MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Fikri *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 19 (2023) 840-854

RESEARCH ARTICLE

Recursive Parameter Estimation and Its Convergence for Multivariate Normal Hidden Markov Inhomogeneous Models

Miftahul Fikri[a,b], Zulkurnain Abdul-Malek[a]*, Mona Riza Mohd Esa[a], Eko Supriyanto[c]

[a]High Voltage and High Current Institute, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia; [b]Faculty of Electricity and Renewable Energy, Institut Teknologi PLN,11750 Jakarta Barat, Jakarta, Indonesia; [c]School of Biomedical Engineering and Health Sciences, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor Malaysia

Abstract In this paper, will discussed parameter estimation and convergence analysis of multivariate normal hidden inhomogeneous Markov models. The results of this research show that by using the expectation maximization algorithm, a sequence of parameter estimators converges to a stationary point of the likelihood function in a monotonically increasing manner.

Keywords: Hidden Markov model inhomogeneous, Multivariat normal, Likelihood function, Expectation Maximization, Monotone convergence.

Introduction

The hidden Markov model (HMM) is a pair of stochastic processes, namely the observation process and the process causing the observation [1]. It is assumed that the stochastic processes that influence these observations are not observed and form a Markov chain, that is, the probability of the effect of an observation at one time only depends on the effect of the observation several units of time before. The effect of this observation is usually called state [2]. HMM is widely applied in various problem areas in the form of time series data such as air pollution problems [3], [4], weather predictions [5], [6], stock price predictions [7], [8], [9], [10], [11], speech recognition [12], [13], [14], prediction of DNA sequences [15] [16], and and It is hoped that it can be applied to diagnose partial discharge acoustic in insulation [17], [18], [19], [20]. This is because HMM maintains relevance to the issues discussed and offers simplification in calculations (memoryless properties) [21]. As for the application to longitudinal data, even though it offers efficiency, it is still very little. This is because the required analyzes are not as easy as when applied to time series data.

To apply to longitudinal data, multivariate assumptions are needed in the model which is the focus of this study, namely the normal hidden Markov multivariate model (MNHMM). The normal hidden Markov multivariate model (MNHMM) is one of the hidden Markov models where the probability of an observation if the state is known is assumed to be a normal multivariate distribution [22], [23], [24]. In previous studies [22], [23], [24], parameter estimation and convergence analysis of MNHMM or simulation have been carried out, but the Markov chain is still assumed to be homogeneous so that the objective function in the form of the likelihood function will obtained less than maximum results. Therefore, in this study, it is assumed that the Markov chain is not homogeneous and the proposed model is the multivariate normal hidden Markov model inhomogeneous at one time before (MNHMM-I) which is expected to later be used to increase the maximum likelihood function obtained and can be used for clustering or predict more accurately.

The novelty in this research is constructing the MNHMM-I, estimating parameters and analyzing the convergence of the parameter estimators. Model construction was carried out using a combination of

***For correspondence:**
miftahul@itpln.ac.id

Received: 25 May 2023
Accepted: 21 Sept. 2023

homogeneous multivariate [24] and multiple non-homogeneous [25], [26]. Parameter estimation is done by maximizing the likelihood function. The likelihood function is calculated using the forward-backward algorithm [27], [28], which is then recursively maximized using the Expectation Maximization algorithm (EM algorithm) to obtain the formula for estimating model parameters with primary references [25], [26], [24] . Due to the estimation and convergence of the parameters of the covariance matrix, it has its own complexity and analysis (multivariate analysis), so it will be published separately. This complexity can be seen in several studies related to covariance matrices [31], [32], [33], [34], [35], [36]. Therefore, this research will discuss recursive parameter estimation and its convergence for multivariate normal hidden Markov inhomogeneous models.

Multivariate Normal Hidden Markov Inhomogeneous Model

The normal hidden Markov inhomogeneous multivariate model (MNHMM-I) is a discrete-time model consisting of a pair of stochastic processes $\{X_t, Y_t\}_{t\in\mathbb{N}}$ [1], where $\{X_t\}_{t\in\mathbb{N}}$ is the cause the assumed event is not observed and forms a Markov chain that is inhomogeneous at one time before and ergodic (irreducible, positive recurrent and aperiodic) [2] with state space $S_X = \{1,2,\cdots,m\}$, whereas $\{Y_t\}_{t\in\mathbb{N}}$ is the observation process which depends only on $\{X_t\}_{t\in\mathbb{N}}$. Then the random variable Y_t is known X_t assumed to be a normal multivariate distribution, for every $t \in \mathbb{N}$ [22], [23], [37].

To simplify writing, the following 10 points are symbolized for further writing:

1. Observation process: $Y = \{Y_t\}_{t=1}^{T}$;
2. Markov chain: $X = \{X_t\}_{t=1}^{T}$;
3. MNHMM-I: $Z = \{X_t, Y_t\}_{t=1}^{T}$;
4. data of process $\{Y_t\}_{t=1}^{T}$: $y = (y_1, y_2, \ldots, y_T)$, (called incomplete data);
5. *state* of process $\{X_t\}_{t=1}^{T}$: $x = (i_1, i_2, \ldots, i_T)$;
6. data of process $\{X_t, Y_t\}_{t=1}^{T}$: $z = (i_1, y_1, \ldots, i_T, y_T) = (x, y)$, (called data complete);
7. the probability mass function of Z: $P(Z = z|\phi) = p(z;\phi) = p(x, y|\phi)$;
8. the probability of Y: $L_T(\phi) = P(Y = y|\phi) = p(y|\phi)$;
9. likelihood function of complete data: $L_T^c(\phi) = p(z|\phi) = p(x, y|\phi)$;
10. the probability mass function of $X = x$ under condition $Y = y$: $P(X = x|Y = y, \phi) = p(x|y, \phi) = \frac{p(z|\phi)}{p(y|\phi)} = \frac{p(x,y|\phi)}{p(y|\phi)} = \frac{L_T^c(\phi)}{L_T(\phi)}$.

Referring to [25][38][22][23][24][26], in this study the form of MNHMM-I is:

$$Y_t - \mu_{s_t^*} = \varphi\left(Y_{t-1} - \mu_{s_{t-1}^*}\right) + \varepsilon_t \tag{1}$$

where

1. $\epsilon_t \sim multivariate\ normal(0, \Sigma)$ and i.i.d for each t;
2. $\{Y_t\}$ is the observed process and scalar (longitudinal data);
3. $\{S_t^*\}$ is Markov chain with state space $S_t^* = \{1,2\}$, with transition matrix $\Gamma^* = \begin{pmatrix} \gamma_{11}^* & \gamma_{21}^* \\ \gamma_{12}^* & \gamma_{22}^* \end{pmatrix}$, $\gamma_{ij}^* = P(S_t^* = j \mid S_{t-1}^* = i)$;
4. The MNHMM-I parameters are $\mu_1, \mu_2, \Sigma, \varphi \in R$.

In this case Y_t does not only depend on S_t^* but also depends on S_{t-1}^* so that in order to comply with Markov properties it is necessary to define a new process S_t where

$$S_t = 1 \text{ if } S_t^* = 1 \text{ and } S_{t-1}^* = 1 \tag{2}$$
$$S_t = 2 \text{ if } S_t^* = 1 \text{ and } S_{t-1}^* = 2 \tag{3}$$
$$S_t = 3 \text{ if } S_t^* = 2 \text{ and } S_{t-1}^* = 1 \tag{4}$$
$$S_t = 4 \text{ if } S_t^* = 2 \text{ and } S_{t-1}^* = 2, \tag{5}$$

so that the transition matrix parameters take the form

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} \\ \gamma_{21} & \gamma_{22} & \gamma_{22} & \gamma_{24} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34} \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} \end{pmatrix}.$$

Lemma 1 (Condition Sufficiently Estimating Parameters of Transition Matrix Γ on MNHMM-I)
If the state of transition matrix is defined in equations $(2)-(5)$, then in order to comply with the Markov property, it is sufficient to estimate the parameters γ_{11}^* and γ_{22}^* where the transition matrix Γ has the following form

$$\Gamma = \begin{pmatrix} \gamma_{11}^* & 0 & \gamma_{12}^* & 0 \\ \gamma_{11}^* & 0 & \gamma_{12}^* & 0 \\ 0 & \gamma_{21}^* & 0 & \gamma_{22}^* \\ 0 & \gamma_{21}^* & 0 & \gamma_{22}^* \end{pmatrix}. \tag{6}$$

Proof, see Appendix 1.
Furthermore,

$$\begin{aligned} F_{Y_t}(y_t) &= P(Y_t \leq y_t) \\ &= P\left(\mu_{s_t^*} + \varphi\left(Y_{t-1} - \mu_{s_t^*}\right) + \varepsilon_t \leq y_t\right) \\ &= P\left(\varepsilon_t \leq y_t - \mu_{s_t^*} - \varphi\left(Y_{t-1} - \mu_{s_{t-1}^*}\right)\right) \\ &= \int_0^{y_t-\mu_{s_t^*}-\varphi\left(Y_{t-1}-\mu_{s_{t-1}^*}\right)} \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\varepsilon_t'{}'\Sigma^{-1}\varepsilon_t}{2}}\, d\varepsilon_t. \end{aligned} \tag{7}$$

So equation (7) can be written as follows

$$F_{Y_t}(y_t) = \int_0^{v} \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\varepsilon_t'{}'\Sigma^{-1}\varepsilon_t}{2}}\, d\varepsilon_t.$$

Consequently,

$$\begin{aligned} f_{Y_t}(y_t) &= \frac{\partial}{\partial y_t} F_{Y_t}(y_t) \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\mathrm{v_{ti}}'\Sigma^{-1}\mathrm{v_{tj}}}{2}} \frac{\partial v}{\partial y_t} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\mathrm{v_{ti}}'\Sigma^{-1}\mathrm{v_{tj}}}{2}} 1 \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\mathrm{v_{ti}}'\Sigma^{-1}\mathrm{v_{ti}}}{2}}. \end{aligned}$$

So that the conditional probability density function in MNHMM-I can be written as equation (8)

$$f_{Y_t}(y_t) = \begin{pmatrix} f(y_t|S_t = 1, \mathcal{Y}_{t-1}; \phi) \\ f(y_t|S_t = 2, \mathcal{Y}_{t-1}; \phi) \\ f(y_t|S_t = 3, \mathcal{Y}_{t-1}; \phi) \\ f(y_t|S_t = 4, \mathcal{Y}_{t-1}; \phi) \end{pmatrix} = \begin{pmatrix} \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\mathrm{v_{t1}}'\Sigma^{-1}\mathrm{v_{t1}}}{2}} \\ \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\mathrm{v_{t1}}'\Sigma^{-1}\mathrm{v_{t2}}}{2}} \\ \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\mathrm{v_{t2}}'\Sigma^{-1}\mathrm{v_{t1}}}{2}} \\ \frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\, e^{-\frac{\mathrm{v_{t2}}'\Sigma^{-1}\mathrm{v_{t2}}}{2}} \end{pmatrix}, \tag{8}$$

where $v_{ti} = y_t - \mu_{s_t^*} - \varphi\left(Y_{t-1} - \mu_{s_{t-1}^*}\right)$, that is

$$v_{t1} = \begin{pmatrix} v_{1t1} \\ v_{2t1} \\ \vdots \\ v_{pt1} \end{pmatrix} = \begin{pmatrix} y_{1t} - \mu_1 - \varphi(Y_{1t-1} - \mu_1) \\ y_{2t} - \mu_1 - \varphi(Y_{2t-1} - \mu_1) \\ \vdots \\ y_{pt} - \mu_1 - \varphi\left(Y_{pt-1} - \mu_1\right) \end{pmatrix}; \quad v_{t2} = \begin{pmatrix} v_{1t2} \\ v_{2t2} \\ \vdots \\ v_{pt2} \end{pmatrix} = \begin{pmatrix} y_{1t} - \mu_1 - \varphi(Y_{1t-1} - \mu_2) \\ y_{2t} - \mu_1 - \varphi(Y_{2t-1} - \mu_2) \\ \vdots \\ y_{pt} - \mu_1 - \varphi\left(Y_{pt-1} - \mu_2\right) \end{pmatrix};$$

$$v_{t3} = \begin{pmatrix} v_{1t3} \\ v_{2t3} \\ \vdots \\ v_{pt3} \end{pmatrix} = \begin{pmatrix} y_{1t} - \mu_2 - \varphi(Y_{1t-1} - \mu_1) \\ y_{2t} - \mu_2 - \varphi(Y_{2t-1} - \mu_1) \\ \vdots \\ y_{pt} - \mu_2 - \varphi\left(Y_{pt-1} - \mu_1\right) \end{pmatrix}; \quad v_{t4} = \begin{pmatrix} v_{1t4} \\ v_{2t4} \\ \vdots \\ v_{pt4} \end{pmatrix} = \begin{pmatrix} y_{1t} - \mu_2 - \varphi(Y_{1t-1} - \mu_2) \\ y_{2t} - \mu_2 - \varphi(Y_{2t-1} - \mu_2) \\ \vdots \\ y_{pt} - \mu_2 - \varphi\left(Y_{pt-1} - \mu_2\right) \end{pmatrix}.$$

Here's a basic MNHMM-I resume:

1. $y_1 = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{p1} \end{pmatrix}, y_2 = \begin{pmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{p2} \end{pmatrix}, \ldots, y_T = \begin{pmatrix} y_{1T} \\ y_{2T} \\ \vdots \\ y_{pT} \end{pmatrix}$ is the longitudinal data to be modeled, where p is the number of cross data and T is the amount of time series data. The MNHMM-I parameters are $\mathrm{M}, \Sigma, \Gamma, \varphi, \delta$ [22], with

   $\mathrm{M} = (\mu_1, \mu_2)$, with $\mu_1 = \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{pmatrix}; \mu_2 = \begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{pmatrix}$.

   $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$,

   $\Gamma = \begin{pmatrix} \gamma_{11}^* & 0 & \gamma_{12}^* & 0 \\ \gamma_{11}^* & 0 & \gamma_{12}^* & 0 \\ 0 & \gamma_{21}^* & 0 & \gamma_{22}^* \\ 0 & \gamma_{21}^* & 0 & \gamma_{22}^* \end{pmatrix}$, with $\Gamma^* = \begin{pmatrix} \gamma_{11}^* & \gamma_{21}^* \\ \gamma_{12}^* & \gamma_{22}^* \end{pmatrix}$,

   $\varphi \in \mathbb{R}$, and $\delta \in \mathbb{R}^{m\times 1}$.

2. The conditional probability Y_t if known $X_t = i$ ($t \in \mathbb{N}$) is a normal multivariate random variable with the mean parameter $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. For each $y \in \mathbb{R}^p$, the probability matrix from the observation process $\Pi = [\pi_{yi}]$ in [39] [22] is

   $$\pi_{y_t i_t} = P(Y_t = y_t | X_t = i_t, \phi) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{v_{ti}{}'\Sigma^{-1}v_{ti}}{2}},$$

   for $i_t = 1,2,\ldots,4$ and $t = 1,2,\ldots,T$.

   $$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{v_{ti}{}'\Sigma^{-1}v_{ti}}{2}} \, dy_1 \, dy_2 \ldots dy_p = 1.$$

3. Transition probability matrix $\Gamma = [\gamma_{ij}]$, where Γ matrix is of size $m \times m$ and $i, j \in S_X$, satisfies:
   - $\gamma_{ij} = P(X_t = j | X_{t-1} = i) = P(X_2 = j | X_1 = i)$,
   - $\gamma_{ij} \geq 0$,
   - $\sum_{j=1}^{m} \gamma_{ij} = 1$, for each $i = 1,2,\ldots,m$.

4. $\varphi \in R$, is a scalar-valued parameter contained in equation (1).

5. Let $\delta = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_m \end{pmatrix}$ be the initial state distribution and the long-term proportion δ is usually called the stationary distribution. Based on [2] the Markov chain $\{X_t\}_{t\in\mathbb{N}}$ which is assumed to be ergodic, the stationary distribution δ can be obtained uniquely, that is, it satisfies

   $$\Gamma\delta = \delta. \tag{9}$$

   with

   $$\delta_i = P(X_1 = i), \qquad \forall i \in S_X$$
   $$\sum_{i=1}^{m} \delta_i = 1.$$

   For each $t \in \mathbb{N}$ and $y \in \mathbb{R}^p$, the marginal distribution function of Y_t, ie

   $$P(Y_t = y) = \sum_{i=1}^{m} P(Y_t = y | X_t = i) P(X_t = i) = \sum_{i=1}^{m} \delta_i \pi_{yi}\,.$$

Based on the discussion above, MNHMM-I $\{X_t, Y_t\}_{t\in\mathbb{N}}$ is characterized by $\delta, \mathrm{M}, \Sigma, \Gamma, \varphi$. The most important thing in MNHMM-I is to estimate the parameters of this model by maximizing its likelihood function. Furthermore, equation (9) informs that δ will be obtained when Γ is obtained (eigen vector) so that δ is not a parameter that must be estimated. In addition, for the transition matrix based on Lemma 1, it is sufficient to estimate the diagonal elements γ_{11}^* and γ_{22}^* which are symbolized by $\hat{\Gamma}$. Due to the estimation and convergence of the parameters of the covariance matrix, it has its own complexity and analysis (multivariate analysis), so it will be published separately. The parameter estimation in this study is limited to $\boldsymbol{\phi} = (\boldsymbol{M}, \hat{\boldsymbol{\Gamma}}, \boldsymbol{\varphi})$. In order to estimate the parameters and analyze the convergence of this model, it is necessary to clarify the parameter space and its assumptions, the likelihood function and the parameter estimation process which will be discussed in the following discussion.

Parameter Estimation

Suppose the number of observation times T, the number of cross-sectional data p, the number of states m, and the observation sequence $y = (y_1, y_2, \dots, y_T)$ are defined. Given that any $\varepsilon > 0$ is small enough to approach 0, define the MNHMM-I parameter space: $\boldsymbol{\Phi} = \left\{\phi = \left(M, \hat{\Gamma}, \varphi\right) : M \in \left[\varepsilon, \frac{1}{\varepsilon}\right]^{p\times m}, \hat{\Gamma} \in [0,1]^{1\times 2}, \varphi \in \mathbb{R} \right\}$. For each $\phi \in \boldsymbol{\Phi}$, $\delta(\phi) = \left(\delta_i(\phi)\right)$, $M(\phi) = \left(\mu_{ij}(\phi)\right)$, $\Sigma(\phi) = \left(\sigma_{ij}(\phi)\right)$, $\Gamma(\phi) = \left(\gamma_{ij}(\phi)\right)$, $\varphi(\phi) = \varphi(\phi)$ [22][37], the following five continuity points are assumed to be satisfied.

1. $\delta_i : \boldsymbol{\Phi} \to \mathbb{R}$, with $\delta_i(\phi) = \delta_i$ is a function that is continuous in $\boldsymbol{\Phi}$, $\forall\ i \in S_X$.
2. $M_i : \boldsymbol{\Phi} \to \mathbb{R}$, with $M_i(\phi) = M_i$ is a function that is continuous in $\boldsymbol{\Phi}$, $\forall\ i \in S_X$,
3. $\Sigma_i : \boldsymbol{\Phi} \to \mathbb{R}$, with $\Sigma_i(\phi) = \Sigma_i$ is a function that is continuous in $\boldsymbol{\Phi}$, $\forall\ i \in S_X$,
4. $\gamma_{ij} : \boldsymbol{\Phi} \to \mathbb{R}$, with $\gamma_{ij}(\phi) = \gamma_{ij}$ is a function that is continuous in $\boldsymbol{\Phi}$, $\forall\ i,j \in S_X$,
5. $\varphi : \boldsymbol{\Phi} \to \mathbb{R}$, with $\varphi(\phi) = \varphi$ is a function that is continuous in $\boldsymbol{\Phi}$.

The likelihood function of the Y observation process is defined in equation (10):

$$\begin{aligned} L_T(\phi) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_T = y_T|\phi) \\ &= p(y_1, y_2, \dots, y_T|\phi) \\ &= p(y|\phi) \\ &= \sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} \left(\pi_{y_1 i_1}\pi_{y_2 i_2} \dots \pi_{y_T i_T}\right) \times \left(\delta_{i_1}\gamma_{i_1 i_2}\gamma_{i_2 i_3} \dots \gamma_{i_{T-1} i_T}\right) \\ &= \sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} \delta_{i_1}\pi_{y_1 i_1} \prod_{t=2}^{T} \gamma_{i_{t-1} i_t}\pi_{y_t i_t}. \end{aligned} \tag{10}$$

In the previous discussion, it has been explained that the main problem in MNHMM-I is finding the parameter $\phi^* \in \boldsymbol{\Phi}$ which maximizes the likelihood function $L_T(\phi)$. For large enough T observation data, calculating the $L_T(\phi)$ function takes quite a long time. To deal with this problem, the forward-backward algorithm is used. The working principle of the forward-backward algorithm is to calculate recursively, thus speeding up computation time. This algorithm is divided into two, namely the forward algorithm and the backward algorithm. Baum *et al*. [40] define forward probability as follows:

$$\alpha_t(i|\phi) = P(Y_1 = y_1, Y_2 = y_2, \dots, Y_t = y_t, X_t = i|\phi),$$

and backward probability:

$$\beta_t(i|\phi) = P(Y_{t+1} = y_{t+1}, \dots, Y_T = y_T|X_t = i, \phi),$$

for $t = 1,2, \dots T$, and $i \in S_X$.

The formulation for forward probability and backward probability recursively [27][28] which is commonly called the forward algorithm is as follows:

$$\alpha_1(i|\phi) = \pi_{y_1 i}\delta_i,$$

$$\alpha_{t+1}(j|\phi) = \left(\sum_{i\in S_X} \alpha_t(i|\phi)\gamma_{ij}\right)\pi_{y_{t+1} j},$$

and backward algorithm

$$\beta_T(j|\phi) = 1,$$

$$\beta_t(j|\phi) = \sum_{i\in S_X} \beta_{t+1}(i|\phi)\pi_{y_{t+1} i}\,\gamma_{ji},$$

for $t = 1, \dots, T-1$, and $i,j \in S_X$.

Then [27][28] uses forward and backward algorithms to calculate the likelihood function $L_T(\phi)$, which is commonly called the forward-backward algorithm and obtains:

$$L_T(\phi) = \sum_{i \in S_X} \alpha_t(i|\phi)\beta_t(i|\phi),$$

for any $t = 1,2,. . .,T$, and $i \in S_X$.
The likelihood function of the complete data can be seen in equation (11)

$$L_T^c(\phi) = \delta_{i_1}\pi_{y_1 i_1}\prod_{t=2}^{T}\gamma_{i_{t-1}i_t}\pi_{y_t i_t}. \tag{11}$$

Based on equations (10) and (11), the relationship between the likelihood function of incomplete data and complete data is as follows:

$$L_T(\phi) = p(y|\phi) = \sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} \delta_{i_1}\pi_{y_1 i_1}\prod_{t=2}^{T}\gamma_{i_{t-1}i_t}\pi_{y_t i_t} = \sum_{x} p(y, x|\phi) = \sum_{x} L_T^c(\phi).$$

To obtain $\phi^* \in \mathbf{\Phi}$ which maximizes $L_T(\phi)$ is a difficult problem. $\phi^* \in \mathbf{\Phi}$ which maximizes $\ln L_T(\phi)$ will also maximize $L_T(\phi)$. For $\phi \in \mathbf{\Phi}$, holds

$$\ln p(x|y,\phi) = \ln \frac{L_T^c(\phi)}{L_T(\phi)} \quad \Rightarrow \ln L_T(\phi) = \ln L_T^c(\phi) - \ln p(x|y,\phi).$$

Note that for any $\hat{\phi} \in \mathbf{\Phi}$ also holds

$$E_{\hat{\phi}}(\ln L_T(\phi)\,|y) = E_{\hat{\phi}}(\ln L_T^c(\phi)\,|y) - E_{\hat{\phi}}(\ln p(x|y,\phi)|y), \tag{12}$$

and

$$\begin{aligned} E_{\hat{\phi}}(\ln L_T(\phi)\,|y) &= \sum_{x} \ln L_T(\phi)\, p(x|y,\hat{\phi}) = \sum_{x} \ln p(y|\phi)\, p(x|y,\hat{\phi}) = \sum_{x} \ln p(y|\phi) \frac{p(x,y|\hat{\phi})}{p(y|\hat{\phi})} \\ &= \frac{\ln p(y|\phi)}{p(y|\hat{\phi})}\sum_{x} p(x,y|\hat{\phi}) = \frac{\ln p(y|\phi)}{p(y|\hat{\phi})} p(y|\hat{\phi}) = \ln p(y|\phi) = \ln L_T(\phi), \end{aligned} \tag{13}$$

so that based on equations (12) and (13) is obtained

$$\ln L_T(\phi) = Q(\phi|\hat{\phi}) - H(\phi|\hat{\phi}), \tag{14}$$

with $Q(\phi|\hat{\phi}) = E_{\hat{\phi}}(\ln L_T^c(\phi)|y)$ dan $H(\phi|\hat{\phi}) = E_{\hat{\phi}}(\ln p(x|y,\phi)\,|y)$.
To get ϕ^* which maximizes $\ln L_T(\phi)$, the first step is to solve the equation $\partial_\phi(\ln L_T(\phi)) = 0$ to get a stationary point. By following the pattern of equation (12), it will be obtained directly

$$\partial_\phi(\ln L_T(\phi)) = E_{\hat{\phi}}\big(\partial_\phi(\ln L_T(\phi))|y\big) \tag{15}$$

Corollary of equations (14) and (15), then

$$\partial_\phi(\ln L_T(\phi)) = E_{\hat{\phi}}\big(\partial_\phi(\ln L_T(\phi))|y\big) = E_{\hat{\phi}}\big(\partial_\phi \ln L_T^c(\phi)\,|y\big) - E_{\hat{\phi}}\big(\partial_\phi \ln p(x|y,\phi)|y\big). \tag{16}$$

Define[29]

$$D^{10}Q(\phi|\hat{\phi}) = E_{\hat{\phi}}\left(\frac{\partial}{\partial\phi}\ln L_T^c(\phi)|y\right), \tag{17}$$

and

$$D^{10}H(\phi|\hat{\phi}) = E_{\hat{\phi}}\left(\frac{\partial}{\partial\phi}\ln p(x|y,\phi)\,|y\right), \tag{18}$$

so by substituting equations (17) and (18) into equation (16), will obtained

$$\partial_\phi(\ln L_T(\phi)) = D^{10}Q(\phi|\hat{\phi}) - D^{10}H(\phi|\hat{\phi}). \tag{19}$$

Lemma 2 (see [29]**)**
Suppose $D^{10}H(\phi|\hat{\phi}) = E_{\hat{\phi}}\left(\frac{\partial}{\partial\phi}\ln p(x|y,\phi)\,|y\right)$, then $D^{10}H(\hat{\phi}|\hat{\phi}) = 0$, for every $\hat{\phi} \in \mathbf{\Phi}$.
Proof, see Appendix 2.

Lemma 3 (see [29]**)**
Suppose $H(\phi|\hat{\phi}) = E_{\hat{\phi}}(\ln p(x|y,\phi)\,|y)$, then $H(\phi|\hat{\phi}) \le H(\hat{\phi}|\hat{\phi})$, for every $\phi, \hat{\phi} \in \mathbf{\Phi}$.
Proof, see Appendix 3.

To get the stationary point of $\ln L_T(\phi)$, based on equation (14), Lemma 2 and Lemma 3, it is enough to find the stationary point from $Q(\phi|\hat{\phi})$ to $\phi \in \mathbf{\Phi}$. However, $D^{10}Q(\phi|\hat{\phi})$ is a non-linear function and is difficult to solve explicitly with respect to the parameter $\phi \in \mathbf{\Phi}$, as a result to obtain a stationary point from $Q(\phi|\hat{\phi})$ to $\phi \in \mathbf{\Phi}$ is an analytical problem difficult, so this problem is solved using the Expectation Maximization algorithm.

The Expectation Maximization (EM) algorithm is a recursive algorithm which consists of two steps in each iteration, namely step E and step M. The steps in the EM algorithm are, take $\phi^{(k)}$ as an estimate for the MNHMM-I parameter obtained at k^{th} iteration. In the $(k+1)^{\text{th}}$ iteration, step E and step M are defined as follows:

1. Give error tolerance, maximum iteration and initial value of parameter $\phi^{(k)}$ for $k = 0$,
2. **E Step** – Given $\phi^{(k)}$, compute

$$Q(\phi;\phi^{(k)}) = E_{\phi^{(k)}}(\ln L_T^c(\phi)|Y = y)$$

$$= \sum_{i\in S_X} \frac{\alpha_1(i|\phi^{(k)})\beta_1(i|\phi^{(k)})}{\sum_{l\in S_X}\alpha_t(l|\phi^{(k)})\beta_t(l|\phi^{(k)})}\ln \delta_i(\phi) + \sum_{i\in S_X}\frac{\sum_{t=1}^{T}\alpha_t(i|\phi^{(k)})\beta_t(i|\phi^{(k)})}{\sum_{l\in S_X}\alpha_t(l|\phi^{(k)})\beta_t(l|\phi^{(k)})}\ln\left(\frac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\,e^{-\frac{v_{ti}'\Sigma^{-1}v_{ti}}{2}}\right)$$

$$+ \sum_{i\in S_X}\sum_{j\in S_X}\frac{\sum_{t=1}^{T-1}\gamma_{ij}(\phi^{(k)})\alpha_t(i|\phi^{(k)})P(Y_{t+1} = y_{t+1}|X_{t+1} = j,\phi^{(k)})\beta_{t+1}(j|\phi^{(k)})}{\sum_{l\in S_X}\alpha_t(l|\phi^{(k)})\beta_t(l|\phi^{(k)})}\ln\gamma_{ij}(\phi).$$

3. **M Step** – Finding the $\phi^{(k+1)}$ that maximizes $Q(\phi;\phi^{(k)})$, so that

$$Q(\phi^{(k+1)}|\phi^{(k)}) \geq Q(\phi|\phi^{(k)}),$$

For every $\phi \in \mathbf{\Phi}$,

4. Replace k with $k+1$ and repeat step 2 to step 4 until $|\ln L_T(\phi^{(k+1)}) - \ln L_T(\phi^{(k)})|$ is less than the given error (in other words $\{\ln L_T(\phi^{(k)})\}$ converges) or the maximum iteration is reached.

In M Step, the estimation of the average parameter MNHMM-I is obtained by method $\frac{\partial Q(\phi|\phi^{(k)})}{\partial \mu_{uw}(\phi)} = 0$, so that for $w = 1$ obtained

$$\mu_{uw} = -\frac{\sum_{t=1}^{T}A_{1tw} + \sum_{t=1}^{T}A_{2tw} + \sum_{t=1}^{T}B_{1tw+1} + \sum_{t=1}^{T}B_{2tw+1} + \sum_{t=1}^{T}C_{1tw+2} + \sum_{t=1}^{T}C_{2tw+2}}{2s_{uu}\left((1-2\varphi+\varphi^2)\sum_{t=1}^{T}\alpha_t(w|\phi^{(k)})\beta_t(w|\phi^{(k)}) + \sum_{t=1}^{T}\alpha_t(w+1|\phi^{(k)})\beta_t(w+1|\phi^{(k)}) + \varphi^2\sum_{t=1}^{T}\alpha_t(w+2|\phi^{(k)})\beta_t(w+2|\phi^{(k)})\right)},$$

with

$$A_{1tw} = \alpha_t(w|\phi^{(k)})\beta_t(w|\phi^{(k)})\left(\sum_{\substack{i=1\\i\neq u}}^{p} s_{iu}\, v_{itw}W^*(w,w) + \sum_{\substack{i=1\\i\neq u}}^{p} s_{ui}\, v_{itw}W^*(w,w)\right);$$

$$A_{2tw} = \alpha_t(w|\phi^{(k)})\beta_t(w|\phi^{(k)})(2s_{uu}(-1+\varphi)(y_{ut} - \varphi\, Y_{ut-1}));$$

$$B_{1tw+1} = \alpha_t(w+1|\phi^{(k)})\beta_t(w+1|\phi^{(k)})\left(\sum_{\substack{i=1\\i\neq u}}^{p} s_{iu}\, v_{itw+1}W^*(w+1,w) + \sum_{\substack{i=1\\i\neq u}}^{p} s_{ui}\, v_{itw+1}W^*(w+1,w)\right);$$

$$B_{2tw+1} = \alpha_t(w+1|\phi^{(k)})\beta_t(w+1|\phi^{(k)})(2s_{uu}(-1)(y_{ut} - \varphi Y_{ut-1} + \varphi\cancel{\mu_{uw}}));$$

$$C_{1tw+2} = \alpha_t(w+2|\phi^{(k)})\beta_t(w+2|\phi^{(k)})\left(\sum_{\substack{i=1\\i\neq u}}^{p} s_{iu}\, v_{itw+2}W^*(w+2,w) + \sum_{\substack{i=1\\i\neq u}}^{p} s_{ui}\, v_{itw+2}W^*(w+2,w)\right);$$

$$C_{2tw+2} = \alpha_t(w+2|\phi^{(k)})\beta_t(w+2|\phi^{(k)})(2s_{uu}\varphi(y_{ut} - \varphi Y_{ut-1} - \cancel{\mu_{uw}})).$$

For $w = 2$

$$\mu_{uw} = -\frac{\sum_{t=1}^{T}A_{1tw} + \sum_{t=1}^{T}A_{2tw} + \sum_{t=1}^{T}B_{1tw+1} + \sum_{t=1}^{T}B_{2tw+1} + \sum_{t=1}^{T}C_{1tw+2} + \sum_{t=1}^{T}C_{2tw+2}}{(2s_{uu})(\varphi^2\sum_{t=1}^{T}\alpha_t(w|\phi^{(k)})\beta_t(w|\phi^{(k)}) + \sum_{t=1}^{T}\alpha_t(w+1|\phi^{(k)})\beta_t(w+1|\phi^{(k)}) + (1-2\varphi+\varphi^2)\sum_{t=1}^{T}\alpha_t(w+2|\phi^{(k)})\beta_t(w+2|\phi^{(k)}))},$$

with

$$A_{1tw} = \alpha_t(w|\phi^{(k)})\beta_t(w|\phi^{(k)})\left(\sum_{\substack{i=1\\i\neq u}}^{p} s_{iu}\, v_{itw}W^*(w,w) + \sum_{\substack{i=1\\i\neq u}}^{p} s_{ui}\, v_{itw}W^*(w,w)\right);$$

$$A_{2tw} = \alpha_t(w|\phi^{(k)})\beta_t(w|\phi^{(k)})(2s_{uu}(\varphi)(y_{ut} - \varphi\, Y_{ut-1} - \cancel{\mu_{uw}}))$$

$$B_{1tw+1} = \alpha_t\left(w+1|\phi^{(k)}\right)\beta_t\left(w+1|\phi^{(k)}\right)\left(\sum_{\substack{i=1\\ i\neq u}}^{p} s_{iu}\, v_{itw+1}W^*(w+1,w) + \sum_{\substack{i=1\\ i\neq u}}^{p} s_{ui}\, v_{itw+1}W^*(w+1,w)\right);$$

$$B_{2tw+1} = \alpha_t\left(w+1|\phi^{(k)}\right)\beta_t\left(w+1|\phi^{(k)}\right)\left(2s_{uu}(-1)(y_{ut} - \varphi Y_{ut-1} + \varphi\mu_{uw})\right);$$

$$C_{1tw+2} = \alpha_t\left(w+2|\phi^{(k)}\right)\beta_t\left(w+2|\phi^{(k)}\right)\left(\sum_{\substack{i=1\\ i\neq u}}^{p} s_{iu}\, v_{itw+2}W^*(w+2,w) + \sum_{\substack{i=1\\ i\neq u}}^{p} s_{ui}\, v_{itw+2}W^*(w+2,w)\right);$$

$$C_{2tw+2} = \alpha_t\left(w+2|\phi^{(k)}\right)\beta_t\left(w+2|\phi^{(k)}\right)\left(2s_{uu}(-1+\varphi)(y_{ut} - \varphi Y_{ut-1})\right);$$

where $u = 1,2,\dots,p$, and $\Sigma^{-1} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$.

To obtain the parameter $\gamma_{uv}^*\left(\phi^{(k+1)}\right)$ which maximizes $Q\left(\phi|\phi^{(k)}\right)$ toward $\phi \in \mathbf{\Phi}$, used the Lagrange multiplier method with the constraint $\sum_{j=1}^{m}\gamma_{ij}^*(\phi) = 1$, for $i = 1,2$. In order for the parameter Γ to still fulfill the inhomogeneous Markov property one time before, then based on Lemma 1 it is enough to update γ_{11}^* and γ_{22}^*. Suppose $G\left(\phi|\phi^{(k)}\right) = Q\left(\phi|\phi^{(k)}\right) - \sum_{i=1}^{m}\theta_i\left(\sum_{\forall j}\gamma_{ij}^*(\phi) - 1\right)$, for any $\theta_i \in \mathbb{R}$. Then $\frac{\partial G(\phi|\phi^{(k)})}{\partial\gamma_{uu}^*(\phi)} = 0$ (for $u = 1,2$) implies

$$\gamma_{uu}^*\left(\phi^{(k+1)}\right) = \frac{\sum_{t=1}^{T-1}\gamma_{uu}^*\left(\phi^{(k)}\right)\alpha_t\left(u|\phi^{(k)}\right)P\left(Y_{t+1} = y_{t+1}|X_{t+1} = u, \phi^{(k)}\right)\beta_{t+1}\left(u|\phi^{(k)}\right)}{\sum_{t=1}^{T-1}\alpha_t\left(u|\phi^{(k)}\right)\beta_t\left(u|\phi^{(k)}\right)}.$$

As for γ_{uv}^* (for $u, v = 1,2$ and $u \neq v$), it is updated using the probability property for the transition matrix, namely

$$\gamma_{uv}^* = 1 - \gamma_{uu}^*$$

So that the MNHMM-I transition matrix is obtained by means of $\Gamma = \begin{pmatrix} \gamma_{11}^* & 0 & \gamma_{12}^* & 0 \\ \gamma_{11}^* & 0 & \gamma_{12}^* & 0 \\ 0 & \gamma_{21}^* & 0 & \gamma_{22}^* \\ 0 & \gamma_{21}^* & 0 & \gamma_{22}^* \end{pmatrix}$.

The estimation of the φ parameter is obtained by method $\frac{\partial Q(\phi|\phi^{(k)})}{\partial\varphi(\phi)} = 0$, so that obtained

$$\varphi = \frac{\sum_{t=1}^{T}A_{1t} + \sum_{t=1}^{T}A_{2t} + \sum_{t=1}^{T}A_{3t} + \sum_{t=1}^{T}A_{4t}}{2\left(\sum_{t=1}^{T}B_{1t} + \sum_{t=1}^{T}B_{2t} + \sum_{t=1}^{T}B_{3t} + \sum_{t=1}^{T}B_{4t}\right)},$$

with

$$A_{1t} = \alpha_t\left(1|\phi^{(k)}\right)\beta_t\left(1|\phi^{(k)}\right)\left(\sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{it-1} - \mu_{i1})(y_{jt} - \mu_{j1}) + \sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{jt-1} - \mu_{j1})(y_{it} - \mu_{i1})\right)$$

$$A_{2t} = \alpha_t\left(2|\phi^{(k)}\right)\beta_t\left(2|\phi^{(k)}\right)\left(\sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{it-1} - \mu_{i2})(y_{jt} - \mu_{j1}) + \sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{jt-1} - \mu_{j2})(y_{it} - \mu_{i1})\right)$$

$$A_{3t} = \alpha_t\left(3|\phi^{(k)}\right)\beta_t\left(3|\phi^{(k)}\right)\left(\sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{it-1} - \mu_{i1})(y_{jt} - \mu_{j2}) + \sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{jt-1} - \mu_{j1})(y_{it} - \mu_{i2})\right)$$

$$A_{4t} = \alpha_t\left(4|\phi^{(k)}\right)\beta_t\left(4|\phi^{(k)}\right)\left(\sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{it-1} - \mu_{i2})(y_{jt} - \mu_{j2}) + \sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{jt-1} - \mu_{j2})(y_{it} - \mu_{i2})\right)$$

$$B_{1t} = \alpha_t\left(1|\phi^{(k)}\right)\beta_t\left(1|\phi^{(k)}\right)\left(\sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{jt-1} - \mu_{j1})(Y_{it-1} - \mu_{i1})\right)$$

$$B_{2t} = \alpha_t\left(2|\phi^{(k)}\right)\beta_t\left(2|\phi^{(k)}\right)\left(\sum_{j=1}^{p}\sum_{i=1}^{p}s_{ij}(Y_{jt-1} - \mu_{j2})(Y_{it-1} - \mu_{i2})\right)$$

$$B_{3t} = \alpha_t\left(3|\phi^{(k)}\right)\beta_t\left(3|\phi^{(k)}\right)\left(\sum_{j=1}^{p}\sum_{i=1}^{p} s_{ij}\left(Y_{jt-1}-\mu_{j1}\right)\left(Y_{it-1}-\mu_{i1}\right)\right)$$
$$B_{4t} = \alpha_t\left(4|\phi^{(k)}\right)\beta_t\left(4|\phi^{(k)}\right)\left(\sum_{j=1}^{p}\sum_{i=1}^{p} s_{ij}\left(Y_{jt-1}-\mu_{j2}\right)\left(Y_{it-1}-\mu_{i2}\right)\right).$$

Parameter Estimator Sequence Convergence MNHMM-I

Furthermore, will proved that the sequence $\{\ln L_T(\phi^{(k)})\}$ converges to $\ln L_T(\phi^*)$ using the EM algorithm, where $\phi^{(k)}$ is the MNHMM-I parameter estimator in the k^{th} iteration and ϕ^* is a stationary point of the function $\ln L_T(\phi)$. This will be discussed in Wu's Theorem (Theorem 2). Before discussing Wu's Theorem, the following symbols are exemplified to simplify writing:

1. Let k denotes the iteration of the EM algorithm, namely $k \in \{0,1,2,3,\dots\}$;
2. Let $\boldsymbol{\Psi} = \{\phi \in \text{int}\,\boldsymbol{\Phi} :\ \phi \text{ stationary point of } \ln L_T(\phi)\}$;
3. Let T T be the set-valued function defined at $\boldsymbol{\Phi}$ and with the range $\boldsymbol{\Phi}$ such that for any $\hat{\phi} \in \boldsymbol{\Phi}$ satisfies
$$T(\hat{\phi}) = \{\varphi' \in \boldsymbol{\Phi} : Q(\varphi'|\hat{\phi}) \geq Q(\varphi|\hat{\phi}) \text{ for every } \varphi \in \boldsymbol{\Phi}\}.$$
As a result, the EM algorithm applies $\phi^{(k+1)} \in T(\phi^{(k)})$;
4. Let $\boldsymbol{\Phi}_{\phi^{(0)}} = \{\phi \in \boldsymbol{\Phi}: \ln L_T(\phi) \geq \ln L_T(\phi^{(0)})\}$.

Theorem 1 (WU Conditional on MNHMM-I [24] [41] [37] **)**
If $\boldsymbol{\Phi}$ is the MNHMM-I parameter space, then the following 4 conditions are fulfilled.

1. $\boldsymbol{\Phi}$ is a finite subset of $\mathbb{R}^{p\times m+3}$,
2. $\ln L_T(\phi)$ is continuous in $\boldsymbol{\Phi}$ and differentiable in the interior $\boldsymbol{\Phi}$,
3. $\boldsymbol{\Phi}_{\phi^{(0)}}$ is a compact set, for any $\phi^{(0)} \in \boldsymbol{\Phi}$, with $\ln L_T(\phi^{(0)}) > -\infty$,
4. $Q(\varphi|\phi)$ is a continuous function with respect to φ and ϕ at $\boldsymbol{\Phi} \times \boldsymbol{\Phi}$.

Proof, see Appendix 4.

Before entering the Wu Theorem (Theorem 2), will proved the following Lemmas:

Lemma 4 (see [24] [29] [41])
If $\phi^{(k)} \in \boldsymbol{\Psi}$, then $\ln L_T(\phi^{(k+1)}) \geq \ln L_T(\phi^{(k)})$ for every $\phi^{(k+1)} \in T(\phi^{(k)})$.
Proof
Determine $k \in \{0,1,2,\dots\}$, and take any $\phi^{(k)} \in \boldsymbol{\Psi}$. Note that
$$\begin{aligned}\ln L_T(\phi^{(k+1)}) - \ln L_T(\phi^{(k)}) &= \left(Q(\phi^{(k+1)}|\phi^{(k)}) - H(\phi^{(k+1)}|\phi^{(k)})\right) - \left(Q(\phi^{(k)}|\phi^{(k)}) - H(\phi^{(k)}|\phi^{(k)})\right)\\ &= \left(Q(\phi^{(k+1)}|\phi^{(k)}) - Q(\phi^{(k)}|\phi^{(k)})\right) - \left(H(\phi^{(k+1)}|\phi^{(k)}) - H(\phi^{(k)}|\phi^{(k)})\right).\end{aligned} \quad (20)$$
Based on the definition of the M Step in the EM algorithm,
$$Q(\phi^{(k+1)}|\phi^{(k)}) \geq Q(\phi^{(k)}|\phi^{(k)}).$$
Corollary,
$$Q(\phi^{(k+1)}|\phi^{(k)}) - Q(\phi^{(k)}|\phi^{(k)}) \geq 0. \quad (21)$$
Based on Lemma 3
$$H(\phi^{(k+1)}|\phi^{(k)}) \leq H(\phi^{(k)}|\phi^{(k)}),$$
as a result
$$H(\phi^{(k+1)}|\phi^{(k)}) - H(\phi^{(k)}|\phi^{(k)}) \leq 0. \quad (22)$$
From $(20), (21)$, dan (22) are obtained
$$\ln L_T(\phi^{(k+1)}) - \ln L_T(\phi^{(k)}) \geq 0.$$
So
$$\ln L_T(\phi^{(k+1)}) \geq \ln L_T(\phi^{(k)}).$$

Lemma 5 (see [24] [29] [41] [42])
If $\phi^{(k)} \notin \boldsymbol{\Psi}$, then $\ln L_T(\phi^{(k+1)}) > \ln L_T(\phi^{(k)})$ for all $\phi^{(k+1)} \in T(\phi^{(k)})$,
Proof
Determine $k \in \{0,1,2,\dots\}$, and take any $\phi^{(k)} \notin \boldsymbol{\Psi}$. Using equation (19), will obtained
$$\partial_{\phi^{(k)}}\left(\ln L_T(\phi^{(k)})\right) = D^{10}Q(\phi^{(k)}|\phi^{(k)}) - D^{10}H(\phi^{(k)}|\phi^{(k)}). \quad (23)$$

Furthermore, based on Lemma 2, $D^{10}H\left(\phi^{(k)}|\phi^{(k)}\right) = 0$. Then equation (23) becomes

$$\partial_{\phi^{(k)}}\left(\ln L_T\left(\phi^{(k)}\right)\right) = D^{10}Q\left(\phi^{(k)}|\phi^{(k)}\right). \tag{24}$$

However $\phi^{(k)} \notin \boldsymbol{\Psi}$, so $\partial_{\phi^{(k)}}\left(\ln L_T\left(\phi^{(k)}\right)\right) \neq 0$. As a result,

$$D^{10}Q\left(\phi^{(k)}|\phi^{(k)}\right) \neq 0.$$

Therefore, $\phi^{(k)}$ is not a local maximum of $Q\left(\phi|\phi^{(k)}\right)$ toward $\phi \in \boldsymbol{\Phi}$, that is $\forall\ \Theta \subset \boldsymbol{\Phi}$ which contains $\phi^{(k)}, \exists\ \bar{\phi} \in \Theta$ such that

$$Q\left(\phi^{(k)}|\phi^{(k)}\right) < Q\left(\bar{\phi}|\phi^{(k)}\right). \tag{25}$$

However according to the definition of M step in the EM algorithm,

$$Q\left(\phi^{(k+1)}|\phi^{(k)}\right) \geq Q\left(\phi|\phi^{(k)}\right),$$

For every $\phi \in \boldsymbol{\Phi}$. So this is also true for $\phi = \bar{\phi}$, that is

$$Q\left(\phi^{(k+1)}|\phi^{(k)}\right) \geq Q\left(\bar{\phi}|\phi^{(k)}\right). \tag{26}$$

From (25) and (26), obtained

$$Q\left(\phi^{(k)}|\phi^{(k)}\right) < Q\left(\phi^{(k+1)}|\phi^{(k)}\right). \tag{27}$$

From (20), (27), and Lemma 3 $\left(H\left(\phi^{(k+1)}|\phi^{(k)}\right) \leq H\left(\phi^{(k)}|\phi^{(k)}\right)\right)$, obtained

$$\ln L_T\left(\phi^{(k+1)}\right) > \ln L_T\left(\phi^{(k)}\right).$$

Lemma 6 (see [24] [42])
The function T is closed in $\boldsymbol{\Phi}\backslash\boldsymbol{\Psi}$.
Proof, see Appendix 5.

Theorem 2 (Wu Theorem on MNHMM-I [24] [29] [41] [42]**)**
Let the $Q(\varphi|\phi)$ is continuous function with respect to φ, ϕ in $\boldsymbol{\Phi} \times \boldsymbol{\Phi}$ Let $\{\phi^{(k)}\}$ be a parameter estimators sequence of MNHMM-I obtained using the EM algorithm. If $\lim\limits_{k\to\infty} \phi^{(k)} = \phi^*$ then,

1. ϕ^* is the stationary point of the function $\ln L_T(\phi)$,
2. $\lim\limits_{k\to\infty} \ln L_T\left(\phi^{(k)}\right) = \ln L_T(\phi^*)$, where the convergence increases monotone.

Proof,

1. Let $\lim\limits_{k\to\infty} \phi^{(k)} = \phi^*$. Suppose ϕ^* is not a stationary point, which is $\phi^* \notin \boldsymbol{\Psi}$. Determine the sequence $\{\phi^{(k+1)}\}_{k=1}^{\infty}$, which is for every k, $\phi^{(k+1)} \in T\left(\phi^{(k)}\right)$. Under the 3rd Wu Condition in Theorem 1, the sequence $\{\phi^{(k+1)}\}_{k=1}^{\infty}$ is in the compact set $\boldsymbol{\Phi}_{\phi^{(0)}}$. Consequently there is a subsequence $\{\phi^{(k+1)_m}\}_{m=1}^{\infty}$ such that $\phi^{(k+1)_m} \to \hat{\phi}$ when $m \to \infty$. A sequence converges to a point if and only if its subsequence converge to that point, consequently,

$$\phi^{(k+1)} \to \hat{\phi} \text{ if } k \to \infty. \tag{28}$$

   Based on Lemma 6 above, T is closed in $\boldsymbol{\Phi}\backslash\boldsymbol{\Psi}$ and by the assumption $\phi^* \notin \boldsymbol{\Psi}$, so that $\hat{\phi} \in T(\phi^*)$. Consequently, based on Lemma 5 then

$$\ln L_T\left(\hat{\phi}\right) > \ln L_T(\phi^*). \tag{29}$$

   Based on (28) and the continuity function $\ln L_T(\phi)$ in $\boldsymbol{\Phi}$ then

$$\lim_{k\to\infty} \ln L_T\left(\phi^{(k+1)}\right) = \lim_{k\to\infty} \ln L_T\left(\hat{\phi}\right), \tag{30}$$

   besides that because $\ln L_T(\phi)$ is a continuous function and the assumption is $\lim\limits_{k\to\infty} \phi^{(k)} = \phi^*$ then

$$\lim_{k\to\infty} \ln L_T\left(\phi^{(k)}\right) = \ln L_T(\phi^*) \tag{31}$$

   and

$$\lim_{k\to\infty} \ln L_T\left(\phi^{(k)}\right) = \lim_{k\to\infty} \ln L_T\left(\phi^{(k+1)}\right). \tag{32}$$

   From (30), (31) and (32) will obtained

$$\ln L_T\left(\hat{\phi}\right) = \ln L_T(\phi^*). \tag{33}$$

   However (29) and (33) are contradict, so that ϕ^* is stationary point.

2. Based on the 1st Wu Theorem, will get ϕ^* as the stationary point of the function $\ln L_T(\phi)$. So it only remains to prove the monotony of $\{\ln L_T\left(\phi^{(k)}\right)\}$. Based on Lemma 4 and Lemma 5 above, $\{\ln L_T\left(\phi^{(k)}\right)\}$ is an ascending monotone sequence, which immediately proves this theorem.

Proven

Conclusions

In conclusion, the multivariate normal hidden Markov model which assumed the Markov chain are inhomogeneous, ergodic and fulfills the assumption of continuity of parameters, then

1. Parameter Estimation of MNHMM-I using the EM algorithm produces a formula that maximizes the likelihood function,
2. The obtained parameter estimator sequence algorithm is converges to the stationary point of the likelihood function monotonically increasing.

Conflicts of Interest

The authors declares that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful suggestions and critical reading of the manuscript.

References

[1] O. Cappe. (2005). *Inference in Hidden Markov Models*. Springer.
[2] S. M. Ross. (2019). *Introduction to Probability Models*. Academic Press Elsevier. Doi: 10.1016/b978-0-12-814346-9.00006-8.
[3] H. Tao and X. Lu. (2019). Smoke vehicle detection based on multi-feature fusion and hidden Markov model. *J. Real-Time Image Process.*, *17*(3), 745-758. Doi: 10.1007/s11554-019-00856-z.
[4] R. Paroli and L. Spezia. (1999). Gaussian Hidden Markov Models: Parameters estimation and applications to air pollution data. Department of Statistical Science, Faculty of Economics, Univ. Cattol. Del Sacro Cuore.
[5] D. Khiatani and U. Ghose. (2017).Weather forecasting using Hidden Markov Model. *2017 Int. Conf. Comput. Commun. Technol. Smart Nation, IC3TSN 2017*. 220-225. Doi: 10.1109/IC3TSN.2017.8284480.
[6] M. Fikri, Samsurizal, Christiono, and K. T. M. (2020). Pemodelan cuaca menggunakan Model Hidden Markov untuk pemanfaatan energi surya. *Kiilat*, *9*(2), 217-224.
[7] Y. Trichilli, M. Boujelbène Abbes, and A. Masmoudi. (2020). Predicting the effect of Googling investor sentiment on Islamic stock market returns: A five-state hidden Markov model. *Int. J. Islam. Middle East. Financ. Manag.*, *13*(2), 165-193. Doi: 10.1108/IMEFM-07-2018-0218.
[8] M. Zhang, X. Jiang, Z. Fang, Y. Zeng, and K. Xu. (2019). High-order Hidden Markov Model for trend prediction in financial time series. *Phys. A Stat. Mech. its Appl.*, *517*, 1-12. Doi: 10.1016/j.physa.2018.10.053.
[9] N. Nguyen. (2018). Hidden Markov Model for stock trading. *Int. J. Financ. Stud.*, *6*(36), 1-17. Doi: 10.3390/ijfs6020036.
[10] P. Somani, S. Talele, and S. Sawant. (2014). Stock market prediction using hidden Markov Model. *2014 IEEE 7th Jt. Int. Inf. Technol. Artif. Intell. Conf*. 89-92.
[11] A. Gupta and B. Dhingra. (2012). Stock market prediction using Hidden Markov Models. *Students Conf. Eng. Syst.* 1-4.
[12] L. R. Rabiner. (1989). A tutorial on Hidden Markov models and selected applications in Speech Re. *Proc. IEEE*, *77*(2), 257-288. Doi: 10.25300/MISQ/2017/41.3.08.
[13] M. Cutajar, E. Gatt, I. Grech, O. Casha, and J. Micallef. (2013). Comparative study of automatic speech recognition techniques. *IET Signal Process.*, *7*(1), 25-46. Doi: 10.1049/iet-spr.2012.0151.
[14] B. Mouaz, B. H. Abderrahim, and E. Abdelmajid. (2019). Speech recognition of moroccan dialect using hidden markov models. *IAES Int. J. Artif. Intell.*, *8*(1), 7-13. Doi: 10.11591/ijai.v8.i1.pp7-13.
[15] A. Luck, P. Giehr, K. Nordstrom, J. Walter, and V. Wolf. (2019). Hidden markov modelling reveals neighborhood dependence of dnmt3a and 3b activity. *IEEE/ACM Trans. Comput. Biol. Bioinforma*, *16*(5), 1598-1609. Doi: 10.1109/TCBB.2019.2910814.
[16] N. Zarrabi, P. Schluesche, M. Meisterernst, M. Börsch, and D. C. Lamb. (2018). Analyzing the dynamics of single TBP-DNA-NC2 complexes using hidden Markov Models. *Biophys. J.*, *115*(12), 2310-2326. Doi: 10.1016/j.bpj.2018.11.015.
[17] M. Fikri and Z. Abdul-Malek. (2023). Partial discharge diagnosis and remaining useful lifetime in XLPE extruded power cables under DC voltage: A review. *Electr. Eng.* Doi: 10.1007/s00202-023-01935-y.
[18] M. Fikri *et al.* (2023). Clustering suara corona discharge berdasarkan tegangan menggunakan metode Fuzzy C-Mean. *ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron.*, *11*(3), 609-624.
[19] N. Pasra, M. Fikri, K. T. Mauriraya, T. Rijanto, and I. G. P. A. Buditjahjanto. (2023). Deteksi suara corona discharge berdasarkan noise menggunakan metode LPC dan Euclidean distance. *ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron.*, *11*(1), 72-85.
[20] M. Fikri, Christiono, and I. G. M. K. (2022). Clustering fenomena corona discharge berdasarkan suara menggunakan metode LPC dan Euclidean distance. *ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron.*, *10*(3), 689-701.

[21] V. S. Barbu and N. Limnios. (2009). *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications*. Springer.

[22] L. Spezia. (2010). Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *J. Time Ser. Anal.*, *31*, 1-11. Doi: 10.1111/j.1467-9892.2009.00635.x.

[23] L. Spezia, M. N. Futter, and M. J. Brewer. (2011). Periodic multivariate normal hidden Markov models for the analysis of water quality time series,”.*nvironmetrics*, *2010*(July), 304-317. Doi: 10.1002/env.1051.

[24] M. Fikri, Z. Abdul-Malek, and K. Atmadja. (1990). Recursive parameter estimation and its convergence for multivariate normal hidden Markov Model.*Thail. Stat.*

[25] J. D. Hamilton. (1990). Analysis of time series subject to changes in regime. *J. Ecomometrics*, *45*, 39-70.

[26] B. Setiawaty, D. H. Santoso, and N. K. K. Ardana. (2007). Pemodelan nilai tukar rupiah terhadap $Us menggunakan deret waktu hidden MARKOV satu waktu sebelumnya. *J. Math. Its Appl.*

[27] L. E. Baum. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Proc Third Symp. Inequalities, ed. O Shisha*, 1-8.

[28] I. L. Macdonald and Walter Zucchini. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall,

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, 1-22. Doi: 10.1111/j.2517-6161.1977.tb01600.x.

[30] M. Fikri, B. Setiawaty, and I. G. P. Purnaba. (2016). Pendugaan parameter dan kekonvergenan penduga parameter model poisson hidden Markov. *J. Math. Its Appl.*, *15*(1) 1-10.

[31] M. Pourahmadi. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, *86*(3), 677-690.

[32] O. Ledoit and M. Wolf. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, *88*, 365-411. Doi: 10.1016/S0047-259X(03)00096-4.

[33] A. J. Rothman, E. Levina, and J. I. Zhu. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, *97*(3), 539-550. Doi: 10.1093/biomet/asq022.

[34] Mohsen Pourahmadi. (2011). Covariance estimation: The GLM and regularization perspectives. *Stat. Sci.*, *26*(3), 369-387. Doi: 10.1214/11-STS358.

[35] C. Lam. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Ann. Stat.*, *44*(3), 928-953. Doi: 10.1214/15-AOS1393.

[36] O. Ledoit and M. Wolf. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Ann. Stat.*, *48*(5), 3043-3065.

[37] R. Paroli, G. Redaelli, and L. Spezia. (2000). Poisson hidden Markov Models for time series of overdispersed insurance counts. *Astin Colloqium*, 461-474.

[38] J. D. Hamilton. (1994). *Time Series Analysis*. Princeton, New Jerse: Princeton University Press.

[39] D. R. S. Geoffrey R. Grimmett. (2001). *Probability and random processes*. Oxford University Press.

[40] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, *41*(1), 64-171.

[41] J. Wu. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.*, *11*(1), 95-103.

[42] W. I. Zangwill. (1969). *Nonlinear Programming: A Unified Approach*. New Jersey: Prentice Hall.

Appendix 1 (Proof of Lemma 1)

Suppose the transition matrix is defined as equations $(2)-(5)$, then consider the following equations

$$\begin{aligned}\gamma_{11} &= P(S_t=1|S_{t-1}=1)\\ &= P(S_t^*=1,S_{t-1}^*=1|S_{t-1}^*=1,S_{t-2}^*=1)\\ &= \frac{P(S_t^*=1,S_{t-1}^*=1,S_{t-1}^*=1,S_{t-2}^*=1)}{P(S_{t-1}^*=1,S_{t-2}^*=1)}\\ &= \frac{P(S_t^*=1,S_{t-1}^*=1,S_{t-2}^*=1)}{P(S_{t-1}^*=1,S_{t-2}^*=1)}\\ &= P(S_t^*=1|S_{t-1}^*=1,S_{t-2}^*=1)\\ &= P(S_t^*=1|S_{t-1}^*=1)\\ &= \gamma_{11}^*\end{aligned}$$

$$\begin{aligned}\gamma_{12} &= P(S_t=2|S_{t-1}=1)\\ &= P(S_t^*=1,S_{t-1}^*=2|S_{t-1}^*=1,S_{t-2}^*=1)\\ &= \frac{P(S_t^*=1,S_{t-1}^*=2,S_{t-1}^*=1,S_{t-2}^*=1)}{P(S_{t-1}^*=1,S_{t-2}^*=1)}\\ &= 0\end{aligned}$$

$$\begin{aligned}\gamma_{13} &= P(S_t=3|S_{t-1}=1)\\ &= P(S_t^*=2,S_{t-1}^*=1|S_{t-1}^*=1,S_{t-2}^*=1)\\ &= \frac{P(S_t^*=2,S_{t-1}^*=1,S_{t-1}^*=1,S_{t-2}^*=1)}{P(S_{t-1}^*=1,S_{t-2}^*=1)}\\ &= \frac{P(S_t^*=2,S_{t-1}^*=1,S_{t-2}^*=1)}{P(S_{t-1}^*=1,S_{t-2}^*=1)}\\ &= P(S_t^*=2|S_{t-1}^*=1,S_{t-2}^*=1)\\ &= P(S_t^*=2|S_{t-1}^*=1)\\ &= \gamma_{12}^*\end{aligned}$$

$$\begin{aligned}\gamma_{14} &= P(S_t=4|S_{t-1}=1)\\ &= P(S_t^*=2,S_{t-1}^*=2|S_{t-1}^*=1,S_{t-2}^*=1)\\ &= \frac{P(S_t^*=2,S_{t-1}^*=2,S_{t-1}^*=1,S_{t-2}^*=1)}{P(S_{t-1}^*=1,S_{t-2}^*=1)}\\ &= 0\end{aligned}$$

$$\begin{aligned}\gamma_{21} &= P(S_t=1|S_{t-1}=2)\\ &= P(S_t^*=1,S_{t-1}^*=1|S_{t-1}^*=1,S_{t-2}^*=2)\\ &= \frac{P(S_t^*=1,S_{t-1}^*=1,S_{t-1}^*=1,S_{t-2}^*=2)}{P(S_{t-1}^*=1,S_{t-2}^*=2)}\\ &= \frac{P(S_t^*=1,S_{t-1}^*=1,S_{t-2}^*=2)}{P(S_{t-1}^*=1,S_{t-2}^*=2)}\\ &= P(S_t^*=1|S_{t-1}^*=1,S_{t-2}^*=2)\\ &= P(S_t^*=1|S_{t-1}^*=1)\\ &= \gamma_{11}^*\end{aligned}$$

$$\begin{aligned}\gamma_{22} &= P(S_t=2|S_{t-1}=2)\\ &= P(S_t^*=1,S_{t-1}^*=2|S_{t-1}^*=1,S_{t-2}^*=2)\\ &= \frac{P(S_t^*=1,S_{t-1}^*=2,S_{t-1}^*=1,S_{t-2}^*=2)}{P(S_{t-1}^*=1,S_{t-2}^*=2)}\\ &= 0\end{aligned}$$

$$\begin{aligned}\gamma_{23} &= P(S_t=3|S_{t-1}=2)\\ &= P(S_t^*=2,S_{t-1}^*=1|S_{t-1}^*=1,S_{t-2}^*=2)\\ &= \frac{P(S_t^*=2,S_{t-1}^*=1,S_{t-1}^*=1,S_{t-2}^*=2)}{P(S_{t-1}^*=1,S_{t-2}^*=2)}\\ &= \frac{P(S_t^*=2,S_{t-1}^*=1,S_{t-2}^*=2)}{P(S_{t-1}^*=1,S_{t-2}^*=2)}\\ &= P(S_t^*=2|S_{t-1}^*=1,S_{t-2}^*=2)\\ &= P(S_t^*=2|S_{t-1}^*=1)\\ &= \gamma_{12}^*\end{aligned}$$

$$\begin{aligned}\gamma_{24} &= P(S_t=4|S_{t-1}=2)\\ &= P(S_t^*=2,S_{t-1}^*=2|S_{t-1}^*=1,S_{t-2}^*=2)\\ &= \frac{P(S_t^*=2,S_{t-1}^*=2,S_{t-1}^*=1,S_{t-2}^*=2)}{P(S_{t-1}^*=1,S_{t-2}^*=2)}\\ &= 0\end{aligned}$$

$$\begin{aligned}\gamma_{31} &= P(S_t=1|S_{t-1}=3)\\ &= P(S_t^*=1,S_{t-1}^*=1|S_{t-1}^*=2,S_{t-2}^*=1)\\ &= \frac{P(S_t^*=1,S_{t-1}^*=1,S_{t-1}^*=2,S_{t-2}^*=1)}{P(S_{t-1}^*=2,S_{t-2}^*=1)}\\ &= 0.\end{aligned}$$

$$\begin{aligned}\gamma_{32} &= P(S_t=2|S_{t-1}=3)\\ &= P(S_t^*=1,S_{t-1}^*=2|S_{t-1}^*=2,S_{t-2}^*=1)\\ &= \frac{P(S_t^*=1,S_{t-1}^*=2,S_{t-1}^*=2,S_{t-2}^*=1)}{P(S_{t-1}^*=2,S_{t-2}^*=1)}\\ &= \frac{P(S_t^*=1,S_{t-1}^*=2,S_{t-2}^*=1)}{P(S_{t-1}^*=2,S_{t-2}^*=1)}\\ &= P(S_t^*=1|S_{t-1}^*=2,S_{t-2}^*=1)\\ &= P(S_t^*=1|S_{t-1}^*=2)\\ &= \gamma_{21}^*\end{aligned}$$

$$\begin{aligned}\gamma_{33} &= P(S_t=3|S_{t-1}=3)\\ &= P(S_t^*=2,S_{t-1}^*=1|S_{t-1}^*=2,S_{t-2}^*=1)\\ &= \frac{P(S_t^*=2,S_{t-1}^*=1,S_{t-1}^*=2,S_{t-2}^*=1)}{P(S_{t-1}^*=2,S_{t-2}^*=1)}\\ &= 0\end{aligned}$$

$$\begin{aligned}\gamma_{34} &= P(S_t=4|S_{t-1}=3)\\ &= P(S_t^*=2,S_{t-1}^*=2|S_{t-1}^*=2,S_{t-2}^*=1)\\ &= \frac{P(S_t^*=2,S_{t-1}^*=2,S_{t-1}^*=2,S_{t-2}^*=1)}{P(S_{t-1}^*=2,S_{t-2}^*=1)}\\ &= \frac{P(S_t^*=2,S_{t-1}^*=2,S_{t-2}^*=1)}{P(S_{t-1}^*=2,S_{t-2}^*=1)}\\ &= P(S_t^*=2|S_{t-1}^*=2,S_{t-2}^*=1)\\ &= P(S_t^*=2|S_{t-1}^*=2)\\ &= \gamma_{22}^*\end{aligned}$$

$$\begin{aligned}\gamma_{41} &= P(S_t=1|S_{t-1}=4)\\ &= P(S_t^*=1,S_{t-1}^*=1|S_{t-1}^*=2,S_{t-2}^*=2)\\ &= \frac{P(S_t^*=1,S_{t-1}^*=1,S_{t-1}^*=2,S_{t-2}^*=2)}{P(S_{t-1}^*=2,S_{t-2}^*=2)}\\ &= 0\end{aligned}$$

$$\begin{aligned}\gamma_{42} &= P(S_t=2|S_{t-1}=4)\\ &= P(S_t^*=1,S_{t-1}^*=2|S_{t-1}^*=2,S_{t-2}^*=2)\\ &= \frac{P(S_t^*=1,S_{t-1}^*=2,S_{t-1}^*=2,S_{t-2}^*=2)}{P(S_{t-1}^*=2,S_{t-2}^*=2)}\\ &= \frac{P(S_t^*=1,S_{t-1}^*=2,S_{t-2}^*=2)}{P(S_{t-1}^*=2,S_{t-2}^*=2)}\\ &= P(S_t^*=1|S_{t-1}^*=2,S_{t-2}^*=2)\\ &= P(S_t^*=1|S_{t-1}^*=2)\\ &= \gamma_{21}^*\end{aligned}$$

$$\begin{aligned}\gamma_{43} &= P(S_t=3|S_{t-1}=4)\\ &= P(S_t^*=2,S_{t-1}^*=1|S_{t-1}^*=2,S_{t-2}^*=2)\\ &= \frac{P(S_t^*=2,S_{t-1}^*=1,S_{t-1}^*=2,S_{t-2}^*=2)}{P(S_{t-1}^*=2,S_{t-2}^*=2)}\\ &= 0\end{aligned}$$

$$\begin{aligned}\gamma_{44} &= P(S_t=4|S_{t-1}=4)\\ &= P(S_t^*=2,S_{t-1}^*=2|S_{t-1}^*=2,S_{t-2}^*=2)\\ &= \frac{P(S_t^*=2,S_{t-1}^*=2,S_{t-1}^*=2,S_{t-2}^*=2)}{P(S_{t-1}^*=2,S_{t-2}^*=2)}\\ &= \frac{P(S_t^*=2,S_{t-1}^*=2,S_{t-2}^*=2)}{P(S_{t-1}^*=2,S_{t-2}^*=2)}\\ &= P(S_t^*=2|S_{t-1}^*=2,S_{t-2}^*=2)\\ &= P(S_t^*=2|S_{t-1}^*=2)\\ &= \gamma_{22}^*\end{aligned}$$

Based on the above equations and the nature of the probability for the transition matrix that the sum in one row must be worth 1, then to estimate the transition matrix Γ in MNHMM-I it is enough to estimate γ_{11}^* and γ_{22}^* and acquire a transition matrix

$$\Gamma=\begin{pmatrix}\gamma_{11}^* & 0 & \gamma_{12}^* & 0\\ \gamma_{11}^* & 0 & \gamma_{12}^* & 0\\ 0 & \gamma_{21}^* & 0 & \gamma_{22}^*\\ 0 & \gamma_{21}^* & 0 & \gamma_{22}^*\end{pmatrix}.$$

Appendix 2 (Proof of Lemma 2)

Take any $\hat{\phi}\in\mathbf{\Phi}$,

$$D^{10}H(\hat{\phi}|\hat{\phi})=\sum_x \partial_{\hat{\phi}}\left(\ln p(x|y,\hat{\phi})\right)p(x|y,\hat{\phi})=\sum_x\frac{\partial_{\hat{\phi}}p(x|y,\hat{\phi})}{\cancel{p(x|y,\hat{\phi})}}\cancel{p(x|y,\hat{\phi})}=\partial_{\hat{\phi}}\left(\sum_x p(x|y,\hat{\phi})\right)=\partial_{\hat{\phi}}(1)=0.$$

Appendix 3 (Proof of Lemma 3)

Take any $\phi,\hat{\phi}\in\mathbf{\Phi}$. If $f(x)=\ln\frac{1}{x}$, then from Jensen's inequality it is obtained,

$$\ln\left(\frac{1}{E_{\hat{\phi}}\left(\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}|y\right)}\right)\le E_{\hat{\phi}}\left(\ln\left(\frac{1}{\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}}\right)|y\right)$$

$$\Leftrightarrow -\ln\left(E_{\hat{\phi}}\left(\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}|y\right)\right)\le -E_{\hat{\phi}}\left(\ln\left(\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}\right)|y\right)\qquad \Leftrightarrow E_{\hat{\phi}}\left(\ln\left(\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}\right)|y\right)\le\ln\left(E_{\hat{\phi}}\left(\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}|y\right)\right)$$

$$\Leftrightarrow E_{\hat{\phi}}\left(\ln\left(\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}\right)|y\right)\le\ln\left(\sum_x\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}p(x|y,\hat{\phi})\right)\qquad \Leftrightarrow E_{\hat{\phi}}\left(\ln\left(\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}\right)|y\right)\le\ln(1)$$

$$\Leftrightarrow E_{\hat{\phi}}\left(\ln\left(\frac{p(x|y,\phi)}{p(x|y,\hat{\phi})}\right)|y\right)\le 0\qquad \Leftrightarrow E_{\hat{\phi}}(\ln p(x|y,\phi)|y)-E_{\hat{\phi}}\left(\ln p(x|y,\hat{\phi})|y\right)\le 0$$

$$\Leftrightarrow E_{\hat{\phi}}(\ln p(x|y,\phi)|y)\le E_{\hat{\phi}}\left(\ln p(x|y,\hat{\phi})|y\right)\qquad \Leftrightarrow H(\phi|\hat{\phi})\le H(\hat{\phi}|\hat{\phi}).$$

Appendix 4 (Proof of Theorem 1)

1. Suppose that T,p,m, and $\varepsilon>0$ are sufficiently small that close to 0 are given. Define the set diameter

$$\mathbf{diam}\ \mathbf{\Phi}=\sqrt{\underbrace{\left(\tfrac{1}{\varepsilon}-\varepsilon\right)^2+\left(\tfrac{1}{\varepsilon}-\varepsilon\right)^2+\cdots+\left(\tfrac{1}{\varepsilon}-\varepsilon\right)^2}_{p\times m}+\underbrace{1^2+1^2}_{1\times 2}+\underbrace{\left(\tfrac{1}{\varepsilon}-\varepsilon\right)^2}_{1\times 1}}$$

$$=\sqrt{(pm+1)\left(\tfrac{1}{\varepsilon}-\varepsilon\right)^2+2}<\sqrt{(pm+1)\left(\tfrac{1}{\varepsilon}\right)^2+2}<\frac{pm+1}{\varepsilon}+2<\infty.$$

As a result, Φ is a finite subset of $\mathbb{R}^{p\times m+3}$.

2. $\ln L_T(\phi)$ is the sum from the multiplication of continuous functions in $\mathbf{\Phi}$ and differentiable in $\mathbf{\Phi}$, then $\ln L_T(\phi)$ is continuous in $\mathbf{\Phi}$ and differentiable in interior $\mathbf{\Phi}$.

3. Take any $\phi^{(0)}\in\mathbf{\Phi}$. It will be proved that $\mathbf{\Phi}_{\phi^{(0)}}$ is compact, i.e. $\mathbf{\Phi}_{\phi^{(0)}}$ is finite and closed. $\mathbf{\Phi}_{\phi^{(0)}}\subset\mathbf{\Phi}$, while $\mathbf{\Phi}$ is finie **(**based on the 1st Wu condition). Consequently, $\mathbf{\Phi}_{\phi^{(0)}}$ is finite. To show $\mathbf{\Phi}_{\phi^{(0)}}$ is closed, sufficient proof $\overline{\mathbf{\Phi}_{\phi^{(0)}}}\subset\mathbf{\Phi}_{\phi^{(0)}}$. Take any $\phi^*\in\overline{\mathbf{\Phi}_{\phi^{(0)}}}$. Then ϕ^* is the limit point of $\mathbf{\Phi}_{\phi^{(0)}}$. Since the point ϕ^* is the limit point of the set $\mathbf{\Phi}_{\phi^{(0)}}$ if and only if there is a distinct sequence in $\mathbf{\Phi}_{\phi^{(0)}}$ which converging to ϕ^*, then $\exists$ the sequence $\{\phi^{(k)}\}$ in $\mathbf{\Phi}_{\phi^{(0)}}$ is such that $\lim_{k\to\infty}\phi^{(k)}\to\phi^*$, with $\phi^{(k)}\neq\phi^*$ for every k.

   Suppose $\phi^*\notin\mathbf{\Phi}_{\phi^{(0)}}$, then $\ln L_T(\phi^*)<\ln L_T\left(\phi^{(0)}\right)$. Determine $\varepsilon=\ln L_T\left(\phi^{(0)}\right)-\ln L_T(\phi^*)>0$. Since $\lim_{k\to\infty}\phi^{(k)}\to\phi^*$ and $\ln L_T(\phi)$ are continuous in $\mathbf{\Phi}$, then $\lim_{k\to\infty}\ln L_T\left(\phi^{(k)}\right)=\ln L_T(\phi^*)$. For >0 above, then $\exists\, k^* \in \mathbb{N}$ such that for $\geq k^*$ it satisfies

$$\left|\ln L_T\left(\phi^{(k)}\right) - \ln L_T(\phi^*)\right| < \varepsilon$$

$$\Rightarrow \ln L_T\left(\phi^{(k)}\right) - \ln L_T(\phi^*) < \varepsilon \Rightarrow \ln L_T\left(\phi^{(k)}\right) - \ln L_T(\phi^*) < \ln L_T\left(\phi^{(0)}\right) - \ln L_T(\phi^*)$$

$$\Rightarrow \ln L_T\left(\phi^{(k)}\right) < \ln L_T\left(\phi^{(0)}\right).$$

This is contradicts with $\phi^{(k)} \in \mathbf{\Phi}_{\phi^{(0)}}$. So $\mathbf{\Phi}_{\phi^{(0)}}$ is a closed set.

4. Because $Q(\varphi|\phi)$ is the addition and multiplication of the functions $\alpha_t(i|\phi), \beta_t(i|\phi), \gamma_{ij}(\phi), \mu_{ij}(\phi)$, $\sigma_{ijk}(\phi), \ln\delta_i\,(\varphi), \ln\mu_{ij}(\varphi)\,, \ln\sigma_{ijk}(\varphi)\,, \ln\gamma_{ij}(\varphi)$ which are continuous in $\mathbf{\Phi}\times\mathbf{\Phi}$, for $t = 1,2,\dots,T$, and $i,j \in \{1,2,3,\dots,m\}$. As a result $Q(\varphi|\phi)$ is a continuous function with respect to φ,ϕ in $\mathbf{\Phi}\times\mathbf{\Phi}$.

Appendix 5 (Proof of Lemma 5)

By using the definition of the set-value function T, from the function $Q(\varphi'|\phi')$ the information is obtained that $\varphi' \in T(\phi')$, with $\varphi',\phi' \in \mathbf{\Phi}$. Take any $\bar{\phi} \in \mathbf{\Phi}\backslash\mathbf{\Psi}$. According to the 4th Wu condition $Q(\varphi|\phi)$ is a continuous function with respect to φ and ϕ at $\mathbf{\Phi}\times\mathbf{\Phi}$, i.e.

if $\phi^{(k)} \to \bar{\phi}$ and $\varphi^{(k)} \to \bar{\varphi}$, then $Q\left(\varphi^{(k)}|\phi^{(k)}\right) \to Q(\bar{\varphi}|\bar{\phi})$,

when $k \to \infty$.

Consequently, obtained $\varphi^{(k)} \in T\left(\phi^{(k)}\right)$ for $k = 0,1,2,\dots$, and satisfy

if $\phi^{(k)} \to \bar{\phi}$ and $\varphi^{(k)} \to \bar{\varphi}$, then $\bar{\varphi} \in T(\bar{\phi})$,

when $k \to \infty$.

As a result the T function is closed, the EM algorithm is a special case by replacing $\varphi^{(k)}$ to $\phi^{(k+1)}$.