Thailand Statistician
July 2025; 23(3): 657-676
http://statassoc.or.th
Contributed paper

Recursive Parameter Estimation and Its Convergence for Multivariate Normal Hidden Markov Model

Miftahul Fikri*[a,b], Zulkurnain Abdul-Malek [b] and Kurniawan Atmadja [c]
[a] Faculty of Electricity and Renewable Energy, Institut Teknologi PLN, Jakarta, Indonesia
[b] High Voltage and High Current Institute, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Johor, Malaysia
[c] Mathematics Study Program, Institut Sains dan Teknologi Nasional, Jakarta, Indonesia
*Corresponding author; e-mail: miftahul@itpln.ac.id

Received: 24 December 2020
Revised: 30 September 2022
Accepted: 18 January 2023

Abstract

Hidden Markov models (HMMs) are models consisting pair of stochastic process which are commonly called observation process and a process that affect observation. Stochastic processes that affect this observation is assumed unobserving and form a Markov chain. HMM is often applied in time series data but still little application to longitudinal data because it requires more complex analysis. One of the HMMs is the multivariate normal hidden Markov model (MNHMM). The MNHMM is a HMMs which the probability of observation if the affect is known and assumed as multivariate normal distribution. This multivariate assumption causes the MNHMM applicable to longitudinal data. The main problem of MNHMM is parameter estimation and the convergence of the parameter estimator sequences. The novelty of this research is the method of estimating the MNHMM parameters used and the analysis of its convergence. Estimation of parameters is done by maximizing the likelihood function. The likelihood function is calculated using the forward-backward algorithm, then maximized recursively using the expectation maximization algorithm (EM algorithm) for obtain a model parameter estimator formula. The MNHM parameter estimator sequence obtained using the EM algorithm converges to the stationary point of the likelihood function monotonically increasing.

Keywords: Multivariate analysis, Markov chain, maximum likelihood, expectation maximization algorithm, monotonically increasing.

1. Introduction

There are many events or phenomena in everyday life that are uncertain. This uncertainty can be modeled by a stochastic process. This is because the stochastic process is a model built with probability rules (Cinlar 2011; 2013). The factors causing this uncertainty are often unobserved. The hidden Markov models (HMMs) can be relied to modeling such a problem. This model can be applied to various problems such as stock price prediction (Trichilli et al. 2020; Zhang et al. 2019; Nguyen 2018) and even gets better results than other methods (Somani et al. 2014; Gupta and Dhingra 2012). Besides that, HMM also can be applied for speech recognition (Rabiner 1989; Cutajar et al. 2013; Mouaz et al. 2019), DNA sequence prediction (Luck et al. 2019; Zarrabi et al. 2018), weather prediction (Khiatani and Ghose 2017; Fikri et al. 2020), application to air pollution (Tao and Lu 2019; Paroli and Spezia 1999), high voltage diagnosis (Fikri et al. 2024), and detect for multivariate time series anomalies (Li et al. 2017). This is because HMMs also offers simplification in calculations (memoryless property) but still preserve relevance to the various application (Barbu and Limnios 2009). The HMM often is applied to data time series as the examples above. However, it can be said that the application for longitudinal data is still rare even though it offers efficiency. It is because the analyzes are not required as easy as when applied to data time series.

HMM consists of a pair of stochastic processes, namely the observation process and a process that affect observation (Cappe 2005). Stochastic processes that affect this observation is assumed unobserving and form a Markov chain. This is probability of effect of observation at any time depends on the effect an observation in several unit's time before. The effect of this observation is usually called the state (Ross 2019). Multivariate normal hidden Markov model (MNHMM) is one of the HMM in which the probability of observation if the state is known and assumed to be multivariate normal distribution (Spezia 2010; Spezia et al. 2011). This multivariate assumption will cause in the event can be modeled form longitudinal data.

The main problem of MNHMM and the objective of this research is parameter estimation and the convergence of its parameter estimator. In previous research, parameter estimation was carried out using Markov chain Monte Carlo (MCMC) (Spezia 2010; Spezia et al. 2011), whereas in this research parameter estimation was carried out which maximized the likelihood function. The likelihood function is calculated using the forward-backward algorithm (Baum 1972; Macdonald and Walter Zucchini 1997), which is then maximized recursively using the expectation maximization algorithm (EM algorithm) to obtain a model parameter estimator formula with main references (Fikri et al. 2016; Wu 1983). Because the estimation and convergence of the covariance matrix parameters has its own complexity and analysis (multivariate analysis) so it will be published separately. This complexity can be seen in several studies related to the covariance matrix (Pourahmadi 1999; Ledoit and Wolf 2004; Rothman et al. 2010; Mohsen Pourahmadi 2011; Lam 2016; Ledoit and Wolf 2020). Estimating the parameters that maximize the likelihood function in longitudinal data using MNHMM is not as easy as in time series data using NHMM. However, behind these difficulties there is an advantage, namely the guarantee of convergence of the parameter estimator sequence and the analysis can be carried out on many data simultaneously. Therefore, after estimating the parameters, we will discuss the convergence of the MNHMM parameter estimator sequence with the main reference (Fikri et al. 2016) (Wu 1983) and simulations accompanied by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) formulations (Leroux and Puterman 1992).

2. Multivariate Normal Hidden Markov Model

The multivariate normal hidden Markov model (MNHMM) is a discrete time model consisting of a pair of stochastic processes $\{X_t, Y_t\}_{t\in\mathbb{N}}$ (Cappe 2005). $\{X_t\}_{t\in\mathbb{N}}$ is the effect of observation which is assumed unobserving and forms a Markov chain. Whereas $\{Y_t\}_{t\in\mathbb{N}}$ is the observation process that only depends on $\{X_t\}_{t\in\mathbb{N}}$. Then the random variable Y_t given a state of X_t is assumed to be multivariate normal distribution, for every $t\in\mathbb{N}$ (Spezia 2010; Spezia et al. 2011; Paroli et al. 2000). In this research $\{X_t\}_{t\in\mathbb{N}}$ is assumed a homogeneous and Ergodic Markov Chain (irreducible, positive recurrent and aperiodic) (Ross 2019) with state space $S_X = \{1, 2, ..., m\}$.

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. $y=(y_1,y_2,...,y_T)$ is longitudinal data of the process $\{Y_t\}_{t=1}^{T}$ (commonly called incomplete data),
5. $x=(i_1,i_2,...,i_T)$ is effect observation y which unobserved and is the state of process $\{X_t\}_{t=1}^{T}$,
6. $z=(i_1,y_1,...,i_T,y_T)=(x,y),$ data and state of the process $\{X_T,Y_T\}_{t=1}^{T}$ (commonly called complete data),
7. $P(Z=z\mid\phi)=p(z\mid\phi)=p(x,y\mid\phi),$ which is the probability mass function of Z,
8. $P(Y=y\mid\phi)=p(y\mid\phi),$ which is the probability function of Y,
9. $L_T^c(\phi)=p(z\mid\phi)=p(x,y\mid\phi),$ which is the likelihood function of the complete data,
10. $P(X=x\mid Y=y,\phi)=p(x\mid y,\phi),$ which is the probability mass function of X with the condition $Y=y,$ i.e. $p\left(x\mid y,\phi\right)=\dfrac{p(z\mid\phi)}{p(y\mid\phi)}=\dfrac{p(x,y\mid\phi)}{p(y\mid\phi)}=\dfrac{L_T^c(\phi)}{L_T(\phi)}.$

The following is a brief explanation of MNHMM:

1. $y_1=\begin{pmatrix}y_{11}\\y_{21}\\\vdots\\y_{p1}\end{pmatrix},y_2=\begin{pmatrix}y_{12}\\y_{22}\\\vdots\\y_{p2}\end{pmatrix},...,y_T=\begin{pmatrix}y_{1T}\\y_{2T}\\\vdots\\y_{pT}\end{pmatrix}$ 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 $\Sigma=(\Sigma_1,\Sigma_2,...,\Sigma_m)$ 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 $i=1,2,\cdots,m$ (Spezia 2010).
2. Transition probability matrix $\Gamma=\left[\gamma_{ij}\right]$, where size of Γ matrix is $m\times m$ and $i,j\in S_X$, satisfies: $\gamma_{ij}=P(X_t=j\mid X_{t-1}=i)=P(X_2=j\mid X_1=i),\ \gamma_{ij}\geq 0,\ \sum_{j=1}^{m}\gamma_{ij}=1,$ for every $i=1,2,...,m.$.
3. The conditional probability Y_t if it is known that $X_t=i\ (t\in\mathbb{N})$ is a normal multivariate random variable with mean μ and covariance matrix $\Sigma.$ For every $y\in\mathbb{R}^p,$ the conditional probability of the observation process $\Pi=\left[\pi_{yi}\right]$ in Geoffrey and Grimmett (2001), Spezia (2010) and Fikri et al. (2023) is

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

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma_i|}}e^{\left(-\frac{(y-\mu_i)'\Sigma_i^{-1}(y-\mu_i)}{2}\right)}dy_1dy_2\cdots dy_p=1$$

4. Let $\delta=\begin{pmatrix}\delta_1\\ \delta_2\\ \vdots\\ \delta_m\end{pmatrix}$ is the initials of the state distribution and the long-run proportion δ is called the stationary distribution. Based on Ross (2019), the Markov chain $\{X_t\}_{t\in\mathbb{N}}$ which is assumed to be ergodic then the stationary distribution δ can be obtained uniquely, namely fulfilling

$$\Gamma\delta=\delta, \tag{1}$$

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

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

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

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

$$\delta=[\delta_i]\text{ for } i\in S_X,\ \Gamma=[\gamma_{ij}]\text{ for } i,j\in S_X,\ \mu=(\mu_1,\mu_2,...,\mu_m),$$

with $\mu_i=\begin{pmatrix}\mu_{1i}\\ \mu_{2i}\\ \vdots\\ \mu_{pi}\end{pmatrix}$, for $i=1,2,...,m.$ $\Sigma=(\Sigma_1,\Sigma_2,...,\Sigma_m)$ 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 $i=1,2,...,m.$

Based on equation (1), δ will be obtained when Γ is obtained so that δ is not a parameter. In addition, for every the diagonal elements of the transition matrix it can be calculated by $\gamma_{ii}=1-\sum_{\substack{j\in S_X\\ j\neq i}}\gamma_{ij}$ so that the corresponding parameter is Γ without its diagonal elements, which is a matrix of size $m\times(m-1)$ and symbolized by $\hat{\Gamma}.$ Because the estimation and convergence of the covariance matrix parameters has its own complexity and analysis (multivariate analysis) so it will be published separately. The parameter estimation in this study is limited to $\phi=(\hat{\Gamma},M).$ In order to estimate this parameter, it is necessary to clarify the parameter space along with the assumptions that accompany which will be discussed in the next chapter.

3. Parameter Estimation and Its Convergence

Let T be the number of times of observation, p is the number of cross data at any time, m is the number of states and $y=(y_1,y_2,...,y_T)$ is the sequence of observations. Given that $\varepsilon>0$ is small enough to approach 0. $\Phi = \left\{ \phi = (\hat{\Gamma}, M) : \hat{\Gamma} \in [0,1]^{m^2-m}, M \in \left[\varepsilon, \frac{1}{\varepsilon} \right]^{p\times m} \right\}$ is the parameter space for the MNHMM. For every $\phi \in \Phi$, $\Gamma(\phi) = (\gamma_{ij}(\phi)), M(\phi) = (\mu_{ij}(\phi)), \Sigma(\phi) = (\sigma_{ijk}(\phi)), \delta(\phi) = (\delta_i(\phi))$ is assumed to fulfill the following four points of continuity (Spezia 2010; Paroli et al. 2000):

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

3.1. Parameter estimation

The likelihood function of the observation process Y is defined as follows:

$$L_T(\phi) = P(Y_1 = y_1, Y_2 = y_2, ..., Y_T = y_T \mid \phi) = p(y_1, y_2, ..., y_T \mid \phi) = p(y \mid \phi)$$

$$= \sum_{i_T=1}^{m} ... \sum_{i_1=1}^{m} \left(\pi_{y_1 i_1} \pi_{y_2 i_2} ... \pi_{y_T i_T} \right) \times \left(\delta_{i_1} \gamma_{i_1 i_2} \gamma_{i_2 i_3} ... \gamma_{i_{T-1} i_T} \right) = \sum_{i_T=1}^{m} ... \sum_{i_1=1}^{m} \delta_{i_1} \pi_{y_1 i_1} \prod_{t=2}^{T} \gamma_{i_{t-1} i_t} \pi_{y_t i_t}. \quad (2)$$

As previously explained, the main problem in the MNHMM is to find the parameter $\phi^* \in \Phi$ which maximizes the likelihood function $L_T(\phi)$. For the number of observation data T which is quite large, calculating the likelihood function takes a long time. A forward-backward algorithm can be used to solve this problem. The forward-backward algorithm is used to calculate the probability of a sequence of observations $(y_1, y_2, ..., y_T)$, recursively, this is very useful to speed up computation time. These algorithms are divided into two, namely the forward algorithm and the backward algorithm. Baum et al. (1970) defined a forward probability as follows:

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

and backward probability

$$\beta_t(i \mid \phi) = P(Y_{t+1} = y_{t+1}, ..., Y_T = y_T \mid X_t = i, \phi), \text{ for } t = 1, 2, \cdots, T \text{ and } i \in S_X.$$

The formulation of recursive forward and backward probabilities which are usually called the forward algorithm, are as follows (Baum 1972; (Macdonald and Zucchini 1997):

$$\alpha_1(i \mid \phi) = \pi_{y_1 i} \delta_i, \ \alpha_{t+1}(j \mid \phi) = \left(\sum_{i \in S_X} \alpha_t(i \mid \phi) \gamma_{ij} \right) \pi_{y_{t+1} j},$$

and backward algorithm

$$\beta_T(j \mid \phi) = 1, \ \beta_t(j \mid \phi) = \sum_{i \in S_X} \beta_{t+1}(i \mid \phi) \pi_{y_{t+1} i} \gamma_{ji}, \text{ for } t = 1, ..., T-1 \text{ and } i, j \in S_X.$$

Then, Baum (1972) and Macdonald and Zucchini (1997) used the forward algorithm and the backward algorithm to calculate the likelihood function $L_T(\phi)$, which is commonly called the forward-backward algorithm and obtained:

$$L_T(\phi) = \sum_{i \in S_X} \alpha_t(i \mid \phi) \beta_t(i \mid \phi), \text{ for any } t = 1, 2, ..., T, \text{ and } i \in S_X.$$

The likelihood function of the complete data is as follows:

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

Based on Equations (2) and (3), the relationship between the incomplete data likelihood function and complete data are as follows:

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

To get $\phi^*\in\Phi$ which maximizes $L_T(\phi)$ is a difficult problem. $\phi^*\in\Phi$ which maximizes $\ln L_T(\phi)$ will also maximize $L_T(\phi)$. For $\phi^*\in\Phi,$ apply

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

Note that for any case $\hat{\phi}\in\Phi$ be valid also

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

and

$$\begin{aligned}E_{\hat{\phi}}\left(\ln L_T(\phi)\mid y\right)&=\sum_x \ln L_T(\phi)p(x\mid y,\hat{\phi})=\sum_x \ln p(y\mid\phi)p(x\mid y,\hat{\phi})\\&=\sum_x \ln p(y\mid\phi)\frac{p(x,y\mid\hat{\phi})}{p(y\mid\hat{\phi})}=\frac{\ln p(y\mid\phi)}{p(y\mid\hat{\phi})}\sum_x p(x,y\mid\hat{\phi})\\&=\frac{\ln p(y\mid\phi)}{p(y\mid\hat{\phi})}p(y\mid\hat{\phi})=\ln p(y\mid\phi)\quad=\ln L_T(\phi),\end{aligned} \tag{5}$$

so based on Equations (4) and (5) are obtained

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

with $Q(\phi\mid\hat{\phi})=E_{\hat{\phi}}\left(\ln L_T^c(\phi)\mid y\right)$ and $H(\phi\mid\hat{\phi})=E_{\hat{\phi}}\left(\ln p(x\mid y,\phi)\mid y\right).$

The first step to obtain ϕ^* which maximizes $\ln L_T(\phi)$ is to solve the equation $\partial_\phi\left(\ln L_T(\phi)\right)=0$ to get a stationary point. By following the pattern of Equation (4), it will be obtained directly

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

As a result of Equations (6) and (7) then

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

Define (Dempster et al. 1977)

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

and

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

then with substituting Equations (9) and (10) into Equation (8) will be obtained

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

Lemma 1 (see Dempster et al. 1977)

Suppose $D^{10}H(\phi\mid\hat{\phi})=E_{\hat{\phi}}\left(\frac{\partial}{\partial\phi}\ln p(x\mid y,\phi)\mid y\right),$ *then* $D^{10}H(\hat{\phi}\mid\hat{\phi})=0,$ *for every* $\hat{\phi}\in\Phi.$

Proof: Take any $\hat{\phi} \in \Phi,$

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

Lemma 2 (see Dempster et al. 1977)
Suppose $H(\phi \mid \hat{\phi}) = E_{\hat{\phi}}\left(\ln p(x \mid y,\phi) \mid y\right),$ *then* $H(\phi \mid \hat{\phi}) \le H(\hat{\phi} \mid \hat{\phi}),$ *for every* $\phi,\hat{\phi} \in \Phi.$

Proof: Take any $\phi,\hat{\phi} \in \Phi,$ if taken $f(x) = \ln(1/x),$ then from Jensen inequality obtained,

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

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

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

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

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

Based on Equation (6), Lemma 1 and Lemma 2 to find the stationary point of $\ln L_T(\phi)$ just find the stationary point of $Q(\phi \mid \hat{\phi})$ with respect to $\phi \in \Phi.$ However, $D^{10}Q(\phi \mid \hat{\phi})$ is a non-linear function and difficult to solve explicitly for the parameter $\phi \in \Phi,$ consequently to obtain a stationary point of $Q(\phi \mid \hat{\phi})$ respect to $\phi \in \Phi$ which analytic is a problem difficult, so this problem is solved by a numerical approach. This research used the expectation maximization algorithm.

The expectation maximization (EM) algorithm is a recursive algorithm that consists of two steps in each iteration, namely step E and step M. The steps in the EM algorithm are to take $\phi^{(k)}$ as an estimator of the MNHMM parameters obtained in the iteration k. In the iteration $(k+1),$ step E and step M are defined as follows:

1. Give error tolerance, maximum iteration and initial parameter value $\phi^{(k)}$ for $k = 0,$
2. E step – count
$Q\left(\phi;\phi^{(k)}\right) = E_{\phi^{(k)}}\left(\ln L_T^c(\phi) \mid Y = y\right)$

$$= \sum_{i \in S_X} \frac{\alpha_1\left(i \mid \phi^{(k)}\right)\beta_1\left(i \mid \phi^{(k)}\right)}{\sum_{l \in S_X} \alpha_t\left(l \mid \phi^{(k)}\right)\beta_t\left(l \mid \phi^{(k)}\right)} \ln \delta_i(\phi) + \sum_{i \in S_X} \frac{\sum_{t=1}^{T} \alpha_t\left(i \mid \phi^{(k)}\right)\beta_t\left(i \mid \phi^{(k)}\right)}{\sum_{l \in S_X} \alpha_t\left(l \mid \phi^{(k)}\right)\beta_t\left(l \mid \phi^{(k)}\right)} \ln P\left(Y_t = y_t \mid X_t = i,\phi\right)$$

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

3. M step – find $\phi^{(k+1)}$ which maximizes $Q\left(\phi;\phi^{(k)}\right)$, so that

$$Q\left(\phi^{(k+1)}\,|\,\phi^{(k)}\right)\geq Q\left(\phi\,|\,\phi^{(k)}\right),\ \text{for every }\ \phi\in\Phi,$$

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

In M step, to obtain the parameter $\gamma_{uv}\left(\phi^{(k+1)}\right)$ which maximizes $Q\left(\phi\,|\,\phi^{(k)}\right)$ respect to $\phi\in\Phi$ is to use the Lagrange multiplier method with the constraint $\sum_{j=1}^{m}\gamma_{ij}(\phi)=1,$ for $u,v,i=1,2,...,m.$ 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 $\dfrac{\partial G\left(\phi\,|\,\phi^{(k)}\right)}{\partial\gamma_{uv}(\phi)}=0$ (for $u,v=1,2,...,m$) implies

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

Estimation of the mean parameter is obtained by $\dfrac{\partial Q\left(\phi\,|\,\phi^{(k)}\right)}{\partial\mu_{uv}(\phi)}=0,$ so it will be obtained

$$\mu_{uv}\left(\phi^{(k+1)}\right)=\frac{\sum_{t=1}^{T}\alpha_t\left(v\,|\,\phi^{(k)}\right)\beta_t\left(v\,|\,\phi^{(k)}\right)\left(2s_{vuu}y_{ut}+\sum_{\substack{k=1\\k\neq u}}^{p}s_{vku}(y_{kt}-\mu_{kv})+\sum_{\substack{k=1\\k\neq u}}^{p}s_{vuk}(y_{kt}-\mu_{kv})\right)}{2s_{vuu}\sum_{t=1}^{T}\alpha_t\left(v\,|\,\phi^{(k)}\right)\beta_t\left(v\,|\,\phi^{(k)}\right)},$$

for $u=1,2,...,m;\ v=1,2,...,p,$ and $\Sigma_i^{-1}=\begin{pmatrix}s_{i11}&s_{i12}&\cdots&s_{i1p}\\s_{i21}&s_{i22}&\cdots&s_{i2p}\\\vdots&\vdots&\ddots&\vdots\\s_{ip1}&s_{ip2}&\cdots&s_{ipp}\end{pmatrix},$ for $i=1,2,...,m.$

3.2. Parameter estimator sequence convergence of MNHMM

Furthermore, it will be proved that the sequence $\left\{\ln L_T(\phi^{(k)})\right\}$ converges to $\ln L_T(\phi^*)$ using the EM algorithm, where $\phi^{(k)}$ is the estimator of the MNHMM parameter in iteration k and ϕ^* are stationary points of the function $\ln L_T(\phi).$ This will be discussed in the Wu theorem. Before discussing the Wu theorem, the following symbols are considered to simplify writing

- Let k denotes the iteration of the EM algorithm, namely $k\in\{0,1,2,3,...\}.$
- Let $\left\{\Psi=\{\phi\in\operatorname{int}\Phi:\text{stationary point of }\ln L_T(\phi)\right\}.$

- Let T be the set-valued function defined at Φ and with the range Φ such that for any $\hat{\phi} \in \Phi$ satisfies

$$T(\hat{\phi}) = \{\varphi' \in \Phi : Q(\varphi' \mid \hat{\phi}) \geq Q(\varphi \mid \hat{\phi}) \text{ for every } \varphi \in \Phi\}.$$

As a result, the EM algorithm applies $\phi^{(k+1)} \in T(\phi^{(k)})$.

- Let $\Phi_{\phi^{(0)}} = \left\{\phi \in \Phi : \ln L_T(\phi) \geq \ln L_T(\phi^{(0)})\right\}$.

Theorem 1 (WU conditional on MNHMM**) (**Wu 1983; Paroli et al. 2000**)**
If Φ is the MNHMM parameter space, then the following 4 conditions are met.

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

Proof:

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

$$diam\Phi = \sqrt{1^2 + 1^2 + \cdots + 1^2 + \left(\frac{1}{\varepsilon} - \varepsilon\right)^2 + \left(\frac{1}{\varepsilon} - \varepsilon\right)^2 + \cdots + \left(\frac{1}{\varepsilon} - \varepsilon\right)^2}$$
$$= \sqrt{m^2 - m + (p\times m)\left(\frac{1}{\varepsilon} - \varepsilon\right)^2} < \sqrt{m^2 + (p\times m)\left(\frac{1}{\varepsilon}\right)^2} < m + \frac{\sqrt{p\times m}}{\varepsilon} < \infty.$$

Consequently Φ is a finite subset of $\mathbb{R}^{m^2-m+p\times m}$.

2. $\ln L_T(\phi)$ is the sum from the multiplication of continuous functions in Φ and differentiable at Φ, then $\ln L_T(\phi)$ is continuous at Φ and differentiable in interior Φ.

3. Take any $\phi^{(0)} \in \Phi$. It will be proven that $\Phi_{\phi^{(0)}}$ is compact, that is, $\Phi_{\phi^{(0)}}$ is finite and closed. $\Phi_{\phi^{(0)}} \subset \Phi$, while Φ is finite (based on the first Wu condition). As a result, $\Phi_{\phi^{(0)}}$ is finite. To show $\Phi_{\phi^{(0)}}$ closed, simply show $\overline{\Phi_{\phi^{(0)}}} \subset \Phi_{\phi^{(0)}}$. Take arbitrarily $\phi^* \in \overline{\Phi_{\phi^{(0)}}}$. Then ϕ^* is the limit point of $\Phi_{\phi^{(0)}}$. Since the point ϕ^* is the limit point of the set $\Phi_{\phi^{(0)}}$ if and only if there is a distinct sequence in $\Phi_{\phi^{(0)}}$ which converging to ϕ^*, then $\exists$ the sequence $\left\{\phi^{(k)}\right\}$ in $\Phi_{\phi^{(0)}}$ is such that $\lim_{k\to\infty} \phi^{(k)} \to \phi^*$, with $\phi^{(k)} \neq \phi^*$ for every k.

4. Suppose $\phi^* \notin \Phi_{\phi^{(0)}}$, then $\ln L_T(\phi^*) < \ln L_T(\phi^{(0)})$. Take $\varepsilon = \ln L_T(\phi^{(0)}) - \ln L_T(\phi^*) > 0$. Since $\lim_{k\to\infty} \phi^{(k)} \to \phi^*$ and $\ln L_T(\phi)$ are continuous in Φ, then $\lim_{k\to\infty} \ln L_T(\phi^{(k)}) = \ln L_T(\phi^*)$. For $\varepsilon > 0$ above, then $\exists k^* \in \mathbb{N} \ni k \geq k^*$ it satisfies

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

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

This contradicts with $\phi^{(k)} \in \Phi_{\phi^{(0)}}$. So $\Phi_{\phi^{(0)}}$ is a closed set. Because $Q(\varphi \mid \phi)$ is the addition and multiplication of the functions $\alpha_t(i \mid \phi)$, $\beta_t(i \mid \phi)$, $\gamma_{ij}(\phi), \mu_{ij}(\phi), \sigma_{ijk}(\phi), \ln \delta_i(\varphi)$, $\ln \mu_{ij}(\varphi)$, $\ln \sigma_{ijk}(\varphi)$, $\ln \gamma_{ij}(\varphi)$ which are continuous in $\Phi \times \Phi$, for $t = 1, 2, ..., T$ and $i, j \in \{1, 2, \cdots, m\}$. Then $Q(\varphi \mid \phi)$ is a continuous function with respect to φ, ϕ in $\Phi \times \Phi$. Before entering the Wu Theorem, will proved the following lemmas:

Lemma 3 (see Dempster et al. 1977; Wu 1983)
If $\phi^{(k)} \in \Psi$, *then* $\ln L_T\left(\phi^{(k+1)}\right) \geq \ln L_T\left(\phi^{(k)}\right)$ *for all* $\phi^{(k+1)} \in T\left(\phi^{(k)}\right)$.

Proof:
Determine $k \in \{0, 1, 2, ...\}$, and take any $\phi^{(k)} \in \Psi$. Note that

$$\ln L_T\left(\phi^{(k+1)}\right) - \ln L_T\left(\phi^{(k)}\right) = \left(Q\left(\phi^{(k+1)} \mid \phi^{(k)}\right) - H\left(\phi^{(k+1)} \mid \phi^{(k)}\right)\right) - \left(Q\left(\phi^{(k)} \mid \phi^{(k)}\right) - H\left(\phi^{(k)} \mid \phi^{(k)}\right)\right)$$
$$= \left(Q\left(\phi^{(k+1)} \mid \phi^{(k)}\right) - Q\left(\phi^{(k)} \mid \phi^{(k)}\right)\right) - \left(H\left(\phi^{(k+1)} \mid \phi^{(k)}\right) - H\left(\phi^{(k)} \mid \phi^{(k)}\right)\right). \quad (9)$$

Based on the definition of the M-step in the EM algorithm,

$$Q\left(\phi^{(k+1)} \mid \phi^{(k)}\right) \geq Q\left(\phi^{(k)} \mid \phi^{(k)}\right).$$

As a result,

$$Q\left(\phi^{(k+1)} \mid \phi^{(k)}\right) - Q\left(\phi^{(k)} \mid \phi^{(k)}\right) \geq 0. \quad (10)$$

Based on Lemma 2,

$$H\left(\phi^{(k+1)} \mid \phi^{(k)}\right) \leq H\left(\phi^{(k)} \mid \phi^{(k)}\right),$$

as a result

$$H\left(\phi^{(k+1)} \mid \phi^{(k)}\right) - H\left(\phi^{(k)} \mid \phi^{(k)}\right) \leq 0. \quad (11)$$

From (12), (13) and (14), we obtained

$$\ln L_T\left(\phi^{(k+1)}\right) - \ln L_T\left(\phi^{(k)}\right) \geq 0.$$

So

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

Lemma 4 (see Dempster et al. 1977; Wu 1983; Zangwill 1969)
If $\phi^{(k)} \notin \Psi$, *then* $\ln L_T\left(\phi^{(k+1)}\right) \geq \ln L_T\left(\phi^{(k)}\right)$ *for all* $\phi^{(k+1)} \in T\left(\phi^{(k)}\right)$.

Proof:
Determine $k \in \{0, 1, 2, \ldots\}$, and take any $\phi^{(k)} \notin \Psi$. Using Equation (11), it is obtained

$$\partial_{\phi^{(k)}}\left(\ln L_T\left(\phi^{(k)}\right)\right) = D^{10} Q\left(\phi^{(k)} \mid \phi^{(k)}\right) - D^{10} H\left(\phi^{(k)} \mid \phi^{(k)}\right). \quad (12)$$

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

$$\partial_{\phi^{(k)}}\left(\ln L_T\left(\phi^{(k)}\right)\right) = D^{10}Q\left(\phi^{(k)} \mid \phi^{(k)}\right). \tag{13}$$

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

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

Therefore $\phi^{(k)}$ is not a local maximum of $Q\left(\phi \mid \phi^{(k)}\right)$ with respect to $\phi \in \Phi,$i.e.$\forall \Theta \subset \Phi$ which contains $\phi^{(k)}, \exists \overline{\phi} \in \Theta$ such that

$$Q\left(\phi^{(k)} \mid \phi^{(k)}\right) < Q\left(\overline{\phi} \mid \phi^{(k)}\right). \tag{14}$$

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

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

for each $\phi \in \Phi.$ So this is also true for $\phi = \overline{\phi},$ that is

$$Q\left(\phi^{(k+1)} \mid \phi^{(k)}\right) \geq Q\left(\overline{\phi} \mid \phi^{(k)}\right). \tag{15}$$

From (17) and (18), we obtained

$$Q\left(\phi^{(k)} \mid \phi^{(k)}\right) < Q\left(\phi^{(k+1)} \mid \phi^{(k)}\right). \tag{16}$$

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

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

Lemma 5 (see Zangwill 1969) *The function T is closed in $\Phi \setminus \Psi$.*

Proof:

By using the definition of a function with set value $T,$ from the function $Q(\varphi' \mid \varphi')$ obtained the information that $\varphi' \in T\left(\phi'\right)$ with $\varphi', \phi' \in \Phi.$ Take any $\overline{\phi} \in \Phi \setminus \Psi.$ Under the 4th Wu condition $Q(\varphi \mid \phi)$ is a continuous function with respect to φ and ϕ in $\Phi \times \Phi,$ i.e.

$$\text{if } \phi^{(k)} \to \overline{\phi} \text{ and } \varphi^{(k)} \to \overline{\varphi} \text{ then } Q\left(\varphi^{(k)} \mid \phi^{(k)}\right) \to Q\left(\overline{\varphi} \mid \overline{\phi}\right), \text{ when } k \to \infty.$$

As a result, we get $\varphi^{(k)} \in T\left(\phi^{(k)}\right)$ for $k = 0, 1, 2, \cdots$ and it fulfill

$$\text{if } \phi^{(k)} \to \overline{\phi} \text{ and } \varphi^{(k)} \to \overline{\varphi}, \text{ then } \overline{\varphi} \in T\left(\overline{\phi}\right), \text{ when } k \to \infty.$$

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

Theorem 2 (Wu Theorem on MNHMM**)** (see Dempster et al. 1977; Wu 1983; Zangwill 1969)
Let the $Q(\varphi \mid \phi)$ is continuous function with respect to φ, ϕ in $\Phi \times \Phi.$ Let $\left\{\phi^{(k)}\right\}$ be a parameter estimators sequence of MNHMM obtained using the EM algorithm. If $\lim_{k \to \infty} \phi^{(k)} = \phi^*$, then

1. *ϕ^* is the stationary point of the function $\ln L_T(\phi),$*

2. $\lim_{k\to\infty} \ln L_T\left(\phi^{(k)}\right) = \ln L_T\left(\phi^*\right),$ *where the convergence increases monotone.*

Proof:

1. Let $\lim_{k\to\infty} \phi^{(k)} = \phi^*$. Suppose ϕ^* is not a stationary point, which is $\phi^* \notin \Psi$. Specify the sequence $\left\{\phi^{(k+1)}\right\}_{k=1}^{\infty}$, which is for every $k, \phi^{(k+1)} \in T\left(\phi^{(k)}\right)$. Under the 3rd Wu condition, the sequence $\left\{\phi^{(k+1)}\right\}_{k=1}^{\infty}$ is in the compact set $\Phi_{\phi^{(0)}}$. Consequently there is a subsequence $\left\{\phi^{(k+1)_m}\right\}_{m=1}^{\infty}$ 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,

$$\phi^{(k+1)} \to \hat{\phi} \text{ if } k \to \infty. \tag{17}$$

Based on Lemma 5 above, T is closed in $\Phi \setminus \Psi$ and by the assumption $\phi^* \notin \Psi$, so that $\hat{\phi} \in T\left(\phi^*\right)$. As a result, based on Lemma 4 then

$$\ln L_T\left(\hat{\phi}\right) > \ln L_T\left(\phi^*\right). \tag{21}$$

Based on Equation (20) and the continuity of the function $\ln L_T(\phi)$ in Φ then

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

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

$$\lim_{k\to\infty} \ln L_T\left(\phi^{(k)}\right) = \ln L_T\left(\phi^*\right), \tag{23}$$

and

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

From Equations (22) (23) and (24), we obtained

$$\ln L_T\left(\hat{\phi}\right) = \ln L_T\left(\phi^*\right). \tag{25}$$

However, Equations (21) and (25) are contradict, so that ϕ^* is stationary point.

2. Based on the 1st Wu theorem, we get ϕ^* as the stationary point of the function $\ln L_T(\phi)$. So it only remains to prove the monotony of $\left\{\ln L_T\left(\phi^{(k)}\right)\right\}$. Based on Lemma 3 and Lemma 4 above, $\left\{\ln L_T\left(\phi^{(k)}\right)\right\}$ is an ascending monotone sequence, which immediately proves this theorem.

Based on the discussion in this chapter, the convergence of the likelihood function obtained will only lead to the stationary point of the likelihood function, monotonically increasing. As a result, it is very important to determine the initial value of the MNHMM parameter estimator in the EM algorithm.

4. Simulation

Estimation and convergence of the parameters discussed above were then simulated on random data with value intervals [10−100], time series 50, 75, 100 and a cross-section of $\pm 0, \pm 25$ from time series data. Because the covariance matrix is not estimated, it is determined that all k covariance matrices have the same value for $k = 1, 2, \ldots, m$ which is the covariance matrix of the generated random data. In the study of multivariate analysis, it is required that the covariance matrix is a positive definite matrix and is well-conditioned so that this covariance matrix is transformed using the formula by (Huang et al. 2017) (Young et al. 2017) so that the covariance matrix meets positive definite and well-conditioned (not discussed in this study). Some sample data are presented as follows:

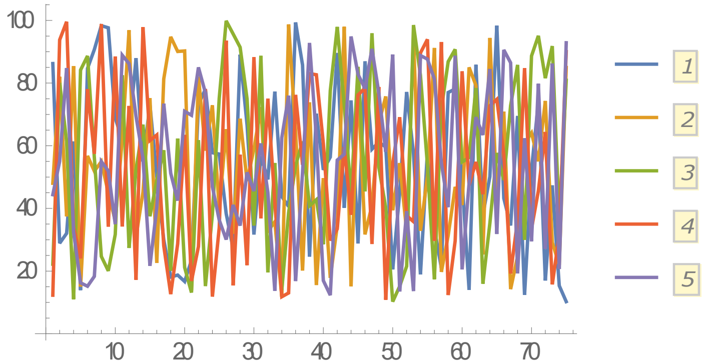

Figure 1 Random panel data graph with 50 time series and 25 cross-section (5 samples)

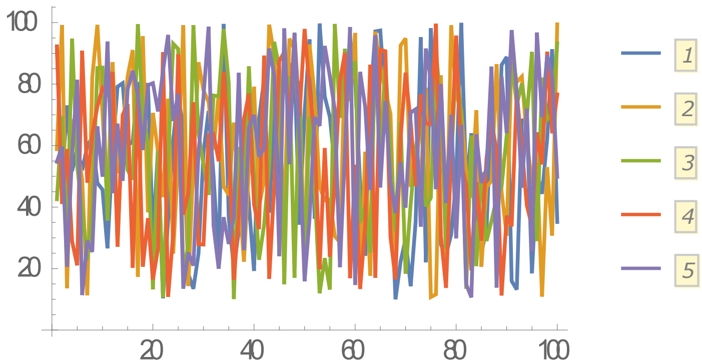

Figure 2 Random panel data graph with 75 time series and 50 cross-section (5 samples)

From the random data above, a simulation is carried out using the parameter estimator formula. In addition, the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) calculations were also carried out using the formula (Leroux and Puterman 1992), namely $AIC: l_m - d_m$ and $BIC: \left(l_m - \left(\log T \times p\right) d_m\right)/2,$ where l_m is log-likelihood maximized with m state, $T \times p$ is sample size of data, d_m is the number of free parameters in the model with m state. The following iteration of the estimated function likelihood according to the data in Figures 1-3 above.

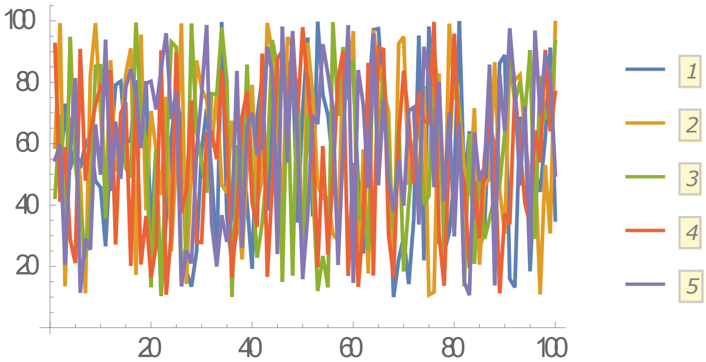

Figure 3 Random panel data graph with 100 time series and 75 cross-section (5 samples)

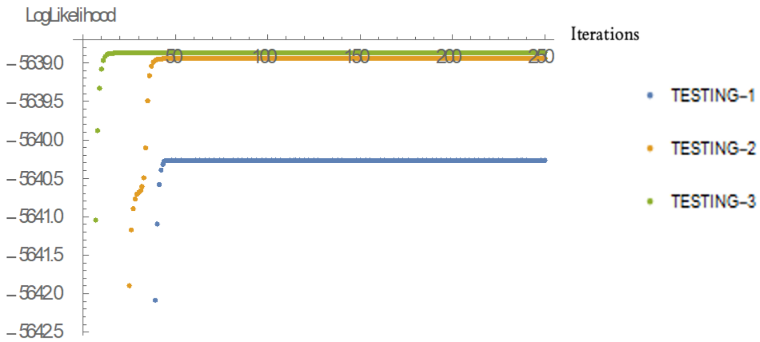

Figure 4 loglikelihood iteration according 50 time series and 25 cross-section (5 samples)

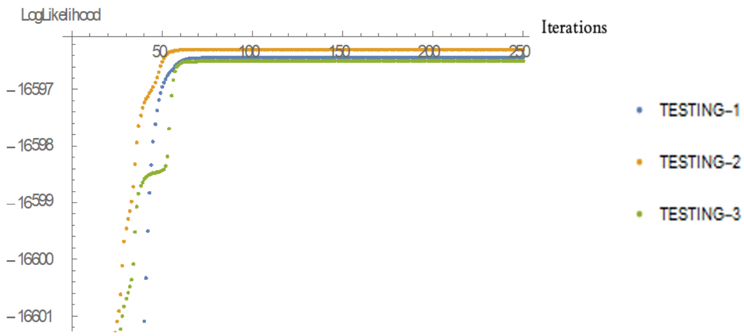

Figure 5 loglikelihood iteration according 75 time series and 50 cross-section (5 samples)

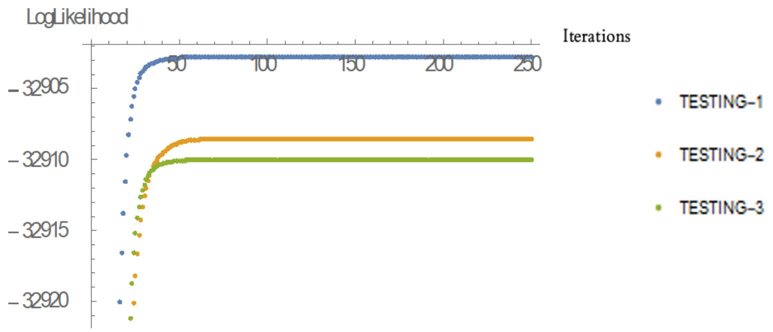

Figure 6 loglikelihood iteration according 100 time series and 75 cross-section (5 samples)

The results obtained are summarized and presented in the following table:

Table 1 MNHM models simulation for number of 50 data time series ($T = 50$)

p	m	Testing	Parameters μ	γ	LogLikelihood start	LogLikelihood final	AIC	BIC
25	2	1	50	2	−6926.24	−5640.26	−5692.26	−5825.66
		2	50	2	−6688.51	−5638.95	−5690.95	−5824.35
		3	50	2	−6754.52	−5638.87	**−5690.87**	**−5824.27**
	3	1	75	6	−6766.18	−5630.49	−5711.49	−5919.29
		2	75	6	−6385.64	−5629.25	−5710.25	−5918.05
		3	75	6	−6457.55	−5625.92	−5706.92	−5914.72
	4	1	100	12	−6755.35	−5619.77	−5731.77	−6019.10
		2	100	12	−6389.26	−5619.68	−5731.68	−6019.01
		3	100	12	−6444.14	**−5617.36**	−5729.36	−6016.69
50	2	1	100	2	−31234.60	−10410.60	−10512.60	−10809.63
		2	100	2	−23069.20	−10408.30	−10510.30	−10807.33
		3	100	2	−27361.70	−10408.20	**−10510.20**	**−10807.23**
	3	1	150	6	−38906.70	−10394.40	−10550.40	−11004.68
		2	150	6	−26127.70	−10391.40	−10547.40	−11001.68
		3	150	6	−28516.00	−10414.60	−10570.60	−11024.88
	4	1	200	12	−20771.20	−10395.90	−10607.90	−11225.25
		2	200	12	−17158.50	−10409.10	−10621.10	−11238.45
		3	200	12	−30187.00	**−10389.90**	−10601.90	−11219.25
75	2	1	150	2	−134712.00	−13339.20	−13491.20	−13964.64
		2	150	2	−121342.00	−13329.90	**−13481.90**	**−13955.34**
		3	150	2	−100817.00	−13338.90	−13490.90	−13964.34
	3	1	225	6	−137256.00	−13334.70	−13565.70	−14285.21
		2	225	6	−78193.00	−13334.10	−13565.10	−14284.61
		3	225	6	−161572.00	−13329.50	−13560.50	−14280.01

Table 1 (Continued)

m	Testing	Parameters μ	γ	LogLikelihood start	LogLikelihood final	AIC	BIC
4	1	300	12	−94940.00	−13332.40	−13644.40	−14616.20
	2	300	12	−122086.00	−13331.00	−13643.00	−14614.80
	3	300	12	−122619.00	**−13329.40**	−13641.40	−14613.20

Table 2 MNHM models simulation for number of 75 data time series ($T = 75$)

p	m	Testing	Parameters μ	γ	LogLikelihood start	LogLikelihood final	AIC	BIC
50	2	1	100	2	−20490.70	−16596.40	−16698.40	−17016.11
		2	100	2	−22454.90	−16596.30	**−16698.30**	**−17016.01**
		3	100	2	−20646.50	−16596.50	−16698.50	−17016.21
	3	1	150	6	−22316.70	−16579.80	−16735.80	−17221.70
		2	150	6	−20144.70	−16573.00	−16729.00	−17214.90
		3	150	6	−21419.90	−16575.40	−16731.40	−17217.30
	4	1	200	12	−20386.10	−16556.60	−16768.60	−17428.93
		2	200	12	−20549.90	−16554.20	−16766.20	−17426.53
		3	200	12	−19885.70	**−16551.20**	−16763.20	−17423.53
75	2	1	150	2	−78644.30	−23464.20	**−23616.20**	**−24120.46**
		2	150	2	−64377.80	−23479.50	−23631.50	−24135.76
		3	150	2	−86579.70	−23468.30	−23620.30	−24124.56
	3	1	225	6	−86050.60	−23443.10	−23674.10	−24440.44
		2	225	6	−66731.90	−23465.90	−23696.90	−24463.24
		3	225	6	−66539.60	**−23442.90**	−23673.90	−24440.24
	4	1	300	12	−66931.70	−23444.80	−23756.80	−24791.86
		2	300	12	−71197.80	−23453.10	−23765.10	−24800.16
		3	300	12	−67117.90	−23463.40	−23775.40	−24810.46
100	2	1	200	2	−284059.00	−28077.60	−28279.60	−28978.79
		2	200	2	−258135.00	−28055.40	−28257.40	−28956.59
		3	200	2	−243800.00	−28048.50	**−28250.50**	**−28949.69**
	3	1	300	6	−318420.00	−28052.90	−28358.90	−29418.07
		2	300	6	−269263.00	−28049.60	−28355.60	−29414.77
		3	300	6	−255696.00	−28048.70	−28354.70	−29413.87
	4	1	400	12	−256261.00	−28043.50	−28455.50	−29881.57
		2	400	12	−231878.00	−28059.20	−28471.20	−29897.27
		3	400	12	−276319.00	**−28034.10**	−28446.10	−29872.17

Table 3 MNHM models simulation for number of 100 data time series $(T = 100)$

p	m	Testing	Parameters μ	γ	LogLikelihood start	LogLikelihood final	AIC	BIC
75	2	1	150	2	–45090.90	–32902.80	–33054.80	–33580.92
		2	150	2	–48680.80	–32908.60	–33060.60	–33586.72
		3	150	2	–45786.10	–32910.00	–33062.00	–33588.12
	3	1	225	6	–41197.00	–32880.20	–33111.20	–33910.77
		2	225	6	–42179.20	–32877.70	–33108.70	–33908.27
		3	225	6	–45186.10	–32889.80	–33120.80	–33920.37
	4	1	300	12	–45478.40	–32857.40	–33169.40	–34249.34
		2	300	12	–39464.10	–32838.40	–33150.40	–34230.34
		3	300	12	–44167.40	–32849.00	–33161.00	–34240.94
100	2	1	200	2	–102057.00	–41766.70	–41968.70	–42696.94
		2	200	2	–151539.00	–41798.10	–42000.10	–42728.34
		3	200	2	–138825.00	–41756.50	–41958.50	–42686.74
	3	1	300	6	–136727.00	–41724.60	–42030.60	–43133.78
		2	300	6	–89290.20	–41753.40	–42059.40	–43162.58
		3	300	6	–119623.00	–41761.00	–42067.00	–43170.18
	4	1	400	12	–126262.00	–41727.80	–42139.80	–43625.13
		2	400	12	–91507.90	–41755.10	–42167.10	–43652.43
		3	400	12	–104235.00	–41752.50	–42164.50	–43649.83
125	2	1	250	2	–338726.00	–48374.70	–48626.70	–49563.32
		2	250	2	–348007.00	–48346.20	–48598.20	–49534.82
		3	250	2	–376975.00	–48351.40	–48603.40	–49540.02
	3	1	375	6	–345004.00	–48338.30	–48719.30	–50135.38
		2	375	6	–284638.00	–48342.20	–48723.20	–50139.28
		3	375	6	–380462.00	–48344.90	–48725.90	–50141.98
	4	1	500	12	–264409.00	–48325.00	–48837.00	–50739.97
		2	500	12	–259946.00	–48333.40	–48845.40	–50748.37
		3	500	12	–338281.00	–48319.40	–48831.40	–50734.37

The results of the MNHM model simulation for all data (attached) whose results are summarized and presented in Tables 1-3 above show that the likelihood function is increasing in each iteration. However, because the likelihood function obtained is a local maximum value, the parameter initialization is important because it affects the parameter estimator obtained, this can be seen from 9 simulations of the final result of the maximum likelihood function at $m = 4$ state is 7 times, $m = 3$ state is 2 times, and $m = 2$ state is 0 times. While the best model of 27 simulations is entirely at $m = 2$ state which can be seen in the AIC and BIC values are always maximum at $m = 2$ state, this is because even though the final maximum likelihood is at $m = 3$ and $m = 4$ states but the penalty due to the increase in independent parameters and the amount of data has a greater effect than the increase in the likelihood function.

5. Conclusions

The multivariate Normal hidden Markov model (MNHMM) which assumed the Markov chain is homogeneous, ergodic and fulfills the assumption of continuity of parameters, then

1. Parameter Estimation of MNHMM using the EM algorithm produces a formula that maximizes the likelihood function,

2. The obtained parameter estimator sequence algorithm converges to the stationary point of the likelihood function monotonically increasing.

Acknowledgements

The research was funded by the Institut Teknologi PLN, and the first author would like to thank Dr. Berlian Setiawaty who has supervised him during his master's studies (10 years ago) regarding the foundations of the hidden Markov models.

References

Barbu VS, Limnios N. Semi-Markov chains and hidden semi-Markov models: Toward applications. Springer; 2009.

Baum LE. An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. In: Proceedings of the Third Symposium on Inequalities. 1972. p. 1-8.

Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat. 1970; 41(1): 164-171.

Cappe O. Inference in hidden Markov models. New York: Springer; 2005.

Cinlar E. Probability and stochastics. New York: Springer; 2011.

Cinlar E. Introduction to stochastic processes. Mineola: Dover Publications; 2013.

Cutajar M, Gatt E, Grech I, Casha O, Micallef J. Comparative study of automatic speech recognition techniques. IET Signal Process. 2013; 7(1): 25-46.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B. 1977; 39(1): 1-22.

Fikri M, Abdul-Malek Z, Esa MRM, Supriyanto E. Recursive parameter estimation and its convergence for multivariate normal hidden Markov inhomogeneous models. Malays J Fundam Appl Sci. 2023;19(5): 840-854.

Fikri M, Christiono MKIG, Abdul-Malek Z, Romadhonid ML, Esa MRM, et al. Comparison of corona discharge identification in 20 kV cubicles based on voltage and noise using ED, HMM, and FCM. J Teknol (Sci Eng). 2024; 86(5): 11-22.

Fikri M, Samsurizal, Christiono MKT. Pemodelan cuaca menggunakan model hidden Markov untuk pemanfaatan energi surya (Weather modeling using hidden Markov models for solar energy utilization). Kilat. 2020; 9(2): 217-224.

Fikri M, Setiawaty B, Purnaba IGP. Pendugaan parameter dan kekonvergenan penduga parameter model Poisson hidden Markov (Parameter Estimation and Convergence of Parameter Estimators of Poisson Hidden Markov Models). J Math Its Appl. 2016; 15(1): 1-10.

Grimmett GR, Stirzaker DR. Probability and random processes. Oxford University Press; 2001.

Gupta A, Dhingra B. Stock market prediction using hidden Markov models. In: 2012 Students Conference on Engineering and Systems (SCES). 2012. p. 1-4.

Huang C, Farewell D, Pan J. A calibration method for non-positive definite covariance matrix in multivariate data analysis. J Multivar Anal. 2017; 157: 45-52.

Khiatani D, Ghose U. Weather forecasting using hidden Markov model. In: 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN). 2017. p. 220-225.

Lam C. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. Ann Stat. 2016; 44(3): 928-953.

Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal. 2004; 88(2): 365-411.

Ledoit O, Wolf M. Analytical nonlinear shrinkage of large-dimensional covariance matrices. Ann Stat. 2020; 48(5): 3043-3065.

Leroux G, Puterman ML. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics. 1992; 48(2): 545-565.

Li J, Pedrycz W, Jamal I. Multivariate time series anomaly detection: A framework of hidden Markov models. Appl Soft Comput. 2017; 60: 229-240.

Luck A, Giehr P, Nordstrom K, Walter J, Wolf V. Hidden Markov modelling reveals neighborhood dependence of Dnmt3a and 3b activity. IEEE/ACM Trans Comput Biol Bioinform. 2019; 16(5): 1598-1609.

Macdonald IL, Zucchini W. Hidden Markov and other models for discrete-valued time series. Chapman & Hall; 1997.

Pourahmadi M. Covariance estimation: The GLM and regularization perspectives. Stat Sci. 2011; 26(3): 369-387.

Mouaz B, Abderrahim BH, Abdelmajid E. Speech recognition of Moroccan dialect using hidden Markov models. Procedia Comput Sci. 2019; 151: 985-991.

Nguyen N. Hidden Markov model for stock trading. Int J Financ Stud. 2018; 6(36): 1-17.

Paroli R, Redaelli G, Spezia L. Poisson hidden Markov models for time series of overdispersed insurance counts. Astin Colloq. 2000: 461-474.

Paroli R, Spezia L. Gaussian hidden Markov models: Parameters estimation and applications to air pollution data. Univ Cattol Sacro Cuore. 1999; 94: 1-24.

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

Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989; 77(2): 257-288.

Ross SM. Introduction to probability models. 12th ed. Cambridge: Academic Press; 2019.

Rothman AJ, Levina E, Zhu J. A new approach to Cholesky-based covariance regularization in high dimensions. Biometrika. 2010; 97(3): 539-550.

Somani P, Talele S, Sawant S. Stock market prediction using hidden Markov model. In: 2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference. 2014. p. 89-92.

Spezia L. Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. J Time Ser Anal. 2010; 31: 1-11.

Spezia L, Futter MN, Brewer MJ. Periodic multivariate normal hidden Markov models for the analysis of water quality time series. Environmetrics. 2011; 22: 304-317.

Tao H, Lu X. Smoke vehicle detection based on multi-feature fusion and hidden Markov model. J Real-Time Image Process. 2019; 17(3): 745-758.

Trichilli Y, Boujelbène Abbes M, Masmoudi A. 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. 2020; 13(2): 165-193.

Wu J. On the convergence properties of the EM algorithm. Ann Stat. 1983;11(1): 95-103.

Young C, Chen C, Wu W. Numerical instability of calculating inverse of spatial covariance matrices. Stat Prob Lett. 2017; 129: 182-188.

Zangwill WI. Nonlinear programming: A unified approach. New Jersey: Prentice Hall; 1969.

Zarrabi N, Schluesche P, Meisterernst M, Börsch M, Lamb DC. Analyzing the dynamics of single TBP-DNA-NC2 complexes using hidden Markov models. Biophys J. 2018; 115(12): 2310-2326.

Zhang M, Jiang X, Fang Z, Zeng Y, Xu K. High-order hidden Markov model for trend prediction in financial time series. Phys A. 2019; 517: 1-12.