ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

APLICACIONES DE MÉTODOS VARIACIONALES PARA INFERENCIA ESTADÍSTICA BAYESIANA.

COMPONENTE: INFERENCIA VARIACIONAL PARA ESTIMACIÓN DE PARÁMETROS DE MODELOS CON ESTADOS OCULTOS DE MÁRKOV

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

DANIEL ARTURO DÍAZ QUICHIMBO

daniel.diaz@epn.edu.ec

DIRECTOR: CARLOS ALBERTO ALMEIDA RODRIGUEZ

carlos.almeidar@epn.edu.ec

DMQ, SEPTIEMBRE 2022

CERTIFICACIONES

Yo, DANIEL ARTURO DÍAZ QUICHIMBO, 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.

DANIEL ARTURO DÍAZ QUICHIMBO

Certifico que el presente trabajo de integración curricular fue desarrollado por DANIEL ARTURO DÍAZ QUICHIMBO, bajo mi supervisión.

Firmado electrónicamente por:
CARLOS ALBERTO ALMEIDA RODRIGUEZ

CARLOS ALBERTO ALMEIDA RODRIGUEZ
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el producto resultante del mismo, es público y estará 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.

DANIEL ARTURO DÍAZ QUICHIMBO

CARLOS ALBERTO ALMEIDA RODRIGUEZ

RESUMEN

El análisis de series de tiempo consiste en predecir el siguiente valor en una secuencia dada en función de lo observado anteriormente. La predicción puede ser la continuación: un símbolo, un número, el clima del día siguiente, el siguiente término en el habla, etc.

Los Modelos Ocultos de Márkov, (HMM) son modelos matemáticos de procesos de Márkov con estados ocultos. Además, es un modelo estadístico que se usa ampliamente para datos que tienen continuidad y extensibilidad, como el análisis de mercado de valores de series temporales [20],[18],[26],[9].

La inferencia variacional (VI) es un método para aproximar una densidad condicional de variables latentes o parámetros dadas las variables observadas. Se utiliza ampliamente para aproximar las densidades posteriores de los modelos Bayesianos [3],[12],[23]. Conceptualmente, VI funciona eligiendo una familia de funciones de densidad de probabilidad y luego encontrando la más cercana a la densidad de probabilidad real, a menudo usando la divergencia Kullback-Leibler (KL) como la métrica de optimización.

Los modelos Bayesianos brindan herramientas para analizar datos de series temporales. Sin embargo, su aplicación en series temporales no ha sido estudiada con mucha frecuencia.

En este documento, presentamos los conceptos de la Inferencia Variacional (VI) y Modelos ocultos de Márkov (HMM), luego describimos un procedimiento Bayesiano de estimación e inferencia para series temporales financieras basándonos en el uso de Inferencia Variacional en Modelos Ocultos de Márkov. Utilizando probabilidades de transición y probabilidades de emisión, se ajusta un modelo para series de tiempo financieras. Se ajusta un modelo oculto de Márkov y se estiman sus parámetros.

Palabras clave: Modelos Ocultos de Markov, Inferencia Variacional, Divergencia Kullback-Leibler, Modelos Bayesianos, Mercado de Valores.

ABSTRACT

Time series analysis consists of predicting the next value in a given sequence based on what has been observed previously. The prediction can be the continuation: a symbol, a number, the next day's weather, the next term in speech, etc.

Hidden Markov Models, (HMM) are mathematical models of Markov processes with hidden states. In addition, it is a statistical model that is widely used for data that have continuity and extensibility, such as time series stock market analysis [20],[18],[26],[9].

Variational inference (VI) is a method for approximating a conditional density of latent variables or parameters given observed variables. It is widely used to approximate the posterior densities of Bayesian [3],[12],[23] models. Conceptually, VI works by choosing a family of probability density functions and then finding the one closest to the actual probability density, often using Kullback-Leibler (KL) divergence as the optimization metric.

Bayesian models provide tools for analyzing time series data. However, their application in time series has not been studied very often.

In this paper, we introduce the concepts of Variational Inference (VI) and Hidden Markov Models (HMM), then describe a Bayesian estimation and inference procedure for financial time series based on the use of Variational Inference in Hidden Markov Models. Using transition probabilities and emission probabilities, a model for financial time series is fitted. A hidden Markov model is fitted and its parameters are estimated.

Keywords: Hidden Markov Model, Variational Inference, Kullback-Leibler Divergence, Bayessian Models, Stock market analysis.

Índice general

Índice de figuras

Capítulo 1

Descripción del componente desarrollado

En este documento, presentamos el concepto de inferencia variacional (VI, por sus siglas en inglés). Conceptualmente, VI funciona eligiendo una familia de funciones de densidad de probabilidad condicional de las variables latentes (ocultas) o parámetros dadas las observadas y luego encontrando la más cercana a la densidad de probabilidad a posterior del modelo, a menudo usando la divergencia Kullback-Leibler (KL) como la métrica de optimización. Además es ampliamente usada para aproximar las densidades posteriores de los modelos Bayesianos [14]. Con la VI se puede trabajar de forma más rápida en grandes cantidades de datos [5], se ha aplicado a problemas de neurociencia computacional y la visión por computadora [6]

Por otro lado, los modelos ocultos de Márkov (HMM, por sus siglas en inglés) se utilizan ampliamente para modelar datos de series temporales; una de sus aplicaciones incluye la predicción de series temporales. Entre los primeros trabajos con HMM se encuentran el propuesto por [19] y una tesis doctoral de [35]; y se han aplicado en la segmentación de series temporales [37]. Los modelos de Márkov son frecuentemente usados en las ciencias sociales, en diferentes áreas y aplicaciones. En psicología, han sido usados para modelar procesos de aprendizaje [24],[4],[24],[29]. En Economía, los modelos latentes de Márkov han sido usados para modelar el cambio de régimen [16]. La predicción de tendencias de series de precios financieros es un problema esencial que se ha discutido amplia-

mente utilizando herramientas y técnicas de Física, Economía y Aprendizaje Automático [14],[24],[4]. La dependencia del tiempo y los problemas de volatilidad han hecho que los HMM sean una herramienta útil para predecir los estados de mercado de valores [20].

La Inferencia Variacional en Modelos Ocultos de Márkov se ha estudiado con menos frecuencia, a pesar que esta tiene propiedades interesantes con respecto a los métodos de estimación, de tal forma que vale la pena explorar más a fondo, se ha desarrollado modelos ocultos de Márkov con Inferencia variacional en Mercadeo por [7], se propone la IV en HMM aplicada a la finanzas.

1.1. Objetivo general

Utilizar los métodos de inferencia Variacional para ajustar un modelo oculto de Márkov aplicada a las finanzas.

1.2. Objetivos específicos

1. Realizar una investigación sobre la inferencia variacional aplicada a métodos ocultos de Márkov.
2. Estudiar los métodos que se han empleado [9], [15], [20], el modelo propuesto por [7], [23] y [13]
3. Realizar un análisis descriptivo de las series de tiempo S&P 500, DAX y Coca Cola.
4. Ajustar un modelo oculto de Márkov con los métodos de inferencia variacional y aplicarlo a series de tiempo financieras.
5. Validar el modelo que se construye

1.3. Alcance

Para alcanzar nuestro objetivo, es necesario adquirir un conocimiento en Inferencia variacional para modelos ocultos de Márkov, así que se

comenzará a estudiar los conceptos y resultados de modelos ocultos de Márkov con Inferencia variacional aplicados a series de tiempo. Se buscará los algoritmos que han sido implementados, uno de ellos es [13] que nos habla sobre inferencia Variacional estocástica para modelos bayesianos de series de tiempo, además [9] nos presenta modelos bayesianos en series de tiempo.

El algoritmo que se implementará con la Inferencia Variacional en Modelos Ocultos de Márkov aproximará la densidad a posterior, tomando en cuenta tres estados ocultos.

Los datos históricos se obtendrán de los siguientes enlaces https://finance.yahoo.com/quote/%5EGSPC/history índice bursátil S&P 500, https://www.eoddata.com/quote.aspx para el Índice DAX y https://finance.yahoo.com/quote/KO para las acciones de Coca Cola; cada uno de ellos obtenido en un periodo mensual desde Enero del 2000 a diciembre del 2021, una vez obtenida la base de datos de los índices bursátiles, vamos a tomar los datos desde Enero del 2000 hasta diciembre del 2020 en una base de datos y otra de Enero del 2021 hasta diciembre del 2021 la que se utilizará para la validación del modelo.

Realizaremos un análisis de datos de los índices S&P 500, DAX y los precios de las acciones de Coca Cola.

- Análisis previo de los datos.

  Previo a implementar los algoritmos es necesario realizar un análisis de los datos con el fin de que se encuentren en las mejores condiciones y específicas, Para ello se realiza lo siguiente:

  Análisis exploratorio de los datos

  Verificar que se cumplan las condiciones para aplicar los algoritmos, es decir, (estacionariedad) y si esto no se cumple, usar los recursos necesarios para tenerlas aplicables al modelo, todo lo mencionado anteriormente se aplicará a la base de datos en el periodo (enero 2000 – diciembre 2020).

- Métodos de Estudio.

  Para seleccionar el modelo a utilizar nos guiaremos en trabajos anteriores [7] y [13], entre otros, para luego aplicarlo a nuestra base

(enero 2000-diciembre 2021).

Una vez obtenido el modelo se aplicará a los índices bursátiles S&P 500, DAX y las acciones de Coca Cola. Para verificar su efectividad; se usarán los criterios de información de Akaike y Bayesiano; y se considerará el criterio de desviación de información propuesto por [23].

1.4. Marco teórico

La predicción del mercado de valores ha sido una área de investigación mas activas, Los HMM son capaces de modelar transacciones de estado ocultas a partir de los datos secuenciales observados.

1.4.1. Series de Tiempo

En esta parte del documento, mostraremos los aspectos más generales de una serie de tiempo y describiremos brevemente los índices bursátiles que serán utilizados en el presente trabajo, basándonos en las notaciones de [23], [7]

Una serie de Tiempo es una sucesión de observaciones correspondientes a una variable tomada en periodos de tiempo regulares y de duración constante. Wikipedia

Componentes de una serie de Tiempo

- Tendencia
- Ciclos
- Estacional
- Aleatoriedad

Las series de tiempo tienen una infinidad de aplicaciones en diversas áreas, como por ejemplo.

- Economía

- Demografía
- Meteorología
- Finanzas (Los índices bursátiles diarios de los últimos 10 años)

El ultimo será de nuestro interés, dado que a lo largo de este trabajo serán utilizados los índices bursátiles (S&P 500 y DAX)

Series de Tiempo Financieras

Son una sucesión de observaciones en periodo de tiempo determinado, las cuales pueden ser, por ejemplo, precios, rendimiento de activos financieros: la volatilidad es una característica de las series de tiempo. Se puede destacar que la volatilidad es la desviación estándar de los rendimientos del proceso [17]

Como vamos ha trabajar con datos financieros, generalmente se utiliza el *return*, dado que los retornos presentan características más interesantes que los precios, como por ejemplo la estacionariedad.

Sea P_t el precio de un activo al tiempo t; el rendimiento se define como:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{1.1}$$

Pero esta fórmula es para un solo periodo.

Ahora vamos a introducir el concepto del logaritmo natural del rendimiento simple (Continuosly Compounted Return o log return) y se define como:

$$r_t = \log\left[\frac{P_t}{P_{t-1}}\right] \tag{1.2}$$

Para más información de esta transformación dentro de una serie de tiempo financiera, ir a [1]

A partir de este momento vamos ha definir como:

$$X_t = \log\left[\frac{P_{t+1}}{P_t}\right] \tag{1.3}$$

a los datos observados, que más adelante serán usados en el modelamiento de los HMM con IV.

1.4.2. Índices bursátiles

Un índice bursátil es un valor promedio de un conjunto de acciones determinado. La finalidad de estos índices es evidenciar los cambios en el tiempo de los precios de las acciones que lo componen. Usualmente, estas acciones comparten ciertas características como pertenecer a una misma bolsa de valores o pertenecer a la misma industria.

Standard $\&$ Poor 500

El índice bursátil que vamos ha estudiar es el Standard $\&$ Poor's 500 $(S\&P500)$ que es uno de los índices mas importantes de Estados Unidos. Se basa en la capitalización de mercado de 500 empresas que cotizan en New York Stock Exchange (NYSE). El $(S\&P500)$ comenzó a registrarse el 04 de junio de 1968: este índice es de suma importancia a nivel mundial ya que juega en la plaza financiera más importante del mundo, es decir, Wall Street.

El cálculo se lo realiza de la siguiente manera:

$$\text{Index} = \frac{\sum_{i=1}^{n}(P_i * Q_i)}{Divisor} \tag{1.4}$$

donde:

- P_i corresponde al precio de cada acción
- Q_i corresponde el número de acciones disponibles para cada acción.
- DIVISOR corresponde a una cifra patentada por Standard $\&$ Poor's, que se ajusta dependiendo de si se trata de divisiones de acciones, dividendos especiales o una escisión, esto con el fin de que no se altere el índice por factores no económicos.

Deutscher Aktienindex, DAX

El otro índice ha analizarse será el DAX, el cuál es el más importante de Alemania, que cotiza la bolsa de Fráncfort, que se compone de las 30 empresas más importantes de Alemania, [8] describe de forma mas clara y precisa este índice.

Comenzó a ser registrado el 1 de julio de 1998, y desde el 21 de junio de 1999 las operaciones comenzaron a realizarse en una plataforma

electrónica llamada XETRA, Para la elaboración del indice se utilizan los precios registrados de la plataforma XETRA, con una condición, que la ponderación de cada una no sea mayor al $10\,\%$.

Para calcular el DAX se utiliza la formula del índice de Laspeyres, de la siguiente manera:

$$\text{Index}_t = K_t \left(\frac{\sum_{i=1}^{n} p_{it} q_{iT} c_{it}}{\sum_{i=1}^{n} p_{i0} q_{i0}} \text{Base} \right) \tag{1.5}$$

donde:

- c_{it} corresponde al factor de ajuste de la empresa i al tiempo t.
- ff_{iT} corresponde al factor *free float* de la clase de acciones i al tiempo t.
- p_{i0} corresponde al precio de cierre de la acción i
- p_{it} corresponde al precio de la acción i al tiempo t.
- q_{i0} corresponde al precio de cierre de la acción i
- q_{iT} corresponde al precio de la acción i al tiempo T.
- K_T corresponde al factor de encadenamiento a partir del tiempo T
- *Base* corresponde a una constante.

Coca Cola

La Compañía Coca-Cola es una corporación multinacional estadounidense de bebidas, con su sede en Atlanta, Georgia, la misma tiene intereses en la fabricación, venta minorista y comercialización de concentrados y jarabes para bebidas no alcohólicas. La compañía produce Coca-Cola, inventada en 1886 por el farmacéutico John Stith

El precio histórico de las acciones será evaluado en nuestro modelo.

1.4.3. Modelos Ocultos de Markov

Se define formalmente los Modelos Ocultos de Markov (HMM) y se explica sus propiedades. La teoría y notación en base a estos autores

[36], [2] y [37]

Cadenas de Markov

Definición 1 (Cadenas de Markov) Una secuencia de variables aleatorias discretas $\{C_t\}_{t\in\mathbf{N}}$, se dice que es una cadena de Markov (CM) a tiempo discreto, si para todo $t \in \mathrm{N}$, satisface la siguiente propiedad de Markov.

$$P(C_t|C_1, C_2, C_3, \ldots, C_{t-1}) = P(C_t|C_{t-1}) \tag{1.6}$$

Una cadena de Markov es un proceso estocástico que satisface (1.6). Esto significa que si una serie de variables aleatorias constituye una Cadena de Markov, el estado futuro de cada una de esas variables solo depende directamente de su estado actual.

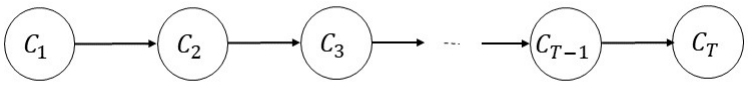


Figura 1.1: Grafo de una Cadena de Markov

HMM

El modelo oculto de Markov (HMM) puede superar algunas de las limitaciones que enfrentan las cadenas de Markov. Como sugiere el nombre, el estado no es directamente observable en este modelo. Sin embargo, las observaciones (que son directamente observables, por definición) están ligadas a cada estado por una distribución de probabilidad. Además los HMM satisfacen la propiedad de Markov (1.6),

Los modelos ocultos de Markov (HMM) son una secuencia de observaciones x_t junto a una secuencia estados ocultos $z_t \in \{1, \ldots, T\}$, los cuales están generados por un Procesos de Markov.

La siguiente notación se usará formalmente para describir los HMMs,

$T =$ longitud de lo observado y secuencia de estados ocultos

$N =$ número de estados ocultos

$S =$ distintos estados ocultos $\{1, 2, 3, \ldots, N\}$

$X =$ Secuencia de observaciones $X = \{x_1, x_2, \ldots, x_t\}$

$Z =$ Secuencia de estados ocultos $Z = \{z_1, z_2, \ldots, z_t\}$

$\pi =$ Probabilidades iniciales de los estados ocultos

$A =$ Probabilidades de transición de los estados ocultos

$B =$ Distribuciones de probabilidad de observación o distribuciones de emisión

Los parámetros de HMM son denotados como θ, con $\theta = (\pi, A, B)$, cada uno tiene las siguientes notaciones.

$$\pi = \{\pi_i\} : \pi_i = P(z_1 = i)$$
$$A = \{a_{ij}\} : a_{ij} = P(z_t = j | z_{t-1} = i)$$
$$B = \{b_i(x_j)\} : b_i(x_j) = P(X_t \leq x_j | Z_t = z_i)$$

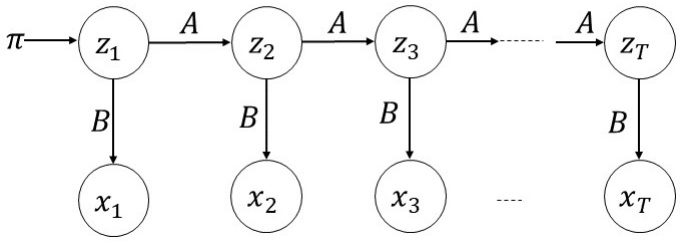


Figura 1.2: Representación gráfica de un Modelo Oculto de Markov

La probabilidad conjunta de una secuencia de longitud T viene dada por:

$$P\left(\theta, \mathbf{Z}, \mathbf{X}\right) = p(\theta)p\left(z_1, \pi\right)p\left(x_1 \mid z_1, B\right)\prod_{t=2}^{T} p\left(z_t \mid z_{t-1}, A\right)p\left(x_t \mid z_t, B\right) \tag{1.7}$$

Estimación de parámetros iniciales

Un HMM requiere valores de parámetros iniciales para $\theta = (\pi, A, B)$. La mayoría de las fuentes recomiendan inicializar π y A con valores uniformes iguales a $\frac{1}{N}$ o valores extraídos al azar de un distribución uniforme [29],[4].

Para las distribuciones de emisiones, es importante que las distribuciones de emisiones se inicialicen al azar, o infiriendo una estimación a partir de los datos. Para distribuciones de emisiones continuas como una Poisson, Gaussiana o una combinación de gaussianas, normalmente se utiliza el agrupamiento de K-Means para proporcionar estimaciones [37].

1.4.4. Enfoque Bayesiano para HMM

Un enfoque bayesiano para calcular $P(\theta, Z \mid X)$, la distribución posterior de los parámetros del modelo y la secuencia de estados ocultos dadas las observaciones, puede proporcionar respuestas a algunos problemas. Al calcular la distribución sobre los parámetros del modelo, es sencillo lograr una medida de la incertidumbre que rodea a los valores de los parámetros. Esto tiene el beneficio adicional de eliminar el problema de los mínimos locales para los valores de los parámetros. La selección de modelos bayesianos se puede realizar mediante el criterio de desviación de información (DIC, por sus siglas en inglés), que ayuda a seleccionar el mejor modelo de un conjunto de candidatos [32],[23]. Un método para calcular HMM de manera bayesiana se basa en el implementación de Cadenas de Markov Monte Carlo (MCMC, por sus siglas en Ingles). El entrenamiento de MCMC de los HMM como una alternativa entre el muestreo de parámetros del modelo en función de los datos observados y el muestreo de los estados ocultos no observados, lo realizó [30], [31].

Un enfoque alternativo al calculo de MCMC es usar VI para estimar las distribuciones posteriores de los parámetros, intercambiando un poco

de precisión por una velocidad computacional mucho mejor [7]. Además, los HMM calculados a través de VI muestran una tendencia a eliminar estados ocultos innecesarios, proporcionando un beneficio adicional para seleccionar un modelo apropiado. Se pueden encontrar excelentes introducciones a VI en Murphy [25],[4], y [5]. MacKay [21] fue uno de los primeros en introducir un enfoque variacional para los HMM, y Beal [3] amplía este trabajo al evaluar los HMM estimados con VI en un conjuntos de datos de muestra.

McGrory y Titterington [23] calcularon HMM con probabilidades de emisiones gaussianas utilizando la inferencia VI para una variedad de conjuntos de datos de muestra, lo que ha demuestrado el efecto de eliminación de estado de los HMM entrenados de forma variable. B. y Carin [11] proporcionaron una derivación para HMM con emisiones de modelo de mezcla gaussiana. Watanabe y Minami [34] utilizaron técnicas variacionales para aplicar HMM al problema del reconocimiento de voz con miras a facilitar el proceso de selección del modelo. Johnson [12] proporcionó una comparación robusta de HMM implementado con Expectation Maximization. Este trabajo encontró que los HMM calculados con VI se desempeñaron mejor.

La mayor parte de este trabajo existente con VI y HMM se ha realizado utilizando una combinación de estilos de notación.

1.4.5. Inferencia Variacional

Revisaremos como se puede aplicar los métodos variacionales para aproximar las probabilidades a posterior de los HMM. Asumimos un modelo paramétrico con parámetros θ, Z las variables no observadas y X las variables observadas. La Inferencia Bayessiana se enfoca en la distribución posterior $P(\theta|X)$. La distribución posterior $P(\theta|X)$ es la distribución marginal apropiada de $P(\theta, Z|X)$. El enfoque variacional de Bayes nos permite aproximar la cantidad compleja $P(\theta, Z|X)$ por una distribución más simple, $Q(\theta, Z)$.

La optimización directa de $P(X|\theta)$ es difícil, pero la optimización de la función de probabilidad conjunta $P(X, Z, \theta)$ se escribe:

$$P(\theta, Z, X) = P(\theta, Z \mid X)P(X) \tag{1.8}$$

El $P(X)$ representa la evidencia proporcionada por los datos observados.

$$P(X) = \int_{\theta,Z} p(\theta, Z, X)d(\theta, Z) \tag{1.9}$$

El enfoque de la VI para calcular $P(\theta, Z|X)$ evita esta integral complicada de obtener, calculando en su lugar una distribución aproximada a $Q(\theta, Z)$ de una familia de distribuciones $\mathcal{Q}$ tal que este lo más cerca posible a la distribución posterior $P(\theta, Z|X)$ como se puede observar en la figura 1.3. $Q(\theta, Z)$ definida sobre las variables ocultas, se hace lo más cercano posible a la distribución deseada $P(\theta, Z|X)$ usando la divergencia $KL(Q(\theta, Z)||P(Z|X))$ de Kullback-Leibler (KL). Se utilizara la notación de [4]. Además, se puede ver más acerca de (KL) en el anexo A.

Se define la ecuación de $Q^*(\theta, Z)$ distribución aproximada como:

$$Q^*(\theta, \mathbf{Z}) = \arg \min_{Q(\theta,\mathbf{Z})\in\mathcal{Q}} D_{\mathrm{KL}}[Q(\theta, \mathbf{Z})\|P(\theta, \mathbf{Z} \mid \mathbf{X})] \tag{1.10}$$

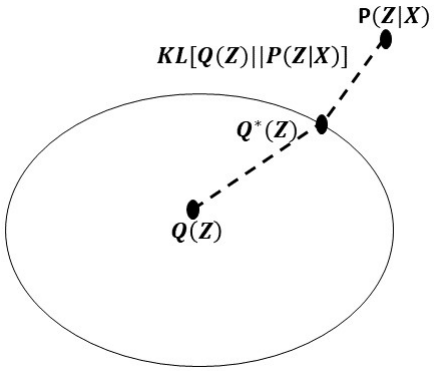


Figura 1.3: Diagrama de la Inferencia Variacional

Notemos que,

$$\log P(X|\theta) = \mathcal{L}(Q(\theta, Z)) + KL[Q(\theta, Z)||P(\theta, Z|X)] \tag{1.11}$$

donde definimos

$$\mathcal{L}(Q(\theta, Z)) = \int q(\theta, Z) \log \left\{ \frac{p(\theta, Z, X)}{q(\theta, Z)} \right\} dZ \tag{1.12}$$

$$KL(Q||P) = -\int q(\theta, Z) \log \left\{ \frac{p(X, Z|\theta)}{q(\theta, Z)} \right\} dZ \tag{1.13}$$

La meta de minimizar $KL(Q(Z)\|P(Z \mid X))$ es equivalente a maximizar $\mathcal{L}(Q(Z))$, como se puede observar en 1.11, cuando $Q(\theta, Z)$ se aproxima mejor a la $P(\theta, Z|X)$ posterior, el valor de KL podría desaparecer. Verifiquemos 1.11

$$\begin{aligned}
\log P(X) &= \mathcal{L}(Q(\theta, Z)) + \mathbb{KL}(Q(\theta, Z)\|P(\theta, Z \mid X)) \\
&= \int_{\theta,Z} q(\theta, Z) \log \left\{ \frac{p(\theta, Z, X)}{q(\theta, Z)} \right\} d(\theta, Z) \\
&- \int_{\theta,Z} q(\theta, Z) \log \left\{ \frac{p(\theta, Z \mid X)}{q(\theta, Z)} \right\} d(\theta, Z)
\end{aligned}$$

Usando la regla de la cadena:

$$\begin{aligned}
&= \int_{\theta,Z} Q(\theta, Z) \log \left\{ \frac{P(\theta, Z \mid X)P(X)}{Q(\theta, Z)} \right\} d(\theta, Z) \\
&- \int_{\theta,Z} Q(\theta, Z) \log \left\{ \frac{P(\theta, Z \mid X)}{Q(\theta, Z)} \right\} d(\theta, Z) \\
&= \int_{\theta,Z} Q(\theta, Z) \log \left\{ \frac{P(\theta, Z \mid X)}{Q(\theta, Z)} \right\} d(\theta, Z) \\
&+ \int_{\theta,Z} Q(\theta, Z) \log P(X) d(\theta, Z) \\
&- \int_{\theta,Z} Q(\theta, Z) \log \left\{ \frac{P(\theta, Z \mid X)}{Q(\theta, Z)} d(\theta, Z) \right\}
\end{aligned}$$

$\int_{\theta,Z} Q(\theta, Z) \log P(X) d(\theta, Z)$ es una constante:

$$\begin{aligned}
&= \int_{\theta,Z} q(\theta, Z) \log \left\{ \frac{p(\theta, Z \mid X)}{q(\theta, Z)} \right\} d(\theta, Z) + \log P(X) \\
&- \int_{\theta,Z} q(\theta, Z) \log \left\{ \frac{p(\theta, Z \mid X)}{q(\theta, Z)} \right\} d(\theta, Z) \\
&= \log P(X)
\end{aligned}$$

Así $\mathcal{L}(Q(\theta, Z))$ seria:

$$\mathcal{L}(q(\theta, Z)) = \int_{\theta,Z} Q(\theta, Z) \log \left\{ \frac{P(X, \theta, Z)}{q(\theta, Z)} \right\} d(\theta, Z) \tag{1.14}$$

$$= \int_{\theta,Z} q(\theta, Z) \log q(X, \theta, Z) d(\theta, Z) - \int_{\theta,Z} q(\theta, Z) \log q(\theta, Z) d(\theta, Z) \tag{1.15}$$

$$= \int_{\theta,Z} q(\theta, Z) \log p(X \mid \theta, Z) d(\theta, Z) + \int_{\theta,Z} q(\theta, Z) \log p(\theta, Z) d(\theta, Z) \tag{1.16}$$

$$- \int_{\theta,Z} q(\theta, Z) \log q(\theta, Z) d(\theta, Z) \tag{1.17}$$

$$= \mathbb{E}_{\theta,Z}[\log p(X \mid \theta, Z)] + \mathbb{E}_{\theta,Z}[\log p(\theta, Z)] - \mathbb{E}_{\theta,Z}[\log q(\theta, Z)] \tag{1.18}$$

$$= -\mathbb{KL}(q(\theta, Z) \| p(\theta, Z)) + \mathbb{E}_{\theta,Z}[\log p(X \mid \theta, Z)] \tag{1.19}$$

Limite Inferior de la Evidencia (Evidence Lower Bound, ELBO)

El límite inferior de la evidencia ahora se puede interpretar como la suma de la probabilidad logarítmica esperada de los datos dados los parámetros del modelo y la divergencia (KL) entre la distribución aproximada $Q(\theta, Z)$ y la distribución previa $P(\theta, Z)$. Maximizar el límite inferior fomentará valores de Z que maximicen la probabilidad de registro de datos y minimicen la divergencia (KL) con respecto al anterior [5].

De la definición de la divergencia KL:

$$\begin{aligned} \mathrm{KL}[Q(\theta, \mathbf{Z}) \| P(\theta, \mathbf{Z} \mid \mathbf{X})] &= \int_{\theta,Z} q(\theta, \mathbf{Z}) \log \frac{q(\theta, \mathbf{Z})}{p(\theta, \mathbf{Z} \mid \mathbf{X})} d(\theta, Z) \\ &= \mathbb{E}_{\theta,Z} \left[\log \frac{q(\theta, \mathbf{Z})}{p(\theta, \mathbf{Z} \mid \mathbf{X})} \right] \\ &= \mathbb{E}_{\theta,Z}[\log q(\theta, \mathbf{Z})] - \mathbb{E}_{\theta,Z}[\log p(\theta, \mathbf{Z}, \mathbf{X})] + \mathbb{E}[\log p(\mathbf{X})] \end{aligned} \tag{1.20}$$

Se define al ELBO como:

$$\mathrm{ELBO}(q(\theta, Z)) = \mathbb{E}_{\theta,Z}[\log p(\theta, \mathbf{Z}, \mathbf{X})] - \mathbb{E}_{\theta,Z}[\log q(\theta, \mathbf{Z})] \tag{1.21}$$

Así la ecuación 1.20, puede ser reescrita,

$$log(p(X)) = \text{ELBO}(q(\theta, Z)) + KL[q(\theta, Z) \| p(\theta, Z|X)] \tag{1.22}$$

por la no negatividad de la divergencia de KL,

$$log(p(X)) \geq \text{ELBO}(q(\theta, Z)) \tag{1.23}$$

Por lo tanto, el $log(p(X))$, una cantidad que es fija para cualquier conjunto de observaciones X no puede ser menos que el ELBO.

Aproximación del Campo Medio

Nosotros estamos interesados en la distribución posterior dada una secuencia de observaciones denotada por $P(\theta, Z|X)$.

A continuación, se elige una familia de distribuciones de aproximación $Q(\theta, Z)$. Dado que (θ, Z) representa el conjunto de parámetros desconocidos y estados latentes, se supone que $Q(\theta, Z)$ puede factorizarse en grupos separados de parámetros o estados recopilados. Esta es una suposición que proviene de la física estadística, donde se conoce como teoría del campo medio. Por lo tanto, $Q(\theta, Z)$ se puede realizar como el producto de varios subconjuntos de (θ, z_i):

$$q(\theta, Z) = \prod_{i=1}^{I} q_i(\theta, z_i) \tag{1.24}$$

donde cada variable latente (θ, z_i) se rige en su propia densidad $q_i(\theta, z_i)$

Debemos tomar en cuenta que la familia variacional no es un modelo de los datos observados, los datos aparecen en la ecuación 1.21. Es decir, el ELBO y el problema de minimización KL correspondiente, conecta la densidad variacional ajustada de los datos y el modelo.

Se usará la notación simplificada $q_i = q_i(\theta, z_i)$. Para demostrar lo siguiente; 1.24 se sustituirá en 1.12, y luego se factorizará un q_j individual, como lo realiza [7].

$$\mathcal{L}(q) = \int_{\theta, Z_i} \prod_i q_i \log \left\{ \frac{p(X, \theta, Z)}{\prod_i q_i} \right\} d(\theta, Z_i) \tag{1.25}$$

Usaremos todos los cálculos descritos en [7] y [33].

$$\mathbb{E}_{i \neq j}[\log p(X, \theta, Z)] = \int_{\substack{\theta, Z_i \\ i \neq j}} \prod_i q_i \log p(\theta, Z, X) d(\theta, Z_i) \tag{1.26}$$

$$\begin{aligned} \mathcal{L}(q) &= \int_{\theta, Z_j} q_j \mathbb{E}_{i \neq j}[\log p(X, \theta, Z)] d(\theta, Z_j) - \int_{\theta, Z_j} q_j \log q_j d(\theta, Z_j) + \text{ const} \\ &= -\mathbb{KL}\left(q_j \| q_j^*\right) + \text{ const.} \end{aligned} \tag{1.27}$$

Así se obtiene la nueva distribución q_j^* esta definida para aproximar la esperanza de 1.26, representada como:

$$\log q_j^*(\theta, Z_j) = \mathbb{E}_{i \neq j}[\log p(\theta, Z, X)] + \text{ const.} \tag{1.28}$$

Inferencia Variacional de Campo Medio de ascenso de coordenadas, Coordinate Ascent mean-field Variational Inference, CAVI

Usando el ELBO y la familia de campo medio se convertirá la inferencia condicional aproximada en un problema de optimización, en esta sección se describirá uno de los algoritmos mas utilizados para resolver este problema de optimización, como lo hace [4].

Algorithm 1 Coordinate Ascent Variational Inferennce, CAVI

Require: Un modelo $p(\theta, X, Z)$, un conjunto de datos X

Ensure: Una densidad variacional $q(\theta, Z) = \prod_{i=1}^{N} q_i(\theta, z_i)$

1: **while** ELBO no ha convergido **do**

2: **for** i $\in$ {1,. . . ,N} **do**

3: Set $q_i(z_i)) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid z_{-j}, X)]\}$

4: Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\theta, \mathbf{Z}, \mathbf{X})] - \mathbb{E}[\log q(\theta, \mathbf{Z})]$

5: Return $q(\theta, Z)$

1.4.6. HMM Gassianos

Otra derivación de los modelos Ocultos de Markov con emisiones Gaussianas esta basada en [23], para la notación vamos ha tomar el estilo y convención de [4]

La distribución conjunta de un HMM con emisiones Gausianas descrita por la media μ y la precisión τ.

$$p(\pi, A, \mu, \tau, Z, X) = p(\pi)p(A)p(\mu \mid \tau)p(\tau)p(Z, X \mid \pi, A, \mu, \tau) \tag{1.29}$$

donde

$$p(Z, X \mid \pi, A, \mu, \tau) = p(z_1|\pi) \prod_{t=2}^{T} p(z_t|z_{t-1}, A) \prod_{t=1}^{T} p(x_t|z_t, \mu, \tau) \tag{1.30}$$

$$p\left(x_t \mid z_t, \mu, \tau\right) = \prod_{i=1}^{N} \mathcal{N}\left(x_t \mid \mu_i, \tau_i\right)^{\delta(z_t, i)} \tag{1.31}$$

$$\mathcal{N}\left(x_t \mid \mu_i, \tau_i\right) = \sqrt{\frac{\tau_i}{2\pi}} \exp\left(\frac{-\tau_i}{2}\left(x_t - \mu_i\right)^2\right) \tag{1.32}$$

$$p(\tau_i) = \text{Gamma}\left(\tau_i \mid \frac{a_i^0}{2}, \frac{b_i^0}{2}\right) \tag{1.33}$$

$$p\left(\mu_i \mid \tau_i\right) = \mathcal{N}\left(\mu_i \mid m_i^0, \left(\beta_i^0 \tau_i\right)^{-1}\right) \tag{1.34}$$

Capítulo 2

Metodología

2.1. Inferencia Variacional en Modelos Ocultos de Markov

En esta sección describiremos el modelo que se ha escogido, con sus respectivas notaciones, las cuales fueron dadas en el capítulo anterior. Comenzaremos indicando las probabilidades a priori elegidas, luego se calcula la verosimilitud.

2.1.1. Parámetros Iniciales

La mayoría de las fuentes recomiendan inicializar π y A con valores uniformes iguales a $\frac{1}{N}$ o valores extraídos aleatoriamente de una distribución uniforme, como lo hacen [29], [4] y [7]. Para este trabajo tomaremos las entradas de la diagonal de la matriz de transición A como $\frac{1}{2}$ y las entradas fuera de la diagonal como $\frac{1}{2(N-1)}$ [10]. Para las distribuciones de emisión continuas serán inicializadas con emisiones Gaussianas.

Dados los conceptos anteriores, vamos ha empezar con el enfoque de la inferencia Variacional en los Modelos Ocultos de Markov. Redefinimos a θ, como $\theta = (\pi, A, \mu, \tau)$.

La distribución conjunta del modelo y de los datos es:

$$p(\theta, Z, X) = p(\theta)p(X, Z|\theta) \tag{2.1}$$
$$= p(\pi, A, \mu, \tau)p(Z, X|\pi, A, \mu, \tau) \tag{2.2}$$
$$= p(\pi)p(A)p(\mu \mid \tau)p(\tau)p(Z, X|\pi, A, \mu, \tau) \tag{2.3}$$
$$= p(\pi)p(A)p(\mu|\tau)p(\tau)p(Z, X|\theta) \tag{2.4}$$

La verosimilitud se calcula de la siguiente manera.

$$p(Z, X|\theta) = p(z_1|\pi)\prod_{t=2}^{T} p(z_t|z_{t-1}, A)\prod_{t=1}^{T} p(x_t|z_t, \mu, \tau) \tag{2.5}$$

Las siguientes igualdades son las ecuaciones de actualización variacional

$$p(z_1|\pi) = \prod_{i=1}^{N} \pi_i^{\delta(z_1,i)} \tag{2.6}$$

$$p(z_t|z_{t-1}, A) = \prod_{i=1}^{N}\prod_{j=1}^{N} a_{ij}^{\delta(z_{t-1},i)\delta(z_t,j)} \tag{2.7}$$

para la actualización de $p(x_t|z_t, \mu, \tau)$, se toma la siguiente ecuación

$$p\left(x_t \mid z_t, \mu, \tau\right) = \prod_{i=1}^{N} \mathcal{N}\left(x_t \mid \mu_i, \tau_i\right)^{\delta(z_t,i)} \tag{2.8}$$

donde $\mathcal{N}$ representa la distribución normal con sus respectivos parámetros.

La notación $\delta(z_t, i)$ representa una función δ-Kronecker indicando el valor del estado escondido de la secuencia en el tiempo t

$$\delta(z_t, i) = \begin{cases} 1 & si \quad z_t = i \\ 0 & si \quad z_t \neq i \end{cases} \tag{2.9}$$

Las distribuciones a priori son introducidas sobre π, A μ y τ, para este trabajo vamos a elegir la distribución de Dirichlet, representado como Dir, la que es asumida para π, y a_j. Para la estructura a priori de μ_i

elegimos una distribución normal y para τ_i una distribución gamma considerando sus respectivos parámetros; como se lo hace en los trabajos de [7], [4] y [3].

A continuación se presenta la ecuaciones

$$p(\pi, A, \mu, \tau) = p(\pi)p(A)p(\mu|\tau)p(\tau) \tag{2.10}$$

$$p(\pi) = Dir(\{\pi_1, \ldots, \pi_N\}|\{\pi_1^0, \ldots, \pi_N^0\}) \tag{2.11}$$

$$p(A) = \prod_{i=1}^{N} Dir(\{a_{i1}, \ldots, a_{iN}\}|\{a_{i1}^0, \ldots, a_{iN}^0\}) \tag{2.12}$$

$$p(\mu|\tau) = \prod_{i=1}^{N} \mathcal{N}\left(\mu_i|m_i^0, (\beta_i^0\tau_i)^{-1}\right) \tag{2.13}$$

$$p(\tau) = \prod_{i=1}^{N} Gamma\left(\tau_i \mid \frac{a_i^0}{2}, \frac{b_i^0}{2}\right) \tag{2.14}$$

La distribución de Dirichlet tiene la siguiente forma para las probabilidades iniciales,

$$p(\pi|\pi^0) = C(\pi^0) \prod_{i=1}^{N} \pi_i^{\pi_i^0 - 1} \tag{2.15}$$

donde C es la constante de Normalización,

$$C\left(\pi^0\right) = \frac{\Gamma\left(\sum_{i=1}^{N} \pi_i^0\right)}{\prod_{i=1}^{N} \Gamma\left(\pi_i^0\right)} \tag{2.16}$$

En este trabajo tomaremos los siguientes hiperparámetros para la distribución de Dirichlet, como lo hace [28]:

$$\pi_i^0 = \frac{1}{N} \tag{2.17}$$

La probabilidad que estamos intentando calcular es $p(\theta, Z|X)$. El enfoque variacional busca una distribución de probabilidad aproximada $q(\theta, Z)$, seguimos con la notación de [4], considerando la suposición del

campo medio, la aproximación variacional puede factorizarse en grupos separados de variables y parámetros latentes:

$$q(\theta, Z) = q(\theta)q(Z) \tag{2.18}$$

2.1.2. Familias de aproximación para los parámetros

El siguiente paso es calcular las familias de aproximación para $q(\pi, A, \mu, \tau)$ y $q(Z)$.

Familia de aproximación para $q(\theta) = q(\pi, A, \mu, \tau)$

Para encontrar la familia de aproximación para $q(\pi, A, \mu, \tau)$, vamos a mostrar la distribución factorizada sobre cada uno de los parámetros, empezamos de la ecuación 1.28,

$$\begin{aligned}
\log q^*(\pi, A, \mu, \tau) &= \mathbb{E}_Z[\log p(\pi, A, \mu, \tau, Z, X)] + \text{ const} \\
&= \mathbb{E}_Z \left[\log \left(p(\pi)p(A)p(\mu|\tau)p(\tau)p(z_1 \mid \pi) \prod_{t=2}^{T} p(z_t \mid z_{t-1}, A) \prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau) \right) \right] \\
&+ \text{ const} \\
&= \log p(\pi) + \mathbb{E}_Z \left[\log p\left(s_1 \mid \pi\right) \right] \\
&+ \log p(A) + \mathbb{E}_Z \left[\log \prod_{t=1}^{T} p(z_t \mid z_{t-1}, A) \right] \\
&+ \log p(\mu|\tau) + \log p(\tau) + \mathbb{E}_Z \left[\log \prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau) \right] + \text{const.}
\end{aligned} \tag{2.19}$$

Como se había mencionado anteriormente, para la teoría del campo medio asumimos que los términos π, A, μ, τ son independientes uno de otro, así la factorización de la distribución tiene la siguiente forma.

$$\begin{aligned}
q(\theta, Z) &= q(\theta)q(Z) \\
&= q(\pi, A, \mu, \tau)q(Z) \\
&= q(\pi)q(A)q(\mu|\tau)q(\tau)q(Z)
\end{aligned} \tag{2.20}$$

Esta factoriazación que se acaba de escribir, es un resultado de las propiedades de la distribución condicional independiente, que se la conoce como una factorizacion inducida [4]. Ahora encontraremos las familias de aproximación para cada una de las distribuciones desconocidas sobre parámetros y estados ocultos.

Familia de aproximación para π

Empezamos de 1.28, y tomando la esperanza se sigue que

$$\begin{aligned}\log q^*(\pi) &= \mathbb{E}_{A,\mu,\tau,Z}[\log p(\pi, A, \mu, \tau, Z, X)] + \text{ const} \\ &= \mathbb{E}_{A,\mu,\tau,Z}\left[\log\left(p(\pi)p(A)p(\mu|\tau)p(\tau)p\left(z_1 \mid \pi\right)\prod_{t=2}^{T} p(z_t \mid z_{t-1}, A)\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau)\right)\right] \\ &+ \text{ const}\end{aligned}$$

Usando las propiedades de la función logarítmica, se sigue que,

$$\begin{aligned}\log q^*(\pi) = \mathbb{E}_{A,\mu,\tau,Z}[&\log p(\pi) + \log p(A) + \log p(\mu|\tau) + p(\tau) + \log p(z_1 \mid \pi) \\ &+ \log\left(\prod_{t=1}^{T} p(z_t \mid z_{t-1}, A)\right) + \log\left(\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau)\right)] + \text{const}\end{aligned}$$

aplicando la suma de las esperanzas y además, notemos que todos los términos que son constantes son absorbidos por la constante

$$\begin{aligned}\log q^*(\pi) &= \mathbb{E}_{A,\mu,\tau,Z}\left[\log p(\pi) + \log p\left(z_1 \mid \pi\right)\right] + \text{ const} \\ &= \log p(\pi) + \mathbb{E}_Z[\log p(z_1|\pi)] + \text{ const} \\ &= \log p(\pi) + \mathbb{E}_Z\left[\log \prod_{i=1}^{N} \pi_i^{\delta(z_1,i)}\right] + \text{ const}\end{aligned}$$

nuevamente usamos las definiciones de logaritmo

$$= \log p(\pi) + \mathbb{E}_Z \left[\sum_{i=1}^{N} \delta(z_1, i) \log \pi_i \right] + \text{ const}$$
$$= \log p(\pi) + \sum_{i=1}^{N} \mathbb{E}_Z[\delta(z_1, i) \log \pi_i] + \text{ const}$$

Finalmente aplicando la función exponencial a los dos lados,

$$\begin{aligned} q^*(\pi) &= p(\pi) \prod_{i=1}^{N} \pi_i^{\mathbb{E}_Z[\delta(z_1,i)]} \text{ const} \\ &= Dir(\pi_i | \pi_i + \mathbb{E}_Z[\delta(z_1, i)]) \end{aligned} \tag{2.21}$$

Como se puede observar, la familia de aproximación de las probabilidades iniciales esta dada por las distribuciones de Dirichlet, con su nueva actualización de los hiperparámetros.

Familia de aproximación para A

La derivación de la matriz de probabilidades de transición es similar a la realizada para π

$$\begin{aligned} \log q^*(A) &= \mathbb{E}_{A,\mu,\tau,Z}[\log p(\pi, A, \mu, \tau, Z, X)] + \text{ const} \\ &= \mathbb{E}_{\pi,\mu,\tau,Z} \left[\log \left(p(\pi)p(A)p(\mu|\tau)p(\tau)p\left(z_1 \mid \pi\right) \prod_{t=2}^{T} p(z_t \mid z_{t-1}, A) \prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau) \right) \right] \\ &+ \text{ const} \end{aligned}$$

$$\begin{aligned} \log q^*(A) = \mathbb{E}_{\pi,\mu,\tau,Z}[&\log p(\pi) + \log p(A) + \log p(\mu|\tau) + p(\tau) + \log p(z_1 \mid \pi) \\ &+ \log \left(\prod_{t=1}^{T} p(z_t \mid z_{t-1}, A) \right) + \log \left(\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau) \right)] + \text{const} \end{aligned}$$

aplicando la suma de las esperanzas y además, notemos que todos los términos que son constantes son absorbidos por la constante; se sigue

que,

$$\begin{aligned}\log q^*(A) &= \mathbb{E}_{\pi,\mu,\tau,Z}\left[\log p(A) + \log\left(\prod_{t=2}^{T} p(z_t \mid z_{t-1}, A)\right)\right] + \text{ const} \\ &= \log p(A) + \mathbb{E}_Z\left[\log\left(\prod_{t=1}^{T} p(z_t \mid z_{t-1}, A)\right)\right] + \text{ const}\end{aligned}$$

Usando las definiciones y propiedades de la función logarítmica.

$$\begin{aligned}\log q^*(A) &= \log p(A) + \mathbb{E}_Z\left[\sum_{t=2}^{T} \log p(z_t \mid z_{t-1}, A)\right] + \text{const} \\ &= \log p(A) + \mathbb{E}_Z\left[\sum_{t=2}^{T} \log \prod_{i=1}^{N}\prod_{j=1}^{N} a_{ij}^{\delta(z_{t-1},i)\delta(z_t,j)}\right] + \text{const} \\ &= \log p(A) + \mathbb{E}_Z\left[\sum_{t=2}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} \delta(z_{t-1},i)\delta(z_t,j)\log a_{ij}\right] + \text{const} \\ &= \log p(A) + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_Z\left[\sum_{t=2}^{T} \delta(z_{t-1},i)\delta(z_t,j)\right]\log a_{ij} + \text{const}\end{aligned}$$

$$\begin{aligned}&= \log p(A) + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T} \mathbb{E}_Z\left[\delta(z_{t-1},i)\delta(z_t,j)\right]\log a_{ij} + \text{const} \\ &= \log p(A) + \sum_{i=1}^{N}\sum_{j=1}^{N} \log a_{ij}^{\sum_{t=2}^{T}\mathbb{E}_Z[\delta(z_{t-1},i)\delta(z_t,j)]} + \text{const}\end{aligned}$$

Finalmente se aplica la función exponencial a ambos lados.

$$\begin{aligned}q^*(A) &= p(A)\prod_{i=1}^{N}\prod_{j=1}^{N} \log a_{ij}^{\sum_{t=2}^{T}\mathbb{E}_Z[\delta(z_{t-1},i)\delta(z_t,j)]}\text{const} \\ &= \prod_{i=1}^{N} Dir(a_{ij}|a_{ij} + \sum_{t=2}^{T}\mathbb{E}_Z\left[\delta(z_{t-1},i)\delta(z_t,j)\right])\end{aligned} \tag{2.22}$$

De forma similar, la familia de aproximación para la matriz de tran-

sición esta dada por las distribuciones de Dirichlet, con su actualización de los hiperparámetros.

Familia de aproximación para μ

Partiendo de la ecuación 1.28

$$\begin{aligned}\log q(\mu|\tau) &= \mathbb{E}_{\pi,A\tau,Z}[\log p(\pi, A, \mu, \tau, Z, X)] + \text{ const} \\ &= \mathbb{E}_{\pi\tau,Z}\left[\log\left(p(\pi)p(A)p(\mu|\tau)p(\tau)p\left(z_1 \mid \pi\right)\prod_{t=2}^{T} p(z_t \mid z_{t-1}, A)\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau)\right)\right] \\ &+ \text{ const}\end{aligned}$$

Aplicando las propiedades de la función logarítmica, la suma de esperanzas y tomando en cuenta que los términos constates son absorbidos por la constante. tenemos que

$$\log q(\mu|\tau) = \mathbb{E}_{\tau,Z}\left[\log p(\mu|\tau) + \log\left(\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau)\right)\right] + \text{const}$$

Realizando los cálculos pertinentes, basándonos en [4] y [7], usando la distribuciones Gaussianas independientes, y aplicando la función exponencial.

$$q(\mu|\tau) = \prod_{i=1}^{N} \mathcal{N}(\mu_i|m_i, (\mathbb{E}_\tau[\tau_i]\beta_i)^{-1}) \tag{2.23}$$

donde

$$m_i = \frac{\beta_i^0 m_i^0 + \sum_{t=1}^{T} \mathbb{E}_S\left[\delta\left(s_t, i\right)\right] x_t}{\beta_i^0 + \sum_{t=1}^{T} \mathbb{E}_S\left[\delta\left(s_t, i\right)\right]} \tag{2.24}$$

$$\beta_i = \beta_i^0 + \sum_{t=1}^{T} \mathbb{E}_S\left[\delta\left(s_t, i\right)\right] \tag{2.25}$$

Para más detalles ver [7], el cuál desarrolla en su tesis doctoral todos los cálculos para obtener la familia de aproximación para μ

La familia de aproximación de $q(\mu|\tau)$ está dada por una distribución normal con una media igual a m_i y una varianza $\frac{1}{\tau_i \beta_i}$ con sus respectivas actualizaciones.

Familia de aproximación para τ

Dada la ecuación 1.28

$$\begin{aligned}\log q(\tau) &= \mathbb{E}_{\pi,A\mu,Z}[\log p(\pi, A, \mu, \tau, Z, X)] + \text{ const} \\ &= \mathbb{E}_{\pi,A,\mu,Z}\left[\log\left(p(\pi)p(A)p(\mu|\tau)p(\tau)p\left(z_1 \mid \pi\right)\prod_{t=2}^{T} p(z_t \mid z_{t-1}, A)\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau)\right)\right] \\ &+ \text{ const}\end{aligned}$$

Con las propiedades de la función logarítmica, la suma de esperanzas y teniendo en cuenta que los términos constates son absorbidos por la constante. se sigue.

$$\log q(\tau) = \mathbb{E}_{\tau,Z}\left[\log p(\mu|\tau) + p(\tau) + \log\left(\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau)\right)\right] + \text{const}$$

Nuevamente aplicando la definición y propiedades de la función logarítmica.

$$\begin{aligned}\log q(\tau) = \mathbb{E}_{\mu}\left[\log\prod_{i=1}^{N}\mathcal{N}(\mu_i|m_i^0(\beta_i^0\tau_i)^{-1})\right] + \log\prod_{i=1}^{N}\text{Gamma}(\tau_i|\frac{a_i^0}{2}, \frac{b_i^0}{2}) + \\ + \mathbb{E}_{\mu,Z}\left[\log\prod_{t=1}^{T}\prod_{i=1}^{N}\mathcal{N}(x_t|\mu_i, \tau_i)^{\delta(z_t,i)})\right] + \text{const}\end{aligned}$$

Basándonos en lo cálculos realizados en [7], se sigue que,

$$\log q(\tau) = \frac{\log \tau_i}{2} - \frac{\beta_i^0 \tau_i}{2}\mathbb{E}_\mu[(\mu_i - m_i^0)^2] + (\frac{a_i^0}{2} - 1)\log \tau_i - \frac{b_i^0 \tau_i}{2} + \sum_{t=1}^{T}\left(\mathbb{E}_Z[\delta(z_t, i)\frac{\log \tau}{2}] - \frac{\tau_i}{2}\mathbb{E}_\mu[(x_t - \mu_i)^2]\right) + \text{const}$$

Finalmente se obtiene que la familia de aproximaciones para $q(\tau)$.

$$q(\tau_i) = \text{Gamma}(\tau_i|\frac{a_i}{2}, \frac{b_i}{2}) \tag{2.26}$$

donde,

$$a_i = a_i^0 + \sum_{t=1}^{T}\mathbb{E}_Z[\delta(z_t, i)]$$

$$b_i = b_i^0 + \beta_i^0 m_i^{0^2} + \sum_{t=1}^{T}\mathbb{E}_Z[\delta(z_t, i)]x_t^2 - \beta_i m_i^2$$

La cuál esta dada por la distribución Gamma con el parámetro de forma a_i y el parámetro de escala b_i

2.1.3. Familias de aproximación para los estados ocultos

Familia de aproximación para Z

Para encontrar la familia de aproximación de los estados escondidos Z con emisiones Gaussianas, se realiza los siguientes cálculos, basándonos en [4] y [7].

Dada la ecuación 1.28

$$
\begin{aligned}
\log q^*(Z) &= \mathbb{E}_{\pi,A,\mu,\tau}[\log p(\pi, A, \mu, \tau, Z, X)] - \text{ const} \\
&= \mathbb{E}_{\pi,\mu,\tau}\left[\log\left(p(\pi)p(A)p(\mu|\tau)p(\tau)p\left(z_1 \mid \pi\right)\prod_{t=2}^{T} p(z_t \mid z_{t-1}, A)\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau)\right)\right] \\
&- \text{ const} \\
&= \mathbb{E}_{\pi,\mu,\tau}[\log p(\pi) + \log p(A) + \log p(\mu|\tau) + p(\tau) + \log p(z_1 \mid \pi) \\
&+ \log\left(\prod_{t=1}^{T} p(z_t \mid z_{t-1}, A)\right) + \log\left(\prod_{t=1}^{T} p(x_t \mid z_t, \mu, \tau)\right)] - \text{const}
\end{aligned}
$$

Vamos asumir la constante como $\log \mathcal{Z}$

$$
\begin{aligned}
\log q^*(Z) = \mathbb{E}_{\pi}[\log p(z_1 \mid \pi)] + \mathbb{E}_A\left(\log\prod_{t=1}^{T} p(z_t \mid z_{t-1}, A)\right) \\
+ \mathbb{E}_{\mu,\tau}\left(\log\prod_{t=1}^{T} p(x_t \mid \mu, \tau)\right) - \log \mathcal{Z}
\end{aligned}
$$

Reemplazando los valores de cada uno de los términos y aplicando la función exponencial a ambos lados de la ecuación se tiene que,

$$
\begin{aligned}
q(Z) = &\prod_{i=1}^{N} \exp\left(\mathbb{E}_{\pi}[\log(\pi_i)]\right)^{\delta(z_1,i)} \\
&\prod_{t=2}^{T}\prod_{i=1}^{N}\prod_{j=1}^{N} \exp\left(\mathbb{E}_A[\log(a_{i,j})]\right)^{\delta(z_{t-1},i)*\delta(z_t,j)} \\
&\prod_{t=1}^{T}\prod_{i=1}^{N} \exp\left(\mathbb{E}_{\mu,\tau}[\log(\mathcal{N}(x_t|\mu_i, \tau_i))]\right)^{\delta(z_t,i)} \frac{1}{\mathcal{Z}}
\end{aligned} \tag{2.27}
$$

Introduciremos una nueva notación:

$$
\tilde{\pi_i} = \exp\left(\mathbb{E}_{\pi}[\log(\pi_i)]\right) \tag{2.28}
$$

$$
\tilde{a}_{ij} = \exp\left(\mathbb{E}_A[\log(a_{ij})]\right) \tag{2.29}
$$

La siguiente propiedad de la distribución de Dirichlet se usará para los valores de $\tilde{\pi}$ y $\tilde{A}$

$$
\mathbb{E}_{x_t}[\log x_t] = \psi(x_t) - \psi(\sum x_t) \tag{2.30}
$$

Donde el símbolo ψ es la función digamma [4]

$$\tilde{\pi} = \tilde{\pi}_i = \exp\left(\psi(\pi_i) - \psi(\sum_{i=1}^{N} \pi_i)\right) \tag{2.31}$$

$$\tilde{A} = \tilde{a}_{ij} = \exp\left(\psi(a_{ij}) - \psi(\sum_{j=1}^{N} a_{ij})\right) \tag{2.32}$$

$$\tag{2.33}$$

Por otro lado notemos que

$$\begin{aligned}
\mathbb{E}_{\mu,\tau}[\log(\mathcal{N}(x_t|\mu_i,\tau_i))] &= \mathbb{E}_{\mu,\tau}[\log\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{\tau_i}{2}(x_t-\mu_i)^2\right)] \\
&= \mathbb{E}_{\mu,\tau}[\frac{\log\tau_i}{2} - \log(2\pi) - \frac{\tau_i}{2}(x_t-\mu_i)^2] \\
&= \mathbb{E}_{\tau}[\frac{\log\tau_i}{2}] - \log(2\pi) - \mathbb{E}_{\tau}[\frac{\tau_i}{2}]\mathbb{E}_{\mu}[(x_t-\mu_i)^2] \\
&= \mathbb{E}_{\tau}[\frac{\log\tau_i}{2}] - \log(2\pi) - \mathbb{E}_{\tau}[\frac{\tau_i}{2}]\mathbb{E}_{\mu}[x_t^2] - \mathbb{E}_{\mu}[2x_t\mu_i] + \mathbb{E}_{\mu}[\mu_i^2]
\end{aligned}$$

por las propiedades de la distribución Gassiana se sabe que $\mathbb{E}_{\mu}[\mu_i] = m_i$ y $\mathbb{E}_{\mu}[\mu_i^2] = m_i^2 + \sigma_i^2$, entonces

$$\begin{aligned}
\mathbb{E}_{\mu,\tau}[\log(\mathcal{N}(x_t|\mu_i,\tau_i))] &= \frac{\mathbb{E}_{\tau}[\log\tau_i]}{2} - \log(2\pi) - \frac{\mathbb{E}_{\tau}\tau_i}{2}[x_t^2 - 2x_t m_i + m_i^2 + \frac{1}{\beta_i\tau_i}] \\
&= \frac{\mathbb{E}_{\tau}[\log\tau_i]}{2} - \log(2\pi) - \frac{\mathbb{E}_{\tau}\tau_i}{2}[x_t^2 - 2x_t m_i + m_i^2 + \frac{1}{\beta_i\mathbb{E}_{\tau}\tau_i}] \\
&= \frac{\mathbb{E}_{\tau}[\log\tau_i]}{2} - \log(2\pi) - \frac{\mathbb{E}_{\tau}\tau_i}{2}[x_t - m_i]^2 - \frac{1}{2\beta_i}
\end{aligned}$$

como $\mathbb{E}_{\tau}[\log\tau_i] = \psi(\frac{a_i}{2}) - \log(\frac{b_i}{2})$ y $\mathbb{E}_{\tau}[\tau_i] = \frac{a_i}{b_i}$ se sigue que,

$$\mathbb{E}_{\mu,\tau}[\log(\mathcal{N}(x_t|\mu_i,\tau_i))] = \frac{1}{2}\left(\psi(\frac{a_i}{2}) - \log(\frac{b_i}{2}) - \frac{a_i}{b_i}(x_t-m_i)^2 - \frac{1}{\beta_i}\right) \tag{2.34}$$

por simplicidad vamos a notar

$$\tilde{b}_{ti} = \exp\left[\frac{1}{2}\left(\psi(\frac{a_i}{2}) - \log(\frac{b_i}{2}) - \frac{a_i}{b_i}(x_t-m_i)^2 - \frac{1}{\beta_i}\right)\right] \tag{2.35}$$

por tanto,

$$q(Z) = \prod_{i=1}^{N} \tilde{\pi}_i^{\delta(z_1,i)} \prod_{t=2}^{T} \prod_{i=1}^{N} \prod_{j=1}^{N} \tilde{a}_{ij}^{\delta(z_{t-1},i)\delta(z_t,j)} \prod_{t=1}^{T} \prod_{i=1}^{N} \tilde{b}_{ti}^{\delta(z_t,i)} \frac{1}{\mathcal{Z}} \tag{2.36}$$

2.1.4. Calculo del ELBO

Dados las ecuaciones anteriores, vamos a calcular el limite inferior de Evidencia (ELBO).

$$\text{ELBO}(q(\theta, Z)) = \mathbb{E}_{\theta,Z}[\log p(\theta, \mathbf{Z}, \mathbf{X})] - \mathbb{E}_{\theta,Z}[\log q(\theta, \mathbf{Z})] \tag{2.37}$$

Aplicando propiedades de la función logarítmica y la suma de las esperanzas

$$\begin{aligned}
\text{ELBO}(q(\theta, Z)) &= \mathbb{E}_{\theta,Z}[\log p(X, Z|\theta)p(\theta)] - \mathbb{E}_{\theta,Z}[\log q(\theta)q(Z)] && (2.38)\\
&= \mathbb{E}_{\theta,Z}[\log p(X, Z|\theta)] + \log p(\theta)] - \mathbb{E}_{\theta,Z}[\log q(\theta) + \log q(Z)] && (2.39)\\
&= \mathbb{E}_Z[\log p(X, Z|\theta)] + \mathbb{E}_\theta[\log p(\theta)] - \mathbb{E}_\theta[\log q(\theta)] - \mathbb{E}_Z[\log q(Z)] && (2.40)
\end{aligned}$$

Por otro lado, notemos que:

$$\begin{aligned}
\mathbb{E}_Z[\log q(Z)] &= \sum_Z q(Z) \log(q(Z)) && (2.41)\\
&= \sum_Z q(Z) \mathbb{E}_\theta \log p(X, Z|\theta) - \log(\mathcal{Z}) && (2.42)\\
&= \sum_Z q(Z) \mathbb{E}_\theta \log p(X, Z|\theta) - \log(\mathcal{Z}) && (2.43)\\
&= \mathbb{E}_\theta \log p(X, Z|\theta) - \log(\mathcal{Z}) && (2.44)
\end{aligned}$$

Si reemplazamos en la ecuación que calcula el ELBO,

$$\begin{aligned}
\text{ELBO}(q(\theta, Z)) &= \mathbb{E}_\theta[\log p(\theta)] - \mathbb{E}_\theta[\log q(\theta)] + \log q(\mathcal{Z}) && (2.45)\\
&= \mathbb{E}_\pi[\log p(\pi)] - \mathbb{E}_\pi[\log q(\pi)] && (2.46)\\
&+ \mathbb{E}_A[\log p(A)] - \mathbb{E}_A[\log q(A)] && (2.47)\\
&+ \mathbb{E}_\mu[\log p(\mu|\tau)] - \mathbb{E}_\mu[\log q(\mu)] && (2.48)\\
&+ \mathbb{E}_\tau[\log p(\tau)] - \mathbb{E}_\pi[\log q(\tau)] + \log \mathcal{Z} && (2.49)
\end{aligned}$$

Usando la definición de la divergencia de Kullback-Leibler

$$\begin{aligned}
\text{ELBO}(q(\theta, Z)) = &-\text{KL}(q(\pi)||p(\pi)) - \text{KL}(q(A)||p(A))\\
&- \text{KL}(q(\mu)||p(\mu|\tau)) - \text{KL}(q(\tau)||p(\tau)) + \log \mathcal{Z}
\end{aligned} \tag{2.50}$$

$\text{KL}(q(\pi)||p(\pi))$ y $\text{KL}(q(A)||p(A))$ son la divergencia KL entre el estado dependiente posterior y a priori de las distribuciones Dirichlet, $\text{KL}(q(\mu)||p(\mu|\tau))$ es la divergencia KL entre el estado dependiente posterior y a priori de las distribuciones Gaussianas; y $\text{KL}(q(\tau)||p(\tau))$ es la divergencia KL entre el estado dependiente posterior y a prior de las distribuciones Gamma. Penny [27] puede ser consultado para más información.

Desarrollo de las divergencias de KL para el calculo del límite inferior de Evidencia

$$\begin{aligned}
\text{KL}(q(\pi)||p(\pi)) &= \int q(\pi) \log\left(\frac{q(\pi)}{p(\pi)}\right) d\pi\\
&= \int Dir(\pi_i|\pi_i^0) \log\left(\frac{Dir(\pi_i|\pi_i^{act}))}{Dir(\pi_i|\pi_i^0))}\right) d\pi\\
&= \mathbb{E}_\pi\left[\log\left(\frac{Dir(\pi_i|\pi_i^{act}))}{Dir(\pi|\pi_i^0)}\right)\right]\\
&= \mathbb{E}_\pi\left[\log\left(\frac{\frac{\Gamma(\sum \pi_i^{act})\prod_{i=1}^N \pi_i^{\pi_i^{act}-1}}{\prod_{i=1}^N \Gamma(\pi_i^{act})}}{\frac{\Gamma(\sum \pi_i^0)\prod_{i=1}^N \pi_i^{\pi_i^0-1}}{\prod_{i=1}^N \Gamma(\pi_i^0)}}\right)\right]\\
&= \mathbb{E}_\pi\left[\log\left(\frac{\Gamma(\sum \pi_i^{act})}{\Gamma(\sum \pi_i^0)}\frac{\prod_{i=1}^N \Gamma(\pi_i^0)}{\prod_{i=1}^N \Gamma(\pi_i^{act})}\prod_{i=1}^N \pi_i^{\pi_i^{act}-\pi_i^0}\right)\right]
\end{aligned}$$

$$\begin{aligned}\mathrm{KL}(q(\pi)||p(\pi)) &= \mathbb{E}_\pi\left[\log\left(\frac{\Gamma(\sum \pi_i^{act})}{\Gamma(\sum \pi_i^0)}\right) + \sum\left(\frac{\Gamma(\pi_i^0)}{\Gamma(\pi_i^{act})}\right) + \sum_{i=1}^{N}(\pi_i^{act} - \pi_i^0)\log \pi_i\right] \\ &= \log\left(\frac{\Gamma(\sum \pi_i^{act})}{\Gamma(\sum \pi_i^0)}\right) + \sum\left(\frac{\Gamma(\pi_i^0)}{\Gamma(\pi_i^{act})}\right) + \sum_{i=1}^{N}(\pi_i^{act} - \pi_i^0)\mathbb{E}_\pi\left[\log \pi_i\right]\end{aligned}$$

Recordemos que

$$\mathbb{E}[\log(x)] = \psi(x) - \psi(\sum x)$$

Así

$$\begin{aligned}\mathrm{KL}(q(\pi)||p(\pi)) = \log\left(\frac{\Gamma(\sum \pi_i^{act})}{\Gamma(\sum \pi_i^0)}\right) + \sum\left(\frac{\Gamma(\pi_i^0)}{\Gamma(\pi_i^{act})}\right) + \\ \sum_{i=1}^{N}(\pi_i^{act} - \pi_i^0)\left(\psi(\pi_i^{act}) - \psi\left(\sum \pi_i^{act}\right)\right)\end{aligned}$$

Para calcular $\mathrm{KL}(q(A)||p(A))$ se realiza de forma similar a $\mathrm{KL}(q(\pi)||p(\pi))$.

Ahora calcularemos $\mathrm{KL}(q(\tau)||p(\tau))$

$$\begin{aligned}\mathrm{KL}(q(\tau)||p(\tau)) &= \int q(\tau)\log\left(\frac{q(\tau)}{p(\tau)}\right)d\pi \\ &= \int Gamm\left(\tau_i|\frac{a_j}{2},\frac{b_j}{2}\right)\log\frac{Gamm\left(\tau_i|\frac{a_j}{2},\frac{b_j}{2}\right)}{Gamm\left(\tau_j|\frac{a_j^0}{2},\frac{b_j^0}{2}\right)}d\tau \\ &= \mathbb{E}\left[\log\frac{Gamm\left(\tau_j|\frac{a_j}{2},\frac{b_j}{2}\right)}{Gamm\left(\tau_i|\frac{a_j^0}{2},\frac{b_j^0}{2}\right)}\right] \\ &= \mathbb{E}\left[\log\frac{\frac{(b_j/2)^{(a_j/2)}}{\Gamma(a_j/2)}\tau_j^{(a_j/2)-1}\exp -(b_j/2)\tau_j}{\frac{(b_j^0/2)^{(a_j^0/2)}}{\Gamma(a_j^0/2)}\tau_j^{(a_j^0/2)-1}\exp -(b_j^0/2)\tau_j}\right] \\ &= \mathbb{E}\left[\log(\frac{(b_j/2)^{(a_j/2)}}{\Gamma(a_j/2)}\tau_j^{(a_j/2)-1}\exp -(b_j/2)\tau_j)\right] \\ &- \mathbb{E}\left[\log(\frac{(b_j^0/2)^{(a_j^0/2)}}{\Gamma(a_j^0/2)}\tau_j^{(a_j^0/2)-1}\exp -(b_j^0/2)\tau_j)\right]\end{aligned}$$

$$\begin{aligned}\text{KL}(q(\tau)||p(\tau)) &= \mathbb{E}\left[(a_j/2)\log\frac{b_j}{2} - \log\Gamma(a_j/2) + [(a_j/2)-1]\log\tau_j - (b_j/2)\tau_j\right] \\ &- \mathbb{E}\left[(a_j^0/2)\log\frac{b_j^0}{2} - \log\Gamma(a_j^0/2) + [(a_j^0/2)-1]\mathbb{E}[\log\tau_j] - (b_j^0/2)\tau_j\right] \\ &= (a_j/2)\log\frac{b_j}{2} - \log\Gamma(a_j/2) + [(a_j/2)-1]\mathbb{E}[\log\tau_j] - (b_j/2)\mathbb{E}[\tau_j] \\ &- (a_j^0/2)\log\frac{b_j^0}{2} + \log\Gamma(a_j^0/2) - [(a_j^0/2)-1]\mathbb{E}[\log\tau_j] + (b_j^0/2)\mathbb{E}[\tau_j]\end{aligned}$$

Se sabe que $\mathbb{E}[\tau] = a_j/b_j$ y $\mathbb{E}[\log\tau] = \psi(a_j/2) - \log(b_j/2)$

Se sigue que

$$\begin{aligned}\text{KL}(q(\tau)||p(\tau)) &= \frac{a_j}{2}\log\frac{b_j}{2} - \log\Gamma(\frac{a_j}{2}) + [(\frac{a_j}{2})-1](\psi(\frac{a_j}{2}) - \log\frac{b_j}{2}) - \frac{b_j}{2}\frac{a_j}{b_j} \\ &- \frac{a_j^0}{2}\log\frac{b_j^0}{2} + \log\Gamma(\frac{a_j^0}{2}) - [(\frac{a_j}{2})-1](\psi(\frac{a_j}{2}) - \log\frac{b_j}{2}) + \frac{b_j^0}{2}\frac{a_j}{b_j} \\ &= \frac{a_j^0}{2}\log\frac{b_j}{b_j^0} - \log\Gamma(\frac{a_j}{2}) + \log\Gamma(\frac{a_j^0}{2}) + [\frac{a_j}{2} - \frac{a_j^0}{2}]\psi(\frac{a_j}{2}) \\ &- (\frac{b_j}{2} - \frac{b_j^0}{2})\frac{a_j}{b_j}\end{aligned}$$

Finalmente $\text{KL}(q(\mu)||p(\mu|\tau))$

$$\text{KL}(q(\mu)||p(\mu|\tau)) = \frac{1}{2}\left(\frac{(m_j - m_j^0)^2}{\sigma_j^{02}} - \frac{\sigma_j^2}{\sigma_j^{02}} - \log\frac{\sigma_j^2}{\sigma_j^{02}} - 1\right)$$

Notemos que $\sigma_j^{02} = 1/(\tau\beta_j^0)$ y $\sigma_j^2 = 1/(\tau\beta_j)$

Por tanto

$$\text{KL}(q(\mu)||p(\mu|\tau)) = \frac{1}{2}\left(\tau\beta_j^0(m_j - m_j^0)^2 - \frac{\beta_j^0}{\beta_j} - \log\frac{\beta_j^0}{\beta_j} - 1\right)$$

2.1.5. Algoritmo Variacional Bayes EM

Para la implementación final se ha tomado los conceptos del algoritmo Expectection-Maximization, es decir, los algoritmos de Fordward-Backward y además las propiedades de Inferencia Variacional.

Para ello, se divide en dos pasos:

Paso E Variacional

- Usar el algoritmo forward-backward, descrito en B; para calcular $\tilde{\alpha}_t(i)$ y $\tilde{\beta}_t(i)$ usando $\tilde{\pi}_i$, $\tilde{a}_{ij}$ y $\tilde{b}_{ti}$.
- Calcular los valores para $\tilde{\gamma}_t(i)$ y $\tilde{\gamma}_t(i,j)$; los cuales prevén la estimación de:

$$\mathbb{E}_Z[\delta(z_1,i)] = \tilde{\gamma}_1(i) \tag{2.51}$$

$$\mathbb{E}_Z[\delta(z_t,i)] = \tilde{\gamma}_t(i) \tag{2.52}$$

$$\mathbb{E}_Z[\delta(z_{t-1},i)\delta(z_t,j)] = \tilde{\gamma}_{t-1}(i,j) \tag{2.53}$$

  Donde:

$$\tilde{\gamma}_t(i) = \frac{\tilde{\alpha}_t(i)\tilde{\beta}_t(i)}{c_t} \tag{2.54}$$

$$\tilde{\gamma}_{t-1}(i,j) = \frac{\tilde{\alpha}_{t-1}(i)a_{ij}b_{tj}\tilde{\beta}_t(j)}{c_t} \tag{2.55}$$

  $\gamma_t(i,j)$ esta definida como la probabilidad de estado i en el tiempo t y simultáneamente trasicionando al estado j en el tiempo $t+1$

  $\gamma_t(j)$ esta definida para almacenar la probabilidad del estado j en el tiempo t

- Calcular $\mathcal{Z}$

$$\mathcal{Z} = \sum_{t=1}^{T} \log c_t \tag{2.56}$$

Paso M Variacional

- Calcular el limite de inferior de evidencia, ELBO, presentado en 2.50

- Actualizar los hiper parámetros

$$\pi_j = \pi_j^0 + \tilde{\gamma}_1(i) \tag{2.57}$$

$$a_{ij} = a_{ij}^0 + \sum_t \tilde{\gamma}_{t-1}(i,j) \tag{2.58}$$

$$m_j = \frac{\beta_j^0 m_j^0 + \sum_t \tilde{\gamma}_t(i) x_t}{\beta_j} \tag{2.59}$$

$$\beta_j = \beta_j^0 + \sum_t \tilde{\gamma}_1(i) \tag{2.60}$$

$$a_j = a_j^0 + \sum_t \tilde{\gamma}_t(i) \tag{2.61}$$

$$b_j = b_j^0 + \beta_i^0 m_j^{0^2} + \sum_{t=1}^{T} \tilde{\gamma}_t(j) x_t^2 - \beta_j m_j^2 \tag{2.62}$$

- Reestimar las distribuciones $\tilde{\pi}$, $\tilde{A}$ y $\tilde{B}$

2.1.6. Selección del Modelo

Para la selección del modelo se evalúa al mismo con los siguientes criterios:

Criterio de Información de Akaike (AIC) es una medida de la calidad relativa de los modelos, y viene dado por la siguiente formula.

$$\text{AIC} = -2 \log L + 2p \tag{2.63}$$

donde $\log L$ es el logaritmo de la verosimilitud del modelo construido y p denota el número de parámetros del modelo.

Criterio de información bayesiano (BIC) es un índice utilizado en las estadísticas bayesianas para elegir entre dos o más modelos alternativos. Dado por la siguiente formula.

$$\text{BIC} = -2 \log L + p \log T \tag{2.64}$$

donde $\log L$ y p están definidos de forma similar que en el AIC y T es el número de observaciones.

Además, McGrory y Titterington [23] han demostrado que el criterio de desviación de información (DIC, por sus siglas en Inglés) para HMM

con emisiones gaussinas asiste con la selección del modelo.

El DIC esta definido como:

$$\text{DIC} = \bar{D(\theta)} + p_D \tag{2.65}$$

donde $\bar{D(\theta)}$ mide el ajuste del modelo y p_D es la medida de complejidad del modelo.

$$D(\theta) = -2 \log p(X|\theta) \tag{2.66}$$

y $\bar{D(\theta)}$ corresponde a la esperanza con respecto a $p(X|\theta)$. La medida de complejidad esta definida como:

$$p_D = \bar{D(\theta)} - D(\bar{\theta}) \tag{2.67}$$

$$= \mathbf{E}[-2 \log p(X|\theta)] + 2 \log p(X|\bar{\theta}) \tag{2.68}$$

Para más detalles ver el anexo C.

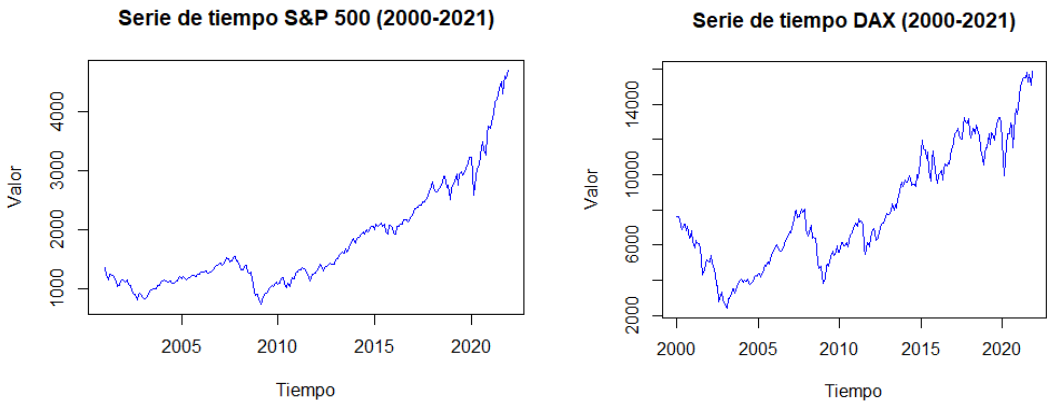
Capítulo 3

Resultados, conclusiones y recomendaciones

En esta sección se va ha describir los resultados que se han obtenido en el proceso de la realización de trabajo de integración curricular, las conclusiones y recomendaciones.

3.1. Resultados

Presentamos las series en estudio, 3.1 y 3.2



(a) S&P 500 (b) DAX

Figura 3.1: Series de Tiempo S&P 500 y DAX

La serie S&P 500 3.1b cuenta con 252 observaciones, la serie DAX 3.1a 263 observaciones y la serie KO 3.2 264 observaciones

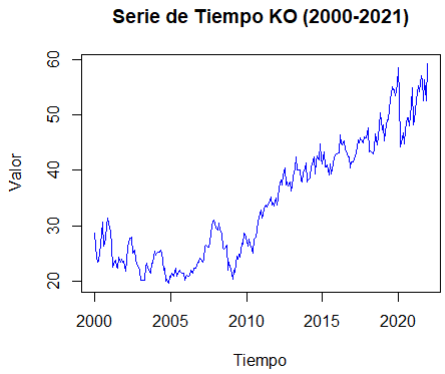
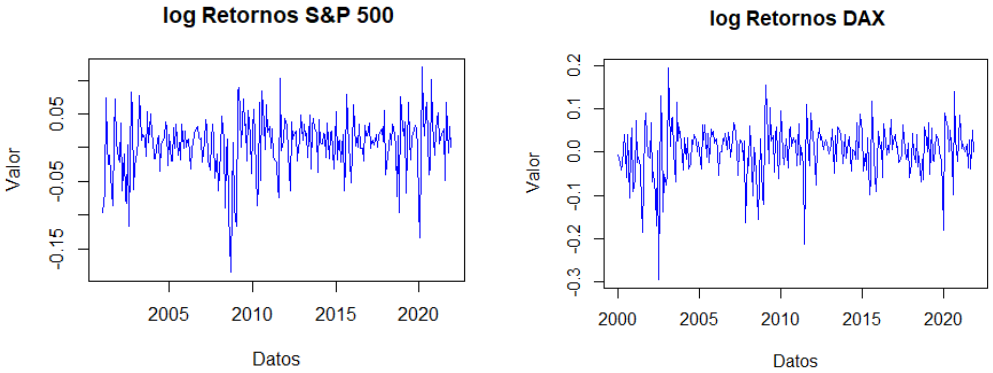


Figura 3.2: Precio histórico de las acciones de Coca Cola

Luego se muestra los log retornos de las series S&P 500, DAX en 3.3 y KO en 3.4.



(a) log return (b) log return

Figura 3.3: log returnos SP 500 y DAX

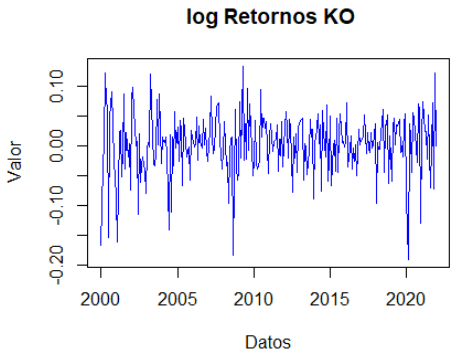


Figura 3.4: log return Acciones de Coca Cola

En la figura 3.4 de log retornos presenta más variabilidad, a compa-

ración de las figuras 3.3a y 3.3b

3.1.1. Performance del Algoritmo

El algoritmo implementado usa una series de códigos que han sido utilizados en los trabajos de [7] y [22], se conserva parte de su estilo.

Para inicializar los hiper parámetros (π_j, a_{ij}, m_j, β_j, a_j,b_j) con $i.j \in \{1, 2, \ldots, N\}$, se consolidó constantes a estos valores, los cuales posteriormente son utilizados para determinar $\tilde{\pi}$, $\tilde{A}$ y $\tilde{B}$, a diferencia de los trabajos realizados por [7], [22] y [23] que toman valores aleatorios.

Para las emisiones Gaussianas se usa los parámetros m_j y β_j con $j \in \{1, 2, \ldots, N\}$, donde N representa el número de estados ocultos. Se fija los valores iniciales de m_j posterior de la siguiente manera:

- Se utiliza la función $kmeans()$ del software estadístico R (3 estados ocultos)
- En el caso de dos estados ocultos $m_j = (0{,}05, -0{,}05)$, se tomó estos valores ya que los log-retornos en estudio varían aproximadamente entre $-0{,}18$ y $0{,}11$

Así el algoritmo converge a su óptimo local.

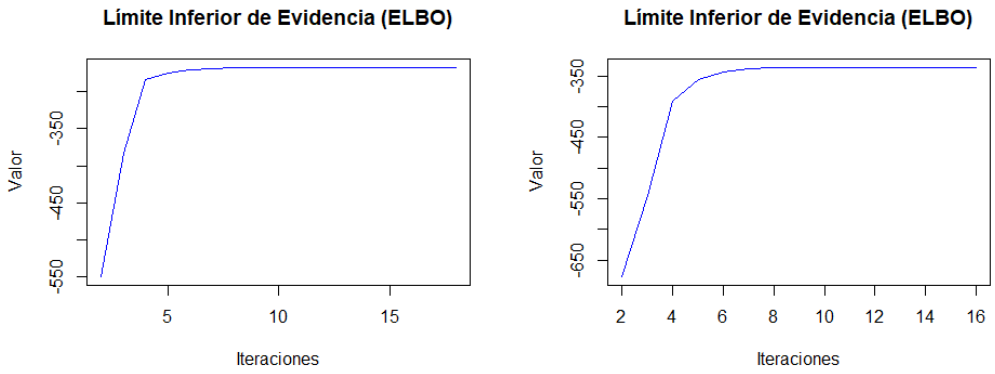
Se intento inicializar los valores de m_j como un vector de ceros, pero el algoritmo presento problemas de convergencia, de igual manera para los hiper parámetros de la familia de distribución Gamma, si se toma valores iguales para el parámetro de forma a_j los resultados no son convincentes.

El tiempo de convergencia es relativamente rápido, puede ser porque no se tiene un conjunto de datos suficientemente grande, es por eso que el algoritmo implementado funciona en un computador personal.

3.1.2. S&P 500

La convergencia del limite inferior de Evidencia (ELBO), presenta los siguientes resultados.

Como se puede apreciar en la figura 3.5 la convergencia con dos estados ocultos en más rápida, por lo que se tiene menos iteraciones hasta



(a) ELBO dos estados (b) ELBO tres estados

Figura 3.5: Convergencia ELBO para SP500

	2 estados	**3 estados**
logLik	-222.85	-334.99
AIC	1455.71	2191.99
BIC	3238.08	4877.88
DIC	451.13	672.19

Cuadro 3.1: Criterios AIC, BIC y DIC para la serie S&P 500

llegar al óptimo, y los resultados de la tabla 3.1 nos muestran que el criterio Bayesiano prefiere el modelo con dos estados ocultos.

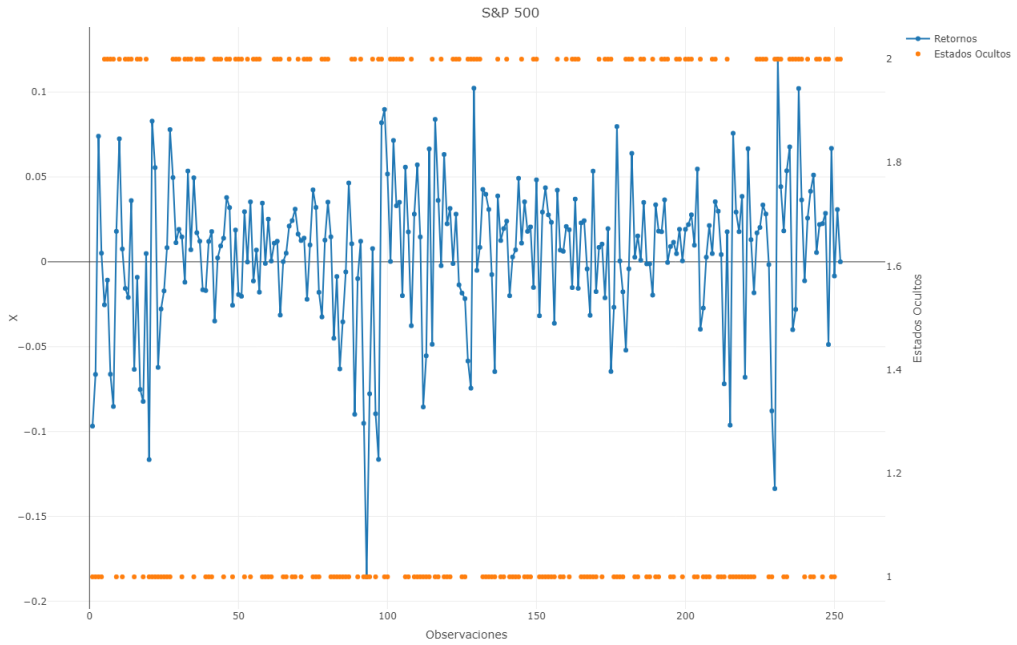


Figura 3.6: log retornos S&P 500 y dos estados ocultos

Veamos los valores de las probabilidades iniciales posteriores, la ma-

triz de transición posterior, los parámetros posteriores de las emisiones Gausianas; es decir, el vector de medias y precisión. Todo lo mencionado anteriormente para dos estados ocultos.

$$\pi = \begin{bmatrix} 0{,}5000098 \\ 0{,}4999902 \end{bmatrix} A = \begin{bmatrix} 0{,}4999689 & 0{,}5000311 \\ 0{,}4999683 & 0{,}5000317 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0{,}00627 \\ 0{,}05113 \end{bmatrix} \tau = \begin{bmatrix} 0{,}01048 \\ 0{,}08840 \end{bmatrix}$$

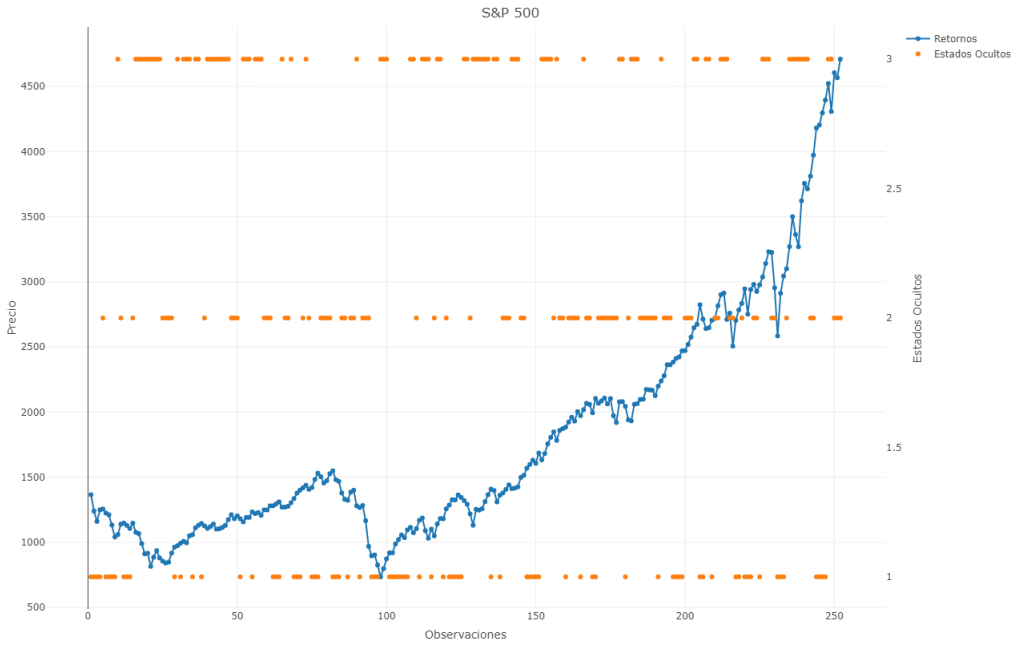


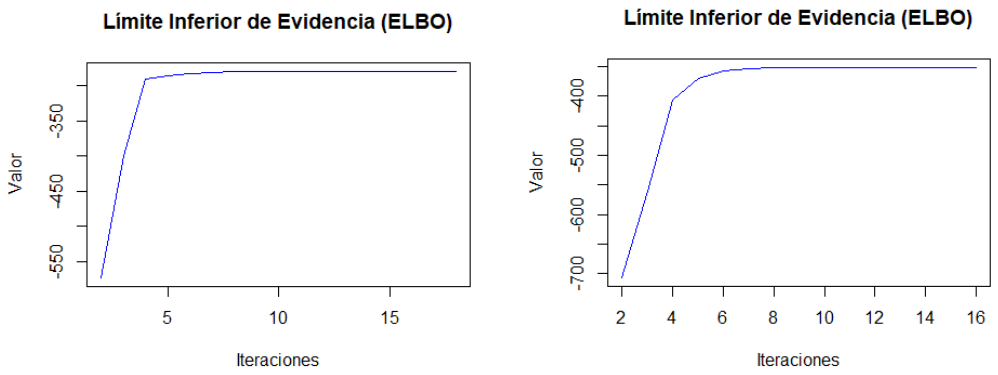
Figura 3.7: Precios S&P 500 y tres estados ocultos

La figura 3.6 muestra dos estados ocultos posteriores en conjunto con los log-retornos de la serie $S\&P500$, mientras que en la figura 3.7 indica el resultado de tres estados ocultos posteriores con el precio de las acciones de la serie Standard & Poor's durante Enero del 2001 hasta diciembre del 2021

El código que se implementó para obtener los resultados presentados se encuentra en el anexo D

3.1.3. DAX

Ahora se analiza la convergencia del limite inferior de Evidencia (ELBO) para la serie de tiempo Deutscher Aktienindex (DAX).



(a) ELBO dos estados (b) ELBO tres estados

Figura 3.8: Convergencia ELBO para DAX

	2 estados	**3 estados**
logLik	-232.62	-349.834
AIC	588.74	888.332
BIC	2471.27	3724.62
DIC	470.66	701.87

Cuadro 3.2: Criterios AIC, BIC y DIC para la serie DAX

Como se puede apreciar en la figura 3.8 la convergencia con dos estados ocultos en más rápida, por lo que se tiene menos iteraciones hasta llegar al óptimo, y los resultados del cuadro 3.2 nos siguen indicando que el criterio Bayesiano elige al modelo con dos estados ocultos y el algoritmo converge con un valor del logaritmo de verosimilitud de $-277{,}71$ en el caso de dos estados ocultos.

A continuación se presenta los valores de las probabilidades iniciales posteriores, la matriz de transición posterior, los parámetros posteriores de las emisiones Gausianas; es decir, los vectores de medias y precisión. Todo lo mencionado anteriormente para dos estados ocultos.

$$\pi = \begin{bmatrix} 0{,}462 \\ 0{,}538 \end{bmatrix} \quad A = \begin{bmatrix} 0{,}5010348 & 0{,}4989652 \\ 0{,}5010344 & 0{,}4989656 \end{bmatrix}$$

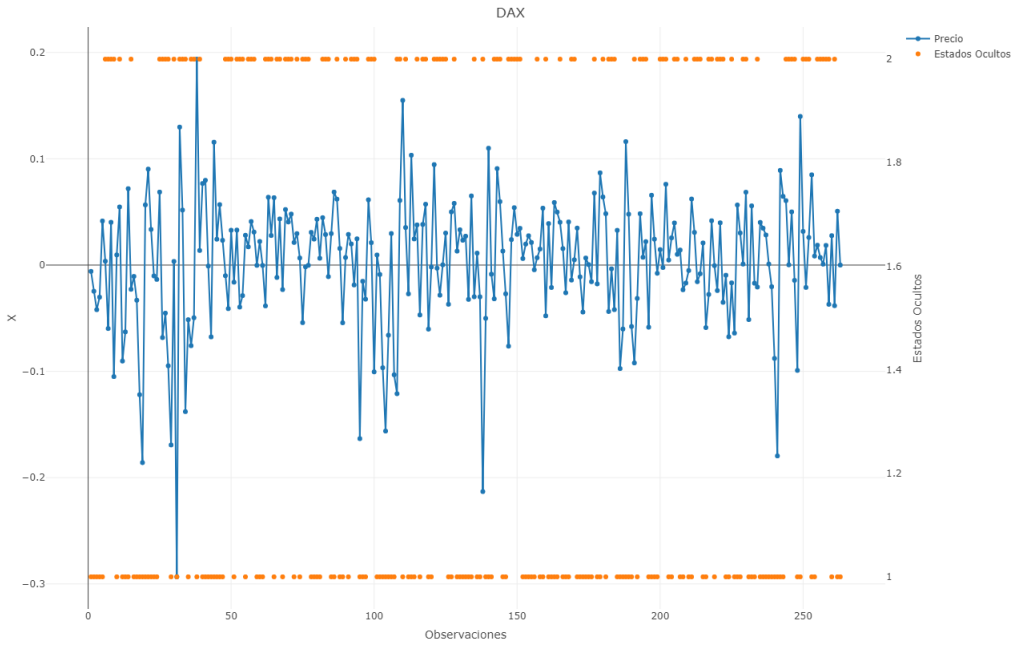


Figura 3.9: log retornos DAX y dos estados ocultos

$$\mu = \begin{bmatrix} 0{,}00629 \\ 0{,}05096 \end{bmatrix} \tau = \begin{bmatrix} 0{,}01051 \\ 0{,}08833 \end{bmatrix}$$

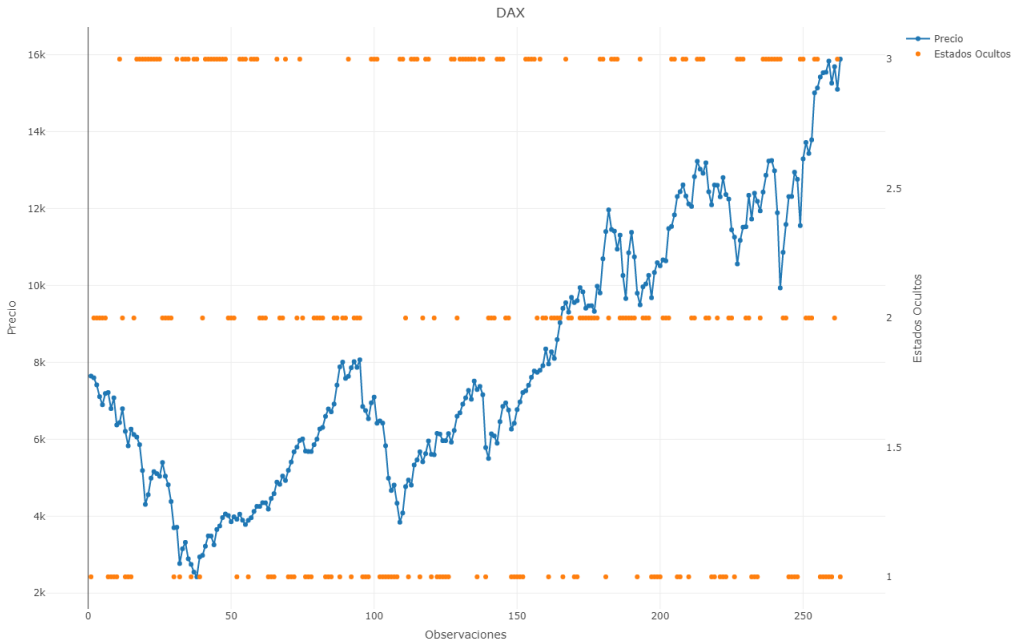


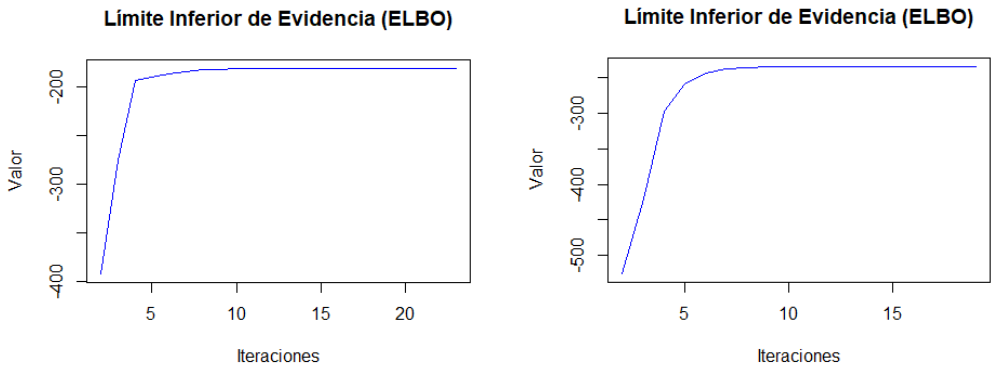
Figura 3.10: Precios DAX y tres estados ocultos

La figura 3.9 muestra dos estados ocultos posteriores en conjunto con los log-retornos de la serie DAX, mientras que en la figura 3.10 indica el resultado de tres estados ocultos posteriores con el precio de las accio-

nes de la serie Deutscher Aktienindex (DAX) durante Enero del 2000 y diciembre del 2021

3.1.4. Acciones Coca Cola

A continuación se presenta los resultados obtenidos para los log-retornos de las acciones de Coca Cola



(a) ELBO dos estados (b) ELBO tres estados

Figura 3.11: Convergencia ELBO para KO

	2 estados	**3 estados**
logLik	-179.88	-233.30
AIC	590.99	1127.4
BIC	3416.68	4910.63
DIC	472.41	469.86

Cuadro 3.3: Criterios AIC, BIC y DIC para la serie KO

Como se había mencionado anteriormente el criterio Bayesiano elige los modelos con menos estados ocultos; además en el cuadro 3.3 que las diferencias de los valores entre estados ocultos presentan una diferencia considerable.

Presentamos los valores de las probabilidades iniciales posteriores, la matriz de transición posterior, los parámetros posteriores de las emisiones Gausianas; es decir, el vector de medias y precisión. Todo lo mencionado anteriormente para dos estados ocultos.

$$\pi = \begin{bmatrix} 0{,}4620916 \\ 0{,}5379084 \end{bmatrix} \quad A = \begin{bmatrix} 0{,}5031588 & 0{,}4968412 \\ 0{,}5031712 & 0{,}4968288 \end{bmatrix}$$

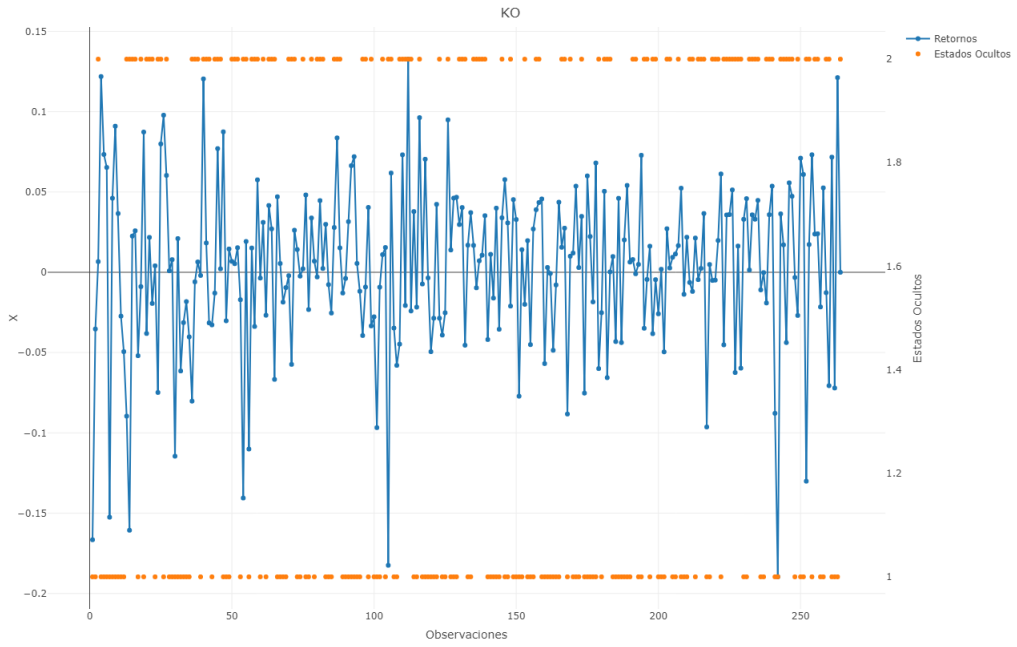


Figura 3.12: log retornos de Coca Cola y dos estados ocultos

$$\mu = \begin{bmatrix} 0{,}00626 \\ 0{,}05100 \end{bmatrix} \tau = \begin{bmatrix} 0{,}01047 \\ 0{,}08839 \end{bmatrix}$$

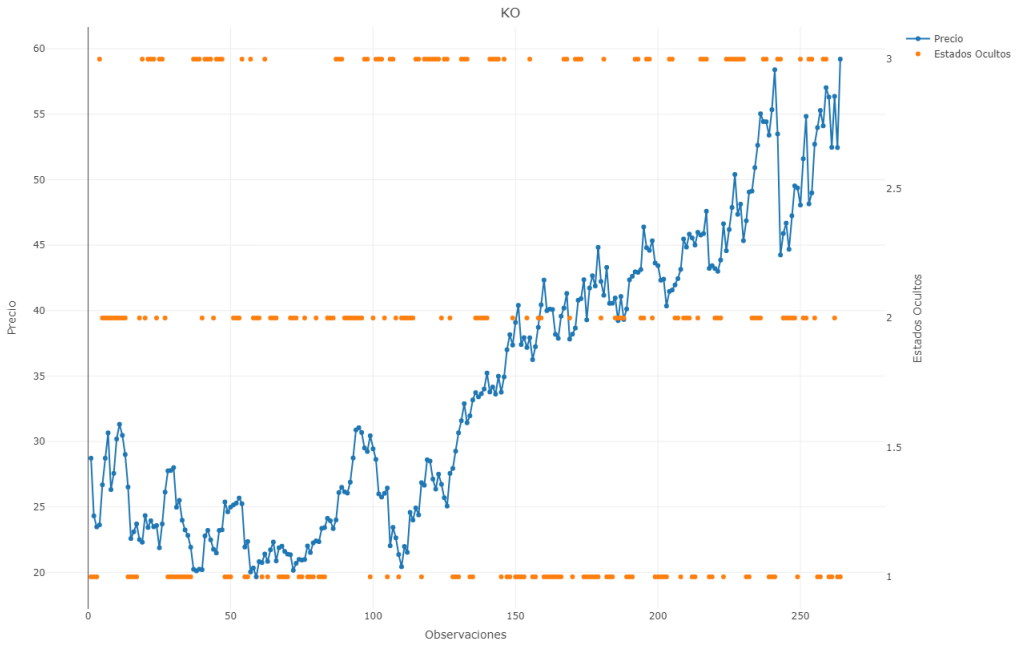


Figura 3.13: Precios de las acciones de KO y tres estados ocultos

La figura 3.4 muestra dos estados ocultos posteriores en conjunto con los log-retornos de la serie KO, mientras que en la figura 3.10 indica el resultado de tres estados ocultos posteriores con el precio de las acciones

de la compañía Coca Cola (KO) durante Enero del 2000 y diciembre del 2021

Para obtener los resultados de la serie (DAX) y los precios de las acciones (KO) se usa el mismo código de (S&P 500), tomando en cuenta que se debe cargar los datos de las series en estudio y definir su propia X.

3.2. Conclusiones y recomendaciones

La Inferencia Variacional (VI) para el calculo de Modelos Ocultos de Markov, (HMM) en series financieras, ha sido muy útil ya que se han obtenido buenos resultados y el tiempo de convergencia para el algoritmo planteado ha sido relativamente rápido.

- Se uso los métodos de Inferencia Variacional, considerando la teoría del campo medio para ajustar un modelo oculto de Markov aplicado a las finanzas, además, cabe recalcar que esta teoría asume que todos los parámetros son independientes.
- Se estudio algunos de los modelos que se han implementado [7], [23], [22] y parte del paquete *depmixS4* ejecutado en el software estadístico R, para crear el modelo presentado.
- Se validó el modelo con los diferentes criterios que han sido considerados para evaluar a los Modelos Ocultos de Markov, tales como AIC, BIC y DIC que ha sido utilizado por [23].
- En este trabajo adaptamos el concepto de técnicas de Inferencia Variacional en Modelos Ocultos de Markov. Se ha definido un algoritmo siguiendo las ideas de la Inferencia Bayesiana Variacional, la Divergencia de Kullback Leibler y el limite inferior de Evidencia ELBO.
- Las familias elegidas para el realizar el trabajo fueron distribuciones discretas y continuas; Dirichlet y Gaussianas respectivamente.
- Para la ejecución de los algoritmos presentados se usó un computador personal, sin presentar ningún inconveniente en los tiempos de ejecución, puede ser por la cantidad de observaciones con las que se trabajo o la complejidad computacional del algoritmo.

- Se puede interpretar al modelo con dos estados ocultos, como dos regímenes, es decir, un periodo de normalidad y un estado de crisis.
- Se interpreta a los modelos con tres estados ocultos como: un periodo de crisis, normalidad e hipercrisis.
- Se debe tener en cuenta que; para ejecutar el modelo con dos estados ocultos, los hiper parámetros iniciales se mantienen para las tres series en estudio, sin presentar problemas de convergencia.
- Para futuros trabajos se puede implementar un paquete en R Studio, para modelos ocultos de Markov con Inferencia Variacional.
- Una recomendación, al momento de inicializar los valores de los hiper parámetros se debe tomar en cuenta que no converge con cualquiera, se busco diferentes valores hasta obtener el adecuado, cabe recalcar que en algunos de los trabajos presentados se usa números aleatorios. Para variar en los resultados, se hizo una búsqueda manual de los valores que podían ser utilizados en el presente trabajo de integración,
- Se puede crear a futuro un algoritmo para la detección de anomalías en series temporales financieras usando la técnica que hemos propuesto; Modelos de Ocultos de Markov con emisiones Gaussianas con técnicas de Inferencia Variacional.
- El lector si esta interesado en desarrollar más de lo que se ha obtenido en el presente trabajo, puede explorar la técnica de la Inferencia Variacional Estocástica (Stocastic Variational Inference, SVI), Johnson [13] un articulo acerca de este método para obtener Modelos Ocultos de Markov (HMM).

Capítulo A

Divergencia de Kullback Leibler

En Teoría de Probabilidades la Divergencia de Kullback Leibler (KL), o también conocida como divergencia de la información, es una medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad P y Q.

A.1. Definición

Para dos distribuciones de probabilidad P y Q de una variable aleatoria discreta su divergencia KL se define en el mismo espacio de probabilidades, $\mathcal{X}$, como:

$$KL(P||Q) = \sum_{x \in \mathcal{X}} P(x) \ln \left(\frac{P(x)}{Q(x)} \right) \tag{A.1}$$

o su vez,

$$KL(P||Q) = -\sum_{x \in \mathcal{X}} P(x) \ln \left(\frac{Q(x)}{P(x)} \right) \tag{A.2}$$

A.2. Propiedades

- Es siempre positiva $KL(P||Q) \geq 0$
- Es nula si y solo si $P == Q$.
- No es simétrica

Capítulo B

Algoritmo Forward-Backward

En los Modelos Ocultos de Markov uno de los problemas es calcular la probabilidad de las secuencias observadas X dado un modelo $\theta = (\pi, A, \mu, \tau)$, con el objetivo de calcular $P(X|\theta)$ de la siguiente manera

$$P(X|\theta) = \sum_{i=1}^{N} \alpha_T(i) \tag{B.1}$$

La variable Forward esta definida con una probabilidad conjunta de la secuencia de observaciones parciales a un tiempo t y un estado s_t

$$\begin{aligned} \alpha_t(i) &= p(x_1, \ldots, x_t | z_t = i) \\ &= \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} b_i(x_t) \end{aligned} \tag{B.2}$$

Este es calculado usando la recursividad. nos basaremos en el escalamiento propuesto por [29]

- **Inicialización**

  1. $\alpha_1(i) = \pi_i b_i(x_1) = \pi_i p(x_1|z_t = i)$
  2. Sea $c_1 = \frac{1}{\sum_{i=1}^{N} \alpha_1(i)}$
  3. Sea $\tilde{\alpha}_t(i) = c_1 \alpha_t(i)$

- **Recursión**

1. $\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(i) a_{ji} b_i(x_t) = \sum_{j=1}^{N} \alpha_{t-1}(i) a_{ji} p(x_t|z_t = i)$
2. Sea $c_t = \frac{1}{\sum_{i=1}^{N} \alpha_t(i)}$
3. Sea $\tilde{\alpha}_t(i) = c_t \alpha_t(i)$

La variable Backward esta definida como la probabilidad de generar los últimos $T - t$ observaciones dada.

$$\beta_t(i) = p(x_{t+1}, \ldots, x_T|z_t = i) \tag{B.3}$$

- **Inicialización** en $t = T$
  1. $\beta_T(i) = c_T$
- **Recursión** para $1 \leq t < T$
  1. $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_i(x_{t+1}) \beta_{t+1}(i) = \sum_{j=1}^{N} a_{ij} p(x_{t+1}|z_{t+1} = i) \beta_{t+1}(i)$
  2. Sea $\tilde{\beta}_t(i) = c_t \beta_t(i)$

Capítulo C

Aproximación Variacional para p_D y DIC

Como había mencionado durante este trabajo, que un criterio para la selección del modelo es el criterio de Desviación de Información (DIC), se va ha detallar más acerca del mismo.

Nuestra aproximación variacional para p_D

$$\begin{aligned} p_D &= \mathbb{E}_\theta[-2\log(p(X|\theta))] + 2\log(p(X|\tilde{\theta})) \\ &= -2\int q_\theta(\theta)\log\left(\frac{q_\theta(\theta)}{p(\theta)}\right)d\theta + 2\log\left(\frac{q_\theta(\tilde{\theta})}{p(\tilde{\theta})}\right) \\ &= -2(\sum_i\sum_j\sum_t \gamma_t(i,j)\left(\psi(a_{ij}) - \psi(\sum_j a_{ij})\right) \\ &+ \sum_j\sum_t \gamma_t(j)\left(\frac{1}{2}(\psi(\frac{a_j}{2}) - \log(\frac{b_j}{2}) - \frac{1}{\beta_j}\right)) \\ &- 2\left(\sum_i\sum_j\sum_t \gamma_t(i,j)\log\left(\frac{a_{ij}}{\sum_j a_{ij}}\right) + \frac{1}{2}\sum_j\sum_t \gamma_t(j)\log\left(\frac{a_j}{b_j}\right)\right) \end{aligned} \tag{C.1}$$

El valor de DIC puede ser encontrado usando la siguiente fórmula

$$\text{DIC} = 2p_D - 2\log(p(X|\tilde{\theta})) \tag{C.2}$$

$\log(p(X|\tilde{\theta}))$ se calcula usando el algoritmo Fordward.

$$\log(p(X|\tilde{\theta}) = -\sum_{t=1}^{T} \log(c_t) \tag{C.3}$$

Capítulo D

Códigos

D.1. Recursión Fordward - Backward

```
# Funciones de recursion Forward
forward_backward <- function(T, N, initDist = ai_j, A, B)
{
  ct <- rep(1,T)
  fvar <- matrix(nrow = T,ncol = N)
  fvar_tilde <- matrix(nrow = T,ncol = N)
  bvar <- matrix(nrow = T,ncol = N)
  bvar_tilde <- matrix(nrow = T,ncol = N)
  fvar[1,] <- initDist*B[1,]
  ct[1] <- 1/sum(fvar[1,])
  fvar_tilde[1,] <- fvar[1,]*ct[1]

  for(t in 2:T){
    for(i in 1:N){
      fvar[t,i] <- sum(fvar_tilde[t-1,i]*A[i,])*B[t,i]
    }
    ct[t] <- 1/sum(fvar[t,])
    fvar_tilde[t,] <- fvar[t,]*ct[t]
  }

  bvar_tilde[T,] <- ct[T]
```

```
  for(t in (T-1):1){
    for(i in 1:N){
      bvar[t,i] <- sum(A[i,]*bvar_tilde[t+1,]*B[t+1,])
    }
    bvar_tilde[t,] <- bvar[t,]*ct[t]
  }
  output <- list('fvar_tilde'=fvar_tilde,
  'bvar_tilde'=bvar_tilde,'ct'= ct )

  return(output)
}
```

D.2. Criterio de Desviación de Información (DIC)

```
# Calculo de DIC (Criterio de Desviacion de Informacion)
uDIC <- function(N,  stateProb, initProb, jtTransMat, ct,
xi, alpha,a_j,b_j,beta_j){
  pd0 <- 0
  pd1 <- 0
  for(j in 1:N)
  {
    pd0 <- pd0 + sum(stateProb[,j])*(1/2)*(digamma(a_j[j])-
    log(b_j[j]/2) + 1/beta_j[j] )
    for(k in 1:N)
      pd1 <- pd1 + rowSums(jtTransMat,dims = 2)[j,k]*(
      log(alpha[j,k]) - log(sum(alpha[j,]))
      - digamma(alpha[j,k]) + digamma(sum(alpha[j,])) )

  }
  pd3 <- sum(initProb*(log(xi) - log(sum(xi))
  - digamma(xi) + digamma(sum(xi))))
  pd <- pd0+sum(pd1)+pd3
  dic <- 4*(pd)  + 2*sum(log(ct))
  output = list('pd' = pd, 'dic' = dic)
  return(output)
}
```

D.3. Límite Inferior de Evidencia (ELBO)

```
# Calculo de ELBO (Limite inferior de la Evidencia)
uELBO = function(N, stateProb, initProb, jtTransMat, ct, xi_0,xi,alpha_0, alph
  kl_pi  <- 0
  kl_A   <- 0
  kl_mu  <- 0
  kl_tao <- 0
  for(j in 1:N){

    kl_tao <- kl_tao + (a[j]/2)*(log(b0[j]/b[j]))-lgamma(a0[j]/2)+
      lgamma(a[j]/2) - sum(stateProb[,j])*( digamma(a0[j]/2)) +
      (beta0[j]*m0[j]*m0[j]+
         sum(stateProb[,j]*obs*obs)-beta_j[j]*m_j[j]*m_j[j])*(a0[j]/b0[j] )

    kl_A <- kl_A + sum(rowSums(jtTransMat,dims = 2)[j,]*( digamma(alpha[j,]) -
      lgamma(sum(alpha[j,])) - sum(lgamma(alpha[j,])) -lgamma(sum(alpha_0[j,])

    kl_mu <- kl_mu - 0.5*log(beta_j[j]/beta0[j]) +
      0.5*(tao[j]*beta_j[j]*(m0[j]-m_j[j])**2)+  0.5*(beta_j[j])/beta0[j] - 0
  }
  kl_pi <- kl_pi + sum(apply(initProb, 2, sum)*(digamma(xi) - digamma(sum(xi))
    lgamma(sum(xi)) - sum(lgamma(xi)) - lgamma(sum(xi_0)) + sum(lgamma(xi_0))

  elbo <-  -sum(log(ct))  - kl_A  - kl_pi - kl_tao -kl_mu
  return(elbo)
}
```

D.4. Actualizaciones

```
## Actualizacion variables latentes

latentes <- function(N,T,xi_j,alpha_j,m_j,beta_j,a_j,b_j,X){
  #### Estimadores iniciales de las variables latentes
  # matriz de trasicion posterior
  a_jk        <- matrix(NA, nrow = N, ncol = N)
```

```
  # matriz de emision posterior
  b_tj        <- matrix(NA, nrow = T, ncol = N)
  # probabilidad inicial posterior
  a_1j        <- rep(NA,N)

  a_1j <- exp(digamma(xi_j) - digamma(sum(xi_j)))
  for(j in 1:N){
    a_jk[j,] <- exp(digamma(alpha_j[j,]) -
    digamma(sum(alpha_j[j,])))
    b_tj[,j] <- exp((1/2)*(digamma(a_j[j]/2)-
    log(b_j[j]/2)-(a_j[j]/b_j[j])*(X-m_j[j])**2 -
    1/beta_j[j]))
  }
  output <- list('pi_tilde'=a_1j,'A_tilde'=a_jk,'B_tilde'=b_tj)
  return(output)
}
```

q_{jk} esta definida como la probabilidad de estado j en el tiempo t y simultáneamente trasicionando al estado j en el tiempo $t+1$

q_{tj} esta definida para almacenar la probabilidad del estado j en el tiempo t

```
probabilidades <- function(ct,bvar,fvar,N,T){
  # distribucion estacionaria de la probabilidad posterior
  q_tj        <- matrix(nrow = T, ncol=N)
  # matrix de probabilidad conjunta posterior
  q_jk        <- array(dim=c(N, N,T-1))
  # probabilidad inicial posterior
  q_1j        <- rep(0,N)

  q_tj <- (fvar*bvar)/sum(bvar*fvar)
  for(j in 1:N){
    for(l in 1:N){
      q_jk[j,l,] <- fvar[-T,j]*a_jk[j,l]*b_tj[-1,l]*bvar[-1,l]
    }}

  q_jk <- q_jk/sum(bvar*fvar)
  q_1j = t(q_tj[1,])
  output <- list('gamma_1j'=q_1j,'gamma_tj'=q_tj,
```

```
  'gamma_tij'=q_jk)
  return(output)
}


hiperparametros <- function(N,xi_0,q_1j,q_tj,q_jk,alpha_0,
                            m0,beta0,a0,b0,X){
  xi_j         <- NULL
  alpha_j      <- matrix(NA,nrow=N,ncol=N)
  m_j          <- NULL
  beta_j       <- NULL
  a_j          <- NULL
  b_j          <- NULL
  for(j in 1:N){
    xi_j[j] <- xi_0[j] + q_1j[,j]
    alpha_j[j,] <- alpha_0[j,] + rowSums(q_jk,dims = 2)[j,]
    m_j[j] = (beta0[j]*m0[j]+sum(q_tj[,j]*X))/(beta0[j]+
    sum(q_tj[,j]))
    beta_j[j] = beta0[j]+ sum(q_tj[,j])
    a_j[j] = a0[j] + sum(q_tj[,j])
    b_j[j] = b0[j] + beta0[j]*m0[j]*m0[j] + sum(q_tj[,j]*X*X) +
    beta_j[j]*m_j[j]*m_j[j]
  }
  output <- list('x_1j'=xi_j,'alpha_j'=alpha_j,'m_j'=m_j,
                 'beta_j'=beta_j,'a_j'=a_j,'b_j'=b_j)
  return(output)
}
```

D.5. S&P 500

```
## Librerias a usarse
library(ConnMatTools)
library(readr)
library(readxl)
library(LaplacesDemon)
library(PerformanceAnalytics)
```

```r
library(plotly)

## Cargar funciones
source('Script/Funciones.R')

## Cargar los datos
data <- read_excel("Data/Datos_historicos_S&P_500.xlsx")
Precio <- data$Ultimo
## log returns
X <- log_return(Precio)
## Fijar la semilla
set.seed(123)
## Numero de estados
n_estados = 2

  T          <- length(X)           # Longitud de los datos
  N        <- n_estados             # Numero de estados
  #### Criterio de Convergencia
  maxiter    <- 500                 # Maximo nro de iteraciones
  dic        <- rep(0,maxiter)
  dic_old    <- 20000
  dic[1]     <- 10000
  elbo       <- rep(0,maxiter)
  elbo_old   <- -20000
  elbo[1]    <- -10000
  tol        <- 10^(-6)
  iter       <- 1

  #Parametros para dos estados ocultos
  m0         <- rep(0,N)  ## parametros de la funcion normal
  beta0      <- rep(1,N)  ## parametros de la funcion normal
  a0         <- rep(1,N)  ## parametros iniciales de la funcion Gamma
  b0         <- rep(2,N)  ## parametros iniciales de la funcion Gamma
  mu         <- rep(mean(X),N)
  tao        <- rep(1,N)
  improvement_dic   <- (dic_old-dic[1])/dic_old
  improvement_elbo  <- (elbo_old-elbo[1])/elbo_old
```

```
#### Estimacion de hyper parametros iniciales
 # Matriz de transicion inicial
alpha_0      <- obt_A(N)
# Matriz de Emision Gaussiana a Priori
emi_0        <- obt_B(X,mu,tao)
# Probalidades iniciales a priori
xi_0         <- rep(1,N)/N

#### Estimadores iniciales de las variables latentes
# matriz de transicion posterior
a_jk         <- matrix(NA, nrow = N, ncol = N)
# matriz de emision posterior
b_tj         <- matrix(NA, nrow = T, ncol = N)
# probabilidad inicial posterior
a_1j         <- rep(NA,N)

ct           <- rep(NA,T)
fvar         <- matrix(nrow= T, ncol =  N)      # forward
bvar         <- matrix(nrow= N, ncol =  T)      # backward
a_iter       <- rep(0,maxiter)
# distribucion estacionaria de la probabilidad posterior
q_tj         <- matrix(nrow = T, ncol=N)
# matrix de probabilidad conjunta posterior
q_jk         <- array(dim=c(N, N,T-1))
# probabilidad inicial posterior
q_1j         <- rep(NA,N)

#### variables posteriores de los hiperparametros

alpha_j      <- alpha_0  # matriz de transicion
emision_tj   <- emi_0    # Matriz de emision
# Parametros Dirichlet porteriores para distribucion inicial
xi_j         <- xi_0
m_j          <- as.vector(kmeans(X,N)[[2]])
beta_j       <- rep(1,N)
a_j          <- a0              # funcion gamma
b_j          <- b0              # funcion gamma
mu_post      <- rep(NA,N)
```

```
 tao_post     <- rep(NA,N)

gama_1t  <- matrix(nrow = maxiter, ncol = N)

 #### Optimizacion variacional
 while((abs(improvement_elbo) > tol | improvement_elbo < 0) &
           iter < maxiter){

   #### Actualizacion de variables Latentes
   var_latentes <-
          latentes(N,T,xi_j,alpha_j,m_j,beta_j,a_j,b_j,X)

   a_1j <- var_latentes$pi_tilde
   a_jk <- var_latentes$A_tilde
   b_tj <- var_latentes$B_tilde

   #### Forward and Backward recursions

   fvar_out <- forward_backward(T,N, a_1j, a_jk, b_tj)
   fvar <- fvar_out[[1]]
   bvar <- fvar_out[[2]]
   ct   <- fvar_out[[3]]

   #### Actualizacion de probabilidades de estado posteriores
   prob_state_post <- probabilidades(ct,bvar,fvar,N,T)
   q_1j <- prob_state_post$gamma_1j
   q_tj <- prob_state_post$gamma_tj
   q_jk <- prob_state_post$gamma_tij

   #### Actualizacion de hiperparametros
   hiper_out <-
   hiperparametros(N,xi_0,q_1j,q_tj,q_jk,alpha_0,m0,
   beta0,a0,b0,X)
   xi_j    <- hiper_out$x_1j
   alpha_j <- hiper_out$alpha_j
   m_j     <- hiper_out$m_j
   beta_j  <- hiper_out$beta_j
   a_j     <- hiper_out$a_j
```

```
    b_j     <- hiper_out$b_j

    for (j in 1:N) {
      mu_post[j]  <- sum(q_tj[,j]*X)/sum(q_tj[,j])
      tao_post[j] <- sum(q_tj[,j]*(X-m_j[j])**2)/sum(q_tj[,j])
    }

    emision_tj <- obt_B(X,mu_post,tao_post)

    #### DIC y ELBO
    dic[iter+1]     <- uDIC(N, q_tj, initProb = q_1j,
       jtTransMat = q_jk,ct = ct, xi = xi_j, alpha = alpha_j,
       a_j, b_j, beta_j)[[2]]
    dic_old         <- dic[iter]
    improvement_dic <- (dic_old-dic[iter+1])/dic_old
    #

    elbo[iter+1]    <- uELBO(N,q_tj,q_1j,q_jk,ct,xi_0,xi_j,
        alpha_0,alpha_j,X,a0, a_j,b0,b_j,tao_post,beta0,beta_j,
        m0,m_j)
    elbo_old        <- elbo[iter]

    improvement_elbo <- (elbo[iter+1]-elbo_old)/abs(elbo_old)

    iter <- iter+1
  }


 params <- post_param(numobs = T,N, alpha = alpha_j, xi = xi_j)
 pi.post     <- params$InitDist
 trans.post   <- params$TransMat
 emision.post   <- emision_tj


plot_elbo <- plot(2:iter,elbo[2:iter],'l',
main = 'Limite␣Inferior␣de␣Evidencia␣(ELBO)',
xlab = 'Iteraciones', ylab = 'Valor',col="blue")
```

```
plot_dic <- plot(2:iter,dic[2:iter],'l',
main = 'Criterio de desviacion de Informacion (DIC)',
xlab = 'Iteraciones', ylab = 'Valor',col="blue")



G = (n_estados-1) + (n_estados)*(n_estados-1) +
(n_estados)*(length(X)-1)
Z  <- estados_ocultos(N,T,trans.post)
logLikelihood <- log(verosimilitud(X,Z,pi.post,trans.post,emision_tj))
logLik <- -sum(log(ct))
AIC.t <- -2*(logLik)+2*G
BIC.t <- -2*(logLik)+G*log(T)
DIC.t <- dic[iter]



plot(ts(X,start = c(2001,1,1),freq=12),xlab = 'Datos',
ylab = 'Valor',col="blue",main = ' log Retornos S&P 500')

# Resumen Criterios
criterios <- function(logLik,AIC.t,BIC.t,DIC.t){
  output <- list('logLik'= logLik, 'AIC'=AIC.t,
  'BIC'=BIC.t,'DIC'=DIC.t)
  return(output)
}
criterios(logLik,AIC.t,BIC.t,DIC.t)
```

Referencias bibliográficas

[1] Torben G Andersen. The econometrics of financial markets: John y. campbell, andrew w. lo, and a. craig mackinlay, princeton university press, 1997. *Econometric Theory*, 14(5):671–685, 1998.

[2] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[3] Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom), 2003.

[4] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[5] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians, 2016. *ISSN 1537274X*, 2018.

[7] Matthew Danielson. *Hidden Markov models with variational inference in marketing science*. PhD thesis, University of St Andrews, 2021.

[8] AG Deutsche Boerse. Guide to the equity indizes of deutsche boerse. *URL: http://www. dax-indices. com/DE/MediaLibrary/Document/Equity_L_6_19_e. pdf, accessed on October*, 13:2014, 2013.

[9] Rafael Eduardo Diaz Bonilla. Métodos bayesianos para modelos ocultos de markov en series de tiempo con conteo. *Departamento de Estadística*, 2019.

[10] Christian Gruhl and Bernhard Sick. Variational bayesian inference for hidden markov models with multivariate gaussian output distributions. *arXiv preprint arXiv:1605.08618*, 2016.

[11] Shihao Ji, Balaji Krishnapuram, and Lawrence Carin. Variational bayes for continuous hidden markov models and its application to active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):522–532, 2006.

[12] Mark Johnson. Why doesn't em find good hmm pos-taggers? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 296–305, 2007.

[13] Matthew Johnson and Alan Willsky. Stochastic variational inference for bayesian time series models. In *International Conference on Machine Learning*, pages 1854–1862. PMLR, 2014.

[14] Michael Irwin Jordan. *Learning in graphical models*. MIT press, 1999.

[15] Sylvia Kaufmann. Hidden markov models in time series, with applications in economics. In *Handbook of Mixture Analysis*, pages 309–341. CRC Press, 2016.

[16] Chang-Jin Kim. Dynamic linear models with markov-switching. *Journal of Econometrics*, 60(1-2):1–22, 1994.

[17] Sangjoon Kim, Neil Shephard, and Siddhartha Chib. Stochastic volatility: likelihood inference and comparison with arch models. *The review of economic studies*, 65(3):361–393, 1998.

[18] Jaroslav Lajos, KM George, and N Park. A six state hmm for the s&p 500 stock market index. In *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods (MSV)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2011.

[19] Paul Felix Lazarsfeld. *The use of mathematical models in the measurement of attitudes*. Rand Corporation, 1951.

[20] Stephen HT Lihn. Hidden markov model for financial time series and its application to s&p 500 index. *Quantitative Finance, Forthcoming*, 2017.

[21] David JC MacKay. Ensemble learning for hidden markov models. Technical report, Citeseer, 1997.

[22] Reetam Majumder, Matthias K Gobbert, Amita Mehta, and Nagaraj K Neerchal. Variational bayes estimation of hidden markov models for daily precipitation with semi-continuous emissions. *UMBC Joint Center for Earth Systems Technology (JCET)*, 2021.

[23] Clare A McGrory and DM Titterington. Variational bayesian analysis for hidden markov models. *Australian & New Zealand Journal of Statistics*, 51(2):227–244, 2009.

[24] George A Miller. Finite markov processes in psychology. *Psychometrika*, 17(2):149–167, 1952.

[25] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[26] Nguyet Nguyen and Dung Nguyen. Hidden markov model for stock selection. *Risks*, 3(4):455–473, 2015.

[27] William D Penny, Nicho Menghi, and Louis Renoult. Cluster-based inference for memory-based cognition. *bioRxiv*, 2022.

[28] Wilfred Perks. Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, 73(2):285–334, 1947.

[29] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[30] Christian P Robert, Tobias Ryden, and David M Titterington. Bayesian inference in hidden markov models through the reversible jump

markov chain monte carlo method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):57–75, 2000.

[31] Tobias Rydén. Em versus markov chain monte carlo for estimation of hidden markov models: A computational perspective. *Bayesian Analysis*, 3(4):659–688, 2008.

[32] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

[33] Yousuke Takada. More prml errata, 2018.

[34] Shinji Watanabe, Yasuhiro Minami, Atsushi Nakamura, and Naonori Ueda. Application of variational bayesian approach to speech recognition. *Advances in Neural Information Processing Systems*, 15, 2002.

[35] Lee Manning Wiggins. *Mathematical models for the interpretation of attitude and behavior change: The analysis of multi-wave panel*. Columbia University, 1955.

[36] Yingjian Zhang. *Prediction of financial time series with Hidden Markov Models*. PhD thesis, Applied Sciences: School of Computing Science, 2004.

[37] Walter Zucchini and Iain L MacDonald. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC, 2009.