ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

INFERENCIA VARIACIONAL EN MODELOS GRÁFICOS PROBABILÍSTICOS PARA EL ESTUDIO DE LA RELACIÓN ENTRE ENFERMEDADES Y SÍNTOMAS

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERA MATEMÁTICA

KAREN VANESSA SALAZAR CAIZA

karen.salazar@epn.edu.ec

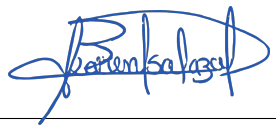
DIRECTOR: CARLOS ALBERTO ALMEIDA RODRÍGUEZ

carlos.almeidar@epn.edu.ec

DMQ, SEPTIEMBRE 2022

CERTIFICACIONES

Yo, KAREN VANESSA SALAZAR CAIZA, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.



Karen Vanessa Salazar Caiza

Certifico que el presente trabajo de integración curricular fue desarrollado por Karen Vanessa Salazar Caiza, bajo mi supervisión.

Carlos Alberto Almeida Rodríguez

DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

Karen Vanessa Salazar Caiza

Carlos Alberto Almeida Rodríguez

RESUMEN

Los problemas de diagnóstico médico en la atención primaria, donde se debe plantear la relación que tienen las enfermedades y los síntomas implica asumir una suma significativa de incertidumbre y una gran cantidad de variables inmersas. En escenarios en los que intervienen estos dos componentes, los modelos gráficos son una alternativa, ya que, al combinar la teoría de probabilidades y la teoría de grafos se convierte en un fuerte modelado estadístico multivariante. Por lo que en este trabajo plantearemos un modelo gráfico probabilista para las enfermedades: tuberculosis, neumonía, alergia, asma bronquial y gripe común, y sus respectivos síntomas.

Los modelos gráficos juegan un papel importante al explorar distribuciones de probabilidad en variables interrelacionadas, a partir de una serie de datos; sin embargo, debido a que existe la dificultad al calcular una distribución de probabilidad condicional sobre las variables latentes, teniendo en cuenta la evidencia, en nuestro caso las enfermedades dados los síntomas introduciremos la inferencia variacional que es parte de los métodos que acerca las densidades a través de la optimización, proporcionando aproximaciones a las probabilidades marginales y condicionales.

Palabras clave: Inferencia Variacional, Modelo gráfico probalístico, ELBO, CAVI, Mean-Field, diagnóstico, enfermedades, síntomas.

ABSTRACT

Medical diagnosis problems in primary care, where the relationship between diseases and symptoms must be considered, entails taking a significant amount of uncertainty and a large number of variables involved. In scenarios involving these two components, graphical models are an excellent alternative, since combining probability theory and graph theory results in strong multivariate statistical modelling. Therefore, in this work we will propose a probabilistic graphic model for the diseases: tuberculosis, pneumonia, allergy, bronchial asthma and the common flu, and their respective symptoms.

Graphical models play an important role when exploring probability distributions of interrelated variables from a series of data; however, due to the difficulty in calculating a conditional probability distribution on the latent variables, taking into account the evidence, in our case the diseases given the symptoms we will introduce the variational inference that is part of the methods that brings the densities closer to through optimization, providing approximations for the marginal and conditional probabilities.

Key words: Variational Inference, Probabilistic graphical models, ELBO, CAVI, Mean-Field, diagnosis, disease, symptoms.

Índice general

Índice de figuras

Capítulo 1

Descripción del componente desarrollado

El problema al que nos enfrentamos ha sido ampliamente estudiado por la Inferencia Bayesiana, especialmente por la Inferencia Exacta y Teoría de Grafos. En el presente trabajo abordaremos un enfoque dado por los Modelos Gráficos Probabilístas y resuelto por los Métodos Variacionales, con el fin de reducir el costo computacional.

Para alcanzar el objetivo del proyecto es necesario enfocarnos en la evidencia que es el conjunto observable de datos, es decir, en nuestro caso son los síntomas dada la enfermedad.

Crearemos un Modelo Gráfico Probabilístas a partir de los datos. Se explorará las relaciones entre las variables; es decir, las conexiones de causa y efecto, donde las causalidades están relacionadas con el conocimiento de los eventos dados.

Luego de realizar el modelo, la estrategia consistirá en definir una familia de distribuciones parametrizadas y mediante el algoritmo CAVI optimizar los parámetros para obtener el elemento más cercano al objetivo con respecto a un error bien definido y analizar la convergencia del ELBO.

1.1. Objetivo general

Resolver un problema de diagnóstico medico modelado previamente por Modelos Gráficos Probabilísticos (PGM) aplicando Inferencia variacional.

1.2. Objetivos específicos

1. Formular un modelo gráfico probabilista que explique la relación de las enfermedades: tuberculosis, neumonía, alergia, asma bronquial y gripe común, con sus respectivos síntomas.

2. Plantear distribuciones variacionales que se ajusten a los parámetros.

3. Determinar las funciones de densidad para los parámetros variacionales mediante el algoritmo CAVI.

1.3. Alcance

Determinar las distribuciones variacionales de probabilidad del modelo gráfico probabilístico propuesto, que sirvan como sustitutos para la estimación de las variables latentes, mediante el algoritmo *Coordinate Ascent Variational Inference*.

1.4. Marco teórico

Para analizar el alcance que tienen los modelos gráficos como una herramienta de modelización es importante tener en cuenta la concepción de incertidumbre. La Real Academia Española (2001), define a la incertidumbre como «Falta de certidumbre». En la vida diaria nos encontramos con situaciones de incertidumbre; intentar comprender lo que está sucediendo en un sistema donde la información es parcial. Por ejemplo, predecir la densidad del tráfico, la cotización en la bolsa de valores, etc. (Evans, 2018).

Considerando que muchos problemas y aplicaciones en el mundo indagan sobre información incompleta, compleja y ambigua, debido a la limitación de las observaciones, ya que las relaciones con frecuencia no son deterministas, por lo menos en correspondencia con la capacidad de modelamiento. Así, es coherente pensar que las observaciones obtenidas vienen mezcladas con ruido y una gran cantidad de variables inmersas. Con esto encontramos que modelar el mundo es un asunto de incertidumbre y complejidad Koller y Friedman, 2009, pg. 2.

1.4.1. Grafos

Los conceptos básicos que se presentan en esta subsección tienen como referencia general a Sucar (2021).

Definición 1.1 (Grafo)**.** Un grafo G corresponde a un par ordenado (V, E), donde V es el conjunto de vértices, también son conocidos como nodos o variables (de aquí en adelante los llamaremos nodos), y a E como el conjunto de aristas.

Definición 1.2 (Grafos dirigidos y no dirigidos)**.** Sea $G = (V, E)$ un grafo tal que $V \neq \emptyset$.

- G es un grafo no dirigido si $E \subseteq \{(A_i, A_j) \in V \times V : \ (A_i, A_j) = (A_j, A_i), \forall i, j\}$.
- G es un grafo dirigido si $E \subseteq \{(A_i, A_j) \in V \times V : \ A_i \neq A_j, \forall i, j\}$.

En grafos no dirigidos, las aristas son enlaces simples sin ningún precepto y es denotado como $A_i - A_j$; mientras que las aristas en grafos dirigidos son conocidas como arcos. Se denota $A_i \longrightarrow A_j$ al arco que se dirige desde el nodo A_i (nodo padre o nodo origen) hasta el nodo A_j (nodo sucesor) (Benferhat *et al.*, 2020).

Además de los grafos dirigidos y no dirigidos hay otro grupo de grafos.

Definición 1.3 (Tipos de grafos)**.** Tenemos los siguientes tipos de grafos:

- **Grafo cadena:** un grafo que mezcla aristas dirigidas y no dirigidas.
- **Grafo simple:** no contiene ciclos, ni arcos paralelos.
- **Grafo múltiple:** es un grafo con varios subgrafos que cada borde no tiene aristas con los otros componentes, es decir, no están conectados.
- **Grafo completo:** es un grafo que tiene aristas entre cada par de vértices.
- **Grafo bipartito:** grafo en el que los vértices se pueden dividir en dos subconjuntos, G_1 y G_2, tal que, todas las aristas conectan a un vértice de G_1 con otro vértice de G_2.
- **Grafo ponderado:** aquel grafo que tiene pesos en las aristas.

1.4.2. Modelos Gráficos Probabilísticos

De acuerdo con Koller y Friedman (2009), define a los modelos gráficos probabilísticos (PGM por sus siglas en inglés, *Probabilistic Graphical Models*) como una herramienta para auscultar un sistema en distribuciones complejas, sintetizar la información y explicarla de forma compacta. Los modelos gráficos probabilísticos emplean grafos como fundamento para estructurar una distribución compleja sobre un espacio de alta dimensión.

Además, conforme a Sucar (2021), los Modelos Gráficos Probabilísticos son una representación sólida de una distribución de probabilidad conjunta, de la cual se obtiene las probabilidades condicionales y marginales; además, facilitan el manejo de la incertidumbre apoyándose en la teoría de probabilidad de una manera computacionalmente eficaz, considerando sólo aquellas relaciones de independencia que son factibles para determinado problema e insertándolos en el modelo para reducir la complejidad en términos de memoria y costo computacional. Las correspondencias de dependencia e independencia se representan mediante grafos. Si las variables son dependientes están conectadas y la representación de independencia es tácita.

Un modelo gráfico probabilístico tiene dos aspectos:

1. Un grafo $G = (V, E)$ que define al modelo.
2. Las funciones locales, $f(Y_i)$, donde Y_i es un subconjunto de X, que define los parámetros del modelo.

La probabilidad conjunta está dada por el producto de las funciones locales antes mencionadas, esto es,

$$P(X_1, X_2, X_3, \cdots, X_n) = K\left[\prod_{i=1}^{M} f(Y_i)\right], \tag{1.1}$$

donde K es una constante de normalización (hace que las probabilidades sumen uno).

1.4.3. Redes Bayesianas

Las redes Bayesianas son modelos gráficos probabilísticos dirigidos que muestran la distribución conjunta de las variables aleatorias inmersas en el modelo.

Definición 1.4 (Red Bayesiana)**.** Una Red Bayesiana (BN por sus siglas en inglés) es una representación de las distribuciones conjuntas de n variables discretas, $X_1, X_2, \ldots, X_n$,

como un grafo dirigido acíclico y un conjunto de probabilidades condicionales. Los nodos corresponden a una variable y tiene asociada la probabilidad del estado de la variable dados sus nodos padres. La estructura del grafo expresa un conjunto de independencias condicionales para las variables del modelo.

Lozano (2011) afirma que una red bayesiana es una representación gráfica y de fácil interpretación de las dependencias para razonamiento probabilístico, en donde los nodos representan variables aleatorias y los arcos simbolizan relaciones de dependencia entre las variables.

1.4.4. Inferencia Bayesiana

Según Ryder (2021), la inferencia bayesiana permite la expresión de la incertidumbre usando una medida de creencia previa, está basada en expresar D y θ a través de la densidad de probabilidad conjunta de la forma:

$$p(D,\theta) = p(D|\theta)\, p(\theta), \tag{1.2}$$

donde θ es la variable latente sobre la cual queremos inferir (conocida también como hipótesis) y D las observaciones (datos).

El objetivo de la inferencia es calcular la distribución posterior, que está dada por el condicionamiento de (1.2) en D, este resultado es conocido como Teorema de Bayes.

Teorema 1.1 (Teorema de Bayes)**.** *Sean θ y D dos eventos tales que $P(D) \neq 0$. Entonces,*

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)}, \tag{1.3}$$

donde θ es la variable latente a inferir o hipótesis, $p(\theta)$ es la probabilidad a priori, D es la evidencia, $p(\theta|D)$ es la probabilidad a posteriori, $p(D|\theta)$ es la verosimilitud (likelihood), y $p(D)$ es la probabilidad marginal.

La probabilidad marginal $p(D)$ está dada por

$$p(D) = \int p(D|\theta) \cdot p(\theta)\, d\theta. \tag{1.4}$$

En la práctica, la integral de la ecuación (1.4) es intratable o conlleva mucho tiempo para estimarla. Para resolver este problema existen algoritmos de Inferencia exacta, como por ejemplo los publicados por Jensen (1996), Shachter *et al.* (1994) o Shenoy (1992).

Aunque estos en algunos casos nos brindan una respuesta satisfactoria, en muchos otros la complejidad de tiempo y espacio para obtener una respuesta exacta es inaceptable (Jordan *et al.*, 1999). Un recurso común para tratar de encarar estos problemas es resolverlo mediante métodos de aproximación, tales como los métodos de Montecarlo basados en cadenas de Markov (MCMC, por sus siglas en inglés). El MCMC a pesar de que nos asegura convergencia teóricamente, no resuelve el problema computacional del todo. Una metodología computacional alternativa para la inferencia estadística es la Inferencia Variacional (conocida por sus siglas en inglés como VI).

1.4.5. Inferencia Variacional

La inferencia variacional convierte el problema de inferencia en un problema de optimización; restringe la distribución de probabilidad que tratamos de encontrar a una familia de distribuciones que tiene buenas propiedades computacionales y trata de encontrar la solución analítica para la hallar distribución óptima de esa familia, esta solución analítica suele ser difícil de resolver pero podría ser simplificada a un punto donde las estrategias de optimización habituales pueden dar una buena solución (Shwe *et al.*, 1991).

La inferencia variacional nos indica que debemos tomar una familia de distribuciones Q, sobre las variables latentes y hallar una distribución $q^* \in Q$ cercana al objetivo: la distribución *a posteriori* $p(\theta|D)$. En otras palabras, en lugar de calcular directamente la probabilidad posterior, intentamos encontrar el parámetro θ de la distribución q^* que mejor nos aproxime a la distribución *a posteriori* real. La «cercanía» (proximidad) entre dos distribuciones de probabilidad se puede medir con la divergencia de Kullback-Leibler (Blei *et al.*, 2017).

Definición 1.5 (Divergencia de Kullback-Leibler)**.** Dadas q y p dos distribuciones de probabilidad continuas, la divergencia de Kullback-Leibler[1] KL se define como

$$\mathrm{KL}\left(q||p\right) = \int q(z) \ln\left[\frac{q(z)}{p(z)}\right] \mathrm{d}z. \tag{1.5}$$

Así, nuestro objetivo es encontrar el parámetro θ de una nueva distribución q^* tal que

$$q^*(\theta) = \arg\min_{\theta} \mathrm{KL}\left[q(\theta)||p(\theta|D)\right]. \tag{1.6}$$

[1] La divergencia de Kullback-Leibler no es simétrica, por ende, no es una métrica.

Intuitivamente, podemos pensar que calcular la divergencia KL es una tarea sencilla. Sin embargo, notemos que al reescribirla en función de la esperanza con respecto a $q(\theta)$, se obtiene

$$\begin{aligned}\mathrm{KL}\left[q(\theta)\,||\,p(\theta|D)\right] &= \mathbb{E}_q\left[\ln q(\theta)\right] - \mathbb{E}_q\left[\ln p(\theta|D)\right] \\ &= \mathbb{E}_q\left[\ln q(\theta)\right] - \mathbb{E}_q\left[\ln p(\theta, D)\right] + \ln p(D)\,. \end{aligned} \tag{1.7}$$

La ecuación (1.7) manifiesta que la divergencia KL depende de $\ln p(D)$, y este término puede ser intratable. Por tanto, no podemos calcularla.

Límite inferior de evidencia (ELBO)

Debido al impedimento mencionado anteriormente, debemos buscar un objetivo alterno.

Teorema 1.2 (Desigualdad de Jensen)**.** *Sean X una variable aleatoria continua y f una función convexa, entonces $f\left[\mathbb{E}(X)\right] \leq \mathbb{E}\left[f(X)\right]$.*

Apliquemos la desigualdad de Jensen al cálculo del logaritmo de la probabilidad marginal $\ln p(D)$,

$$\begin{aligned}\ln p(D) &= \ln\left[\int p(\theta, D)\ \mathrm{d}\theta\right] \\ &= \ln\left[\int p(\theta, D)\,\frac{q(\theta)}{q(\theta)}\,\mathrm{d}\theta\right] \\ &= \ln\left\{\mathbb{E}_q\left[\frac{p(\theta, D)}{q(\theta)}\right]\right\} \\ &\geq \mathbb{E}_q\left[\ln p(\theta, D)\right] - \mathbb{E}_q\left[\ln q(\theta)\right]. \end{aligned} \tag{1.8}$$

La cota inferior encontrada se conoce como el límite inferior de la evidencia (*Evidence lower bound*, abreviado como ELBO):

$$\mathrm{ELBO}\left(q\right) = \mathbb{E}_q\left[\ln p(\theta, D)\right] - \mathbb{E}_q\left[\ln q(\theta)\right]. \tag{1.9}$$

De las ecuaciones 1.7 y 1.9, notemos que

$$\begin{aligned}\mathrm{KL}\left[q(\theta)\,||\,p(\theta|D)\right] &= -\left\{\mathbb{E}_q\left[\ln p(\theta, D)\right] - \mathbb{E}_q\left[\ln q(\theta)\right]\right\} + \ln p(D) \\ &= -\mathrm{ELBO}\left(q\right) + \ln p(D)\,. \end{aligned} \tag{1.10}$$

De manera que buscar una aproximación de q que maximice el ELBO es equivalente a encontrar una aproximación de q que minimice la divergencia KL. Por ende, el ELBO es un objetivo alterno válido.

Aproximación de Campo-Medio

Una vez ya descrita la función objetivo variacional (ELBO) en el problema de optimización, es necesario caracterizar la familia variacional Q. La complejidad de la familia determinará que tan compleja será la resolución del problema.

El método de aproximación de Campo-Medio (*Mean-Field Approximation*) supone que las variables latentes son mutuamente independientes entre sí, y que cada una de ellas está gobernada por un factor distinto en la densidad variacional. Esto significa que podemos factorizar fácilmente las distribuciones variacionales en grupos, es decir,

$$q(\theta) = \prod_k q_k(\theta_k). \tag{1.11}$$

La variable latente θ_k se rige por su propio factor de variación: la densidad $q_k(\theta_k)$. En el problema de optimización, estos factores variacionales se eligen para maximizar el ELBO de la ecuación 1.9.

Inferencia variacional de ascenso de coordenadas (CAVI)

Nuestros esfuerzos se centrarán en escoger los parámetros variacionales que maximicen el ELBO. El método de ascenso de coordenadas (*Coordinate Ascent Variational Inference*, abreviado como CAVI) es un enfoque muy popular en Inferencia Variacional. El método optimiza la aproximación variacional de cada variable latente, mientras que las demás variables las mantiene fijas.

Teorema 1.3 (Ley de la esperanza total)**.** *Sean p y q variables aleatorias. Entonces,*

$$\mathbb{E}(p) = \mathbb{E}\left[\mathbb{E}(p|q)\right]. \tag{1.12}$$

Como consecuencia de las ecuaciones 1.11, 1.12 y la definición de esperanza, tenemos que

$$\text{ELBO} = \mathbb{E}_q\left[\ln p(\theta, D)\right] - \mathbb{E}_q\left[\ln q(\theta)\right]$$

$$
\begin{aligned}
&= \mathbb{E}_q \left[\ln p\left(\theta_j, \theta_{-j}, D\right) \right] - \sum_{q_i} \mathbb{E}_{q_i} \left[q_i \left(\theta_i\right) \right] \\
&= \mathbb{E}_{q_j} \left\{ \mathbb{E}_{q_{-j}} \left[\ln p\left(\theta_j, \theta_{-j}, D\right) | \theta_j \right] \right\} - \mathbb{E}_{q_j} \left(q_j\right) + \text{constante} \\
&= \mathbb{E}_{q_j} \left\{ \mathbb{E}_{q_{-j}} \left[\ln p\left(\theta_j, \theta_{-j}, D\right) \right] \right\} - \mathbb{E}_{q_j} \left(q_j\right) + \text{constante} \\
&= \mathbb{E}_{q_j} \left\{ \mathbb{E}_{q_{-j}} \left[\ln p\left(\theta, D\right) \right] \right\} - \mathbb{E}_{q_j} \left(q_j\right) + \text{constante.}
\end{aligned} \tag{1.13}
$$

Ahora, podemos escribir la solución óptima como

$$
q_j * \propto \exp \left\{ E_{q_{-j}} \left[\ln p\left(\theta, D\right) \right] \right\}. \tag{1.14}
$$

Algoritmo 1: Inferencia variacional de ascenso de coordenadas (CAVI)

Entrada: Un modelo $p(\theta, D)$, una base de datos D
Salida : Densidad variacional $q(\theta) = \prod_{j=1}^{m} q_j\left(\theta_j\right)$
Inicio : Parámetros variacionales $q_j\left(\theta_j\right)$

mientras *ELBO no converja* **hacer**
  para $j \in \{1, \ldots m\}$ **hacer**
    Fijar $q_j\left(\theta_j\right) \propto \exp \left\{ \mathbb{E}_{q_{-j}} \left[\ln p\left(\theta_j | \theta_{-j}, D\right) \right] \right\}$
  fin
  Calcular ELBO$\left(q\right) = \mathbb{E}\left[\ln p\left(\theta, D\right)\right] - \mathbb{E}\left[\ln q\left(\theta\right)\right]$
fin

Observación 1.1. La expresión 1.14 encontrada para la actualización de los parámetros variacionales parece fácil de calcular, pero no es así en todos los casos. Por lo que existe una alternativa, que consiste en calcular las derivadas parciales del ELBO con respecto a los parámetros variacionales, para simplificar el cálculo. En ambos casos, se usará el ELBO para dar seguimiento a la convergencia (Blei *et al.*, 2017).

Capítulo 2

Metodología

En este capítulo se muestra un enfoque para el cálculo de la probabilidad condicional (conocido como el problema de la Inferencia Bayesiana) para el modelo gráfico probabilístico de las enfermedades:

- Tuberculosis,
- Neumonía,
- Alergia,
- Asma bronquial,
- Gripe común,

y sus respectivos síntomas. Dicho enfoque, transforma este cálculo en un problema de optimización, mediante el uso de inferencia variacional. Presentándose como una alternativa a la inferencia exacta.

En primera instancia, construimos el Modelo Gráfico Probabilístico que se ajusta a los datos disponibles. A partir de este, hallaremos la función variacional asociada al problema con la ayuda del método de Aproximación de Campo-Medio (*Approximation Mean-Field*). Finalmente, calcularemos el Límite Inferior de Evidencia (ELBO, por sus siglas en inglés): parte fundamental de la implementación del algoritmo *Coordinate Ascent Variational Inference* (CAVI) aplicado al problema.

2.1. Base de datos

Disease Symptom Prediction es una base de datos que tiene como objetivo proporcionar a los estudiantes una fuente para desarrollar modelos afines a la atención médica. La base de datos es de acceso público bajo la licencia CC BY-SA 4.0, y de actualización mensual. Para nuestro caso de estudio, procedimos a descargarla del sitio web `https://www.kaggle.com/` el 20 de junio de 2022.

Cada fila de la base de datos corresponde a un paciente. La primera columna de cada observación describe la enfermedad diagnosticada, mientras que el resto de columnas describen los síntomas presentados. La estructura de la base de datos se muestra en la Tabla 2.1, y está conformada por 600 filas y 18 columnas. Originalmente, los nombres de las enfermedades y síntomas se encuentran en inglés. En los Anexos A.3 y A.2 se muestran la traducción al español de cada síntoma y enfermedad, así como la indización que se usará posteriormente. La base contiene un total de 5 enfermedades, 31 síntomas, y 600 observaciones (120 por cada enfermedad).

Tabla 2.1

Vistazo de la base de datos Disease Symptom Prediction.

	Disease	**Symptom1**	**Symptom2**	⋯	**Symptom17**
1	Allergy	continuous sneezing	shivering	⋯	
2	Allergy	shivering	chills	⋯	
3	Allergy	continuous sneezing	shivering	⋯	
4	Tuberculosis	chills	vomiting	⋯	blood in sputum
5	Tuberculosis	vomiting	fatigue	⋯	
⋮	⋮	⋮	⋮	⋮	⋮
600	Common Cold	continuous sneezing	chills	⋯	muscle pain

Fuente: `https://www.kaggle.com/`.

Con el fin de que la base de datos sea más favorable para el desarrollo del modelo que se pretende, realizamos un proceso de limpieza y codificación binaria. Comenzamos por marcar en la primera columna la enfermedad diagnosticada (esta columna tiene la etiqueta *Disease*), el resto de columnas tendrán como etiquetas los nombres de los síntomas analizados (*continuous sneezing*, *shivering*, *chills*, etc.). Así, la base resultante tiene 600 filas y 32 columnas. Para cada observación, marcamos con 1 si presentó el síntoma de la columna correspondiente; caso contrario, lo marcamos con 0. En la Tabla 2.2 se presenta un leve vistazo de la estructura resultante. La preparación de los datos se efectuó con la ayuda del lenguaje de programación estadística `R` y la interfaz `RStudio`.

Tabla 2.2

Vistazo de la base de datos Disease Symptom Prediction procesada.

	Disease	continuous sneezing	shivering	⋯	blood in sputum
1	Allergy	1	1	⋯	0
2	Allergy	0	1	⋯	0
3	Allergy	1	0	⋯	0
4	Tuberculosis	0	0	⋯	1
5	Common Cold	1	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮
600	Bronchial Asthma	0	0	⋯	0

Fuente: Elaboración propia.

2.2. Modelo gráfico probabilístico

En esta sección, buscamos representar visualmente la relación que tiene cada una de las enfermedades con los síntomas mediante un grafo acíclico dirigido. Trabajaremos bajo los supuestos de causalidad enfermedad-síntoma e independencia de las enfermedades. Los **nodos padres** representan las variables de las enfermedades, mientras que los **nodos hijos** los síntomas. Las aristas dirigidas simbolizan la dependencia entre la enfermedad y el síntoma.

Realizamos el siguiente procedimiento para la construcción del gráfico probabilístico dirigido:

1. Colocamos en los nodos superiores (nodos padres) las enfermedades: tuberculosis, neumonía, alergia, asma bronquial, y gripe común.
2. Situamos en los nodos inferiores (nodos hijos) los síntomas.
3. Dividimos los datos por enfermedad.
4. Sumamos en cada división el total de observaciones por síntoma.
5. Si la suma por síntoma es mayor o igual a uno colegimos que hay relación entre el síntoma y la enfermedad. Caso contrario, no existe relación.

El resultado del procedimiento descrito anteriormente aplicado a nuestra base de datos se detalla en la Tabla 2.3.

A manera de ejemplo, una representación del modelo gráfico probabilístico reducido para 5 enfermedades y 12 síntomas se muestra en la Figura 2.1. En tanto que el modelo gráfico completo que describe nuestra base de datos (5 enfermedades y 31 síntomas) se presenta en el Anexo A.1.

Tabla 2.3

Relaciones entre las enfermedades y los síntomas.

Enfermedad	**Síntomas**
Alergia	Estornudos continuos, temblores, escalofríos, lagrimeo de ojos.
Asma Bronquial	Fatiga, tos, fiebre alta, dificultad para respirar, antecedentes familiares, esputo mucoide.
Gripe Común	Estornudos continuos, escalofríos, fatiga, tos, fiebre alta, dolor de cabeza, ganglios hinchados, malestar general, flema, irritación de la garganta, enrojecimiento de los ojos, presión sinusal, secreción nariz, congestión, dolor pecho, pérdida de olfato dolor muscular.
Neumonia	Escalofríos, fatiga, tos, fiebre alta, falta de aliento, sudoración, malestar general, flema, dolor en el pecho, taquicardia, esputo oxidado.
Tuberculosis	Escalofríos, vómitos, fatiga, pérdida de peso, tos, fiebre alta, falta de aliento, sudoración, pérdida del apetito, fiebre leve, ojos amarillos, ganglios hinchados, malestar general, flema, dolor pecho, sangre en esputo.

Fuente: Elaboración propia.

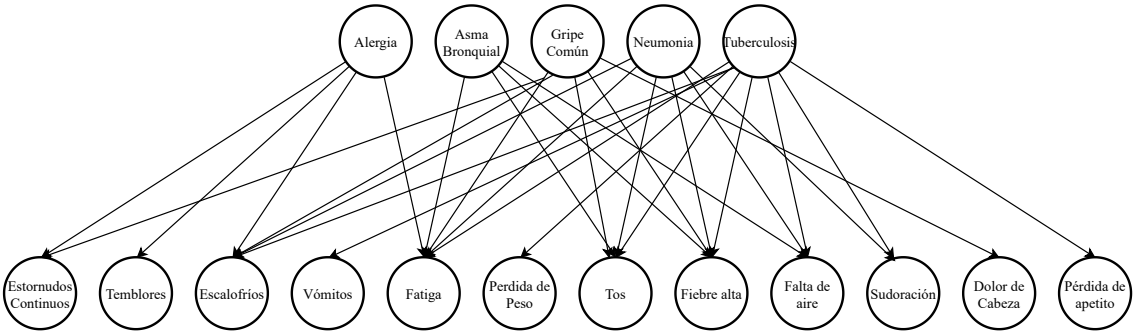


Figura 2.1. Representación del Modelo Gráfico Probabilístico para 5 enfermedades y 12 síntomas. Fuente: Elaboración propia.

2.3. Inferencia Variacional

En la sección anterior, modelamos un grafo dirigido acíclico para el estudio de la relación de los datos, en esta se agregará un modelo probabilístico sobre los parámetros del modelo, tal como se hizo para los datos.

Establezcamos un marco general para la inferencia bayesiana. Para ello, consideremos un conjunto de datos S y las variables latentes M. El objetivo es aprender de las variables ocultas dados los datos.

Sean $m \in M$ y $s \in S$ cualesquiera. Por el Teorema de Bayes, tenemos que

$$p(m|s) = \frac{p(m)\ p(s|m)}{p(s)} \tag{2.1}$$

El problema de la inferencia es calcular la probabilidad condicional de las variables latentes m dadas las observaciones s, esto es, condicionando los datos y calculando la probabilidad posterior $p(m|s)$.

Ahora ajustaremos el modelo gráfico probabilístico para un modelo de mezcla Beta-Bernoulli (*Beta-Bernoulli mixture model*). Comenzamos denotando a cada variable observada por $s_{ijk} \in \{0,1\}$ para la observación $i \in \{1, \ldots, 120\}$ del síntoma $j \in \{1, \ldots, 31\}$, dada la enfermedad $k \in \{1, \ldots, 5\}$. La parametrización completa es la siguiente:

$$\pi \sim \text{Dirichlet}\,(1/5, 1/5, 1/5, 1/5, 1/5) \tag{2.2}$$

$$z_k|\pi \sim \text{Categorical}(\pi) \tag{2.3}$$

$$\theta_{jk} \sim \text{Beta}\,(1/2, 1/2) \tag{2.4}$$

$$s_{ijk}|z_k \sim \text{Bernoulli}\left(\theta_{jk}\right) \tag{2.5}$$

Ahora para lograr aprender de los parámetros, aprenderemos de las distribuciones descritas. Así, centraremos nuestros esfuerzos en encontrar

$$p\,(z, \pi, \theta|s) \tag{2.6}$$

A partir de la aproximación mediante distribuciones uniformes.

2.3.1. Aproximación de Campo-Medio

Por la aproximación del campo-medio de la inferencia variacional (*Mean-Field Approximation*), introduciremos distribuciones variacionales para las variables latentes. De acuerdo a esto, optamos por las siguientes distribuciones de probabilidad para los parámetros π y θ:

$$q(\pi, \theta) = q(\pi) \cdot q(\theta) \tag{2.7}$$

$$= \prod_k q(\pi_k|a_k, b_k) \cdot \prod_j \prod_k q(\theta_{jk}|a_{jk}, b_{jk}) \tag{2.8}$$

donde

$$\pi_k|a_k, b_k \sim \text{Unif}\,(a_k, b_k) \tag{2.9}$$

y

$$\theta_{jk}|a_{jk},b_{jk} \sim \text{Unif}\left(a_{jk},b_{jk}\right), \tag{2.10}$$

con $0 < a_k < 1$, $0 < b_k < 1$, $0 < a_{jk} < 1$ y $0 < b_{jk} < 1$.

2.3.2. Límite inferior de evidencia (ELBO)

El algoritmo CAVI funciona estableciendo primero valores iniciales para los parámetros de densidad de variación y luego actualizándolos repetidamente hasta que el ELBO converja. Para esto, necesitamos calcular el ELBO en cada actualización. Recordemos que el ELBO está dado por

$$\text{ELBO} = \mathbb{E}_q\left[\ln\left(p(z,\pi,\theta,s)\right)\right] - \mathbb{E}_q\left[\ln\left(p(z)\right)\right]. \tag{2.11}$$

Logaritmo de la probabilidad conjunta

Para el cálculo del ELBO, necesitamos encontrar el logaritmo de la probabilidad conjunta de las variables latentes dados los datos.

$$\begin{aligned}
\ln\left[p(s,z,\theta,\pi)\right] &= \ln\left[p(\pi)\cdot p(z|\pi)\cdot p(\theta)\cdot p(x|z,\theta)\right] \\
&= \ln\left\{\left[\prod_{k=1}^{5} p\left(z_k|\pi\right)\right]\cdot\left[\prod_{j=1}^{31}\prod_{k=1}^{5} p\left(\theta_{jk}\right)\right]^{\delta(\text{padres } j,\ k)}\right\} \\
&\quad + \ln\left\{\left[\prod_{i=1}^{120}\prod_{j=1}^{31}\prod_{k=1}^{5} p\left(\theta_{jk}\right)\right]\right\} \\
&= \ln\left[p(\pi)\right] + \sum_{k=1}^{5}\ln\left[p(z_k|\pi)\right] + \sum_{j=1}^{31}\sum_{k=1}^{5}\delta(\text{padres } j,\ k)\ln\left[p\left(\theta_{jk}\right)\right] \\
&\quad + \sum_{i=1}^{120}\sum_{j=1}^{31}\sum_{k=1}^{5}\ln\left[p\left(s_{ijk}|z_k,\theta_{jk}\right)\right].
\end{aligned} \tag{2.12}$$

Para este cálculo utilizamos la delta de Kronecker $\delta(\cdot,\cdot)$, la cual tomará el valor de 1 si uno de los nodos padres del síntoma j es la enfermedad k, y 0 caso contrario.

Ahora, trataremos por separada cada término de la ecuación 2.12:

- De la ecuación 2.12 y la ecuación 2.2, colegimos que

$$\ln\left[p(\pi)\right] = \ln\left[\frac{1}{\text{B}\left(1/5,1/5,1/5,1/5,1/5\right)}\prod_{k=1}^{5}(\pi_k)^{\frac{1}{5}-1}\right]$$

$$
\begin{aligned}
&= \ln\left[\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma(\sum_{k=1}^{5}1/5)}}\prod_{k=1}^{5}(\pi_k)^{-\frac{4}{5}}\right] \\
&= \ln\left(\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma(\sum_{k=1}^{5}1/5)}}\right) + \ln\left(\prod_{k=1}^{5}(\pi_k)^{-\frac{4}{5}}\right) \\
&= \ln\left(\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma(\sum_{k=1}^{5}1/5)}}\right) - \frac{4}{5}\ln\left(\prod_{k=1}^{5}(\pi_k)\right) \\
&= \ln\left(\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma(\sum_{k=1}^{5}1/5)}}\right) - \frac{4}{5}\sum_{k=1}^{5}\ln(\pi_k) \\
&= -\frac{4}{5}\left[-\frac{5}{4}\ln\left(\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma(\sum_{k=1}^{5}1/5)}}\right) + \sum_{k=1}^{5}\ln(\pi_k)\right] \\
&\propto \frac{5}{4}\ln\left(\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma(\sum_{k=1}^{5}1/5)}}\right) - \sum_{k=1}^{5}\ln(\pi_k), \qquad (2.13)
\end{aligned}
$$

donde $\mathrm{B}(\cdot)$ y $\Gamma(\cdot)$ son las funciones beta multinomial y gamma, respectivamente.

- Por otro lado, de las ecuaciones 2.12 y 2.3, obtenemos

$$\ln\left[p(z_k|\pi)\right] = \ln\left[(\pi_k)^{z_k}\right] = z_k \ln\left[\pi_k\right]. \qquad (2.14)$$

- También, de las ecuaciones 2.12 y 2.4, inferimos que

$$
\begin{aligned}
\ln\left[p\left(\theta_{jk}\right)\right] &= \ln\left[\frac{(\theta_{jk})^{-\frac{1}{2}}\cdot(1-\theta_{jk})^{-\frac{1}{2}}}{\mathrm{B}(1/2,1/2)}\right] \\
&= \ln\left[\frac{(\theta_{jk})^{-\frac{1}{2}}\cdot(1-\theta_{jk})^{-\frac{1}{2}}}{\frac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1)}}\right] \\
&= \ln\left[\frac{(\theta_{jk})^{-\frac{1}{2}}\cdot(1-\theta_{jk})^{-\frac{1}{2}}}{\pi}\right] \\
&= \ln\left[(\theta_{jk})^{-\frac{1}{2}}\cdot(1-\theta_{jk})^{-\frac{1}{2}}\right] - \ln\left[\pi\right] \\
&= -\frac{1}{2}\left[\ln(\theta_{jk}) + \ln(1-\theta_{jk})\right] - \ln(\pi) \\
&= -\frac{1}{2}\left[\ln(\theta_{jk}) + \ln(1-\theta_{jk}) + 2\ln(\pi)\right]
\end{aligned}
$$

$$\propto -\ln(\theta_{jk}) - \ln(1-\theta_{jk}) - 2\ln(\pi), \tag{2.15}$$

donde $\mathrm{B}(\cdot)$ y $\Gamma(\cdot)$ son las funciones beta multinomial y gamma, respectivamente.

- Además, las ecuaciones 2.12 y 2.5 nos ayudan a concluir que

$$\begin{aligned}\ln\left[p\left(s_{ijk}|z_k,\theta_{jk}\right)\right] &= \ln\left[\theta_{jk}^{s_{ijk}} \cdot (1-\theta_{jk})^{1-s_{ijk}}\right] \\ &= \ln\left[\theta_{jk}^{s_{ijk}}\right] + \ln\left[(1-\theta_{jk})^{1-s_{ijk}}\right] \\ &= s_{ijk}\ln\left[\theta_{jk}\right] + (1-s_{ijk})\ln\left[(1-\theta_{jk})\right]. \end{aligned} \tag{2.16}$$

Finalmente, combinando todo lo anterior, el logaritmo de la probabilidad conjunta se puede escribir como

$$\begin{aligned}\ln\left[p(s,z,\theta,\pi)\right] \propto\ & \frac{5}{4}\ln\left(\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma(\sum_{k=1}^{5}1/5)}}\right) - \sum_{k=1}^{5}\ln(\pi_k) + \sum_{k=1}^{5} z_k \ln\left[\pi_k\right] \\ & - \sum_{j=1}^{31}\sum_{k=1}^{5}\delta(\text{padres } j,k)\left[\ln(\theta_{jk}) + \ln(1-\theta_{jk}) + 2\ln(\pi)\right] \\ & + \sum_{i=1}^{120}\sum_{j=1}^{31}\sum_{k=1}^{5}\left\{s_{ijk}\ln\left[\theta_{jk}\right] + (1-s_{ijk})\ln\left[(1-\theta_{jk})\right]\right\}. \end{aligned} \tag{2.17}$$

Entropía de distribuciones variacionales

Como resultado del supuesto de la aproximación de campo-medio (*Mean-Field Approximation*), tenemos que

$$\begin{aligned}\ln\left[q(\pi,\theta)\right] &= \ln\left[q(\pi)\right] + \ln\left[q(\theta)\right] \\ &= \sum_{k=1}^{5}\ln\left[p(\pi_k|a_k,b_k)\right] + \sum_{j=1}^{31}\sum_{k=1}^{5}\ln\left[p(\theta_{jk}|a_{jk},b_{jk})\right]. \end{aligned} \tag{2.18}$$

Ahora, desarrollamos cada término por separado:

- Como resultado de las ecuaciones 2.18 y 2.7, obtenemos

$$\begin{aligned}\ln\left[p(\pi_i|a_i,b_i)\right] &= \ln\left[\frac{1}{b_i-a_i}\right] \\ &= \ln[1] - \ln[b_i-a_i] \\ &= -\ln[b_i-a_i], \end{aligned} \tag{2.19}$$

con $0 < a_i < 1$, $0 < b_i < 1$ y $a_i < b_i$.

- De las ecuaciones 2.18 y 2.8, vemos que

$$\begin{aligned} \ln\left[p\left(\theta_{jk}|a_{jk},b_{jk}\right)\right] &= \ln\left[\frac{1}{b_{jk}-a_{jk}}\right] \\ &= \ln[1] - \ln\left[b_{jk}-a_{jk}\right] \\ &= -\ln\left[b_{jk}-a_{jk}\right], \end{aligned} \tag{2.20}$$

con $0 < a_i < 1$, $0 < b_i < 1$ y $a_i < b_i$.

En definitiva, logramos la expresión

$$\ln\left[q(\pi,\theta)\right] \propto \sum_{k=1}^{5} \ln\left[b_k - a_k\right] + \sum_{j=1}^{31}\sum_{k=1}^{5} \ln\left[b_{jk}-a_{jk}\right]. \tag{2.21}$$

Distribuciones importantes

En este punto, nos encontramos con la necesidad de caracterizar la función de densidad de cada parámetro variacional, y así calcular la actualización del ELBO.

Sea X una variable aleatoria continua uniformemente distribuida en el intervalo (a,b), con $0 < a < b \leq 1$.

- $Y = -\ln(X)$
  - **Dominio:** $y \in [-\ln(b), -\ln(a)]$
  - **Función de distribución**

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(-\ln(X) \leq y) \\ &= P(\ln(X) \geq -y) \\ &= P(X \geq \exp(-y)) \\ &= 1 - P(X \leq \exp(-y)) \\ &= 1 - \frac{\exp(-y)-a}{b-a} \\ &= \frac{b-a-\exp(-y)+a}{b-a} \\ &= \frac{b-\exp(-y)}{b-a} \end{aligned} \tag{2.22}$$

- **Función de densidad**

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}\left[\frac{b-\exp(-y)}{b-a}\right] = \frac{\exp(-y)}{b-a} \tag{2.23}$$

- **Esperanza**

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[-\ln(X)] \\ &= \int_{-ln(b)}^{-\ln(a)} y \cdot \frac{\exp(-y)}{b-a}\, dy \\ &= \frac{a\ln(a) - a - b\ln(b) + b}{b-a}\end{aligned} \tag{2.24}$$

- $Z = 1 - X$

  - **Dominio:** $y \in [1-b, 1-a]$
  - **Función de distribución**

$$\begin{aligned}F_Z(z) &= P(Z \leq z) \\ &= P(1 - X \leq y) \\ &= P(-X \leq y - 1) \\ &= P(X \geq 1 - y) \\ &= P(X \geq 1 - y) \\ &= 1 - P(X \leq 1 - y) \\ &= 1 - \frac{1-y-a}{b-a} \\ &= \frac{b-a-1+y+a}{b-a} \\ &= \frac{y-(1-b)}{b-a}\end{aligned} \tag{2.25}$$

    Así, concluimos que $Z \sim \text{Unif}(1-b, 1-a)$.

  - **Función de densidad**

$$f_Z(z) = \frac{1}{b-a} \tag{2.26}$$

  - **Esperanza**

$$\mathbb{E}[Z] = \mathbb{E}[1-X] = \frac{2-b-a}{2} \tag{2.27}$$

- $T = -\ln(1-X)$
  - **Dominio:** $t \in [-\ln(1-a), -\ln(1-b)]$
  - **Función de distribución**

$$\begin{aligned} F_T(t) &= P(T \leq t) \\ &= P(-\ln(1-X) \leq t) \\ &= P(\ln(1-X) \geq -t) \\ &= P(1-X \geq \exp(-t)) \\ &= P(-X \leq \exp(-t) - 1) \\ &= P(X \geq -\exp(-t) + 1) \\ &= \frac{1 - e^{(}-z) - a}{b-a} \\ &= \frac{b - a - \exp(-y) + a}{b-a} \\ &= \frac{b - \exp(-y)}{b-a} \end{aligned} \tag{2.28}$$

  - **Función de densidad**

$$f_Y(y) = F'_Y(y) = \frac{d}{dz}\left(\frac{1-\exp(-z)-a}{b-a}\right) = \frac{\exp(-z)}{b-a} \tag{2.29}$$

  - **Esperanza**

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[-\ln(1-X)] \\ &= \int_{-\ln(1-a)}^{-\ln(1-b)} t\frac{\exp(-t)}{b-a}\, dt \\ &= \frac{\ln(-b+1) - b\ln(-b+1) + b + a\ln(-a+1) - a - \ln(-a+1)}{b-a} \end{aligned} \tag{2.30}$$

Cálculo del ELBO

Combinando los resultados anteriores, tenemos que el ELBO es:

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_q\left\{\ln\left[p(z,\pi,\theta,s)\right]\right\} - \mathbb{E}_q\left\{\ln\left[p(z)\right]\right\} \\ &\propto \mathbb{E}\left\{\frac{5}{4}\ln\left[\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma(\sum_{k=1}^{5}1/5)}}\right] - \sum_{k=1}^{5}\ln(\pi_k) + \sum_{k=1}^{5} z_k \ln(\pi_k)\right\} \end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}_q \left\{ \sum_{j=1}^{31} \sum_{k=1}^{5} \delta\left(\text{padres } j,k\right) \left[\ln\left(\theta_{jk}\right) + \ln\left(1-\theta_{jk}\right) + 2\ln\left(\pi_k\right) \right] \right\} \\
&+ \mathbb{E}_q \left\{ \sum_{i=1}^{120} \sum_{j=1}^{31} \sum_{k=1}^{5} \left[s_{ijk} \ln\left(\theta_{jk}\right) + \left(1-s_{ijk}\right) \ln\left(1-\theta_{jk}\right) \right] \right\} \\
&+ \mathbb{E}_q \left[\sum_{i=1}^{5} \ln\left(b_i - a_i\right) + \sum_{j=1}^{31} \sum_{k=1}^{5} \ln\left(b_{jk} - a_{jk}\right) \right] \\
\propto\ & \frac{5}{4} \ln \left[\frac{1}{\frac{\prod_{k=1}^{5} \Gamma(1/5)}{\Gamma\left(\sum_{k=1}^{5} 1/5\right)}} \right] + \sum_{k=1}^{5} \mathbb{E}_q \left[-\ln\left(\pi_k\right) \right] - \sum_{k=1}^{5} z_k \mathbb{E}_q \left[-\ln\left(\pi_k\right) \right] \\
&+ \sum_{j=1}^{31} \sum_{k=1}^{5} \delta\left(\text{padres } j,k\right) \left\{ \mathbb{E}_q \left[-\ln\left(\theta_{jk}\right) \right] + \mathbb{E}_q \left[-\ln\left(1-\theta_{jk}\right) \right] - 2\ln\left(\pi_k\right) \right\} \\
&+ \sum_{i=1}^{120} \sum_{j=1}^{31} \sum_{k=1}^{5} \left\{ -s_{ijk} \mathbb{E}_q \left[-\ln\left(\theta_{jk}\right) \right] - \left(1-s_{ijk}\right) \mathbb{E}_q \left[-\ln\left(1-\theta_{jk}\right) \right] \right\} \\
&+ \sum_{k=1}^{5} \ln\left(b_k - a_k\right) + \sum_{j=1}^{31} \sum_{k=1}^{5} \ln\left(b_{jk} - a_{jk}\right)
\end{aligned} \tag{2.31}
$$

Definimos $A_k := \mathbb{E}_q \left[-\ln\left(\pi_k\right) \right]$, $B_{jk} := \mathbb{E}_q \left[-\ln\left(\theta_{jk}\right) \right]$ y $C_{jk} := \mathbb{E}_q \left[-\ln\left(\theta_{jk}\right) \right]$, para $j \in \{1, \ldots, 31\}$ y $k \in \{1, \ldots, 5\}$. Ajustando las ecuaciones 2.24 y 2.30 a las distribuciones de las variables π y θ, tenemos

$$
A_k = \frac{\ln\left(a_k\right) a_k - a_k - \ln\left(b_k\right) b_k + b_k}{-a_k + b_k}, \tag{2.32}
$$

$$
B_{jk} = \frac{a_{jk} \ln\left(a_{jk}\right) - a_{jk} - b_{jk} \ln\left(b_{jk}\right) + b_{jk}}{-a_{jk} + b_{jk}}, \tag{2.33}
$$

y

$$
\begin{aligned}
C_{jk} = & \frac{a_{jk} \ln\left(-a_{jk}+1\right) - \ln\left(-a_{jk}+1\right) + b_{jk}}{-a_{jk} + b_{jk}} \\
& + \frac{\ln\left(-b_{jk}+1\right) - a_{jk} - b_{jk} \ln\left(-b_{jk}+1\right)}{-a_{jk} + b_{jk}}.
\end{aligned} \tag{2.34}
$$

Por consiguiente, el ELBO se puede escribir como

$$
\begin{aligned}
\text{ELBO} \propto & \frac{5}{4} \ln \left[\frac{1}{\frac{\prod_{k=1}^{5} \Gamma(1/5)}{\Gamma\left(\sum_{k=1}^{5} 1/5\right)}} \right] + \sum_{k=1}^{5} A_k - \sum_{k=1}^{5} z_k \cdot A_k \\
& + \sum_{j=1}^{31} \sum_{k=1}^{5} \delta\left(\text{padres } j,k\right) \left[B_{jk} + C_{jk} - 2\ln\left(\pi_k\right) \right]
\end{aligned}
$$

$$+\sum_{i=1}^{120}\sum_{j=1}^{31}\sum_{k=1}^{5}\left[-s_{ijk}B_{jk}-\left(1-s_{ijk}C_{jk}\right)\right]$$
$$+\sum_{k=1}^{5}\ln\left(b_k-a_k\right)+\sum_{j=1}^{31}\sum_{k=1}^{5}\ln\left(b_{jk}-a_{jk}\right) \tag{2.35}$$

2.3.3. Actualización de los parámetros

Para encontrar el valor óptimo de las actualizaciones de los parámetros, necesitamos resolver la ecuación $\nabla(\text{ELBO}) = 0$. En nuestro caso, no es posible encontrar una expresión explícita de la actualización. Por ello, determinaremos solamente la condición que debe cumplir la actualización para ser considerada óptima.

- a_k

$$\begin{aligned}\frac{\partial}{\partial a_k}(\text{ELBO}) \propto\ & \frac{\partial}{\partial a_k}\left\{\frac{5}{4}\ln\left[\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma\left(\sum_{k=1}^{5}1/5\right)}}\right]+A_k-z_k\cdot A_k\right\}\\ &+\frac{\partial}{\partial a_k}\left\{\sum_{j=1}^{31}\sum_{k=1}^{5}\delta\left(\text{padres } j,k\right)\left[B_{jk}+C_{jk}-2\ln\left(\pi_k\right)\right]\right\}\\ &+\frac{\partial}{\partial a_k}\left\{\sum_{i=1}^{120}\sum_{j=1}^{31}\sum_{k=1}^{5}\left[-s_{ijk}B_{jk}-\left(1-s_{ijk}C_{jk}\right)\right]\right\}\\ &+\frac{\partial}{\partial a_k}\left[\sum_{k=1}^{5}\ln\left(b_k-a_k\right)+\sum_{j=1}^{31}\sum_{k=1}^{5}\ln\left(b_{jk}-a_{jk}\right)\right]\\ =\ & \frac{\partial}{\partial a_k}\left(A_k-z_k\cdot A_k\right)+\frac{\partial}{\partial a_k}\left[\ln\left(b_k-a_k\right)\right]\\ =\ & \frac{-b_kz_k+b_kz_k\ln\left(b_k\right)+b_k\ln\left(a_k\right)-b_kz_k\ln\left(a_k\right)-b_k\ln\left(b_k\right)+a_kz_k}{\left(b_k-a_k\right)^2}.\end{aligned} \tag{2.36}$$

  El valor óptimo para la actualización a_k^* es tal que

$$-b_kz_k+b_kz_k\ln\left(b_k\right)+b_k\ln\left(a_k^*\right)-b_kz_k\ln\left(a_k^*\right)-b_k\ln\left(b_k\right)+a_k^*z_k=0. \tag{2.37}$$

- b_k

$$\frac{\partial}{\partial b_k}(\text{ELBO}) \propto \frac{\partial}{\partial b_k}\left\{\frac{5}{4}\ln\left[\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma\left(\sum_{k=1}^{5}1/5\right)}}\right]+A_k-z_k\cdot A_k\right\}$$

$$
\begin{aligned}
&+\frac{\partial}{\partial b_k}\left\{\sum_{j=1}^{31}\sum_{k=1}^{5}\delta\left(\text{padres } j,k\right)\left[B_{jk}+C_{jk}-2\ln\left(\pi_k\right)\right]\right\}\\
&+\frac{\partial}{\partial b_k}\left\{\sum_{i=1}^{120}\sum_{j=1}^{31}\sum_{k=1}^{5}\left[-s_{ijk}B_{jk}-\left(1-s_{ijk}C_{jk}\right)\right]\right\}\\
&+\frac{\partial}{\partial b_k}\left[\sum_{k=1}^{5}\ln\left(b_k-a_k\right)+\sum_{j=1}^{31}\sum_{k=1}^{5}\ln\left(b_{jk}-a_{jk}\right)\right]\\
&=\frac{\partial}{\partial b_k}\left(A_k-z_k\cdot A_k\right)+\frac{\partial}{\partial b_k}\left[\ln\left(b_k-a_k\right)\right]\\
&=\frac{b_k z_k+a_k\ln\left(b_k\right)-a_k z_k\ln\left(b_k\right)-a_k z_k-a_k\ln\left(a_k\right)+a_k z_k\ln\left(a_k\right)}{\left(b_k-a_k\right)^2}.
\end{aligned}
\tag{2.38}
$$

El valor óptimo para la actualización b_k^* es tal que

$$
b_k^* z_k+a_k\ln\left(b_k^*\right)-a_k z_k\ln\left(b_k^*\right)-a_k z_k-a_k\ln\left(a_k\right)+a_k z_k\ln\left(a_k\right)=0. \tag{2.39}
$$

Para facilitar la escritura de las actualizaciones óptimas de a_{jk} y b_{jk}, definimos $c_{ijk}:=\sum_{i=1}^{120}s_{ijk}$ y $d_{ijk}:=\sum_{i=1}^{120}\left(1-s_{ijk}\right)$.

- a_{jk}

$$
\begin{aligned}
\frac{\partial}{\partial a_{jk}}\left(\text{ELBO}\right)\propto&\ \frac{\partial}{\partial a_{jk}}\left\{\frac{5}{4}\ln\left[\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma\left(\sum_{k=1}^{5}1/5\right)}}\right]+\sum_{k=1}^{5}A_k-\sum_{k=1}^{5}z_k\cdot A_k\right\}\\
&+\frac{\partial}{\partial a_{jk}}\left\{\delta\left(\text{padres } j,k\right)\left[B_{jk}+C_{jk}-2\ln\left(\pi_k\right)\right]\right\}\\
&+\frac{\partial}{\partial a_{jk}}\left\{\sum_{i=1}^{120}\left[-s_{ijk}\,B_{jk}-\left(1-s_{ijk}\,C_{jk}\right)\right]\right\}\\
&+\frac{\partial}{\partial a_{jk}}\left[\sum_{k=1}^{5}\ln\left(b_k-a_k\right)+\ln\left(b_{jk}-a_{jk}\right)\right]\\
=&\ \delta\left(\text{padres } j,k\right)\left[\frac{b_{jk}\ln\left(a_{jk}\right)-a_{jk}-b_{jk}\ln\left(b_{jk}\right)+b_{jk}}{\left(b_{jk}-a_{jk}\right)^2}\right]\\
&+\delta\left(\text{padres } j,k\right)\left[\frac{-a_{jk}+b_{jk}\ln\left(-a_{jk}+1\right)-\ln\left(-a_{jk}+1\right)+b_{jk}}{\left(b_{jk}-a_{jk}\right)^2}\right]\\
&+\delta\left(\text{padres } j,k\right)\left[\frac{\ln\left(-b_{jk}+1\right)-b_{jk}\ln\left(-b_{jk}+1\right)}{\left(b_{jk}-a_{jk}\right)^2}\right]\\
&+\sum_{i=1}^{120}\left\{-s_{ijk}\left[\frac{b_{jk}\ln\left(a_{jk}\right)-a_{jk}-b_{jk}\ln\left(b_{jk}\right)+b_{jk}}{\left(b_{jk}-a_{jk}\right)^2}\right]\right\}
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{i=1}^{120}\left\{-\left(1-s_{ijk}\right)\left[\frac{-a_{jk}+b_{jk}\ln\left(-a_{jk}+1\right)-\ln\left(-a_{jk}+1\right)}{\left(b_{jk}-a_{jk}\right)^2}\right]\right\}\\
&+\sum_{i=1}^{120}\left\{-\left(1-s_{ijk}\right)\left[\frac{b_{jk}+\ln\left(-b_{jk}+1\right)-b_{jk}\ln\left(-b_{jk}+1\right)}{\left(b_{jk}-a_{jk}\right)^2}\right]\right\}\\
&-\frac{1}{b_{jk}-a_{jk}}.
\end{aligned}
\tag{2.40}
$$

El valor óptimo para la actualización a_{jk}^* verifica que

$$
\begin{aligned}
0 = {} & b_{jk}c_{ijk}\ln\left(b_{jk}\right)-b_{jk}c_{ijk}-b_{jk}d_{ijk}+b_{jk}d_{ijk}\ln\left(-b_{jk}+1\right)+b_{jk}\ln\left(a_{jk}^*\right)\\
&+b_{jk}d_{ijk}\ln\left(-b_{jk}+1\right)+b_{jk}\ln\left(a_{jk}^*\right)+b_{jk}\ln\left(-a_{jk}^*+1\right)-b_{jk}c_{ijk}\ln\left(a_{jk}^*\right)\\
&-b_{jk}d_{ijk}\ln\left(-a_{jk}^*+1\right)-b_{jk}\ln\left(b_{jk}\right)+b_{jk}-b_{jk}\ln\left(-b_{jk}+1\right)\\
&-d_{ijk}\ln\left(\frac{-b_{jk}+1}{-a_{jk}^*+1}\right)+\ln\left(\frac{-b_{jk}+1}{-a_{jk}^*+1}\right)-a_{jk}^*+a_{jk}^*c_{ijk}+a_{jk}^*d_{ijk}.
\end{aligned}
\tag{2.41}
$$

- b_{jk}

$$
\begin{aligned}
\frac{\partial}{\partial b_{jk}}\left(\text{ELBO}\right) \propto {} & \frac{\partial}{\partial b_{jk}}\left\{\frac{5}{4}\ln\left[\frac{1}{\frac{\prod_{k=1}^{5}\Gamma(1/5)}{\Gamma\left(\sum_{k=1}^{5}1/5\right)}}\right]+\sum_{k=1}^{5}A_k-\sum_{k=1}^{5}z_k\cdot A_k\right\}\\
&+\frac{\partial}{\partial b_{jk}}\left\{\delta\left(\text{padres } j,k\right)\left[B_{jk}+C_{jk}-2\ln\left(\pi_k\right)\right]\right\}\\
&+\frac{\partial}{\partial b_{jk}}\left\{\sum_{i=1}^{120}\left[-s_{ijk}\,B_{jk}-\left(1-s_{ijk}\,C_{jk}\right)\right]\right\}\\
&+\frac{\partial}{\partial b_{jk}}\left[\sum_{k=1}^{5}\ln\left(b_k-a_k\right)+\ln\left(b_{jk}-a_{jk}\right)\right]\\
= {} & \delta\left(\text{padres } j,k\right)\left[\frac{-b_{jk}+a_{jk}\ln\left(b_{jk}\right)+a_{jk}-a_{jk}\ln\left(a_{jk}\right)}{\left(b_{jk}-a_{jk}\right)^2}\right]\\
&+\delta\left(\text{padres } j,k\right)\left[\frac{-b_{jk}+a_{jk}\ln\left(-b_{jk}+1\right)-\ln\left(-b_{jk}+1\right)+a_{jk}}{\left(b_{jk}-a_{jk}\right)^2}\right]\\
&+\delta\left(\text{padres } j,k\right)\left[\frac{-a_{jk}\ln\left(-a_{jk}+1\right)+\ln\left(-a_{jk}+1\right)}{\left(b_{jk}-a_{jk}\right)^2}\right]\\
&+\sum_{i=1}^{120}\left\{-s_{ijk}\left[\frac{-b_{jk}+a_{jk}\ln\left(b_{jk}\right)+a_{jk}-a_{jk}\ln\left(a_{jk}\right)}{\left(b_{jk}-a_{jk}\right)^2}\right]\right\}\\
&+\sum_{i=1}^{120}\left\{-\left(1-s_{ijk}\right)\left[\frac{-b_{jk}+a_{jk}\ln\left(-b_{jk}+1\right)-\ln\left(-b_{jk}+1\right)+a_{jk}}{\left(b_{jk}-a_{jk}\right)^2}\right]\right\}
\end{aligned}
$$

$$+\sum_{i=1}^{120}\left\{-\left(1-s_{ijk}\right)\left[\frac{-a_{jk}\ln\left(-a_{jk}+1\right)+\ln\left(-a_{jk}+1\right)}{\left(b_{jk}-a_{jk}\right)^2}\right]\right\}$$
$$+\frac{1}{b_{jk}-a_{jk}}. \tag{2.42}$$

El valor óptimo para la actualización b^*_{jk} cumple que

$$\begin{aligned}0 = {} & a_{jk}\ln\left(b^*_{jk}\right)+a_{jk}\ln\left(-b^*_{jk}+1\right)+a_{jk}-a_{jk}\ln\left(a_{jk}\right)-a_{jk}\ln\left(-a_{jk}+1\right)\\ & -c_{ijk}\left[a_{jk}\ln\left(b^*_{jk}\right)+a_{jk}-a_{jk}\ln\left(a_{jk}\right)-b^*_{jk}\right]+\ln\left(\frac{-a_{jk}+1}{-b^*_{jk}+1}\right)\\ & -d_{ijk}\left[a_{jk}\ln\left(-b^*_{jk}+1\right)+a_{jk}-a_{jk}\ln\left(-a_{jk}+1\right)+\ln\left(\frac{-a_{jk}+1}{-b^*_{jk}+1}\right)\right].\end{aligned} \tag{2.43}$$

2.4. Implementación del algoritmo CAVI

El algoritmo 2 introduce la Inferencia Variacional de Ascenso de Coordenadas (CAVI, por sus siglas en inglés) al problema de diagnóstico médico al que nos enfrentamos. Este será implementado en el lenguaje de programación estadística `R`.

El CAVI combina las actualizaciones variacionales de las ecuaciones 2.37, 2.39, 2.41 y 2.43 para encontrar las densidades variacionales que mejor se ajusten. Además, el algoritmo necesita calcular el ELBO de la ecuación 2.11. Utilizamos el ELBO para realizar un seguimiento del progreso del algoritmo y evaluar cuándo se ha producido la convergencia. Los criterios de parada que empleamos son:

- Hacer un máximo de 350 iteraciones.
- Si el error absoluto del ELBO es menor 0.001, esto es,

$$|\text{ELBO}_n - \text{ELBO}_{n-1}| < 0.001 \tag{2.44}$$

La función objetivo del algoritmo es el ELBO, en nuestro problema es no convexa (generalmente, no lo suele ser). Por esta razón, el algoritmo CAVI solo garantiza la convergencia a óptimos locales y es sensible a la inicialización de los parámetros variacionales. Justamente para este punto optamos por escoger los parámetros de inicialización de forma que siguen una distribución aleatoria uniforme, y para asegurar la reproducibilidad debemos fijar una semilla aleatoria `setseed(2511)`.

Por otro lado, las actualizaciones variacionales no se pudieron encontrar explícitamente, por lo que debemos calcular la solución numérica de ecuaciones no lineales restringidas al intervalo $(0,1)$. La función `unitroot.all` de la biblioteca `solveRoot` de `R` nos permite encontrar las raíces de funciones no lineales utilizando el método de *Newton-Raphson*. Sin embargo, pueden existir múltiples raíces en el intervalo $(0,1)$. Una solución a esto es seleccionar la raíz más grande encontrada para la actualización.

Algoritmo 2: CAVI aplicado al problema de diagnóstico médico.

Entrada: Datos s_{ijk}
Salida : Densidades variacionales uniformes $q(\pi_k|a_k,b_k)$ y $q\left(\theta_{jk}|a_{jk},b_{jk}\right)$

Inicialización:
  Parámetros variacionales **para** $k \in \{1,\ldots,5\}$ **hacer**
    $a_k \leftarrow \text{Unif}(0,0.5)$;
    $b_k \leftarrow \text{Unif}(0.5,1)$;
  fin
  para $j \in \{1,\ldots,31\}$ **hacer**
    para $k \in \{1,\ldots,5\}$ **hacer**
      $a_{jk} \leftarrow \text{Unif}(0,0.5)$;
      $b_{jk} \leftarrow \text{Unif}(0.5,1)$;
    fin
  fin
fin
mientras *ELBO no converja* **hacer**
  para $k \in \{1,\ldots,5\}$ **hacer**
    Resolver la ecuación 2.37 en el intervalo $(0,1)$ y actualizar a_k^*;
    Resolver la ecuación 2.39 en el intervalo $(0,1)$ y actualizar b_k^*;
  fin
  para $j \in \{1,\ldots,31\}$ **hacer**
    para $k \in \{1,\ldots,5\}$ **hacer**
      Resolver la ecuación 2.41 en el intervalo $(0,1)$ y actualizar a_{jk}^*;
      Resolver la ecuación 2.43 en el intervalo $(0,1)$ y actualizar b_{jk}^*;
    fin
  fin
  Calcular ELBO $\left(a_k^*, b_k^*, a_{jk}^*, b_{jk}^*\right)$;
fin

Capítulo 3

Resultados, conclusiones y recomendaciones

En este capítulo exhibiremos los resultados obtenidos de los datos aplicados a la metodología que se presento en el capítulo anterior. Estos resultados mostrarán la convergencia del ELBO y el valor de los parámetros variacionales, obtenidos de la aplicación del algoritmo CAVI.

3.1. Resultados

Como se indicó en el anterior capítulo, se implementó el algoritmo CAVI adaptado a nuestro problema en el lenguaje de programación `R` (ver en los anexos). Para la ejecución del programa se utilizó una computadora Dell Latitude E6330 con procesador Intel(R) Core(TM) i5-3320M CPU @ 2.60GHz y memoria RAM de 8 GB. El tiempo total de ejecución del programa fue 31.6 segundos. Este tiempo parece bastante razonable con respecto al tamaño del problema al que nos enfrentamos.

3.1.1. Convergencia del ELBO

El ELBO y los parámetros variacionales han convergido en 152 iteraciones. El error absoluto alcanzado fue de 0.0004346.

La Figura 3.1 muestra la trayectoria del ELBO con las inicializaciones aleatorias. Como se había mencionado el ELBO es sensible a los valores iniciales. Esto lo podemos apreciar en las primeras iteraciones. A pesar de ello, al algoritmo le toma menos de 25

iteraciones encontrar el camino que conduce a la maximización del ELBO.

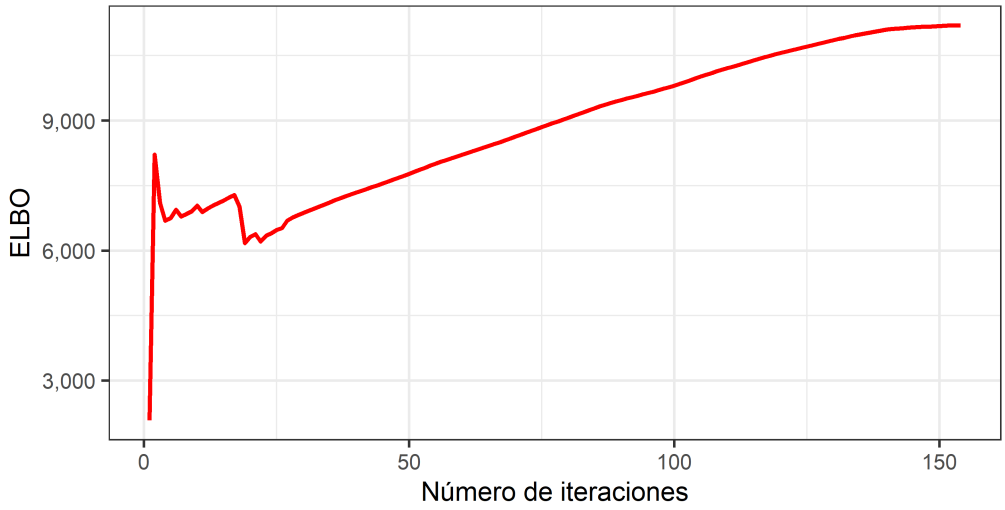


Figura 3.1. Evolución de la convergencia del ELBO (152 iteraciones).
Fuente: Elaboración propia.

3.1.2. Actualización de los parámetros

La probabilidad de tener la tuberculosis sigue una distribución uniforme de parámetros 0.284 y 0.919 (Ver en la Figura 3.2). La esperanza de esta distribución es 0.601 con varianza 0.034.

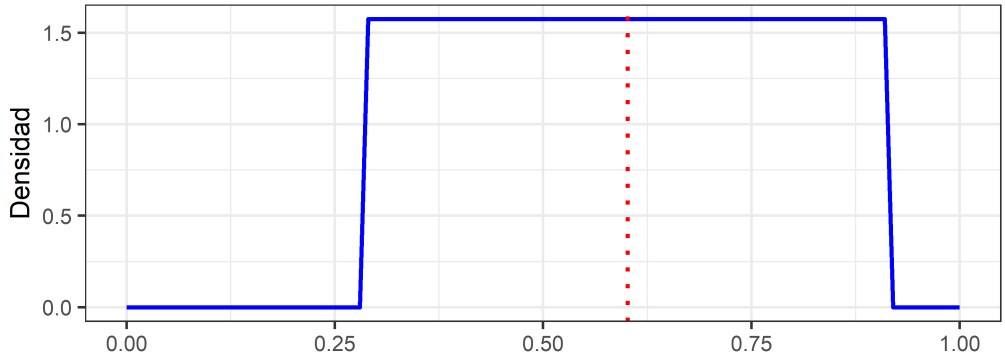


Figura 3.2. Función de densidad de probabilidad del parámetro π_1.
Fuente: Elaboración propia.

De igual importancia, la probabilidad de presenta neumonía sigue una distribución uniforme de parámetros 0.311 y 0.863 (Ver en la Figura 3.3). La probabilidad de tener neumonía tiene una esperanza de 0.587 con varianza 0.025.

Así mismo, la probabilidad de padecer alergia tiene una distribución uniforme de

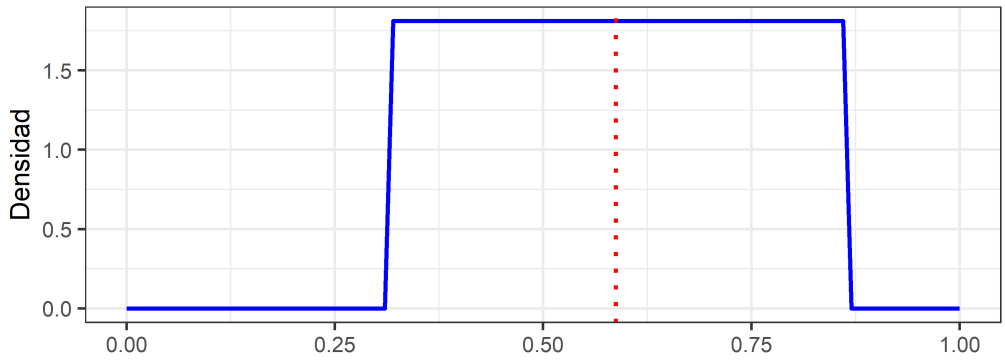


Figura 3.3. Función de densidad de probabilidad del parámetro π_2.
Fuente: Elaboración propia.

parámetros 0.240 y 0.864 (Ver en la Figura 3.4), con una esperanza de 0.552 y con varianza 0.032.

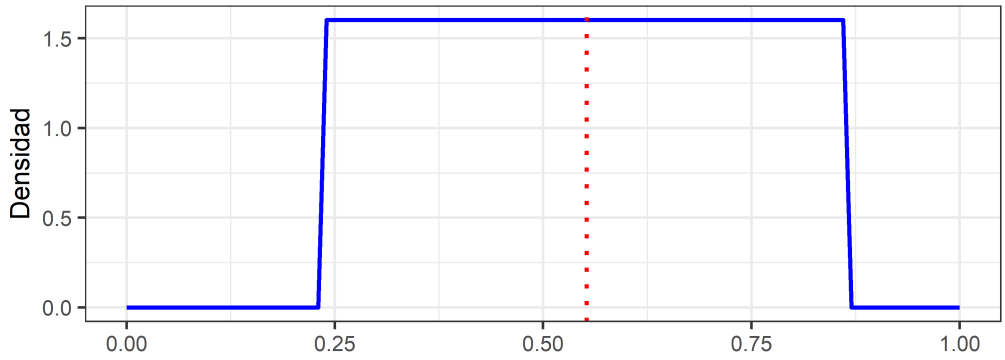


Figura 3.4. Función de densidad de probabilidad del parámetro π_3.
Fuente: Elaboración propia.

De la Figura 3.5, observamos que la probabilidad de tener asma sigue una distribución uniforme entre 0.066 y 0.651, con una ezperanza de e 0.359 y varianza 0.029.

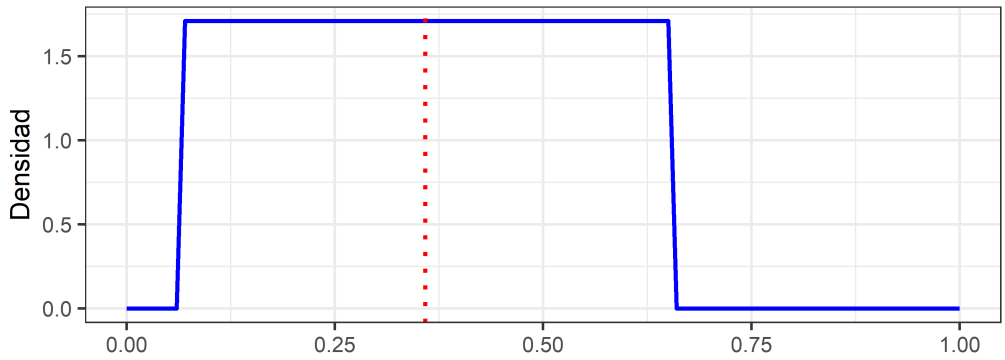


Figura 3.5. Función de densidad de probabilidad del parámetro π_4.
Fuente: Elaboración propia.

Por último, la probabilidad de tener gripe común se distribuye uniformemente con parámetros 0.270 y 0.540 (ver la Figura 3.6). La esperanza de esta distribución es de 0.405 y con varianza 0.006.

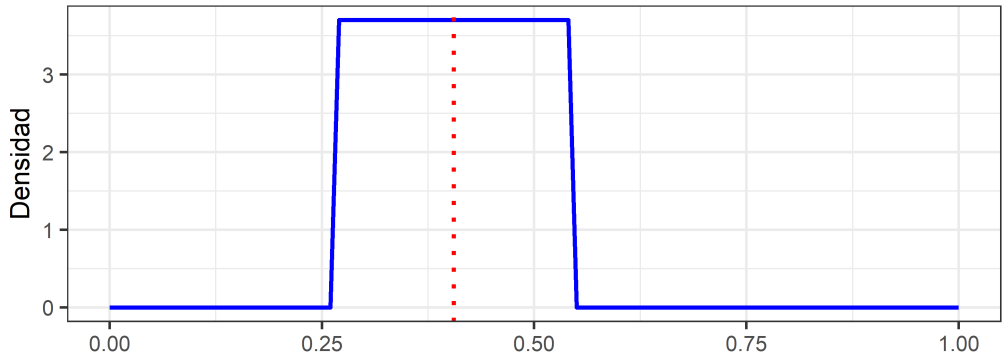


Figura 3.6. Función de densidad de probabilidad del parámetro π_5.
Fuente: Elaboración propia.

En la tabla 3.1 observamos que ningún síntoma asociado con la tuberculosis es predominante, es decir, ninguno de los síntomas presentes indican por si solos la presencia de la enfermedad.

Tabla 3.1

Resultados de los parámetros variacionales para la enfermedad tuberculosis.

Síntoma	**Parámetro**	**Distribución de densidad**
Escalofríos	$\theta_{3,1}$	Unif(0.032, 0.033)
Vómitos	$\theta_{4,1}$	Unif(0.032, 0.033)
Fatiga	$\theta_{5,1}$	Unif(0.032, 0.033)
Pérdida de peso	$\theta_{6,1}$	Unif(0.032, 0.033)
Tos	$\theta_{7,1}$	Unif(0.027, 0.028)
Fiebre alta	$\theta_{8,1}$	Unif(0.032, 0.033)
Falta de aliento	$\theta_{9,1}$	Unif(0.032, 0.033)
Sudoración	$\theta_{10,1}$	Unif(0.032, 0.033)
Pérdida de apetito	$\theta_{12,1}$	Unif(0.034, 0.035)
Fiebre sueve	$\theta_{13,1}$	Unif(0.034, 0.035)
Ojos amarillos	$\theta_{14,1}$	Unif(0.034, 0.035)
Ganglios inflamados	$\theta_{15,1}$	Unif(0.034, 0.035)
Malestar general	$\theta_{16,1}$	Unif(0.034, 0.035)
Flema	$\theta_{17,1}$	Unif(0.034, 0.035)
Dolor de pecho	$\theta_{23,1}$	Unif(0.034, 0.035)
Sangre en esputo	$\theta_{31,1}$	Unif(0.034, 0.035)

Fuente: Elaboración propia.

Por otro lado, la tabla 3.2 nos indica que los síntomas asociados con esta no podrían ayudarnos a obtener un diagnóstico favorable para neumonía, debido que el intevalo

al que pertenece el paràmetro tiene valores muy bajos.

Tabla 3.2

Resultados de los parámetros variacionales para la enfermedad neumonía.

Síntoma	**Parámetro**	**Distribución de densidad**
Escalofríos	$\theta_{3,2}$	Unif(0.032, 0.033)
Fatiga	$\theta_{5,2}$	Unif(0.032, 0.033)
Tos	$\theta_{7,2}$	Unif(0.032, 0.033)
Fiebre alta	$\theta_{8,2}$	Unif(0.032, 0.033)
Falta de aliento	$\theta_{9,2}$	Unif(0.032, 0.033)
Sudoración	$\theta_{10,2}$	Unif(0.032, 0.033)
Malestar general	$\theta_{16,2}$	Unif(0.032, 0.033)
Flema	$\theta_{17,2}$	Unif(0.032, 0.033)
Dolor de pecho	$\theta_{23,2}$	Unif(0.034, 0.035)
Ritmo cardiaco rápido	$\theta_{24,2}$	Unif(0.034, 0.035)
Esputo oxidado	$\theta_{30,2}$	Unif(0.034, 0.035)

Fuente: Elaboración propia.

Se observa que en la tabla 3.3, las cosas cambian. Las enfermedades tienen un comportamiento similar, por lo tanto, todos los síntomas podrían llevarnos a tomar en consideración a la alergia como un posible diagnóstico.

Tabla 3.3

Resultados de los parámetros variacionales para la enfermedad alergia.

Síntoma	**Parámetro**	**Distribución de densidad**
Estornudos continuos	$\theta_{1,3}$	Unif(0.440, 0.560)
Temblores	$\theta_{2,3}$	Unif(0.441, 0.559)
Escalofríos	$\theta_{3,3}$	Unif(0.440, 0.560)
Lagrimeo de los ojos	$\theta_{27,3}$	Unif(0.440, 0.560)

Fuente: Elaboración propia.

Seguidamente de la tabla 3.4 sabemos que los síntomas tienen una probabilidad de entre el 44 % y el 56 %. Al estar en un intervalo donde la probabilidad es alta, suponemos que el padecimiento es Asma Bronquial.

Finalmente, en la tabla 3.5 podemos ver como todos los síntomas son predominantes excepto la tos, la fatiga y la irritación de garganta. Aunque algunas observaciones llegan a presentar estos síntomas, estos no nos indican la presencia de la gripe común.

Tabla 3.4

Resultados de los parámetros variacionales para la enfermedad asma bronquial.

Síntoma	Parámetro	Distribución de densidad
Fatiga	$\theta_{5,4}$	Unif(0.440, 0.560)
Tos	$\theta_{7,4}$	Unif(0.440, 0.560)
Fiebre alta	$\theta_{8,4}$	Unif(0.442, 0.558)
Falta de aliento	$\theta_{9,4}$	Unif(0.442, 0.558)
Historia familiar	$\theta_{28,4}$	Unif(0.442, 0.558)
Esputo mucoide	$\theta_{29,4}$	Unif(0.442, 0.558)

Fuente: Elaboración propia.

Tabla 3.5

Resultados de los parámetros variacionales para la enfermedad gripe común.

Síntoma	Parámetro	Distribución de densidad
Estornudos continuos	$\theta_{1,5}$	Unif(0.442, 0.558)
Escalofríos	$\theta_{3,5}$	Unif(0.442, 0.558)
Fatiga	$\theta_{5,5}$	Unif(0.027, 0.028)
Tos	$\theta_{7,5}$	Unif(0.010, 0.010)
Fiebre alta	$\theta_{8,5}$	Unif(0.442, 0.558)
Dolor de cabeza	$\theta_{11,5}$	Unif(0.442, 0.558)
Ganglios linfáticos inflamados	$\theta_{15,5}$	Unif(0.442, 0.558)
Malestar general	$\theta_{16,5}$	Unif(0.442, 0.558)
Flema	$\theta_{17,5}$	Unif(0.443, 0.557)
Irritación de garganta	$\theta_{18,5}$	Unif(0.033, 0.034)
Enrojecimiento de ojos	$\theta_{19,5}$	Unif(0.443, 0.557)
Presión sinusal	$\theta_{20,5}$	Unif(0.444, 0.556)
Nariz que moquea	$\theta_{21,5}$	Unif(0.443, 0.557)
Congestión	$\theta_{22,5}$	Unif(0.443, 0.557)
Dolor de pecho	$\theta_{23,5}$	Unif(0.443, 0.557)
Pérdida de olfato	$\theta_{25,5}$	Unif(0.443, 0.557)
Dolor muscular	$\theta_{26,5}$	Unif(0.443, 0.557)

Fuente: Elaboración propia.

3.2. Conclusiones y recomendaciones

3.2.1. Conclusiones

- El objetivo de este trabajo de integración curricular era resolver un problema de diagnóstico médico en la atención primaria. En este objetivo se buscaba modelar la relación entre la enfermedad y el síntoma, mediante un modelo gráfico pro-

babilístico, para lo cual se realizó un estudio de cada enfermedad y el número de veces que aparece el síntoma en la misma, logrando así tener un sistema que explique dichas relaciones.

- Nos interesa saber si dado un síntoma, se tiene cierta enfermedad. Debido a que el cálculo de la probabilidad *a posteriori* puede ser intratable, se propuso convertirlo en un problema de optimización mediante la inferencia variacional. Bajo el supuesto de que las densidades variacionales son independientes, derivamos las densidades variacionales óptimas para el modelo gráfico probabilístico. Además, se planteó un algoritmo y se realizó la implementación para estimar los parámetros de las densidades variacionales.

- Para las densidades variacionales se propuso trabajar con distribuciones uniformes continuas, donde los parámetros estuvieran entre 0 y 1, ya que tratamos de aproximar probabilidades. Los parámetros resultantes nos muestran el intervalo en el que se encuentra la probabilidad de tener cierta enfermedad, dado que la observación presenta el síntoma.

- La aplicación de modelos matemáticos en el diagnóstico médico se encuentra en auge. Dentro de varios enfoques que se adoptan podemos mencionar que la Inferencia Bayesiana resalta, ya que muestra ventajas con respecto a los otros, tales como la facilidad para entrenar el modelo, la inclusión de nueva información, la robustez frente a datos atípicos, entre otros.

- De la atención medica primaria no se puede deducir la enfermedad que tiene una observación solo por los síntomas, se necesitan estudios más profundos. Sin embargo, conocer la probabilidad que tiene la observación de poseer enfermedad, dado que presentó un determinado síntoma puede ser de gran utilidad para tener un esquema más claro de las posibles enfermedades que puede presentar el paciente, y así llegar a un diagnóstico más rápido.

3.2.2. Recomendaciones

- Dentro de un proyecto siempre se desea que haya una mejora continua; por lo que se recomienda a personas que tengan interés en el proyecto, estudiar a los síntomas en conjunto, ya que un solo síntoma puede tener una probabilidad insignificante para deducir si es posible tener la enfermedad o no, pero el conjunto

de síntomas puede formar una fuerte evidencia de la existencia del padecimiento en la observación.

- Se debe realizar la elección de las distribuciones variacionales con cuidado, debido a que estamos aproximando probabilidades; por ello se recomienda usar distribuciones donde el dominio esté entre 0 y 1.
- Ampliar la información de entrenamiento es algo vital. El modelo planteado se encuentra limitado en variables y observaciones. Sin embargo, la fuente de donde fue extraía la información se va ampliando mensualmente, así posteriores interesados podrán aplicar el modelo presentado a una base más robusta.

Referencias bibliográficas

Benferhat, S., Leray, P., y Tabia, K. (2020). Belief Graphical Models for Uncertainty representation and reasoning. En *A Guided Tour of Artificial Intelligence Research*, pp. 209–246. Springer.

Blei, D. M., Kucukelbir, A., y McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Española, R. A. y Madrid, E. (2001). *Diccionario de la lengua española*, volumen 22. Real academia española Madrid.

Evans, M. J. (2018). *Probabilidad y estadística: la ciencia de la incertidumbre*. Reverté.

Jensen, F. V. (1996). *An introduction to Bayesian networks*, volumen 210. UCL press London.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., y Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Koller, D. y Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lozano, M. R. (2011). El papel de las redes bayesianas en la toma de decisiones. *Recuperado de https://docplayer. es/13694103-El-papel-de-las-redesbayesianas-en-la-toma-de-decisiones-miller-rivera-lozano-miller-riveraurosario-edu-co. html*.

Ryder, T. (2021). *Variational inference for stochastic processes*. Tesis doctoral, Newcastle University.

Shachter, R. D., Andersen, S. K., y Szolovits, P. (1994). Global conditioning for probabilistic inference in belief networks. En *Uncertainty Proceedings 1994*, pp. 514–522. Elsevier.

Shenoy, P. P. (1992). Valuation-based systems for Bayesian decision analysis. *Operations research*, 40(3):463–484.

Shwe, M. A., Middleton, B., Heckerman, D. E., Henrion, M., Horvitz, E. J., Lehmann, H. P., y Cooper, G. F. (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of information in Medicine*, 30(04):241–255.

Sucar, L. E. (2021). *Probabilistic Graphical Models: principles and applications*. Springer.

Capítulo A

Anexos

A.1. Distribuciones estadísticas para modelar

Tabla A.1

Distribuciones de probabilidad usadas en el modelo.

Distribución	Función de densidad de probabilidad	Espacio de parámetros	Media	Varianza
Uniforme	$f(x) = \frac{1}{b-a}$	$-\text{ínf} < a < b < \text{ínf}$	$\frac{a+b}{2}$	$\frac{(a+b)^2}{12}$
Beta	$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$	$a > 0$ $b > 0$	$\frac{a}{a+b}$	$\frac{ab}{(a+b+1)(a+b)^2}$
Bernoulli	$f(x) = p^x \cdot q^{1-x}$	$0 \leq p \leq 1$ $q = 1-p$	p	pq
Dirichlet	$f(x) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - a}$	$k \leq 2$ α_i>0, donde $i = 1, \cdots, k$	$\frac{\alpha_i}{\sum_{j=1}^{k} \alpha_k}$	$\frac{\widehat{\alpha_i}(1-\widehat{\alpha_i})}{\alpha_0+1}$ donde $\widehat{\alpha_i} = \frac{\alpha_i}{\alpha_0}$, $\alpha_0 = \sum_{i=1}^{k} \alpha_i$
Categórica	$p(x=k) = p_k$	$k > 0$	p	p

Fuente: Elaboración propia.

A.2. Descripción de las variables de la base de datos

Tabla A.2

Indezación y descripción de las variables de síntomas.

Índice j	Etiqueta de variable	Descripción en español
1	continuous_sneezing	Estornudos continuos
2	shivering	Temblor
3	chills	Escalofríos
4	vomiting	Vómitos
5	fatigue	Fatiga
6	weight_loss	Pérdida de peso
7	cough	Tos
8	high_fever	Fiebre alta
9	breathlessness	Falta de aliento
10	sweating	Sudoración
11	headache	Dolor de cabeza
12	loss_of_appetite	Pérdida de apetito
13	mild_fever	Fiebre suave
14	yellowing_of_eyes	Ojos amarillentos
15	swelled_lymph_nodes	Ganglios linfáticos inflamados
16	malaise	Malestar general
17	phlegm	Flema
18	throat_irritation	Irritación de garganta
19	redness_of_eyes	Enrojecimiento de ojos
20	sinus_pressure	Presión sinusal
21	runny_nose	Nariz que moquea
22	congestion	Congestión
23	chest_pain	Dolor de pecho
24	fast_heart_rate	Ritmo cardíaco rápido
25	loss_of_smell	Pérdida de olfato
26	muscle_pain	Dolor muscular
27	watering_from_eyes	Lagrimeo de los ojos
28	family_history	Historia familiar
29	mucoid_sputum	Esputo mucoide
30	rusty_sputum	Esputo oxidado
31	blood_in_sputum	Sangre en esputo

Fuente: Elaboración propia.

Tabla A.3

Indezación y descripción de las variables de enfermedades.

Índice k	**Etiqueta de variable**	**Descripción en español**
1	Tuberculosis	Tuberculosis
2	Pneumonia	Neumonia
3	Allergy	Alergia
4	Bronchial Asthma	Asma Bronquial
5	Common Cold	Gripe común

Fuente: Elaboración propia.

A.3. Modelo Gráfico Probabilístico

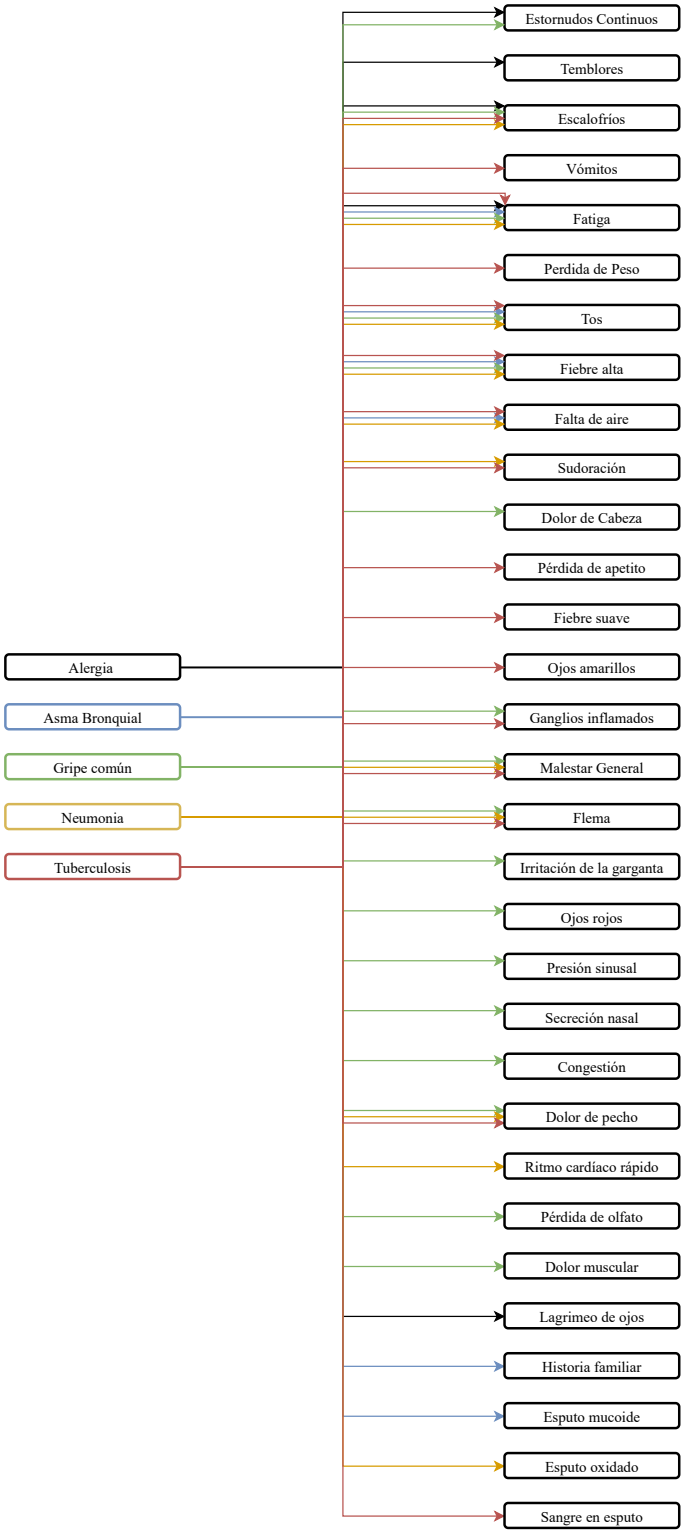


Figura A.1. Representación del Modelo Gráfico Probabilístico para 5 enfermedades y 31 síntomas
Fuente: Elaboración propia.

A.4. Implementación del algoritmo CAVI

```
rm(list = ls())

#LIBRERIAS-----------------------------------------------------------------
library(readxl)
library(tidyverse)
library(extraDistr)
library(tidyverse)
library(rootSolve)
library(latex2exp)
library(writexl)

#SEMILLA -------------------------------------------------------------------
set.seed(2511)

# LECTURA Y TRATAMIENTO DE LOS DATOS ---------------------------------------
data_entre <- read_excel("Datos/data_entre.xlsx")
unique.cols <- which(colSums(data_entre[1:132]) != 0)
datos <- data_entre %>%dplyr::select(names(unique.cols),133)
tuberculosis_1 <- datos %>% filter(prognosis=="Tuberculosis")
pneumonia_2 <- datos %>% filter(prognosis=="Pneumonia")
allergy_3 <- datos %>% filter(prognosis=="Allergy")
bronchial_asthma_4 <- datos %>% filter(prognosis=="Bronchial Asthma")
common_cold_5 <- datos %>% filter(prognosis=="Common Cold")
rm(data_entre, unique.cols, datos)

#VARIABLES INICIALES ALEATORIAS--------------------------------------------
b_i=matrix(runif(5,0.5,1),1,5)
a_i=matrix(runif(5,0,0.5),1,5)
b_jk=matrix(runif(155,0.5,1),31,5)
a_jk=matrix(runif(155,0,0.5),31,5)

#DELTA DE KRONECKER--------------------------------------------------------
delta_kro_enf <- matrix(ncol = 5, nrow = 31)
for (i in 1:31) {
  delta_kro_enf[i,1]=ifelse(sum(tuberculosis_1[i])>0,1,0)
  delta_kro_enf[i,2]= ifelse(sum(pneumonia_2[i])>0,1,0)
  delta_kro_enf[i,3]= ifelse(sum(allergy_3[i])>0,1,0)
  delta_kro_enf[i,4]= ifelse(sum(bronchial_asthma_4[i])>0,1,0)
  delta_kro_enf[i,5]= ifelse(sum(common_cold_5[i])>0,1,0)
}

```

```
#CONSTANTES c_jk Y d_jk-------------------------------------------------------
c_data= matrix(nrow = 31, ncol = 5)
tuberculosis.1=tuberculosis_1[1:31]
tuberculosis.1=colSums(tuberculosis.1)
pneumonia.1=pneumonia_2[1:31]
pneumonia.1=colSums(pneumonia.1)
allergy.1=allergy_3[1:31]
allergy.1=colSums(allergy.1)
bronchial_asthma.1=bronchial_asthma_4[1:31]
bronchial_asthma.1=colSums(bronchial_asthma.1)
common_cold.1=common_cold_5[1:31]
common_cold.1= colSums(common_cold.1)

for(j in 1:31){
  c_data[j,1]=tuberculosis.1[j]
  c_data[j,2]=pneumonia.1[j]
  c_data[j,3]=allergy.1[j]
  c_data[j,4]=bronchial_asthma.1[j]
  c_data[j,5]= common_cold.1[j]
}
c_data
d_data= matrix(nrow = 31, ncol = 5)
tuberculosis.2=1-tuberculosis_1[1:31]
tuberculosis.2=colSums(tuberculosis.2)
pneumonia.2=1-pneumonia_2[1:31]
pneumonia.2=colSums(pneumonia.2)
allergy.2=allergy_3[1:31]
allergy.2=colSums(allergy.2)
bronchial_asthma.2=bronchial_asthma_4[1:31]
bronchial_asthma.2=colSums(bronchial_asthma.2)
common_cold.2=common_cold_5[1:31]
common_cold.2= colSums(common_cold.2)


for(j in 1:31){
  d_data[j,1]=tuberculosis.2[j]
  d_data[j,2]=pneumonia.2[j]
  d_data[j,3]=allergy.2[j]
  d_data[j,4]=bronchial_asthma.2[j]
  d_data[j,5]= common_cold.2[j]
}

d_data
```

```
85
86 #ELBO ------------------------------------------------------------------------
87 ELBO <- function(a_enf,b_enf,a_sint,b_sint){
88   a=(a_enf*log(a_enf)-b_enf*log(b_enf)-a_enf+b_enf)/(b_enf-a_enf)
89   a=-sum(a, na.rm = TRUE)
90   b.1=1*((a_enf*log(a_enf)-b_enf*log(b_enf)-a_enf+b_enf)/(b_enf-a_enf) )
91   b.2=2*((a_enf*log(a_enf)-b_enf*log(b_enf)-a_enf+b_enf)/(b_enf-a_enf) )
92   b.3=3*((a_enf*log(a_enf)-b_enf*log(b_enf)-a_enf+b_enf)/(b_enf-a_enf) )
93   b.4=4*((a_enf*log(a_enf)-b_enf*log(b_enf)-a_enf+b_enf)/(b_enf-a_enf) )
94   b.5=5*((a_enf*log(a_enf)-b_enf*log(b_enf)-a_enf+b_enf)/(b_enf-a_enf) )
95   b=-sum(b.1,na.rm = TRUE)+sum(b.2,na.rm = TRUE)+sum(b.3, na.rm = TRUE)+
96     sum(b.4, na.rm = TRUE)+sum(b.5, na.rm = TRUE)
97   c.1=delta_kro_enf*((a_sint*log(a_sint)-a_sint-b_sint*log(b_sint)+b_sint)/
98                        (b_sint-a_sint))
99   c.1=ifelse(c.1==-Inf,0,ifelse(c.1==Inf,1,c.1))
100  c.2=delta_kro_enf*((a_sint*log(1-a_sint)-log(1-a_sint)+b_sint+log(1-b_sint)-
101                       a_sint-b_sint*log(1-b_sint))/(b_sint-a_sint))
102  c.2=ifelse(c.2==Inf,1,c.2)
103  c.3=delta_kro_enf*2*log(pi)
104  c=-sum(c.1,na.rm = TRUE)+sum(c.2,na.rm = TRUE)-sum(c.3)
105  c=0
106  d.1=-c_data*((a_sint*log(a_sint)-b_sint*log(b_sint)-a_sint+b_sint)/(b_sint-a_
       sint))
107  d.2=-d_data*(log(1-b_sint)-b_sint*log(1-b_sint)+b_sint+a_sint*log(1-a_sint)-
108                a_sint-log(1-a_sint))/(b_sint-a_sint)
109  d=-sum(delta_kro_enf*d.1,na.rm = TRUE)+sum(delta_kro_enf*d.2, na.rm = TRUE)
110  e=ifelse(abs(b_enf-a_enf)==0,0,log(abs(b_enf-a_enf)))
111  e=-sum(e, na.rm = TRUE)
112  f.1=ifelse(abs(b_sint-a_sint)==0,0,log(abs(b_sint-a_sint)))
113  f=delta_kro_enf*f.1
114  f=-sum(f, na.rm = TRUE)
115  return(-9.52638+a-b+c+d-e-f)
116 }
117
118 #FUNCIONES DE LAS ACTUALIZACIONES-----------------------------------------------
119 a_1 <- function(x,b=b_i[1],z=1){
120   a=-b*z+b*z*log(b)+b*log(x)-b*z*log(x)-b*log(b)+x*z
121   return(a)
122 }
123 a_2 <- function(x,b=b_i[2],z=2){
124   a=-b*z+b*z*log(b)+b*log(x)-b*z*log(x)-b*log(b)+x*z
125   return(a)
126 }
```

```
a_3 <- function(x,b=b_i[3],z=3){
  a=-b*z+b*z*log(b)+b*log(x)-b*z*log(x)-b*log(b)+x*z
  return(a)
}
a_4 <- function(x,b=b_i[4],z=4){

  a=-b*z+b*z*log(b)+b*log(x)-b*z*log(x)-b*log(b)+x*z
  return(a)
}
a_5 <- function(x,b=b_i[5],z=5){
  a=-b*z+b*z*log(b)+b*log(x)-b*z*log(x)-b*log(b)+x*z
  return(a)
}



b_1 <- function(x,a=a_i[1],z=1){
  x*z+a*log(x)-a*z*log(x)-a*z-a*log(a)+a*z*log(a)
}
b_2 <- function(x,a=a_i[2],z=2){
  x*z+a*log(x)-a*z*log(x)-a*z-a*log(a)+a*z*log(a)
}
b_3 <- function(x,a=a_i[3],z=3){
  x*z+a*log(x)-a*z*log(x)-a*z-a*log(a)+a*z*log(a)
}
b_4 <- function(x,a=a_i[4],z=4){
  x*z+a*log(x)-a*z*log(x)-a*z-a*log(a)+a*z*log(a)
}
b_5 <- function(x,a=a_i[5],z=5){
  x*z+a*log(x)-a*z*log(x)-a*z-a*log(a)+a*z*log(a)
}

a_jk.funcion <- function(x,b,c,d){
  a=c*b*log(b)-b*c-b*d+b*d*log(1-b)+0.5*b*log(x)+0.5*b*log(1-x)-b*c*log(x)-b*d*
      log(1-x)-
    0.5*b*log(b)-0.5*b*log(1-b)-d*log(1-b)+d*log(1-x)+0.5*log(1-b)-0.5*log(1-x)+
        x*c+x*d
  return(round(a,2))
}


b_jk.function <- function(x,a,c,d){
  b=0.5*a*log(x)+0.5*a*log(1-x)-0.5*a*log(a)-0.5*a*log(1-a)-c*(a*log(x)+
```

```
    a-a*log(a)-x)-d*(a*log(1-x)+a-a*log(1-a)+log((-a+1)/(-x+1))-x)+0.5*log(1-a)-
    0.5*log(1-x)
  return(round(b,2))
}

#CAVI ----------------------------------------------------------------------
iter=350
elbos= rep(NA, iter+1)
diferencia = rep(NA, iter)
elbos[1]=ELBO(a_enf = a_i, b_enf = b_i,a_sint = a_jk, b_sint = b_jk)
a_enfermedad=rep(NA, iter+1)
a_enfermedad[1]=a_i[2]
for (k in 1:iter) {
  b_1.act=max(uniroot.all(b_1,c(0,1)))
  b_2.act=max(uniroot.all(b_2, c(0,1)))
  b_3.act=max(uniroot.all(b_3,c(0,1)))
  b_4.act=max(uniroot.all(b_4, c(0,1)))
  b_5.act=max(uniroot.all(b_5, c(0,1)))
  a_1.act=max(uniroot.all(a_1, c(0,1)))
  a_2.act=max(uniroot.all(a_2, c(0,1)))
  a_3.act=max(uniroot.all(a_3, c(0,1)))
  a_4.act=max(uniroot.all(a_4, c(0,1)))
  a_5.act=max(uniroot.all(a_5, c(0,1)))
  b_i[1]=b_1.act
  b_i[2]=b_2.act
  b_i[3]=b_3.act
  b_i[4]=b_4.act
  b_i[5]=b_5.act

  a_i[1]=a_1.act
  a_i[2]=a_2.act
  a_i[3]=a_3.act
  a_i[4]=a_4.act
  a_i[5]=a_5.act

  b_jk.new=matrix(nrow = 31,ncol = 5)
  a_jk.new=matrix(nrow = 31,ncol = 5)

  for (i in 1:31) {
    for (j in 1:5) {
      m=b_jk.function(seq(0,1,0.001),a = a_jk[i,j],c = c_data[i,j], d = d_data[i
          ,j])
      m.1=seq(0,1,0.001)[max(which(m==0))]
```

```
        b_jk.new[i,j]=m.1
        n=a_jk.funcion(seq(0,1,0.001),b = b_jk.new[i,j],c = c_data[i,j], d = d_
            data[i,j])
        n.1=seq(0,1,0.001)[min(which(n==0))]
        a_jk.new[i,j]=n.1

      }
    }
    a_jk=a_jk.new
    b_jk=b_jk.new
    elbos[k+1]=ELBO(a_enf = a_i, b_enf = b_i,a_sint = a_jk, b_sint = b_jk)
    diferencia[k]=abs(elbos[k+1]-elbos[k])
    a_enfermedad[k+1]=a_i[2]
    cat("Iteration: ",k, "Elbo: ", elbos[k+1], "\n")
    if(abs(elbos[k+1]-elbos[k])<0.1) break
}
a_jk=delta_kro_enf*a_jk
b_jk=delta_kro_enf*b_jk

#RESULTADOS-----------------------------------------------------------------
ELBO
diferencia

###ACTUALIZACIONES PARA PI-------------------------------------------------
a_i
b_i

###ACTUALIZACIONES PARA THETA----------------------------------------------
a_jk
b_jk
```

Listing A.1: Implementación en R del algoritmo CAVI