWE=results[[i]]$AWE,
93     CLC=results[[i]]$CLC)
94 }
95
96 results <- round(data.frame(G = c(2,3,4,5,6),bind_rows(fitstats)),1)
97 write.csv(results,"ajustes.csv")
98
99 #-----#
100 # distribucion normal contaminada multivariada: G2
101 #-----#
102 fit2$pi
103 fit2$mu
104 #fit2$sigma
105 fit2$alpha
106 fit2$eta
107 fit2$clusters
108 fit2$outliers
109 summary(fit2)
110
111 local <- data.frame(fit2$mu)
112 local$Conglomerado <- c(1,2)
113 local <- local %>% pivot_longer(!Conglomerado,names_to="Constructo",values_to = "Media")
114 local$Conglomerado <- as.factor(local$Conglomerado)
115
116 #nombres <- c("Autoeficacia/nAca")
117 plotg2 <-
118 ggplot(data=local, aes(x=Constructo, y=Media, group=Conglomerado)) +
119   geom_line(aes(color=Conglomerado))+
120   geom_point(aes(color=Conglomerado))+theme_bw()+
121   theme(axis.text.x=element_text(size=6),legend.position = "bottom")
122
123 ggsave("mu_conglomerados_g2.png",plotg2,dpi=300,width = 10,height = 5.5)
124 #[1] 9.952381 5.438095
125
126 #-----#
127 # distribucion normal contaminada multivariada: G4
128 #-----#
129 fit4$pi
130 #fit4$mu
131 table(fit4$clusters)
132 summary(fit4)
133 #fit2$sigma
134 fit4$alpha
135 fit4$eta
136 #fit2$outliers
137
138
139 local <- data.frame(fit4$mu)
140 local$Conglomerado <- c(1,2,3,4)
141 local <- local %>% pivot_longer(!Conglomerado,names_to="Constructo",values_to = "Media")
142 local$Conglomerado <- as.factor(local$Conglomerado)
143
144 #nombres <- c("Autoeficacia/nAca")
145 plotg4 <-
146 ggplot(data=local, aes(x=Constructo, y=Media, group=Conglomerado)) +
147   geom_line(aes(color=Conglomerado))+
148   geom_point(aes(color=Conglomerado))+theme_bw()+
149   theme(axis.text.x=element_text(size=6),legend.position = "bottom")
150
151 ggsave("mu_conglomerados_g4.png",plotg4,dpi=300,width = 10,height = 5.5)
152 #[1] 9.952381 5.438095

```

Bibliografía

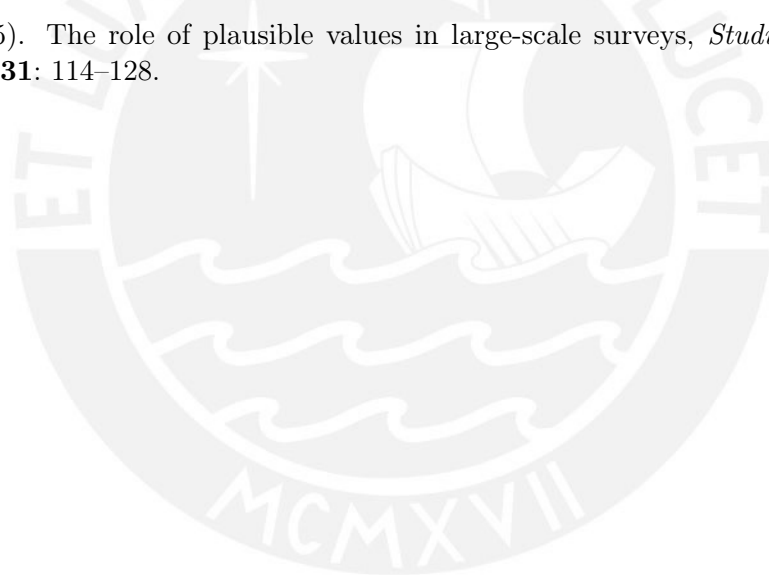
- Adams, R. J., Wu, M. L. y Wilson, M. (2012). The Rasch Rating Model and the Disordered Threshold Controversy, *Educational and Psychological Measurement* **72**: 547—573.
- Akogul, S. y Erisoglu, M. (2016). A Comparison of Information Criteria in Clustering Based on Mixture of Multivariate Normal Distributions, *Mathematical and Computational Applications* **21**: 34.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models , *Psychometrika* **42**: 69–81.
- Andrich, D. (1978). Rating Formulation for Ordered Response Categories, *Psychometrika* **43**: 561–573.
- Andrich, D. (1988). *Rasch models for measurement*, Sage.
- Andrich, D. y Marais, I. (2020). *A course in Rasch measurement theory. Measuring in the educational, social and health sciences*, Springer.
- Baker, F. B. y Kim, S. (2017). *The Basics of Item Response Theory Using R*, Springer.
- Birnbaum, S. A. (1968). Some latent trait models and their use in inferring an examinee's ability, *Journal of Applied Probability* **5**: 392–398.
- Bock, R. D. y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm, *Psychometrika* **46**: 443–459.
- Bock, R. D. y Gibbons, R. D. (2021). *Item Response Theory*, Wiley.
- Bock, R. D. y Lieberman, M. (1970). Fitting a response model for n dichotomously scored items, *Psychometrika* **35**: 179—197.
- Bock, R. D. y Mislevy, R. J. (1982). Adaptive eap estimation of ability in a microcomputer environment, *Applied Psychological Measurement* **6**: 36.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences, *Annual Review of Psychology* **53**: 605–634.
- Bond, T. G. y Fox, C. M. (2017). *Applying the Rasch model: Fundamental measurement in the human sciences*, 3 edn, Lawrence Erlbaum.
- Bond, T. G., Zi, Y. y Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*, 4 edn, Routledge.
- Borsboom, D., Mellenbergh, G. J. y van Heerden, J. (2003). The theoretical status of latent variables , *Psychological Review* **110**: 203–219.
- Borsboom, D. y Scholten, A. Z. (2008). The Rasch model and Conjoint Measurement theory from the perspective of psychometrics , *Theory & Psychology* **18**: 111–117.

- Bouveyron, C. y Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review , *Computational Statistics and Data Analysis* **71**: 52–78.
- Bouveyron, C., Celeux, G., Murphy, T. B. y Raftery, A. E. (2019). *Model-based clustering and classification for data science with applications in R*, Cambridge University Press.
- Christensen, S. A., Farr, M. T. y Williams, D. M. (2021). Assessment and novel application of N-mixture models for aerial surveys of wildlife , *Ecosphere* **12**: e03725.
- Compiani, G. y Kitamura, Y. (2016). Using Mixtures in Econometric Models: A Brief Review and Some New Results , *The Econometrics Journal* **19**: 95—127.
- Demg, H. y Han, J. (2014). *Data clustering: Algorithms and applications*, CRC Press.
- Dempster, A. P., Laird, N. M. y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)* **39**: 1–22.
- Dodge, Y. (2008). *The Concise Encyclopedia of Statistics*, Springer.
- Dragow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model, *Applied Psychological Measurement* **13**.
- Duckworth, A. L. y Yeager, D. S. (2021). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes, *Educational Researcher* **44**: 237—251.
- Engelhard, G. y Wang, J. (2021). *Rasch Models for Solving Measurement Problems: Invariant Measurement in the Social Sciences (Quantitative Applications in the Social Sciences)*, Sage.
- Engelhard, G. y Wind, S. (2018). *Invariant measurement with raters and rating scales*, Routledge.
- Everitt, B. S. y Skrondal, A. (2010). *The Cambridge dictionary of statistics*, Cambridge University Press.
- Finch, W. H. y French, B. F. (2019). *Educational and psychological measurement*, Routledge.
- Fisher, W. P. (1994). *Objective measurement: Theory into practice*, Vol. 2, Ablex, chapter The Rasch debate: Validity and revolution in educational measurement, pp. 36—72.
- García-Cabrero, B. (2018). Las habilidades socioemocionales, no cognitivas o “blandas”: aproximaciones a su evaluación, *Revista Digital Universitaria* **19**: 1—17.
- Gianola, D., Boettcher, P. J., Odegard, J. y Heringstad, B. (2007). Mixture models in quantitative genetics and applications to animal breeding, *Revista Brasileira de Zootecnia* **36**: 172–183.
- Giordani, P., Ferraro, M. B. y Martella, F. (2020). *An introduction to clustering with R*, Springer.
- Golam Kibria, B. M. y Joarder, A. H. (2006). A short review of the Multivariate t-Distribution, *Journal of Statistical Research* **40**: 59–72.
- Grami, A. (2020). *Probability, Random Variables, Statistics, and Random Processes: Fundamentals & Applications*, Wiley.
- Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology , *Frontiers in Psychology* **4**: 1–4.

- Hunter, D. R. y Lange, K. (2012). A Tutorial on MM Algorithms, *The American Statistician* **58**: 30–37.
- Khine, M. S. (2020). *Rasch Measurement Applications in Quantitative Educational Research*, Springer.
- Kuhn, M. A. y Feigelson, E. D. (2018). *Handbook of Mixture Analysis*, CRC Press, chapter Mixture Models in Astornomy, pp. 463–490.
- Kyngdon, A. (2008). Conjoint Measurement, error and the Rasch model , *Theory & Psychology* **18**: 125–131.
- Lamprianou, I. (2020). *Applying the Rasch Model in Social Sciences Using R and BlueSky Statistics*, Routledge.
- Lim, R. y Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning, *Journal of Applied Psychology* **75**.
- Lin, P.-E. (1972). Some characterizations of the multivariate t distribution, *Journal of Multivariate Analysis* **2**(3): 339–344.
URL: <https://www.sciencedirect.com/science/article/pii/0047259X72900218>
- Linacre, J. M. (1999). Understanding rasch measurement: estimation methods for rasch, *Journal of Outcome Measurement* **3**: 381–405.
- Loehlin, J. C. y Beaujean, A. A. (2017). *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*, Taylor & Francis.
- Lord, M. (1980). *Applications of Item Response Theory To Practical Testing Problems*, Routledge.
- Luce, R. D. y Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement , *Journal of Mathematical Psychology* **1**: 1–27.
- Mavrakakis, M. C. y Penzer, J. (2021). *Probability and Statistical Inference From Basic Principles to Advanced Models*, CRC Press.
- McLachlan, G. J. y Krishnan, T. (2008). *The EM algorithm and extensions*, Wiley.
- McLachlan, G. J. y Peel, D. (2000). *Finite Mixture Models*, Wiley.
- McNicholas, P. D. (2016). Model-based Clustering, *Journal of Classification* **33**: 331–373.
- Meng, X. L. y Rubin, D. B. (1993). Maximum likelihood esstimation via the ECM algorithm: A general framework, *Biometrika* **80**: 267–278.
- Michell, J. (1985). *Measurement and personality assessment*, Elsevier, chapter Additivity in psychological measurement, pp. 101—112.
- Michell, J. (2000). Normal science, pathological science and psychometrics , *Theory & Psychology* **10**: 639–667.
- Michell, J. (2007). *Philosophy of anthropology and sociology*, Elsevier, chapter Measurement, pp. 71—120.
- Michell, J. (2008). Conjoint measurement and the Rasch paradox , *Theory & Psychology* **18**: 119–124.

- Michell, J. (2014). The Rasch paradox, conjoint measurement, and psychometrics: Response to Humphry and Sijtsma , *Theory & Psychology* **24**: 11–123.
- Miettunen, J., Nordstrom, T., Kaakinen, T. y Ahmed, A. O. (2016). Latent variable mixture modeling in psychiatric research A review and application, *Psychological medicine* **36**: 457–467.
- Nielsen, S. F. (2000). The Stochastic EM Algorithm: Estimation and Asymptotic Results, *Bernoulli* **6**: 457–489.
- Paek, I. y Cole, K. (2020). *Using R for Item Response Theory model applications*, Routledge.
- Perline, R., Wright, B. D. y Wainer, H. (1979). The Rasch model as Additive Conjoint Measurement , *Applied Psychological Measurement* **3**: 237–255.
- Punzo, A., Mazza, A. y D, M. P. (2018). ContaminatedMix: An R Package for Fitting Parsimonious Mixtures of Multivariate Contaminated Normal Distributions, *Journal of Statistical Software* **85**: 1–25.
- Punzo, A. y McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions, *Biometrical Journal* **58**: 1506–1537.
- Punzo, A. y Tortora, C. (2021). Multiple scaled contaminated normal distribution and its application in clustering, *Statistical Modelling* **21**: 332–358.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*, Copenhagen.
- Raykov, T. y Marcoulides, G. A. (2011). *Introduction to psychometric theory*, Routledge.
- Rubin, D. B. (1976). Inference and Missing Data, *Biometrika* **63**: 581–592.
- Rust, B., Kosinski, M. y Stillwell, D. (2020). *Modern Psychometrics The Science of Psychological Assessment*, Routledge.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometrika Monograph Supplement* **34**: 100.
- Schlattmann, P. (2009). *Medical applications of Finite Mixture Models*, Springer.
- Schwarz, G. (1978). Estimating the Dimension of a Model, *The Annals of Statistics* **6**: 461–464.
- Sijtsma, K. (2011). Introduction to the measurement of psychological attributes , *Measurement* **44**: 1209–1219.
- Tanner, M. A. y Wong, H. (1987). The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association* **82**: 528–540.
- Tinsley, H. E. A. y Brown, S. D. (2000). *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, Academic Press.
- Tong, H. y Tortora, C. (2022). Model-based clustering and outlier detection with missing data, *Advances in Data Analysis and Classification* **16**: 5–30.
- Tukey, J. W. (1960). *Contributions to Probability and Statistics*, Stanford University Press, chapter A Survey of Sampling from Contaminated Distributions.

- Tuma, M. N. y Decker, R. (2013). Finite mixture models in market segmentation: A review and suggestions for best practices, *The Electronic Journal of Business Research Methods* **11**: 2.
- Vila, S., Gilar-Corbí, R. y Pozo-rico, T. (2021). Effects of Student Training in Social Skills and Emotional Intelligence on the Behaviour and Coexistence of Adolescents in the 21st Century, *International journal of environmental research and public health* **18**: 5498.
- Warm, T. (1989). Weighted likelihood estimation of ability in item response theory, *Psychometrika* **54**: 427—450.
- Wolfe, J. H. (1967). NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions, *Research Memo. SRM* **68-2**.
- Wright, B. D. y Douglas, G. A. (1975). *Best test design and self-tailored testing*, MESA.
- Wright, B. D. y Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval, *Archives of Physical Medicine and Rehabilitation* **70**: 857–860.
- Wright, B. D. y Masters, G. N. (1979). *Best test design*, MESA.
- Wright, B. D. y Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*, MESA.
- Wu, M. (2005). The role of plausible values in large-scale surveys, *Studies in Educational Evaluation* **31**: 114–128.



Anexo 1

A. Modelo Rasch

En psicometría, el objeto de estudio corresponde a la medición de constructos psicológicos como habilidades cognitivas, estados de ánimo, motivación o rasgos de personalidad (Rust et al., 2020). Debido a la naturaleza latente de estos atributos, los modelos psicométricos plantean una medición indirecta al establecer una relación entre una estructura latente y una serie de variables observables.

En los instrumentos psicométricos, las variables observables corresponden a un conjunto de respuestas a los ítems que contienen muestras del comportamiento teóricamente asociado a un atributo psicológico (Sijtsma, 2011). De esta manera, las respuestas a los ítems se consideran como manifestaciones observables del nivel del rasgo latente de un individuo (Raykov y Marcoulides, 2011). En psicología, esta relación entre constructos no observables e indicadores manifiestos es modelada a través de modelos de variables latentes (Borsboom et al., 2003).

A través de los años, el estudio sobre los modelos de variables latentes se ha ampliado considerablemente (Bollen, 2002). En la actualidad coexiste una amplia cantidad de modelos de variables latentes empleados en psicometría (Finch y French, 2019; Loehlin y Beaujean, 2017). En el presente anexo se introducen los principales modelos probabilísticos de la familia Rasch en el modelamiento de respuestas a ítems dicotómicos y politómicos ordenados, con énfasis en la propuesta de Andrich (1978).

B. Modelo Rasch para ítems dicotómicos

George Rasch, matemático danés, reconoció que la interacción entre un individuo y un ítem corresponde a un proceso estocástico en donde no es posible determinar con certeza cuál será el resultado (Bond et al., 2020). Por estos motivos, Rasch (1960) propuso un modelo de medición para ítems dicotómicos no determinista, en donde la probabilidad de acertar a un ítem está gobernada principalmente por la diferencia entre la habilidad del individuo y la dificultad del ítem (Andrich y Marais, 2020). La familia de modelos Rasch se circunscribe en un marco filosófico de medición objetiva e invariante que requiere del cumplimiento de estrictas condiciones para construir medidas en ciencias sociales con propiedades similares a las de las ciencias físicas (Bond et al., 2020; Wright y Masters, 1979). En otras palabras, el modelo Rasch establece un marco prescriptivo de medición al cual los datos deben ajustarse para producir medidas de personas cuyas comparaciones son invariantes ante los ítems empleados para estimarlas y; de la misma forma, calibraciones de ítems independientes del conjunto evaluado de personas (Engelhard y Wang, 2021). El supuesto principal del modelo establece que siempre que un individuo tenga mayor nivel del rasgo latente que otro, su probabilidad de acertar a cualquier ítem será mayor. Del mismo modo, siempre que un ítem tenga mayor dificultad que otro, la probabilidad de que cualquiera de los individuos lo acierte será menor (Wright y Masters, 1979). Esta relación se expresa en la siguiente función de respuesta:

$$P(X_{ni} = 1|\theta_n, b_i) = \frac{e^{(\theta_n - b_i)}}{1 + e^{(\theta_n - b_i)}}. \quad (1)$$

En la formulación presentada del modelo, la probabilidad de que la respuesta X del individuo n al enfrentarse al ítem i resulte en un acierto $X_{ni} = 1$, se encuentra condicionada por sus respectivos parámetros de localización θ_n y b_i (Engelhard y Wang, 2021). La diferencia entre ambos parámetros $\theta_n - b_i$ delimita la probabilidad de acierto a través de una función de enlace logística (Khine, 2020). En este modelo, θ se interpreta como el nivel del individuo en el rasgo latente y b como la dificultad del ítem (Bond et al., 2020). Las medidas de localización de ítems y personas se expresan en una métrica común basada en el logaritmo de odds ratio, o logits que expresan el posicionamiento del individuo o ítem en el continuum latente (Bond et al., 2020). La escala se establece en un rango de $-\infty$ a ∞ ; por ello, este es el rango de posibles valores tanto para los parámetros θ como b . Como resultado, la relación entre ambas medidas y la probabilidad de acierto expresada en la función de respuesta puede ser representada a partir de Curvas Características de Ítems (CCI), funciones que expresan el cambio en la probabilidad de acierto a cada ítem en distintos niveles del rasgo latente (Baker y Kim, 2017).

La forma de la función para cada ítem se denomina sigmoideal o s-shape y denota la relación monotónica entre la probabilidad de acierto y la habilidad latente θ . Dicha relación monotónica es una realización del supuesto principal del modelo, pues mientras mayor sea la habilidad latente de un individuo, mayor será la probabilidad de acertar a cualquier ítem al que se enfrente. Asimismo, el parámetro de localización de los ítems corresponde al punto en el eje de las abscisas en donde la probabilidad de acierto al ítem es igual a ,50, denominado punto de inflexión (Bock y Gibbons, 2021). Como todas las funciones de respuesta para cada ítem tienen la misma pendiente, la probabilidad de acertar a un ítem con mayor dificultad siempre será menor en comparación a un ítem con menor dificultad.

En la Figura B.1 se presentan las CCI para tres ítems hipotéticos con parámetros de dificultad definidos como $b_1=-1,00$, $b_2=0,00$ y $b_3=1,00$, considerando la función de respuesta del modelo Rasch para ítems dicotómicos, tal y como se expresa en (1).

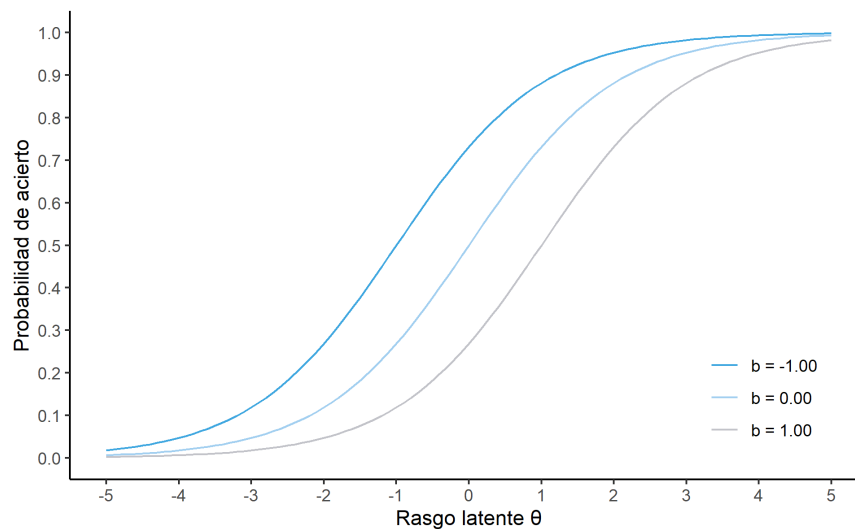


Figura B.1: Curvas Características para tres ítems hipotéticos con parámetros de dificultad definidos como $b_1=-1,00$, $b_2=0,00$ y $b_3=1,00$, bajo el modelo Rasch para ítems dicotómicos

C. Modelo de Escala de Valoración

Rasch (1960) propuso una aproximación para el modelamiento de escalas de valoración que fue posteriormente ampliada en los trabajos de Andersen (1977) y Andrich (1978) dando como resultado al modelo de Escala de Valoración (RSM). Este modelo es considerado como parte de la familia de modelos Rasch y una extensión del modelo para ítems dicotómicos. El RSM se aplica a situaciones en donde todos los ítems tienen las mismas categorías de respuesta y estas denotan un orden con respecto al posicionamiento del individuo en el continuum latente; por ejemplo, en escalas Likert (Lamprianou, 2020). En el RSM, cada categoría de respuesta es calificada con un número entero comenzando por el 0. A partir de ello se asume que el continuum del rasgo latente puede ser particionado en las respectivas categorías de respuesta ordinales a través de una serie de puntos sucesivos denominados umbrales τ_j . Para representar esta idea, a continuación, se presenta en la Figura C.1 una escala de respuesta Likert de acuerdo con $m=5$ categorías de respuesta y $m-1$ umbrales.

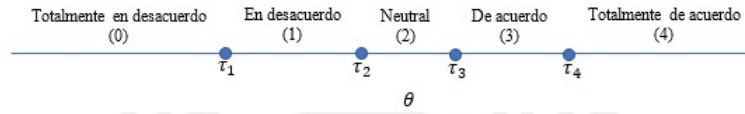


Figura C.1: Umbrales en un ítem tipo Likert de 5 alternativas de respuesta

La noción de umbral puede tener diferentes interpretaciones dependiendo de cómo se definen en el modelo. Particularmente, los modelos Rasch emplean umbrales definidos como categorías adyacentes (Andrich, 1978). En esta definición, el umbral o punto de intersección es el valor de θ en donde la probabilidad de responder a la categoría $k+1$ es igual a la probabilidad de responder a la categoría k , pues el umbral corresponde al límite entre dos categorías adyacentes. Los umbrales también son denominados parámetros estructurales de los ítems que denotan las distancias relativas entre categorías de respuesta adyacentes (Wright y Douglas, 1975). En el RSM se asume que la estructura de umbrales es común para todos los ítems que componen la escala, siempre y cuando compartan el mismo formato de respuesta (Engelhard y Wind, 2018). En otras palabras, el modelo asume que, si todos los ítems tienen las mismas categorías de respuesta, entonces la distancia entre las categorías también debería ser la misma. Esta cuestión no necesariamente implica que los umbrales de todos los ítems tengan la misma ubicación en el continuum latente. Cada ítem, además de contar con parámetros estructurales, también cuenta con un parámetro de localización δ que representa su ubicación en el continuum latente, análogo a la dificultad del ítem (Bock y Gibbons, 2021). En este sentido, los umbrales son desviaciones con respecto a la ubicación δ del ítem, con la restricción de que la suma de sus valores debe ser cero; i.e., $\sum_{j=1}^{m-1} \tau_j = 0$.

En síntesis, el RSM establece que los ítems de una escala de valoración tienen un parámetro de localización y una serie de parámetros estructurales comunes. De este modo, el modelo integra ambos parámetros en la formulación.

$$P(X_{ni} = x | \theta_n, \delta_i, \boldsymbol{\tau}) = \frac{e^{\sum_{j=0}^x (\theta_n - (\delta_i + \tau_j))}}{\sum_{k=0}^{m_i} e^{\sum_{j=0}^k (\theta_n - (\delta_i + \tau_j))}}, \quad (2)$$

En el modelo presentado en 2, $P(X_{ni} = x|\theta_n, \delta_i, \boldsymbol{\tau})$ representa la probabilidad P de obtener una de las categorías de respuesta x (en el ejemplo: 0, 1, 2, 3 o 4) dados el parámetro de localización θ del individuo y δ del ítem, con un vector de parámetros de umbrales $\boldsymbol{\tau}$ (Paek y Cole, 2020; Andrich y Marais, 2020). Como se observa en la formulación del modelo, cada posible categoría tendrá su propia función de respuesta o CCI. En este punto es importante indicar que para modelar el caso en donde $P(X_{ni} = x|\theta_n, \delta_i, \boldsymbol{\tau})$ es posible introducir un umbral $\tau_0 = 0$ que permite representar la situación en donde el individuo no supera ninguno de los umbrales y obtiene la categoría con el nivel más bajo en el continuum latente (Andrich y Marais, 2020). En la Figura C.2 se presentan las CCIs para cinco categorías de respuesta de un ítem hipotético con parámetro de localización definido como $\delta=0$ y parámetros de umbrales equidistantes $\tau_1 = -2,00$, $\tau_2 = -1,00$, $\tau_3 = 1,00$, $\tau_4 = 2,00$.

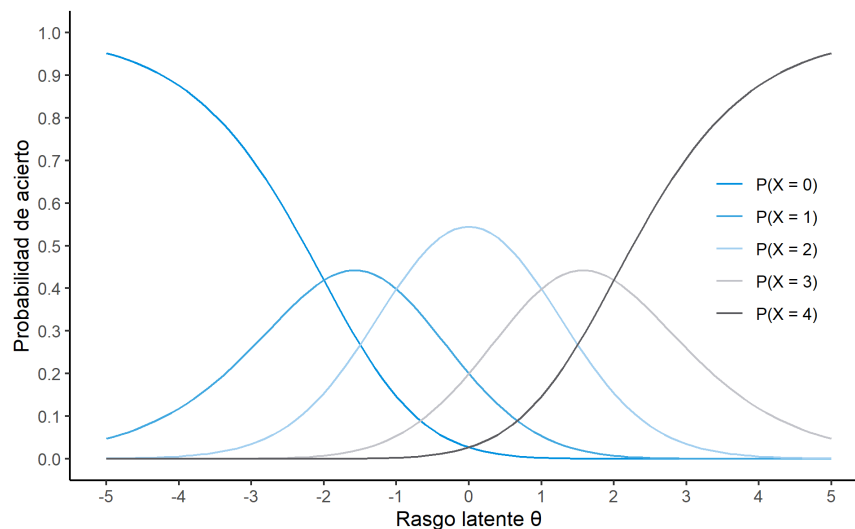


Figura C.2: Curvas Características para para cinco categorías de respuesta de un ítem hipotético con parámetro de localización definido como $\delta=0$ y parámetros de umbrales equidistantes $\tau_1 = -2,00$, $\tau_2 = -1,00$, $\tau_3 = 1,00$, $\tau_4 = 2,00$

En la figura es posible observar cómo cada categoría de respuesta tiene una función de respuesta que representa la probabilidad de que un individuo alcance dicha categoría, tal como se expresa en (2). Asimismo, la probabilidad de obtener una categoría de respuesta con mayor jerarquía incrementa conforme más nivel se tenga en el rasgo latente θ . Además, el punto de intersección entre dos categorías adyacentes corresponde al umbral τ_j (Bond et al., 2020). Es importante mencionar que existe una controversia con respecto al orden esperado de los umbrales (Adams et al., 2012). Aunque el modelo no impone ninguna restricción con respecto al ordenamiento de parámetros de umbrales, autores como Andrich y Marais (2020) sugieren esperar un orden creciente, de modo que $\tau_1 < \tau_2 < \dots < \tau_{m-1}$. En esta perspectiva, el modelo sirve para contrastar si un orden a priori de las categorías de respuesta puede ser observado a posteriori al implementar el análisis. Así, observar umbrales desordenados implicaría un funcionamiento no esperado de la escala. No obstante, Adams et al. (2012) indican que el ordenamiento de las categorías de respuesta y de los parámetros de umbrales son fenómenos distintos. Si los datos ajustan apropiadamente al modelo de Escala de Valoración, entonces las categorías estarán ordenadas. Mientras que, un desorden en los parámetros de umbrales se limita a ser un indicador de la cantidad relativa de respuestas en cada categoría.

D. Métodos de estimación de parámetros

Los parámetros del modelo Rasch son desconocidos y deben estimarse sobre la base de patrones de respuesta observadas de manera empírica. Para ello existen una amplia diversidad de métodos en la literatura que pueden ser clasificados como una estimación conjunta (i.e., tanto parámetros de ítems como de personas son estimados de manera simultánea), o dividida (i.e., primero se estiman los parámetros de ítems y luego de personas). En la Tabla D.1 se presentan los principales métodos.

Cuadro D.1: Métodos de Estimación de Parámetros en Modelos Rasch

| Categoría | Método | Descripción |
|---|--|--|
| Estimación simultánea de parámetros de ítems y personas | Máxima Verosimilitud Conjunta (JMLE) (Birnbaum, 1968) | JMLE es el método más popular para estimar parámetros del modelo Rasch; consiste en un proceso iterativo entre la calibración de ítems y medición de personas que considera a ambos parámetros como efectos fijos. |
| | Métodos Bayesianos (BAYES) | Los métodos BAYES pertenecen a la perspectiva estadística bayesiana, en donde la información previa sobre la distribución de los parámetros a estimar es utilizada en conjunto con los datos empíricos para estimar con mayor precisión a los parámetros. |
| Estimación de parámetros de ítems | Máxima Verosimilitud Condicional (CMLE) (Andersen, 1977) | CMLE fue uno de los métodos sugeridos por Rasch (1960), que consiste en utilizar los puntajes brutos de las personas como un estadístico suficiente de los parámetros de habilidad, de modo que no sean requeridos para la calibración de los ítems. |
| | Máxima Verosimilitud Marginal MMLE (Bock y Lieberman, 1970) | (MMLE) es uno de los métodos más utilizados en la estimación de parámetros, su característica principal es asumir un supuesto distribucional (usualmente una distribución normal) sobre la distribución de la variable latente, tratándola como un efecto aleatorio. De esta manera, remueve a las personas del proceso de calibración de ítems. |
| Estimación de parámetros de personas | Máxima Verosimilitud (MLE) (Birnbaum, 1968; Lord, 1980). | Utiliza un proceso iterativo para estimar parámetros de personas. La metodología es similar al JMLE, pero los parámetros de ítems son establecidos a priori. Asigna parámetros a personas de modo que se maximice la verosimilitud de los datos observados enfocando las estimaciones en la moda de la función de verosimilitud. |
| | Máxima Verosimilitud Ponderada (WLE) (Warm, 1989) | El método surge como respuesta ante el sesgo en las estimaciones del MLE; proponiendo una función de verosimilitud ponderada para corregir el sesgo enfocando las estimaciones en la media de la función. |
| | Posterior (MAP) (Samejima, 1969) | Método bayesiano con una metodología equivalente al MLE, pero incluyendo una distribución a priori de los parámetros de personas. Si la distribución es uniforme, debería generar los mismos parámetros que el MLE. El valor del parámetro de habilidad asignado a cada persona es la moda de su distribución posterior. |
| | Posterior Esperada (EAP) (Bock y Aitkin, 1981; Bock y Mislevy, 1982) | A diferencia del MAP, este estimador utilizar la media aritmética de la distribución posterior de cada persona para asignar un parámetro de habilidad. |
| | Valores Plausibles (Wu, 2005) | Los valores plausibles corresponden a una aproximación bayesiana que, a diferencia de la estimación de un solo valor como representación de la habilidad de cada persona en MAP y EAP, una serie de valores de la distribución posterior son extraídos de manera aleatoria para cada individuo. |
| | | |

La investigación ha demostrado que el método MMLE tiende a ser el más preciso y consistente para la estimación de parámetros de ítems (i.e., calibración), pues a diferencia de otras técnicas iterativas, el MMLE permite la estimación de parámetros incluso cuando existe información insuficiente para estimar a algunos individuos; por ejemplo, puntajes extremos en donde se acertó o falló a todos los ítems; además, es robusto ante valores perdidos (Linacre, 1999). En efecto, en estudios de simulación se ha demostrado que el MMLE es superior a otros métodos tradicionales como el JMLE en diversas condiciones (Drasgow, 1989; Lim y Drasgow, 1990). Esto ocurre porque el algoritmo estima los parámetros de los ítems excluyendo a los parámetros de personas de la ecuación de verosimilitud (Tinsley y Brown, 2000). En su lugar, se asume un supuesto distribucional sobre la habilidad de las personas, usualmente una distribución normal (Linacre, 1999).

Para el caso de la estimación de parámetros de personas, Wu (2005) demostró que el método de valores plausible permitía recuperar la varianza poblacional de manera más precisa y consistente. Dichos valores corresponden a múltiples estimaciones de un rasgo o habilidad latente que se consideran estadísticamente plausibles o razonables dadas las respuestas observadas a los ítems. Específicamente, corresponden a valores extraídos aleatoriamente de la distribución posterior de cada persona. Aunque en la práctica suelen extraerse cinco valores para cada individuo, en el estudio de simulación de Wu (2005), la autora demuestra que incluso contar con un solo valor plausible brinda estimaciones más precisas de la varianza poblacional en comparación con los métodos tradicionales MLE, WLE, EAP y MAP.



E. Una nota sobre las medidas Rasch

En las ciencias del comportamiento es usual emplear las sumas totales obtenidas de pruebas psicométricas como insumos para análisis estadísticos que tienden a asumir medidas continuas como input. En efecto, las puntuaciones totales son derivadas de datos que son en esencia categóricos y son empleadas en análisis posteriores dada su simplicidad de cálculo (Wright y Linacre, 1989). Los modelos Rasch reconocen la naturaleza categórica de las respuestas a los ítems y en lugar de realizar sumas directas, se realiza un modelamiento de los patrones de respuesta observados con el fin de estimar parámetros de un modelo que representen la habilidad de los individuos y la dificultad de los ítems.

Para ejemplificar cómo las medidas derivadas del modelo Rasch pueden ser tratadas como continuas, es necesario remontarse a los trabajos de Robert Duncan Luce y John Wilder Tukey en 1964. Estos autores propusieron una teoría congruente con la perspectiva de medición de las ciencias físicas que permitiría demostrar la estructura cuantitativa de los atributos psicológicos, la teoría de la Medición Conjunta (Conjoint Measurement). En general, la teoría de la Medición Conjunta se aplica cuando el ordenamiento de un atributo dependiente P cambia con el efecto conjunto de dos atributos independientes A y X (Heene, 2013; Perline et al., 1979).

La Medición Conjunta también postula axiomas que delimitan la estructura cuantitativa de los tres atributos implicados (presentados en la Tabla E.1).

Cuadro E.1: Requerimientos de la Medición Conjunta

| N | Requerimiento |
|-----|---|
| I | La variable P posee un número infinito de valores |
| II | $P = f(A, X)$ |
| III | Existe un orden simple entre los valores de P |
| IV | Los valores de A y X pueden identificarse |
| V | P , A y X son cuantitativos |
| V | f es una función no interactiva |

Nota. Adaptado de *An introduction to the logic of psychological measurement* (p. 69) de Michell (2014)).

Sin embargo, esta aproximación es compleja y una descripción completa de los axiomas puede ser consultada en la obra *Simultaneous Conjoint Measurement* (Luce y Tukey, 1964).

En este contexto, el modelo Rasch surge como una aproximación a la Medición Conjunta de Luce y Tukey (1964). En este modelo de medición se establece que la probabilidad de acertar a un ítem se encuentra determinada por la diferencia entre la habilidad de las personas y la dificultad de los ítems (Wright y Masters, 1979). Esta formulación es estructuralmente equivalente con la Medición Conjunta, pues la probabilidad de acierto representa a un atributo dependiente que es una función aditiva de la habilidad de las personas y la dificultad de los ítems, dos atributos independientes (Borsboom y Scholten, 2008).

Dada la equivalencia estructural entre ambas perspectivas, algunos autores consideran que el modelo Rasch es un caso especial de la teoría de Medición Conjunta (Borsboom y Scholten, 2008; Perline et al., 1979). Mientras que otros niegan esta posibilidad, argumentando que el modelo Rasch asume que los atributos son de naturaleza cuantitativa; mientras que, la teoría de medición conjunta especifica condiciones ordinales necesarias o suficientes para que los tres componentes sean considerados como cuantitativos (Kyngdon, 2008; Michell, 2008, 2000).

Ciertamente, el modelo Rasch no es exactamente una aplicación directa de la teoría de Medición Conjunta, Bond y Fox (2017) reconocen las limitaciones del modelo y; por ello, lo denominan como una Medición Conjunta Probabilística, en donde el carácter probabilístico hace referencia a que es imposible determinar con certeza qué sucederá durante la interacción entre una persona y un ítem, pues siempre se encuentra implicado un componente de incertidumbre (Wright y Masters, 1982). Por el contrario, la teoría de la Medición Conjunta es considerada como determinista (Michell, 2007).

A pesar de ello, diversos autores (Andrich, 1988; Fisher, 1994; Michell, 1985) han demostrado que el modelo Rasch permite obtener mediciones en ciencias sociales cuyas propiedades se asemejan a las medidas de las ciencias físicas. En otras palabras, la evidencia empírica sugiere que este modelo probabilístico permite producir estimaciones con utilidad práctica con una rigurosidad que se aproxima a los principios que rigen la medición científica (Bond y Fox, 2017).

En síntesis, el modelo Rasch se circunscribe en un panorama histórico de cuestionamientos sobre la posibilidad de realizar mediciones en psicología (Andrich y Marais, 2020). Aunque el modelo no implica una respuesta concreta y consensuada ante las limitaciones de la medición en las ciencias del comportamiento (Borsboom y Scholten, 2008), sus postulados representan una aproximación más cercana a los principios que rigen la medición en las ciencias físicas (Bond y Fox, 2017).

Finalmente, emplear medidas Rasch es que estos son estimados a través de algoritmos de máxima verosimilitud lo que deriva en propiedades estadísticas favorables como la consistencia (1), pues tras aumentar el tamaño de muestra los parámetros del modelo convergen a los valores poblacionales; eficiencia (2), pues presenta errores estándar relativamente más pequeños que otros estimadores; y normalidad del error de estimación, lo que permite derivar indicadores como los intervalos de confianza asumiendo dicha distribución (Bond et al., 2020).

