MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Fikri *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 21 (2025) 1987-1997

RESEARCH ARTICLE

Recursive Estimation of The Covariance Matrix and Its Convergence for Multivariate Normal Hidden Markov Models

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

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

Abstract This work discusses the covariance matrix estimates and convergence analysis for multivariate normal hidden Markov models. This study findings a series of covariance matrix estimators converges monotonically increasing to a stationary point of the likelihood function through the application of the expectation maximization algorithm.

Keywords: Multivariate hidden Markov model, Covariance matrix, Expectation maximization algorithm, Monotone convergence.

Introduction

Two stochastic processes, the process that causes the observation and the process that is the subject of the observation, make up the hidden Markov model (HMM) [1]. The probability of an observation's effect at a given time solely depends on the effect of an observation made several units of time earlier, which it is assumed that the stochastic processes influencing these observations are not observable and form a Markov chain. State is the common term used to describe this observation's effect [2]. HMM is frequently used to analyse time series data in a variety of problem domains, including stock price forecast [3][4][5][6][7], DNA sequence prediction [8][9], issues with air pollution [10][11], speech recognition [12][13][14], forecasts for the weather [15][16], and it is intended to be utilized for the diagnosis of insulation-related partial discharge acoustic [17][18][19][20][21], and among other applications. This is due to the fact that HMM provides calculation simplification (memoryless properties) to the difficulties raised while maintaining relevance [22]. Whereas regarding the utilization to longitudinal data, it is still quite limited despite offering efficiency. This is a result of the necessary analyses being more difficult than they would be with time series data.

The multivariate normal hidden Markov model (MNHMM) for this subject study, requires multivariate assumptions to be applied to longitudinal data. The MNHMM is one of the HMM that assumes if the state is known, then the probability of observation is multivariate normal distribution [23][24][25] [26]. While parameter estimation and convergence analysis of MNHMM have been done in previous research [23][24][25], it does not discuss the covariance matrix because the covariance matrix has its complexity during the estimation and convergence of its parameter values (multivariate analysis). Numerous studies pertaining to covariance matrices [27][28][29][30][31][32] demonstrate this intricate. This research aims to complement previous research, namely covariance matrix estimation and its convergence for the MNHMM with the assumption that the covariance matrix is well-conditioned at each iteration.

The construction of the MNHMM is refer to [25], then parameter estimation and convergence analysis of the parameter estimators (covariance matrix) are the novelty aspects of this study. Furthermore the likelihood function is maximized for the process of covariance matrix estimation. The method for estimating model parameters with main references [25][33][34] is obtained by recursively maximizing the likelihood function, which is computed using the forward-backwards algorithm [35][36]. This is done using the Expectation Maximization (EM) algorithm. As a result, the convergence of recursive covariance matrix estimation for the multivariate normal hidden Markov models will be covered in this study.

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

Received: 19 May 2024
Accepted: 14 April 2025

Multivariate Normal Hidden Markov Model

The MNHMM is a discrete-time model made up of two stochastic processes $\{X_t, Y_t\}_{t\in\mathbb{N}}$ [1], with $\{Y_t\}_{t\in\mathbb{N}}$ is the observation process that depends only on $\{X_t\}_{t\in\mathbb{N}}$, and $\{X_t\}_{t\in\mathbb{N}}$ is the cause of observation process that assumed forms a Markov chain homogeneous and ergodic (aperiodic, positive recurrent, and irreducible) [2] with $S_X = \{1,2,\cdots,m\}$ is state space. For each $t \in \mathbb{N}$, the random variable Y_t if X_t is known assumed to have a multivariate normal distribution [23], [24], [37].

To simplify the next writing, symbolized the following 10 points:

1. $Y = \{Y_t\}_{t=1}^{T}$, which is a process of observation,
2. $X = \{X_t\}_{t=1}^{T}$, which is the Markov chain,
3. $Z = \{X_t, Y_t\}_{t=1}^{T}$, which is the HMM,
4. $y = (y_1, y_2, \dots, y_T)$ is longitudinal data of the process $\{Y_t\}_{t=1}^{T}$ (commonly called incomplete data),
5. $x = (i_1, i_2, \dots, i_T)$ is effect observation y which unobserved and is the state of process $\{X_t\}_{t=1}^{T}$,
6. $z = (i_1, y_1, \dots, i_T, y_T) = (x, y)$, data and state of the process $\{X_t, Y_t\}_{t=1}^{T}$ (commonly called complete data),
7. $P(Z = z|\phi) = p(z;\phi) = p(x, y|\phi)$, which is the probability mass function of Z,
8. $P(Y = y|\phi) = p(y|\phi)$, which is the probability function of Y,
9. $L_T^c(\phi) = p(z|\phi) = p(x, y|\phi)$, which is the likelihood function of the complete data,
10. $P(X = x|Y = y, \phi) = p(x|y, \phi)$, which is the probability mass function of X with the condition Y = y, i.e. $p(x|y,\phi) = \frac{p(z|\phi)}{p(y|\phi)} = \frac{p(x,y|\phi)}{p(y|\phi)} = \frac{L_T^c(\phi)}{L_T(\phi)}$.

The following is a brief explanation of MNHMM:

1. $y_1 = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{p1} \end{pmatrix}, y_2 = \begin{pmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{p2} \end{pmatrix}, \dots, y_T = \begin{pmatrix} y_{1T} \\ y_{2T} \\ \vdots \\ y_{pT} \end{pmatrix}$ is the longitudinal data will be modeled, where T is the number of time series data and p is the number of cross data. Parameter $M = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1m} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{p1} & \mu_{p2} & \cdots & \mu_{pm} \end{pmatrix}$, and $\sum = (\Sigma_1, \Sigma_2, \dots, \Sigma_m)$ with $\Sigma_i = \begin{pmatrix} \sigma_{i11} & \sigma_{i12} & \cdots & \sigma_{i1p} \\ \sigma_{i21} & \sigma_{i22} & \cdots & \sigma_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{ip1} & \sigma_{ip2} & \cdots & \sigma_{ipp} \end{pmatrix}$, for $i = 1,2,\dots,m$ [23].
2. Transition probability matrix $\Gamma = [\gamma_{ij}]$, with size of Γ matrix is $m \times m$ and $j, i \in S_X$, 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 every $i = 1,2,\dots,m$.
3. The conditional probability Y_t if it is known that $X_t = i$ ($t \in \mathbb{N}$) is a multivariate random variable that is normal, with a covariance matrix $\boldsymbol{\Sigma}$ and a mean $\boldsymbol{\mu}$. The conditional probability of the observation process $\Pi = [\pi_{yi}]$ (for each $y \in \mathbb{R}^p$) is

$$\pi_{yi} = P(Y_t = y|X_t = i) = \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{\Sigma_i}|}}\, e^{\left(-\frac{(\boldsymbol{y}-\boldsymbol{\mu_i})'\boldsymbol{\Sigma_i}^{-1}(\boldsymbol{y}-\boldsymbol{\mu_i})}{2}\right)},$$

$$\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{\Sigma_i}|}}\, e^{\left(-\frac{(\boldsymbol{y}-\boldsymbol{\mu_i})'\boldsymbol{\Sigma_i}^{-1}(\boldsymbol{y}-\boldsymbol{\mu_i})}{2}\right)}\, dy_1\, dy_2 \dots\, dy_p = 1.$$ [38] [39]

4. The stationary distribution is defined as the long-run proportion δ, where $\delta = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_m \end{pmatrix}$ is the initials of the state distribution. Given [2] and the ergodic assumption made on the Markov chain $\{X_t\}_{t\in\mathbb{N}}$, it is possible to uniquely acquire the stationary distribution δ, i.e. fulfilling

$$\Gamma\delta = \delta. \qquad (1)$$

with

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

5. For every $y \in \mathbb{R}^p$ and $t \in \mathbb{N}$, the marginal distribution function of Y_t is

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

Something which very important on MNHMM is estimating model parameters and its convergence. Based on the discussion above, the MNHMM $\{X_t, Y_t\}_{t\in\mathbb{N}}$ is characterized by $\delta, \Gamma, \mu, \Sigma$, with

$$\delta = [\delta_i] \quad i \in S_X,$$
$$\Gamma = [\gamma_{ij}] \quad i,j \in S_X,$$
$$\mu = (\mu_1, \mu_2, \dots, \mu_m), \text{ with } \mu_i = \begin{pmatrix} \mu_{1i} \\ \mu_{2i} \\ \vdots \\ \mu_{pi} \end{pmatrix}, \text{ for } i = 1,2,\dots,m.$$
$$\Sigma = (\Sigma_1, \Sigma_2, \dots, \Sigma_m), \text{ with } \Sigma_i = \begin{pmatrix} \sigma_{i11} & \sigma_{i12} & \cdots & \sigma_{i1p} \\ \sigma_{i21} & \sigma_{i22} & \cdots & \sigma_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{ip1} & \sigma_{ip2} & \cdots & \sigma_{ipp} \end{pmatrix}, \text{ for } i = 1,2,\dots,m.$$

Based on equation (1), δ will be obtained when Γ is obtained so that δ is not a parameter. Then in previous research [25], the parameters Γ and μ were estimated so that only the covariance matrix parameter Σ remained. Therefore, this research will estimate the parameter Σ, so the MNHMM parameter is $\phi = (\Sigma)$. The parameter space and the underlying assumptions must be made clear in order to estimate this parameter, which will be covered in the next chapter.

Parameter Estimation

Let T be the amount of observation time, p is the amount of cross-data at any time, m is the amount of states and $y = (y_1, y_2, \dots, y_T)$ is the sequence of observations. Given that $\varepsilon > 0$ is small enough to approach 0. $\boldsymbol{\Phi} = \left\{\phi = (\Sigma) : \ \Sigma \in \left[\varepsilon, \frac{1}{\varepsilon}\right]^{m\times p^2}\right\}$ is the parameter space for the MNHMM in this research. For every $\phi \in \boldsymbol{\Phi}$, $\Sigma(\phi) = \left(\sigma_{ijk}(\phi)\right), \Gamma(\phi) = \left(\gamma_{ij}(\phi)\right), M(\phi) = \left(\mu_{ij}(\phi)\right), \delta(\phi) = \left(\delta_i(\phi)\right)$ [23][37], is assumed to fulfil the following five points follows:

1. $\gamma_{ij}: \boldsymbol{\Phi} \to \mathbb{R}$ is continuous function in $\boldsymbol{\Phi}$, where $\gamma_{ij}(\phi) = \gamma_{ij}$, $\forall\, i,j \in S_X$,
2. $M_i: \boldsymbol{\Phi} \to \mathbb{R}$ is continuous function in $\boldsymbol{\Phi}$, where $M_i(\phi) = M_i$, $\forall\, i \in S_X$,
3. $\delta_i: \boldsymbol{\Phi} \to \mathbb{R}$ is continuous function in $\boldsymbol{\Phi}$, where $\delta_i(\phi) = \delta_i$, $\forall\, i \in S_X$,
4. $\Sigma_i: \boldsymbol{\Phi} \to \mathbb{R}$ is continuous function in $\boldsymbol{\Phi}$, where $\Sigma_i(\phi) = \Sigma_i$, $\forall\, i \in S_X$,
5. Σ_i is assumed to be a well-conditioned matrix at each iteration, $\forall\, i \in S_X$.

Equation (2) defines the observation process Y as likelihood function:

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

As was mentioned in the previous discussion, the primary challenge with MNHMM is figuring out which parameter $\phi^* \in \boldsymbol{\Phi}$ maximizes the likelihood function $L_T(\phi)$. Calculating the $L_T(\phi)$ function requires significant time for sufficiently large observation data T. The forward-backward algorithm is employed to solve this issue. Recursive calculation is the forward-backward algorithm's basic method of operation, which reduces computing time. There are two parts this algorithm: the forward algorithm and the backward algorithm. The following is previous study define forward probability [40]:

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

and probability of backward:

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

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

The forward and backward algorithms, often known as the recursive formulation for forward probability and backward probability [35][36]. Forward algorithm 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 $i, j \in S_X$.

Next, [35][36] computes the likelihood function $L_T(\phi)$ using forward and backward methods; this process is known as the forward-backward algorithm, and the following results are obtained:

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

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

Equation (3) shows the likelihood function of the complete data.

$$L_T^c(\phi) = \delta_{i_1}\pi_{y_1 i_1}\prod_{t=2}^{T}\gamma_{i_{t-1}i_t}\pi_{y_t i_t}. \tag{3}$$

The following is the probability function relationship between complete data and incomplete data, based on equations (2) and (3):

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

Finding $\phi^* \in \mathbf{\Phi}$ that maximizes $L_T(\phi)$ is a difficult problem. $L_T(\phi)$ will automatically be maximized by $\phi^* \in \mathbf{\Phi}$ when maximizes $\ln L_T(\phi)$. For $\phi \in \mathbf{\Phi}$, holds

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

Pay attention for any $\hat{\phi} \in \mathbf{\Phi}$ also fulfilled

$$E_{\hat{\phi}}(\ln L_T(\phi)\,|y) = E_{\hat{\phi}}(\ln L_T^c(\phi)\,|y) - E_{\hat{\phi}}(\ln p(x|y,\phi)|y), \tag{4}$$

and

$$E_{\hat{\phi}}(\ln L_T(\phi)\,|y) = \sum_x \ln L_T(\phi)\,p\left(x|y,\hat{\phi}\right) = \sum_x \ln p(y|\phi)\,p\left(x|y,\hat{\phi}\right) = \sum_x \ln p(y|\phi)\frac{p\left(x,y|\hat{\phi}\right)}{p\left(y|\hat{\phi}\right)}$$
$$= \frac{\ln p(y|\phi)}{p\left(y|\hat{\phi}\right)}\sum_x p\left(x,y|\hat{\phi}\right) = \frac{\ln p(y|\phi)}{p\left(y|\hat{\phi}\right)}p\left(y|\hat{\phi}\right) = \ln p(y|\phi) = \ln L_T(\phi), \tag{5}$$

so that based on equations (4) and (5) is obtained

$$\ln L_T(\phi) = Q\left(\phi|\hat{\phi}\right) - H\left(\phi|\hat{\phi}\right), \tag{6}$$

with $Q\left(\phi|\hat{\phi}\right) = E_{\hat{\phi}}(\ln L_T^c(\phi)|y)$ dan $H\left(\phi|\hat{\phi}\right) = E_{\hat{\phi}}(\ln p(x|y,\phi)\,|y)$.
The first step in obtaining ϕ^*, which maximizes $\ln L_T(\phi)$, is to solve the equation $\partial_\phi(\ln L_T(\phi))$ equal to zero for obtain a stationary point. Following equation (4) steps, will result in its direct acquisition:

$$\partial_\phi(\ln L_T(\phi)) = E_{\hat{\phi}}\left(\partial_\phi(\ln L_T(\phi))|y\right) \tag{7}$$

Consequent of equations (6) and (7),

$$\partial_\phi(\ln L_T(\phi)) = E_{\hat{\phi}}\left(\partial_\phi(\ln L_T(\phi))|y\right) = E_{\hat{\phi}}\left(\partial_\phi \ln L_T^c(\phi)\,|y\right) - E_{\hat{\phi}}\left(\partial_\phi \ln p(x|y,\phi)|y\right). \tag{8}$$

Define[41]

$$D^{10}Q\left(\phi|\hat{\phi}\right) = E_{\hat{\phi}}\left(\frac{\partial}{\partial\phi}\ln L_T^c(\phi)|y\right), \tag{9}$$

and

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

Consequently, substituting equations (9) and (10) into equation (8) would be:

$$\partial_{\phi}(\ln L_T(\phi)) = D^{10}Q(\phi|\hat{\phi}) - D^{10}H(\phi|\hat{\phi}). \tag{11}$$

Lemma 1 [41]
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 each $\hat{\phi} \in \mathbf{\Phi}$.
Proof, refer to Appendix 1.

Lemma 2 [41]
Suppose $H(\phi|\hat{\phi}) = E_{\hat{\phi}}(\ln p(x|y,\phi)\,|y)$, then $H(\phi|\hat{\phi}) \leq H(\hat{\phi}|\hat{\phi})$, for every $\phi, \hat{\phi} \in \mathbf{\Phi}$.
Proof, refer to Appendix 2.

Based on Lemma 1-2 and equation (14), the stationary point from $Q(\phi|\hat{\phi})$ to $\phi \in \mathbf{\Phi}$ is sufficient to get it for $\ln L_T(\phi)$. A non-linear function $D^{10}Q(\phi|\hat{\phi})$ is intricate to solve analytically concerning the covariance matrix parameter $\phi \in \mathbf{\Phi}$. Corollary, finding a stationary point from $Q(\phi|\hat{\phi})$ to $\phi \in \mathbf{\Phi}$ is a problematic intricate problem. The Expectation Maximization approach is used to solve this problem.

Step E and Step M are the two steps that make up each iteration of the recursive Expectation Maximization (EM) algorithm. Taking $\phi^{(k)}$ as covariance matrix estimator of MNHMM obtained at the k^{th} iteration, which is the first step in the EM algorithm. The definition of steps E and M in the $(k+1)^{\text{th}}$ iteration are:

1. Determine: error tolerance value, the initial value of the covariance matrix estimator $\phi^{(k)}$ for $k = 0$, and the maximum iteration value,
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 $Q(\phi;\phi^{(k)})$, that is
$$Q(\phi^{(k+1)}|\phi^{(k)}) \geq Q(\phi|\phi^{(k)}),$$
for each $\phi \in \mathbf{\Phi}$,
4. Steps 2 through 4 are repeated after substituting k for $k+1$. This process continues until either the maximum iteration is reached or $\left|\ln L_T(\phi^{(k+1)}) - \ln L_T(\phi^{(k)})\right|$ is smaller than the given error, indicating that $\{\ln L_T(\phi^{(k)})\}$ converges.

Estimation of covariance matrix parameters in M step is obtained by $\frac{\partial Q(\phi|\phi^{(k)})}{\partial\sigma_{uvw}(\phi)}$=0 (for $u = 1,2,\ldots,m\ and\ v,w = 1,2,\ldots,p$), so it will be obtained

$$\sigma_{uvw} = \frac{a-b}{-(\Sigma_{uvw})^2\sum_{t=1}^{T}\alpha_t(u|\phi^{(k)})\beta_t(u|\phi^{(k)})},$$

with

$$a = \left(\sum_{\substack{j=1\\ j\neq w}}^{p}(-1)^{v+j}\sigma_{uvj}\Sigma_{uvj}\right)\sum_{t=1}^{T}\alpha_t(u|\phi^{(k)})\beta_t(u|\phi^{(k)})\left(((-1)^{v+w}\Sigma_{uvw}) + (y_t-\mu_u)'\,A\,(y_t-\mu_u)\right),$$

$$b = \sum_{t=1}^{T}\alpha_t(u|\phi^{(k)})\beta_t(u|\phi^{(k)})\left((y_t-\mu_u)'\,B\,(-1)^{v+w}\Sigma_{uvw}\,(y_t-\mu_u)\right),$$

$$A=\begin{pmatrix} (-1)^{1+1}(-1)^{v^*+w^*}\Sigma_{u11v^*w^*} & (-1)^{2+1}(-1)^{v^*+w^*}\Sigma_{u21v^*w^*} & \dots & 0 & \dots & (-1)^{p+1}(-1)^{v^*+w^*}\Sigma_{up1v^*w^*} \\ (-1)^{1+2}(-1)^{v^*+w^*}\Sigma_{u12v^*w^*} & (-1)^{2+2}(-1)^{v^*+w^*}\Sigma_{u22v^*w^*} & \dots & 0 & \dots & (-1)^{p+2}(-1)^{v^*+w^*}\Sigma_{up2v^*w^*} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (-1)^{1+p}(-1)^{v^*+w^*}\Sigma_{u1pv^*w^*} & (-1)^{2+p}(-1)^{v^*+w^*}\Sigma_{u2pv^*w^*} & \dots & 0 & \dots & (-1)^{p+p}(-1)^{v^*+w^*}\Sigma_{uppv^*w^*} \end{pmatrix},$$

$$v^*=\begin{cases} v-1, & j<v \\ 0, & j=v, \\ v, & j>v \end{cases}$$

$$w^*=\begin{cases} w-1, & i<w \\ 0, & i=w. \\ w, & i>w \end{cases}$$

Σ_{ijklm}: The Determinant of a matrix form by manner (removing row j and column k from the matrix Σ_i), then deleting row l and column m.

Covariance Matrix Estimator Sequence Convergence

Additionally, it will prove that the EM method converges to $\ln L_T(\phi^*)$ for the sequence $\{\ln L_T(\phi^{(k)})\}$, with ϕ^* is a stationary point of the function $\ln L_T(\phi)$ and $\phi^{(k)}$ is covariance matrix estimator of the MNHMM in iteration k^{th}. Wu's Theorem (Theorem 2) will address this. To make writing easier, the following symbols are explained before talking about Wu's Theorem:

1. Let k represent the EM algorithm iteration, which is $k \in \{0,1,2,3,\dots\}$,
2. Let $\boldsymbol{\Phi}_{\phi^{(0)}} = \{\phi \in \boldsymbol{\Phi}: \ln L_T(\phi) \geq \ln L_T(\phi^{(0)})\}$,
3. Let T be the set-valued function defined in $\boldsymbol{\Phi}$ and the range $\boldsymbol{\Phi} \ni$ for any $\hat{\phi} \in \boldsymbol{\Phi}$ fulfill
$$T(\hat{\phi}) = \{\varphi' \in \boldsymbol{\Phi} : Q(\varphi'|\hat{\phi}) \geq Q(\varphi|\hat{\phi}) \text{ for every } \varphi \in \boldsymbol{\Phi}\}.$$
Corollary, the EM algorithm applies $\phi^{(k+1)} \in T(\phi^{(k)})$,
4. Let $\boldsymbol{\Psi} = \{\phi \in \text{int } \boldsymbol{\Phi} : \ \phi \ is \text{ stationary point of } \ln L_T(\phi)\}$.

Theorem 1 [25][42][37] **(WU Conditional on MNHMM)**
If $\boldsymbol{\Phi}$ is the covariance matrix parameter space of MNHMM, then the following four conditions are provable:

1. $\boldsymbol{\Phi}$ is a finite subset of $\mathbb{R}^{m\times p^2}$,
2. $\ln L_T(\phi)$ is differentiable in interior $\boldsymbol{\Phi}$, and continuous in $\boldsymbol{\Phi}$,
3. For any $\phi^{(0)} \in \boldsymbol{\Phi}$, will obtain $\boldsymbol{\Phi}_{\phi^{(0)}}$ is a compact set, where $\ln L_T(\phi^{(0)}) > -\infty$,
4. The function $Q(\varphi|\phi)$ is continuous in $\varphi, \phi \in \boldsymbol{\Phi} \times \boldsymbol{\Phi}$.

Proof, refer to Appendix 3.

Prior to introducing Wu's Theorem, the following lemmas will be proved:

Lemma 3 [25][41][42]
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
Take any $\phi^{(k)} \in \boldsymbol{\Psi}$ for $k \in \{0,1,2,\dots\}$, and. Note that

$$\ln L_T(\phi^{(k+1)}) - \ln L_T(\phi^{(k)}) = \left(Q(\phi^{(k+1)}|\phi^{(k)}) - H(\phi^{(k+1)}|\phi^{(k)})\right) - \left(Q(\phi^{(k)}|\phi^{(k)}) - H(\phi^{(k)}|\phi^{(k)})\right)$$
$$= \left(Q(\phi^{(k+1)}|\phi^{(k)}) - Q(\phi^{(k)}|\phi^{(k)})\right) - \left(H(\phi^{(k+1)}|\phi^{(k)}) - H(\phi^{(k)}|\phi^{(k)})\right). \quad (12)$$

Using the EM algorithm's definition of the M Step,
$$Q(\phi^{(k+1)}|\phi^{(k)}) \geq Q(\phi^{(k)}|\phi^{(k)}).$$
Corollary,
$$Q(\phi^{(k+1)}|\phi^{(k)}) - Q(\phi^{(k)}|\phi^{(k)}) \geq 0. \quad (13)$$
Based on Lemma 2
$$H(\phi^{(k+1)}|\phi^{(k)}) \leq H(\phi^{(k)}|\phi^{(k)}),$$
as a result
$$H(\phi^{(k+1)}|\phi^{(k)}) - H(\phi^{(k)}|\phi^{(k)}) \leq 0. \quad (14)$$
From (12), (13), dan (14) 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 4 [25][41][42][43]
If $\phi^{(k)} \notin \mathbf{\Psi}$, then $\ln L_T(\phi^{(k+1)}) > \ln L_T(\phi^{(k)})$ for all $\phi^{(k+1)} \in T(\phi^{(k)})$.
Proof
Take any $\phi^{(k)} \notin \mathbf{\Psi}$ for $k \in \{0,1,2,\dots\}$. Based on equation (11), will obtained

$$\partial_{\phi^{(k)}}(\ln L_T(\phi^{(k)})) = D^{10}Q(\phi^{(k)}|\phi^{(k)}) - D^{10}H(\phi^{(k)}|\phi^{(k)}). \quad (15)$$

Furthermore, based on Lemma 1, $D^{10}H(\phi^{(k)}|\phi^{(k)}) = 0$. Then equation (15) becomes

$$\partial_{\phi^{(k)}}(\ln L_T(\phi^{(k)})) = D^{10}Q(\phi^{(k)}|\phi^{(k)}). \quad (16)$$

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

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

Consequently, $\phi^{(k)}$ is not a local maximum of $Q(\phi|\phi^{(k)})$ toward $\phi \in \mathbf{\Phi}$, that is $\forall\, \Theta \subset \mathbf{\Phi}$ which contains $\phi^{(k)}, \exists\, \bar{\phi} \in \Theta\, \ni$

$$Q(\phi^{(k)}|\phi^{(k)}) < Q(\bar{\phi}|\phi^{(k)}). \quad (17)$$

However the EM algorithm's definition of the M Step,

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

for each $\phi \in \mathbf{\Phi}$. Corollary this is also true for $\phi = \bar{\phi}$, i.e.

$$Q(\phi^{(k+1)}|\phi^{(k)}) \geq Q(\bar{\phi}|\phi^{(k)}). \quad (18)$$

From (17) and (18), obtained

$$Q(\phi^{(k)}|\phi^{(k)}) < Q(\phi^{(k+1)}|\phi^{(k)}). \quad (19)$$

From (12), (19), and Lemma 2 $\left(H(\phi^{(k+1)}|\phi^{(k)}) \leq H(\phi^{(k)}|\phi^{(k)})\right)$, obtained

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

Lemma 5 [25][43]
In $\mathbf{\Phi}\backslash\mathbf{\Psi}$**,** the function T is closed.
Proof, refer to Appendix 4.

Theorem 2 [25][41][42][43] **(Wu Theorem on MNHMM)**
Assume that $Q(\varphi|\phi)$ is a continuous function for $\varphi, \phi \in \mathbf{\Phi} \times \mathbf{\Phi}$. Using the EM algorithm, we will obtain the parameter estimators' sequence of MNHMM $\{\phi^{(k)}\}$. If $\lim_{k\to\infty} \phi^{(k)} = \phi^*$ then,

1. The stationary point of the function $\ln L_T(\phi)$ is denoted by ϕ^*.
2. $\lim_{k\to\infty} \ln L_T(\phi^{(k)}) = \ln L_T(\phi^*)$, with the convergence increases monotone.

Proof,

1. Suppose $\lim_{k\to\infty} \phi^{(k)} = \phi^*$. Let ϕ^* is not a stationary point for $\phi^* \notin \mathbf{\Psi}$. The sequence $\{\phi^{(k+1)}\}_{k=1}^{\infty}$ is determined, which is $\phi^{(k+1)} \in T(\phi^{(k)})$ for every k. Under the 3rd Wu Condition in Theorem 1, the compact set $\mathbf{\Phi}_{\phi^{(0)}}$ contains the sequence $\{\phi^{(k+1)}\}_{k=1}^{\infty}$. Corollary there is a subsequence $\{\phi^{(k+1)_m}\}_{m=1}^{\infty}$ such that $\phi^{(k+1)_m} \to \hat{\phi}$ if $m \to \infty$. A sequence converges to a point if and only if its subsequence converge to that point, corollary,

$$\phi^{(k+1)} \to \hat{\phi} \quad \text{if} \quad k \to \infty. \quad (20)$$

T is closed in $\mathbf{\Phi}\backslash\mathbf{\Psi}$ based on Lemma 5, and from the condition $\phi^* \notin \mathbf{\Psi}$, meaning that $\hat{\phi} \in T(\phi^*)$. Thus, in light of Lemma 4, then

$$\ln L_T(\hat{\phi}) > \ln L_T(\phi^*). \quad (21)$$

With reference to the continuity function $\ln L_T(\phi)$ in $\mathbf{\Phi}$ and equation (20), corollary

$$\lim_{k\to\infty} \ln L_T(\phi^{(k+1)}) = \lim_{k\to\infty} \ln L_T(\hat{\phi}), \quad (22)$$

In addition, since the assumption is $\lim_{k\to\infty} \phi^{(k)} = \phi^*$ and the function $\ln L_T(\phi)$ is continuous, corollary

$$\lim_{k\to\infty} \ln L_T(\phi^{(k)}) = \ln L_T(\phi^*) \quad (23)$$

and

$$\lim_{k\to\infty} \ln L_T(\phi^{(k)}) = \lim_{k\to\infty} \ln L_T(\phi^{(k+1)}). \quad (24)$$

From (22), (23) and (24) to be

$$\ln L_T(\hat{\phi}) = \ln L_T(\phi^*). \quad (25)$$

On the other hand, since (21) and (25) contradict, ϕ^* is a stationary point.

2. The stationary point of the function $\ln L_T(\phi)$ is ϕ^*, obtained by the first Wu Theorem. Therefore, it remains only to prove that $\{\ln L_T(\phi^{(k)})\}$ converges monotonically increasing. Based on Lemma 3 and Lemma 4 above, $\{\ln L_T(\phi^{(k)})\}$ is a monotone increasing sequence. So, this theorem is proven.

Conclusions

The MNHMM which assumed ergodic, fulfills the assumption of continuity of parameters, and covariance matrix is well-condition then

1. The likelihood function is maximized by the covariance matrix estimation of MNHMM using the EM algorithm.
2. The covariance matrix estimator sequence algorithm that is obtained converges to the stationary point of the likelihood function that is monotonically increasing.

Conflicts of Interest

The authors state that there isn't any conflict of interest with this paper's publishing.

Acknowledgement

The authors thank the anonymous reviewers for their critical reading of the text and thoughtful suggestions.

References

[1] Cappé, O. (2005). *Inference in hidden Markov models*. Springer.
[2] Ross, S. M. (2019). *Introduction to probability models*. Academic Press. https://doi.org/10.1016/b978-0-12-814346-9.00006-8
[3] Trichilli, Y., Boujelbène Abbes, M., & Masmoudi, A. (2020). Predicting the effect of Googling investor sentiment on Islamic stock market returns: A five-state hidden Markov model. *International Journal of Islamic and Middle Eastern Finance and Management, 13*(2), 165–193. https://doi.org/10.1108/IMEFM-07-2018-0218
[4] Zhang, M., Jiang, X., Fang, Z., Zeng, Y., & Xu, K. (2019). High-order hidden Markov model for trend prediction in financial time series. *Physica A: Statistical Mechanics and Its Applications, 517*, 1–12. https://doi.org/10.1016/j.physa.2018.10.053
[5] Nguyen, N. (2018). Hidden Markov model for stock trading. *International Journal of Financial Studies, 6*(36), 1–17. https://doi.org/10.3390/ijfs6020036
[6] Somani, P., Talele, S., & Sawant, S. (2014). Stock market prediction using hidden Markov model. *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference* (pp. 89–92). IEEE.
[7] Gupta, A., & Dhingra, B. (2012). Stock market prediction using hidden Markov models. *Students Conference on Engineering Systems* (pp. 1–4).
[8] Luck, A., Giehr, P., Nordstrom, K., Walter, J., & Wolf, V. (2019). Hidden Markov modelling reveals neighborhood dependence of DNMT3a and 3b activity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16*(5), 1598–1609. https://doi.org/10.1109/TCBB.2019.2910814
[9] Zarrabi, N., Schluesche, P., Meisterernst, M., Börsch, M., & Lamb, D. C. (2018). Analyzing the dynamics of single TBP-DNA-NC2 complexes using hidden Markov models. *Biophysical Journal, 115*(12), 2310–2326. https://doi.org/10.1016/j.bpj.2018.11.015
[10] Tao, H., & Lu, X. (2019). Smoke vehicle detection based on multi-feature fusion and hidden Markov model. *Journal of Real-Time Image Processing, 17*(3), 745–758. https://doi.org/10.1007/s11554-019-00856-z
[11] Paroli, R., & Spezia, L. (1999). *Gaussian hidden Markov models: Parameters estimation and applications to air pollution data* [Technical Report No. 94]. Università Cattolica del Sacro Cuore.
[12] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286. https://doi.org/10.25300/MISQ/2017/41.3.08
[13] Cutajar, M., Gatt, E., Grech, I., Casha, O., & Micallef, J. (2013). Comparative study of automatic speech recognition techniques. *IET Signal Processing, 7*(1), 25–46. https://doi.org/10.1049/iet-spr.2012.0151
[14] Mouaz, B., Abderrahim, B. H., & Abdelmajid, E. (2019). Speech recognition of Moroccan dialect using hidden Markov models. *IAES International Journal of Artificial Intelligence, 8*(1), 7–13. https://doi.org/10.11591/ijai.v8.i1.pp7-13
[15] Khiatani, D., & Ghose, U. (2017). Weather forecasting using hidden Markov model. *2017 International Conference on Computing, Communication and Automation* (pp. 220–225). https://doi.org/10.1109/IC3TSN.2017.8284480
[16] Fikri, M., Samsurizal, Christiono, & Mauriraya, K. T. (2020). Pemodelan cuaca menggunakan model hidden

Markov untuk pemanfaatan energi surya. *Kiilat, 9*(2), 217–224.
[17] Fikri, M., & Abdul-Malek, Z. (2023). Partial discharge diagnosis and remaining useful lifetime in XLPE extruded power cables under DC voltage: A review. *Electrical Engineering*. https://doi.org/10.1007/s00202-023-01935-y
[18] Fikri, M., *et al.* (2023). Clustering suara corona discharge berdasarkan tegangan menggunakan metode fuzzy c-mean. *Elkomika: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika, 11*(3), 609–624.
[19] Pasra, N., Fikri, M., Mauriraya, K. T., Rijanto, T., & Buditjahjanto, I. G. P. A. (2023). Deteksi suara corona discharge berdasarkan noise menggunakan metode LPC dan Euclidean distance. *Elkomika: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika, 11*(1), 72–85.
[20] Fikri, M., Christiono, & Mulyadi, I. G. K. (2022). Clustering fenomena corona discharge berdasarkan suara menggunakan metode LPC dan Euclidean distance. *Elkomika: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika, 10*(3), 689–701.
[21] Fikri, M., *et al.* (2024). Comparison of corona discharge identification in 20 kV cubicles based on voltage and noise using ED, HMM, and FCM. *Jurnal Teknologi, 86*(5), 11–22.
[22] Barbu, V. S., & Limnios, N. (2009). *Semi-Markov chains and hidden semi-Markov models toward applications*. Springer.
[23] Spezia, L. (2010). Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis, 31*, 1–11. https://doi.org/10.1111/j.1467-9892.2009.00635.x
[24] Spezia, L., Futter, M. N., & Brewer, M. J. (2011). Periodic multivariate normal hidden Markov models for the analysis of water quality time series. *Environmetrics, 22*(3), 304–317. https://doi.org/10.1002/env.1051
[25] Fikri, M., Abdul-Malek, Z., & Atmadja, K. (2023). Recursive parameter estimation and its convergence for multivariate normal hidden Markov model. *Thai Statistician*.
[26] Fikri, M., Abdul-Malek, Z., Esa, M. R. M., & Supriyanto, E. (2023). Recursive parameter estimation and its convergence for multivariate normal hidden Markov inhomogeneous models. *Malaysian Journal of Fundamental and Applied Sciences, 19*(5), 840–854.
[27] Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika, 86*(3), 677–690.
[28] Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis, 88*, 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4
[29] Rothman, A. J., Levina, E., & Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika, 97*(3), 539–550. https://doi.org/10.1093/biomet/asq022
[30] Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science, 26*(3), 369–387. https://doi.org/10.1214/11-STS358
[31] Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Annals of Statistics, 44*(3), 928–953. https://doi.org/10.1214/15-AOS1393
[32] Ledoit, O., & Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics, 48*(5), 3043–3065.
[33] Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics, 45*, 39–70.
[34] Setiawaty, B., Santoso, D. H., & Ardana, N. K. K. (2007). Pemodelan nilai tukar rupiah terhadap $US menggunakan deret waktu hidden Markov satu waktu sebelumnya. *Jurnal Matematika dan Aplikasinya*.
[35] Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Proceedings of the Third Symposium on Inequalities* (pp. 1–8).
[36] MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall.
[37] Paroli, R., Redaelli, G., & Spezia, L. (2000). Poisson hidden Markov models for time series of overdispersed insurance counts. *ASTIN Colloquium* (pp. 461–474).
[38] Spezia, L. (2010). *Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes* (Vol. 31, pp. 1–11). https://doi.org/10.1111/j.1467-9892.2009.00635.x
[39] Grimmett, G. R., & Stirzaker, D. R. (2001). *Probability and random processes*. Oxford University Press.
[40] Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics, 41*(1), 164–171.
[41] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
[42] Wu, J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, 11*(1), 95–103.
[43] Zangwill, W. I. (1969). *Nonlinear programming: A unified approach*. Prentice-Hall.

Appendix 1 (Lemma 1's proof)

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 2 (Lemma 2's proof)

Take any $\phi, \hat{\phi} \in \mathbf{\Phi}$. Given $f(x) = \ln\frac{1}{x}$, Jensen's inequality yields the following:

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

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

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

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

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

Appendix 3 (Theorem 1's proof)

1. Assume that T, p, m, and $\varepsilon > 0$ are small enough to yield values that are almost equal to zero. Specify the desired diameter:

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

$$= \sqrt{mp^2\left(\frac{1}{\varepsilon}-\varepsilon\right)^2} < \sqrt{mp^2\left(\frac{1}{\varepsilon}\right)^2} < \frac{p\sqrt{m}}{\varepsilon} < \infty.$$

   Corollary, $\mathbf{\Phi}$ is a finite subset of $\mathbb{R}^{m\times p^2}$.
2. When the sum of continuous functions in $\mathbf{\Phi}$ is multiplied by a function that is differentiable in $\mathbf{\Phi}$, we obtain $\ln L_T(\phi)$, which is differentiable in interior $\mathbf{\Phi}$ and continuous in $\mathbf{\Phi}$.
3. Consider any $\phi^{(0)} \in \mathbf{\Phi}$. The compactness of $\mathbf{\Phi}_{\phi^{(0)}}$ will be proved, meaning that it is closed and finite.

   $\mathbf{\Phi}_{\phi^{(0)}} \subset \mathbf{\Phi}$ and $\mathbf{\Phi}$ is finite (refer to the first Wu condition). As a result, $\mathbf{\Phi}_{\phi^{(0)}}$ is finite. It suffices to demonstrate that $\mathbf{\Phi}_{\phi^{(0)}}$ is closed by showing that $\overline{\mathbf{\Phi}_{\phi^{(0)}}} \subset \mathbf{\Phi}_{\phi^{(0)}}$. Consider arbitrary $\phi^* \in \overline{\mathbf{\Phi}_{\phi^{(0)}}}$. The limit point of $\mathbf{\Phi}_{\phi^{(0)}}$ is hence ϕ^*. 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. This is because the point ϕ^* is the limit point of the set $\mathbf{\Phi}_{\phi^{(0)}}$ if and only if there is a distinct sequence in $\mathbf{\Phi}_{\phi^{(0)}}$ that converges to ϕ^*.

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

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

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

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

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

4. Since $Q(\varphi|\phi)$ is the product (multiplication and addition) of the continuous functions $\alpha_t(i|\phi), \beta_t(i|\phi), \gamma_{ij}(\phi), \mu_{ij}(\phi), \sigma_{ijk}(\phi), \ln \delta_i(\varphi), \ln \mu_{ij}(\varphi), \ln \sigma_{ijk}(\varphi), \ln \gamma_{ij}(\varphi)$ in $\boldsymbol{\Phi} \times \boldsymbol{\Phi}$, for $t = 1,2,\dots,T$, and $i,j \in \{1,2,3,\dots,m\}$. Consequently, $Q(\varphi|\phi)$ is a continuous function for φ, ϕ in $\boldsymbol{\Phi} \times \boldsymbol{\Phi}$.

Appendix 4 (Lemma 4's proof)

Based on the definition of the set-value function T, and information of the function $Q(\varphi'|\phi')$ will obtained $\varphi' \in T(\phi')$, with $\varphi', \phi' \in \boldsymbol{\Phi}$. Take any $\bar{\phi} \in \boldsymbol{\Phi} \backslash \boldsymbol{\Psi}$, based on the 4th Wu condition $Q(\varphi|\phi)$ is a continuous function with respect to φ, ϕ in $\boldsymbol{\Phi} \times \boldsymbol{\Phi}$, that is

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

when $k \to \infty$.

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

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

when $k \to \infty$.

Corollary, the T function is closed. The EM algorithm, which changes $\varphi^{(k)}$ to $\phi^{(k+1)}$, is a special case.