

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO



Análisis Bayesiano de modelos de clases latentes
para variables politómicas:
Confianza hacia instituciones públicas

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN
ESTADÍSTICA

Presentado por:

Marylía Paola Cruz Sarmiento

Asesor:

Dr. Luis Hilmar Valdivieso Serrano

Miembros del jurado:

Dr. Luis Hilmar Valdivieso Serrano
Dr. Victor Giancarlo Sal y Rosas Celi
Mg. Enver Gerald Tarazona Vargas

Lima, 2018

Dedicatoria

A mi familia , Yolanda Sarmiento, Julio Cruz y Carlos Cruz por su apoyo incondicional y ánimo.

A mis compañeros y profesores de la maestría quienes me motivaron constantemente en este proceso.

A mis colegas y amigos de Ciencia Política que constantemente me dieron aliento y ánimo durante mis estudios de maestría, especialmente a Jorge Aragón .



Agradecimientos

En primer lugar, agradecer al Doctor Luis Valdivieso por haberme brindado la oportunidad de recurrir a su capacidad y conocimiento científico, así como también brindarme sus orientaciones, paciencia y esfuerzo en este proceso de la tesis.

A mis profesores de la maestría quienes siempre estuvieron dispuestos a atender mis dudas.

A mis compañeros y amigos de la maestría : Francisco Calderón , Elio Peralta y Cesar Masgo.



Resumen

El modelo de análisis de clases latentes tiene como finalidad describir una variable no observable a través del agrupamiento de los individuos en base a sus patrones de respuestas. La estimación en este modelo se puede realizar mediante el algoritmo de Esperanza-Maximización (EM) y su desarrollo para el caso politómico se encuentra implementado en el paquete `poLCA` de R. Desde el punto de vista bayesiano, esta estimación ha sido hasta el momento implementada sólo para el caso de variables dicotómicas. En este trabajo, se busca extender este último aporte para el caso politómico, haciendo uso del muestrador de Gibbs.

La aplicación del modelo de análisis de clases latentes, bajo el enfoque bayesiano aquí desarrollado, se realizó sobre un conjunto de datos reales relacionados con la confianza hacia 21 instituciones públicas en una encuesta para Lima Metropolitana. En general, se identificaron tres grupos de encuestados según sus niveles de confianza institucional, los cuales se analizaron luego en relación a otras variables.

Palabras-clave: variables latentes, análisis bayesiano, variables politómicas, confianza institucional.

Índice general

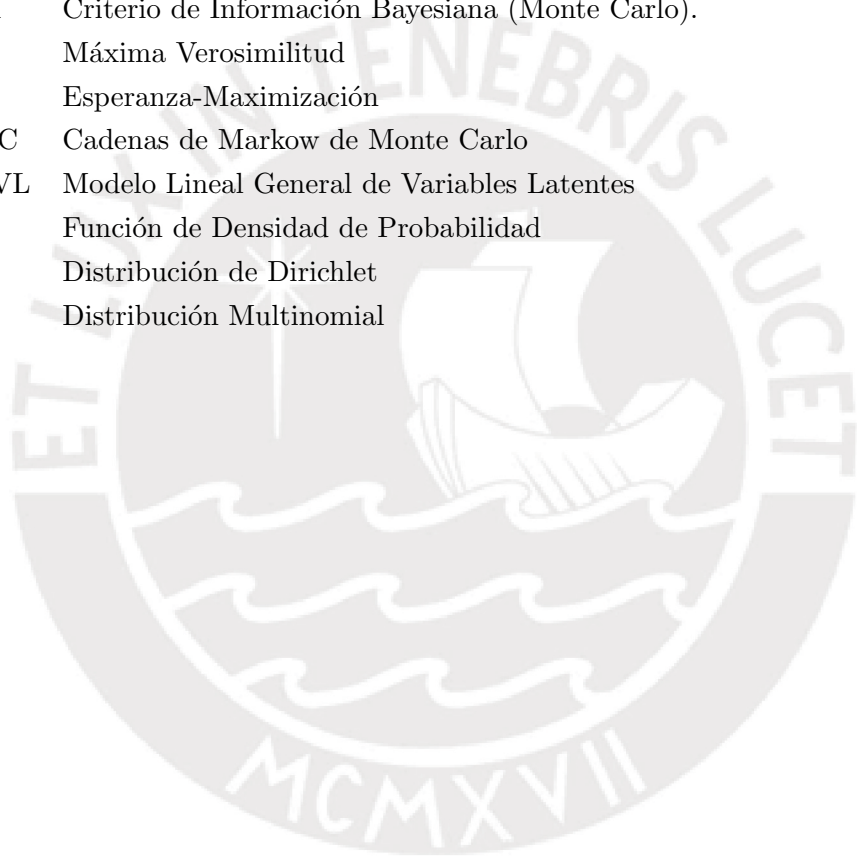
Lista de Abreviaturas	VII
Índice de cuadros	VIII
1. Introducción	1
1.1. Consideraciones Preliminares	1
1.2. Análisis de modelos de clases latentes	2
1.3. Relevancia de la aplicación	3
1.4. Objetivos	3
1.5. Organización del Trabajo	4
2. El Modelo de Clases Latentes	5
2.1. Introducción	5
2.2. Estimación por máxima verosimilitud para el modelo de clases latentes	7
3. Estimación Bayesiana del modelo de clases latentes	10
3.1. Introducción	10
3.2. Distribuciones de Dirichlet	10
3.3. Estimación bayesiana para el modelo de clases latentes	11
3.3.1. Descripción del modelo	11
3.3.2. Máxima verosimilitud aumentada del modelo	12
3.3.3. Distribuciones condicionales completas	13
3.4. El muestrador de Gibbs	15
3.5. El problema de cambio de etiquetas	16
3.6. Criterios de información	16
4. Aplicación	18
4.1. Objetivos	18
4.2. Aplicación a la base de datos <i>Alzheimer</i> : el caso dicotómico	18
4.2.1. Descripción de la base de datos	18
4.2.2. Estimación de parámetros	19
4.3. Aplicación para la confianza institucional	22
4.3.1. Instrumentos de Medición	23
4.3.2. Diseño muestral	25
4.3.3. Descripción de los datos	25
4.3.4. Caso dicotómico	26

<i>ÍNDICE GENERAL</i>	VI
4.3.5. Caso politómico	28
4.3.6. Confianza institucional en relación a otras variables	33
5. Conclusiones	35
5.1. Conclusiones	35
5.2. Sugerencias para investigaciones futuras	36
A. Código en R	37
Bibliografía	56



Lista de Abreviaturas

ACL	Análisis de Clases Latentes.
AIC	Criterio de Información de Akaike.
BIC	Criterio de Información Bayesiana.
AICM	Criterio de Información Akaike (Monte Carlo).
BICM	Criterio de Información Bayesiana (Monte Carlo).
MV	Máxima Verosimilitud
EM	Esperanza-Maximización
MCMC	Cadenas de Markow de Monte Carlo
MLGVL	Modelo Lineal General de Variables Latentes
fdp	Función de Densidad de Probabilidad
Dir	Distribución de Dirichlet
Mul	Distribución Multinomial



Índice de cuadros

1.1. Modelo de Variables Latentes	1
4.1. Distribución de frecuencias de los patrones de respuesta a los síntomas	19
4.2. Probabilidad de pertenencia a una clase (η) - Aplicación en la base de datos <i>Alzheimer</i>	19
4.3. Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase (π_{ij})- Aplicación en la base de datos <i>Alzheimer</i>	20
4.4. Comparación entre el promedio de los niveles de confianza en instituciones en Sudamérica y el promedio en Perú	24
4.5. Porcentajes de respuestas de confianza en las instituciones	26
4.6. Criterios de información para el ajuste del modelo	27
4.7. Probabilidad de pertenencia a una clase (η) - Aplicación en la base de datos <i>Confianza(versión dicotómica)</i>	27
4.8. Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase (π_{ij}) - Aplicación en la base de datos <i>Confianza(versión dicotómica)</i>	28
4.9. Los tres patrones de respuestas de mayor frecuencia	29
4.10. Criterios de información para el ajuste del modelo	29
4.11. Probabilidad de pertenencia a la clase o grupo (η) - Aplicación de la base de datos <i>Confianza (versión politómica)</i>	30
4.12. Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase (π_{ijc}) - Aplicación de la base de datos <i>Confianza (versión politómica)</i>	31

Capítulo 1

Introducción

1.1. Consideraciones Preliminares

En el ámbito de las ciencias sociales, la ciencia política y la psicología, los conceptos no son fijos; es decir, son empleados de diversas maneras de acuerdo al contexto (Gerring, 2017). En este escenario, la forma de medir un proceso se denomina operacionalización. La tarea se hace más compleja si en la operacionalización de los conceptos teóricos se identifican constructos que no son posibles de observar de forma directa. Por ejemplo, no se observa directamente la democracia, el autoritarismo, las creencias, los prejuicios, las percepciones, etc. Por esta razón, existen una amplia cantidad de indicadores imperfectos que pretenden medir estos conceptos amplios. Se han desarrollado diversas metodologías sobre la aproximación para estos fenómenos a través de un conjunto de variables medibles directamente. (McCutcheon, 1987). La covariación o relación entre las variables directamente observables darán como resultado características subyacentes denominadas variables latentes, como la manera más óptima para aproximarse al concepto teórico (Collins y S., 2010).

Existen distintas técnicas para condensar un conjunto de variables observables en una o más variables no observables, las cuales dependen de la naturaleza de estas variables. Estas se muestran en el siguiente cuadro:

		Variables Latentes	
		Continua	Categorica
Variables Manifiestas	Continua	Análisis Factorial	Modelo de Perfiles Latentes
	Categorica	Modelos de Trazo Latente	Modelos de clases latentes

Cuadro 1.1: Modelo de Variables Latentes

Cuando las variables manifiestas son categóricas; es decir, refieren a una cualidad y no a un valor numérico como el peso, la edad, el salario o el número de hijos y se desea aproximar una variable latente categórica, se emplean los modelos de clases latentes (LCA). En particular, esta técnica de clasificación no presupone una escala de medición cuantitativa como el análisis factorial, el análisis de conglomerados o el análisis discriminante (Hagenaars y Halman, 1989). Por ello, los modelos de clases latentes tienen un importante rol para crear grupos de acuerdo a valoraciones, comportamiento y actitudes de los sujetos, por ejemplo: actitudes políticas y religiosas.

1.2. Análisis de modelos de clases latentes

Los modelos de clases latentes suponen una estructura subyacente que permite establecer relaciones entre variables manifiestas. A través de los patrones de respuesta de las variables manifiestas, se definen el número de clases y las probabilidades de que una observación pertenezca a cada clase o grupo. En este sentido, la variable categórica latente es definida como aquella donde existen diferencias cualitativas entre clases o grupos conformado por observaciones, personas, objetos o elementos (Ruscio y Ruscio, 2008).

Los estudios de los modelos de clases latentes se inician con Lazarsfeld y Henry (1968), quienes sugieren la idea de una variable latente categórica con indicadores dicotómicos. Si bien en esta tesis se considera que las variables manifiestas serán en general politómicas, supongamos por simplicidad para introducir los conceptos que tenemos p variables dicotómicas y sea Y la variable latente categórica subyacente con $j = 1, 2, \dots, J$ clases. Si denotamos por $\eta_j = P(Y = j)$ a la probabilidad de que un elemento de estudio, al cual llamaremos por simplicidad en adelante un individuo, pertenezca a la clase latente j , se debe de cumplir naturalmente que

$$\sum_{j=1}^J \eta_j = 1.$$

La probabilidad de responder positivamente a una variable manifiesta i , condicionada por la pertenencia a la clase latente j , se representará en adelante por π_{ij} ; es decir, si X_i denota a la posible respuesta dada por un individuo a la variable manifiesta i ,

$$\pi_{ij} = P(X_i = 1 | Y = j).$$

Sea $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ el vector de respuestas de un individuo a las p variables manifiestas y sea $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ el valor observado o patrón de respuesta correspondiente a este individuo. La probabilidad de observar este patrón particular de respuestas condicionada a la pertenencia a una clase latente j , se obtendrá como:

$$P(\mathbf{X}=\mathbf{x}|Y = j) = \prod_{i=1}^p \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i},$$

donde se asume aquí un supuesto de independencia condicional. Para calcular luego la probabilidad de pertenencia a una clase latente j , dado un vector de respuestas \mathbf{x} , podríamos invocar al teorema de Bayes y obtener la siguiente probabilidad de clasificación a posteriori:

$$P(Y = j|\mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x}|Y = j)\eta_j}{P(\mathbf{X}=\mathbf{x})}, \quad (1.1)$$

donde

$$P(\mathbf{X}=\mathbf{x}) = \sum_{j=1}^J \eta_j P(\mathbf{X}=\mathbf{x}|Y = j)$$

o más compactamente

$$P(Y = j | \mathbf{X} = \mathbf{x}) = \frac{\eta_j \prod_{i=1}^p \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i}}{\sum_{j=1}^J \eta_j \prod_{i=1}^p \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i}}.$$

En el análisis de clases latentes será de interés estimar estas probabilidades, para lo cual requeriremos estimar las probabilidades π_{ij} y las de pertenencia a las clases η_j . Estas estimaciones las realizaremos desde el enfoque bayesiano, donde recordemos que bajo este enfoque podremos considerar distribuciones a priori para los parámetros, de tal manera que se reduzcan el nivel de incertidumbre en la agrupación.

En la actualidad, se ha incrementado el uso del enfoque bayesiano en diferentes ciencias de estudio. Una de las principales razones a este incremento es la consideración de lo subjetivo al considerar la información previa de los parámetros. Además, en muchos casos las estimaciones resultan ser más precisas, sobretodo para tamaños de muestras no muy grandes (Gill, 2014).

1.3. Relevancia de la aplicación

El análisis de los modelos de clases latentes ha sido ampliamente aplicado en distintas áreas de investigación como por ejemplo en: psicología con el trabajo de (Chung et al., 2006) sobre las actitudes de adolescentes a las drogas; en marketing con la identificación de tipos de consumidores (Bhatnagar y Ghose, 2004); en ciencias sociales con el análisis de la influencia de roles de género en el apoyo al fundamentalismo islámico (Blaydes y Linzer, 2008); en ciencias políticas con los trabajos sobre la tolerancia en la sociedad americana (McCutcheon, 1985) y el apoyo a partidos extremistas (Breen, 2000).

Luego del retorno del Perú a la democracia en el año 2001, ha sido de interés el estudio de los factores asociados al apoyo ciudadano del régimen democrático. De esta manera, el crecimiento económico, la cobertura de la capacidad del Estado al brindar servicios básicos, el desempeño de las autoridades, los niveles de corrupción y delincuencia, y la confianza hacia las instituciones son factores que influyen en el sostenimiento del sistema político.

En este trabajo nos abocaremos a estudiar la confianza institucional entendida como la relación entre ciudadanos e instituciones. Nos interesa analizar el modo en que los ciudadanos confían en las instituciones públicas, y cómo este se relaciona con la percepción sobre corrupción y el apoyo a la democracia. Consideramos que la confianza hacia las instituciones no es uniforme, y por tanto, una técnica de agrupamiento como el modelo de clases latentes puede brindar una aproximación que tome en cuenta las distintas valoraciones otorgada a las instituciones.

1.4. Objetivos

El objetivo general de esta tesis es estudiar la estimación de los modelos de clases latentes con variables manifiestas politómicas desde la aproximación bayesiana empleando el algoritmo de Gibbs. Se busca realizar la implementación computacional de este algoritmo y aplicarlo a un conjunto de datos reales.

Los objetivos específicos que orientan la presente son los siguientes:

- Revisar la literatura acerca de los modelos de clases latentes. Presentar las formulaciones de los modelos de clases latentes tanto desde la aproximación clásica como bayesiana.
- Estudiar las propiedades y la implementación bayesiana para los modelos de clases latentes, en específico para el caso de variables manifiestas politómicas.
- Comparar la estimación de los modelos de clases latentes por el método bayesiano y el método clásico.
- Aplicar el modelo de clases latentes a conjuntos de datos reales sobre Confianza Institucional en Lima Metropolitana.

1.5. Organización del Trabajo

En el Capítulo 2, se presenta una introducción al modelo de clases latentes en variables manifiestas politómicas y analizamos su estimación con el método de máxima verosimilitud. En el Capítulo 3, se estudiará la estimación del modelo de clases latentes con el enfoque bayesiano, haciendo uso del muestreador de Gibbs. Para ello se derivan las distribuciones condicionales completas necesarias para la implementación del algoritmo. En el Capítulo 4, se mostrará la aplicación del modelo de clases latentes a un conjunto de datos reales relacionados a la confianza hacia instituciones públicas en Lima Metropolitana. Además, se mostraré la relación entre los grupos de los individuos según sus niveles de confianza institucional y las percepciones de corrupción y el apoyo a la democracia. Finalmente, en el Capítulo 5 discutimos algunas conclusiones obtenidas en este trabajo. Analizamos la ventajas y desventajas de los métodos propuestos y futuras extensiones de este trabajo.

En el anexo presentamos el código de programación desarrollado en R para la estimación bayesiana del modelo de clases latentes en variables manifiestas politómicas.

Capítulo 2

El Modelo de Clases Latentes

2.1. Introducción

El estudio de constructos teóricos a las que denominaremos variables latentes se realiza a través de un conjunto de variables observables denominadas variables manifiestas. Entre las definiciones de variable latente encontramos las intuitivas que suelen considerar a esta como una variable hipotética e irreal; es decir, como constructos elaborados por los investigadores para analizar fenómenos reales. Otra definición sobre variable latente señala a esta variable como inobservable e imposible de medir directamente. En contraste, también se considera a la variable latente como un resumen de variables observables que tiene el propósito de la reducción de dimensionalidad, al igual de lo que realiza el Análisis Factorial. Esta definición presupone una variable subyacente que permite explicar la asociación entre dos o más variables observadas (Bollen, 2002).

El modelo de clases latentes (ACL) considera que las variables manifiestas y latentes son categóricas. El hecho que consideremos a la variable latente como categórica puede ser producto de una evidencia a priori o teórica, o simplemente tomada así por una cuestión práctica. Las nociones preliminares sobre el modelo de clases latentes se desarrollaron con Lazarsfeld (1950), Green (1951) y Lazarsfeld y Henry (1968). Más adelante, con el trabajo de Goodman (1974) se desarrolla el modelo estadístico con la incorporación del método de Máxima Verosimilitud para la estimación de sus parámetros. Los trabajos de McCutcheon (1987), Hagenaars (1993), Clogg (1995), contribuyeron a ampliar el desarrollo del estudio del modelo de clases latentes.

Para los modelos de clases latentes se tiene un conjunto de individuos, los cuales deberán de responder a un conjunto de preguntas con respuestas categóricas, a las cuales llamaremos las variables manifiestas. La idea detrás del modelo de clases latentes subyace en la posibilidad de encontrar grupos de individuos, a los que llamaremos clases latentes, en base a sus patrones de respuestas a las variables manifiestas. El modelo de clases latentes se basa en los dos supuestos siguientes:

1. Dentro de cada clase, todos los individuos tienen las mismas probabilidades de respuesta a las variables manifiestas.
2. Se cumple la independencia condicional; es decir, las respuestas entre los individuos son independientes dado que estos pertenecen a una misma clase.

Para formalizar estas ideas, sea $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ el vector columna p -dimensional de variables manifiestas politómicas asociadas a las respuestas de los individuos, Y una variable latente que presenta J categorías o clases latentes, las cuales denotaremos con el índice $j = 1, 2, \dots, J$. Consideremos, por otro lado, a π_{ijc} como la probabilidad condicional de que un individuo cualesquiera de una respuesta c a la variables manifiesta X_i dado que este pertenezca a la clase j ; es decir,

$$\pi_{ij}(c) = P(X_i(c) = 1 | Y = j)$$

donde

$$X_i(c) = \begin{cases} 1, & \text{si la respuesta de un individuo a la variable manifiesta } i \text{ está en la categoría } c, \\ 0, & \text{en caso contrario} \end{cases}$$

Sea, de otro lado, η_j la probabilidad que un individuo pertenezca a la clase latente j . A lo largo de este trabajo asumiremos que X_i es una variable aleatoria politómica con C_i posibles respuestas excluyentes y exhaustivas, las cuales estamos denotando con el índice $c = 1, 2, \dots, C_i$. En tal sentido se debe de cumplir que $\sum_{c=1}^{C_i} X_i(c) = 1$.

Si identificamos la respuesta de un individuo a la variable manifiesta i con el vector columna $\mathbf{X}_i = [X_i(0), \dots, X_i(C_i)]'$, el vector de respuestas de un individuo se podrá representar en el vector columna $\mathbf{X} = [\mathbf{X}'_1, \mathbf{X}'_2, \dots, \mathbf{X}'_p]'$. La función de probabilidad conjunta de este vector o patrón de respuesta \mathbf{X} vendrá dada por:

$$P(\mathbf{X} = \mathbf{x}) = \sum_{j=1}^J P(\mathbf{X} = \mathbf{x} | Y = j) P(Y = j) = \sum_{j=1}^J P(\mathbf{X} = \mathbf{x} | Y = j) \eta_j, \quad (2.1)$$

donde

$$P(\mathbf{X} = \mathbf{x} | Y = j) = \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ij}(c)^{x_i(c)} \quad (2.2)$$

y $x_i(c)$ es el valor observado de la variable $X_i(c)$.

En resumen, esto nos conduce a la siguiente función de verosimilitud individual:

$$P(\mathbf{X} = \mathbf{x}) = \sum_{j=1}^J \eta_j \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ij}(c)^{x_i(c)}. \quad (2.3)$$

Más aún, por el teorema de Bayes, la distribución a posteriori de que un individuo pertenezca a la clase j , observado su patrón de respuesta anterior viene dado por:

$$P(Y = j | \mathbf{X} = \mathbf{x}) = \frac{\eta_j \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ij}(c)^{x_i(c)}}{\sum_{j=1}^J \eta_j \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ij}(c)^{x_i(c)}}. \quad (2.4)$$

Seguidamente, desarrollaremos la estimación por máxima verosimilitud de este modelo. En el capítulo 3 siguiente nos centraremos en la metodología bayesiana.

2.2. Estimación por máxima verosimilitud para el modelo de clases latentes

En la estimación clásica, dado que se desconoce la distribución de la variable latente, se sugiere emplear el algoritmo de Esperanza Maximización (Dempster et al., 1977). Este es un procedimiento iterativo que busca maximizar la función de verosimilitud del modelo aún cuando esta sea una función numéricamente difícil de optimizar. El algoritmo de Esperanza Maximización se divide en el paso de *esperado* referente al cálculo de las probabilidades de las clases o los valores esperados de las clases y la *maximización*, que busca maximizar monótonamente la verosimilitud según los datos.

Desde el aspecto computacional, este algoritmo se ha desarrollado en el paquete *poLCA* del software R, incluso cuando existan algunos datos perdidos para las variables manifiestas (Drew y Lewis, 2011).

El objetivo de este procedimiento consiste en maximizar la función de verosimilitud, la cual está definida por:

$$\prod_{n=1}^N P(\mathbf{X}_n = \mathbf{x}_n) = \prod_{n=1}^N \sum_{j=1}^J \eta_j \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ij}(c)^{x_{ni}(c)}$$

donde estamos suponiendo de que existen n individuos en la muestra de estudio y $x_{ni}(c)$ denota al valor observado de la variable $X_i(c)$ para el n -ésimo individuo.

La correspondiente función de log-verosimilitud viene dada por

$$\mathcal{L} = \sum_{n=1}^N \log P(\mathbf{X}_n = \mathbf{x}_n) = \sum_{n=1}^N \log \left(\sum_{j=1}^J \eta_j P(\mathbf{X}_n = \mathbf{x}_n | Y_n = j) \right), \quad (2.5)$$

donde

$$P(\mathbf{X}_n = \mathbf{x}_n | Y_n = j) = \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ij}(c)^{x_{ni}(c)}$$

Dado que buscamos maximizar esta función bajo las siguientes restricciones:

$$\sum_{j=1}^J \eta_j = 1$$

y

$$\sum_{c=1}^{C_i} \pi_{ij}(c) = 1, \quad ,$$

para $i = 1, 2, \dots, p$ y $c_i = 1, 2, \dots, C_i$, el problema equivale a maximizar:

$$L_g = \mathcal{L} + \lambda \left(\sum_{j=1}^J \eta_j - 1 \right) + \sum_{j=1}^J \sum_{i=1}^p \phi_{ij} \left(\sum_{c=1}^{C_i} \pi_{ij}(c) - 1 \right) \quad , \quad (2.6)$$

donde λ y ϕ_{ij} son multiplicadores de Lagrange indeterminados.

Si denotamos $g(\mathbf{x}_n | j) = P(\mathbf{X}_n = \mathbf{x}_n | Y_n = j)$ y por $h(j | \mathbf{x}_n) = P(Y_n = j | \mathbf{X}_n = \mathbf{x}_n)$ a la expresión descrita en (2.4) para el sujeto n , las derivadas parciales con respecto a η_j vienen dadas por:

$$\frac{\partial L_g}{\partial \eta_j} = \sum_{n=1}^N \frac{g(\mathbf{x}_n | j)}{f(\mathbf{x}_n)} + \lambda = \sum_{n=1}^N \frac{h(j | \mathbf{x}_n)}{\eta_j} + \lambda \quad (2.7)$$

Igualando a 0 esta expresión y utilizando la restricción para los η_j , obtendremos que:

$$\hat{\eta}_j = \frac{1}{N} \sum_{n=1}^N h(j | \mathbf{x}_n) \quad (2.8)$$

Similarmente de

$$\frac{\partial \mathcal{L}}{\partial \pi_{ij}(c)} = \sum_{n=1}^N \frac{1}{f(\mathbf{x}_n)} \hat{\eta}_j \frac{\partial g(\mathbf{x}_n | j)}{\partial \pi_{ij}(c)} \quad (2.9)$$

y

$$\frac{\partial g(\mathbf{x}_n | j)}{\partial \pi_{ij}(c)} = \frac{\partial}{\partial \pi_{ij}(c)} \exp \left(\sum_{i=1}^p \sum_{c=1}^{C_i} x_{ni}(c) \log(\pi_{ij}(c)) \right) = \frac{g(\mathbf{x}_n | j) x_{ni}(c)}{\pi_{ij}(c)} \quad (2.10)$$

se tiene que

$$\frac{\partial L_g}{\partial \pi_{ij}(c)} = \eta_j \sum_{n=1}^N \frac{g(\mathbf{x}_n | j) x_{ni}(c)}{f(\mathbf{x}_n) \pi_{ij}(c)} + \phi_{ij} = \sum_{n=1}^N \frac{h(j | \mathbf{x}_n) x_{ni}(c)}{\pi_{ij}(c)} + \phi_{ij} \quad (2.11)$$

Igualando a 0

$$0 = \sum_{n=1}^N h(j | \mathbf{x}_n) x_{ni}(c) + \pi_{ij}(c) \phi_{ij} \quad (2.12)$$

y sumando ambos lados sobre c obtendremos que

$$0 = \sum_{n=1}^N h(j | \mathbf{x}_n) + \phi_{ij}$$

y

$$\hat{\phi}_{ij} = - \sum_{n=1}^N h(j | \mathbf{x}_n)$$

Sustituyendo en (2.12) se obtienen entonces las ecuaciones de estimación.

$$\hat{\pi}_{ij}(c) = \frac{\sum_{n=1}^N h(j | \mathbf{x}_n) x_{ni}(c)}{\sum_{n=1}^N h(j | \mathbf{x}_n)} = \frac{\sum_{n=1}^N h(j | \mathbf{x}_n) x_{ni}(c)}{N \hat{\eta}_j} \quad (2.13)$$

En el caso que las variables manifiestas sean politómicas ordenadas, las probabilidades también pueden escribirse como (Wiener, 2015) :

$$\hat{\Pi}_{ij}(c) = \hat{\Pi}_{ij}(c+1) + \frac{\sum_{n=1}^N x_{ni}(c) h(j | \mathbf{x}_n)}{N \hat{\eta}_j} \quad (2.14)$$

donde $\hat{\Pi}_{ij}(c) = \sum \pi_{ij}(c)$, $i = 1, 2, \dots, p$; $j = 1, \dots, J$ y $c = 1, \dots, C_i$

Así en resumen el algoritmo EM se define por:

1. Elegir una distribución inicial para las probabilidades a posteriori $h(j|\mathbf{x}_n)$.
2. Usar (2.8) y (2.13) ó (2.14) para obtener una primera aproximación de los estimadores de máxima verosimilitud para los $\hat{\eta}_j$ y $\hat{\pi}_{ij}$.
3. Substituir las estimaciones anteriores en (2.4) para mejorar la distribución de las probabilidades a posteriori $h(j|\mathbf{x}_n)$.
4. Retomar el paso 2 para obtener una segunda aproximación de los parámetros y continuar el ciclo hasta alcanzar la convergencia.

Capítulo 3

Estimación Bayesiana del modelo de clases latentes

3.1. Introducción

En el capítulo previo se ha revisado la estimación clásica del modelo de clases latentes, en la que se asume que los parámetros son desconocidos pero fijos. En el enfoque bayesiano; se considera que los parámetros son aleatorios y se incorpora una distribución a priori para ellos. La estimación bayesiana realiza la inferencia de los parámetros del modelo a través de la distribución a posteriori calculada a partir de la distribución a priori y la información de máxima verosimilitud. A diferencia del enfoque clásico, el bayesiano supone que se tiene alguna información previa sobre los parámetros, que se denomina distribución a priori.

A continuación se va a realizar una breve introducción de la distribución Dirichlet, la cual se empleará más adelante como distribución a priori en la estimación bayesiana para los modelos de clases latentes.

3.2. Distribuciones de Dirichlet

La distribución de Dirichlet, nombrada en honor al matemático alemán Johann Peter Gustav Lejeune Dirichlet y denotada como $Dir(\boldsymbol{\alpha})$, proviene de la familia de distribuciones de probabilidad continuas multivariadas, parametrizadas por un vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K)$ cuyos componentes pertenecen al conjunto de los números reales positivos.

La distribución de Dirichlet se puede concebir como una generalización multivariada de la distribución beta y se expresa a través de la siguiente función de probabilidad conjunta:

$$f(x_1, x_2, \dots, x_K; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K x_k^{\alpha_k - 1},$$

donde el soporte de esta distribución recae sobre el simplejo de orden $K - 1$:

$$\Delta^{K-1} = \left\{ (x_1, x_2, \dots, x_K) \in [0, 1]^K / \sum_{k=1}^K x_k = 1 \right\}$$

Note que en el caso particular de $K = 2$ se tiene la conocida distribución Beta.

La distribución de Dirichlet se utiliza frecuentemente como una distribución a priori en estadística bayesiana, y de hecho ella es la conjugada a priori de una distribución categórica. En adelante, utilizaremos la notación $\mathbf{X} = (X_1, \dots, X_K) \sim Dir(\boldsymbol{\alpha})$ para referirnos a una distribución Dirichlet con vector de parámetros $\boldsymbol{\alpha}$. En ella notemos que la última componente de este vector aleatorio satisface $X_K = 1 - \sum_{k=1}^{K-1} X_k$. Si se define $\alpha_0 = \sum_{k=1}^K \alpha_k$, no es difícil

probar que las distribuciones marginales de este vector satisfacen (Balakrishnan, 2003):

$$E(X_k) = \frac{\alpha_k}{\alpha_0}$$

$$Var(X_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$$

$$Cov(X_k, X_j) = \frac{-\alpha_k\alpha_j}{\alpha_0^2(\alpha_0 + 1)}.$$

3.3. Estimación bayesiana para el modelo de clases latentes

3.3.1. Descripción del modelo

Si tenemos p variables manifiestas politómicas y N individuos, denotemos por

$$\mathbf{X}_n = (X_{n1}, X_{n2}, \dots, X_{np})$$

al vector de respuestas que el individuo n da a las p variables manifiestas, donde cada variable X_{ni} del individuo n podría tomar cualquiera de las C_i categorías, con que cuenta la i -ésima variable manifiesta. El modelo asume que los N individuos están agrupados en J clases latentes disjuntas que no podemos observar. En este sentido, resulta conveniente introducir como antes una variable latente Y_{nj} que valdrá 1 si el individuo n pertenece a la clase j y 0 en caso contrario.

La distribución de probabilidad de esta variable aleatoria vendrá dada por

$$\eta_j = P(Y_{nj} = 1), \quad j = 1, 2, \dots, J$$

mientras que la probabilidad de que un individuo n muestre un patrón de respuestas

$$\mathbf{x}_n = (x_{n1}, x_{n2}, \dots, x_{np})$$

viene dada por:

$$P(\mathbf{X}_n = \mathbf{x}_n) = \sum_{j=1}^J P(\mathbf{X}_n = \mathbf{x}_n | Y_{nj} = 1)P(Y_{nj} = 1)$$

$$= \sum_{j=1}^J P(\mathbf{X}_n = \mathbf{x}_n | Y_{nj} = 1)\eta_j, \quad (3.1)$$

donde

$$P(\mathbf{X}_n = \mathbf{x}_n | Y_{nj} = 1) = \prod_{i=1}^p P(X_{ni} = x_{ni} | Y_{nj} = 1) = \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)}, \quad (3.2)$$

siendo π_{ijc} la probabilidad condicional de que un individuo cualesquiera de a la variable

manifiesta i una respuesta c sabiéndose de que él pertenece a la clase latente j y

$$x_{ni}(c) = \begin{cases} 1 & , \text{si la respuesta del sujeto } n \text{ al ítem } i \text{ presenta la categoría } c \\ 0 & , \text{en caso contrario} \end{cases}$$

es una variable indicadora.

Note que estamos aquí asumiendo, como lo hemos hecho en todo el trabajo, de que las respuestas de los individuos a las p variables manifiestas son condicionalmente independientes al interior de cada clase j .

3.3.2. Máxima verosimilitud aumentada del modelo

La verosimilitud aumentada de este modelo para el individuo n , en base a su respuesta $\mathbf{x}_n = (\mathbf{x}_{n1}, \mathbf{x}_{n2}, \dots, \mathbf{x}_{np})$ y su clasificación de pertenencia a alguna de las clases latentes la denotaremos por:

$$P(\mathbf{X}_n = \mathbf{x}_n, \mathbf{Y}_n = \mathbf{y}_n \mid \boldsymbol{\pi}, \boldsymbol{\eta}),$$

donde $\mathbf{Y}_n = (Y_{n1}, Y_{n2}, \dots, Y_{nJ})$, $\mathbf{y}_n = (y_{n1}, y_{n2}, \dots, y_{nJ})$, $\boldsymbol{\pi}$ es un arreglo de todas las probabilidades π_{ijc} y $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_J)$ es la distribución de probabilidades de pertenecer a alguna clase latente. Esta verosimilitud aumentada para el individuo n viene dada explícitamente por:

$$\begin{aligned} P(\mathbf{X}_n = \mathbf{x}_n, \mathbf{Y}_n = \mathbf{y}_n \mid \boldsymbol{\pi}, \boldsymbol{\eta}) &= \prod_{j=1}^J (P(\mathbf{X}_n = \mathbf{x}_n \mid Y_{nj} = 1) \eta_j)^{y_{nj}} \\ &= \prod_{j=1}^J \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)} \eta_j \right)^{y_{nj}} \end{aligned} \quad (3.3)$$

La función de verosimilitud aumentada para toda la muestra se define por:

$$L(\boldsymbol{\pi}, \boldsymbol{\eta}) = \prod_{n=1}^N P(\mathbf{X}_n = \mathbf{x}_n, \mathbf{Y}_n = \mathbf{y}_n \mid \boldsymbol{\pi}, \boldsymbol{\eta}) \quad (3.4)$$

Puesto que estamos trabajando aquí desde un enfoque bayesiano, nuestro interés se centrará en la distribución a posteriori

$$P(\boldsymbol{\pi}, \boldsymbol{\eta} \mid \mathbf{x}, \mathbf{y}) \propto L(\boldsymbol{\pi}, \boldsymbol{\eta}) P(\boldsymbol{\pi}, \boldsymbol{\eta}), \quad (3.5)$$

donde estamos denotando por \mathbf{x} e \mathbf{y} al arreglo total de las respuestas e indicadores de clase para los individuos. Para esto debemos especificar la distribución a priori para $\boldsymbol{\pi}$ y $\boldsymbol{\eta}$ en el lado derecho de (3.5). Nosotros asumiremos, como es usual en la literatura, prioris independientes para $\boldsymbol{\pi}$ y $\boldsymbol{\eta}$. De acuerdo a lo sugerido por [White y Murphy \(2014\)](#), utilizaremos distribuciones de Dirichlet a priori tanto para $\boldsymbol{\eta}$ como para los vectores $\boldsymbol{\pi}_{ij} = (\pi_{ij1}, \pi_{ij2}, \dots, \pi_{ijC_i})$.

Más concretamente estas tendrán la forma:

$$P(\boldsymbol{\eta}) \propto \prod_{j=1}^J \eta_j^{\delta_j - 1}$$

$$P(\boldsymbol{\pi}_{ji}) \propto \prod_{c=1}^{C_i} \pi_{ijc}^{\alpha_{ji} - 1},$$

donde

$$\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_J) \in \mathbb{R}_+^J$$

y $\alpha_{ji} \in \mathbb{R}_+$ son todos hiperparámetros del modelo; mientras que el símbolo \propto denota a que las expresiones del lado derecho en las distribuciones a priori anteriores son iguales a los de la izquierda, salvo por una constante de proporcionalidad que no dependen de los argumentos de las funciones a izquierda.

Para la estimación del caso binario se ha desarrollado en R el paquete *BayesLCA*. Este paquete permite agrupar las observaciones, obtener la distribución a posteriori, graficar el comportamiento de los parámetros y obtener criterios de selección. Contiene para la estimación tanto el algoritmo EM, el muestreador de Gibbs y una aproximación variacional de Bayes (White y Murphy, 2014). El paquete *BayesLCA*; sin embargo, sólo considera variables manifiestas dicotómicas, por lo que será objetivo de nuestro estudio extender este al caso politómico.

3.3.3. Distribuciones condicionales completas

A fin de estimar los parámetros del modelo de clases latentes desde la perspectiva bayesiana, haremos uso en este trabajo del algoritmo de Gibbs. Este requiere una forma concreta de las distribuciones condicionales completas, las cuales derivamos a continuación.

La distribución conjunta a posteriori de $\boldsymbol{\pi}$ y $\boldsymbol{\eta}$ viene dada, según (3.4), por:

$$\begin{aligned} P(\boldsymbol{\pi}, \boldsymbol{\eta} \mid \mathbf{x}, \mathbf{y}) &\propto L(\boldsymbol{\pi}, \boldsymbol{\eta})P(\boldsymbol{\eta})P(\boldsymbol{\pi}) \\ &= \prod_{n=1}^N \prod_{j=1}^J \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)} \eta_j \right)^{y_{nj}} \prod_{j=1}^J \eta_j^{\delta_j - 1} \prod_{j=1}^J \prod_{i=1}^p P(\boldsymbol{\pi}_{ij}) \\ &= \prod_{n=1}^N \prod_{j=1}^J \eta_j^{y_{nj}} \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)} y_{nj} \right) \prod_{j=1}^J \eta_j^{\delta_j - 1} \prod_{j=1}^J \prod_{i=1}^p P(\boldsymbol{\pi}_{ij}) \quad , \end{aligned} \quad (3.6)$$

donde recordemos que los y_{nj} son indicadores dicotómicos que toman el valor de 1 cuando el individuo n pertenece a la clase latente j y valen 0 en caso contrario.

Las distribución condicional para $\boldsymbol{\pi}$ viene definida por:

$$\begin{aligned}
 P(\boldsymbol{\pi} \mid \mathbf{x}, \mathbf{y}, \boldsymbol{\eta}) &\propto P(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\pi}, \boldsymbol{\eta})P(\boldsymbol{\pi}) \\
 &= \prod_{n=1}^N \prod_{j=1}^J \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)} \eta_j \right)^{y_{nj}} \prod_{j=1}^J \prod_{i=1}^p P(\boldsymbol{\pi}_{ij}) \\
 &= \prod_{n=1}^N \prod_{j=1}^J \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)} \eta_j \right)^{y_{nj}} \prod_{j=1}^J \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{\alpha_{ji}-1} \\
 &= \prod_{j=1}^J \eta_j^{m_j} \prod_{n=1}^N \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)y_{nj}} \right) \prod_{j=1}^J \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{\alpha_{ji}-1} \\
 &= \prod_{j=1}^J \eta_j^{m_j} \prod_{n=1}^N \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)y_{nj}} \prod_{j=1}^J \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{\alpha_{ji}-1} \\
 &= \prod_{j=1}^J \prod_{i=1}^p \prod_{c=1}^{C_i} \eta_j^{m_j} \prod_{c=1}^{C_i} \pi_{ijc}^{\sum_{n=1}^N x_{ni}(c)y_{nj} + \alpha_{ji}-1},
 \end{aligned} \tag{3.7}$$

donde $m_j = \sum_{n=1}^N y_{nj}$ denota el número de individuos que pertenecen a la clase j . Así, la distribución condicional completa para $\boldsymbol{\pi}_{ij}$ vendrá dada por:

$$\boldsymbol{\pi}_{ji} \sim Dir(\mathbf{a})$$

donde $\mathbf{a} = (a_1, a_2, \dots, a_{C_i})$ y $a_c = \sum_{n=1}^N x_{ni}(c)y_{nj} + \alpha_{ji}$.

De otro lado, para el vector de parámetros $\boldsymbol{\eta}$ se tiene que:

$$\begin{aligned}
 P(\boldsymbol{\eta} \mid \mathbf{x}, \mathbf{y}, \boldsymbol{\pi}) &\propto P(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\pi}, \boldsymbol{\eta})P(\boldsymbol{\eta}) \\
 &= \prod_{n=1}^N \prod_{j=1}^J \eta_j^{y_{nj}} \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)y_{nj}} \right) \prod_{j=1}^J \eta_j^{\delta_j-1} \\
 &= \prod_{j=1}^J \eta_j^{m_j + \delta_j - 1} \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)y_{nj}} \right)
 \end{aligned} \tag{3.8}$$

Por tanto, la distribución condicional completa para $\boldsymbol{\eta}$ vendrá dada por:

$$\boldsymbol{\eta} \sim Dir(\mathbf{m} + \boldsymbol{\delta}), \tag{3.9}$$

donde $\mathbf{m} = (m_1, m_2, \dots, m_J)$.

Finalmente para \mathbf{Y}_n se cumple que:

$$P(\mathbf{y}_n \mid \mathbf{x}, \boldsymbol{\pi}, \boldsymbol{\eta}) \propto P_n(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\pi}, \boldsymbol{\eta}) = \prod_{j=1}^J \left(\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)} \eta_j \right)^{y_{nj}}$$

Así la distribución condicional completa para el vector de pertenencia del sujeto n a las clases

latentes vendrá dada por la distribución multinomial:

$$\mathbf{Y}_n \sim \text{Mul}(1, q_1, q_2, \dots, q_J),$$

donde

$$q_j = \frac{\prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)} \eta_j}{\sum_{j=1}^J \prod_{i=1}^p \prod_{c=1}^{C_i} \pi_{ijc}^{x_{ni}(c)} \eta_j}.$$

3.4. El muestrador de Gibbs

El muestrador de Gibbs fue nombrado en honor al físico Josiah Willard Gibbs y su algoritmo fue descrito por los hermanos Stuart y Donald German ([German y German, 1984](#)). Este es un método interactivo por Cadenas de Markov de Monte Carlo (MCMC). El muestreo se realiza a partir de las distribuciones condicionales completas de cada parámetro, las cuales se asumen que deben ser conocidas. Este proceso iterativo genera de cada distribución condicional muestras de la distribución a posteriori conjunta.

[White et al. \(2016\)](#) ha explorado la estimación bayesiana de clases latentes para el caso politómico empleando un enfoque marginal que aborda la incertidumbre en la determinación del número de clases J y la selección de variables. El autor emplea un vector que indica a las variables manifiestas usadas para la clasificación y a otro conjunto que no. Nosotros optaremos por un enfoque más clásico fijando los valores de J y analizando el ajuste de estos modelos por métodos de información Bayesiana al hacer variar los valores de J . El muestrador de Gibbs se basa en la suposición de Markov, de que las muestras en la iteración $k + 1$ dependen sólo de los valores de los parámetros muestreados durante la iteración anterior k . Empezando con un valor inicial $\boldsymbol{\pi}^0$, el algoritmo genera iterativamente para cada individuo n una estimación en el paso k mediante la simulación de valores de las distribuciones condicionales completas:

$$\begin{aligned} \boldsymbol{\pi}^{k+1} &\sim P(\boldsymbol{\pi} | \mathbf{x}_n, \mathbf{y}_n^k, \boldsymbol{\eta}^k) \\ \boldsymbol{\eta}^{k+1} &\sim P(\boldsymbol{\eta} | \mathbf{x}_n, \mathbf{y}_n^k, \boldsymbol{\pi}^{k+1}) \\ \mathbf{y}_n^{k+1} &\sim P(\mathbf{y}_n | \mathbf{x}_n, \boldsymbol{\eta}^{k+1}, \boldsymbol{\pi}^{k+1}) \end{aligned} \quad (3.10)$$

donde $\boldsymbol{\pi}^k = [\pi_{ijc}^k]$, $\boldsymbol{\eta}^k = (\eta_1^k, \eta_2^k, \dots, \eta_J^k)$ e $\mathbf{y}_n^k = (y_{n1}^k, \dots, y_{nJ}^k)$ son los arreglos de los parámetros de interés para la k -ésima interacción, entonces los parámetros simulados para la siguiente iteración se obtendrán de :

$$\begin{aligned} \boldsymbol{\eta}^{k+1} &\sim \text{Dir}(\mathbf{m}^k + \boldsymbol{\delta}) \\ \pi_{ijc}^{k+1} &\sim \text{Dir}\left(\sum_{n=1}^N x_{ni}(c) \mathbf{y}_{nj}^k + \alpha_{ji}\right) \\ \mathbf{y}_n^{k+1} &\sim \text{Mul}(1, q_1, q_2, \dots, q_J) \end{aligned} \quad (3.11)$$

donde $\mathbf{m}^k = (m_1^k, m_2^k, \dots, m_J^k) = (\sum_{n=1}^N y_{n1}^k, \sum_{n=1}^N y_{n2}^k, \dots, \sum_{n=1}^N y_{nJ}^k)$ denota a la distribución de pertenencia de los individuos a las J clases en la k -ésima iteración.

3.5. El problema de cambio de etiquetas

Tal como se señaló en la sección anterior, el muestrador de Gibbs permite la simulación de una cadena de Markov desde la distribución a posteriori conjunta de los parámetros del modelo y de clases latentes. Sin embargo, la estimación de la probabilidad es invariante al orden de las clases latentes, fenómeno conocido como cambio de etiqueta (labels switching) y genera problemas en la estimación (Papastamoulis, 2015).

Para solucionar este problema se utilizará en la implementación de la estimación del modelo la función *matchClasses*, la cual está implementada en el paquete *BayesLCA* y cuyo código se muestra en el apéndice. La idea de esta función es generar una matriz cuadrada en cada transición de la cadena donde las filas y las columnas representan la pertenencia de cada patrón de respuesta a una clase en cada iteración. La función realiza un ordenamiento (permutación) de esta matriz para lograr que los valores de la diagonal principal sumen el máximo valor, el cual aproximará al número total de patrones de respuesta, y será elegido como corrector del *labelswitching* en esa iteración.

3.6. Criterios de información

Para encontrar el modelo estadístico más parsimonioso, se necesitan indicadores que midan la capacidad explicativa del modelo y penalicen su grado de complejidad. Si bien la log-verosimilitud se utiliza como criterio de bondad de ajuste, para el enfoque clásico, el ajuste se suele analizar a través del Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC), los cuales incorporan términos adicionales que toman en cuenta el número de parámetros y la cantidad de datos. Por otro lado, bajo el enfoque bayesiano, se trabaja con la distribución a posteriori de los modelos obtenidos empleándose usualmente las cadenas de Markov generadas. Para seleccionar los modelos bayesianos, como menciona Bazan y Bayes (2010) y Chen et al. (2014), se utilizan distintos criterios como el Criterio de información del Desvío (DIC), el cual es una generalización jerárquica del Criterio de Información Akaike (AIC) (Akaike, 1973) y del Criterio de Información Bayesiano (BIC) (Raftery et al., 2007). En los modelos con casos perdidos o variables latentes, sin embargo, el DIC admite diferentes variantes (Celeux et al., 2006). Una de estas, la que usaremos en este trabajo, es el DIC condicional. Si $\theta = (\pi, \eta)$ denota a nuestro arreglo de interés este criterio de información se basa en el desvío

$$D(\theta) = -2 \log \left(\prod_{n=1}^N P(\mathbf{X}_n = \mathbf{x}_n \mid \mathbf{Y}_n = \mathbf{y}_n, \theta) \right)$$

y viene definido por

$$DIC = Dbar + p\widehat{D} = 2Dbar - Dhat,$$

donde

$$Dbar = \frac{1}{K} \sum_{k=1}^K D(\theta^k)$$

siendo θ^k la k-ésima observación o simulación del conjunto de parámetros θ obtenidos en la cadena para las K iteraciones sin periodo de quemado y $p\widehat{D}$ el número efectivo de parámetros

dado por: $\widehat{pD} = Dbar - Dhat$, siendo

$$Dhat = D \left(\frac{1}{K} \sum_{k=1}^K \theta^k \right).$$

Cuando se contrastan modelos desde el enfoque bayesiano se prefieren aquellos modelos con un menor valor de DIC.



Capítulo 4

Aplicación

4.1. Objetivos

El presente capítulo tiene como objetivo aplicar, a un conjunto de datos reales, la estimación en el modelo de clases latentes, tanto por el método bayesiano como por el método clásico. Para ello, primero se realizará una aplicación en el caso dicotómico con la base de datos *Alzheimer* contenida en la librería *BayesLCA* (White y Murphy, 2014). Se compararán las estimaciones obtenidas por esta librería con las aquí desarrolladas a cuyo procedimiento denominaremos como *BayesGibbsP*. Como comentamos el paquete *BayesLCA* realiza la estimación bayesiana desarrollada para este trabajo sólo para el caso dicotómico.

En segundo lugar, se realizará una aplicación para el caso politómico con un conjunto de datos reales sobre la confianza institucional en Lima Metropolitana. Aquí usaremos tanto los métodos bayesianos como clásicos. Previamente a lo anterior, se presentan los criterios con que se evaluarán las estimaciones de los parámetros generados.

4.2. Aplicación a la base de datos *Alzheimer* : el caso dicotómico

El crecimiento demográfico y el aumento de la expectativa de vida hacen que aumente el número de población senil, y con ello un aumento del número de patologías asociadas al envejecimiento. Según el Reporte Mundial de Alzheimer 2016 (Prince et al., 2016), 47 millones de personas a nivel mundial tienen Alzheimer y se estima que la cifra seguirá en aumento. El Alzheimer es una enfermedad cerebral que causa problemas con la memoria, el pensamiento, el carácter y el comportamiento. En este sentido, la investigación sobre la enfermedad ha abordado la relación de los síntomas de comportamiento y psicológicos de la demencia (BPSD, por sus siglas en inglés) con la misma enfermedad (Moran et al., 2004). De esta manera, se analiza si los pacientes con Alzheimer podrían clasificarse de manera significativa, según las relaciones entre los síntomas de comportamiento y psicológicos de la demencia (BPSD).

4.2.1. Descripción de la base de datos

La bases de datos *Alzheimer* fue descrita por White y Murphy (2014) y Walsh (2006), proviene de la clínica St James's Hospital en Dublín, Irlanda. La muestra contiene la información de 240 individuos quienes por primera vez asistieron al centro médico y se les realizó un examen sobre síntomas de la enfermedad con la presencia o ausencia de : Alucinación, Actividad, Agresión, Agitación, Diurnal y Afectividad. Dado que los pacientes presentaban los síntomas de la enfermedad en etapas leves, se recodificó la escala ordinal del instrumento

original a una versión binaria.

Dado que el patrón de síntomas puede ser ahora descrito por variables binarias, se presentan tales combinaciones o patrones de respuesta en la forma $\{0, 1\}^6$. Por ejemplo si la persona presenta todos los síntomas al patrón de respuesta lo identificamos con $\{1, 1, 1, 1, 1, 1\}$ y la ausencia de síntomas con $\{0, 0, 0, 0, 0, 0\}$. La distribución de frecuencia de los patrones encontrados se muestra en el cuadro 4.1.

Cuadro 4.1: Distribución de frecuencias de los patrones de respuesta a los síntomas

Patrón	Nº de casos	Patrón	Nº de casos	Patrón	Nº de casos
111111	3	011101	14	001101	2
111011	1	011011	2	001011	1
111001	1	011010	3	001010	1
110101	1	011001	9	001001	4
110011	2	011000	1	001000	2
110001	5	010111	11	000111	3
110000	2	010101	24	000110	1
101001	1	010100	3	000101	9
100101	1	010011	11	000100	3
100001	1	010010	2	000011	6
100000	1	010001	35	000010	1
011111	6	010000	20	000001	25
011110	1	001111	3	000000	18

4.2.2. Estimación de parámetros

Para la estimación de los parámetros, se muestran tablas comparativas con los resultados de la librería *BayesLCA* bajo la función *bayes.gibbs* y con los parámetros estimados por el código desarrollado al que estamos llamando *BayesGibbsP*. Se realizó la prueba con 5000 iteraciones y con burn-in de 100. En este caso, se muestran los resultados asumiendo que existen $J = 3$ clases, tal como lo sugiere [White et al. \(2016\)](#), en base a los criterios de información y parsimonia. Este autor indica también que se podrían usar 2 clases. Los códigos en R del algoritmo *BayesGibbsP* se encuentran descritos en el anexo.

Probabilidad de pertenencia a una clase

Los resultados de la estimación del parámetro η , que recordemos denota a la distribución de probabilidades de pertenencia a las clases, se muestran en el cuadro 4.2.

Cuadro 4.2: Probabilidad de pertenencia a una clase (η) - Aplicación en la base de datos *Alzheimer*

	<i>Método BayesLCA</i>			<i>Método BayesGibbsP</i>		
	Clase 1	Clase 2	Clase 3	Clase 1	Clase 2	Clase 3
η	0.493	0.424	0.083	0.513	0.398	0.089

Como se observa en el cuadro 4.2, los resultados de la estimación de los parámetros son muy similares con ambos métodos. Se tiene una mayor probabilidad de pertenencia a la clase 1 (0.50), mientras que la probabilidad de pertenencia a la clase 2 es 0.40 y a la clase 3 es 0.08.

Adicionalmente, en las figuras 4.1 y 4.2 se muestran las iteraciones de los parámetros η según el algoritmo de Gibbs, donde se evidencian gráficamente la convergencia de los parámetros descritos.

Figura 4.1: Iteraciones del parámetro η generado por la función `blca.gibbs` del paquete *BayesLCA*- Aplicación en la base de datos *Alzheimer*

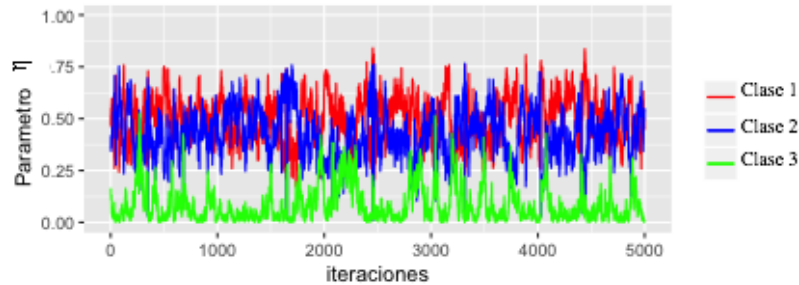
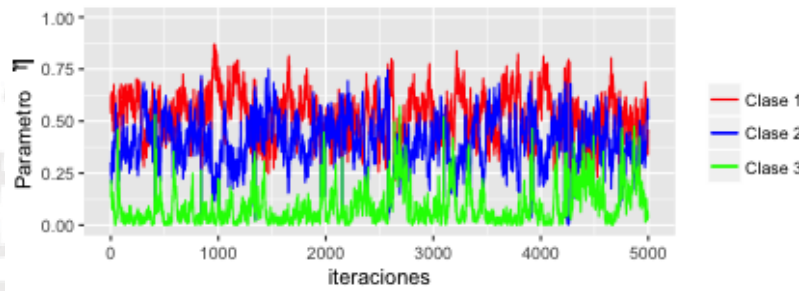


Figura 4.2: Iteraciones del parámetro η generado por el código construido *BayesGibbsP*- Aplicación en la base de datos *Alzheimer*



Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase

La estimación de los parámetros π_{ij} , que recordemos denotan a la probabilidad de que un individuo de a la variable manifiesta i una respuesta positiva sabiéndose que pertenece a la clase j , se muestran en el siguiente cuadro tanto por el método *BayesLCA* como por el método *BayesGibbsP* implementado.

Cuadro 4.3: Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase (π_{ij})- Aplicación en la base de datos *Alzheimer*

	<i>Método BayesLCA</i>			<i>Método BayesGibbsP</i>		
	Clase 1	Clase 2	Clase 3	Clase 1	Clase 2	Clase 3
Alucinación	0.076	0.086	0.266	0.075	0.090	0.266
Actividad	0.532	0.784	0.642	0.538	0.793	0.641
Agresión	0.095	0.375	0.485	0.101	0.392	0.483
Agitación	0.130	0.639	0.381	0.141	0.657	0.395
Diurnal	0.117	0.357	0.519	0.124	0.373	0.514
Afectividad	0.580	0.937	0.630	0.599	0.938	0.643

Como se observa en el cuadro 4.3, en la clase 1 se agrupan los pacientes con menor proba-

bilidad de presentar síntomas de Alzheimer. En la clase 2, las probabilidades de presentar los síntomas de Alzheimer son mayores que en la clase 1, siendo Actividad, Agitación y Afectividad los síntomas con más altas probabilidades. Asimismo, en la clase 3, también las probabilidades de presentar los síntomas de Alzheimer son también mayores que en la clase 1, siendo los síntomas de Alucinación, Agresión y Diurnal, los que registran más alta probabilidad.

Asimismo, en las figuras 4.3 y 4.4 se observan las iteraciones de los parámetros π_{ij} generadas por el algoritmo de Gibbs.

Figura 4.3: Iteraciones de los parámetros π_{ij} generados por el método *BayesLCA*- Aplicación en la base de datos *Alzheimer*

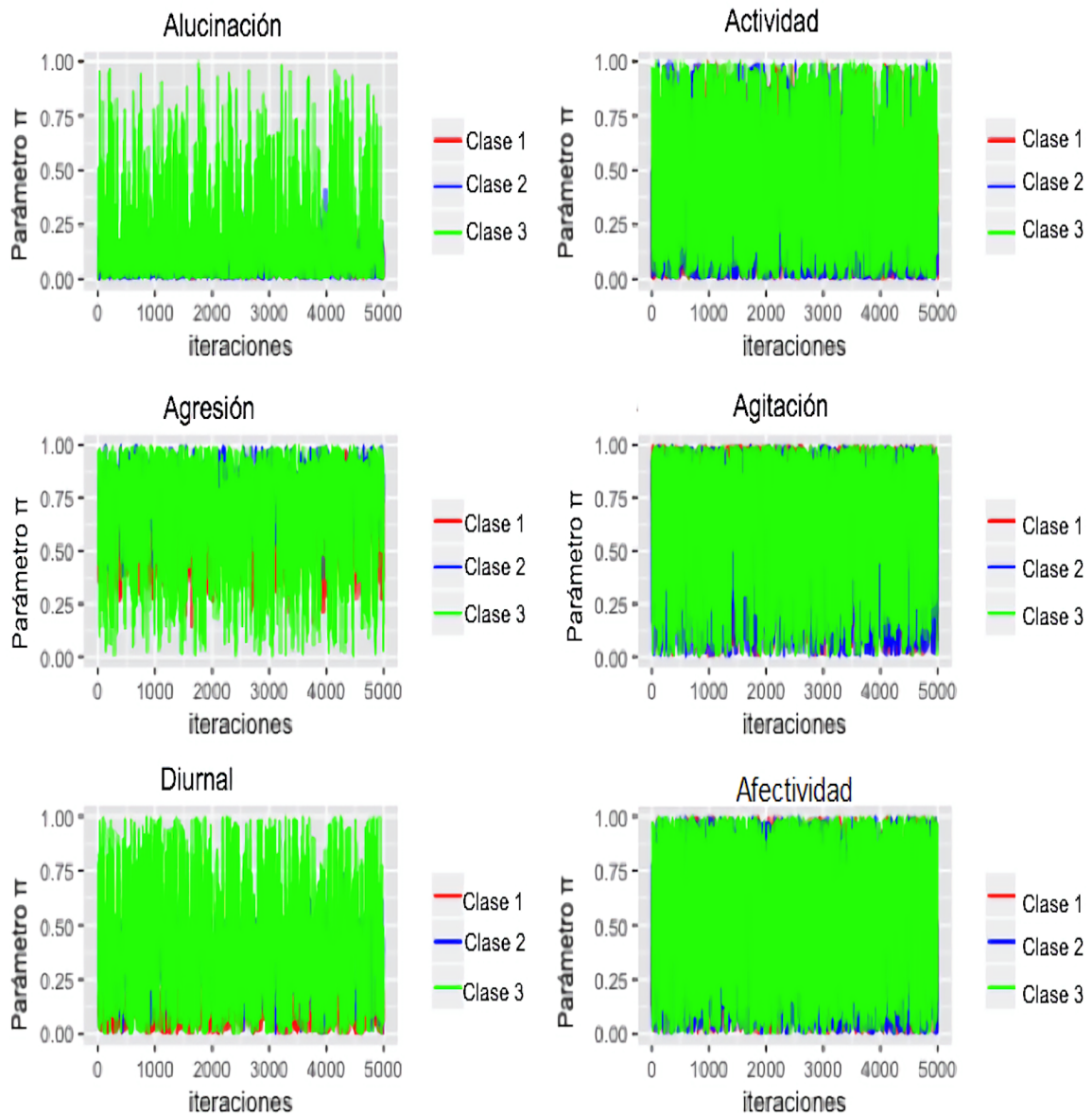
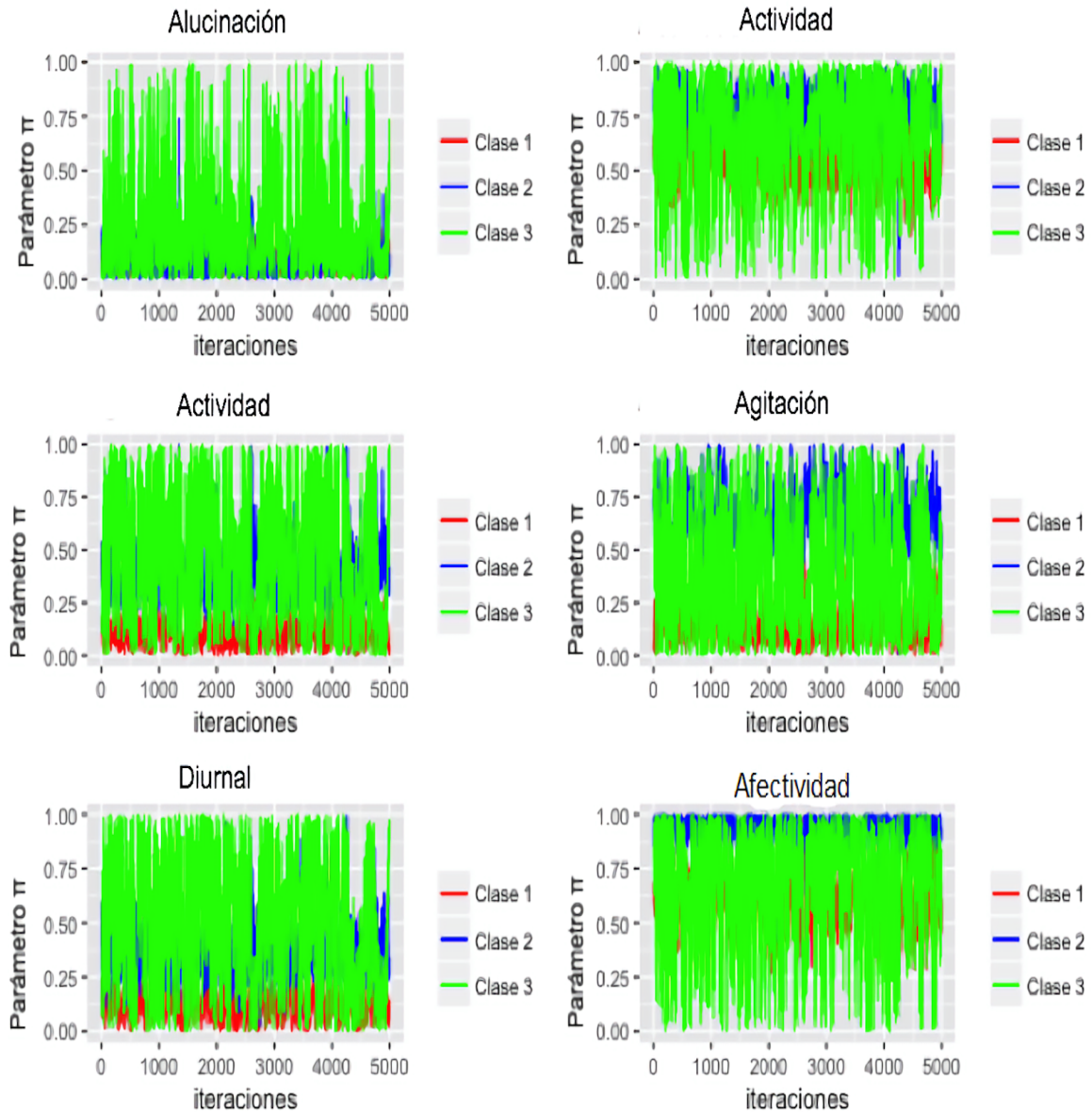


Figura 4.4: Iteraciones de los parámetros π_{ij} generados por el método *BayesGibbsP*- Aplicación en la base de datos *Alzheimer*



4.3. Aplicación para la confianza institucional

La aplicación presentada en este trabajo se encuentra relacionada a las percepciones sobre política, específicamente sobre confianza institucional. El objetivo de esta aplicación será el estimar y analizar a través del modelo de clases latentes a la confianza institucional usando el enfoque clásico y bayesiano, siendo este último el implementado en este trabajo a través del algoritmo de Gibbs.

En el estudio del comportamiento político y las relaciones entre los gobernantes y gobernados se ha abordado el concepto de «Confianza», entendida como la fe que tienen los ciudadanos en los actores políticos y organizaciones públicas (Levi y Stoker, 2000). El modo como se da la confianza en los ciudadanos con las instituciones genera la base para la legitimidad de las instituciones y también de la democracia. Definimos a la confianza insti-

tucional como la percepción individual sobre el desempeño de las instituciones, basando sus juicios a criterios racionales vinculados al desempeño y criterios socioculturales exógenos al sistema político (Hakhverdian y Mayne, 2012). Existe una amplia literatura que relaciona a la confianza institucional con la corrupción, la rendición de cuentas, el crimen, el sistema democrático y económico como los trabajos de Hakhverdian y Mayne (2012), Chang y Chu (2006) y Mishler y Rose (2001). A esta literatura se suma el análisis de confianza institucional en países como Canadá (Crête et al., 2006); China (Yang, 2007) ; Guyana (Thomas et al., 2015) ; Finlandia (Kestila-Kekkonen y Soderlund, 2016) , etc.

De manera general, se ha abordado el concepto de confianza desde la teoría de elección racional, la teoría de los sistemas y la teoría cognitiva (Gordon, 2005). La teoría de la elección racional indica que la confianza se establece en la medida que se da un intercambio específico, es decir, el individuo realiza un análisis de costo y beneficio, en el que obtiene algún beneficio, si establece una relación de confianza. La teoría de sistemas sostiene que la confianza reduce la incertidumbre o el riesgo de una acción, en otras palabras, el individuo confía para de alguna manera predecir un comportamiento y reducir su incertidumbre. La teoría cognitiva sostiene que la confianza depende de un conocimiento previo sobre las características y desempeño en el comportamiento sobre lo cual se va a confiar (Cohaila, 2015).

Específicamente, las teorías que han abordado el concepto de confianza institucional son la teoría culturalista e institucionalista. La primera menciona que la confianza institucional es exógena y se forma a través de las normas culturales y es la extensión de la confianza interpersonal como se aprecia en los estudios de Inglehart (1988) y Almond y Verba (1963) . En contraste la teoría institucionalista enfatiza que es necesario la evaluación del desempeño de cada institución para construir una relación de confianza. Los autores que han abordado las explicaciones sobre desempeño asociada a la confianza institucional son: Hiskey y Seligson (2003) , Levi (1998) y Hardin (1998) .

Los estudios bianuales sobre cultura política en América Latina realizados desde 1996 por el Barómetro de la Américas LAPOP de la Universidad de Vanderbilt (Carrión et al., 2015), sostienen que nuestro país registra los niveles más bajos de confianza institucional (en una escala del 0 al 100).

Como se observa en el cuadro 4.4, los promedios de confianza hacia las instituciones en Perú están por debajo del promedio de Sudamérica.

Hasta el momento, los estudios sobre confianza institucional en el Perú han sido descriptivos por cada institución y no han abordado algún análisis que involucre a la confianza institucional en conjunto. En este sentido, la aplicación del modelo de clases latentes permite agrupar las percepciones hacia las instituciones públicas e identificar perfiles de confianza institucional según los patrones de respuestas de los encuestados. Adicionalmente, se realizará un análisis de las clases resultantes de confianza institucional en función de otras variables complementarias.

4.3.1. Instrumentos de Medición

Entre los instrumentos de medición de la confianza institucional encontramos cuestionarios de una escala Likert de 1 a 10, donde el mínimo valor indique que no hay nada de confianza y el máximo valor que hay confianza plena (Cuestionario de la Encuesta Social

Cuadro 4.4: Comparación entre el promedio de los niveles de confianza en instituciones en Sudamérica y el promedio en Perú

	Promedio en Sudamérica	Promedio en Perú
Fuerzas Armadas	57.8	52.6
Jurado Nacional de Elecciones(*)	49	46.9
Defensoría del Pueblo	50.4	46.8
Presidente	50.1	38.8
Policía	38.5	38.5
Gobierno Locales	48.9	36.6
Poder Judicial	43.8	34.2
Congreso de la República	43.3	27.4
Partidos Políticos	31.1	24.7

(*) Sólo Bolivia/Perú/Ecuador

Fuente: Barómetro de las Américas LAPOP 2014 (Carrión et al., 2015)

Europea (ESE, 2018)), una escala Likert de 1 a 7 (Cuestionario de la Encuesta de Latino-barómetro (LAPOP, 2018)) y una reducción de la escala Likert a 4 opciones, siendo 1 Nada, 2 Poco, 3 Algo y 4 Mucho (Cuestionario de la Encuesta Mundial de Valores (WVS, 2018)).

En este trabajo usaremos la Encuesta Nacional de Hogares (INEI, 2018). La sección del cuestionario que empleamos se encuentra en el módulo de Gobernabilidad, Democracia y Transparencia. El encuestador registro la información sobre los niveles de confianza de cada una de las veintiuno instituciones públicas que se mencionan a continuación, mediante una escala Likert de cuatro opciones: *Nada* (1), *Poco* (2), *Suficiente* (3) y *Bastante* (4).

Instituciones Públicas

1. JNE : El Jurado Nacional de Elecciones JNE
2. ONPE : La Oficina Nacional de Procesos Electorales ONPE
3. RENIEC: Registro Nacional de Identificación y Estado Civil RENIEC
4. MP : La Municipalidad Provincial
5. MD: La Municipalidad Distrital
6. PNP: La Policía Nacional del Perú
7. FFAA: Las Fuerzas Armadas
8. GR: El Gobierno Regional
9. PJ: El Poder Judicial
10. MINEDU: El Ministerio de Educación
11. DP : La Defensoría del Pueblo

12. CR : El Congreso de la República
13. PP: Los Partidos Políticos
14. PE: La Prensa Escrita
15. RTV : La Radio o Televisión
16. IC : La Iglesia Católica
17. PA : La Procuraduría Anticorrupción
18. MPU : Ministerio Público
19. CGR : Contraloría General de la República
20. SUNAT: Superintendencia Nacional de Aduanas y de Administración Tributaria
21. CANA : La Comisión de Alto Nivel Anticorrupción

El cuestionario sobre confianza presentó un coeficiente de alfa de Cronbach superior a 0.7, lo cual asegura la confiabilidad de este instrumento.

4.3.2. Diseño muestral

El diseño muestral empleado fue probabilístico, estatificado, multietápico, independiente a cada departamento. El tamaño de la muestra 2017 fue de 36 996 viviendas particulares, correspondiendo 23 348 viviendas al área urbana y 13 648 viviendas al área rural. Asimismo, el tamaño de la muestra panel fue de 11 900 viviendas particulares mientras que el tamaño de la muestra no panel es fue 25 096 viviendas particulares. La muestra de conglomerados en el ámbito nacional fue de 5 359, correspondiendo 3 3653 conglomerados al área urbana y 1 718 conglomerados al área rural. Con respecto al tamaño de la muestra panel, esta fue de 1 718 conglomerados mientras que el tamaño de la muestra no panel fue de 3 426 conglomerados (INEI, 2018). En este estudio se ha determinado trabajar con la muestra registrada para Lima Metropolitana y Callao que asciende a un total de 1 119 individuos que contestaron todas las preguntas de interés sobre la confianza institucional.

Para la recolección de la muestra, se realizaron entrevistas directas en la vivienda a informantes como jefes de hogares, cónyuges y personas mayores de 18 años a más.

4.3.3. Descripción de los datos

De manera general, se observa que para la mayoría de instituciones se tienen porcentajes altos en las categorías de *Nada* y *Poca* confianza. Las instituciones percibidas como menos confiables son: el Congreso de la República, los Partidos Políticos. Estas instituciones tienen porcentajes que no superan el 10 % de recepción de confianza.

Cabe resaltar que la alta confianza institucional en todas las instituciones mencionadas, a excepción de la Iglesia Católica (12 %), no son mayores al 5 %. Por otro lado, como se muestra en el cuadro 4.5, las instituciones más confiables son: la Iglesia Católica (*Suficiente* 39 %, *Bastante* 12 %), las Fuerzas Armadas (*Suficiente* 31 %, *Bastante* 5 %), el Ministerio de Educación (*Suficiente* 31 %, *Bastante* 4 %) y el Registro Nacional de Identificación y Estado Civil (*Suficiente* 51 %, *Bastante* 10 %).

Cuadro 4.5: Porcentajes de respuestas de confianza en las instituciones

	Nada	Poco	Suficiente	Bastante
JNE	32 %	43 %	24 %	2 %
ONPE	29 %	42 %	27 %	3 %
RENIEC	15 %	24 %	51 %	10 %
MP	43 %	40 %	16 %	1 %
MD	41 %	38 %	18 %	3 %
PNP	40 %	42 %	16 %	2 %
FFAA	28 %	36 %	31 %	5 %
GR	48 %	39 %	12 %	0 %
PJ	58 %	32 %	9 %	1 %
MINEDU	25 %	41 %	31 %	4 %
DP	32 %	42 %	23 %	2 %
CR	65 %	28 %	6 %	1 %
PP:	70 %	26 %	4 %	0 %
PE:	35 %	46 %	17 %	2 %
RTV	26 %	47 %	24 %	3 %
IC	24 %	25 %	39 %	12 %
PA	50 %	38 %	11 %	1 %
MPU	44 %	41 %	13 %	1 %
CGR	44 %	41 %	14 %	1 %
SUNAT:	36 %	38 %	25 %	2 %
CANA	48 %	39 %	12 %	1 %

4.3.4. Caso dicotómico

En esta sección, se presentará la aplicación considerando a las variables manifiestas dadas como binarias con el propósito de comparar el algoritmo *blca.gibbs* de paquete *BayesLCA* (White y Murphy, 2014) y el *BayesGibbsP* del algoritmo implementado. Para la aplicación de la base de datos de *Confianza Institucional*, primero se recodifico a las variables manifiestas originales politómicas en variables dicotómicas, en esta se agrupan las categorías *Nada* y *Poco* como 0 y *Suficiente* y *Bastante* como 1. Teniendo como patrones de respuesta a las combinaciones de la forma $\{0, 1\}^{21}$, los cuales serán utilizados para la conformación de clases latentes.

Si apreciamos los estadísticos del Criterio de información de Desviación (DIC) en el cuadro 4.6, se observa que a mayor número de grupos o clases, los valores de DIC son menores por el método *BayesGibbsP*. Se observa los valores disminuyen rápidamente y se estabilizan a partir del modelo de 3 clases.

Para elegir el número de clases latentes, se consideró la parsimonia (más simple mejor) y facilidad de interpretación (Collins y S., 2010). Si bien los resultados del cuadro 4.6 nos sugiere que un modelo de 5 clases, se ajustan ligeramente mejor que los modelos con 3 o 4 clases, por un tema de interpretación y de acuerdo a la distribución de la base de datos mostrada en el cuadro 4.5, la frecuencia es mayor en las categorías *Poco*, *Nada*. Una agrupación en 3 clases nos brinda una mayor diferenciación entre los niveles de confianza hacia las instituciones, como se verá más adelante.

En este caso, se mostrará la aplicación con 5 000 iteraciones con un quemado (burn-in) de 100.

Cuadro 4.6: Criterios de información para el ajuste del modelo

<i>Método BayesGibbsP</i>	
Clases	DIC
2	19405.65
3	18621.22
4	18327.39
5	18034.3

4.3.4.1. Estimación de parámetros

En esta sección se mostrará la estimación de los parámetros de probabilidad de pertenencia a una clase (η) y las probabilidades de respuesta a las variables manifiestas de acuerdo a la clase (π_{ijc}).

Probabilidad de pertenencia a una clase

Los resultados de la estimación de los parámetro η que recordemos denota a la probabilidad de pertenencia a la clase, se muestran en el cuadro 4.7. La probabilidad de pertenecer a la clase 1 es 0.47, a la clase 2 es 0.41 y a la clase 3 es 0.11. Las probabilidades de pertenencia a cada clase son las mismas para ambos métodos *BayesLCA* y *BayesGibbsP* implementado.

Cuadro 4.7: Probabilidad de pertenencia a una clase (η) - Aplicación en la base de datos *Confianza*(versión dicotómica)

	<i>Método BayesLCA</i>			<i>Método BayesGibbsP</i>		
	Clase 1	Clase 2	Clase 3	Clase 1	Clase 2	Clase 3
η	0.475	0.417	0.107	0.474	0.417	0.108

Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase

La estimación de los parámetros π_{ij} que denotan a las probabilidades de que un individuo, que pertenece a la clase j , de a la variable manifiesta i de una respuesta positiva, se muestran en el cuadro 4.8. Se realizó la prueba con 5 000 repeticiones descartando las 100 primeras iteraciones.

Como se observa en el cuadro 4.8, las probabilidades más altas de tener confianza en las instituciones se presenta en la clase 3, con mayor énfasis en el Registro Nacional de Identificación y Estado Civil (RENIEC), las Fuerzas Armadas (FFAA), el Ministerio de Educación (MINEDU), la Oficina Nacional de Procesos Electorales (ONPE), la Defensoría del Pueblo (DP) y la Iglesia Católica,(IC) que son las instituciones con más alta probabilidad. En la clase 2, se observa alta probabilidad en RENIEC, seguidas de probabilidades entre 0.4 y 0.6 en las Fuerzas Armadas, MINEDU, ONPE, JNE e Iglesia Católica. En la clase 1, se puede apreciar la probabilidad de la confianza en las instituciones públicas es baja, dentro

Cuadro 4.8: Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase (π_{ij}) - Aplicación en la base de datos *Confianza*(versión dicotómica)

	<i>Método BayesLCA</i>			<i>Método BayesGibbsP</i>		
	Clase 1	Clase 2	Clase 3	Clase 1	Clase 2	Clase 3
JNE	0.023	0.402	0.718	0.024	0.404	0.718
ONPE	0.025	0.469	0.855	0.026	0.47	0.856
RENIEC	0.322	0.85	0.927	0.323	0.851	0.927
MP	0.031	0.21	0.616	0.031	0.211	0.617
MD	0.069	0.258	0.631	0.069	0.259	0.633
PNP	0.019	0.232	0.728	0.02	0.233	0.729
FFAA	0.114	0.507	0.885	0.115	0.509	0.885
GR	0.014	0.113	0.684	0.014	0.114	0.687
PJ	0.007	0.076	0.62	0.007	0.077	0.622
MINEDU	0.135	0.441	0.88	0.135	0.444	0.88
DP	0.042	0.348	0.847	0.042	0.349	0.848
CR	0.004	0.059	0.42	0.004	0.06	0.423
PP	0.008	0.035	0.275	0.008	0.035	0.279
PE	0.048	0.265	0.539	0.049	0.265	0.54
RTV	0.106	0.353	0.694	0.107	0.354	0.694
IC	0.386	0.582	0.789	0.387	0.583	0.791
PA	0.010	0.115	0.687	0.010	0.115	0.691
MPU	0.016	0.128	0.773	0.016	0.129	0.776
CGR	0.008	0.164	0.702	0.008	0.166	0.704
SUNAT	0.035	0.395	0.782	0.035	0.397	0.786
CANA	0.010	0.136	0.617	0.010	0.136	0.621

de las cuales la Iglesia Católica tiene la probabilidad más alta; es decir en el grupo más desconfiado, aún se mantiene la confianza en la Iglesia Católica.

Una de las limitaciones de utilizar las variables binarias es la pérdida de información y variabilidad. Para el caso de confianza, es necesario realizar una distinción en quien confía poco y quien no confía nada en las instituciones. En el caso dicotómico, no se realiza una clara caracterización de las clases según los ítems, ya que la alta probabilidad de estos se concentran en una clase y las bajas probabilidades de todos los ítems se agrupan en otra. Por estas razones es pertinente realizar la aplicación de modelos de clases latentes para variables politómicas.

4.3.5. Caso politómico

En esta sección se comparan los parámetros estimados en el enfoque clásico y el bayesiano. Para la estimación clásica se utilizará el paquete *poLCA* (Drew y Lewis, 2011) que desarrolla la estimación por el método de máxima verosimilitud y para el método bayesiano por el código desarrollado con el algoritmo de Gibbs *BayesGibbsP* implementado.

En este caso, se tiene como patrón de respuesta a las combinaciones posibles en la forma $\{1, 2, 3, 4\}^{21}$. Según la muestra trabajada se encontraron 214 patrones distintos de respuestas. El patrón más común tiene 35 repeticiones en el cual respondieron Nada (1) en los 21 ítems, a este le sigue el patrón donde se obtuvo Nada (1) en 20 ítems y Suficiente (3) en la Iglesia Católica.

Cuadro 4.9: Los tres patrones de respuestas de mayor frecuencia

Patrón	Nº de casos
11111111111111111111111111111111	35
1111111111111111111131111111	13
11111111111111111112111111	8

Para elegir el número de clases, se han considerado los criterios de información del cuadro 4.9, la parsimonia y la facilidad de interpretación. En este caso, por el enfoque clásico, los resultados de AIC , BIC (*poLCA*) y DIC (*BayesGibbsP*) nos sugieren que un modelo de 5 clases se ajusta mejor que los demás modelos. No existe una clara una diferenciación entre los niveles de confianza hacia las instituciones con una interpretación de 4 y 5 clases. Por esta razón, además por un tema interpretación y simpleza, preferimos quedarnos con un modelo de 3 clases.

En este caso, se mostrará la aplicación con 5 000 iteraciones con un quemado (burn-in) de 100.

Cuadro 4.10: Criterios de información para el ajuste del modelo

Clases	<i>Método poLCA</i>		<i>Método BayesGibbsP</i>
	AIC	BIC	DIC
2	44266.7	44904.3	44134.68
3	42354.5	43313.3	42163.97
4	41251.2	42531.4	41006.94
5	40822.9	42422.3	40701.95

Por el método *poLCA* se observa que los valores de AIC y BIC un punto de inflexión en el modelo de clases latentes que toma en cuenta 3 grupos, y posteriormente a medida que se incrementan el número de clases, los valores se estabilizan. De la misma manera, por el método *BayesGibbsP* se observa un punto de inflexión en los valores de DIC en el modelo de clases latentes que considera 3 grupos, por esta razón y la facilidad en la interpretación optamos por un modelo de 3 clases latentes.

4.3.5.1. Estimación de parámetros

En esta sección se mostrará la estimación de los parámetros de probabilidad de pertenencia a una clase (η) y las probabilidades de respuesta a las 21 variables manifiestas de acuerdo a la clase (π_{ijc}) . Se mostrará la aplicación con 5 000 iteraciones con un quemado (burn-in) de 100 y como mencionamos antes la conformación de 3 clases.

Probabilidad de pertenencia a la clase

En el cuadro 4.11 observamos la estimación de la probabilidad de pertenencia a la clase denotada por η . La probabilidad de pertenencia a la clase1 es 0.42 , a la clase 2 es 0.39 y a la clase 3 es 0.18, es decir, es más probable de que una institución pertenezca a la clase 1.

Cuadro 4.11: Probabilidad de pertenencia a la clase o grupo (η) - Aplicación de la base de datos *Confianza (versión politómica)*

	<i>Método poLCA</i>			<i>Método BayesGibbsP</i>		
	Clase 1	Clase 2	Clase 3	Clase 1	Clase 2	Clase 3
η	0.422	0.392	0.186	0.422	0.4	0.178

Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase

A diferencia de la aplicación dicotómica, en el caso politómico se realiza una distinción entre las categorías de respuesta a las variables manifiestas. Los parámetros π_{ijc} que denotan a las probabilidades de que un individuo pertenezca a la clase j de a la variable manifiesta (o ítem) i una respuesta $c \in \{1, 2, 3, 4\}$.

En el cuadro 4.12 se muestra que en la clase 1, se agrupan quienes confían poco en las instituciones públicas, las probabilidades más altas se presentan en las categorías «Poca Confianza» en el Ministerio Público, la Contraloría General de la República, la Defensoría del Pueblo, la Prensa Escrita, la Procuraduría Anticorrupción, el Gobierno Regional y la Comisión de Alto Nivel Anticorrupción, Radio y TV, Policía Nacional. A esta clase la denominaremos como el grupo de poca confianza institucional.

En la clase 2 se agrupan por quienes en mayor proporción respondieron a la categoría «Nada Confianza» en los Partidos Políticos, Congreso de la República y Poder Judicial, la Comisión de Alto Nivel Anticorrupción, la Contraloría General de la República, Ministerio Público. Se denominará a esta clase como el grupo de ninguna confianza institucional. Es necesario notar que las variables son las referentes a instituciones relacionadas a la representación política y de justicia.

En la clase 3 que agrupa a los que confían lo suficiente en las instituciones públicas, los ítems que mayor respuestas en la categoría «Suficiente Confianza» son la confianza al Ministerio de Educación, la Defensoría del Pueblo, la SUNAT, el RENIEC, la ONPE, las Fuerzas Armadas y la Iglesia Católica. Por tanto, denominaremos a esta clase como la de suficiente confianza institucional.

De esta manera, indicamos que las instituciones percibidas como más confiables son las que tienen mayor alcance territorial en el país y son más cercanas a la población en términos de su atención como también sobre las labores que realizan. Por ejemplo, el Ministerio de Educación a través de los centros educativos, la Defensoría del Pueblo a través de las oficinas defensoriales y módulos de atención, la SUNAT a través de sus centros contribuyentes y su plataforma virtual, la RENIEC a través sus agencias, oficinas y su plataforma virtual, la ONPE por su labor durante las elecciones, la Iglesia Católica a través de los centros parroquiales, y las Fuerzas Armadas a través de las labores de ayuda y protección a la población. De otro lado, las instituciones percibidas como menos confiables son organizaciones de representación política y de justicia como el Congreso, los Partidos Políticos, el Poder Judicial, la Comisión de Alto Nivel Anticorrupción, la Contraloría General de la República y Ministerio Público. La ejecución de las funciones de estas instituciones no son cercanas a la población.

Cuadro 4.12: Probabilidades de respuesta a las variables manifiestas de acuerdo a la clase (π_{ijc}) - Aplicación de la base de datos *Confianza (versión politómica)*

		<i>Método poLCA</i>			<i>Método BayesGibbsP</i>		
		Clase 1	Clase 2	Clase 3	Clase 1	Clase 2	Clase 3
JNE	Nada	0.155	0.612	0.041	0.159	0.621	0.036
	Poco	0.64	0.259	0.291	0.644	0.254	0.299
	Suficiente	0.201	0.120	0.585	0.196	0.118	0.586
	Bastante	0.004	0.009	0.084	0.001	0.007	0.079
ONPE	Nada	0.119	0.576	0.031	0.122	0.585	0.027
	Poco	0.642	0.278	0.176	0.646	0.274	0.187
	Suficiente	0.231	0.133	0.671	0.226	0.13	0.669
	Bastante	0.009	0.013	0.122	0.006	0.011	0.117
RENIEC	Nada	0.021	0.350	0.025	0.019	0.357	0.019
	Poco	0.326	0.235	0.049	0.331	0.233	0.045
	Suficiente	0.579	0.365	0.66	0.576	0.363	0.677
	Bastante	0.074	0.050	0.266	0.074	0.046	0.259
MP	Nada	0.247	0.762	0.132	0.252	0.772	0.129
	Poco	0.639	0.165	0.348	0.641	0.159	0.358
	Suficiente	0.105	0.07	0.494	0.100	0.068	0.491
	Bastante	0.009	0.003	0.026	0.008	0.000	0.021
MD	Nada	0.242	0.729	0.105	0.249	0.738	0.102
	Poco	0.592	0.155	0.369	0.59	0.151	0.381
	Suficiente	0.141	0.101	0.441	0.137	0.100	0.437
	Bastante	0.024	0.015	0.086	0.024	0.012	0.079
PNP	Nada	0.223	0.718	0.078	0.228	0.728	0.074
	Poco	0.658	0.215	0.310	0.660	0.210	0.324
	Suficiente	0.102	0.057	0.526	0.097	0.054	0.522
	Bastante	0.017	0.010	0.087	0.014	0.008	0.081
FFAA	Nada	0.111	0.576	0.033	0.113	0.586	0.027
	Poco	0.544	0.249	0.150	0.551	0.243	0.155
	Suficiente	0.316	0.152	0.661	0.31	0.15	0.668
	Bastante	0.029	0.023	0.156	0.027	0.021	0.150
GR	Nada	0.275	0.869	0.103	0.284	0.88	0.100
	Poco	0.663	0.099	0.391	0.662	0.092	0.405
	Suficiente	0.06	0.03	0.477	0.055	0.028	0.472
	Bastante	0.002	0.002	0.029	0.000	0.000	0.024
PJ	Nada	0.420	0.919	0.173	0.428	0.932	0.173
	Poco	0.533	0.063	0.396	0.529	0.057	0.410
	Suficiente	0.044	0.014	0.397	0.041	0.011	0.388
	Bastante	0.003	0.004	0.034	0.002	0.000	0.029
MINEDU	Nada	0.077	0.523	0.048	0.078	0.531	0.042
	Poco	0.586	0.321	0.173	0.59	0.318	0.181
	Suficiente	0.294	0.145	0.696	0.289	0.142	0.7
	Bastante	0.044	0.011	0.083	0.043	0.008	0.077
DP	Nada	0.09	0.683	0.048	0.095	0.693	0.043
	Poco	0.703	0.236	0.181	0.709	0.231	0.187
	Suficiente	0.184	0.076	0.698	0.176	0.073	0.702
	Bastante	0.023	0.005	0.073	0.021	0.003	0.068

		<i>Método poLCA</i>			<i>Método BayesGibbsP</i>		
		Clase 1	Clase 2	Clase 3	Clase 1	Clase 2	Clase 3
CR	Nada	0.542	0.943	0.244	0.556	0.95	0.243
	Poco	0.433	0.046	0.427	0.427	0.043	0.437
	Suficiente	0.023	0.009	0.29	0.018	0.007	0.287
	Bastante	0.002	0.002	0.039	0.000	0.000	0.034
PP	Nada	0.584	0.950	0.379	0.598	0.957	0.377
	Poco	0.391	0.037	0.429	0.383	0.034	0.44
	Suficiente	0.023	0.009	0.178	0.019	0.007	0.173
	Bastante	0.002	0.004	0.015	0.000	0.002	0.01
PE	Nada	0.156	0.643	0.141	0.16	0.651	0.136
	Poco	0.678	0.255	0.399	0.678	0.25	0.414
	Suficiente	0.158	0.092	0.400	0.156	0.089	0.396
	Bastante	0.008	0.011	0.061	0.006	0.009	0.054
RTV	Nada	0.093	0.522	0.064	0.095	0.529	0.061
	Poco	0.661	0.321	0.338	0.66	0.321	0.350
	Suficiente	0.225	0.132	0.518	0.226	0.126	0.516
	Bastante	0.021	0.026	0.080	0.019	0.024	0.073
IC	Nada	0.182	0.38	0.082	0.185	0.383	0.079
	Poco	0.346	0.18	0.152	0.348	0.179	0.15
	Suficiente	0.352	0.344	0.597	0.348	0.344	0.604
	Bastante	0.12	0.096	0.169	0.118	0.094	0.167
PA	Nada	0.279	0.88	0.142	0.292	0.887	0.139
	Poco	0.665	0.099	0.32	0.661	0.095	0.333
	Suficiente	0.05	0.015	0.493	0.044	0.013	0.488
	Bastante	0.005	0.007	0.046	0.003	0.005	0.04
MPU	Nada	0.19	0.872	0.08	0.198	0.882	0.076
	Poco	0.748	0.091	0.331	0.746	0.085	0.35
	Suficiente	0.06	0.033	0.54	0.056	0.03	0.53
	Bastante	0.002	0.004	0.049	0.000	0.002	0.043
CGR	Nada	0.174	0.88	0.100	0.183	0.889	0.096
	Poco	0.740	0.098	0.315	0.738	0.092	0.331
	Suficiente	0.084	0.021	0.525	0.079	0.018	0.520
	Bastante	0.002	0.002	0.059	0.000	0.000	0.053
SUNAT	Nada	0.134	0.736	0.046	0.142	0.744	0.042
	Poco	0.608	0.191	0.233	0.609	0.189	0.239
	Suficiente	0.249	0.063	0.638	0.243	0.059	0.642
	Bastante	0.009	0.011	0.084	0.007	0.009	0.077
CANA	Nada	0.254	0.883	0.109	0.264	0.893	0.107
	Poco	0.663	0.102	0.397	0.66	0.097	0.408
	Suficiente	0.079	0.011	0.45	0.073	0.008	0.447
	Bastante	0.004	0.004	0.044	0.002	0.002	0.038

El modelo de análisis de clases latentes, nos ha conformado una variable subyacente a la que hemos denominado Confianza Institucional como ya hemos mencionado. Esta nueva variable resultante será usada ahora para ver la relación de los niveles de confianza institucional con otras percepciones y actitudes de los encuestados. En la siguiente sección, se analizará la confianza institucional en relación a otras variables.

4.3.6. Confianza institucional en relación a otras variables

Con el propósito de analizar a mayor profundidad el rol de la confianza institucional, se realizó un análisis de las tres clases conformadas con variables complementarias que se relacionan a ella según la literatura (Morales, 2008) y que habían sido también registradas en la ENAHO como: la mención de corrupción como problema principal en el país y el apoyo a la democracia.

Sensación de Corrupción

Las denuncias sobre la corrupción generan una sensación de corrupción en el país, que afecta a los niveles de confianza institucional. De manera que si se considera que la dinámica del funcionamiento de las instituciones tienen acciones basadas en la corrupción, los niveles de confianza de los ciudadanos hacia las institucionales se verá afectada. En cuanto a esta variable, se preguntó por la percepción de corrupción y registró si considera a la corrupción como uno de los principales problemas del país, con las opciones de respuesta *Sí* - (Corrupción) y *No* - (No Corrupción).

Apoyo a la democracia

Analizar el apoyo a la democracia es una aproximación a evaluar la legitimidad del régimen político. Se considera que ciudadanos con baja confianza en las instituciones, se asocia con un poco apoyo a la democracia, lo cual puede generar escenarios de crisis del régimen. El apoyo a la democracia se registró mediante la pregunta ¿Con cuál de estas frases está más de acuerdo?, con las siguientes opciones de respuesta :

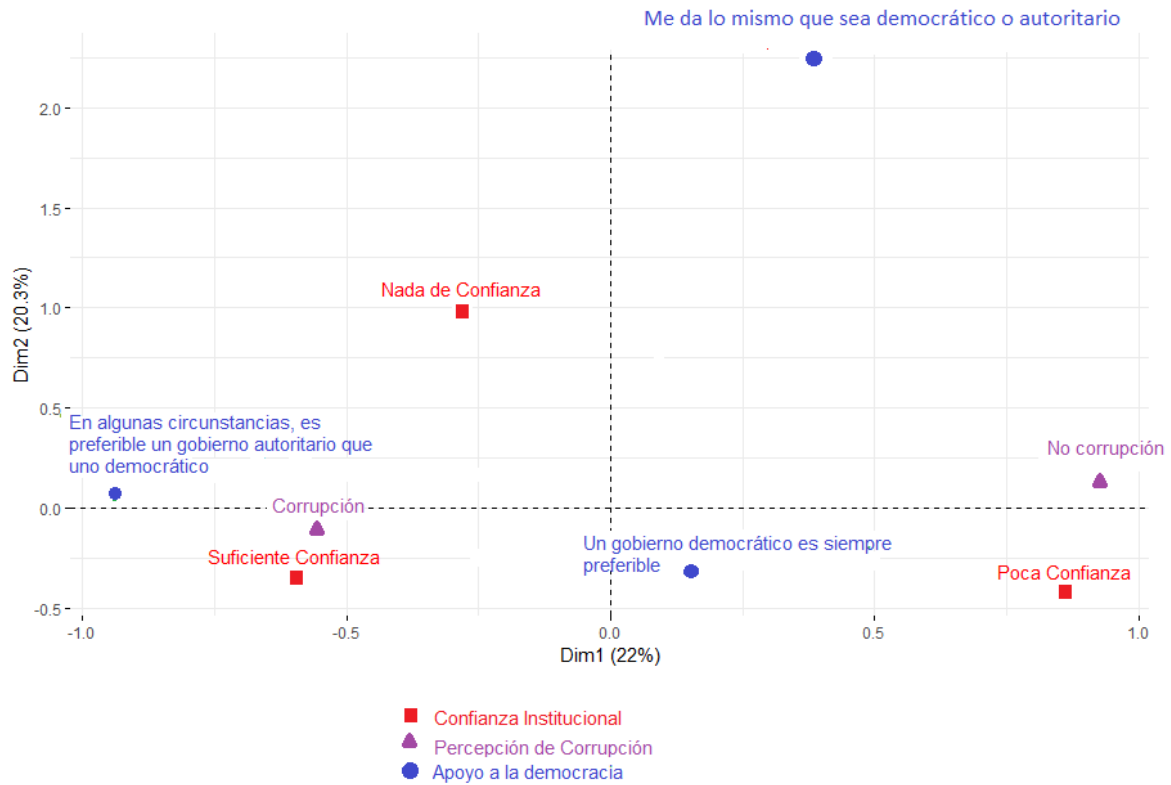
1. Un gobierno democrático es siempre preferible.
2. En algunas circunstancias, es preferible un gobierno autoritario que uno democrático.
3. Me da lo mismo que sea democrático o autoritario.

Análisis de correspondencia múltiple (ACM)

Para observar la relación entre Confianza Institucional y las variables arriba mencionadas, se utilizó el análisis de correspondencia múltiple (ACM). Esta técnica reduce la dimensionalidad de los datos analizados de un conjunto de variables categóricas u ordinales a dos o tres dimensiones, donde sitúan las categorías proyectadas de las variables analizadas y de los sujetos que intervienen en el análisis a través de las frecuencias y modalidades de las variables. Para la interpretación, se observa las relaciones entre los individuos, variables y categorías. Sobre los individuos se observa que si son cercanos es porque respondieron con alta frecuencia a la pregunta de la misma manera. Sobre las variables, se analiza si existe o no asociación entre las variables. Por último, se desea explorar también una caracterización de los grupos de individuos por categorías.

La implementación en R se realizó con la función MCA del paquete FactoMinerR. Este paquete permite una visualización sencilla del mapa de coordenadas y añadir curvas de densidad donde se visualiza las zonas con más concentración de sujetos.

Figura 4.5: Análisis de Correspondencia Múltiple de Confianza Institucional con variables complementarias



Como se observa en la figura 4.5 el grupo de nula confianza institucional está más asociado en su preferencia a un gobierno autoritario que democrático. Recordemos que las instituciones en las que se registran nula confianza son: el Congreso de la República, los partidos políticos, el Poder Judicial y el Ministerio Público. De otro lado, se evidencia una clara relación entre quienes consideran a la corrupción como el principal problema del país y los que registran una suficiente confianza a las instituciones del Estado. Finalmente, el grupo de quienes confían poco tienden a preferir un gobierno democrático y no tener la opinión de que la corrupción sea el principal problema del país.

Capítulo 5

Conclusiones

5.1. Conclusiones

En este trabajo se ha abordado el análisis de clases latentes desde el enfoque bayesiano. La literatura actual (White y Murphy, 2014) ha implementado la estimación de los parámetros de este modelo bajo el enfoque bayesiano sólo para el caso dicotómico, pero no para el politómico. El objetivo principal de esta tesis fue implementar esta estimación bayesiana mediante el algoritmo de Gibbs. A partir del análisis del comportamiento y propiedades del modelo de clases latentes implementado se llegaron a las siguientes conclusiones.

- Los algoritmos desarrollados presentaron estimaciones similares tanto en el caso dicotómico bajo el enfoque bayesiano a través del paquete BayesLCA, como en el politómico bajo el enfoque clásico a través del paquete poLCA.
- La incorporación de la distribución a priori de Dirichlet para los parámetros de pertenencia a la clase y del ítem de respuesta permite tratar la base en su escala original e incorporar, de tenerse un mejor conocimiento sobre los datos, esta en la distribución a priori.
- El análisis de clases latentes es una herramienta estadística que permite analizar una gran cantidad de variables categóricas y dar cuenta de la existencia de un constructo inobservable directamente. Hasta el momento la literatura ha realizado un análisis factorial para abordar el concepto de Confianza Institucional. En este sentido, en el presente trabajo se buscó aproximarse a la confianza institucional en Lima Metropolitana a través de una serie de preguntas sobre cada institución en particular y observar la relación entre ellas. Dentro de los hallazgos, se identificaron a tres grupos según su nivel de confianza (Suficiente, Poca y Nada) en las instituciones del Estado. Las instituciones que registran una confianza suficiente son aquellas que tienen mayor alcance territorial y más cercanía con la población como: el Ministerio de Educación, la Defensoría del Pueblo, la SUNAT, el RENIEC, la ONPE, las Fuerzas Armadas y la Iglesia Católica. Las instituciones que registran menos confianza son las que tienen labores relacionadas con la representación política y la justicia como: los Partidos Políticos, Congreso de la República y Poder Judicial, la Comisión de Alto Nivel Anticorrupción, la Contraloría General de la República y el Ministerio Público.
- De otro lado, el grupo de individuos que registran nada confianza a la vez prefieren

a un gobierno autoritario que democrático en ciertas circunstancias, mientras que el grupo que registra poca confianza, considera que el gobierno democrático es siempre preferible.

5.2. Sugerencias para investigaciones futuras

- Una de las ventajas del enfoque bayesiano recae en la posibilidad de estimar el número óptimo de clases latentes. Esto fue mencionado por [White et al. \(2016\)](#) y sería de mucho interés el extender nuestra metodología a esta posibilidad.
- Un extensión de este trabajo podría ser implementar el análisis bayesiano de clases latentes con variables manifiestas politómicas ordinales.
- Será también interesante analizar la estimación bayesiana para variables manifiestas politómicas en las diversas extensiones del modelo de clases latentes, tales como el análisis de regresión de clases latentes o el análisis de transición de clases latentes.



Apéndice A

Código en R

- Estimación Bayesiana del Modelo de Clases Latentes con el muestrador de Gibbs para variables manifiesta politómicas
- Código : gibbsSampling

```
## >>>>> INSTALACION DE PAQUETES REQUERIDOS <<<<<<<

# install.packages("tidyr")
# install.packages("MCMCpack")
# install.packages("BayesLCA")
# install.packages("poLCA")

# >>>>> LLAMADO DE PAQUETES REQUERIDOS <<<<<<<

library(tidyr)
library(MCMCpack)
library(BayesLCA)
library(poLCA)

rm(list = ls())

# >>>>> FUNCION ALGORITMO PROPIO <<<<<<<

gibbsSampling <- function(mdatos, clases, iteraciones, quemadas){

# >>>>> CREACION DE FUNCION ZSAMP <<<<<<<

Zsamp <- function(prob, counts){
  J <- ncol(prob)
  z <- array(0, dim(prob), dimnames = list(NULL, 1:J))
  for(n in 1:nrow(prob)){
    tab <- table(sample(1:J, counts[n], replace = TRUE, prob[n, ]))
    z[n, names(tab)] <- tab
  }
}
```

```

}
z
}

# >>>>> CREACION DE FUNCION XLOGY <<<<<<

xlogy <- function(x, y){
  if(length(x) == 1) x <- rep(x, length(y))
  z <- x * log(y)
  z[y == 0] <- x[y == 0] * (-230)
  z
}

# >>>>> ASIGNACION DE PARAMETROS INICIALES <<<<<<

alpha <- 1
delta <- 1
iter <- iteraciones
burn.in <- quemadas
J <- clases
C <- length(table(mdatos[, 1]))
P <- ncol(mdatos)
M <- nrow(mdatos)
state <- FALSE

# >>>>> PREPARACION DE MATRIZ DE DATOS <<<<<<

tipo_base <- (rownames(table(mdatos[, 1]))[1] == "0" &&
rownames(table(mdatos[, 1]))[2] == "1")

if(tipo_base){
  Xdat <- data.blca(mdatos)
  X <- Xdat$data
  counts.n <- Xdat$counts.n
  counts.h <- names(counts.n)
} else {
  data <- cbind(mdatos, unite(mdatos, Patrones, c(1:ncol(mdatos))), sep = ""
, remove = TRUE))
  counts.n <- table(data[, P + 1])
  counts.h <- names(counts.n)
  X <- matrix(0, length(counts.h), P)
  for(l in 1:length(counts.h)){

```

```

    for(m in 1:M){
      if(data[m, P + 1] == counts.h[1]){
        for(p in 1:P) X[l, p] <- data[m, p]
      }
    }
  }
  counts.n <- as.vector(counts.n)
}

N <- nrow(X)

# >>>>> CONFIGURACION DE PARAMETROS Y MATRIZ DE DATOS <<<

if(tipo_base){
  state <- TRUE
  alpha <- matrix(alpha, J, P)
  beta <- 1
  beta <- matrix(beta, J, P)
}

if (state == FALSE){
  Xdesc <- array(0, c(N, P, C))
  for (c in 1:C){
    Xdesc[, , c] <- matrix(X == c, N, P)
    Xdesc[, , c][Xdesc[, , c] == "true"] <- 1
    Xdesc[, , c][Xdesc[, , c] == "false"] <- 0
  }
  alpha <- array(alpha, c(J, P, C))
}

delta <- rep(delta, J)

# >> SIMULACION DE VALORES INICIALES DE PARAMETROS <<<

eta <- rep(0, J)

if (state == FALSE){
  pi <- array(0, c(J, P, C))
} else {
  pi <- array(0, c(J, P))
}

```

```

Z <- unMAP(sample(1:J, size = N, replace = TRUE))

Z_init <- Z

eta <- rdirichlet(1, delta + colSums(Z))

for(j in 1:J){
  if(state == FALSE){
    for(p in 1:P) pi[j, p, ] <- rdirichlet(1, t(alpha[j, p, ] +
      colSums(Z[ , j] * Xdesc[, p, ]))
    } else {
      pi[j, ] <- rbeta(P, alpha + colSums(Z[ , j] * X), beta +
        colSums(Z[ , j] * (1 - X)))
    }
  }
}

pi_init <- pi

# >>ELEMENTOS Y CONTENEDORES PARA SECUENCIA DE MUESTREO <<<<<<

maxiter <- iter
hist <- maxiter
contador <- 1

W <- matrix(nrow = N, ncol = J)
labelstore <- matrix(NA, hist, J)
logpost.store <- rep(NA, hist)

eta_store <- matrix(NA, hist, J)

if (state == FALSE){
  pi_store <- list()
  for(count in 1:hist) pi_store[[count]] <- array(NA, c(J, P, C))
} else {
  pi_store <- array(NA, c(hist, J, P))
}

Z_store <- matrix(0, N, J)

# >>>>> INICIO DE SECUENCIA DE MUESTREO <<<<<<

cat("Inicio del muestreo. Comenzando el quemado.\n")

```



```

for (iter in 1:(maxiter + burn.in)){
  if (state == FALSE){
    for(j in 1:J){
      W_aux <- matrix(1, P, N)
      for(c in 1:C) W_aux <- W_aux * pi[j, , c] ^ t(Xdesc[ , , c])
      W[, j] <- eta[j] * apply(W_aux, 2, prod)
    }
  } else {
    for(j in 1:J){
      W[ , j] <- eta[j] * apply(pi[j, ] ^ t(X) * (1 - pi[j, ]) ^ t(1 - X), 2, prod)
    }
  }

  Z <- Zsamp(W, counts.n)

  if(iter == burn.in + 1){
    cat("Quemado completo.\n")
    Zmatch <- counts.n * Z
  }

  eta <- rdirichlet(1, delta + colSums(Z))

  for(j in 1:J){
    if (state == FALSE){
      for(p in 1:P) pi[j, p, ] <- rdirichlet(1, t(alpha[j, p, ]) +
        colSums(Z[ , j] * Xdesc[, p, ]))
    } else {
      pi[j, ] <- rbeta(P, alpha + colSums(Z[ , j] * X), beta +
        colSums(Z[ , j] * (1 - X)))
    }
  }
}

if (iter > burn.in){
  if((iter - burn.in)%100 == 0)
    cat(iter-burn.in, "de", maxiter, "muestras completas.\n")

  match1 <- matchClasses(t(Z)%*%Zmatch, method="exact", verbose = FALSE)
  labelstore[contador, ] <- match1 # match1 = ZPRIVATE

  eta_store[contador, labelstore[contador, ]] <- eta
}

```

```

    if (state == FALSE){
      pi_store[[contador]][labelstore[contador, ], , ] <- pi
    } else {
      pi_store[contador, labelstore[contador, ], , ] <- pi
    }

    Z_store <- ((iter - burn.in - 1) * Z_store + Z[ ,
labelstore[contador, ]])/(iter - burn.in)

    if (state == FALSE){
      sum.xlogy <- 0
      for(c in 1:C) sum.xlogy <- sum.xlogy + xlogy(alpha[ , , c] - 1,
pi[ , , c])
      logpost.store[contador] <- (-2)*sum(log(rowSums(W)) * counts.n)
    } else {
      logpost.store[contador] <- (-2)*sum(log(rowSums(W)) * counts.n)
    }

    contador<- contador + 1

    if (contador > hist){
      cat("Muestreo completo.\n")
      break
    }
  }
}

# >>>>> CALCULO DE RESULTADOS <<<<<<

eta <- apply(eta_store, 2, mean)

if(state == FALSE){
  for(c in 1:C){
    for(k in 1:hist) pi[ , , c] <- pi[ , , c] + pi_store[[k]][ , , c]
    pi[ , , c] <- pi[ , , c] / hist
  }
} else {
  pi <- apply(pi_store, c(2, 3), mean)
}

x <- NULL
o <- order(eta, decreasing = TRUE)

```

```

x$classprob <- eta[o]

if(state == FALSE){
  x$itemprob <- pi[o, , ]
} else {
  x$itemprob <- pi[o, ]
}

if(J == 1){
  if(state == FALSE){
    x$itemprob <- array(x$itemprob, c(J, P, C))
  } else {
    x$itemprob <- matrix(x$itemprob, J, P)
  }
}

if (state == FALSE){
  dum <- array(NA, dim = c(P, N, J))
  for(j in 1:J){
    dum_aux <- matrix(1, P, N)
    for(c in 1:C) dum_aux <- dum_aux * x$itemprob[j, , c] ^
      t(Xdesc[ , , c])
    dum[ , ,j] <- dum_aux
  }
} else {
  dum <- array(apply(x$itemprob, 1, dbinom, size = 1, x = t(X)),
    dim = c(P, N, J))
}

Z1 <- t(x$classprob * t(apply(dum, c(2, 3), prod)))
if (state == FALSE){
  sum.xlogy <- 0
  for(c in 1:C) sum.xlogy <- sum.xlogy + xlogy(alpha[ , , c] - 1,
    x$itemprob[ , , c])
  x$logpost <- (-2)*sum(log(rowSums(W)) * counts.n)
} else {
  x$logpost <- (-2)*sum(log(rowSums(W)) * counts.n)
}

x$Z <- (Z_store / counts.n)[ , o]
if(J == 1) x$Z <- matrix(x$Z, N, J)

```

```

rownames(x$Z) <- counts.h
colnames(x$Z) <- paste("Clase", 1:J)

x$mclassprob <- eta_store[ ,o]

if (state == FALSE){
  pi_store_aux <- list()
  for (k in 1:hist){
    pi_store_aux[[k]] <- pi_store[[k]][o , , ]
  }
  x$mitemprob <- pi_store_aux
} else {
  x$mitemprob <- pi_store[ , o, ]
}

Dbar <- mean(logpost.store)
Dhat <- log(mean(x$logpost))
S2 <- var(logpost.store)
x$DIC <- Dbar - Dhat

# >>>>> MATRIZ X EXTENDIDA <<<<<<

Z_dist <- round(x$Z * counts.n)
X_ext <- matrix(NA, dim(mdatos)[1], P + 1)
iterador <- 1
for(i in 1:dim(Z_dist)[1]){
  for(j in 1:J){
    for(k in 1:Z_dist[i, j]){
      if(iterador <= dim(mdatos)[1]) X_ext[iterador, ] <- c(j, X[i, ])
      iterador <- iterador + 1
    }
  }
}

colnames(X_ext) <- c("clase", colnames(mdatos))
x$dataext <- X_ext

x$Z_init <- Z_init
x$pi_init <- pi_init
x$state <- state
x$hist <- hist
x$J <- J
x$P <- P

```

```
x$C <- C
x$N <- N
x$counts.n <- counts.n
x$counts.h <- counts.h

# >>>>> VALOR RETORNADO <<<<<<

x
}

# >>>>> IMPORTACION DE BD ALZHEIMER <<<<<<

# data(Alzheimer)
# alzh = Alzheimer
# MyData = alzh

# >>>>> IMPORTACION DE BD DICOTOMICA <<<<<<

# >>>>> IMPORTACION DE BD DICOTOMICA <<<<<<

dico = read.csv("confianzadicoLM.csv",header = TRUE, sep = ",")
MyData = dico[-1]

# >>>>> IMPORTACION DE BD POLITOMICA <<<<<<

# poli = read.csv("confianzaLM.csv", header = TRUE, sep = ",")
# MyData = poli[-1]

# >>>>> IMPORTACION DE SEGURIDAD <<<<<<

# seguridad = read.csv("seguridadLM.csv", header = TRUE, sep = ",")
# MyData = seguridad[-1]

# >>>>> ENLACE DE BD MYDATA <<<<<<

attach(MyData)

# >>>>> ENLACE DE BD MYDATA <<<<<<

# >>>>> CORRIDA DE FUNCION CON MyData <<<<<<
```

```

blca <- gibbsSampling(MyData, 3, 5000, 100)

# >>>>> FUNCIONES BLCA.GIBBS Y POLCA <<<<<<

if (blca$state == TRUE){
  blca_sample <- blca.gibbs(MyData, 3, alpha = 1, beta = 1,
    delta = 1, start.vals = blca$Z_init, iter = 5000, burn.in = 1000)
} else {
  pi_probs <- array(0,c(blca$J, blca$C, blca$P))

  for(j in 1:blca$J){
    for(c in 1:blca$C){
      for(p in 1:blca$P){
        pi_probs[j, c, p] <- blca$pi_init[j, p, c]
      }
    }
  }

  start_val <- list()

  for(p in 1:blca$P) start_val[[p]] <- pi_probs[, , p]

  f <- cbind(P1.01, P1.02, P1.03, P1.04, P1.05, P1.06, P1.07,
    P1.08, P1.09, P1.10, P1.11, P1.12, P1.13, P1.14,
    P1.15, P1.16, P1.17, P1.18, P1.19, P1.20, P1.21)~1

  polca <- polCA(f, MyData, nclass = blca$J, maxiter = 100,
    probs.start = start_val)

  polca_pr <- array(0, dim(blca$pi_init))

  for(j in 1:blca$J){
    for(c in 1:blca$C){
      for(p in 1:blca$P){
        polca_pr[j, p, c] <- polca$probs[[p]][j, c]
      }
    }
  }

  polca_o <- order(polca$P, decreasing = TRUE)
  polca_sample <- NULL
  polca_sample$classprob <- polca$P[polca_o]

```

```

    polca_sample$itemprob <- polca_pr[polca_o, , ]
    polca_sample$AIC <- polca$aic
    polca_sample$BIC <- polca$bic
}

### PROBABILIDADES DE LOS PARAMETROS DEL ALGORITMO
blca$itemprob
blca$classprob
blca$logpost
blca$DIC

blca_sample$itemprob
blca_sample$classprob
blca_sample$logpost
blca_sample$DIC

##Guardar la variables de la clase latente

blcapoli=as.data.frame(blca$dataext)
write.csv(blcapoli,"CONFIANZAPOLI.csv")

#variable clase
blcadico=as.data.frame(blca$dataext)
blcadico[1]
write.csv(blcadico,"blca_3_dico.csv")
polca_sample$itemprob
polca_sample$classprob
polca_sample$logpost
polca_sample$AIC
polca_sample$BIC

# install.packages("ggplot")
# install.packages("gridExtra")

library(ggplot2)
library(gridExtra)

# >>>>> FUNCION GRAFICOS 3 CLASES <<<<<<

graficar_clases <- function(clases, hist, z, colores, index_columnas,
etiqueta_x, etiqueta_y, titulo){

```

```

lista_parametros <- as.data.frame(clases)
str(lista_parametros)
historial <- c(1:hist)
nombres_clases <- colnames(z)
colores_clases <- colores

grafico <- ggplot(data = lista_parametros, aes(x = historial)) +
  geom_line(aes(y = lista_parametros[ , index_columnas[1]],
  colour = nombres_clases[1])) +
  geom_line(aes(y = lista_parametros[ , index_columnas[2]],
  colour = nombres_clases[2])) +
  geom_line(aes(y = lista_parametros[ , index_columnas[3]],
  colour = nombres_clases[3])) +
  scale_colour_manual("", breaks = nombres_clases,
  values = colores_clases) +
  xlab(etiqueta_x) +
  scale_y_continuous(etiqueta_y, limits = c(0, 1)) +
  labs(title = titulo)
grafico
}

conversion_matriz_pi <- function(mitemprob, state, hist, J, P, C ){
  if (state == FALSE){
    matriz_pi <- array(0, c(hist, J * P, C))
  } else {
    matriz_pi <- matrix(0, hist, P * J)
  }

  i <- 1
  for(k in 1:hist){
    for(p in 1:P){
      for(j in 1:J){
        if (state == FALSE){
          for(c in 1:C) {
            matriz_pi[k, i, c] <- mitemprob[[k]][j, p, c]
            i <- i + 1
          }
        } else {
          matriz_pi[k, i] <- mitemprob[k, j, p]
          i <- i + 1
        }
      }
    }
  }
}

```



```

    }
    i <- 1
  }
  matriz_pi
}

# >>>>> GRAFICOS PARA BLCA <<<<<<

# CLASSPROB
png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS
BAYES DICO/ETA_blca_dico.png",width = 15,    # Anchura
    height = 6,    # Altura
    res= 72,      # Resolucion 72ppi es un estandar
    units = "cm")
graficar_clases(blca$mclassprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(1, 2, 3), "iteraciones", "Parametro eta", "Clases")
dev.off()
# ITEMPROB

blca_mitemprob <- conversion_matriz_pi(blca$mitemprob, blca$state,
    blca$hist, blca$J, blca$P, blca$C)

png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS
BAYES DICO/PI_blca_dico_1.png",width = 20,    # Anchura
    height = 20,    # Altura
    res= 72,      # Resolucion 72ppi es un estandar
    units = "cm")
blca_mitemprob <- conversion_matriz_pi(blca$samples$mitemprob, blca$state
, blca$hist, blca$J, blca$P, blca$C)

blca_a <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(1, 2, 3), "iteraciones", "Parametro pi", "JNE")
blca_b <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(4, 5, 6), "iteraciones", "Parametro pi", "ONPE")
blca_c <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(7, 8, 9), "iteraciones", "Parametro pi", "RENIEC")
blca_d <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(10, 11, 12), "iteraciones", "Parametro pi", "MP")
blca_e <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(13, 14, 15), "iteraciones", "Parametro pi", "MD")

```

```

blca_f <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(16, 17, 18), "iteraciones", "Parametro pi", "PNP")

grid.arrange(blca_a, blca_b, blca_c, blca_d, blca_e, blca_f)
dev.off()

png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS BAYES DICO
/PI_blca_dico_2.png",width = 20,    # Anchura
    height = 20,    # Altura
    res= 72,    # Resolucion 72ppi es un estandar
    units = "cm")
blca_mitemprob <- conversion_matriz_pi(blca$samples$itemprob,
    blca$state, blca$hist, blca$J, blca$P, blca$C)

blca_a <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(19, 20, 21), "iteraciones", "Parametro pi", "FFAA")
blca_b <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(22, 23, 24), "iteraciones", "Parametro pi", "GR")
blca_c <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(25, 26, 27), "iteraciones", "Parametro pi", "PJ")
blca_d <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(28, 29, 30), "iteraciones", "Parametro pi", "MINEDU")
blca_e <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(31, 32, 33), "iteraciones", "Parametro pi", "DP")
blca_f <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(34, 35, 36), "iteraciones", "Parametro pi", "CR")

grid.arrange(blca_a, blca_b, blca_c, blca_d, blca_e, blca_f)
dev.off()

png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS BAYES DICO/
PI_blca_dico_3.png",width = 20,    # Anchura
    height = 20,    # Altura
    res= 72,    # Resolucion 72ppi es un estandar
    units = "cm")
blca_mitemprob <- conversion_matriz_pi(blca$samples$itemprob, blca$state,
blca$hist, blca$J, blca$P, blca$C)

blca_a <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(37, 38, 39), "iteraciones", "Parametro pi", "PP")
blca_b <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(40, 41, 42), "iteraciones", "Parametro pi", "PE")

```

```

blca_c <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(43, 44, 45), "iteraciones", "Parametro pi", "RTV")
blca_d <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(46, 47, 48), "iteraciones", "Parametro pi", "IC")
blca_e <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(49, 50, 51), "iteraciones", "Parametro pi", "PA")
blca_f <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(52, 53, 54), "iteraciones", "Parametro pi", "MPU")
grid.arrange(blca_a, blca_b, blca_c, blca_d, blca_e, blca_f)
dev.off()

png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/
FIGURAS BAYES DICO/PI_blca_dico_4.png",width = 20,      # Anchura
      height = 20,      # Altura
      res= 72,          # Resolucion 72ppi es un estandar
      units = "cm")
blca_mitemprob <- conversion_matriz_pi(blca$samples$itemprob,
blca$state, blca$hist, blca$J, blca$P, blca$C)

blca_a <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(55, 56, 57), "iteraciones", "Parametro pi", "CGR")
blca_b <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(58, 59, 60), "iteraciones", "Parametro pi", "SUNAR")
blca_c <- graficar_clases(blca_mitemprob, blca$hist, blca$Z,
c("red", "blue", "green"), c(61, 62, 63), "iteraciones", "Parametro pi", "CANA")
grid.arrange(blca_a, blca_b, blca_c)
dev.off()

# >>>>> GRAFICOS PARA BLCA_SAMPLE <<<<<<

# CLASSPROB
png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS BAYES DICO
/ETA_blca_dico_sample.png",width = 15,      # Anchura
      height = 6,      # Altura
      res= 72,          # Resolucion 72ppi es un estandar
      units = "cm")
graficar_clases(blca_sample$samples$classprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(1, 2, 3), "iteraciones", "Parametro eta", "Clases")
dev.off()

# ITEMPROB

```

```

png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS BAYES DICO/
PI_blca_dico_sample.png",width = 20,    # Anchura
    height = 20,    # Altura
    res= 72,      # Resolucion 72ppi es un estandar
    units = "cm")
blca_sample_mitemprob <- conversion_matriz_pi(blca_sample$samples$itemprob,
blca$state, blca$hist, blca$J, blca$P, blca$C)

blca_sample_a <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(1, 2, 3), "iteraciones", "Parametro pi", "JNE")
blca_sample_b <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(4, 5, 6), "iteraciones", "Parametro pi", "ONPE")
blca_sample_c <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(7, 8, 9), "iteraciones", "Parametro pi", "RENIEC")
blca_sample_d <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(10, 11, 12), "iteraciones", "Parametro pi", "MP")
blca_sample_e <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(13, 14, 15), "iteraciones", "Parametro pi", "MD")
blca_sample_f <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(16, 17, 18), "iteraciones", "Parametro pi", "PNP")

grid.arrange(blca_sample_a, blca_sample_b, blca_sample_c, blca_sample_d,
blca_sample_e, blca_sample_f)
dev.off()

png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS BAYES DICO/
PI_blca_dico_sample2.png",width = 20,    # Anchura
    height = 20,    # Altura
    res= 72,      # Resolucion 72ppi es un estandar
    units = "cm")
blca_sample_mitemprob <- conversion_matriz_pi(blca_sample$samples$itemprob,
blca$state, blca$hist, blca$J, blca$P, blca$C)

blca_sample_a <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(19, 20, 21), "iteraciones", "Parametro pi", "FFAA")
blca_sample_b <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(22, 23, 24), "iteraciones", "Parametro pi", "GR")
blca_sample_c <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(25, 26, 27), "iteraciones", "Parametro pi", "PJ")
blca_sample_d <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(28, 29, 30), "iteraciones", "Parametro pi", "MINEDU")
blca_sample_e <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,

```

```

c("red", "blue", "green"), c(31, 32, 33), "iteraciones", "Parametro pi", "DP")
blca_sample_f <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(34, 35, 36), "iteraciones", "Parametro pi", "CR")

grid.arrange(blca_sample_a, blca_sample_b, blca_sample_c, blca_sample_d,
blca_sample_e, blca_sample_f)
dev.off()

```

```

png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS BAYES DICO/
PI_blca_dico_sample3.png",width = 20, # Anchura
  height = 20, # Altura
  res= 72, # Resolucion 72ppi es un estandar
  units = "cm")
blca_sample_mitemprob <- conversion_matriz_pi(blca_sample$samples$mitemprob,
blca$state, blca$hist, blca$J, blca$P, blca$C)

```

```

blca_sample_a <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(37, 38, 39), "iteraciones", "Parametro pi", "PP")
blca_sample_b <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(40, 41, 42), "iteraciones", "Parametro pi", "PE")
blca_sample_c <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(43, 44, 45), "iteraciones", "Parametro pi", "RTV")
blca_sample_d <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(46, 47, 48), "iteraciones", "Parametro pi", "IC")
blca_sample_e <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(49, 50, 51), "iteraciones", "Parametro pi", "PA")
blca_sample_f <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(52, 53, 54), "iteraciones", "Parametro pi", "MPU")
grid.arrange(blca_sample_a, blca_sample_b, blca_sample_c, blca_sample_d,
blca_sample_e, blca_sample_f)
dev.off()

```

```

png("~/Dropbox/Maestría Tesis/Tesis-MCRUZ 2018/FIGURAS BAYES DICO/
PI_blca_dico_sample4.png",width = 20, # Anchura
  height = 20, # Altura
  res= 72, # Resolucion 72ppi es un estandar
  units = "cm")
blca_sample_mitemprob <- conversion_matriz_pi(blca_sample$samples$mitemprob,
blca$state, blca$hist, blca$J, blca$P, blca$C)

```

```

blca_sample_a <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(55, 56, 57), "iteraciones", "Parametro pi", "CGR")

```

```
blca_sample_b <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(58, 59, 60), "iteraciones", "Parametro pi", "SUNAR")
blca_sample_c <- graficar_clases(blca_sample_mitemprob, blca$hist, blca_sample$Z,
c("red", "blue", "green"), c(61, 62, 63), "iteraciones", "Parametro pi", "CANA")
grid.arrange(blca_sample_a, blca_sample_b, blca_sample_c)
dev.off()
```

####ANALISIS DE CORRESPONDENCIA

```
poli = read.csv("CONFIANZAPOLI.csv", header = TRUE, sep = ",")
```

```
#Abir la base de la ENAHO
```

```
library(foreign)
```

```
data=read.spss("Enaho01B-2017-1.sav",to.data.frame=TRUE)
```

```
data=data[data$DOMINO==Lima Metropolitana,]
```

```
data$identificador=paste(data$MES,data$NCONGLOME,data$VIVIENDA,data$HOGAR,data$CODPERSO,d
```

```
seleccion=data[,c(12:32,202:2013,193)]
```

```
rownames(seleccion)=data$identificacion
```

```
ENAHO=merge(poli,enaho)
```

```
col_names <- names(ENAHO)
```

```
ENAHO[,col_names] <- lapply(ENAHO[,col_names] , factor)
```

```
install.packages("factoextra","FactoMineR")
```

```
library(factoextra)
```

```
library(FactoMineR)
```

```
#3 P2_1.01 P2_1.15 P22B1 P6 P7
```

```
names(dataca4)
```

```
dataca2=dataca[,c(25,79,93,239)]
```

```
dataca2=dataca4[,c(25,79,239)]
```

```
str(dataca2)
```

```
dataca2$P2_1.01=factor(dataca2$P2_1.01,labels = c("No Corrupción", "Corrupción"))
```

```
dataca2$P2_1.15=factor(dataca2$P2_1.15,labels=c("No Delincuencia", "Delincuencia"))
```

```
dataca2$P22B1=factor(dataca2$P22B1,labels = c("Inseguro", "Seguro"))
```

```
dataca2$clasepoli=as.factor(dataca2$clasepoli)
```

```
dataca2$clasepoli=factor(dataca2$clasepoli,labels = c("Poca Confianza",
"Nada de Confianza","Confianza Suficiente"))
```

```
dataca2$clasedico=as.factor(dataca2$clasedico)
```

```
colnames(dataca2)[1]="Confianza Institucional"
```

```
colnames(dataca2)[2]="Corrupcion"
```

```
colnames(dataca2)[3]="Delincuencia"
```

```
colnames(dataca2)[4]="Inseguridad"
cats = apply(dataca2, 2, function(x) nlevels(as.factor(x)))
res.mca <- MCA(dataca2, graph = F)
mca1_vars_df = data.frame(res.mca$var$coord, Variable = rep(names(cats), cats))
# data frame with observation coordinates
mca1_obs_df = data.frame(res.mca$ind$coord)
res.mca$eig
res.mca$var
# plot of variable categories

png("C:/Users/MARYLIA/Dropbox/Maestria Tesis/Tesis-MCRUZ 2018/FIGURAS
BAYES POLI/PCM4.png",width = 20, # Anchura
    height = 15, # Altura
    res= 72, # Resolucion 72ppi es un estandar
    units = "cm")

ggplot(data = mca1_obs_df, aes(x = Dim.1, y = Dim.2)) +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(colour = "gray50", alpha = 0.7) +
  geom_density2d(colour = "gray80") +
  geom_text(data = mca1_vars_df,
            aes(x = Dim.1, y = Dim.2,
                label = rownames(mca1_vars_df), colour = Variable)) +
  ggtitle("MCA confianza institucional y seguridad y corrupcion") +
  scale_colour_discrete(name = "Variable")
dev.off()
```

Bibliografía

- Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*, Springer New York, pp. 199–213.
- Almond, G. y Verba, S. (1963). *The Civic Culture: Political Attitudes and Democracy in Five Nations*, Princeton University Press.
- Balakrishnan, N. (2003). *A Primer on Statistical Distributions*, John Wiley Sons, chapter 27. Dirichlet Distribution, p. 274.
- Bazan, J. y Bayes, C. (2010). Inferencia bayesiana en modelos de regresión binaria usando brmuw, *Serie B 25*, Departamento de Ciencias PUCP.
- Bhatnagar, A. y Ghose, S. (2004). A latent class segmentation analysis of e-shoppers, *Journal of Business Research* **57**(7): 758–767.
- Blaydes, L. y Linzer, D. A. (2008). The political economy of women’s support for fundamentalist islam, *World Politics* **60**(4): 576–609.
- Bollen, K. A. (2002). Latent variables in psychology, *Annual Review of Psychology* **53**(605-634).
- Breen, R. (2000). Why is support for extreme parties underestimated by surveys? a latent class analysis, *British Journal of Political Science* **30**(2): 375–382.
- Carrión, J., Zárate, P. y Zeichmeister, E. (2015). *Cultura política de la democracia en Perú y en las Américas, 2014: Gobernalidad democrática a través de 10 años del Barómetro de las Américas*, Agency for International Development (USAID), Instituto de Estudios Peruanos (IEP), Proyecto de Opinión Pública de América Latina (LAPOP).
- Celeux, G., Forbes, F., Robert, C. y Titterton, D. (2006). Deviance information criteria for missing data models, *Bayesian Analysis* **1**(4): 651–674.
- Chang, E. C. C. y Chu, Y. (2006). Corruption and trust: Exceptionalism in asian democracies?, *The Journal of Politics* **68**(2): 259–271.
- Chen, M., Kuo, L. y Lewis, P. O. (2014). *Bayesian Phylogenetics: Methods, Algorithms, and Applications*, first edn, Chapman Hall/CRC Mathematical and Computational Biology.
- Chung, H., Flaherty, B. P. y Schafer, J. L. (2006). Latent class logistic regression: application to marijuana use and attitudes among high school seniors, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**(4): 723–743.
- Clogg, C. C. (1995). *Handbook of statistical modeling for the social and behavioral sciences*, New York: Plenum, chapter Latent class models.
- Cohaila, E. (2015). *La construcción de la confianza en las instituciones políticas: el caso de los distritos de San Martín de Porres y Los Olivos*, Tesis en Doctorado en Ciencia Política. Pontificia Universidad Católica del Perú. Escuela de Posgrado.

- Collins, L. y S., L. (2010). *Latent Class and Latent Transition Analysis*, Wiley, New Jersey.
- Crête, J., Pelletier, R. y Couture, J. (2006). *Political Trust in Canada*, Canadian Political Science Association, Department of Political Science Laval University, New York University, Toronto, Ontario.
- Dempster, A., Laird, N. y Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (Series B. 39): 1–38.
- Drew, A. L. y Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis, *Journal of Statistical Software, Articles* **42**(10).
- ESE (2018). Encuesta social europea.
URL: <https://www.upf.edu/ess/>
- German, S. y German, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6): 721–741.
- Gerring, J. (2017). *Applied Social Science Methodology: An Introductory Guide*, Cambridge University Press.
- Gill, J. (2014). *Bayesian Methods: A Social and Behavioral Sciences Approach*, Statistics in the Social and Behavioral Sciences Series, third edition edn, CRC Press.
- Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika* (61): 215–231.
- Gordon, S. (2005). Confianza, capital social y desempeño de organizaciones, *Revista Mexicana de ciencias políticas y sociales XLVII*(193): 41–55.
- Green, B. (1951). A general solution for the latent class model of latent structure analysis, *Psychometrika* (47): 71–76.
- Hagenaars, J. A. (1993). *Loglinear models with latent variables*, Thousand Oaks, CA: Sage Publications.
- Hagenaars, J. A. y Halman, L. C. (1989). Searching for ideal types: the potentialities of latent class analysis, *European Sociological Review. Oxford University Press* **5**(1): 81–96.
- Hakhverdian, A. y Mayne, Q. (2012). Institutional trust, education, and corruption: A micro-macro interactive institutional trust, education and corruption: A micro-macro interactive approach, *Journal of Politics* **74**(3): 739–750.
- Hardin, R. (1998). *Trust and Government*, Russell Sage Foundation, chapter Trust in Government.
- Hiskey, J. y Seligson, M. (2003). Pitfalls of power to the people: Decentralization, local government performance, and system support in bolivia, *Studies in Comparative International Development* **37**(4): 64–88.
- INEI (2018). Encuesta nacional de hogares enaho.
URL: <http://inei.inei.gob.pe/microdatos/>
- Inglehart, R. (1988). The renaissance of political culture, *The American Political Science Review* **82**(4): 1203–1230.

- Kestila-Kekkonen, E. y Soderlund, P. (2016). Political trust, individual-level characteristic and institutional performance: Evidence from finland 2014-13, *Nordic Political Science Association* **39**(2).
- LAPOP (2018). Latinobarómetro de las américas.
URL: <https://www.vanderbilt.edu/lapop/>
- Lazarsfeld, P. (1950). *Measurement and Prediction: Studies in Social Psychology in World War II*, Vol. IV, Princeton University Press., chapter The Logical and mathematical foundation of latent structure analysis.
- Lazarsfeld, P. y Henry, N. (1968). *Latent structure Analysis*, Houghton Mifflin, Boston.
- Levi, M. (1998). *A State of Trust*, Russell Sage Foundation, chapter Trust in Government.
- Levi, M. y Stoker, L. (2000). Political trust and trustworthiness, *Annual Review of Political Science* **3**: 475–507.
- McCutcheon, A. (1985). A latent class analysis of tolerance for nonconformity in the american public, *The Public Opinion Quarterly* **49**(4): 474–488.
- McCutcheon, A. (1987). *Latent Class Analysis*, Beverly Hills: Sage Publications.
- Mishler, W. y Rose, R. (2001). What are the origins of political trust?, *Comparative Political Studies* **34**(1): 20–62.
- Morales, M. (2008). Evaluando la confianza institucional en chile. una mirada desde los resultados lapop, *Revista de Ciencia Política* **28**(2): 161–186.
- Moran, M., Walsh, C., Lynch, A., Coen, R. F., Coakley, D. y Lawlor, B. A. (2004). Syndromes of behavioural and psychological symptoms in mild alzheimer's disease, *International Journal of Geriatric Psychiatry* **19**(4): 359–364.
- Papastamoulis, P. (2015). label.switching: An r packages for dealing with the labels switching problem in mcmc outputs, *Journal of Statistical Software* **VV**.
- Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M. y Karagiannidou, M. (2016). *World Alzheimer Report 2016*, Alzheimer's Disease International (ADI).
- Raftery, A., Newton, M. A., Satagopan, J. M. y Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity, in M. J. B. J. M. Bernardo y J. O. Berger (eds), *Bayesian Statistics*, Oxford: Oxford University Press, pp. 1–45.
- Ruscio, J. y Ruscio, A. M. (2008). Advancing psychological science through the study of latent structure, *Current Directions in Psychological Science* **17**(203-207).
- Thomas, T. D., Abts, K., Stroeken, K. y Weyden, P. V. (2015). Measuring institutional trust evidence from gurana, *Journal of Politics in Latin America* **3**: 85–115.
- Walsh, C. D. (2006). Latent class analysis identification of syndromes in alzheimer's disease: A bayesian approach, *Metodološki zvezki* **3**(1): 147–162.
- White, A. J., Wyse, J. y Murphy, T. B. (2016). Bayesian variable selection for latent class analysis using a collapsed gibbs sampler, *Statistics and Computing* **26**(1-2): 511–527.
URL: <http://dx.doi.org/10.1007/s11222-014-9542-5>

White, A. y Murphy, T. B. (2014). BayesLCA: An R package for Bayesian latent class analysis, *Journal of Statistical Software* **61**(13).

URL: <http://www.jstatsoft.org/v61/i13>

Wiener, L. (2015). *Modelo de regresión de clases latentes: factores asociados a la valoración de una universidad privada*, Maestría en Estadística. Pontificia Universidad Católica del Perú. Escuela de Posgrado.

WVS (2018). World value survey.

URL: <http://www.worldvaluessurvey.org/wvs.jsp>

Yang, P. (2007). Government behavior and trust: The case of china, *Cato Journal* **27**(3).

